TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Osmo Eerola

Prototype Modeling of Vowel Perception and Production in a Quantity Language

Julkaisu 1230 • Publication 1230

Tampere 2014

Tampereen teknillinen yliopisto. Julkaisu 1230
Tampere University of Technology. Publication 1230

Osmo Eerola

Prototype Modeling of Vowel Perception and Production in a Quantity Language

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Rakennustalo Building, Auditorium RG202, at Tampere University of Technology, on the 22nd of August 2014, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2014

ISBN 978-952-15-3333-4 (printed)
ISBN 978-952-15-3343-3 (PDF)
ISSN 1459-2045

Abstract

Vowel prototypes refer to the psychological memory representations of the best exemplars of a vowel category. This thesis examines the role of prototypes in the perception and production of Finnish short and long vowels. A comparison with German as a linguistically different language with a similar vowel system is also made. The thesis reports on a series of four experiments in which prototypes are examined by means of behavioral psychoacoustic measurements and compared with vowel productions in quiet and in noise. In the perception experiments, Finnish and German listeners were asked to identify and evaluate the goodness of synthesized vowels representing either the entire vowel space or selected subareas of the space. In the production experiments, only Finnish speakers were recruited, but earlier reported production data were used for the comparison of Finnish and German. The new concept of the *weighted prototype* (Pω) is introduced in Study I, and its usability in contrast to absolute prototypes (Pa) and category centroids (Pc) is examined in Study IV.

Generally, the results support the finding that vowel categories are not homogenous in quality, but have an internal structure, and that there are significant quality differences between category members in terms of goodness ratings. The results of Studies I, II and III support the *identity group interpretation* of the Finnish quantity opposition by showing that the differences in the perceived quality and in the produced short and long vowels are not demonstrably dependent on the physical duration of the stimuli, although the production experiments in Studies I and III indicated that the short peripheral vowels, especially /u/ in Study III, are more centralized in the vowel space than the long vowels. On the basis of the results of Study II, the spectral and durational local effective vowel indicators of the *initial auditory theory of vowel perception* appear to be independent of each other, thus suggesting that the auditory vowel space (AVS) is orthogonal in terms of the measures used in the experiment. Furthermore, the reaction time results of Study II indicate that stimulus typicality in terms of vowel quantity affects the categorization process of quality but not its end result. The noise masking of production in Study III indicated that both of the noise types applied in the experiment, pink noise and babble noise, resulted in a prolongation of all vowel durations as reported earlier on the Lombard effect. However, the noise masking did not affect the Euclidean distances between the short and long vowels, but caused a minor systematic drift on F1–F2 space in both vowel types. The minor differences suggest that prototypes act as articulatory targets in a fire-and-forget manner without the auditory feedback affecting the immediate articulation.

The results concerning the different prototype measures indicated that the Pa and Pω differ significantly from the Pc, with the Pa being most peripheral. This gives some support to the adaptive dispersion effect in perception. The individual variations of the

measures were normally distributed, with some exceptions for Pa in Finnish, and were, in terms of the coefficient of variation (CV), of the order of difference limen (DL) of frequency. These results suggest that, for normally distributed prototypes, and especially for Pω, which showed the least variation, two thirds of the subjects detected the best category representatives from a subset of stimuli that lie within the limits of DL of frequency from each other in the F1–F2 space. This finding can be regarded as a strong evidence for prototype theories, in other words, the best category representatives play a role by acting as templates in vowel perception. The listeners were able to recognize quality differences between and within vowel categories, but the majority of them ranked the best category exemplars from a subset of stimuli that were hardly distinguishable from each other.

There were some minor differences in the vowel systems of Finnish and German as indicated by the different prototype measures: the absolute prototypes showed the largest differences between the languages in /e/, / ø/ and /u/. This is in line with the earlier investigations on produced vowels in Finnish and German. Generally, the vowel systems of these two linguistically unrelated languages were strikingly similar, especially in the light of the Pω measure.

As presented in this thesis, the prototype approach provides a feasible tool for research and the results lend support to the idea that speech comprehension on the auditory, phonetic, and even on phonological processing levels is based on the memory representations of typical speech sounds of one's native tongue, formed during the early language acquisition phase, and these representations may be similar for the speakers and listeners of two different languages with comparable vowel systems.

Tiivistelmä

Psykologiassa *prototyypillä* tarkoitetaan tietyn käsiteluokan tyypillisintä edustajaa. Ihmisaivoissa prototyypit muodostuvat automaattisesti aistialtistuksen kautta ja tallentuvat pitkäkestoiseen muistiin. Prototyyppiteorioiden mukaan nämä luokkansa tyyppiedustajat toimivat hahmontunnistuksessa vertailukohteina, joihin uusia havaittuja ärsykkeitä verrataan. Puheen havaitsemisessa prototyyppien oletetaan vaikuttavan äänteiden tunnistuksessa ja ohjaavan artikulaatiota puheen tuottamisessa.

Väitöstyö muodostuu neljästä alkuperäisjulkaisusta, joissa tutkitaan puhesyntetisaattorilla tuotettujen suomen kielen vokaalien laatuerojen havaitsemista psykoakustisin kuuntelukokein sekä verrataan koehenkilöiden mitattuja havaintoprototyyppejä heidän tuottamiinsa vokaaleihin. Julkaisussa I esitetään uusi painotetun prototyypin (Pω) käsite, jonka avulla voidaan laskea usean hyväksi arvioidun vokaaliärsykkeen joukosta prototyyppien formantit. Julkaisuissa I ja IV painotettua prototyyppiä verrataan absoluuttisiin prototyyppeihin (Pa) ja kategorioiden keskuksiin (Pc). Julkaisussa IV vertailukielenä käytetään saksaa, jonka vokaalijärjestelmä on samankaltainen kuin suomen. Julkaisussa II tutkitaan suomen vokaalikeston ja -laadun keskinäisvaikutusta vokaaliparilla /y/ ja /i/ neljällä eri kestolla (50 ms, 100 ms, 250 ms ja 500 ms). Julkaisussa III tutkitaan kahden eri kohinatyypin vaikutusta suomen vokaalien tuottoon.

Tulosten perusteella vokaalikategoriat eivät ole homogeenisia, vaan niiden sisällä on tilastollisesti merkitseviä laatueroja eri allofonien välillä. Koehenkilöiden tuottamat vokaalit /i/, /e/, /y/ ja /ø/ olivat F1–F2-formanttiavaruudessa lähempänä kyseisen vokaalin havaintoprototyyppiä kuin muiden vokaalien prototyyppejä, mutta tuotetut vokaalit olivat sentraalisempia kuin prototyypit. Koehenkilöiden välinen tilastollinen hajonta oli pienempi painotetuilla kuin absoluuttisilla prototyypeillä. Vokaalien keston ja laadun keskinäisvaikutusta ei löydetty, vaan koehenkilöt kuulivat pitkät ja lyhyet vokaalit laadullisesti samankaltaisina, vaikka puhuttuina niiden välillä on mitattu pieniä spektraalisia eroja. Tulos tukee suomen kvantiteettiopposition identiteettiryhmätulkintaa. Reaktioaikamittauksin osoitettiin, että vokaalin laadun tunnistamiseen kuluu enemmän aikaa, kun vokaalin kesto on tyypillisen lyhyen tai pitkän vokaalin väliltä (100 ms). Taustahälyn käyttö sai aikaan Lombard-ilmiön, mutta ei muuten vaikuttanut koehenkilöiden tuottamien lyhyiden ja pitkien vokaalien spektrirakenteisiin. Tästä pääteltiin, että havaintoprototyypit toimivat tuottoa ohjaavina malleina hälystä huolimatta. Suomen ja saksan vokaalijärjestelmät osoittautuivat prototyyppien perusteella samankaltaisiksi ja useimmilla vokaaleilla prototyyppien erot kielten välillä eivät olleet kuultavissa. Prototyypit Pa ja Pω erosivat tilastollisesti merkitsevästi Pc:stä ja Pω oli sentraalisempi kuin Pa. Pω:n keskihajonta oli lähellä psykoakustista formanttitaajuuksien erojen havaintokynnystä (DL), minkä perusteella voidaan päätellä, että painotetut prototyypit toimivat äänteiden havaitsemisen referenssinä.

List of original publications

The following original publications form the empirical part of the thesis and are presented below in the order of their appearance:

I. Eerola, O., Savela, J. (2011) Differences in Finnish front vowel production and weighted perceptual prototypes in the F1–F2 space. In: Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China, 2011. Pages 631-634.

II. Eerola, O., Savela, J., Laaksonen, J-P., Aaltonen, O. (2012) The effect of duration on vowel categorization and perceptual prototypes in a quantity language. Journal of Phonetics, Vol. 40 (2). Pages 315-328.

III. Eerola, O., Savela, J. (2012) Production of short and long Finnish vowels with and without noise masking. Linguistica Uralica, Vol. XLVIII (3). Pages 200-208.

IV. Savela, J., Eerola, O., Aaltonen, O. (2014) Weighted vowel prototypes in Finnish and German. Journal of the Acoustical Society of America, Vol. 135 (3). Pages 1530-1540.

Author's contribution to original publications

Publications I and III

The author devised the concept of weighted prototype and planned the research questions of these publications. The stimuli and experimental setup and arrangements were designed and implemented by the author. The measurement methods for data collection and data conversions for statistical analysis were composed by the author. The essential parts of statistical analysis are made by co-author Janne Savela, who also contributed to the general discussion.

Publication II

The author conceived the idea of testing the possible effect of vowel duration on the perception of vowel quality. The stimuli and experimental setup and arrangements were designed and implemented by the author. The measurement methods for data collection and data conversions for statistical analysis were composed by the author. The essential parts of statistical analysis were made by co-author Janne Savela, who also contributed to the general discussion, together with co-author Juha-Pertti Laaksonen. The report continues the line of research initiated in the 1980s by Olli Aaltonen, who also contributed to the general discussion.

Publication IV

The author has participated essentially in the construction of Publication IV, and in the writing of the report. The author had devised the concept of weighted prototype, and suggested that, as a new measure, it should be compared with absolute prototypes and category centroids in a study initiated by Janne Savela for comparing the vowel systems of two different languages, Finnish and German. Further, the author suggested that the perceptual data of the Turku Vowel Test could be compared with the earlier published production data in order to investigate possible interaction between prototypes and production. The need to explore the distribution and normality of prototypes within a category was envisioned by the author. The report continues the line of research initiated in the 1980s by Olli Aaltonen, who also contributed to the general discussion.

Preface and acknowledgments

The prelude of this thesis was played in the late 1970s during my years as an undergraduate student at Tampere University of Technology. In his course on technical information systems, Professor Matti Karjalainen used the first microprocessor controlled Finnish speech synthesizer SYNTE as a case example (Karjalainen 1978). This course evoked my interest in speech and speech technologies. Professor Karjalainen substantially influenced my thinking about biological and technical information systems. He passed away in 2010, and I remember him as the great pioneer of speech technology in Finland. Professor emeritus Jaakko Malmivuo, the supervisor of my Master's and Licentiate Theses, introduced me to scientific thinking. His lectures on the theory of physiological systems, and the many discussions, especially those on pattern recognition and non-linear modeling in connection with my work for the licentiate degree, were particularly useful for my reasoning how complex biological systems should be modeled. Professor Malmivuo encouraged me to carry on scientific research besides my engineering career.

After graduation in 1980, I again encountered speech technologies in 1988 when I was working at Nokia Mobira as the program manager of the company's first GSM phone development project, "Elmo". In the project, I participated in the standardization work in the Groupe Spécial Mobile Working Party 5 of the CEPT (European Conference of Postal and Telecommunications Administration), which worked on the essential patents in GSM technology, among them the speech coding issues. The 13 kbps RPE-LTP coding became the standard, and is still in use as the full speed codec in GSM systems. Although I was not involved in the actual technical details, two things were eye-opening: the huge amount of redundancy in a speech signal, and the efficiency of reducing it by the signal processors of the time (Kuisma et al. 1988). The years at Nokia included many project challenges but also interesting discussions with colleagues about the auspicious possibilities of modern speech technology, which then became reality and went far beyond expectations. My colleagues, especially Timo Ali-Vehmas (Chief Engineering Manager of the Elmo project), Timo Kolehmainen (Speech Codec Development Specialist of the Elmo project) and Erkki Kuisma (GSM Research Manager, Nokia Fellow), taught me a lot of the implementation of speech processing in communication devices.

In 1990-1994, I acted as Senior Laboratory Engineer in the Centre for Cognitive Neuroscience at the University of Turku. During that time and the following years, I got acquainted with many senior and young researchers who manifest themselves in the pedigree of this thesis. I am especially grateful to Olli Aaltonen, Heikki Lang, Altti Salmivalli, Jyrki Tuomainen, Matti Laine, Esa Uusipaikka, Pirjo Korpilahti, Hannu Mikola, Ilkka Raimo, Unto Laine, Ulla Ruotsalainen, Kalevi Wiik, Teija Kujala, Mari Tervaniemi, Risto Näätänen, Riitta Hari, Istvan Winkler, Åke Hellström, Francisco Lacerda, Patricia Kuhl, Terrance Nearey and Adrianus Houtsma.

The experiments for Studies I, II and III were carried out at the Centre for Cognitive Neuroscience. I express my gratitude to Professor Heikki Hämäläinen and Professor Pirjo Korpilahti who, in their capacity as the head of the research centre, kindly gave the opportunity to use the speech laboratory facilities. I also want to thank the personnel of the centre, especially Mia Ek who kindly helped me in the practical arrangements.

The Turku Vowel Test database has been used as the information source in Study IV. The database was compiled at the Department of Phonetics, and is maintained at the Department of Information Technology, University of Turku. I thank the personnel of these departments for providing access to the database.

I extend my gratitude to my co-authors Janne Savela, PhD, Department of Information Technology, University of Turku, and the late Juha-Pertti Laaksonen, PhD, for their overall contribution to the original publications, and Jyrki Tuomainen, PhD, Research Department of Speech, Hearing and Phonetic Sciences, University College of London for his comments to the thesis manuscript.

I wish to thank my principal supervisor Professor Hannu Eskola, Department of Electronics and Communications Engineering, Tampere University of Technology, for not only acting as the instructor but also for the many encouraging and helpful comments on the manuscript.

I am grateful to my supervisor and co-author Professor emeritus Olli Aaltonen, Institute of Behavioural Sciences, University of Helsinki for the long-lasting co-operation, countless inspiring discussions, shared visits to international congresses and research institutes, and for the critical and encouraging comments during the writing process.

I wish to thank the reviewers, Professor Paavo Alku, Department of Signal Processing and Acoustics, Aalto University, Espoo, and Docent Stefan Werner, Department of Linguistics and Language Technology, University of Eastern Finland, Joensuu, for the pre-examination and constructive comments to the manuscript.

The Cultural Foundation of Finland has supported this thesis by a personal grant.

This thesis work has been a long-term undertaking accomplished mainly besides my daily work. I thank my wife Lea for all her support during the endeavor, and for the revision of the English language of the manuscripts. Finally, I thank our sons Risto, Lauri and Jyrki who have prompted me to finalize the work, and who have grown up to independent young gentlemen during the project.

In Kuusisto, Finland, 15 July 2014

Osmo Eerola

List of terms, symbols and abbreviations

Notations:

[i]	Phone (short speech sound of <i> , allophone of i)
[i:]	Phone (long speech sound of <i>, long duration allophone of i)
/i/	Phoneme (short speech segment, category i)
/i:/	Phoneme (long speech segment, category i)
<i>	Orthographic notation (alphabet "i")

ABX	Comparison test where stimulus X is matched either to reference A or B
AC	Air conducted sound transmission
ART	Adaptive resonance theory
ASR	Automatic speech recognition
AVS	Auditory vowel space
AX	Comparison test where stimulus X is compared to reference A (yes/no)
Bark	Unit of the critical band of hearing (a psychoacoustic scale of hearing)
BC	Bone conducted sound transmission
C	Consonant
CB	Critical band; Category boundary
CP	Categorical perception
CV	Consonant-vowel pair; Coefficient of variation (SD/mean)
d	Physical duration
D	Temporal information on d (perceptual representation of physical duration)
dB	Decibel, logarithmic scale unit used, e.g., in indicating sound pressure levels
Diphone	Adjacent pair of phones, transition between two phones
DL	Difference Limen (often ΔF/F, F= frequency), the smallest detectable frequency difference
E1, E2, E3	Local effective vowel indicators, perceptual representations of formants F1, F2 and F3

EEG	Electroencephalography, measurement of the brain's electric activity on scalp
ERB	Equivalent rectangular bandwidth
ERP	Event related potential, an evoked EEG potential
f0	Fundamental frequency of vowels, oscillation frequency of glottal source
F1, F2, F3	The first, second and third formants of vowels
F4, F5, F6	The fourth, fifth and sixth formants of vowels
fMRI	Functional magnetic resonance imaging
FOXP2	Forkhead box protein P2, FOXP2 transcription gene
IPA	International Phonetic Alphabet
ISI	Interstimulus interval, time from stimulus onset to the next stimulus
JND	Just noticeable difference, the smallest detectable difference in a stimulus parameter
LEVI	Local effective vowel indicator (e.g., D, E1-E3)
mel	Psychoacoustic frequency unit
MMN	Mismatch negativity, an ERP component obtained for oddball stimulus
N1, N100	Negative deflection waveform at around 100 ms from stimulus onset, an ERP component
N2, N200	Negative deflection waveform at around 200 ms from stimulus onset, an ERP component
NP	Non-prototype, a poorly rated category member in goodness evaluation test
P3, P300	Positive deflection waveform at around 300 ms from stimulus onset, an ERP component
P600	Positive deflection waveform at around 600 ms from stimulus onset, an ERP component
P	Prototype (generally), the best rated category member in goodness evaluation test
Pa	Absolute prototype, the highest scored category member in goodness evaluation test
Pa_{est}	Estimate of Pa to emphasize that the true Pa is estimated from data
Pc	Category centroid, a prototype measure signifying the arithmetic mean of a category

Pω	Weighted prototype
PDF	Probability density function
PESQ	Perceptual evaluation of speech quality
PME	Perceptual magnet effect, an effect in which the prototypes shrink the perceptual space
PET	Positron emission tomography
Quality	Vowel quality refers to vowel type (e.g. /i/, /y/, /u/); voice quality refers to the type of voice (e.g. modal, breathy, whispery, tense, lax and creaky voice); technical speech quality refers to transmission bandwidth (e.g. "toll" quality, wide band audio) or quality determined by Perceptual Evaluation of Speech Quality (PESQ)
Quantity	Vowel quantity refers to vowel duration (e.g. short and long vowels /i/ and /i:/), in quantity languages phoneme duration is a distinctive feature for phonological opposition
RMS	Root mean square value
SD	Standard deviation
SPL	Sound pressure level
Triphone	Sequence of three phones
TVT	Turku Vowel Test, experimental setup and database of multi-lingual vowel identifications
UG	Universal grammar
V	Vowel

Contents

1. INTRODUCTION 1

1.1. Origin of speech 1

1.2. From code units to meanings 2

1.3. Vowels as code units and objects of research 4

1.4. Historical background of the thesis 7

1.5. Outline of the thesis 8

2. REVIEW OF THE LITERATURE 9

2.1. Cognitive models in speech perception 9
2.1.1. Serial symbolic information processing models 10
2.1.2. Connectionist parallel processing models 11
2.1.3. Challenges of speech perception models 13

2.2. Modeling speech production 14

2.3. Vowels 19
2.3.1. Finnish vowel system 23
2.3.2. German vowel system 25

2.4. Vowel prototypes and perceptual magnet effect 26

2.5. Internal variation of vowel categories 28

2.6. The initial auditory theory of vowel perception 29

2.7. Detection of small frequency differences of pure tones and vowels 30

2.8. Masking the speech production by noise 32

3. AIMS OF THE RESEARCH 35

3.1 Research hypotheses 35

3.2. Specific research questions 36

4. MATERIALS AND METHODS 39

4.1. Subjects 39

4.2. Vowel stimuli and noise masks 40

4.3. Experimental procedures and data analysis 41

5. SUMMARY OF EXPERIMENTS **44**

5.1. Study I: Vowel prototypes and vowel production **44**

5.2. Study II: Effect of vowel duration on categorization and prototypes **48**

5.3. Study III: Vowel production in noise **52**

5.4. Study IV: Perceptual vowel prototypes in Finnish and German **54**

6. GENERAL DISCUSSION **58**

7. CONCLUSIONS **62**

REFERENCES **64**

APPENDICES: ORIGINAL PUBLICATIONS I - IV **81**

1. Introduction

1.1. Origin of speech

Speech is our species-specific, innate and ancient way to communicate. Speech is much older than writing, and proto-speech may be even older than conscious thinking, the prerequisite for the development of symbol function and language (Hackett 1969; Damasio 2000; Aaltonen 2012). When exactly speech emerged in human evolution, is and may remain unresolved due to the lack of direct evidence (Johansson 2005; p. 85). Fossils and other archaeological findings (Lieberman 1987; Lieberman et al. 1992; Lieberman 2000), the spread, development and relations of contemporary languages and speech sounds (Nichols 1998; Perreault & Mathew 2012), the processes of first and secondary language acquisition (Houston & Jusczyk 2003; Kuhl 2004; Garcia-Sierra et al. 2011), auditory neuroethology (Suga 2006), and evolutionary genetics (Fisher et al. 1998; Lai et al. 2001) have deepened our understanding about the origin of speech (Berwick et al. 2013). However, the estimated time span for the emergence of speech is wide, 50 000–300 000 years ago. This is much earlier than the known emergence of transcription of speech to literal signs 4 000–2 500 B.C.E. (Iivonen 2009), but much later than the known genealogy of hominids, starting 2.5 million years ago. Whether the ability to speak appeared as a sudden stage in evolution or over the course of gradual development, is under debate (Ulbaek 1998). The best known proponent of the biological evolution theory is Steven Pinker, who argues that language and grammatical reasoning result from Darwinian evolution (Pinker & Bloom 1992; Pinker 2010), i.e., they are necessary adaptations for survival, similar to the echolocalization in bats (Suga 1988).

Noam Chomsky originally claimed that humans have a genetically programmed *universal grammar* (UG), that is, an innate *language instinct* that makes language acquisition possible, and that the UG appeared through a mutation about 100 000 years ago. It represents the specific brain structures, often called the language module (Fodor 1983) or language faculty, that can resolve the general rules of other humans' speech and utilize recursion in doing so (Hauser et al. 2002; Chomsky 2004). Chomsky developed his theories on natural and formal languages originally in the 1950s, and since then revised them several times because empirical research has shown some of his original ideas inaccurate or false. For example, Chomsky emphasized the importance of recursion as a language feature that cannot be learned without an innate grammar. However, it has been argued that a connectionist network can 'learn' recursion to the degree[1] needed for human language processing (Johansson 2006).

[1] According to Karlsson, the maximal degree of center-embedding in written language is three, and in spoken language it is practically absent (Karlsson 2007).

The opponents of the sudden appearance theory maintain that speech, as an extremely complicated phenomenon, has evolved and gradually developed during the history of thousands of generations to the current variety of about 7 000 languages (Falk 2004; Masataka 2007; Botha 2008). Scholars in this camp emphasize the role of social interaction (Knight & Power 2012) and slow cultural evolution in language development (Tomasello 1996), and, in principle, do not presume a human language instinct or organ, but rather regard the connectionist plasticity of brain as the enabler of speech and language. They view speech as a strongly social phenomenon, e.g., speech ability does not develop for children who have grown up in entire isolation from other human beings, and a single deaf child does not develop a sign language but only a few distinct beckons. History knows a few documented feral child cases, with Victor of Aveyron, found in a forest in France in 1798, and Genie, found in Los Angeles in 1970 after about 13 years of abuse and isolation, being perhaps the most famous. Victor's story was made widely known by a movie in 1970 (*L'Enfant sauvage* (The Wild Boy) 1970). The rehabilitation of Genie was studied and documented by Curtiss (Curtiss 1977).

Recent research on a genetic mutation in the FOXP2 transcription factor (Fisher et al. 1998; Lai et al. 2001) has shed light on the biological basis of speech. The mutation in FOXP2 resulted in changes in the proteins contributing to brain plasticity and the development of speech organs that strongly favored the evolution of speech ability. This mutation has been timed to have taken place around 100 000 years ago in the human hominid lineage. Interestingly, this finding from genetics research can be interpreted to support both the sudden appearance and continuous development theories of language evolution: the gene mutation improved the basic apparatus needed for speech communication which facilitated the fast evolution of language since spoken language communication formed a strong competitive advantage in the battle of existence.

1.2. From code units to meanings

In the communication system theory, when resolving information of an encoded and modulated message, a demodulator is first needed to separate the code chain from the carrier, then a code book to segregate or decode the elementary units, and finally, a lexicon for mapping the permitted combinations of elementary units to a meaningful message (Carlson 1986; p. 559). This *information processing view* has influenced many theories of human speech communication (Klatt 1979; McClelland & Elman 1986) arguing that, on the abstract level, the same phases are applicable in the message encoding and decoding of speech: Glottal excitation or airflow noise serve as the carrier, which is modulated by the articulators under the control of cerebral commands that are mediated by facial and tongue nerves. The resulting acoustic pressure fluctuations bear all the

information elements[2] that are needed to resolve the message (Pfeifer & Shoup 1976), provided that the receiver has the apparatus to do it properly and is tuned to do it. What then belongs to this decoding apparatus? Ear and peripheral hearing, evidently, but do we need a special cortical *speech module*, or a *language organ* that decodes the abstract code units and combines them into words and meanings, or do we perhaps decode the *articulatory gestures*[3], or the neural commands behind them, as suggested by the *motor theory of speech perception* (Liberman 1985; Galantucci et al. 2006)?

The *double articulation* of speech[4] makes it possible to construct an endless variation of meaningful words and expressions of thoughts based on a finite number of *phonemes*, the elementary code units of speech (Martinet 1984; Studdert-Kennedy 2005). Therefore, a long tradition of research has concentrated on exploring speech communication through understanding how phonemes are encoded and decoded in spoken language. Human brain does this automatically and often seemingly effortlessly; we do not focus on distinct phonemes or words, but the meanings. From the perspective of the communication system theory, the only changes in the code level take place in the articulators, where the neural commands are converted to muscle movements controlling the air flow, and in the cochlea, where the acoustic pressure variation is converted to variation in neural discharge rates, and the subsequent decoding of meaning is then processed by a myriad of interconnected neurons. How this automatic production, recognition and combining of meaningless sound elements into meaningful thoughts happens, has been and continues to be under extensive multi-disciplinary research, and also forms the general framework for this thesis.

In terms of linguistics, it is assumed that the first human speech sounds were short vocal exclamations and mono syllabic consonant-vowel (CV) pairs that were used to warn the fellow men, to signal in hunting, or to express sexual appeal (Johansson 2006). In contemporary languages, the overall range of speech sounds is extensive, approximately 600 consonants and 200 vowels, but each language typically uses only 20–50 of these elementary code units, with wide variability across languages: in Rotokas and Pirahã languages, for example, there are only 11 phonemes, whereas the number of phonemes is 141 in !Xũ language (De Boer 2000; Ladefoged & Disner 2012). *Phoneme* is defined as the smallest contrastive linguistic unit that causes a change in the meaning of a word, for example, the minimum pair words /big/ and /pig/ in English. In articulatory phonetics, phonemes are classified into two main groups: *vowels* are voiced sounds which are produced when the open vocal tract modifies the air flow generated by glottal excitation,

[2] In a real conversation, the message is largely perceived via other channels than hearing (Mehrabian 1968, 1969; Pease 1991): up to 55–65% of the information in a face-to-face conversation is non-verbal (body language), about 35% vocal (tone of voice, prosody), and only about 5% purely verbal (words).

[3] Articulatory gesture refers to the (muscular) act of forming and releasing a constriction of variable location and degree in the vocal tract (Studdert-Kennedy 2005).

[4] Also known as *duality of patterning*.

and *consonants* are sounds that are produced by partially or entirely constricting the air flow through the vocal tract. *Phonetics* studies the production, acoustic properties, and perception of vowels and consonants and their combinations as speech sounds (Ohala 1990). *Phonology* is more concerned on phonemes as abstract linguistic sound systems, and generally regards a single phoneme category as the basic speech unit, whereas phonetics research is also interested in the finer phonemic and supra-segmental cues (Hawkins 2003), intra-categorical variation (Hillenbrand et al. 1990), and the categorical implementation of the different sound variants (*allophones*) (Miller 1997). These questions of categorical variation are in the focus of this thesis.

1.3. Vowels as code units and objects of research

In contemporary languages, simple speech sounds can exceptionally be interpreted both as phonemes and words as is the case, for example, for the Swedish vowels <ö> [œ:] and <å> [o:], which are also nouns meaning 'island' (*en ö*) and 'river' (*en å*) (Fant 1983). The comprehension of normal speech, however, does not focus on the phoneme level, but rather, on the combinations of phones forming longer phonemic segments (*diphones*, *triphones*[5]), syllables, or words (Nearey 1990; Nearey 1992; Goldinger et al. 2003; Hawkins 2003; Port 2007). Research on coarticulated vowels has shown that vowels in CVC context are more accurately identified than the same vowels in isolation, and that important acoustic information for vowel identification also resides in the changing spectral structures of the entire CVC segment, not only in the steady-state part of the central vowel V (Rakerd 1984; Miller 1989; Nearey 1989; Strange 1989). Furthermore, it is known that the identification of vowels in a word may be influenced by the sentence context (Ladefoged & Broadbent 1957; Ganong 1980).

Why then all the effort to examine the perception or production of vowels in isolation, or the distinct vowel classes and their variation within a language or between languages? One obvious answer derives from the history of linguistics. One of the great innovations of mankind is the phonetic transcription of spoken speech sounds as orthographic symbols: ancient Sumerians observed that speech consists of recurring sound units, and for the purposes of describing and transliterating speech, it is enough to mark each sound with a specific symbol, instead of giving symbols for words or syllables (Iivonen 2009; Kemp 2006). This finding reduced the number of necessary symbols from thousands to tens, resulting in the alphabet letters, and suggested that the understanding of the nature of these sound elements is important for the understanding of speech comprehension. However, the finding has indirectly biased the later speech research (Port 2007). The

[5] Diphone = transition between two phones; triphone = a sequence of three phones.

intra-individual and inter-individual variation of spoken phonemes is large, and also the experimentally defined borders of phoneme categories are far from clear-cut, rather, they are fuzzy, partially overlapping, and even unstable over the course of time (Miller 1989). Nevertheless, in orthography, the richness of speech sounds is forced into a limited set of certain alphabetical symbols, and in phonetic transcription, into special phonetic symbol categories (IPA 2005; Vaissière 2011). In the early cognitive modeling of speech, this led to the thinking that speech comprehension is basically comprised of information processing of symbols in a similar manner as computers do pattern recognition, e.g., by implementing the Chomsky hierarchy of formal grammars by automata (Gonzalez & Thomason 1978; p. 29, 96). The validity of this approach will be reviewed in Chapter 2.1.

Another motivation for phoneme studies originates from the discovery of and subsequent research on categorical perception (CP) (Liberman et al. 1957; Burns & Ward 1978; Liberman 1985; Rozsypal et al. 1985; Pastore 1987; Repp & Liberman 1987; Decoene 1993; Schouten et al. 2003; Kurtz 2007). Categories are regarded important for human reasoning and parsing of the world. In CP, the varying sensory information is analyzed and classified into distinct classes or categories, which share similar sensory or conceptual elements and features. According to this view, the pattern recognition is based on stored categories, and the formation of the memorized categories is based on past perception. Within linguistics and phonetics, CP is an interesting object for research since we can create stimulus continua where minor and gradual changes that are psychoacoustically equal (e.g., 30 mels in formant frequency) cause an abrupt change in the interpretation of the percepts between two adjacent categories. Perception is categorical if discrimination peaks in the midrange of a continuum and can be predicated from identifications.

Categories form the basis for linguistic quantity and quality[6] oppositions. The following minimal series of Finnish words demonstrate how a phonemic change in segmental length changes the categorical quantity, and consequently, the meaning of the word: tule-tuule-tulle-tuulle-tuullee-tuulee-tulee-tullee[7] (Karlsson 1983). Similar minimal series for changes in vowel quality are, for example, muuli-mooli-maali or tiili-tyyli-tuuli.[8]

---

[6] The meaning of the term *quality* depends on the context: Vowel quality refers to vowel type (e.g., /i/, /y/, /u/); voice quality refers to the type of voice (e.g., modal, breathy, whispery, tense, lax and creaky voice), technical speech quality refers to transmission bandwidth (e.g., telephone "toll" quality, wideband audio) or quality determined by Perceptual Evaluation of Speech Quality, PESQ.

[7] Word meanings: tule (‘come!’) - tuule (‘blow!’) - ei tulle (‘it may not come’) - ei tuulle (‘it may not blow’) - tuullee (‘it may blow’) - tuulee (‘it blows’) - tulee (‘it comes’) - tullee, (‘it may come’); phonetically with IPA symbols: [tule] - [tuːle] - [tulːe] - [tuːlːe] - [tuːlːeː] - [tuːleː] - [tuleː] - [tulːeː].

[8] Word meanings: muuli (‘mule’) - mooli (‘mole’) - maali (‘paint’); phonetically with IPA symbols: [muːli]-[moːli]-[maːli]; tiili (‘brick’) - tyyli (‘style’) - tuuli (‘wind’); phonetical with IPA symbols: [tiːli]-[tyːli]-[tuːli].

Recent findings indicate that neuron populations in human posterior superior temporal gyrus (pSTG), a part of secondary auditory cortex, respond categorically to linguistically continuous (/ba/ to /da/ to /ga/) stimuli (Chang et al. 2010; Scott & Evans 2010). This suggests that our neural system may be intrinsically tuned to CP. However, although CP is demonstrated in laboratory experiments, it has also shown to be a phenomenon that depends on the experimental setup (Burns & Ward 1978; Schouten et al. 2003). Thus, it is not necessarily the mechanism behind phoneme identification in continuous speech.

Several slightly different models of vowel systems and their perception have been presented during the past 50 years. Liberman and Whalen (Liberman & Whalen 2000) divide the many theories into two main classes, horizontal and vertical. In horizontal theories, speech consists of sound units (phonemes) which are the primary objects of perception and targets of articulation, and it is the mental representations of sounds that are processed. In vertical theories, the primary object is articulation, also in perception; the articulatory gestures needed to produce the sounds are the percepts (i.e., the motor theory). In the *double-weak theory* by Nearey, the objects of speech perception and production are neither primarily auditory nor gestural; to understand speech, both articulation and acoustics are important (Nearey 1992). In the *Dispersion theory*, originally postulated by Liljencrants and Lindblom (Liljencrants & Lindblom 1972), and developed further later on (Lindblom 1986; Lindblom 1992; Schwartz et al. 1997), sufficient perceptual contrasts are emphasized in the forming of categories in the vowel space, hence the name perceptual dispersion. Dispersion theories are examined and discussed in Study IV. In the *Quantal Theory,* originally postulated by Kenneth Stevens in 1972 (Stevens & Keyser 2010), quantal regions in the articulatory-acoustic space are those areas where different articulations produce similar acoustic patterns. It proposes that regions of insensitivity of acoustic attributes to changes in articulation could provide a quantitative basis for defining distinctive features (Pisoni 1980).

In Strange's classification of theories (Strange 1989), the *Simple target model* refers to vowels as canonical targets, which are best represented by the static vocal tract shapes in articulation, and acoustically as points in the formant space. *Elaborated target model* accounts for talker normalization, and uses the Bark scale and the formant differences F1–F0, F2–F1 and F3–F2. One challenge in using monophthongal phonemes is the selection of the time range for calculating the formants (typically this is done from the steady-state part). *Dynamic specification model* prefers the use of co-articulated vowels, since vowels are better identified in the CVC context than in isolation, in other words, "the acoustic information for vowel identification resides in the changing spectral structure" (Strange 1989). *Prototype theories*, the *perceptual magnet effect*, and the *initial auditory theory of vowel perception* are in the focus of this thesis, and they are discussed in more detail in Chapter 2.

1.4. Historical background of the thesis

This thesis work continues the long tradition of speech and vowel research carried out in the Department of Phonetics (Wiik 1965; Aaltonen & Suonpää 1983; Aaltonen 1985; Peltola 2003), in the Centre for Cognitive Neuroscience (Korpilahti 1996; Aaltonen 1997; Krause et al. 1998), and in the Department of Information Technology at the University of Turku (Savela et al. 2002; Eerola et al. 2003; Laaksonen 2006; Savela 2009; Saarni 2010; Ojala 2011). Of the earlier research in the Department of Phonetics, the following two publications are particularly relevant to this thesis work: In 1965, Kalevi Wiik presented, in his doctoral thesis entitled *Finnish and English vowels*, the F1–F3 formant variation areas of produced English and Finnish single and double primary-stressed monophthongs (vowels), which were measured, for the first time, on a new device called spectrum analyzer (Kay Electric Sonagraph) (Wiik 1965). Some of Wiik's results are discussed in Studies II, III and IV. Two decades later, in 1983, Olli Aaltonen and Jouko Suonpää published a computerized model for Finnish vowel identifications (Aaltonen & Suonpää 1983). Based on the identifications of 511 different synthesized Finnish vowels by 32 adult listeners, the two-dimensional map illustrates the vowel distribution in the F1–F2 formant space, and it served as an important guideline for the synthesis parameters used in Studies I and II.

In the late 1980s, the vowel research at the Department of Phonetics became more multidisciplinary along with the initiation of psychophysiological vowel perception studies in co-operation with the Department of Clinical Neurophysiology (University Hospital of Turku). Vowel perception was explored by means of event-related brain potentials (ERP), that were measured by an EEG apparatus synchronized with the stimulus delivery devices (Aaltonen et al. 1987; Sams et al. 1990; Aaltonen et al. 1993). In 1990, a research project funded by the Academy of Finland, *Puheen prosessointi ihmisaivoissa* (Speech Processing in the Human Brain), was initiated, in which the author of this thesis served as a research and laboratory engineer until 1994. The project led to the foundation of the Centre for Hearing, Cognition, and Communication Research in 1991, later known as the Centre for Cognitive Neuroscience. The task of the author of this thesis was to design and establish the laboratory and to develop methods and instrumentation for the research (Eerola 1993). Contributions to publications during those years formed the basis and motivation for the later continuation of the research as reported in this thesis (Aaltonen et al. 1994; Aaltonen 1997; Lang et al. 1995).

1.5. Outline of the thesis

The research focuses on the prototype modeling of the perception and production of monophthongal vowels in the Finnish language. For comparing the different vowel prototype classes (absolute, weighted, and category centroid) between two languages that are linguistically unrelated languages but have a similar vowel system, the Finnish and German[9] languages were used.

The overall structure of this thesis is the following:

Chapter 2, Review of the Literature, first provides an overview of those theories and models related to speech perception (2.1) and production (2.2) that explain why vowels as vocalized phonemes have been and continue being important objects of research. From Chapter 2.3 onwards, the acoustic and articulatory properties and descriptors of vowels are discussed, with a review of the Finnish vowel system and a brief overview of the German vowel system. Theories on vowel perception relevant to this study are presented, together with discussion on earlier research results, in Chapters 2.4, 2.5, and 2.6. Finally, the detection of minor frequency differences in vowel formants and the effect of noise on production are described in Chapters 2.6 and 2.7.

Chapter 3, Aims of the Research, gives an overview of the assumptions behind the experiments, and summarizes the hypotheses and aims of the four experiments included in this thesis (Studies I–IV).

Chapter 4, Materials and Methods, provides a summary of informants, stimuli, procedures, and data analysis methods uses in Studies I–IV. The formula for the weighted prototype is introduced and explained in Chapter 4.4.

Chapter 5, Summary of Experiments, reviews and discusses the main results obtained in Studies I–IV.

Chapter 6 includes the General Discussion, and Conclusions are drawn in **Chapter 7**.

[9] From the 22 languages available in the TVT database German is the only that fulfils these criteria and has a sufficient number of subjects for statistical comparison.

2. Review of the literature

2.1. Cognitive models in speech perception

Cognitive processes refer to human information processing, in other words, the implementation of mental functions, such as perception, memory, learning, believing, deciding or willing, in the brain. This machinery, consisting of biological neural networks, differs from the devices for digital computing, but at a higher abstraction level, the same principles of information theory and processing are applicable to both systems (Lindsay & Norman 1977; p. 594). The basic question of how information, speech and language are represented and stored in the brain still waits for a conclusive answer even though some partial answers and numerous theories exist (Churchland 2004; Port 2007; Cutler 2008).

The functional significance of abstract representations for cognitive processing in general, and for speech processing in particular, is emphasized by Cutler (2008): without abstraction, the instant adaptation to different talkers and talking styles would not be possible. Therefore, human listeners must have different representations of incoming speech at the pre-lexical (phonemic) level and at the lexical level, in which the word form and word meaning dissociate. This is demonstrated by the learning of a new (foreign) word: we need to construct the acoustic-phonemic form, the phonological form, and the meaning representation for the word (Cutler 2008). A new word, once heard, can be repeated on the basis of the acoustic-phonemic representation in episodic memory, but transliterating it into a textual entity presupposes that a conversion from acoustic-phonemic representation to phonological representation is available. Still, this can be accomplished without combining the phonological representation to the word meaning.

Cognitive processes can be viewed either as serial or parallel-connectionist. In the computational serial approach, at least in its most straightforward form, cognitive processing of speech equals parsing, i.e., symbol manipulation in a serial manner by the internal language of thought and by an innate universal syntax (Gonzalez & Thomason 1978; p. 133; Chomsky 2004). In the parallel-connectionist view, cognitive processing is based on the parallel working of interconnected neurons, and mental phenomena are regarded more as states of distributed neural networks than as results of symbol manipulations (McClelland & Rumelhart 1986; Pinker & Prince 1988). The serial symbolic processing and connectionist views are not necessarily exclusive, but they approach the modeling problem from different viewpoints; the symbol manipulation is an abstract-level description of the message decoding in speech comprehension, whereas connectionist models take one step further by describing the possible neural implementation of the message decoding in terms of interconnected processing nodes (Kohonen 1978; Tank 1989; Levine 1991; Kohonen 2001; Scott 2003; Hickok & Poeppel 2007).

2.1.1. Serial symbolic information processing models

The development of the theoretical basis of information processing and formal languages, and the triumph of the early von Neumann computers in the 1950s and 1960s strongly affected the thinking of the model builders of the era (Lindsay & Norman 1977; Von Neumann 2012). According to these models, speech perception is hierarchically organized and the acoustic pressure waves are successively transformed, reduced and stored in a serial manner to more abstract forms of representation (Masterton 1992; pp. 181-199). According to Forster's *Autonomous Search Model theory*, linguistic processing is strictly serial and autonomous from bottom up (Forster 1981). In this data-driven system, there are three linguistic processors: a *lexical* processor, a *syntactic* processor and a *message* processor, which are supervised by the general processing system. In this model, the peripheral perceptual system feeds information to the lexical processor, which matches the visual data to an orthographic file, the auditory data to a phonetic file and both types of data to a syntactic-semantic file. When a match is found in the peripheral files, a pointer to the master lexicon is formed. Subsequently, the items found in the master lexicon are passed on to the syntactic processor in order to form a syntactic structure, and this structure further to the message processor for resolving the information contents of the entire sentence.

The model suggested by Studdert-Kennedy (1993) includes the *auditory*, *phonetic, phonological*, *lexical, syntactic* and *semantic* levels of processing. Pisoni and Luce have added the level of *peripheral auditory analysis* to the model (Pisoni & Luce 1987; p. 27). In this model, the speech perception process is presented as a series of hierarchical and successive processing levels. The acoustic pressure wave is, at first, converted to a neural spectrogram in the cochlea. A phonetic feature matrix is a table of acoustic segments and their features which are mapped to phonemes by the phonological processor. The string of perceived phonemes is used for word recognition through a lexical search, and finally, the syntax and the message of the speech utterance are analyzed by higher level processors. According to *modularism* (Fodor 1983), the acoustic processing of speech and non-speech sounds is similar, but from the phonetic level onwards, speech is processed by a special speech module, which is innate in humans. The speech module is activated when the peripheral auditory level has received enough information to distinguish speech elements from the incoming sound pattern. The Haskins school of thought has promoted this difference in processing speech and non-speech sounds (Galantucci et al. 2006), and several researchers have shown evidence that speech is processed differently from other complex sounds starting from the primary auditory cortex (Whalen & Liberman 1987; Belin et al. 2000; Benson et al. 2001; Whalen et al. 2006).

The *Lexical Access From Spectra (LAFS)* model by Klatt (Klatt 1979) represents a bottom-up process, which is based on the spectral memory representations of all 'allowed' sequences of diphones in the long-term memory. Word recognition is performed by direct mapping of the short-term spectra of the incoming sound cues to neural

spectrogram templates. For ambiguous or new words, there is a separate branch for solving and decomposing phonetic structures that do not immediately match in the spectral mapping. The detailed structure of the LAFS model forms a good foundation for experimental testing (Nearey 1990).

Since speech communication is serial in its nature (Milner 1999), the serial modeling approach makes sense to a certain degree, even though the existence of the modular 'processors' in the brain is debatable. It can be tracked to the finding of the Broca's and Wernicke's areas, the damages of which contribute in certain language disorders such as aphasia (Damasio 2000; p. 106). Certain evidence in favor of this approach exists, e.g., the auditory pathway from cochlea to the auditory cortex is evidently responsible for the auditory processing, such as the sound spectral analysis shown in tonotopy (Delgutte & Kiang 1984; Shamma 1985; Deng & Geisler 1987; Geisler 1988; Sanes & Rubel 1988; Masterton 1992; Suga et al. 2003; Suga 2006), and there are reported results that phonetic (Aaltonen et al. 1987; Sams et al. 1990; Aaltonen et al. 1994; Aaltonen et al. 1997) and even phonological decoding takes place within the auditory cortex and in the brain areas closely associated thereto (Näätänen et al. 1997; Winkler et al. 1999; Cheour et al. 2002; Savela et al. 2002; Jacquemot et al. 2003; Näätänen et al. 2004; Uppenkamp et al. 2006; Ylinen et al. 2006; Näätänen et al. 2007; Obleser & Eisner 2008).

2.1.2. Connectionist parallel processing models

The contemporary connectionist speech perception models include more parallel processing and interaction between the hierarchical functional levels (Plaut & Kello 1999; Plaut 2003; Guenther 2006). Most of the models consist of the *bottom-up* pathway, which decodes the embedded message by decreasing redundancy from the input, and the *top-down* pathway, which controls the decoding process by *contextual information* (Cutler 2008). This kind of two-way approach was represented already in the *analysis-by-synthesis* model suggested by Stevens and Halle, in which hypotheses concerning the form of the message are constructed parallelly with the analysis of the incoming data (Halle & Stevens 1962). The system thus tries to predict the result on the basis of fuzzy data, and the process will continue until exact match is achieved. The *Motor Theory of Speech Perception* (Liberman 1985; Liberman et al. 1967; Galantucci et al. 2006) is an extreme case of the analysis-by-synthesis theories. In this theory, in its newest form, the listener attempts to form, from the incoming acoustic waveform, the same neural representations that are needed to produce the spoken utterance (Liberman & Whalen 2000).

According to the *cohort theory* by Marslen-Wilson, linguistic stimuli are identified as soon as there is sufficient acoustic information (Marslen-Wilson & Welsh 1978; Marslen-Wilson 1987). In most cases, a particular word is recognized even before the entire word is spoken. The word recognition process is partially *autonomous* and partially *interactive.* In the recognition process, the acoustic-phonetic information is first used to form a 'key

address' to all those words in memory that share the same word-initial information. This set of words forms a 'cohort' as a result of a bottom-up autonomous process. The next phase is to select the proper word from the cohort. This process is interactive in the manner that contextual information can be used in reducing the possible alternatives until the address in the lexicon is clear. The emphasis the theory gives to word-initials is also one of its shortcomings: the theory does not specify the *error recovery* procedure in case of mispronunciation or misperception. Some experimental evidence for the cohort theory is offered by reaction time (Ferreira et al. 1996) and psychophysiological ERP measurements (O'Rourke & Holcomb 2002). The cohort theory and a similar but earlier theory known as the *logogen theory* (Morton 1969) apply the ideas of addressing memory by its contents (Kohonen 1978), and have their power in describing how an interactive system may work, but they leave open how the acoustic-phonetic and contextual information actually are processed.

The TRACE model by Elman and McClelland is a connectionist model based on interconnected nodes in the input layer and in three processing layers[10]. Nodes are elementary processing units that have threshold and resting levels, and stand for different features, phonemes and words (McClelland & Elman 1986). The TRACE model is highly interactive as it includes feed-forward, lateral and top-down feedback connectivity between and within the layers. The occurrence of a particular acoustic feature may excite the corresponding node towards its threshold. The activation of this node further excites nodes above and below the level of its own and inhibits the nodes at the same level. This facilitates the exclusion of competing nodes at the same level. Although TRACE is basically a computational parsing model implemented by artificial neural networks, it is difficult to avoid the analogue between the nodes and the neurons, except that a single neuron is most likely too simple an element to perform the tasks the node model suggests for the nodes. At the higher levels, the 'word nodes' might be neuron populations or neural networks. The node model explicitly assumes speech segments, but in the actual processing, the segmentation into phonemes is accomplished 'naturally' as a part of the node activation process itself, and the results are observed at the word level. The model overcomes the problems of contextual effects, e.g., the Ganong effect[11] (Ganong 1980), in a similar manner as the problem of segmentation: the coarticulation activates the relevant nodes and consequently excites and inhibits the whole set of nodes for converging towards the best matching.

The *Adaptive Resonance Theory*[12] (ART) by Grossberg (2003) explains the coding strategy of interleaving acoustic cues by stating that the speech units are emergent properties of perceptual dynamics, and the percepts of any grain size emerge as a result of resonant

[10] For the processing layers of neural networks, see Levine (1991), pp.196-260.
[11] Ganong effect means the perception of an ambiguous phoneme in accordance with the lexical context.
[12] For an overview of ART, see Freeman & Skapura (1992), pp.291-340.

brain states when the bottom-up acoustic stream and top-down information resonate (Goldinger & Azuma 2003; Grossberg 2003): "Processing in ART begins when featural input activates items (feature clusters) in working memory. Items, in turn, activate list chunks in memory. These are products of prior learning (perhaps prototypes) that may correspond to any combination of features. Possible chunks therefore include phonemes, syllables, and words." In this model, the memory representations (templates, prototypes) play a crucial role; they are formed through self-organizing on the basis of earlier exposure to speech. Perceptual inference and learning are associated in the sense that the pre-requisite of perception is sensation *per se* combined with the experience on earlier sensations. ART thus provides a possible general framework for the prototype approach of experiments.

Another approach would be that the primary objects of perception are not the phonemes and syllables, but the entire meaning that would be guessed in a stochastic (Bayesian) top-down manner by making continuous estimates from the very beginning of the speech input until a satisfactory result is achieved. The *Shortlist B* model is based on this approach and it utilizes the Bayesian reasoning for continuous speech perception (Norris & McQueen 2008). The Bayesian predictive coding represents the analysis-by-synthesis approach: the minimizing of prediction error adjusts the generative model until the most likely explanation for the sensory input is achieved. Karl Friston explains the cortical EEG responses (e.g., MMN and P300, which reflect responses to abruptly changing sensory information) as a reflection of the brain's prediction errors, or more generally, as momentary fails in minimizing the brain's free energy (Friston 2005).

The neural representation of the lexicon is not known (Port 2007) but the modern imaging techniques (fMRI, PET, MEG) have revealed where in the brain the pre-lexical and lexical processing takes place (Pulvermüller 2005; Obleser & Eisner 2008). The results support the connectionist, parallel, and simultaneous bottom-up and top-down processing approach rather than the strict serial, feed-forward and hierarchical information processing.

2.1.3. Challenges of speech perception models

The perceptual models of speech recognition have to resolve the challenges of *linearity, invariance* and *segmentation* of the acoustic speech signal (Klatt 1979; Pisoni & Luce 1987). These issues of speech perception are often related to the speech production: speech cannot be reduced to the production of successive but distinct sound segments since the articulation of these segments is affected by the articulation of the neighboring segments via *co-articulation*. The differently articulated variants of the phonemes are called allophones. The *linearity condition* refers to the relation between the acoustic and *phonological* representations; if phonemes /l/ and /y/ occur in this order (/ly/) in the phonemic representation, then the corresponding acoustic cues must occur in the same order in the acoustic signal. In practice, this assumption is not entirely valid since the

acoustic cues of phonemes smear in the acoustic waveform. According to the *principle of invariance,* there is a constant set of acoustic features for each phoneme, which do not depend on the acoustic environment of the phoneme occurrence. This feature set is present always when the phoneme is present and absent when some other phoneme is present. In spite of extensive research, such invariant features have not been found, although their existence would explain the perceptual constancy for speech sounds. Apparently, there must be a decoding strategy that permits the listener to cope with acoustic cues, which depend on the phonetic context. Because of the dissatisfaction of the linearity and invariance conditions, the *segmentation of the audiogram* into elements corresponding the phonetic or higher order linguistic units is difficult and in many cases impossible (Pfeifer & Shoup 1976). Even the word boundaries are difficult to detect in the flow of speech by using simple physical criteria. Typically, the segmentation based on acoustic criteria yields more acoustic units than there are phonemes in the utterance. *Time normalization* is needed for distinguishing the irrelevant durational variants of speech segments from the relevant one; for example, in Finnish, the plain spectral representation of a segment cannot be used for distinguishing a long vowel from a short vowel in words [tiili] ('*brick*') and [tili] ('*account*'). The segmental duration can vary due to increased speaking rate, syllable stress or other similar reasons (Suomi 2006). *Talker normalization* is needed since the speech producing organs of individual speakers differ, for example, for the length and shape of the vocal tract. Vocal tract normalization has also been found to be useful in ASR applications (Molau et al. 2000). Thus, the speech decoding process has to be able to normalize the influence of a speaker's gender, age, physical size and *prosody*[13] on the message (Monahan & Idsardi 2010). The prototype based theories intend to find an answer to some of these challenges (Rosner & Pickering 1994).

2.2. Modeling speech production

Speech perception and production are closely related counterparts that should be studied in parallel for the better understanding of spoken language processing (Blumstein & Stevens 1979; Nearey 1992; Rosner & Pickering 1994; Eerola et al. 2002; Ru et al. 2003; Jacquemot et al. 2007). Language acquisition takes place via exposure to one's native tongue (or parents' tongues in bi- and multilingual children) and through imitations[14], first by babbling and gradually by forming words and expressions of thoughts (Cheour-Luhtanen M. et al. 1995; Cheour et al. 2002; Kuhl 2004). The correct pronunciation of a foreign language can only be learned by listening and emulating the reference utterances of speakers of that language. Congenital deafness prevents proper learning of spoken language in childhood, whereas later deafness does not prevent speaking although it may

[13] Prosodic cue: the effect of the speech intensity and voicing contours on speech.
[14] Parents typically highlight the important features in what is called infant-directed speech (De Boer & Kuhl 2003).

interfere. Deaf children start babbling as the normally hearing do, but without the necessary auditory feedback, they soon cease to mimic the caretaker's speech (Milner 1999). Children with cochlear implants develop better speaking skills if the operation is done during the first 12 months (Miyamoto et al. 2008). These observations suggest that, for acquiring a spoken language, the necessary innate ability is not sufficient, but proper reference models and their imitation are equally essential. Elements of speech and language from phonemes to syntax are learned in a social process by listening and by speaking. A theoretical basis and experimental evidence for the importance of mimicking can be provided by the finding of *mirror neurons* in monkey's ventral premotor cortex area F5, which has a possible homology to the Broca's area of human brain (Rizzolatti et al. 1996; Kohler et al. 2002; Ferrari et al. 2003). Mirror neurons discharge similarly regardless of whether the monkey performs an action or observes another monkey to perform a similar action. It has been debated whether mirror neurons exist in human brain, and some evidence for their existence has been found in fMRI studies (Chong et al. 2008; Kilner et al. 2009; Heyes 2010). The mirror neuron system is suggested to be either an adaptation to action understanding, or a part of neuron network behind sensory-motor associative learning (Heyes 2010).

In the human speech organ, there are six articulators that can be actively manipulated for producing sounds with distinctive acoustic attributes: the vocal folds, pharynx, soft palate, tongue body, tongue blade and lips. In addition, the slackening or stiffening of glottis can alter the acoustic output (Stevens 2000; p. 249). The acoustics of human speech production is generally modeled by means of source-filter models (Fant 1960; Fant et al. 1985; Veldhuis 1998; Stevens 2000) (Figure 1) or articulatory models (Goldstein 1980; Maeda 1990; Perkell et al. 2000) (Figure 5), and since the 1960s, synthetic speech has been produced on the basis of these models (Karjalainen 1978; Klatt 1980; Laine 1989; Alku et al. 1999; Dang & Honda 2002). In the source-filter model of vowel production, the imaginary parts of the poles on the s-plane[15] of the vocal tract transfer function represent the natural frequencies (formants) of the tract, and the real parts give the decay of the oscillations at the natural frequency (see Chapter 2.3 Vowels).

When speech production is modeled in its entirety, the view has to be widened from the plain acoustic or articulatory domain to the neural control of speech (for a review see e.g. (Laaksonen 2006)).

[15]S-plane = frequency domain representation, computed by the Laplace transform.

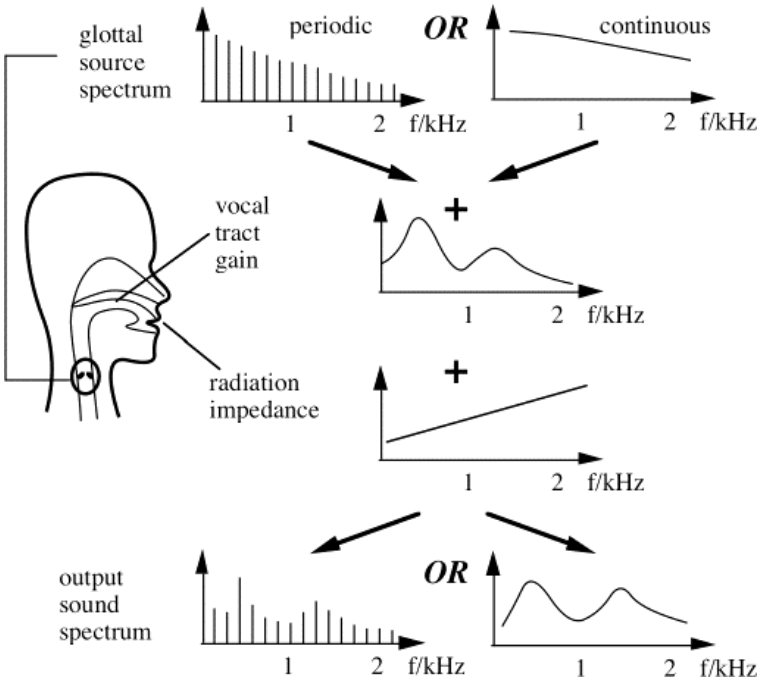


Figure 1. Simplified source-filter model of speech production (Wolfe et al. 2014, with permission)

Johnson and co-workers (Johnson et al. 1993b) have shown that speakers of the same language and dialect may use different articulatory plans, thus suggesting that the universal articulatory hypothesis is wrong. They found large individual differences in the general pattern of articulation, although the speakers produced the words consistently from one trial to another, and concluded that the speakers generate the same acoustic patterns with different articulatory strategies, in other words, speech production is controlled by the acoustic target templates stored in memory and adjusted via perceptual feedback when pronouncing the sound patterns. Moreover, this control seems to perform rather well since, according to the results, the sound patterns generated by an individual speaker are reproducible.

Perkell and co-workers have presented a theory of the segmental component of speech motor control (Perkell et al. 1995; Perkell et al. 2000;). In this theory, auditory feedback is used in forming and maintaining an internal mental model that maps the articulatory movements needed to achieve the auditory goals, represented by auditory templates. The articulatory movements are represented as a sequence of trajectory milestones via which the actual production system drives when speech is produced in a fire-and-forget manner. This facilitates a sound output with relatively stable acoustic properties despite of variation in the motor input. However, the real-time correction of movements between the trajectory milestones is hardly possible with auditory feedback, which

confines the role of immediate auditory feedback to the control of larger units than segmental components during the articulation.

Jones and Munhall have studied the effect of auditory feedback in compensating for the artificial modifications made to the vocal tract by a dental prosthesis (Jones & Munhall 2003). Their subjects were able to partially compensate for the vocal-tract modifications by using auditory feedback information: along with the increased experience of prosthesis use, the distribution of energy across the spectrum moved toward that of normal, unperturbed production. However, the acoustic analysis did not show any significant differences in learning dependent on auditory feedback.

Niemi and co-workers have studied the effects of distorting the somatosensory feedback path or the articulatory musculature control on the acoustics of vowels, diphthongs and sibilant /s/ (Niemi et al. 2002; Niemi et al. 2004; Niemi et al. 2006). They found that distortions affected the speech production, but the acoustic changes were highly individual. The results indicated further that speakers with minor distortions produced speech based on feed-forward mechanism, whereas significant distortions caused the speakers to rely merely on feedback information.

Directions Into Velocities of Articulators (DIVA) is a recent model of speech production presented by Frank Guenther (Guenther et al. 1998; Callan et al. 2000; Guenther 2006). It is a neural network model that consists of a feed-forward control subsystem and a feedback control subsystem (Figure 2). The blocks in the model have hypothesized neuroanatomical equivalents in the human speech system (e.g., Auditory State Map in Superior Temporal Cortex). DIVA also takes into account speech motor skill acquisition: In the learning phase, the neural mappings of the model are adjusted through auditory and somatosensory feedback from self-generated babbling sounds to form relations between the motor actions of the computer-simulated vocal tract and their acoustic and somatosensory consequences. After the learning phase, DIVA is capable of producing arbitrary speech sounds.

Hickok and co-workers suggest in their review article (Hickok et al. 2011) that the interaction between speech production and perception can be studied in a unified framework based on state feedback control (SFC) and sensorimotor integration. The basic idea of the feedback control models of auditory-motor interaction in speech production is presented already by Fairbanks (1954), and in essence, it states that speech goals are represented as sequences of sensory outcomes, and the articulators produce speech directed by a control system that minimizes error between the desired and actual sensory feedback.

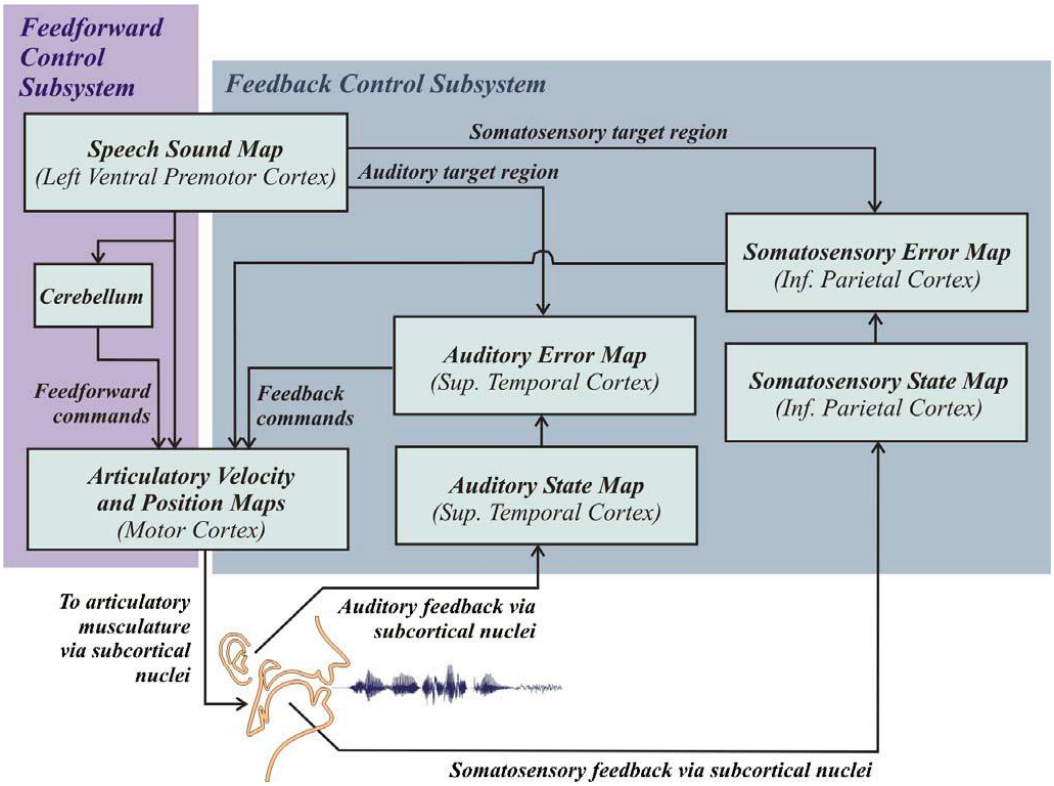


Figure 2. Schematic diagram of the hypothesized neural correlates of the DIVA model (Guenther 2006).

There are, however, known problems in the explanatory power of the real time feedback control of speech articulation; the nerve conduction velocity and processing delay for forming the error function and motor correction do not allow for real time articulation control. Every speaker implicitly knows this by experience: a pronunciation error cannot be corrected on fly although one recognizes that the likely output will be something else what was meant, but reproduction is required to correct the error. In the state feedback control (SFC) theory (Ventura et al. 2009; Hickok et al. 2011) (Figure 3), the feedback correction to motor commands bases on the Kalman filter approach (Kalman 1960). In this model, an internal estimate of the current dynamical state of the vocal tract is updated on the basis of delayed acoustic feedback and an efferent copy of the perceptual representation of the sensorimotor status of the articulators; the online control takes place through internal forward model predictions, and the model itself is trained and updated by actual feedback. The internal model of the vocal tract states (articulatory gestures) is based on previously learned associations between the sensory outcomes and the underlying motor commands. The predictive nature of the state feedback controller is based on these learned associations; the system can predict the likely sensory

consequences to the desired motor commands before the actual sensory feedback is available, and use this predicted state information in forming the error control signal.

The SFC offers a tempting framework for testing the idea that perceptual prototypes play a role in the forming of the internal model of vocal tract states. In this thesis, Studies I and III concentrate on vowel production, and Study I also provides a comparison of the productions with vowel perception (though in separate experiments). In Study III, the auditory feedback loop of production is blocked using either pink or babble noise masks in the experiments. Noise masking is known to cause the so-called Lombard effect (Lane & Tranel 1971). In this thesis, one of the aims of the different studies is to figure out how the Speech Sound Map of the DIVA model (see Figure 2) and the desired speech target and the internal model of vocal apparatus in the SFC model (see Figure 3) are represented in terms of formant frequencies of perceptual vowel prototypes and vowel productions.

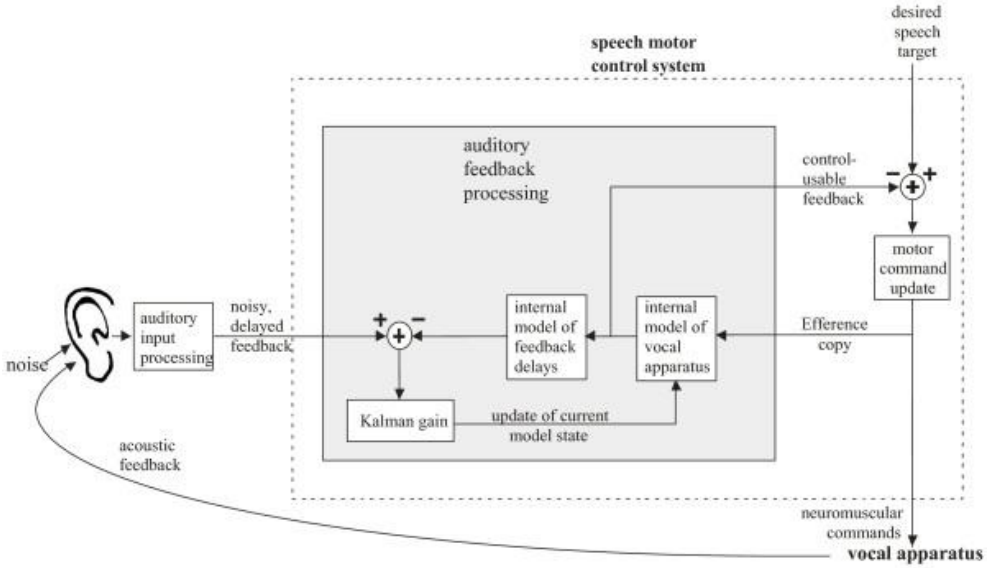


Figure 3. The SFC (State Feedback Control) model of auditory feedback processing for speech as proposed by Houde & Nagarajan (Ventura et al. 2009).

2.3. Vowels

Vowels are voiced sounds which are produced when the vocal tract modifies the air flow generated by glottal excitation[16]. The IPA (International Phonetic Alphabet) vowel diagram is presented in Figure 4. In phonetics, the vowels are generally described in terms of the place or way of articulation (Suomi et al. 2006): high-low (close-open), front-

[16] Some consonants are phonetically vowel-like, i.e., they are produced without constriction, e.g. [j] and [w].

back and rounded-unrounded. The IPA vowel diagram reflects the places of articulation in the oral cavity as given in the sagittal view of head (Figure 5). The distinctive features of vowels in articulatory phonetics are bound to the place and way of articulation (Stevens 2000; pp. 249-255).

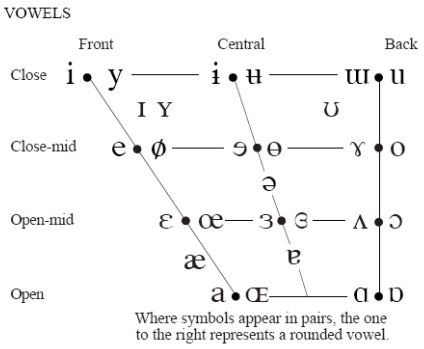


Figure 4. IPA vowel diagram (2005)

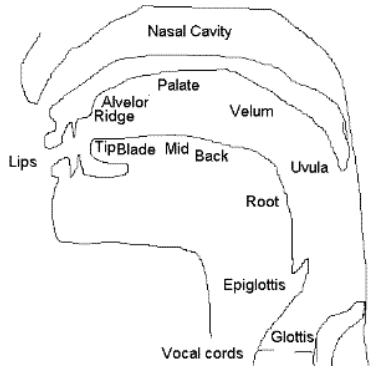


Figure 5. IPA articulators

The so-called neutral or schwa vowel (denoted as /ə/ in IPA) is a voiced speech sound that results when the airflow passes through an open vocal tract, air flows through middle of mouth, and no friction noise is formed. Physically, this can be approximated by a uniform tube without acoustic losses and closed at the one end (infinite impedance) and open at the other end (zero impedance). Mathematically, this can be formulated by a one-dimensional wave equation (1):

$$\frac{d^2p}{dx^2} + k^2p = 0 \qquad (1)$$

p = pressure
x = displacement
k = 2πf/c
f = frequency
c = velocity of sound

The tube resonates at natural frequencies, which can be derived from equation (1) with x=0 at the open end of tube, l is the tube length, and c is sound velocity (Stevens 2000; pp. 138-139). The resonant frequencies F_n are called formants (F1, F2, F3, and so on):

$$F_n = \frac{2n-1}{4}\frac{c}{l} \qquad (2)$$

For a uniform tube, the resonances thus correspond tube lengths with odd multiples of a quarter wavelength, and for an /ə/ sound produced by a male with a 14.1 cm vocal tract, the calculated formants are F1=630 Hz, F2=1880 Hz, and F3=3140 Hz (Stevens 2000; pp.138-140). Apparently, approximating the vocal tract by a uniform tube is not sufficient for describing all the variety of different vowels, but perturbing its shape is needed. In articulatory phonetics this is often done by modeling the oral cavities and positions of articulators by connected tubes with varying lengths and cross sections, i.e., using coupled Helmholtz resonators (Figure 6).

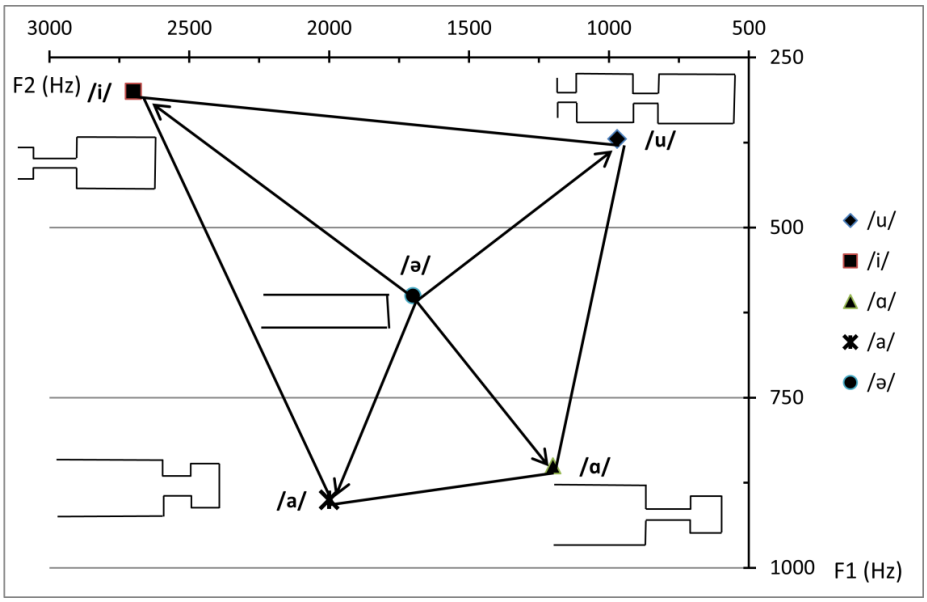


Figure 6. Cardinal vowels in the F1-F2 formant space as approximated by perturbing the shape of a uniform acoustic tube (length 15.4 cm) generating the schwa vowel /ə/ (in the middle of the chart). Figure modified from (Stevens 2000; p. 286).

Of the many different vowel sounds found in various languages in the world, the eight primary cardinal vowels [i], [e], [ɛ], [a], [ɑ], [ɔ], [o] and [u] are used as reference vowels in describing the other vowel sounds (Ohala 1983). The role of the cardinal vowels as reference sounds has made their acoustic and articulatory properties, and their variation, subjects of intensive research (Vaissière 2011). In the IPA vowel chart (IPA 2005), the primary cardinal vowels represent the vowels most right or left on the chart, the axes of which reflect the formant F1 (y-axis) and F2 (x-axis) values of the vowels, /u/ vowel representing the origo with the lowest F1 and F2 values. Anatomical and physiological constraints set the limits for sounds that human talkers can produce; for vowel sounds, these extreme sounds in the front-back and high-low distinction are called corner vowels [i], [u], and [ɑ].

Generally, the first two formant center frequencies, F1 and F2, are regarded as sufficient to provide the essential information on the quality of vowels (Peterson & Barney 1952; Aaltonen & Suonpää 1983), but often the formant F3 and F4 frequencies are employed as well. For example, F3 is involved in the differentiation between rounded and unrounded vowels (Aaltonen 1985), and between speaker gender, physical size and age, and therefore F3 is suggested to be used as a scaling factor for the transformed formant axis F1/F3 and F2/F3 (Monahan & Idsardi 2010). The formants F3 and F4 are important in singing voice analysis, and the higher formants F4–F6 play a role in determining the voice quality. When radiation impedance of the mouth opening, the impedance of the vocal tract walls (compliance), and the heat conduction and viscosity are added to the resonator models, a more realistic model describing the vowel production is available. The real parts of these impedances cause energy losses in the airflow and thereby show as increased bandwidths of the natural frequencies.

The effect of formant amplitudes on vowel categorizing has been studied with slightly contradictory results (Carlson et al. 1970; Chistovich 1979; Aaltonen 1985; Klatt 1985; Schwartz & Escudier 1987), but generally, the influence of formant amplitude is regarded minor, as compared with the formant center frequency: changes up to 25 dB in the formant amplitude do not affect the identification, provided that the center frequencies are unchanged (Rosner & Pickering 1994; p. 161). Formant bandwidths have been reported to have a minor effect on identifications (Carlson et al. 1979).

Several different formant axis conversions and scaling approaches are suggested (Liljencrants & Lindblom 1972; Miller 1989; Strange 1989), among them the two formant models with F1 and F2' where F2' is constructed through a rather complicated calculus from F2, F3, and F4, and from the vocal tract transfer function in the valley between F3 and F4 (Bladon & Fant 1978; Paliwal et al. 1983). The idea behind the F2´ conversion is that the influence of the higher formants on identification is reduced to one characteristic figure, and 2-dimensional mapping can still be used to represent the vowel diagram.

Formant based approaches, especially the two formant F1–F2 representation model, evidently do not utilize all the information available in the acoustic spectrum. Therefore, a whole spectrum approach to vowel identification has been suggested in which the vowels are compared on the basis of their entire auditory loudness density spectra (Bladon & Lindblom 1981; Bladon 1982). This method, however, requires the use of a multidimensional vector space in contrast to the two-dimensional space of the F1–F2 representation. In an experimental study by Carlson and Granström, the whole spectrum model by Bladon and Lindblom performed well in psychophysical distance judgments but poorly in phonetic distance judgments (Carlson et al. 1979). Savela has compared the role of different spectral attributes, the formants and spectral moments (center of gravity, standard deviation, skewness and kurtosis) in the identification and discrimination of vowels in 15 different languages (Savela 2009). Savela's main conclusion was that, while

formants are the primary criteria used in vowel identification, spectral moments suit to be used as a secondary attribute, and neither can be the only explaining factor in perceiving vowels.

2.3.1. Finnish vowel system

The Finnish vowel system (Figure 7) includes eight vowels: /a/, /e/, /i/, /o/, /u/, /y/, /æ/ and /ø/ (Suomi et al. 2006) which all can occur as short (single) or long (double) in any position of a word. In written texts, the short vowels are denoted by the orthographic symbols <a>, <e>, <i>, <o>, <u>, <y>, < ä> and <ö>, while two identical symbols indicate the long vowels <aa>, <ee>, <ii>, <oo>, <uu>, <yy>, <ää> and <öö>. In the Finnish orthography, the conventions to express the quantity oppositions were originally slightly different from modern standard Finnish. However, the orthography stabilized to its present form in the early 19th century and has remained similar since then. The modern orthography of Finnish thus reflects the identity group interpretation of quantity opposition: the long vowel segments of spoken Finnish consist of two similar shorter segments (Karlsson 1983).

The identity group interpretation of Finnish quantity opposition would suggest that the durational ratio of short and long segments is 1:2. However, the segmental length in Finnish is not fixed, but gradient and dependent on contextual parameters, word length, speaking rate, and speaker-specific factors, as shown in several studies (Harrikari 2000). Wiik and Lehtonen have reported durational ratios 1:2.7, and durations 60–100 ms for short vowels and 160–270 ms for long vowels measured from words embedded in sentences (Wiik 1965; Lehtonen 1970). Kukkonen found mean durational ratio 1:2.25, and durations 130-150 ms for short vowels and 250-310 ms for long vowels measured from isolated words (Kukkonen 1990). In a more recent study (Ylinen et al. 2006), /u/ variants with a duration of less than 100 ms were categorized as short, both in a word and in an isolated vowel condition, while vowels with durations of more than 150 ms in a word context and of more than 175 ms in an isolated vowel condition were categorized as long. In this study, the mean durational ratio of produced short and long /u/ vowels was 1:2.03. According to the studies by Suomi and co-workers, there are four statistically distinct, non-contrastive duration degrees for phonologically single vowels: extra short (48 ms), short (58 ms), longish (73 ms) and long (84 ms), and three degrees for double vowels: longish + longish (149 ms), long + extra short (142 ms) and very long (135 ms), indicating that, within the binary quantity opposition, there is a categorical fine structure of duration as well (Suomi et al. 2003; Suomi & Ylitalo 2004; Suomi 2005; Suomi 2006; Suomi 2007).

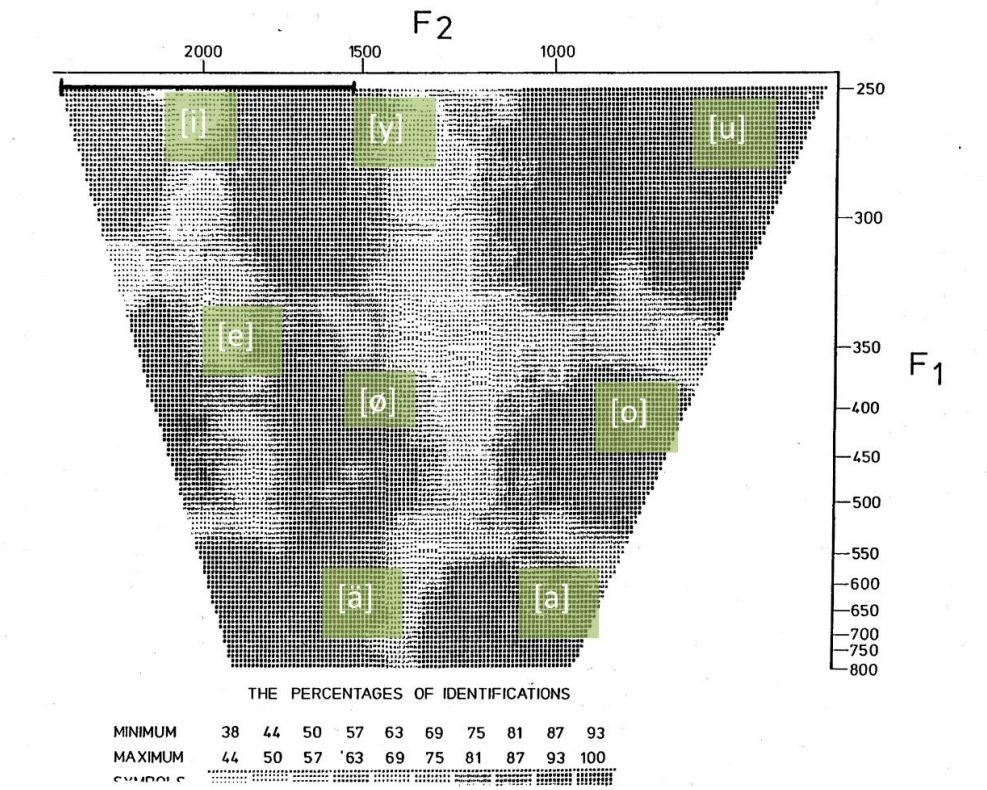


Figure 7. The identification consistency (shown on grey scale) of the Finnish vowel system, averaged data of 32 listeners. The formant F1 and F2 frequency axes are in Hz. Modified from (Aaltonen & Suonpää 1983).

Generally, the two different durational variants are regarded as being similar in perceived quality, but there are reports on minor spectral dissimilarities in the formant frequencies of the produced short and long vowels. For example, Wiik (1965) reported clear differences in the variability ranges of Finnish single and double /y/ and /i/ vowels as measured in terms of F1, F2 and F3, stating that F1 is 40 Hz higher and F2 is 75 Hz lower in [y] than in [y:], and F1 is 65 Hz higher, F2 is 140 Hz lower, and F3 is 265 Hz lower in [i] than in [i:]. In a later study on vowel production by Kukkonen (1990), differences of similar type but smaller magnitude were reported from a normal Finnish-speaking control group (N=4): F1 was 16 Hz higher, and F2 and F3 were 63 Hz and 32 Hz lower in single than in double /i/ vowel. For single and double /y/ vowels the corresponding differences were as follows: F1 was 19 Hz higher, F2 was 75 Hz lower, and F3 was 20 Hz lower. However, only the F1 differences were statistically significant. In some earlier studies (Eerola et al. 2002), a non-significant difference of 109 Hz was found for F2 between the short /i/ (F2=2391 Hz, SD=194 Hz) and long /i:/ (F2=2500 Hz, SD=212 Hz) produced by 26 informants in the first syllables of the Finnish words *tikki* and *tiili*. Iivonen and Laukkanen studied the qualitative variation of the eight Finnish vowels in 352 bisyllabic and trisyllabic words

uttered by one male speaker. They found a clear tendency for the short vowels to be more centralized in the psychoacoustic F1–F2 space, as compared with the long ones. However, except for the /u/ - /uː/ pair, this difference was smaller than one critical band, and thus auditorily negligible (Iivonen & Laukkanen 1993).

2.3.2. German vowel system

According to the classical view, the German language has 15 distinctive vowel categories (e.g., the database study of Heid (1995)) and, in different studies, the number of vowels in modern standard German is varyingly reported as 14–17 (Moosmüller 2007; p.33). However, the quality difference between the German vowels of different quantities is dependent on the syllable structure of the word and on the results of different centralization mechanisms. Becker argues that there are only eight vowel categories in German, and they are in correspondence with the eight orthographic vowel qualities: /ɑ/, /e/, /i/, /o/, /u/, /y/, /æ/, and /ø/ (Figure 8) (Becker 1998).

German speakers are known to be sensitive to the quality difference between short and long vowels (Sendlmeier 1981). They are able to distinguish between the quality of the first vowels in word pairs such as <offen> [**o**f:en] and <Ofen> [o:fen] or <Mitte> [m**ɪ**tte] and <Miete> [mi:te][17]. Although not all German speakers make a difference between /e:/ and /æ:/ when pronouncing words, this distinction exists between the names of the alphabets <e> and <ä> if pronounced separately. In German, there are larger differences in the loci of F1 and F2 in the vowel space between long and short vowels than in Finnish: the differences in the phonological quality of these variants are mainly based on duration. There is also variation in the German dialects spoken in the different German-speaking regions (for example, Northern Germany, Bavaria and Austria) (Iivonen 1987; Winkler et al. 1999).

[17] Word meanings: offen - 'open', Ofen - 'oven', Mitte - 'midpoint', Miete -'rental' (Krech et al. 2009; Mangold, 2005).

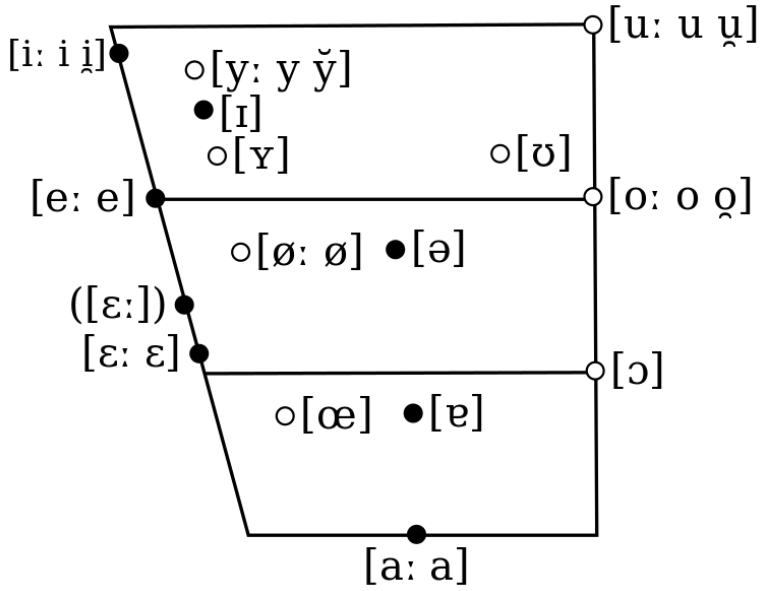


Figure 8. German oral monophthongs[18] (Mangold 2005)

2.4. Vowel prototypes and perceptual magnet effect

The *prototype view* originates from the Gestalt tradition (Eysenck & Keane 1992; pp. 44-55), and from the fundamental work of Rosch (1975), which suggests that perceptual categories are gradual and based on prototypes, and that prototypes play a role in object recognition. Prototype theories assume that perceptual categories consist of classes of stimuli with similar features or characteristic attributes, and that they are organized around central prototypes, that is, the best representatives of the stimuli forming the class (Eysenck & Keane 1992; p. 263). The categories have an internal structure and fuzzy boundaries to neighbouring categories. Fairbanks and Grubb discovered that all category members identified as belonging to a particular category are not equally important representatives of the category: there are preferred samples of a certain vowel class, which are "the most representative samples from among the most readily identified samples" (Fairbanks & Grubb 1961). Prototype based theories of perception assume that new sensory information is first processed, often in a non-linear fashion, into a particular form, which is then compared to the stored memory representations, i.e., the prototypes. Recognition takes place when the best match to a stored representation is achieved. In processing differences in phone quality, the phoneme prototypes are suggested to act as reference templates for individual quality categories.

[18] http://en.wikipedia.org/wiki/File:German_oral_monophthongs_chart.svg (accessed 18.3.2014)

A class of theories dealing with prototype effects in speech perception considers the prototype effects as being emergent phenomena in a learning system, rather than essential representations with exemplars in memory. Such theories are proposed by Guenther and Gjaja (Guenther & Gjaja 1996) and Boersma (Boersma & Hamann 2008; Boersma 2009). In this thesis, it is assumed that the ability to perceive distinct auditory representations for a particular sound precede the ability to intentionally identify them to distinct phonological categories. However, when the sounds are repeatedly heard during the language acquisition phase, memory traces will gradually be formed for such sounds that are keys to further phonological processing of longer segments and words.

In the literature, two separate effects related to phoneme prototypes have been presented: the *phoneme boundary effect*, in which the sensitivity to phone differences peaks at category borders (categorical perception), as shown in phone identification experiments (Miller et al. 1983; Nearey 1989; Strange 1989; Repp & Crowder 1990; Nábelek et al. 1993; Miller 1997), and the *perceptual magnet effect* (PME), in which the least sensitivity occurs in the vicinity of perceptual prototypes, as shown in phone discrimination experiments (Kuhl 1991; Iverson & Kuhl 1995; Guenther & Gjaja 1996; Aaltonen et al. 1997; Iverson & Kuhl 2000). The PME actually suggests that prototypes shrink the perceptual space around them and thereby generalize sensations to preset categories. The existence of an internal structure of phonetic categories and prototypical category representatives has been shown in many reports (Miller 1997), whereas the existence of the PME as an independent phenomenon that is not related to general perceptual contrast effects has been challenged in some reports (Lively & Pisoni 1997; Lotto et al. 1998; Lotto 2000), resulting in counter-arguments (Guenther 2000).

Kuhl and co-workers diverted the emphasis of speech perception from the category boundaries to the centers of categories. According to Kuhl, the discrimination is affected not only by physical factors but also by stimulus typicality: the best category exemplars, or prototypes, are more difficult to discriminate from their neighbors than the less prototypical ones (non-prototypes). Kuhl further suggests that infants develop such prototypes through experience with spoken language. Since Kuhl's findings, the neural and theoretical bases of the perceptual magnet effect have been studied by several research teams. Guenther and co-workers developed a self-organizing neural network model for the magnet effect (Guenther & Gjaja 1996). Aaltonen and co-workers (1997) reported that the perceptual magnet effect was only manifested in the good categorizers, evidenced both by behavioral and psychophysiological discrimination data. Sharma and Dorman (1998) were not able to reproduce Kuhl's finding of the magnet effect: firstly, the discrimination accuracy was not significantly different in the prototype and non-prototype conditions, and secondly, the psychophysiological data from mismatch negativity (MMN) recordings seemed to reflect category acoustic differences rather than stimulus typicality effects.

2.5. Internal variation of vowel categories

Vowel categories are formed during the early language acquisition in childhood under exposure to one's native tongue (Grieser & Kuhl 1989; Kuhl et al. 1992; Jusczyk 1993; Kuhl 2004). The Word Recognition and Phonetic Structure Acquisition (WRAPSA) model assumes that the early proficiencies displayed in many speech perception tasks with infants under 6 months old are the result of general auditory analyzers. However, during the latter part of the first year, infants start to weigh the information available through the auditory analyzers and derive prosody-based syllable-sized featural information which, when matched to utterances encountered previously, leads to word recognition (Jusczyk 1993).

Part of the inter-individual variation in the categories can be attributed to the differences in the caretakers' speech. Obviously the child's own produced speech categories cannot exceed the acquired categories without confusion. Thus, there are two possible ways to solve the consistency demand between own perception and production of vowel categories: either the originally acquired category boundaries have to be adjusted according to one's own production or the own production is affected by the acquired categories. Bart de Boer has investigated the forming of vowel categories by simulations[19], and has shown that it is possible to explain the universal tendencies of vowel systems as a result of self-organization in a population under constraints of perception and production (De Boer 2000; De Boer & Kuhl 2003). For the purpose of simulation, de Boer used autonomously working computer programs to represent agents with the capabilities of realistic articulatory synthesis (Maeda 1990), human like perception, and the ability to learn by imitating. In the simulation, the agents communicate in an imitation game: the initiator articulates first a random vowel from its repertoire with noise added to simulate the communication channel, and the imitator analyzes it against its own vowel repertoire, and then articulates it back to the initiator who decides whether it was the same sound it originally transmitted. During the game, the agents store vowels as prototypes (in the machine learning sense), and the classification of sounds prior to their imitations is based on the nearest prototype rule: the sound to be reproduced is the best matching prototype, not the actually "heard" sound. The prototypes themselves are dynamic in the sense that their place in the vowel space can be moved on the basis of new classification information. By repeating the imitation turns up to 10 000 games in a population of 20 agents, vowel systems emerge that are, with their internal variation, very similar to those existing in natural languages (De Boer 2000; p. 454).

[19] De Boer's computer simulation is an imitation game between population agents inspired by the artificial-life methodology for investigating complex natural phenomena (Emmeche 1996; Langton 1990).

The fuzzy boundaries of the Finnish vowel categories reported by Aaltonen & Suonpää (1983) and Iivonen & Laukkanen (1993) suggest that there has to be large variation in the identifications done by individual listeners, even to the extent that the same phoneme would be classified into different categories by different people. Furthermore, in another study by Aaltonen and co-workers (Aaltonen 1997), individual variation in the location of the vowel boundary manifested variation in the within-category fine structure as well. Although overlapping categories would exist inter-individually, it is hard to imagine overlapping categories intra-individually, that is, between one's own perception and production. In other words, it is supposed that the articulation of a particular vowel by a person would always match best with the prototype of the same perceptual vowel category, not that of a neighboring category, by the person in question. This is investigated in Study I.

2.6. The initial auditory theory of vowel perception

Rosner and Pickering suggest in their initial auditory theory of vowel perception that three local effective vowel indicators (LEVIs), E1, E2 and E3, which are based on the perceptual correlates of the first three physical formants (F1, F2 and F3) of a vowel, and additional temporal information (D) on the physical duration (d) of the vowel, together determine a point (E1, E2, E3, D) in the auditory vowel space (AVS) for a particular speaker (Rosner & Pickering 1994). The theory represents strong auditory theories, since it is based on auditory loci in preference to physical formants. Rosner and Pickering do not present any closed form mathematical formula for the transformation of the time domain acoustic information (as represented by F1, F2, F2, and d) to the LEVIs and D; however, they describe some principles and introduce perceptual processes participating in this conversion (e.g., the auditory conversion of physical frequency to pitch, and the effect of speaking rate on duration in the form D ~ dR, where R is the momentary speaking rate).

In this thesis, the Hz - mel conversion[20] is used as the first approximation for transforming the F2 frequency to the LEVI E2. For the temporal information, the approximation is D = d, i.e., the physical duration is used as such. In the initial auditory theory of vowel perception, the vowel identification rests on the nearest prototype rule: the listener first relies on (and always can back up to) the learnt language-specific prototypes, against which he compares the speaker's AVS points. Identification then results as the best match of the speaker's AVS point with the set of the listener's prototypes. Whenever possible, the listener uses prototypes that reflect the speaker class (gender, age), and during the conversation, the listener also attempts to adjust the prototypes for a particular speaker's

[20] Fant's formula: $f_{mel} = 2410*\log_{10}(1.6*f_{Hz}/1000+1)$ (Lindsay & Norman 1977)

voice, a process that may temporarily move the prototypes away from their initial position.

2.7. Detection of small frequency differences of pure tones and vowels

The meaningful and detectable differences in the parameters of loudness and frequency of the acoustic stimuli are an important issue in the planning of behavioral psychoacoustic experiments for phonetic research. The *difference limen* (DL) or *Just Noticeable Difference* (JND) of a pure tone frequency (f) depends on the absolute frequency, but also on the sound intensity and sound duration. DLs and JNDs are expressed either as a physical frequency (f_{Hz}) or on a psychoacoustic scale (mels, barks or ERBs). Often they are also given as percentage (%) values that actually refer to the Weber ratio Δf/f (Liu & Kewley-Port 2004).

The measuring procedure also affects the DL values. Various studies on the DL of frequency for a 1000 Hz pure tone (SPL > 50 dB) has given a DL range of 1–10 Hz; for example, 6.7 Hz (range 5–11 Hz) (Meurmann 1954), 3.6 Hz (range 2–6 Hz) (König 1957), and 1.2 Hz (Wier et al. 1977). The shortening of the sound duration to less than 200 ms or the lowering of sound pressure level below 20 dB will increase the DL. For complex sounds, the DL of a frequency component depends on the spectral composition of the sound and on the perceptual *masking effects* of the ear (Plomp 1964). For vowels, the DL is 4–5% of the formant frequencies, as stated in the classical study of Flanagan (1955), although some more recent studies have reported both larger (10–15%) (Mermelstein 1978) and smaller (1–3%) values (Hawks 1994; Liu & Kewley-Port 2004). There is also variation in the discrimination capability of individuals. Moore (1976) reported a 1 to 40 ratio between the worst and best subjects (N=20) in a DL experiment for a 1 kHz sinusoidal tone. The mean DL values of two successive sessions were 8.4 Hz (SD 10.0 Hz) and 6.5 Hz (SD 7.1 Hz), with the latter result being an indication of learning effect.

For testing any prototype related theories, one has to have access to the minor perceptual phonemic differences within a phoneme category. While psychoacoustic experiments approach the perceptual processes retrospectively via the subject's conscious responses, psychophysiology offers a more direct view to the neural processes involved in perception, memory and attention. Event related electroencephalography (EEG) or magnetoencephalography (MEG) recordings, in which signal averaging technique is used for detecting minor potentials or fields evoked by auditory stimuli (Regan 1989; p. 47), offer feasible means to study also the pre-attentive perceptual processing. An ERP component, *mismatch negativity* (MMN), has proven to be an especially useful indicator of pre-attentive auditory discrimination, which reflects the activation of echoic memory (Näätänen et al. 2007). The MMN response is elicited when a rare deviant stimulus randomly occurs in a stream of frequent standard stimuli. Some of the clinical and practical constrains and best practices of the method are discussed in (Lang et al. 1995);

the smallest theoretically measureable MMN responses are defined and compared to JNDs of pure tones and complex tones.

Besides MMN, several other ERP components are used to study the auditory and visual language processing: N100[21] (Näätänen & Picton 1987; Pantev et al. 1988; Sams et al. 1990), P200 (Evans & Federmeier 2007; Ross & Tremblay 2009), N200 (Schmitt et al. 2000; Patel & Azzam 2005), P300 (Comerchero & Polich 1999; Patel & Azzam 2005), N400 (O'Rourke & Holcomb 2002), and P600 (Kuperberg et al. 2003). The earlier ERP responses N100, P200 and N200 are merely evoked potentials reflecting sensory processing whereas the later responses are associated with attentional processes (P300) (Polich 2007), language and speech comprehension (N400) (O'Rourke & Holcomb 2002), or grammatical errors and conceptual anomalies (P600) (Coulson et al. 1998; Kuperberg et al. 2003). Some studies report that there may be the same brain mechanisms underlying, e.g., N100 and MMN (May & Tiitinen 2004; May & Tiitinen 2010), and P300 and P600 (Coulson et al. 1998).

Aaltonen et al. (1994) applied the MMN method to study the automatic discrimination of phonetically relevant and irrelevant vowel parameters. The second formant F2 frequency of a synthetic [y] vowel served as the relevant, and the fundamental frequency f0 as the irrelevant parameter. Similar changes in sine wave frequencies were used as a reference. From the acoustical point of view, the experiment was conducted to compare, one frequency component at a time, the automatic processing of complex tones consisting of several frequency components and that of pure sine tones consisting only of one frequency component. From the phonological point of view, the situation is different. Since the f0 represents the glottal source oscillator frequency and the F2 represents a resonance peak of the vocal tract, the discrimination ability of similar relative differences in these frequencies of a phoneme is of interest: f0 is associated with voice quality, gender, and to some extent, the physical size of the speaker (of the same gender), but not with vowel identification, whereas F2 affects strongly the identification of Finnish vowels. Therefore, from the phonological point of view, the automatic processing of similar relative differences in the f0 frequency and in the F2 frequency was expected to be different for vowels than for sine tones, and also different for f0 than for F2.

In the experiment, the fundamental frequency for the standard stimulus was set at 100 Hz and for the deviant stimuli at 104 Hz, 120 Hz and 140 Hz. These correspond to a male voice pitch, except for 140 Hz, which can also be perceived as a low female pitch. These f0 variations were applied to two different vowel variants; for [y] (F1=240 Hz, F2=1700 Hz, F3=3030 Hz, F4=3500 Hz, F5=4200 Hz), and for [y/i] (F1=240 Hz, F2=2050 Hz, F3=3030 Hz, F4=3500 Hz, F5=4200 Hz). Hence, similar f0 standard-deviant pairs were used for a good category [y] vowel and for a boundary area vowel between [y] and [i]. Using the good

[21] N=negative and P=positive waveform deflection; 100, 200 etc. refer to milliseconds from stimulus onset.

category [y] as the standard, the following deviant F2 frequencies were used in the F2 condition: F2=1768 Hz (4% increase), F2=2050 Hz (20% increase), and F2=2400 Hz (40% increase).

All deviant stimuli, even the ones with the smallest frequency difference of 4%, evoked MMN responses (irrespective of the electrode), and the amplitudes of the responses increased with increasing frequency differences. This is in accordance with earlier reported MMN results. Vowel stimuli produced lower MMN amplitudes than the pure tones at 4% and 20% frequency differences, but not anymore at 40% difference; this is probably due to the fact that a larger difference will initiate other brain mechanisms (attention shift), which diminish the difference between vowels and pure tones observed with smaller deviants. Another explanation to the lower amplitudes with vowels than with pure tones is offered by the stimulus intensity that was set equal for all stimuli while, for vowels, one component of the complex stimulus spectrum was altered at a time whereas for pure tones the entire energy is carried at the single frequency. The pure tone frequency change may generate louder sensation than a corresponding change of f0 or F2 across the entire spectrum of a vowel, thus resulting in larger MMN amplitudes.

In vowels, the 4% difference in f0 produced significantly greater MMN amplitudes than the 4% difference in F2. This is in line with the behavioral measurements of just noticeable differences: JND is 0.3–0.5% for f0, and 3–5% for F2 (Flanagan 1955). For the purposes of this thesis, the finding that increasing f0 and F2 frequency differences cause increasing MMN responses and, especially the finding that the smallest F2 difference (4%, 68 Hz, 31 mels) of a good category exemplar of Finnish [y] evoked an MMN response, proved that the ERP component MMN can be used to study the effect of minor changes of formant frequencies on the automatic discrimination between vowel variants within the same category. Further, the F2 difference of about 30 mels[22], when detected pre-attentively, would justify the use of such a difference in behavioral listening tasks with the subjects attentively focusing on minor stimulus differences.

2.8. Masking the speech production by noise

Noise is a stochastic process, n(t), the amplitude of which cannot be precisely predicted at a certain time moment t_i by any analytical function. Instead, in the time domain, the noise amplitude can be described by a probability density function (PDF) which is often assumed to be Gaussian (other PDF types are, e.g., uniform, Rayleigh, and Maxwell distributions) (Bendat & Piersol 2000; p. 66). Noise is a *strongly stationary* process if all possible statistical moments and joint moments are time invariant. For a *weakly stationary* random process, the mean value (RMS) is a constant over any integration

[22] In Studies I and II, 30 mels is used as the stimulus difference. In Study IV, 30 mels is used for F1 and 40 mels for F2.

period, and the autocorrelation function depends only on the time displacement. Noise is *ergodic* if the statistics calculated over any time period from any possible ensemble are the same. In the frequency domain, (white) noise has a flat power versus frequency distribution from DC to infinity. In practice, the noise frequency range is limited by band pass filtering. Noise is always present in communication channels, and there is ample literature on the detection of signals in noise (Green & Swets 1966; Carlson 1986; Rao & Letowski 2006).

In *frequency masking*, the presence of one sound or frequency component of a complex sound affects the perception of other sound or frequency components: a lower frequency tone will mask a higher frequency tone more effectively than vice versa. In *temporal masking*, an immediately preceding or following sound affects the perception of the target sound (Plomp 1964). The pre-stimulatory or *forward masking* is explained by a temporary decrease of sensitivity due to fatigue. The post-stimulatory or *backward masking* is suggested to be a result of late interactions of neural processing at the higher levels of the auditory pathway (Haughton 1980; p. 51). The effective time span of forward masking is less than 200 ms, whereas the stronger backward masking works only with shorter delays of 20 ms or less (O'Shaughnessy 1987). If the repetition rate of the used stimuli is high (ISI < 200 ms), temporal masking may influence the result even though frequency masking is not used in the experiment.

Masking experiments have led to the concept of *critical bands,* which are imaginary band-pass filters around a *center frequency (CF).* Critical bands can be derived by different methods (Greenwood 1961), but one of the most convenient is to mask a pure tone with wideband noise: by reducing the noise bandwidth around the pure tone and by keeping the total noise power constant until the pure tone is audible one can find the critical band around the pure tone frequency. It is noteworthy that the loudness of a band-limited noise within a critical band is heard constant until the noise bandwidth exceeds the critical band. Zwicker and Terhardt have suggested an analytical expression for the mapping of a frequency into critical band rate (Zwicker & Terhardt 1980; Weitzman 1992). The bandwidth of the critical band increases when the center frequency increases, being 100 Hz at 200 Hz, 160 Hz at 1000 Hz and 700 Hz at 4000 Hz. Critical bands are frequently used in psychoacoustic perception experiments along the mel, König and ERB scales to correct the non-linear frequency response of hearing (Traunmüller 1990; Weitzman 1992; Iivonen & Laukkanen 1993; Aaltonen et al. 1997; Iivonen & Harnud 2005).

Over a century ago, it was discovered by Etienne Lombard that speakers attempt to alter their voice level accordingly when the ambient noise level increases or when their hearing of own voice decreases; this is known as the Lombard effect[23] (Lane & Tranel 1971; Lau

[23] Also known as the Lombard reflex.

2008). The Lombard effect has been reported to cause an increase in intensity and pitch, a shift of spectral energy toward the medium frequencies, a decrease of speech rate, articulatory movements of greater amplitude, and phoneme modifications (Garnier et al. 2010). These effects have recently been demonstrated also in speech synthesis by modeling the vocal effort continuum in breathy, normal, and Lombard speech (Raitio et al. 2014).

Van Summers and co-workers studied the effect (Van Summers et al. 1988; Castellanos et al. 1996; Beckford Wassink et al. 2007) on two informants, and found that the mean RMS amplitudes and mean f0 frequencies for words they produced in quiet and with a masking noise at 80, 90, and 100 dB SPL increased significantly ($p<0.01$) along with the increasing noise; the mean amplitudes were 58.5 dB (quiet), 63.1 dB (80 dB), 64.0 (90 dB) and 65.4 (100 dB). Furthermore, significant prolongations (10–25%, $p<0.0001$) of word durations were found in masking conditions. However, no significant main effects on vowel formant frequencies F1 and F2 were found, except for F1 of the first informant. In a subsequent perception experiment, 41 listeners identified the utterances of the two informants produced in quiet and at noise levels 90 dB and 100 dB in three different listening conditions: SNR -5dB, -10 dB, and -15 dB. SNR had a significant main effect on the identifications ($p<0.0001$), as was expected, but more interestingly, the utterances produced at 90 dB and 100 dB masking noise were consistently better identified than the ones produced in quiet, regardless of the talker or SNR. The results of the perception experiment thus indicate that speech produced by the same talkers in noise is better identified in severe listening conditions than the speech produced in quiet.

3. Aims of the research

The articulatory variation of vowels in production and the resolving of this variation in perception form the general framework of research for this thesis. Perceptual prototypes are assumed to play a role in the resolving of allophonic variation, and also in the articulation, in the sense that the prototypes constitute the articulatory targets for production. The role of prototypes is primarily examined in the perception and production of Finnish short and long vowels. German, a linguistically different language but with a similar vowel system, is studied for comparison. A new concept of the *weighted prototype* (Pω) is introduced as an alternative measure of prototypicality, and its usability is compared with absolute prototypes (Pa) and category centroids (Pc) in an experiment on Finnish and German vowels.

3.1 Research hypotheses

On the basis of the earlier research reported in Chapter 2, the following *general assumption*s are made:

> A1. In quantity languages, both the phonological quality and quantity play a role in the message coding. In Finnish, according to the identity group interpretation, the quality and quantity do not interact in such a manner that would influence the phoneme interpretation, but rather, a change in either quality or quantity is sufficient to lead to a change in word meanings (e.g., in minimum pairs such as *tiili - tyyli, tili - tiili*[24]).
>
> A2. Owing to allophonic variation, the internal structures of phoneme categories are not homogenous in terms of perceived quality and, therefore, the best category exemplars of a given category can be experimentally found, and they are denoted as the absolute prototypes, Pa (or their estimates, Pa_{est}).
>
> A3. There is inter-individual variation in the loci of the absolute perceptual prototypes that may be attributed to differences in speaking environments during the language acquisition, differences in hearing or categorization performance, or experimental reasons, e.g., unfamiliarity to synthetic speech. There may also exist intra-individual variation, and several candidates for absolute prototypes within a category may evolve as individuals evaluate several stimuli as the best within the given rating scale.

[24] Word meanings: tiili "brick"; tyyli "style"; tili "account".

A4. The perceptual space shrinks in the vicinity of prototypes, which is manifested in the perceptual magnet effect (PME), i.e., close to the prototypes, the JND is larger than close to the non-prototypes.

The *hypotheses of the research reported in this thesis* are as follows:

B1. Weighted prototype is a better and less varying measure than the absolute prototype since it takes into account all stimuli identified as belonging to a particular category with a sufficiently high certainty (>70%) and rating score (6-7 on scale 1-7). The level of certainty and the rating scores are used as the weighting factors.

B2. The perceptual prototypes of Finnish vowels are solid and do not vary in response to the segmental length (physical duration) although the produced short and long vowels may exhibit such variation.

B3. The variation in the prototypes is presumably smaller than the variation in the production, given that the perceptual prototypes guide the articulation, i.e., act as articulatory targets. Further, if the perceptual prototypes guide the articulation, the result and variation of the articulation do not remarkably depend on noise masking, apart from the known Lombard effect.

B4. The loci of the arithmetic category center (Pc) differ from those of the absolute prototypes (Pa), especially in the case of corner vowels due to the hyper-articulation effect. The differences from weighted prototypes (Pω) should be smaller.

B5. The perceptual vowel spaces of two different and unrelated languages that have similar vowel systems (in IPA notations) with an equal number of vowels should manifest themselves in the loci of prototypes that are spread evenly and similarly in both languages.

These issues are addressed in Studies I–IV with the specific research questions (Q1–11) as explained below.

3.2. Specific research questions

Study I

The concept of the weighted prototype is introduced in this study for the first time. It is assumed to be a more concise measure of prototypicality than the absolute prototype is. In this study, the perception and production of mid-front Finnish vowels are compared with the same subjects serving both as listeners and speakers. Especially,

Q1. Are there differences in the perception and production of Finnish vowels in terms of the F1 and F2 formant frequencies of Pa, Pω, and the articulated vowels?

Q2. What is the inter-individual variation in Pa, Pω, and articulation?

Q3. Does either the Pa or Pω act as the articulatory target in vowel production?

Study II

This study of vowel perception concentrates on the interaction of vowel quality and quantity in Finnish, which is an example of an extreme quantity language. The interaction is studied in the framework of the initial auditory theory of vowel perception (Rosner & Pickering 1994), and the null hypothesis (H_0) was formulated on the basis of the identity group interpretation of the Finnish quantity opposition (Karlsson 1983). Especially,

Q4. Does physical vowel duration affect the perception of vowel quality in terms of the location of category boundary, boundary width, and the location and rating of vowel prototypes?

Q5. Does vowel duration affect the perception process in terms of reaction times when categorizing quality differences?

Study III

The production of the short and long Finnish vowels is studied in the presence of multi-talker babble and pink masking noises, and without a noise mask. It is assumed that masking the production by noise may cause (forced) hyperarticulation (Johnson et al. 1993a; Beckford Wassink et al. 2007), and possibly accentuate the reported minor quality differences between produced short and long Finnish vowels. Further, if prototypes act as articulatory targets in a fire-and-forget manner, the differences between the with and without noise situations should be minimal. Especially,

Q6. Are the short Finnish vowels more centralized in the F1–F2 space than the long vowels, as suggested in some earlier studies (Iivonen & Laukkanen 1993)?

Q7. Does noise masking cause differences in the production of the short and long Finnish vowels?

Study IV

Three different prototypicality measures, the absolute prototype (Pa), the weighted prototype (Pω), and the category centroid (Pc), are compared in this perception study on two different languages, Finnish and German, which have a similar vowel system but are linguistically unrelated. The type of distribution of the prototypes within a category is further investigated since earlier studies have shown large individual variation in the formation of internal structure of the Finnish /i/ category (Aaltonen et al. 1997, Study II). The results of the perception experiment of this study are compared with Finnish and German production data published earlier. Especially,

> Q8. What kind of individual differences may exist in the perceived vowel categories among native speakers of a given language?
>
> Q9. Are there perceivable differences between the various prototype measures (Pa, Pω and Pc), and how are they distributed within a category?
>
> Q10. What kind of similarities and dissimilarities may exist in the category structure and prototypes of Finnish and German?
>
> Q11. Do the perception results, in terms of the different prototype measures obtained in this study, compare to earlier published production data, where the listeners and talkers are not the same as in Study I?

4. Materials and methods

This chapter gives an overview of the materials and research methods used in the experimental part of the thesis (Studies I–IV). Research materials cover the subjects participating in the experiments, the synthesized vowel stimuli, the noise masks used in Study III, the raw data repositories of stimulus responses, and the audio records of produced carrier words. Research methods cover the instrumental setup and the experimental procedures, raw data analyses and statistical analyses. The equation for calculating the weighted prototype is presented in Chapter 4.4. An overview of the methods and instrumentation generally used in speech perception research is given in (Eerola 1993).

4.1. Subjects

Studies I, II and III were typical behavioral laboratory studies used in speech research. The same group of subjects (N=14) was used in Studies I and III but not everyone participated in all of the experiments. This group represented modern educated Finnish spoken in South-West Finland. In Study II, an entirely different group of subjects (N=16) was used. Their dialectal background represented a larger variety of Finnish dialects, although dialect was not used as an explaining variable of the results. In Study II, gender was employed as an independent variable in order to investigate whether there are differences in categorization and goodness rating between male and female listeners when using stimuli synthesized with a male voice. In the other studies, the results were not reported separately by gender.

Study IV was an observatory study in which the data were obtained from larger subject groups (N=68 for Finnish, N=18 for German). In this study, also the dialectal background of the subjects was used as an explaining variable for the results of the Finnish subjects, as it was facilitated by the sample size.

In Studies I, II and III, the listeners were screened for possible hearing impairments by means of a standard audiometric test. In Study IV, only listeners with no self-reported hearing impairments were included. The subjects were typically university students and staff members who spoke modern educated Finnish or German. A summary of the subjects is given in Table I.

Table I. Summary of subjects in Studies I–IV. FI = Finnish, DE = German

	Study I	Study II	Study III	Study IV
N, total	14	16	10	68 (FI) / 18 (DE)
N, male	7	9	7	35 (FI) / 6 (DE)
N, female	7	7	7	33 (FI) / 12 (DE)
Mean age (years)	22	27	22	25.5 / 26.2
Age range (years)	17-31	19-44	17-31	N / A
Mother tongue	FI	FI	FI	FI / DE
Dialect background	South-West	Varied	South-West	See article
Hearing	Screened	Screened	(Screened)	Reported

4.2. Vowel stimuli and noise masks

Vowels are voiced sounds that are produced when the vocal tract modifies the air flow generated by glottal excitation. In order to explore the perceptual vowel space, parameterized speech sounds are needed, and speech synthesis is the only way to generate a well controlled and wide set of varying stimuli. In the vowel perception experiments for Studies I, II and IV, the stimuli were synthesized using Klatt speech synthesis (Klatt 1980). Klatt synthesizers are based on parametric formant synthesis, and the typically approx. 40 different synthesis parameters are software controlled in a timely manner, i.e., at certain time points, a change in the parameter mix causes the desired change in the acoustic output. Klatt synthesis is well established and widely used in speech perception studies (Klatt 1987). In speech synthesis, the excitation in a speech frame is parameterized by using the fundamental frequency (f0), voicing decision, and overall amplitude. All the vowel stimuli used in the experiments discussed in this thesis were synthesized for male voices[25]. For making the voicing most natural, a rise-fall pattern of f0 is typically used; for example, in Study IV, the pitch initially (0–120 ms) rose from 100 Hz to 120 Hz and then fell to 80 Hz during the rest of the stimulus (120–350 ms). Both the physical frequency scale (Hz) and the psychoacoustic mel scale (Stevens et al. 1937) are used in the experiments: The synthesis parameters are fed in Hz, whereas, the perceptually equal steps, for example, are calculated in mels and converted back to Hz by using the mel-Hz and inverse equations (Lindsay & Norman 1977; p. 163). Critical bandwidths (Hz-Bark conversion) are used when appropriate, e.g., in comparing the results to earlier published data (Zwicker & Terhardt 1980; Traunmüller 1990). A summary of the vowel stimuli used in the experiments is given in Table II, and the more

[25] The sparser harmonic structure of female and child voices may affect the perception of vowel prototypes.

detailed synthesis parameters are described in the methods sections of the corresponding study reports. Multi-talker and pink noise masks were used in Study III (for details, see the report).

4.3. Experimental procedures and data analysis

In the vowel identification and goodness evaluation experiments of Studies I, II and IV, computer controlled stimulus delivery and response recordings were used. Various equipment and control software were applied in the different experiments, as explained in Table II.

In Study I, the EMFC (Multiple Forced Choice listening Experiment) tool of Praat software (Boersma & Weenink 2009) was used to randomize and deliver the stimuli and collect the response data of identifications and goodness ratings. The raw data of responses was fed into Excel spreadsheets, and the identification rates and goodness mean scores were calculated for each stimulus. In Study II, the NeuroStim software (NeuroScan Inc.) was used to deliver the stimuli and collect the categorization data and the corresponding reaction times (RT) from the stimulus onset. NeuroStim was interfaced to a PC via a special response keypad that enabled accurate RT measurements. The Probit non-linear curve fitting method (Bliss 1934; Finney 1944) available in the SPSS statistical software was applied for fitting the categorization data, and for determining the category boundary (CB) and boundary width (BW) from the individual categorization data. For collecting the goodness evaluation results (scale 1-7), a form sheet was used, and the results of each listener were manually fed to the S+ statistical analysis software. In the other studies, an ordinary mouse was used as the response device, and no RTs were recorded. In Study IV, the procedural setup of the Turku Vowel Test (TVT) was used (Savela 2009). TVT is an experimental method and database for structured observational studies conducted in a laboratory setting or over the Internet. The TVT has been used for exploring the vowel categories in multiple languages; for example, Raimo and co-workers (2002) used it for the determination of the subsets of the most solid vowel categories in 10 languages, and Savela (2009) showed data for 13 languages. The TVT also provides goodness ratings for the identified vowels.

Standard parametric and non-parametric statistical analysis methods were used, as explained in the study reports. In Studies I, II and IV, those stimuli were denoted as absolute prototypes (Pa) that received the highest goodness scores within a category and differed significantly from the scores of non-prototypes (NP). In Studies I and IV, weighted prototypes (Pω) were calculated according to Equation (3), and in Study IV, the centroid (Pc) was the arithmetic mean of all stimuli identified as belonging to a particular category.

In Studies I and III, the subjects were asked to utter the given carrier words (e.g., /tili/ - /tiili/), each word five times successively, using their normal speech style (in Study III, first without the noise mask, and then with the two noise masks). The recordings were carried

out in an acoustically dampened room by using a high quality microphone that was connected via an amplifier to a PC. The recording level was adjusted so that dynamic range of the analog-to-digital converter was optimized without clipping the signal when it was at its loudest. Praat software was used for both the recordings and analysis. The f0, formants F1–F4, and vowel durations were analyzed from the steady state part of the target vowel within the word.

Table II. Overview of the stimuli, noise masks, stimulus delivery, response collection, recordings, experimental setups, and data conversion and analysis in Studies I–IV.

	Study I	**Study II**	**Study III**	**Study IV**
1. Stimuli				
N of vowels	4 (i, y, e, ö)	2, (i, y)	8 (all)	8 (all)
N of vowel variants	46	19		386
Synthesis method	Klatt, serial	Klatt, parallel		Klatt, parallel
2. Noise masks				
Type of noise			MTB, pink	
Mask duration			Continuous	
Noise levels (SPL)			92 dB (MTB), 83 dB (pink)	
3. Stimulus delivery				
Stimulus control SW	Praat	NeuroStim	Praat	TVT, Java
Playback dynamics	16 bit	12 bit	16 bit	PC sound board
Playback rate	44.1 kHz	10 kHz	44.1 kHz	44.1 kHz
Headphones	Sennheiser PC161	Ear-Tone 3A	Sennheiser PC161	PMB K 800
Play level (SPL)	74 dB (A)	75 dB (A)	83 dB, 92 dB	Adjustable
Calibration	SPL meter	Artificial ear	SPL meter	No
Stimulus repetitions	10 x 46	15 x 19	5 recordings	Self-paced
Interstimulus interval (ISI)	Self-paced	2 s / self-paced		Self-paced
4. Responses				
Identification	Mouse	Reaction pad		Mouse
Goodness rating	Mouse, 1-7	Form sheet, 1-7		Mouse, 1-7
Reaction times		From onset		
5. Recordings				
Control SW	Praat	NeuroStim	Praat	TVT
Microphone	AKG D660S		AKG D660S	
Recording dynamics	16 bit ADC		16 bit ADC	
Sampling rate	44.1 kHz		44.1 kHz	
Booth ambient noise	27 dB	< 40 dB	27 dB	Language center
6. Analysis				
Raw data conversion	Praat to Excel	NeuroStim to S+ and to Excel	Praat to Excel	TVT to Excel
Curve fitting		Probit (SPSS)		
Statistical analysis SW	SPSS	S+, SPSS	SPSS	SPSS
Main measures	Pa, Pω, F1-F2	Pa, CB, BW, RT	D, F1-F4	Pa, Pc, Pω

4.4. Weighted prototype

The weighted prototype Pω(F1,F2) is a novel measure of prototypicality introduced in Study I and compared with other prototype measures in Study IV.

Pω(F1,F2) is calculated using Equation (3)

$$\mathbf{Fi} = (a_1 r_1 Fi_1 + a_j r_j Fi_j + ... + a_n r_n Fi_n) / (a_1 r_1 + a_j r_j + ... + a_n r_n) \quad (3)$$

where:

Fi = weighted formant frequency, i = 1, 2;

Fi_j = formant i of stimulus j, j = 1, 2, ..., n;

a_j = evaluation score (1–7), j = 1, 2, ..., n;

r_j = identification consistency (0.7–1.0), j = 1, 2, ..., n;

n = number of stimuli identified as category members.

Pω(F1,F2) represents a point in the F1–F2 space, either in Hz or mel scale, that is obtained by weighting the F1 and F2 values of each stimulus identified as a category member by the goodness rating value (a) and unanimousity (r). The identification consistency r is 1 when only those stimuli are included that are identified 100% as belonging to a category. Often also stimuli with 70–99% identifications are included; the r-factor weights these stimuli less than the stimuli with 100% identification. Scale 1–7 is used for the goodness evaluation, but other scales (1–5 or 1–10) could be used as well.

5. Summary of experiments

The experimental part of this thesis comprises a series of four experiments in which vowel prototypes are examined by means of behavioral psychoacoustic measurements and comparison with vowel production in quiet and in noise. Study I explores the variation in vowel production and perception by Finnish speakers, Study II focuses on the interaction of vowel quantity and quality in Finnish, Study III investigates the influence of noise masking on articulatory variation in Finnish, and Study IV compares the perception of artificial speech sounds among Finnish and German listeners.

In the experiments, native speakers of Finnish and German identified and evaluated the goodness of synthesized vowels representing either the entire vowel space or selected subareas of the space. The production experiments involved Finnish speakers only, but earlier reported production data were used for comparison with the perception results of German speakers.

5.1. Study I: Vowel prototypes and vowel production

The aim of Study I was to compare the perception and production of the mid-front Finnish vowels /i/, /e/, /y/ and /ø/ in fourteen Finnish-speaking subjects. This study introduces, for the first time, a new concept of the weighted perceptual prototype (Pω), which is compared with the estimated absolute prototypes (Pa) obtained in the perception experiment. Another aim was to test whether the absolute or weighted prototypes serve better as articulatory targets for production, and whether the articulatory targets are better achieved in the long or short versions of the vowels (for perceptual prototypes, the duration represented long vowels). The purpose of this study was to test the hypothesis that the acoustic features (formants F1 and F2) of an individual's perceptual vowel prototype correlate with the same acoustic features of the produced vowel, and to compare the Euclidean distances of perceived and produced vowels in the F1–F2 space. Additionally, since Finnish is an example of an extreme quantity language, the effect of vowel duration on the articulated vowel quality was investigated. It was assumed that the long vowels better achieve the articulatory targets (prototypes), in other words, the distances from the prototypes are smaller than in the case of short vowels.

Results and discussion

The results of Study I show that the listeners were able to accurately identify the stimuli as belonging to one of the given four Finnish vowel categories, and also that they perceived significant quality differences between those vowel variants that they had identified as belonging to one and the same category (Table III, Figure 9). The pair-wise differences of goodness scores between the highest ranked and the lowest ranked stimuli

were highly significant (t-test, <0.001). There were also significant differences in the goodness rating values between the vowels: the goodness ratings of non-labial vowels were higher than those of the labial vowels. These quality differences also reflect the commonness of various vowels in Finnish (KOTUS 2008); the more frequently occurring /i/ and /e/ received significantly higher goodness rates than the uncommon /y/ and the rare /ø/. This finding suggests that there may be a relation between the perceived goodness and the prevalence of a vowel in a language: the memory representations of the more common vowels may be different from those of the rarer ones. Assuming that they are learned differently during the language acquisition process, the more common vowels would have more stable or sharper memory traces (Blomberg 1993; Huotilainen et al. 2001).

Table III. The mean (N=14) goodness rating values (scale 1–7) of the absolute prototypes Pa and non-prototypes NP for Finnish vowels /e/, /i/, /y/ and /ø/. Standard deviations are given in the parentheses. (Unpublished data from Study I.) Prevalence (%) refers to the relative commonness of each vowel in text corpuses (KOTUS 2008).

	/e/	**/i/**	**/y/**	**/ ø/**
Pa score (SD)	6.21 (0.93)	6.07 (0.75)	5.71 (0.83)	5.74 (0.80)
NP score (SD)	4.34 (1.04)	4.24 (0.89)	3.54 (1.09)	3.45 (0.99)
Prevalence %	12.6	19.6	2.1	0.1

The mean differences between the two methods of obtaining the perceptual prototypes (Pa and Pω) are 9–25 mels for F1 and 11–36 mels for F2. As illustrated in Figure 10, the weighted perceptual prototypes are more centralized in the F1–F2 vowel space than the absolute prototypes (Q1)[26]. The Wilcoxon signed ranks test showed that there are significant differences between the estimated absolute and the center-of-gravity type (i.e., weighted) prototypes ($p < 0.05$) in categories /e/, /i/ and /ø/, but not /y/. The obtained differences are smaller than or of the order of difference limens (DL) of frequency, and hardly noticeable, but the individual variation is considerably smaller in the weighted prototypes than in absolute prototypes (Figure 10).

[26] The specific research questions Q1-Q11 of Chapter 3.2 are referred to in the text as (Q1, Q2,...Q11).

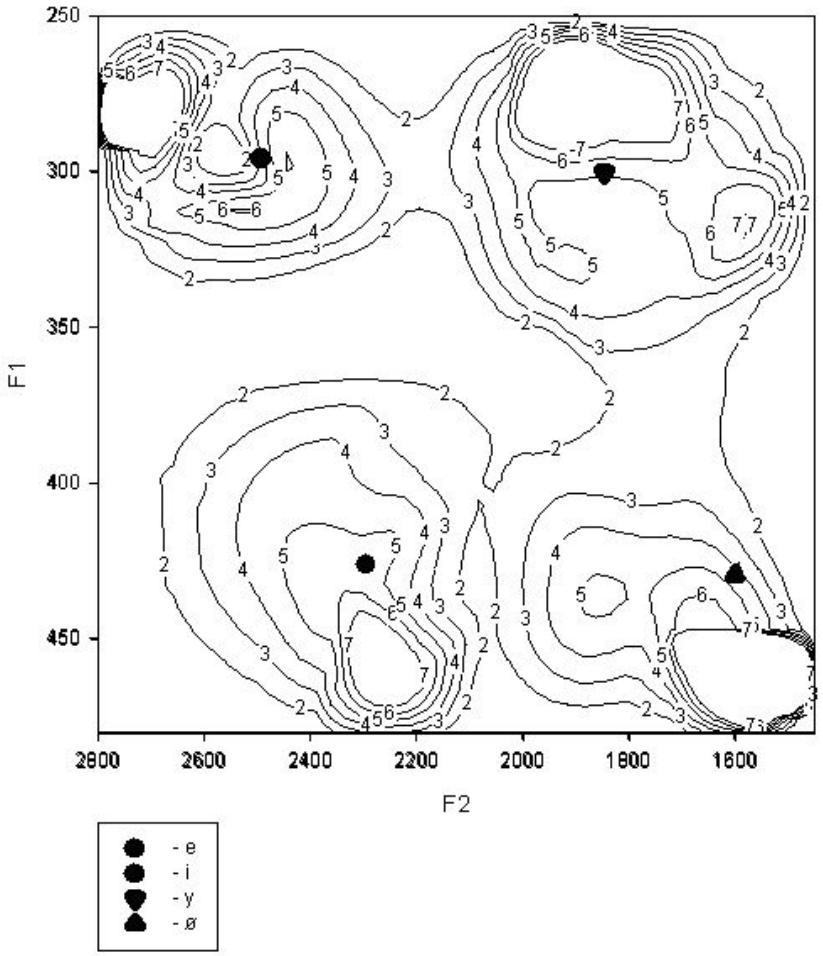


Figure 9. Mean goodness ratings of the Finnish vowels /i/, /y/, /e/ and /ø/ of 14 listeners on scale 1–7 (the contours represent the results for ratings 2–7). The symbols represent the weighted perceptual prototypes (Pω) of the categories. The absolute prototypes (Pa) lie within the contours indicating the highest (7) goodness scores. F1 and F2 formant frequencies are in Hz. Illustration of the raw data of Study I.

In the production experiment, the durations of the short vowels were 120 (SD 32) ms for /e/, 125 (SD 24) ms for /ø/, 103 (SD 25) ms for /i/, and 118 (SD 36) ms for /y/, and of the long vowels 316 (SD 51) ms for /e:/, 326 (SD 71) ms for /ø:/, 301 (SD 59) ms for /i:/, and 329 (SD 58) ms for /y:/ (Q2). The absolute values of both the short and long vowels are longer than reported by, for example, Suomi (2006), but in line with the results for vowels in isolated words as reported by Kukkonen (1990). The durational ratio 1:2.7 between the short and long segments of the four vowels is in line with the earlier investigations of Lehtonen (1970) and Wiik (1965).

The Euclidean distances in the F1–F2 space between the short and long vowels produced by the 14 subjects were 29 (SD 16) mels for /e/, 42 (SD 31) mels for /ø/, 49 (SD 24) mels for /i/, and 51 (SD 44) mels for /y/ (Q2). The distances, especially those for /e/ and /ø/, are of the order of the combined F1 and F2 difference limens (DL) reported in literature (Mermelstein 1978; Hawks 1994), and suggest that the quality differences between the short and long Finnish vowels /e/ and /ø/ uttered in normal speech style are hardly noticeable. The differences for the high vowels /i/ and /y/ are larger and may exceed the DL. The ANOVA showed no effect of the vowel quantity on the F1 or F2 values across the four vowel categories. This result is in line with the identity group interpretation of the phonological quantity opposition of Finnish, i.e., in spoken Finnish, the duration of the long segments of vowels is of the order of the duration of two short segments, and that the psychoacoustic spectral composition of the short and long segments is essentially the same in terms of noticeable Euclidean distances (Karlsson 1983; Suomi et al. 2006).

The Euclidean distances in the F1–F2 space between the produced vowels and their weighted category prototypes were 113 (SD 34) mels for short and 116 (SD 24) mels for long vowels, and between the produced vowels and their absolute prototypes 127 (SD 36) mels for short and 125 (SD 29) mels for long vowels. These differences are of the order of 3–4 DLs. The results indicate that the productions are slightly closer to the weighted than the absolute prototypes, and are clearly more centralized or lower than either of the prototypes, as illustrated in Figure 10 (Q3). At an individual level, the F1 and F2 values of the weighted perceptual prototypes correlated significantly with the F1 and F2 values of the produced short and long vowels. At the group level (N=14), the produced vowels were always closest to the weighted prototypes of the category in question, and productions were generally more central and/or lower than the perceptual targets. The individual variation of the produced vowels is remarkably larger than that of the perceptual prototypes, and slightly smaller in short than in long vowels (Q2).

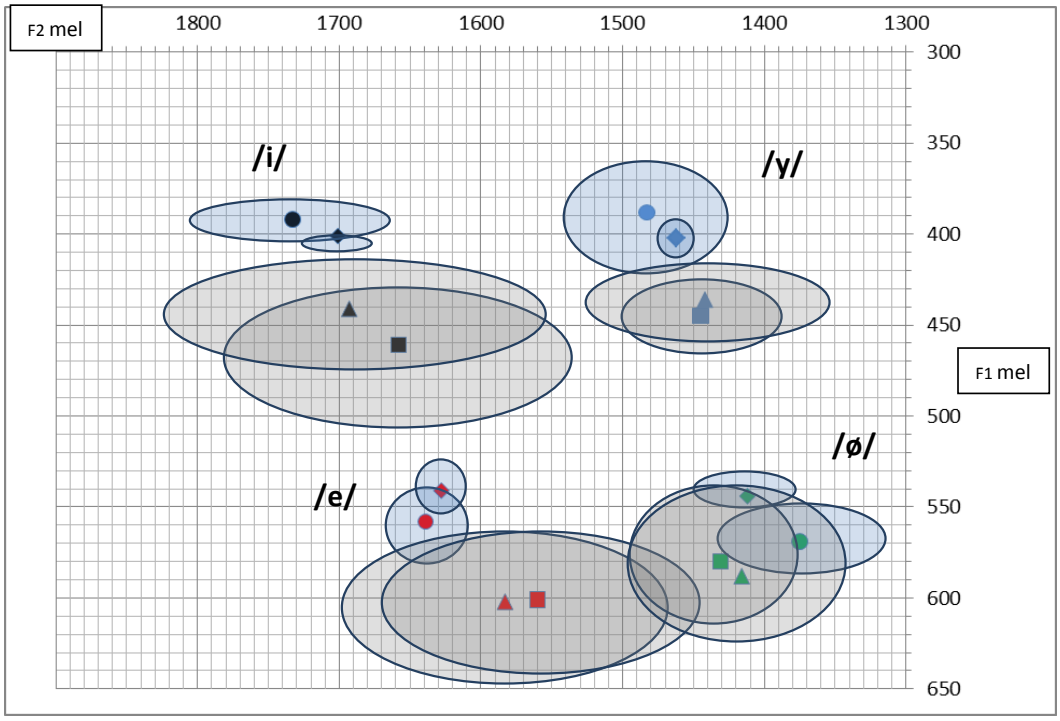


Figure 10. The mean values (N=14) of absolute (filled circle) and weighted (diamond) prototypes, and the produced short (square) and long (triangle) Finnish /i/, /y/, /e/ and /ø/ vowels in F1 and F2 formant space (mel scale). Standard deviations are shown by the elliptical circles. Illustration of data from Study I, Table 2 and Table 3.

5.2. Study II: Effect of vowel duration on categorization and prototypes

Study II explores the interaction between vowel quality and quantity. Vowel quality refers to the perceptual attributes of vowel sounds on the basis of which listeners are able to identify vowels as belonging to distinctive categories. In quantity languages, such as Finnish, not only the spectral quality of phones but also their duration is of importance in making judgments of phonological categories and thereby perceiving the meaning of words correctly. Finnish is an example of a contrastive quantity language where both vowels and consonants may occur independently of each other in short or long oppositions, without the quantity being bound to the word stress. According to the identity group interpretation, the long segments of vowels or consonants consist of two successive and identical short segments (Karlsson 1983). This is generally accepted in textbooks on Finnish phonetics (Suomi et al. 2006; Iivonen & Tella 2009) as the de facto interpretation of the phonological quantity opposition in the Finnish language. Within the AVS framework (Rosner & Pickering 1994), the identity group interpretation of Finnish quantity opposition would mean that the LEVIs and D in the AVS are independent, meaning that the perceptual space is orthogonal in that sense. In Study II, this issue is

explored by focusing on the relationship between E2 (as a function of F2) and D of the Finnish high-front vowels /y/ and /i/. This vowel pair was selected because it allows a gradual shift between the qualities of /y/ and /i/ by varying only the E2 while keeping the E1 and E3 constant (Aaltonen & Suonpää 1983; Aaltonen et al. 1997).

In Study II, the conservative null hypothesis (H_0) of the first research question (Q4) was formulated according to the identity group interpretation: short and long vowels are perceived similarly in terms of their spectral quality and they have similar prototypes. The alternative hypothesis (H_1) was based on reported minor spectral differences in the *produced* short and long Finnish vowels (see Table 2 in Study II), and hence, the assumption that these differences may be reflected in the perception of the short and long vowels.

For the second research question of Study II (Q5), it was assumed that stimulus duration does not affect the perception process itself, i.e., readiness in identifying quality differences between short and long vowels is similar (H_0). Study II consisted of two separate experiments; Experiment I for vowel identification and Experiment II for vowel goodness rating for the individual /i/ category members identified in Experiment I. The time span between the experiments was from days to a week.

Results and discussion

In Experiment I, large individual variation was found in the categorization, but the category boundary (CB) F2 value and the boundary width (BW) were independent of duration at the group level, suggesting that quantity does not affect the category formation between /y/ and /i/ (Q4). The categorization results at different durations are shown in Figure 11. Further, it was found that the listeners were, in general, able to make their judgment within one critical band rate (BW/CBR < 1.0) at all durations, and that the BW was of the order of 2-3 stimulus steps (60-90 mels). This is in line with previous findings (Aaltonen et al. 1997). Only one significant gender difference was found in Experiment 1: the BW differed significantly between male and female listeners for the duration of 50 ms (166 Hz for males and 323 Hz for females). Based on the experimental data, there is no good explanation for such a large difference, especially since the hearing of the listeners was tested by using clinical audiometric tests. One may speculate that women differ from men in their ability to categorize short stimuli on the boundary area when the stimuli are synthesized for a male voice. Another explanation may be the rather limited sample size (N=7 for female listeners).

The reaction times were 0.25–0.30 s longer at the boundary than within a category. The difference was highly significant ($p < 0.001$) for all durations and listeners, and in accordance with the earlier findings concerning categorical perception. For the purpose of comparing the measured reaction times to stimuli of varying lengths, two normalized RT ratios were formed for each listener and each duration: $t_a = t_{CB} / t_{tot,}$ and $t_b = t_{CB} / t_{cat}$.

The former (t_a) is the ratio of the RT at the CB (t_{CB}) to the overall mean RT (t_{tot}), and the latter (t_b) is the ratio of the RT at the CB to the mean within-category RT of the /y/ and /i/ category stimuli. Both normalized measures were significantly dependent on duration, and pair-wise comparisons showed that the categorization was most difficult at 100 ms, that is, a duration that falls between a typical short and long Finnish vowel (Lehtonen 1970; Ylinen et al. 2006;). The result suggests that the quality of vowels with a duration that represents the borderline between the short and long vowels may be perceived differently and processed at a slower rate than the vowels representing more clearly either the short or the long Finnish vowels (Q5).

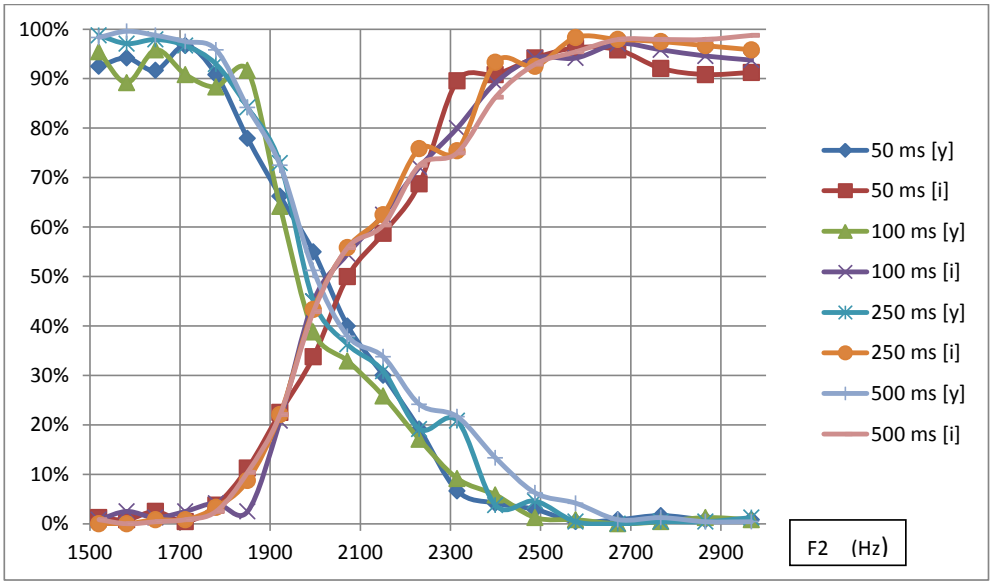


Figure 11. Grand averaged (N=16) categorizations of 19 synthesized vowel stimuli to [y] and [i] phones (y-axis: categorization %) as a function of the second formant (F2, in Hz) at stimulus durations of 50 ms, 100 ms, 250 ms, and 500 ms. The F2 continuum (x-axis) spans from 1520 Hz (1290 mels) to 2968 Hz (1830 mels) in steps of 30 mels (e.g., four stimulus increments correspond to 260 Hz at 1520 Hz, but to 367 Hz at 2400 Hz). Illustration of raw data of Study II.

In Experiment 2, three different types of curves emerged for goodness ratings depending on the location of the highest ranking values on the F2 scale within the /i/ category: the "hill" type (prototypes in the middle of category, illustrated in Figure 12, the "down" type (prototypes near the category border), and the "up" type (prototypes at the end of the continuum). When the results were summed together, no duration-dependent main effect on stimulus duration was found either for prototypical /i/ rating scores or the F2

values of the prototypes (Q4), but pair-wise comparisons showed that there was a significant difference between the goodness ratings for durations of 50 ms and 100 ms.

To summarize the results of Experiment 2, the F2 frequencies of the highest scoring (prototypical) stimuli were statistically independent of duration, suggesting that the phonological quantity categories do not influence significantly the perception of quality differences within a particular vowel category (Q4). The RTs were generally shorter within the category, and had a minimum around the stimuli with F2 of 2672 Hz - 2767 Hz, but these values were 30–60 mels higher than the prototypical stimuli measured in Experiment 2. This result indicates that even if there are differences between the within-category stimuli, as measured by RTs in a categorization task, the stimuli showing the shortest RTs are not necessarily identical with the prototypical stimuli emerging in a dedicated goodness rating setting, but rather, with the stimuli that lie farthest from the category border.

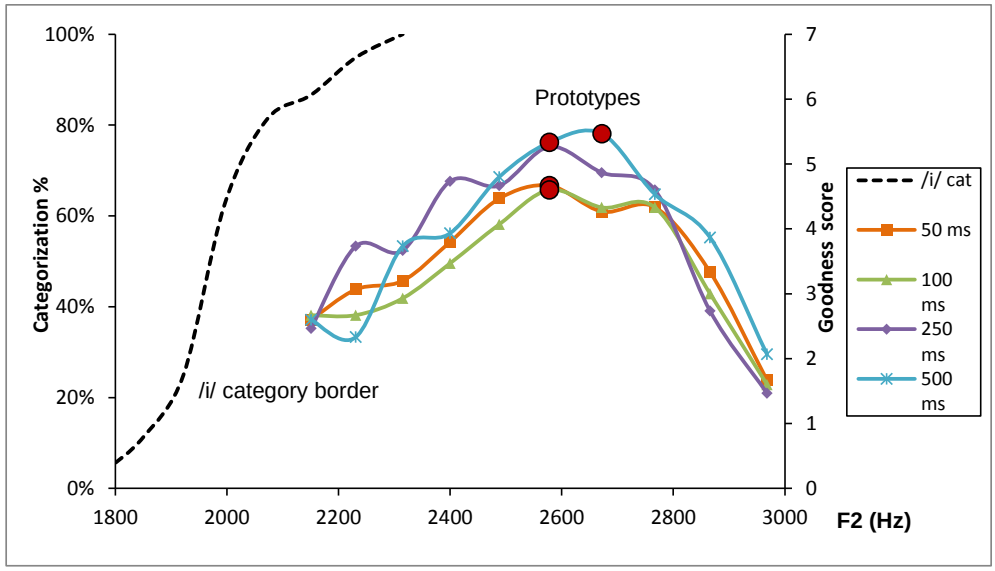


Figure 12. Example of "hill" type goodness ratings (scale 1–7) of stimuli within the individual /i/ category of Listener 2 with stimulus durations of 50 ms, 100 ms, 250 ms, and 500 ms. The /i/ category border is shown as a dotted line. The highest scoring stimuli (perceptual prototypes, marked as circles) are at 2578 Hz (50 ms), 2578 Hz (100 ms), 2578 Hz (250 ms), and 2672 Hz (500 ms). Stimulus step size is 30 mels. [27]

The main results of Study II leave the null hypothesis valid: In Experiment 1, duration had no significant effect on the location and width of the /y/–/i/ category boundary (on the F2

[27] Reprinted with permission from Elsevier.

axis), and in Experiment 2, duration had no significant effect on either the F2 value or the goodness rating value of the prototypical /i/ within the individually determined /i/ categories (however, the quality difference between 50 ms and 100 ms was significant). In other words, the listeners' category boundaries between /y/ and /i/, and the /i/ prototypes (in terms of F2 frequency) were not demonstrably dependent on the stimulus duration. Another important finding in Experiment 1 was that the listener's gender had no effect on the location (F2 frequency) of the category border between /y/ and /i/, although statistical analysis revealed that the category boundary area (BW) was narrower in male listeners at 50 ms. The results of this study indicate that two key characteristics of the initial auditory theory of vowel perception (Rosner & Pickering 1994), namely, the local effective vowel indicator (LEVI) E2 (approximated by the auditory Hz-to-mel frequency conversion of F2) and factor D (representing here directly the physical duration d), are not seemingly dependent on each other, thus suggesting that the AVS is orthogonal for these two variables in the Finnish vowel space of /y/ and /i/. Another noteworthy result of this study was that stimulus typicality (quantity) affects the quality categorization process, but not its end result.

5.3. Study III: Vowel production in noise

Study III further explores the reported minor quality differences between produced short and long vowels (Wiik 1965; Kukkonen 1990; Iivonen & Laukkanen 1993) across the entire Finnish vowel system in two different noise masking conditions, and without any noise mask. It was assumed that noise masking may cause (forced) hyper-articulation (Johnson et al. 1993a) and possibly accentuate the reported minor quality differences between short and long Finnish vowels. Two different types of masking noise were used: multi-talker babble noise at 92 dB SPL, and pink noise at 83 dB. The Lombard effect has been reported to cause measureable differences in vowel intensity and duration, and also in formant frequencies: ambient noise elevates the speech amplitude by 5–10 dB, increases word durations by 10–20%, and increases the F1 and F2 frequencies, thus causing a shift in the vowel space (Lane & Tranel 1971; Van Summers et al. 1988; Castellanos et al. 1996).

Results and discussion

The results of Study III are illustrated in Figure 13. It was found that, in the non-masking situation, the short and long vowels differ in terms of F1 and F2 between the categories with the differences being largest between /u/ and /u:/. The mean individual distance in the F1–F2 space between the long and short vowels without noise masking was 62 mels across all vowel categories. Variation was found between vowel categories: /e/ and /ø/ had distances of 29–39 mels and no centralization tendency was observed, whereas /o/, /u/ and /æ/ showed clearly larger distances, up to 128 mels. The differences in the F1 and F2 values were significant for /i/ in F1 ($p = .013$) and F2 ($p = .005$), for /e/ in F2 ($p = .012$),

for /y/ in F1 (p = .012), and for /u/ in F1 (p = .012). Noticeable centralization of the short vowels was found especially in /i/, /u/, /o/, /a/ and /æ/ (Q6). This is in accordance with the earlier reported findings (Wiik 1965; p. 60; Kukkonen 1990; p. 229; Iivonen & Laukkanen 1993; p. 37; Iivonen & Harnud 2005; p. 66).

The effect of noise on the produced vowel quality was similar in both two masking conditions, and no major differences between babble and pink noise were found (Figure 13). Both noise types seem to cause higher F1 frequencies in the production of the mid-high vowels: On the average, the F1 values of the short and long vowels produced in the masking conditions are about 34 mels higher than without masking. No similar effect was found for the low vowels /a/ and /æ/. The results indicate that noise masking causes a systematic shift of F1 values in the production of mid-high Finnish vowels. By using Wilcoxon signed rank test, the differences in F1 between the quiet (Q) and noise (B=Babble, P=Pink) conditions were highly significant both for short and long vowels ($p<.001$).

The vowel durations were 121 ms (SD 8 ms) for short and 334 ms (SD 94 ms) for long vowels. With babble noise, the durations were 143 ms (SD 37 ms) and 349 ms (SD 76 ms), and correspondingly with pink noise, 130 ms (SD 32 ms) and 341 ms (SD 79 ms). By using Wilcoxon signed rank test, the differences in duration between the quiet (Q) and noise (B=Babble, P=Pink) conditions were highly significant for short vowels in Q versus P ($p=.002$), and in Q versus B ($p=.000$). For long vowels, the differences between the two noise conditions were only marginally significant in B versus P ($p=.039$) (Q7).

The results of this study on the production of the short and long Finnish vowels confirmed, first, the earlier findings that the short vowels /i/, /u/, /o/, /a/ and /æ/ are more centralized in the F1–F2 space than their longer counterparts. Second, the Lombard effect induced by the two different noise masks resulted in a significant increase in the duration of the short vowels, but not the long ones. The increase was larger with the louder babble noise than with the pink noise. Whether this difference was due to the higher amplitude of the babble noise or to the noise type itself, is a subject for further studies. Third, the Lombard effect resulted in an increase in the F1 of the mid-high vowels, but had no effect on the Euclidean distances of the short and long vowels. These results regarding the F1 value and the Euclidean distances are in line with earlier findings (Van Summers et al. 1988; Beckford Wassink et al. 2007). The latter study among Jamaican speakers is particularly interesting, since Jamaican Creole utilizes the phonemic vowel length in a similar manner as Finnish, which, however, is a distinctive quantity language. The vowel quality (in terms of F1 and F2) was affected similarly by the Lombard effect in both of these languages, but a clear durational prolongation of short vowels was only found in Finnish.

As a conclusion, in the light of the results of Study III, the articulations do not depend remarkably on noise masking besides the known Lombard effect.

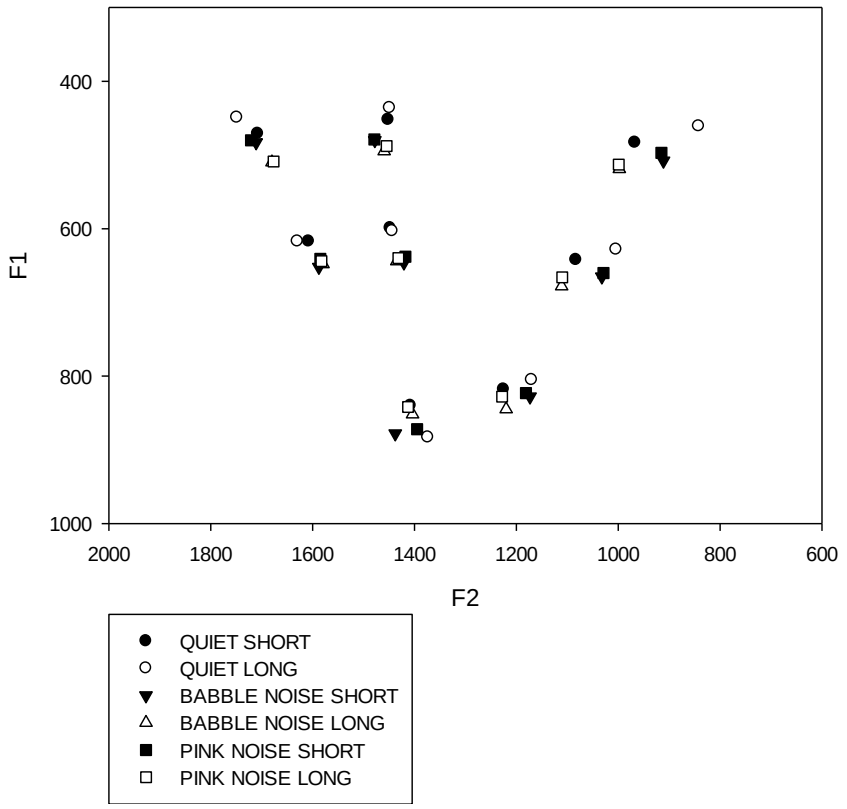


Figure 13. Grand averages of short and long Finnish vowels in the F1–F2 space (mel scale) in the two different masking conditions and without noise masking. The number of subjects varies in different categories. The categories are from top left to right down: /i/, /y/, /u/, /e/, /ø/, /o/, /æ/ and /ɑ/. [28]

5.4. Study IV: Perceptual vowel prototypes in Finnish and German

Study IV explores the areas of high prototypicality in the vowel systems of two linguistically unrelated languages, Finnish and German, the vowel systems of which, however, are phonologically comparable since both have a similar vowel system with eight vowels, /ɑ/ /e/ /i/ /o/ /u/ /y/ /æ~ɛ/ /ø/, which motivates such a comparison. The basic assumption of Study IV is that, within each phonetic category, there are subsets of one or several distinct category members that are more representative of the category than the category members in general. These subareas of high prototypicality were

[28] Reprinted with permission from Linguistica Uralica.

studied by using three different measures: the arithmetic mean (centroid) of the F1–F2 space of the category (Pc), the absolute prototype of the category (Pa), and the weighted prototype of the category (Pω), in which the stimulus formant values are weighted by their goodness rating values. If the general hyper-space effect (Johnson et al. 1993a; Johnson 2000) affects the vowel perception, the most peripheral stimuli should receive the best ratings, and consequently the loci of the absolute and weighted prototypes should differ essentially from those of the centroids.

Earlier studies (Aaltonen et al. 1997; Lively & Pisoni 1997), and Study II have indicated large individual variation in the location of the category prototypes in the F1–F2 space. Therefore, one of the aims of Study IV was to further investigate this variation by means of several different prototypicality measures, with a larger number of subjects, and in two different languages. Further, the statistical distribution of the prototypes within a category was explored since previous research includes certain indications that the prototypes are assumed to be located either evenly within the category[29] (Aaltonen et al. 1997), near the category center (Kuhl 1991; Rendell 1986), or towards the peripheral parts of the category (Johnson et al. 1993a). Finnish and German constitute an interesting language pair for comparative studies since, in general, the produced long vowels of these two languages resemble each other. According to the adaptive dispersion theory (Johnson 2000), the vowel systems of two languages with an equal number of vowels should be similar, and the research question involves possible individual differences between native speakers of a given languages (Q8).

Results and discussion

The results of the goodness ratings of the 386 stimuli clearly indicate that the categories are graded, and that the calculated individual Pωs form clusters in the F1–F2 space that illustrate subsets of stimuli with high rating scores (Figure 14). The areas of 90% identification consistency between listeners (solid line in Figure 15) match with the earlier results of Aaltonen and Suonpää (93-100% areas in Figure 7) (Aaltonen & Suonpää 1983). Of the different prototype measures, Pω had the smallest variation (in terms of CV), whereas Pc and Pa showed larger individual variation (Q9). All prototype measures showed less variation than has been reported in earlier literature, and the inter-individual variation was of the order of DL of formant frequencies. The dialect background of the two major dialect groups of Finnish subjects (Tavastian and South Western) had no effect on the vowel categories. The normality test results showed that the formant distributions of the Pω and Pc measures were normal in the vast majority of the cases in both languages. This was also the case for Pa in German, whereas only one half of the Pa formant values were normally distributed in Finnish. This may reflect the earlier finding

[29] In the study by Aaltonen et al., three prototype classes emerged: prototypes were located either close to the border, in the center, or at the periphery of the category (Aaltonen et al. 1997)

for Finnish /i/ where three distinct absolute prototype classes were found (Aaltonen et al., 1997; Study II). The ANOVA results thus showed that, in both languages, all of the vowel types were distinct from each other in terms of F1 and F2, and the different prototype measures deviated from the median value of the entire vowel grid (605 mels for F1, 1240 mels for F2) in the following order Pa > Pω > Pc, the absolute prototypes being the most peripheral (Figure 15) (Q9).

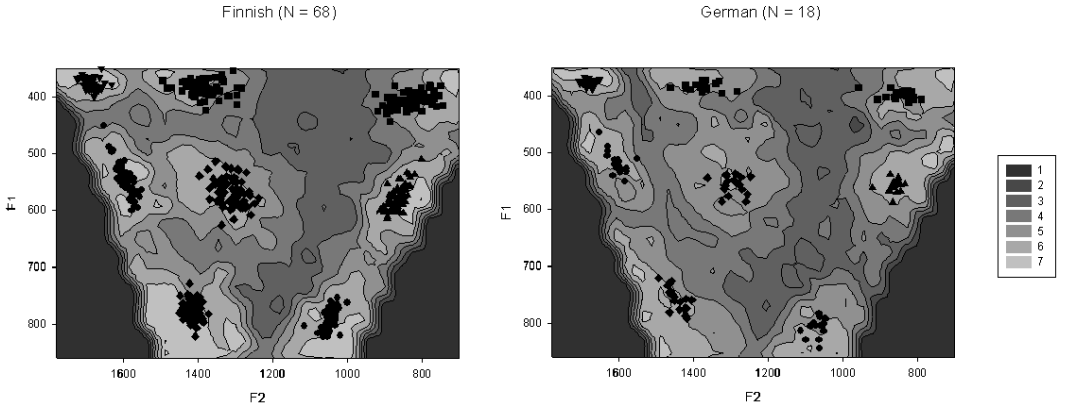


Figure 14. Goodness ratings (on scale 1–7) and individual weighted vowel prototypes Pω(F1,F2) in the F1–F2 formant space (mel scale) obtained in the identification and rating experiments in Finnish (N=68) and German (N=18) listeners.[30]

The absolute prototypes showed the largest differences between the two languages in /e/, /ø/ and /u/. This is in line with the earlier investigations on produced vowels in Finnish and German. However, in general, the vowel systems of the two languages were similar, as suggested by the dispersion theories (e.g., Becker-Kristal 2010). In terms of weighted prototypes, the Euclidean distances of corresponding categories in the two languages varied between 7–34 mels. This result indicates that the acoustical differences of the vowel systems in these two linguistically different languages are strikingly small. The largest differences were observed in the non-closed front vowels, while the other types of vowels showed minor differences, as expected on the basis of their production (Q10).

The differences observed between the various prototype measures and produced vowels, as obtained in earlier studies (Kuronen 2000; Sendlmeier & Seebode 2006) are similar to the findings of Study I, with the mean difference being approximately 110 mels across all categories in both languages: 131 mels for Finnish and 83 mels for German (Q11). In Study I, differences of 113-116 mels were found between the perception and production of the vowels /i/, /e/, /y/ and /ø/ when the same subjects participated in both experiments. When the production results of that study are compared with the

[30] Reprinted with permission from Acoustic Society of America.

perception results of the vowels in Study IV, the difference was 99 mels on the average. This gives further evidence that the differences between perception and production of Finnish vowels are of the order 100–120 mels.

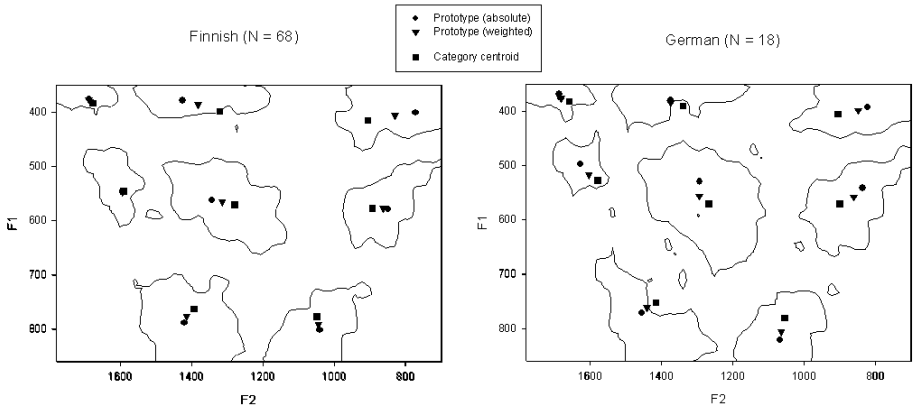


Figure 15. The loci of the absolute prototypes, weighted prototypes, and category centroids in the F1–F2 formant space (mel scale) of the Finnish and German vowels. Areas of >90% identification consistency between listeners are shown by a solid line. [31]

[31] Reprinted with permission from Acoustic Society of America.

6. General discussion

Vowel prototypes are the main focus of this thesis. In the experiments, the prototypes were determined on the basis of simultaneous identifications and goodness ratings (Studies I and IV), or successive identifications and goodness ratings (Study II). Both approaches gave similar results confirming the initial assumption that vowel categories have an internal fine structure (A2[32]), while also showing that, within a category, several separate subareas may receive the highest goodness scores (see Figures 9 and 14), the values of which differ statistically significantly from the goodness score values of the non-prototypes. This implies that the conventional absolute prototype approach may lead to inaccurate or false conclusions if the loci of the absolute prototype are defined on the basis of a single within-category subarea. The new concept of a weighted prototype takes better into account all subareas of high perceived vowel quality within a category, and due to its smaller individual variation, the weighted prototype can be regarded as a more representative candidate for the category prototype than the absolute prototype as hypothesized (B1). The weighted and absolute prototypes were compared with each other in Studies I and IV, and with category centroids in Study IV. Of the different prototype measures, the weighted prototype had the smallest variation (in terms of CV), and the measures differed from the median of the entire vowel space in the order Pa > Pω > Pc within each category (see Figure 15) (B4).

In the production experiment of Study I, the vowel quantity did not affect significantly the F1 or F2 values of the four vowel categories /i/, /e/, /y/ and /ø/. This result implies that the identity group interpretation of the phonological quantity opposition holds true in spoken Finnish, and that the articulated quality differences between short and long vowels are minor (B2). The idea that articulated long vowels would better reach the prototypical targets is not supported by the results of Study I, which indicate that the distances of the short and long produced vowels from the prototypes are nearly identical, and are of the order 115 mels for Pω and 125 mels for Pa when same subjects served as listeners and speakers. In contrast, the results suggest rather that the target is the same for both short and long vowels, but the perceptual prototype does not fully match with the articulation result (O'Dell 2003; Boersma 2009). When the comparison is made (Study IV) between the perception and articulation of different listeners and speakers, the distances were smaller (Pω 64-97 mels for Finnish /i/, /e/, /y/ and /ø/ vowels), but still on the same order, in other words, they exceeded the DL of combined F1 and F2 frequencies.

Bone conduction (BC) of one's own voice may offer an explanation for this result (Békésy 1949; Pörschmann 2000; Reinfeldt 2009). The formation of perceptual prototypes is

[32] For a description of general assumptions A1-A4 and hypotheses B1-B5, see Chapter 3.1.

based on hearing and gross averaging air conducted (AC) phones spoken by others[33] (Jusczyk 1993), whereas the construction of the articulatory sound target for own production is also based on feedback by AC and BC transmitted own speech (babbling in childhood). Could the different conduction mechanisms explain the observed differences between the perceptual prototypes and own articulations? In a recent article, Reinfeldt and co-workers report on average 12 dB higher BC than AC for front vowels /e/ and /i/, and 15 dB higher BC than AC for the back vowel /a/ in the frequency range 1–2 kHz, which is within the range of F1–F2 formants of these vowels (Reinfeldt et al. 2010). However, several mechanisms influence the perception of one's own voice, e.g., the conduction time differences and phase delays from glottal source and vocal tract to cochlea by air, by bone and by soft tissues within the skull, and consequently, more research is needed to explore the effect of BC in the formation of perceptual prototypes and articulation targets.

Study II gave three main results. First, it gave evidence in favor of the identity group interpretation of Finnish quantity opposition (Karlsson 1983) by showing that neither the perceived quality of Finnish /i/ nor the category border and width between /y/ and /i/ depended on vowel duration (A1, B2). Second, in accordance with the above finding, the results showed that two key characteristics of the initial auditory theory of vowel perception (Rosner & Pickering 1994), namely, the local effective vowel indicator (LEVI) E2, and physical duration (parameter D in the model), are seemingly not dependent on each other, thus indicating that the AVS is orthogonal for these two variables in the Finnish vowel space of /y/ and /i/. Third, the normalized reaction times to the stimuli with the duration of 100 ms showed a significant difference in comparison to the other durations of 50 ms, 250 ms and 500 ms. This could be interpreted to signify that the 100 ms stimuli do not represent properly either the short or the long Finnish vowels, and consequently, the normalized reaction times at the boundary of the quality categories are slightly longer. These results indicate that the stimulus typicality (quantity) affects the categorization process of vowel quality, but not its end result.

Meister and Werner (Meister & Werner 2009) have studied the interaction of quality and quantity in vowel category perception for the high-mid vowel pairs /i/-/e/, /y/-/ö/ and /u/-/o/ among Finnish and Estonian listeners. They found that openness correlated positively with the stimulus duration in an ABX setup, where A and B represent the prototypical exemplars of the pair (e.g., /i/ and /e/) and X represents a vowel variant located on the continuum between the pair. The longer the duration of the ambiguous stimulus in the category boundary area, the more likely it was categorized as the more open vowel of the pair, thus suggesting an interaction between duration and perceived quality. Hence, the results of Study II and those of Meister and Werner appear to be different and need discussion. To start with, both the experimental setup and the

[33] See Chapter 2.5. Internal variation of vowel categories

stimulus parameters differ between the studies. First, in Study II, the stimuli varied only for their F2 formant (front-back), whereas Meister and Werner varied primarily the F1 formant (high-low). Second, the ABX setup used by Meister and Werner differed from the categorization setup used in Study II by offering two prototypical references at the opposite ends of the continuum for comparison. Third, the durational range used in Study II was 50–500 ms, whereas Meister and Werner used shorter vowels of 50–100 ms. Fourth, the formant frequencies for Meister and Werner's prototypical /i/ reference were 250 Hz (F1) and 2205 Hz (F2), whereas in Study II, the F1 value was fixed at 250 Hz, resulting in F2 values of about 2500 Hz for a prototypical /i/ regardless of duration. In essence, the ABX setup gives two direct references to which the subject is asked to compare the ambiguous stimulus, whereas in Study II, only the memory representations were available for judging the stimulus being either /i/ or /y/. Given that the F2 value of the reference /i/ used by Meister and Werner is typical of a produced short /i/ (see Table 2, Study II), the prolongation of the ambiguous X stimulus may thus cause a growing mismatch to the typical produced long /i:/. To the author's knowledge, Study II and the study by Meister and Werner are the only perceptual studies with the focus on the interaction of the quality and quantity of Finnish vowels. The differing results of these two studies thus leave the final answer pending, and suggest that a study that includes all Finnish vowels and perhaps all their durational variants, as reported by Suomi (Suomi 2006), is needed to resolve the question.

The result obtained in Study II that vowel quality and quantity are not seemingly dependent may also be interpreted as giving indirect support to the perceptual magnet effect theory (Kuhl, 1991): the reported minor differences in F2 (see Table 2, Study II) between the short and long Finnish /i/ vowels were perceived equally because the perceptual /i/ prototypes generalized the minor differences in vowel quality (A4). This question, however, cannot be fully explored without further behavioral discrimination (AX or ABX type setup) or psychophysiological (ERP) experiments (Iverson & Kuhl 1995; Aaltonen et al. 1997; Sharma & Dorman 1998).

In Study III, the articulation of all Finnish vowels in isolated words was studied. Minor quality differences were found between the short and long vowels in the non-mask condition. The mean individual distance in the F1–F2 space between the long and short vowels without noise masking was 62 mels over all vowel categories. The short and long vowels differed in terms of F1 and F2 between the categories with the differences being the largest between /u/ and /u:/ (B2). Except for vowels /y/ and /ø/, the other categories showed a pattern where short vowels are more centralized than long vowels. This is in accordance with the results of Iivonen and Laukkanen (1993). The effect of noise on the produced vowel quality was similar in both masking conditions, i.e., no major differences between babble and pink noise were found. Noise masking caused the known Lombard effect with a significant prolongation of produced short vowels and a significant increase in the F1 frequency, but had no effect on the Euclidean distances of the short and long

vowels (B3). The results of the noise-masked articulation (i.e., that the Euclidean distances between short and long vowels do not depend on the masking) support the idea that perceptual prototypes guide the articulation, and once initiated, the process follows the fire-and-forget principle, without the influence of auditory feedback, as discussed in Chapter 2.2. This can be interpreted not to challenge the SFC model approach to the control of articulation; noise masking does not cause the auditory feedback loop to correct the articulation in fly. Furthermore, this supports the idea that perceptual prototypes serve as the desired speech targets of the SFC model proposed by Houde & Nagarajan (Ventura et al. 2009).

The main results of Study IV are the following: In both languages, Finnish and German, the inter-subject differences were mostly under the difference limens of F1–F2 frequencies for all of the three measures (A3). At the group level of averaged measures, the Euclidean distances between the weighted prototypes of Finnish and German vowels ranged from 7 to 34 mels, indicating that the vowel systems of these two linguistically different languages were strikingly similar (B5). The absolute prototype method seems to be more sensitive to language differences, but it suffers from larger individual variation in the loci of vowel prototypes, and to some extent, from non-normal density distribution within the category. This was also found in the earlier studies concerning absolute prototypes (Aaltonen et al., 1997). The weighted prototype approach provides a new method for defining the loci of perceptual sound spaces. Furthermore, it seems to provide a robust way for approximating an area within a category where individual results differ from the group mean value to a lesser extent than or equally to the difference limens of F1 and F2 frequencies (B1). This can be interpreted to show that the formation of prototypes is similar among the speakers of a particular language, and even between different languages with similar sound systems. Study IV gave some support to Johnson's theory on the adaptive dispersion effect in perception (Johnson 2000), since there was a main effect indicating that the absolute prototypes were the most peripheral of the various prototypes. Additionally, the weighted prototypes differed from the category centroids, suggesting that the gravity center of a category differs from the arithmetic mean of the category. The average distances between prototypes and articulations across all Finnish and German vowels were 129 mels and 82 mels for Pω, and 140 mels and 86 mels for Pa, respectively (Table IV, Study IV). These results suggest that perceptual vowel prototypes are not completely equal to their articulated counterparts on individual or group level in either language, with the differences being smaller in German (B3). The difference between the two languages is rather large (40–60 mels), given the fact that the vowel systems resemble each other in many respects. This evokes the question whether the result is an artifact (e.g., due to different experimental conditions and inaccuracies), or if it could be explained by other differences between the languages (e.g., by the fact that Finnish is an extreme quantity language whereas German is not). In the light of the results of the current experiments the question remains open and offers a topic for further studies.

7. Conclusions

The main findings are concluded with reference to the research hypotheses (Chapter 3.1) as follows:

B1 and B4. Weighted prototypes in comparison to category centroids and absolute prototypes

Modeling the perception and production of vowels in terms of prototypes seems feasible in the light of the results of the reported experiments. Weighted perceptual prototypes are a tempting way to replace the absolute prototypes since they are a less varying and more normally distributed indicators of good category representatives of various vowel categories.

B2. Interaction of the perceived quality of Finnish vowels with quantity

On the basis of the results of Study II, the category borders and perceptual prototypes of Finnish vowels (only the pair /i/ and /y/ was studied) are solid at the four tested durations of 50 ms, 100 ms, 250 ms and 500 ms, indicating that the quality differences between the short and long Finnish monophthongs are minor and non-significant. The results support the identity group interpretation of the Finnish quantity opposition. Whether this result can be generalized to all Finnish vowels, is a subject for further studies.

B3. Perceptual prototypes and vowel production with and without noise

In Study I, the individual variation (in terms of CV) in the produced vowels was larger than the variation of prototypes, suggesting that, although the weighted mean values (Pω) of good category exemplars may constitute the sound targets for articulation, there are other mechanisms (such as the bone conduction) that cause the actual articulation to differ from the perceptual template.

In Study III, noise masking caused the known Lombard effect but affected minimally the spread of the articulated vowels in the F1–F2 space. This result is in line with the SFC model approach to the control of articulation; noise masking does not cause the auditory feedback loop to correct the articulation in fly during the articulation.

B5. Finnish and German vowel systems in the light of the weighted prototypes

In Study IV, the vowel systems of Finnish and German, two unrelated languages with the same number of vowels, openness levels and secondary rounded vowels, appeared strikingly similar when weighted vowel prototypes were used for their comparison. This raises the question whether the weighted prototype could be used to reveal differences or classify languages with differing vowel systems.

Further research

The question whether perceptual prototypes are an experimental phenomenon that is only found in ideal laboratory conditions, or if they actually play a crucial role in speech acquisition and recognition through prototypical vowels acting as perceptual templates, calls for further research. Unresolved issues include, for example, how the weighted prototypes behave in noisy listening conditions, if they embody the perceptual magnet effect, and whether the sparser harmonic structure of female and child voices affects the perception of vowel prototypes.

References

Aaltonen, O. 1985. The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. Journal of Phonetics 13, 1, pp. 1-9.

Aaltonen, O. 1997. Vowel Perception: Behavioural and Psychophysiological Experiments. Doctoral Thesis. Turku, University of Turku, Painosalama Oy. 172 p.

Aaltonen, O. 2012. Puhe kommunikaatiomuotona ja tutkimuskohteena. Puhe ja kieli 28, 2, pp. 85–94.

Aaltonen, O., Eerola, O., Lang, H., A., Uusipaikka, E. & Tuomainen, J. 1994. Automatic discrimination of phonetically relevant and irrelevant vowel parameters as reflected by mismatch negativity. Journal of the Acoustical Society of America 96, 3, pp. 1489-1493.

Aaltonen, O., Eerola, O., Hellström, Å., Uusipaikka, E. & Lang, H., A. 1997. Perceptual magnet effect in the light of behavioral and psychophysiological data. Journal of the Acoustical Society of America 101, 2, pp. 1090-1103.

Aaltonen, O., Niemi, P., Nyrke, T. & Tuhkanen, M. 1987. Event-related brain potentials and the perception of a phonetic continuum. Biological Psychology 24, 3, pp. 197-207.

Aaltonen, O. & Suonpää, J. 1983. Computerized Two-Dimensional Model for Finnish Vowel Identifications. Audiology 22, pp. 410-415.

Aaltonen, O., Tuomainen, J., Laine, M. & Niemi, P. 1993. Cortical Differences in Tonal versus Vowel Processing as Revealed by an ERP Component Called Mismatch Negativity (MMN). Brain and Language 44, 2, pp. 139-152.

Alku, P., Tiitinen, H. & Näätänen, R. 1999. A method for generating natural-sounding speech stimuli for cognitive brain research. Clinical Neurophysiology 110, 8, pp. 1329-1333.

Becker, T. 1998. Das Vokalsystem der deutschen Standardsprache (Vowel system of Standard German). Frankfurt am Main, Germany: Peter Lang. 200 p.

Beckford Wassink, A., Wright, R.A. & Franklin, A.D. 2007. Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. Journal of Phonetics 35, 3, pp. 363-379.

Békésy, G.V. 1949. The structure of the middle ear and the hearing of one's own voice by bone conduction. Journal of the Acoustical Society of America 21, 3, pp. 217-232.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P. & Pike, B. 2000. Voice-selective areas in human auditory cortex. Nature 403, 6767, pp. 309-312.

Bendat, J.S. & Piersol, A.G. 1971. Random data analysis and measurement procedures. New York, Wiley-Interscience. 407 p.

Benson, R.R., Whalen, D., Richardson, M., Swainson, B., Clark, V.P., Lai, S. & Liberman, A.M. 2001. Parametrically dissociating speech and nonspeech perception in the brain using fMRI. Brain and Language 78, 3, pp. 364-396.

Berwick, R.C., Friederici, A.D., Chomsky, N. & Bolhuis, J.J. 2013. Evolution, brain, and the nature of language. Trends in Cognitive Sciences 17, 2, pp. 89-98.

Bladon, A. & Fant, G. 1978. A two-formant model and the cardinal vowels. Speech Transmission Laboratory Quarterly Progress and Status Report 19, 1, KTH, Stockholm, Sweden, pp. 1-8.

Bladon, A. 1982. Arguments against formants in the auditory representation of speech. In: Carlson, R. & Granström, B. (ed.). *Representation of speech in the peripheral auditory system*. Amsterdam, Elsevier Biomedical Press, pp. 95-102.

Bladon, R. & Lindblom, B. 1981. Modeling the judgment of vowel quality differences. The Journal of the Acoustical Society of America 69, 5, pp. 1414-1422.

Bliss, C.I. 1934. The method of probits. Science 79, pp. 38-39.

Blomberg, M. 1993. Synthetic phoneme prototypes and dynamic voice source adaptation in speech recognition. Speech Transmission Laboratory Quarterly Progress and Status Report 4, KTH, Stockholm, Sweden, pp. 97-140.

Blumstein, S.E. & Stevens, K.N. 1979. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America 66, 4, pp. 1001-1017.

Boersma, P. 2009. Cue constraints and their interactions in phonological perception and production. In: Boersma, P. and Hamann, S. (ed.) *Phonology in perception*, Berlin, Mouton de Gruyter. pp. 55-110.

Boersma, P. & Hamann, S. 2008. The evolution of auditory dispersion in bidirectional constraint grammars. Phonology 25, 02, pp. 217-270.

Boersma, P. & Weenink, D. 2009. Praat: doing phonetics by computer (Computer program V 5.1. 05). Online: http://www.praat.org Accessed 19.2.2014.

Botha, R. 2008. On modelling prelinguistic evolution in early hominins. Language & Communication 28, 3, pp. 258-275.

Burns, E.M. & Ward, D.W. 1978. Categorical perception - phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. Journal of the Acoustical Society of America 63, 2, pp. 456-468.

Callan, D.E., Kent, R.D., Guenther, F.H. & Vorperian, H.K. 2000. An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. Journal of Speech, Language and Hearing Research 43, 3, pp. 721-736.

Carlson, A.B. 1986. Communication Systems, An Introduction to Signals and Noise in Electrical Communication. 3rd ed., Singapore, McGraw-Hill Book Company. 686 p.

Carlson, R., Granström, B. & Fant, G. 1970. Some studies concerning perception of isolated vowels. Speech Transmission Laboratory Quarterly Progress and Status Report 11, 2-3, KTH, Stockholm, Sweden, pp. 19-35.

Carlson, R., Granström, B. & Klatt, D. 1979. Vowel perception: The relative perceptual salience of selected acoustic manipulations. Speech Transmission Laboratories Quarterly Progress Report, 20, 3-4, KTH, Stockholm, Sweden, pp. 73-83.

Castellanos, A., Benedí, J. & Casacuberta, F. 1996. An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. Speech Communication 20, 1, pp. 23-35.

Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M. & Knight, R.T. 2010. Categorical speech representation in human superior temporal gyrus. Nature neuroscience 13, 11, pp. 1428-1432.

Cheour, M., Martynova, O., Näätänen, R., Erkkola, R., Sillanpää, M., Kero, P., Raz, A., Kaipio, M., Hiltunen, J., Aaltonen, O., Savela, J. & Hämäläinen, H. 2002. Speech sounds learned by sleeping newborn. Nature 415, pp. 599-600.

Cheour-Luhtanen, M., Alho, K., Kujala, T., Sainio, K., Reinikainen, K., Renlund, M., Aaltonen, O., Eerola, O. & Näätanen, R. 1995. Mismatch negativity indicates vowel discrimination in newborns. Hearing Research, 82, 1, pp. 53-58.

Chistovich, L.A., Lublinskaya, V.V. 1979. The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. Hearing Research, 1, 3, pp. 185-195.

Chomsky, N. 2004. Language and Mind: Current Thoughts on Ancient Problems (Part I and II). In: Jenkins, L. (ed.). *Variation and Universals in Biolinguistics*. Amsterdam, Elsevier, pp. 379-405.

Chong, T.T., Cunnington, R., Williams, M.A., Kanwisher, N. & Mattingley, J.B. 2008. fMRI adaptation reveals mirror neurons in human inferior parietal cortex. Current Biology 18, 20, pp. 1576-1580.

Churchland, P.S. 2004. Neurofilosofia. Helsinki, Terra Cognita Oy. 524 p.

Comerchero, M.D. & Polich, J. 1999. P3a and P3b from typical auditory and visual stimuli. Clinical Neurophysiology 110, 1, pp. 24-30.

Coulson, S., King, J.W. & Kutas, M. 1998. Expect the unexpected: Event-related brain response to morphosyntactic violations. Language and Cognitive Pprocesses 13, 1, pp. 21-58.

Curtiss, S. 1977. Genie: A Psycholinguistic Study of a Modern Day Wild Child. New York: Academic Press.

Cutler, A. 2008. The abstract representations in speech processing. The Quarterly Journal of Experimental Psychology 61, 11, pp. 1601-1619.

Damasio, A. 2000. Tapahtumisen tunne. Miten tietoisuus syntyy. Helsinki, Terra Cognita. 349 p.

Dang, J. & Honda, K. 2002. Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. Journal of Phonetics 30, 3, pp. 511-532.

De Boer, B. 2000. Self-organization in vowel systems. Journal of Phonetics 28, pp. 441-465.

De Boer, B. & Kuhl, P.K. 2003. Investigating the role of infant-directed speech with a computer model. Acoustics Research Letters Online 4, 4, pp. 129-134.

Decoene, S. 1993. Testing the speech unit hypothesis with the primed matching task: Phoneme categories are perceptually basic. Perception & Psychophysics 53, 6, pp. 601-616.

Delgutte, B. & Kiang, N.Y. 1984. Speech coding in the auditory nerve: I. Vowel-like sounds. Journal of the Acoustical Society of America 75, 3, pp. 866-878.

Deng, L. & Geisler, C.D. 1987. A composite auditory model for processing speech sounds. Journal of the Acoustical Society of America 82, 6, pp. 2001-2012.

Eerola, O. 1993. Laboratory for Speech Perception Research: Methods and Instrumentation. Licentiate Thesis. Tampere. Tampere University of Technology. 122 p.

Eerola, O., Laaksonen, J., Savela, J. & Aaltonen, O. 2002. Suomen [y] / [i] ja [y:] / [i:] -vokaalien tuotto havaintokokeiden tulosten valossa. Fonetiikan Päivät 2002 - Phonetics Symposium 2002, 30.-31.8.2002. Espoo, Otamedia Oy, pp. 109-113.

Eerola, O., Laaksonen, J., Savela, J. & Aaltonen, O. 2003. Perception and production of the short and long Finnish [i] vowels: individuals seem to have different perceptual and articulatory

templates. Proceedings of the 15th International Congress of Phonetics Sciences, 3-9 August 2003.

Emmeche, C. 1996. The garden in the machine: the emerging science of artificial life. New Jersey, Princeton University Press. 214 p.

Evans, K.M. & Federmeier, K.D. 2007. The memory that's right and the memory that's left: Event-related potentials reveal hemispheric asymmetries in the encoding and retention of verbal information. Neuropsychologia 45, 8, pp. 1777-1790.

Eysenck, M.W. & Keane, M.T. 1992. Cognitive psychology: A student's handbook. London, Lawrence Erlbaum Associates. 557 p.

Fairbanks, G. 1954. Systematic research in experimental phonetics: 1. A theory of the speech mechanism as a servosystem. Journal of Speech & Hearing Disorders 19, pp. 133-139.

Fairbanks, G. & Grubb, P. 1961. A psychophysical investigation of vowel formants. Journal of Speech, Language and Hearing Research 4, 3, pp. 203-219.

Falk, D. 2004. Prelinguistic evolution in early hominins: Whence motherese? Behavioral and Brain Sciences 27, 04, pp. 491-503.

Fant, G., Liljencrants, J. & Lin, Q. 1985. A four-parameter model of glottal flow. Speech Transmission Laboratories Quarterly Progress Report 26, 4, KTH, Stockholm, Sweden, pp. 1-13.

Fant, G. 1960. Acoustic theory of speech production. The Hague, Mouton & Co. N.V.

Fant, G. 1983. Feature analysis of Swedish vowels - a revisit. Speech Transmission Laboratories Quarterly Progress Report 24, 2-3, KTH, Stockholm, Sweden, pp. 1-19.

Ferrari, P.F., Gallese, V., Rizzolatti, G. & Fogassi, L. 2003. Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. European Journal of Neuroscience 17, 8, pp. 1703-1714.

Ferreira, F., Henderson, J.M., Anes, M.D., Weeks, P.A. & McFarlane, D.K. 1996. Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving-window technique. Journal of Experimental Psychology: Learning, Memory, and Cognition 22, 2, p. 324.

Finney, D. 1944. The application of probit analysis to the results of mental tests. Psychometrika 9, 1, pp. 31-39.

Fisher, S.E., Vargha-Khadem, F., Watkins, K.E., Monaco, A.P. & Pembrey, M.E. 1998. Localization of a gene implicated in a severe speech and language disorder. Nature 18, 2, pp. 168-170.

Flanagan, J.L. 1955. Difference limens for formant patterns of vowel sounds. Journal of the Acoustical Society of America 27, 3, pp. 613-617.

Fodor, J. 1983. The Modularity of Mind: An Essay on Faculty Psychology. Cambridge, The MIT Press. 145 p.

Forster, K. 1981. Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. Quarterly Journal of Experimental Psychology 33, 4, pp. 465-495.

Freeman, J.A., Skapura, D.M. 1991. Neural Networks: Algorithms, Applications, and Programming Techniques, Reading, MA, Addison-Wesley, 401 p.

Friston, K. 2005. A theory of cortical responses. Philosophical Transactions of the Royal Society 360, pp. 815-836.

Galantucci, B., Fowler, C.A. & Turvey, M.T. 2006. The motor theory of speech perception reviewed. Psychonomic Bulletin & Review 13, 3, pp. 361-377.

Ganong, W.F. 1980. Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance 6, 1, p. 110.

Garcia-Sierra, A., Rivera-Gaxiola, M., Percaccio, C.R., Conboy, B.T., Romo, H., Klarman, L., Ortiz, S. & Kuhl, P.K. 2011. Bilingual language learning: An ERP study relating early brain responses to speech, language input, and later word production. Journal of Phonetics 39, 4, pp. 546-557.

Garnier, M., Henrich, N. & Dubois, D. 2010. Influence of sound immersion and communicative interaction on the Lombard effect. Journal of Speech, Language and Hearing Research 53, 3, pp. 588-608.

Geisler, C. 1988. Representation of speech sounds in the auditory nerve. Journal of Phonetics 16, pp. 19-35.

Goldinger, S.D., Azuma, T. 2003. Puzzle-solving science: the quixotic quest for units in speech perception. Journal of Phonetics 31, pp. 305-320.

Goldstein, U. 1980. An articulatory model for the vocal tracts of growing children. Doctoral Thesis. Massachusetts, U.S.A., Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science. 271 p.

Gonzalez, R.C. & Thomason, M.G. 1978. Syntactic pattern recognition: An introduction. Reading, MA, Addison-Wesley Publishing Company. 283 p.

Green, D.M. & Swets, J.A. 1966. Signal detection theory and psychophysics. New York, John Wiley & Sons.

Greenwood, D.D. 1961. Critical bandwidth and the frequency coordinates of the basilar membrane. Journal of the Acoustical Society of America 33, 10, p. 1344-1356.

Grieser, D. & Kuhl, P.K. 1989. Categorization of speech by infants: Support for speech-sound prototypes. Developmental Psychology 25, 4, p. 577-588.

Grossberg, S. 2003. Resonant neural dynamics of speech perception. Journal of Phonetics 31, pp. 423-445.

Guenther, F.H. 2000. An analytical error invalidates the "depolarization" of the perceptual magnet effect. Journal of the Acoustical Society of America 107, 6, pp. 3576-3577.

Guenther, F.H. 2006. Cortical interactions underlying the production of speech sounds. Journal of Communication Disorders 39, pp. 350-365.

Guenther, F.H. & Gjaja, M.N. 1996. The perceptual magnet effect as an emergent property of neural map formation. Journal of the Acoustical Society of America 100, 2, pp. 1111-1121.

Guenther, F.H., Hampson, M. & Johnson, D. 1998. A theoretical investigation of reference frames for the planning of speech movements. Psychological Review 105, 4, pp. 611-633.

Hackett, C.F. 1969. The origin of speech. Scientific American, 203, pp. 88-111.

Halle, M. & Stevens, K. 1962. Speech recognition: A model and a program for research. IRE Transactions on Information Theory 8, 2, pp. 155-159.

Harrikari, H. 2000. Segmental Length in Finnish - Studies within Constraint-Based Approach. Doctoral Thesis. Helsinki. Department of General Linguistics, University of Helsinki. 151 p.

Haughton, P. 1980. Physical principles of audiology. London, Adam Hilger Ltd. 183 p.

Hauser, M.D., Chomsky, N. & Fitch, W.T. 2002. The faculty of language: what is it, who has it, and how did it evolve? Science 298, 5598, pp. 1569-1579.

Hawkins, S. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. Journal of Phonetics 31, pp. 373-405.

Hawks, J.W. 1994. Difference limens for formant patterns of vowel sounds. Journal of the Acoustical Society of America 95, 2, pp. 1074-1084.

Heid, S., Wesenick, M. & Draxler, C. 1995. Phonetic analysis of vowel segments in the PhonDat database of spoken German. Proceedings of the XIII International Conference of Phonetic Sciences, Stockholm, Sweden, pp. 416-419.

Heyes, C. 2010. Where do mirror neurons come from? Neuroscience & Biobehavioral Reviews 34, 4, pp. 575-583.

Hickok, G., Houde, J. & Rong, F. 2011. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69, 3, pp. 407-422.

Hickok, G. & Poeppel, D. 2007. The cortical organization of speech processing. Nature Reviews Neuroscience 8, 5, pp. 393-402.

Hillenbrand, J., Canter, G.J. & Smith, B.L. 1990. Perception of intraphonemic differences by phoneticians, musicians, and inexperienced listeners. Journal of the Acoustical Society of America 88, 2, pp. 655-662.

Houston, D.M. & Jusczyk, P.W. 2003. Infants' long-term memory for the sound patterns of words and voices. Journal of Experimental Psychology: Human Perception and Performance 29, 6, pp. 1143.

Huotilainen, M., Kujala, A. & Alku, P. 2001. Long-term memory traces facilitate short-term memory trace formation in audition in humans. Neuroscience Letters 310, 2, pp. 133-136.

Iivonen, A. 1987. Regional differences in the realization of standard German vowels. Proceedings of the 11th International Congress of Phonetic Sciences, pp. 161-164.

Iivonen, A. 2009. Tietoisuus puheen ominaisuuksista - puheentutkimuksen kehitys. In: Aaltonen, O., Aulanko, R., Iivonen, A., Klippi, A. & Vainio, M. (ed.). *Puhuva Ihminen - Puhetieteiden perusteet.* 1st ed. Helsinki, Kustannusosakeyhtiö Otava, pp. 39-58.

Iivonen, A. & Harnud, H. 2005. Acoustical comparison of the monophthong systems in Finnish, Mongolian, and Udmurt. Journal of the International Phonetic Association 35, 1, pp. 59-71.

Iivonen, A. & Laukkanen, A. 1993. Explanations for the qualitative variation of Finnish vowels. Studies in Logopedics and Phonetics 4, pp. 29-55.

Iivonen, A. & Tella, S. 2009. Vieraan kielen ääntämisen ja kuulemisen opetus ja harjoittelu. In: Aaltonen, O., Aulanko, R., Iivonen, A., Klippi, A. & Vainio, M. (ed.). *Puhuva Ihminen - Puhetieteiden perusteet.* 1st ed. Helsinki, Kustannusosakeyhtiö Otava, pp. 269-281.

IPA, International Phonetic Alphabet (Rev. 2005): http://www.langsci.ucl.ac.uk/ipa/vowels.html. Accessed 10.3.2014.

Iverson, P. & Kuhl, P.K. 1995. Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. Journal of the Acoustical Society of America 97, 1, pp. 553-562.

Iverson, P. & Kuhl, P.K. 2000. Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? Perception & Psychophysics 62, 4, pp. 874-886.

Jacquemot, C., Dupoux, E. & Bachoud-Lévi, A. 2007. Breaking the mirror: Asymmetrical disconnection between the phonological input and output codes. Cognitive Neuropsychology 24, 1, pp. 3-22.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S. & Dupoux, E. 2003. Phonological Grammar Shapes the Auditory Cortex: A Functional Magnetic Resonance Imaging Study. Journal of Neuroscience 23, 29, pp. 9541-9546.

Johansson, S. 2005. Origins of language: constraints on hypotheses. Amsterdam, John Benjamins Publishing.

Johansson, S. 2006. Working backwards from modern language to proto-grammar. Proceedings of the 6th International Conference on the Evolution of Language, pp. 160-167.

Johnson, K. 2000. Adaptive dispersion in vowel perception. Phonetica 57, 2-4, pp. 181-188.

Johnson, K., Flemming, E. & Wright, R. 1993a. The Hyperspace Effect: Phonetic targets are hyperarticulated. Language 69, 3, pp. 505-528.

Johnson, K., Ladefoged, P. & Lindau, M. 1993b. Individual differences in vowel production. Journal of the Acoustical Society of America 94, 2, pp. 701-714.

Jones, J.A. & Munhall, K.G. 2003. Learning to produce speech with an altered vocal tract: The role of auditory feedback. Journal of the Acoustical Society of America 113, 1, pp. 532-543.

Jusczyk, P.W. 1993. From general to language-specific capacities: The WRAPSA model of how speech perception develops. Journal of Phonetics 21, 1-2, pp. 3-28.

Kalman, R.E. 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82, 1, pp. 35-45.

Karjalainen, M. 1978. An Approach to Hierarchical Information Processes with an Application to Speech Synthesis by Rule. Doctoral Dissertation. Tampere, Tampere University of Technology.

Karlsson, F. 1983. Suomen kielen äänne- ja muotorakenne. Porvoo-Helsinki-Juva, Werner Södesrstöm Oy. 410 p.

Karlsson, F. 2007. Constraints on multiple center-embedding of clauses. Journal of Linguistics 43, 2, pp. 365-392.

Kemp, J.A. 2006. Phonetic transcription: History. In: Brown, K. (ed.). The encyclopedia of language and linguistics. Elsevier, pp. 396-410.

Kilner, J.M., Neal, A., Weiskopf, N., Friston, K.J. & Frith, C.D. 2009. Evidence of mirror neurons in human inferior frontal gyrus. Journal of Neuroscience 29, 32, pp. 10153-10159.

Klatt, D.H. 1979. Speech perception: A model of acoustic-phonetic analysis and lexical access. Journal of Phonetics 7, 3, pp. 279-312.

Klatt, D.H. 1980. Software for Cascade/Parallel Formant Synthesizer. Journal of the Acoustical Society of America 67, 3, pp. 971-995.

Klatt, D.H. 1985. A shift in formant frequencies is not the same as a shift in the center of gravity of a multiformant energy concentration. Journal of the Acoustical Society of America 77, S7.

Klatt, D.H. 1987. Review of text-to-speech conversion for English. Journal of the Acoustical Society of America 82, 3, pp. 737-793.

Knight, C. & Power, C. 2012. Social conditions for the evolutionary emergence of language. In: Tallerman, M. & Gibson, K.R. (ed.). The Oxford Handbook of Language Evolution. New York, Oxford University Press, pp. 346-349.

Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V. & Rizzolatti, G. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. Science 297, 5582, pp. 846-848.

Kohonen, T. 1978. Associative memory: A system-theoretical approach. Berlin, Springer. 176 p.

Kohonen, T. 2001. Self-organizing maps. Berlin, Springer. 501 p.

König, E. 1957. Pitch Discrimination and Age. Acta Oto-Laryngologica 48, pp. 475-489.

Korpilahti, P. 1996. Electrophysiological correlates of auditory perception in normal and language impaired children. Doctoral Thesis. Turku, University of Turku. 61 p.

KOTUS Kotimaisten kielten keskus (The Institute for the Languages of Finland). (2008): Pitkien vokaalien esiintyminen; yleisyystietoja (Prevalence of long Finnish vowels). http://scripta.kotus.fi/visk/sisallys.php?p=18. Accessed 18.10.2012.

Krause, C.M., Korpilahti, P., Pörn, B., Jäntti, J. & Lang, H.A. 1998. Automatic auditory word perception as measured by 40 Hz EEG responses. Electroencephalography and Clinical Neurophysiology 107, 2, pp. 84-87.

Krech, E., Haas, W., Hove, I., Wiesinge, P. & Alvarez, M. 2009. Deutsches Aussprachewörterbuch (German Pronunciation Dictionary). Walter de Gruyter. 1076 p.

Kuhl, P.K. 1991. Human adults and human infants show a "perceptual magnet effect" for prototypes of speech categories, monkeys do not. Perception & Psychophysics 50, 2, pp. 93-107.

Kuhl, P.K. 2004. Early Language Acquisition: Cracking the Speech Code. Nature Reviews Neuroscience 5, pp. 831-843.

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

Kuisma, E., Kolehmainen, T., Renfors, M., Tomberg, J. & Tenhunen, H. 1988. Signal processing requirements in pan-European digital mobile communications. IEEE International Symposium on Circuits and Systems, pp. 1803-1810.

Kukkonen, P. 1990. Patterns of Phonological Disturbances in Adult Aphasia. Doctoral Thesis. Helsinki, University of Helsinki. 231 p.

Kuperberg, G.R., Holcomb, P.J., Sitnikova, T., Greve, D., Dale, A.M. & Caplan, D. 2003. Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. Journal of Cognitive Neuroscience 15, 2, pp. 272-293.

Kurtz, K.J. 2007. The divergent autoencoder (DIVA) model of category learning. Psychonomic Bulletin & Review 14, 4, pp. 560-576.

Laaksonen, J. 2006. Control Mechanisms of Speech Production, Evidence from Acoustic Studies of Speech after Neural and Muscular Manipulations of the Tongue. Doctoral Thesis. Turku, University of Turku. 129 p.

Ladefoged, P. & Disner, S., F. 2012. Vowels and Consonants. 3rd ed. Wiley-Blackwell. 230 p.

Ladefoged, P. & Broadbent, D.E. 1957. Information conveyed by vowels. Journal of the Acoustical Society of America 29, 1, pp. 98-104.

Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F. & Monaco, A.P. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. Nature 413, 6855, pp. 519-523.

Laine, U. 1989. Studies on Modelling of Vocal Tract Acoustics with Applications to Speech Synthesis. Doctoral Thesis. Espoo, Helsinki University of Technology.

Lane, H. & Tranel, B. 1971. The Lombard sign and the role of hearing in speech. Journal of Speech, Language and Hearing Research 14, 4, pp. 677-709.

Lang, A.H., Eerola, O., Korpilahti, P., Holopainen, I., Salo, S. & Aaltonen, O. 1995. Practical Issues in the Clinical Application of Mismatch Negativity. Ear & Hearing 16, 1, pp. 118-130.

Langton, C.G. 1990. Computation at the edge of chaos: Phase transitions and emergent computation. Physica D: Nonlinear Phenomena 42, 1, pp. 12-37.

Lau, P. 2008. The Lombard effect as a communicative phenomenon. UC Berkeley Phonology Lab Report, pp. 1-9.

Lehtonen, J. 1970. Aspects of Quantity in Standard Finnish. Doctoral Thesis. Jyväskylä, University of Jyväskylä. 199 p.

L'Enfant sauvage (The Wild Boy). 1970. United Artists/Les Films du Carrosse.

Levine, D.S. 1991. Introduction to Neural and Cognitive Modelling. New Jersey, Lawrence Erlbaum Associates, Inc. 439 p.

Liberman, A.M. 1985. The motor theory of speech perception revised. Cognition 21, pp. 1-36.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. & Studdert-Kennedy, M. 1967. Perception of the speech code. Psychological Review 74, 6, pp. 431-461.

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, pp. 358-368.

Liberman, A.M. & Whalen, D.H. 2000. On the relation of speech to language. Trends in Cognitive Sciences 4, 5, pp. 187-196.

Lieberman, P. 1987. On the origins of language: An introduction to the evolution of human speech. Lanham, MD, University Press of America. 196 p.

Lieberman, P. 2000. Eeva puhui, Ihmisen kieli ja ihmisen evoluutio. Helsinki, Terra Cognita. 263 p.

Lieberman, P., Laitman, J.T., Reidenberg, J.S. & Gannon, P.J. 1992. The anatomy, physiology, acoustics and perception of speech: essential elements in analysis of the evolution of human speech. Journal of Human Evolution 23, 6, pp. 447-467.

Liljencrants, J. & Lindblom, B. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. Language 48, 4, pp. 839-862.

Lindblom, B. 1986. Phonetic universals in vowel systems. In: Ohala, J. & Jaeger, J. (ed.) *Experimental phonology.* Orlando, FL, Academic Press, pp. 13-44.

Lindblom, B. 1992. Phonetic content in phonology. Phonologica 1988: Proceedings of the 6th International Phonology Meeting, pp. 181-196.

Lindsay, P.H. & Norman, D.A. 1977. Human Information Processing, An Introduction to Psychology. 2nd ed. Academic Press. 777 p.

Liu, C. & Kewley-Port, D. 2004. Formant discrimination in noise for isolated vowels. Journal of the Acoustical Society of America 116, 5, pp. 3119-3129.

Lively, S.E. & Pisoni, D.B. 1997. On prototypes and phonetic categories: a critical assessment of the perceptual magnet effect in speech perception. Journal of Experimental Psychology 23, 6, pp. 1665-1679.

Lotto, A.J. 2000. Reply to "An analytical error invalidates the 'depolarization' of the perceptual magnet effect" [J.Acoust.Soc.Am. 107, 6, 3576-3577 (2000)]. Journal of the Acoustical Society of America 107, 6, pp. 3578-3580.

Lotto, A.J., Kluender, K.R. & Holt, L.L. 1998. Depolarizing the perceptual magnet effect. Journal of the Acoustical Society of America 103, 6, pp. 3648-3655.

Maeda, S. 1990. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W., J. & Marchal, A.

(ed.). *Speech production and speech modelling*. The Netherlands, Kluwer Academic Publishers, pp. 131-149.

Mangold, M. 2005. Duden 06. Das Aussprachewörterbuch ((German) Pronunciation Dictionary). Mannheim, Germany, Bibliographisches Institut. 791 p.

Marslen-Wilson, W.D. 1987. Functional parallelism in spoken word-recognition. Cognition 25, 1, pp. 71-102.

Marslen-Wilson, W.D. & Welsh, A. 1978. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology 10, 1, pp. 29-63.

Martinet, A. 1984. Double articulation as a criterion of linguisticity. Language Sciences 6, 1, pp. 31-38.

Masataka, N. 2007. Music, evolution and language. Developmental Science 10, 1, pp. 35-39.

Masterton, R.B. 1992. Role of the central auditory system in hearing: the new direction. Trends in Neurosciences 15, 8, pp. 280-285.

May, P. & Tiitinen, H. 2004. The MMN is a derivative of the auditory N100 response. Neurology and Clinical Neurophysiology 20, pp. 1-5.

May, P.J. & Tiitinen, H. 2010. Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. Psychophysiology 47, 1, pp. 66-122.

McClelland, J.L. & Elman, J.L. 1986. The TRACE model of speech perception. Cognitive Psychology, 18, pp. 1-86.

McClelland, J. & Rumelhart, D. 1986. Palallel distributed processing. Explorations in the microstructures of cognition (2 Vol). Cambridge, MA, The MIT Press.

Mehrabian, A. 1968. Some referents and measures of nonverbal behavior. Behavior Research Methods & Instrumentation, 1, 6, pp. 203-207.

Mehrabian, A. 1969. Significance of posture and position in the communication of attitude and status relationships. Psychological bulletin, 71, 5, pp. 359-372.

Meister, E. & Werner, S. 2009. Duration affects vowel perception in Estonian and Finnish. Linguistica Uralica 3, pp. 161-177.

Mermelstein, P. 1978. Difference limens for formant frequencies of steady-state and consonant-bound vowels. Journal of the Acoustical Society of America 63, 2, pp. 572-580.

Meurmann, O.H. 1954. The Difference Limen of Frequency in Tests of Auditory Function. Acta Oto-Laryngologica 118, pp. 144-155.

Miller, J.D. 1989. Auditory-perceptual interpretation of the vowel. Journal of the Acoustical Society of America 85, 5, pp. 2114-2134.

Miller, J.L., Connine, C.M., Schermer, T.M. & Kluender, K.R. 1983. A possible auditory basis for internal structure of phonetic categories. Journal of the Acoustical Society of America 73, 6, pp. 2124-2133.

Miller, J.L. 1997. Internal Structure of Phonetic Categories. Language and Cognitive Pprocesses 12, 5/6, pp. 865-869.

Milner, P.M. 1999. The Autonomous Brain: A Neural Theory of Attention and Learning. New Jersey, Lawrence Erlbaum Associates, pp. 57-69.

Miyamoto, R.T., Hay-McCutcheon, M.J., Iler K.K., Houston, D.M. & Bergeson-Dana, T. 2008. Language skills of profoundly deaf children who received cochlear implants under 12 months of age: a preliminary study. Acta Oto-Laryngologica 128, 4, pp. 373-377.

Molau, S., Kanthak, S. & Ney, H. 2000. Efficient vocal tract normalization in automatic speech recognition. In Proceedings of the ESSV'00, Cottbus, Germany.

Monahan, P.J. & Idsardi, W.J. 2010. Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. Language and Cognitive Pprocesses 25, 6, pp. 808-839.

Moore, B. 1976. Comparison of frequency DL's for pulsed tones and modulated tones. British Journal of Audiology 10, 1, pp. 17-20.

Moosmüller, S. 2007. Vowels in Standard Austrian German, An Acoustic-Phonetic and Phonological Analysis. Vienna, Austria, Philologisch-Kulturwissenschaftliche Fakultät, Universität Wien. 271 p.

Morton, J. 1969. Interaction of information in word recognition. Psychological Review 76, 2. pp. 165-178.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R.J., Luuk, A., Allik, J., Sinkkonen, J. & Alho, K. 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385, 30, pp. 432-434.

Näätänen, R., Paavilainen, P., Rinne, T. & Alho, K. 2007. The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clinical Neurophysiology 118, pp. 2544-2590.

Näätänen, R., Pakarinen, S., Rinne, T. & Takegata, R. 2004. The mismatch negativity (MMN): towards the optimal paradigm. Clinical Neurophysiology 115, pp. 140-144.

Näätänen, R. & Picton, T. 1987. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology 24, 4, pp. 375-425.

Nábelek, A.K., Czyzewski, Z. & Crowley, H.J. 1993. Vowel boundaries for steady-state and linear formant trajectories. Journal of the Acoustical Society of America 94, 2, pp. 675-687.

Nearey, T.M. 1989. Static, dynamic, and relational properties in vowel perception. Journal of the Acoustical Society of America 85, 5, pp. 2088-2113.

Nearey, T.M. 1990. The segment as a unit of speech perception. Journal of Phonetics 18, pp. 347-373.

Nearey, T.M. 1992. Context Effects in a Double-weak Theory of Speech Perception. Language and Speech 35, 1,2, pp. 153-171.

Nichols, J. 1998. The origin and dispersal of languages: Linguistic evidence. In: Jablonski, N. & C. Aiello, L., A. (ed.). *The Origin and Diversification of Language*. San Francisco, California Academy of Sciences, pp. 127-170.

Niemi, M., Laaksonen, J.P., Aaltonen, O. & Happonen, R. 2004. Effects of transitory lingual nerve impairment on speech: an acoustic study of diphthong sounds. Journal of Oral and Maxillofacial Surgery 62, 1, pp. 44-51.

Niemi, M., Laaksonen, J., Ojala, S., Aaltonen, O. & Happonen, R. 2006. Effects of transitory lingual nerve impairment on speech: an acoustic study of sibilant sound /s/. International Journal of Oral and Maxillofacial Surgery 35, 10, pp. 920-923.

Niemi, M., Laaksonen, J., Tuomainen, J., Aaltonen, O. & Happonen, R. 2002. Effects of transitory lingual nerve impairment on speech: an acoustic study of vowel sounds. Journal of Oral and Maxillofacial Surgery 60, 6, pp. 647-652.

Norris, D. & McQueen, J.M. 2008. Shortlist B: a Bayesian model of continuous speech recognition. Psychological Review 115, 2. 357 p.

Obleser, J. & Eisner, F. 2008. Pre-lexical abstraction of speech in the auditory cortex. Trends in Cognitive Sciences 13, 1, pp. 14-19.

O'Dell, M. 2003. Intrinsic timing and quantity in Finnish. Doctoral Thesis. Tampere, University of Tampere. 128 p.

Ohala, J.J. 1983. The origin of sound patterns in vocal tract constraints. In: MacNeilage, P. (ed.). *The production of speech*. New York, Springer, pp. 189-216.

Ohala, J.J. 1990. There is no interface between phonology and phonetics: a personal view. Journal of Phonetics 18, 2, pp. 153-172.

Ojala, S. 2011. Towards an Integrative Information Society: Studies on Individuality in Speech and Sign. Doctoral Thesis. Turku, University of Turku. 129 p.

O'Rourke, T.B. & Holcomb, P.J. 2002. Electrophysiological evidence for the efficiency of spoken word processing. Biological Psychology 60, 2, pp. 121-150.

O'Shaughnessy, D. 1987. Speech Communication: Human and Machine. Reading, MA, Addison-Wesley.

Paliwal, K.K., Ainsworth, W.A. & Lindsay, D. 1983. A study of two-formant models for vowel identification. Speech Communication 2, 4, pp. 295-303.

Pantev, C., Hoke, M., Lehnertz, K., Lütkenhöner, B., Anogianakis, G. & Wittkowski, W. 1988. Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields. Electroencephalography and Clinical Neurophysiology 69, 2, pp. 160-170.

Pastore, R.E. 1987. Categorical perception: Some psychophysical models. In: Harnard, S. (ed.). *Categorical Perception, The Groundwork of Cognition.* New York, Cambridge University Press, pp. 29-52.

Patel, S.H. & Azzam, P.N. 2005. Characterization of N200 and P300: Selected Studies of the Event-Related Potential. International Journal of Medical Sciences 2, 4, pp. 147-154.

Pease, A. 1991. Body Language. How to read others' thoughts by their gestures. 19th ed., London, Sheldon Press. 152 p.

Peltola, M.S. 2003. The attentive and preattentive perception of native and non-native vowels. Doctoral Thesis. Turku, University of Turku. 110 p.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Perrier, P., Vick, J., Wilhelms-Tricarico, R. & Zandipour, M. 2000. A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. Journal of Phonetics 28, pp. 233-272.

Perkell, J.S., Matthies, M.L., Svirsky, M.A. & Jordan, M.I. 1995. Goal-based speech motor control: a theoretical framework and some preliminary data. Journal of Phonetics 23, 1, pp. 23-35.

Perreault, C., & Mathew, S. 2012. Dating the Origin of Language Using Phonemic Diversity. PLoS ONE 7, 4.

Peterson, G.E. & Barney, H.L. 1952. Control methods used in a study of the vowels. Journal of the Acoustical Society of America 24,2, pp. 175-184.

Pfeifer, J.E. & Shoup, L.L. 1976. Acoustic Characteristics of Speech Sounds. In: Lass, N.J. (ed.). *Contemporary Issues in Experimental Phonetics*. New York, Academic Press Inc, pp. 171-224.

Pinker, S. 2010. The language instinct: How the mind creates language. HarperCollins.

Pinker, S. & Bloom, P. 1990. Natural language and natural selection. Behavioral and Brain Sciences, 13, 4, pp. 707-727.

Pinker, S. & Prince, A. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition 28, 1, pp. 73-193.

Pisoni, D.B. 1980. Variability of vowel formant frequencies and the quantal theory of speech: a first report. Phonetica 37, 5-6, pp. 285-305.

Pisoni, D.B. & Luce, P.A. 1987. Acoustic-phonetic representations in word recognition. Cognition 25, 1, pp. 21-52.

Plaut, D.C. 2003. Connectionist modeling of language: Examples and implications. In: Banich, M. & Mack, M. (eds.), *Mind, Brain, and Language: Multidisciplinary Perspectives*. pp. 143-167.

Plaut, D.C. & Kello, C.T. 1999. The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In: MacWhinney, B. (ed.) *The emergence of language.* Mahwah, NJ, Erlbaum, pp. 381-415.

Plomp, R. 1964. The Ear as a Frequency Analyser. Journal of the Acoustical Society of America 36, 9, pp. 1628-1636.

Polich, J. 2007. Updating P300: an integrative theory of P3a and P3b. Clinical Neurophysiology 118, 10, pp. 2128-2148.

Pörschmann, C. 2000. Influences of bone conduction and air conduction on the sound of one's own voice. Acta Acustica united with Acustica 86, 6, pp. 1038-1045.

Port, R. 2007. How are words stored in memory? Beyond phones and phonemes. New Ideas in Psychology 25, 2, pp. 143-170.

Pulvermüller, F. 2005. Brain mechanisms linking language and action. Nature Reviews Neuroscience 6, 7, pp. 576-582.

Raimo, I., Savela, J., Launonen, A., Kärki, T., Mattila, M., Uusipaikka, E. & Aaltonen, O. 2002. Multilingual Vowel Perception. Presented at the ISCA Tutorial and Research Workshop (ITRW) on Temporal Integration in the Perception of Speech. Aix-en-Provence, France, 8-10 April 2002.

Raitio, T., Suni, A., Vainio, M. & Alku, P. 2014. Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise. Computer Speech & Language 28, 2, pp. 648-664.

Rakerd, B. 1984. Vowels in consonantal context are perceived more linguistically than are isolated vowels: Evidence from an individual differences scaling study. Perception & Psychophysics 35, 2, pp. 123-136.

Rao, M.D. & Letowski, T. 2006. Callsign Acquisition Test (CAT): Speech Intelligibility in Noise. Ear & Hearing 27, pp. 120-128.

Regan, D. 1989. Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine. New York, Amsterdam, Elsevier. 652 pp.

Reinfeldt, S., Östli, P., Håkansson, B. & Stenfelt, S. 2010. Hearing one's own voice during phoneme vocalization - Transmission by air and bone conduction. Journal of the Acoustical Society of America 128, 2, pp. 751-762.

Reinfeldt, S. 2009. Bone Conduction Hearing in Human Communication - Sensitivity, Transmission, and Applications. Doctoral Thesis. Gothenburg, Sweden, Chalmers University of Technology. 58 p.

Rendell, L. 1986. A general framework for induction and a study of selective induction. Machine Learning 1, 2, pp. 177-226.

Repp, B.H. & Crowder, R.G. 1990. Stimulus order effects in vowel discrimination. Journal of the Acoustical Society of America 88, 5, pp. 2080-2090.

Repp, B.H. & Liberman, A.M. 1987. Phonetic category boundaries are flexible. In: Harnad, S. (ed.). *Categorical Perception, The Groundwork of Cognition*. Cambridge University Press, pp. 89-112.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. 1996. Premotor cortex and the recognition of motor actions. Cognitive Brain Research 3, 2, pp. 131-141.

Rosch, E. 1975. Cognitive Reference Points. Cognitive Psychology 7, pp. 532-547.

Rosner, B.S. & Pickering, J.B. 1994. Vowel Perception and Production. New York, Oxford University Press. 432 p.

Ross, B. & Tremblay, K. 2009. Stimulus experience modifies auditory neuromagnetic responses in young and older listeners. Hearing Research 248, 1, pp. 48-59.

Rozsypal, A.J., Stevenson, D.C. & Hogan, J.T. 1985. Dispersion in Models of Categorical Perception. Journal of Mathematical Psychology 29, 3, pp. 271-288.

Ru, P., Chi, T. & Shamma, S. 2003. The synergy between speech production and perception. Journal of the Acoustical Society of America 113, 1, pp. 498-515.

Saarni, T. 2010. Segmental durations of speech. Doctoral Thesis. Turku, TUCS Dissertations No 126.

Sams, M., Aulanko, R., Aaltonen, O. & Näätänen, R. 1990. Event-related potentials to infrequent changes in synthesized phonetic stimuli. Journal of Cognitive Neuroscience 2, 4, pp. 344-357.

Sanes, D. & Rubel, E. 1988. The development of stimulus coding in the auditory system. In: Jahn, A.F. & Santos-Sachhi, J. (ed.). *Physiology of the Ear*. New York, Raven Press, pp. 431-456.

Savela, J. 2009. Role of Selected Spectral Attributes in the Perception of Synthetic Vowels. Doctoral Thesis. Turku, University of Turku. 82 p.

Savela, J., Kujala, T., Tuomainen, J., Ek, M., Aaltonen, O. & Näätänen, R. 2002. The mismatch negativity and reaction time as indices of the perceptual distance between corresponding vowels of two related languages. Cognitive Brain Research 16, pp. 250-256.

Schmitt, B.M., Münte, T.F. & Kutas, M. 2000. Electrophysiological estimates of the time course of semantic and phonological encoding during implicit picture naming. Psychophysiology 37, 4, pp. 473-484.

Schouten, B., Gerrits, E. & van Hessen, A. 2003. The end of categorical perception as we know it. Speech Communication 41, pp. 71-80.

Schwartz, J., Boë, L., Vallée, N. & Abry, C. 1997. The dispersion-focalization theory of vowel systems. Journal of Phonetics 25, 3, pp. 255-286.

Schwartz, J. & Escudier, P. 1987. Does the human auditory system include large scale spectral integration? In: Schouten, M. (ed.). *The psychophysics of Speech Perception*. Springer, pp. 284-292.

Scott, S.K. 2003. How might we conceptualize speech perception? The view from neurobiology. Journal of Phonetics 31, pp. 417-422.

Scott, S.K. & Evans, S. 2010. Categorizing speech. Nature Neuroscience 13, 11, pp. 1304-1306.

Sendlmeier, W.F. 1981. Der Einfluß von Qualität und Quantität auf die Perzeption betonter Vokale des Deutschen (The Influence of Quality and Quantity on the Perception of Stressed Vowels in German). Phonetica 38, 5-6, pp. 291-308.

Shamma, S.A. 1985. Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. Journal of the Acoustical Society of America 78, 1 pp. 1612-1621.

Sharma, A. & Dorman, M.F. 1998. Exploration of perceptual magnet effect using the mismatch negativity auditory evoked potentials. Journal of the Acoustical Society of America 104, 1, pp. 511-517.

Stevens, K.N. 2000. Acoustic Phonetics. 30th ed. Cambridge, Massachusetts, The MIT Press. 607 p.

Stevens, K.N. & Keyser, S.J. 2010. Quantal theory, enhancement and overlap. Journal of Phonetics 38, 1, pp. 10-19.

Stevens, S.S. Volkmann, J. & Newman, E.B. 1937. A Scale for the Measurement of the Psychological Magnitude Pitch. Journal of the Acoustical Society of America 8, pp. 185-190.

Strange, W. 1989. Evolving theories of vowel perception. Journal of the Acoustical Society of America 85, 5, pp. 2081-2087.

Studdert-Kennedy, M. 1993. Discovering phonetic function. Journal of Phonetics 21, 1-2, pp. 147-155.

Studdert-Kennedy, M. 2005. How did language go discrete? In: Tallerman, M. (ed.) *Language Origins: Perspectives on Evolution.* Oxford University Press, pp. 48-67.

Suga, N. 1988. Auditory neuroethology and speech processing: complex-sound processing by combination-sensitive neurons. In: Edelman, G.M., Gall, W., E. & Cowan, W.M. (ed.). *Functions of the Auditory System.* John Wiley and Sons, pp. 679-720.

Suga, N. 2006. Basic acoustic patterns and neural mechanisms shared by humans and animals for auditory perception. In: Greenberg, S. & Ainsworth, W.A. (ed.). *Listening to Speech: An Auditory Perspective*. Mahwah, New Jersey, U.S.A., Lawrence Erlbaum Associates, pp. 159-181.

Suga, N., Ma, X., Gao, E., Sakai, M. & Chowdhury, S.A. 2003. Descending system and plasticity for auditory signal processing: neuroethological data for speech scientists. Speech Communication 41, 1, pp. 189-200.

Suomi, K. 2005. Temporal conspiracies for a tonal end: Segmental durations and accentual f0 movement in a quantity language. Journal of Phonetics 33, pp. 291-309.

Suomi, K. 2006. Stress, Accent and Vowel Durations in Finnish. Lund, Department of Linguistics & Phonetics, Lund University, Sweden. Working Papers 52, pp. 129-132.

Suomi, K. 2007. On the tonal and temporal domains of accent in Finnish. Journal of Phonetics 35, pp. 40-55.

Suomi, K., Toivanen, J. & Ylitalo, R. 2003. Durational and tonal correlates of accent in Finnish. Journal of Phonetics 31, pp. 113-138.

Suomi, K., Toivanen, J. & Ylitalo, R. 2006. Fonetiikan ja suomen äänneopin perusteet. Helsinki, Gaudeamus Kirja. 270 p.

Suomi, K. & Ylitalo, R. 2004. On durational correlates of word stress in Finnish. Journal of Phonetics 32, pp. 35-63.

Tank, D.W. 1989. What details of neural circuits matter? Seminars in the Neurosci 1, pp. 67-79.

Tomasello, M. 1996. The cultural roots of language. In: Velichkovskiĭ, B.M. & Rumbaugh, D.M. (ed.). *Communicating meaning: The evolution and development of language*. New Jersey, Erlbaum, pp. 275-307.

Traunmüller, H. 1990. Analytical expressions for the tonotopic sensory scale. Journal of the Acoustical Society of America 88, pp. 97-100.

Ulbaek, I. 1998. The origin of language and cognition. In: Hurford, J.R., Studdert-Kennedy, M. & Knight, C. *Approaches to the Evolution of Language*. Cambridge, UK, Cambridge University Press, pp. 30-43.

Uppenkamp, S., Johnsrude, I.S., Norris, D., Marslen-Wilson, W. & Patterson, R.D. 2006. Locating the initial stages of speech-sound processing in human temporal cortex. Neuroimage 31, 3, pp. 1284-1296.

Vaissière, J. 2011. On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII), pp. 52-59.

Van Summers, W., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I. & Stokes, M.A. 1988. Effects of noise on speech production: Acoustic and perceptual analyses. Journal of the Acoustical Society of America 84, 3, pp. 917-928.

Veldhuis, R. 1998. A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. Journal of the Acoustical Society of America 103, 1, pp. 566-571.

Ventura, M.I., Nagarajan, S.S. & Houde, J.F. 2009. Speech target modulates speaking induced suppression in auditory cortex. BMC Neuroscience 10, pp. 10-58.

Von Neumann, J. 2012. The computer and the brain. 3rd ed., Yale University Press.

Weitzman, R.S. 1992. Vowel categorization and the critical band. Language and Speech 35, 1,2, pp. 115-125.

Whalen, D., Benson, R.R., Richardson, M., Swainson, B., Clark, V.P., Lai, S., Mencl, W.E., Fulbright, R.K., Constable, R.T. & Liberman, A.M. 2006. Differentiation of speech and nonspeech processing within primary auditory cortex. Journal of the Acoustical Society of America 119, 1, pp. 575-581.

Whalen, D.H. & Liberman, A.M. 1987. Speech perception takes precedence over nonspeech perception. Science 237, 4811, pp. 169-171.

Wier, C.C., Jesteadt, W. & Green, D.M. 1977. Frequency discrimination as a function of frequency and sensation level. Journal of the Acoustical Society of America 61, 1, pp. 178-184.

Wiik, K. 1965. Finnish and English vowels. Doctoral Thesis. Turku, University of Turku.

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe, V., Ilmoniemi, R.J. & Näätänen, R. 1999. Brain responses reveal the learning of foreign language phonemes. Psychophysiology 36, pp. 638-642.

Wolfe, J., Garnier, M. & Smith, J. Voice acoustics, An Introduction. School of Physics, The University of New South Wales. Sydney, Australia. http://www.phys.unsw.edu.au/jw/voice.html. Accessed 25.2.2014.

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, pp. 175-185.

Zwicker, E. & Terhardt, E. 1980. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. Journal of the Acoustical Society of America 68, 5, pp. 1523-1525.

Appendices: Original publications I - IV

Publication I

Eerola, O., Savela, J. 2011.

Differences in Finnish front vowel production and weighted perceptual prototypes in the F1–F2 space.

In *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, China, 2011. Pages 631-634.

DIFFERENCES IN FINNISH FRONT VOWEL PRODUCTION AND WEIGHTED PERCEPTUAL PROTOTYPES IN THE F1-F2 SPACE

Osmo Eerola[a,b] *& Janne Savela*[c]

[a]Department of Biomedical Engineering, Tampere University of Technology, Tampere, Finland;
[b]Centre for Cognitive Neuroscience, University of Turku, Turku, Finland;
[c]Department of Information Technology, University of Turku, Turku, Finland
osmo.eerola@tut.fi; jansav@utu.fi

ABSTRACT

The perception and production of the mid-front Finnish vowels /i/, /e/, /y/, and /ø/ were investigated in fourteen Finnish-speaking subjects. In the perception experiment, synthesized long vowels were used as stimuli in order to identify category prototypes. For production, the subjects were asked to pronounce words including these vowels as short and long variants. This study introduces a new concept of *weighted perceptual prototype*, which is compared with the estimated absolute prototypes obtained in the perception experiment. The calculated mean Euclidean distance in the F1-F2 space between the produced vowels and their weighted category prototypes was 111 mel for short and 116 mel for long vowels. At an individual level, the F1 and F2 values of the weighted perceptual prototypes correlated significantly with the F1 and F2 values of the produced short and long vowels. Statistically significant differences were found between the mean values of the weighted category prototypes and estimated absolute prototypes for /i/, /e/, and /ø/ but not for /y/.

Keywords: vowel perception and production, weighted prototypes

1. INTRODUCTION

The existence of internal structures of phonetic categories and prototypical category representatives has been shown in many reports [10-16]. The phoneme prototype (P) is traditionally defined as the best representative of a phoneme category, and experimentally determined as the highest ranking category member in goodness evaluation tests [8].

Irrespective of the experimental approach, the measured prototype represents an estimate of the absolute or 'true' category prototype, marked here as Pa_{est}. The goodness of the estimate depends on the number of stimuli used in the grid to cover the investigated vowel space; decreasing the step size of the synthesis parameters will rapidly increase the number of stimuli unpractically large for use in listening experiments. To overcome this problem, novel optimizing methods have been presented [2, 6]. The weighted prototype (Pω) approach enables us to avoid some of these experimental problems. The Pω is robust in the sense that it represents the center of gravity of the category: the absolute prototype can most likely be found within the area of the vowel space where the majority of the stimuli with high goodness values lie.

Phoneme prototypes are the natural candidates for the 'auditory targets', which many models assume to be the elementary neural representations used in the template matching of speech perception, and for control references in speech production [3, 4].

The Finnish vowel system includes eight vowels: /a/, /e/, /i/, /o/, /u/, /y/, /æ/, and /ø/, which all can occur as short (single) or long (double) in any position of a word. This study concerns the perception and production of the Finnish mid-front vowels /i/, /y/, /e/, and /ø/ and consists of two experiments: a combined vowel identification and rating experiment, and a subsequent vowel production experiment. The purpose of this study was to test the hypothesis that the acoustic features (as implemented in F1 and F2) of an individual's perceptual vowel prototype correlate with the same acoustic features of the produced vowel, and to compare the Euclidean distances of perceived and produced vowels in the F1-F2 space. Additionally, since Finnish is an example of an extreme quantity language, the effect of vowel duration on the articulated vowel quality was investigated as well: it was assumed that the long vowels better achieve the articulatory targets (prototypes), in other words, they have smaller distances from the prototypes than the short vowels have.

2. EXPERIMENT 1: PERCEPTION

2.1. Subjects

Fourteen (14) normally hearing young adults aged 17-31 and speaking the modern educated Finnish of South-West Finland volunteered as subjects (7 male, 7 female) in both experiments. All subjects were screened for hearing impairments by means of an audiometer (Amplivox 116).

2.2. Stimuli and procedure

Forty-six (46) vowel variants were synthesized using the Klatt serial mode speech synthesizer [7] to represent the long Finnish /e:/, /i:/, /y:/, and /ø:/ vowels with a duration of 250 ms. On the basis of the earlier reported typical formant values of the relevant vowels occurring in Finnish words [1], a tentative category center was determined for each vowel category (Table 1, upper part).

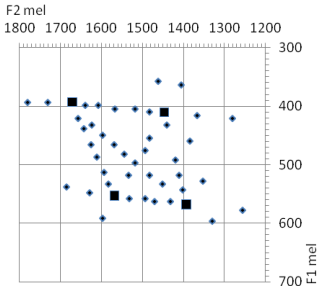
Forty-two (42) vowel variants were then synthesized around the category centers to form the grid of stimuli shown in Figure 1. The F1 and F2 of these stimuli varied in approximately similar steps of 30 mel in the psychoacoustic mel scale (Table 1, lower part). The other formants were fixed: F3 was 2400 Hz for /y:/, 2460 Hz for /ø:/, 2800 Hz for /e:/, and 2980 Hz for /i:/, and F4 was 3200 Hz for /y:/, 3300 Hz for /ø:/, 3800 Hz for /e:/, and 4000 Hz for /i:/.

The f0 contour rose from 112 Hz to 122 Hz during the first 50 ms and then decreased to 102 Hz until the end of the 250 ms stimulus. A linear window of 10 ms was used at the beginning and end of the stimulus in order to prevent audible clicks. The stimuli were presented in an acoustically dampened room (27 dB_A) via Sennheiser PC161 headphones that were calibrated for each session by Brüel & Kjaer Type 2235 SPL meter to deliver 83 +/- 0.5 dB_A.

Table 1: The F1 and F2 values (in Hz and mel) of the tentative category centers, and the range of F1 and F2 variation (in Hz) of the grid of synthesized stimuli presented in Figure 1.

Vowel	F1 Hz	F1 mel	F2 Hz	F2 mel
/e:/	435	553	2170	1568
/i:/	285	393	2460	1671
/y:/	300	410	1865	1447
/ø:/	450	568	1740	1393
Range	F1 min	F1 max	F2 min	F2 max
/e:/	370	475	1980	2500
/i:/	285	335	2170	2800
/y:/	255	340	1500	2040
/ø:/	375	480	1450	1920

Figure 1: The grid of synthesized long vowels in Experiment 1. The category centers (from left to right, and from top to down; /i:/, /y:/, /e:/, and /ø:/) determined on the basis of literature are marked with large squares. The horizontal F2 and vertical F1 axes are in mels.



The EMFC tool of the Praat software was used for stimulus delivery and data collection. The stimuli were presented in 10 blocks of 46 stimuli, each variant occurring 10 times in a random order. After each block, the subject was allowed to take a short break. The test started with a training block consisting of 30 vowels.

In the perception experiment, the subjects were first asked to identify the vowels as belonging to one of the four categories /e:/, /i:/, /y:/, or /ø:/, and then to rate the goodness of each vowel stimulus. A rating scale of 1-7 was employed. The highest score (7) represented a natural sounding, good exemplar of the relevant vowel category, whereas the lowest score (1) represented a poor exemplar. If the subject was not able to categorize the stimulus into the given categories, then the subject was instructed to select the null goodness score (0).

2.3. Analysis and results

For each subject, the identifications of the 46 stimulus variants were counted. This resulted in a categorization rate (%) for each stimulus. For those stimuli that were classified as belonging to one and the same category at a rate of ≥70%, a mean goodness score value was calculated based on the ratings on the scale 1-7. The highest scoring stimulus token in each category signifies an estimate of the absolute prototype Pa_{est}(F1,F2).

The weighted prototype Pω(**F1,F2**) of each category was formed by using the equation

(1) $\mathbf{Fi}=(a_1r_1Fi_1+ a_jr_jFi_j +...+a_nr_nFi_n) /(a_1r_1+a_jr_j +...+ a_nr_n)$

where **Fi** = weighted formant frequency, i=1,2,
Fi_j = formant i of stimulus j, j=1,2, ...,n,
a_j = evaluation mean score (1-7), j=1,2, ...,n,
r_j = identification consistency (0.7-1.0), j=1,2, ...,n,
n=number of stimuli identified as category members

Pω(F1,F2) thus represents a point in the F1-F2 space (mel) that is obtained by weighting the F1 and F2 mel values of each stimulus identified as a category member (≥70%) by the goodness rating value. The mean values and standard deviations of the F1 and F2 frequencies (mel) of the estimated absolute category prototypes (Pa_{est}) and the mean values of the weighted prototypes (Pω) of the 14 listeners are presented in Table 2.

Table 2: The mean F1 and F2 values (mel) of perceived /e:/, /i:/, /y:/, and /øː/ vowels given as the estimated absolute prototypes (Pa_{est}) and weighted prototypes (Pω). Standard deviations are in the parentheses.

Vowel	Pa_{est} F1	Pa_{est} F2	Pω F1	Pω F2
/e:/	558 (24)	1639 (23)	541 (17)	1628 (15)
/i:/	392 (12)	1733 (60)	401 (5)	1701 (23)
/y:/	388 (33)	1483 (48)	402 (12)	1462 (11)
/øː/	569 (20)	1375 (53)	544 (14)	1412 (33)

The mean differences between the two methods of obtaining the prototype are 9-25 mel for F1 and 11-36 mel for F2. The Wilcoxon signed ranks test showed that there were significant differences between the estimated absolute and the center-of-gravity type (i.e. weighted) prototypes ($p < 0.05$) in the categories /e:/, /i:/, and /øː/, but not in /y:/.

3. EXPERIMENT 2: PRODUCTION

3.1. Procedure

In the production experiment, the articulation of the utterances [tili], [ti:li], [teli], [te:li] [tyli], [ty:li], and [tøli], [tøːli] (Finnish words and non-words) was recorded from the subjects of Experiment 1. They were asked to utter each word five times successively using their normal speech style. The recording was carried out in an acoustically dampened room by using a high quality AKG D660S microphone that was connected via an amplifier to a PC. The recordings were made at a sampling rate of 44.1 kHz, and saved as sound files for later analysis. Praat SW was used both for the recordings and analysis.

3.2. Analysis and results

The sound samples were automatically analyzed using a text grid in which the steady state part of each target vowel was windowed varying between utterances. Five vowel formants (F1-F5) were analyzed by using the Burg method in which short-term LPC coefficients are averaged for the length of an entire sound. The Praat formant analysis settings were 0.025 s for the Window length, and 5000 Hz (male) and 5500 Hz (female) for the Maximum formant.

The mean values and standard deviations of the F1 and F2 frequencies (mel) of the produced short and long /e/, /i/, /y/, and /ø/ vowels of the 14 listeners are presented in Table 3. ANOVA showed no effect of the vowel quantity on the F1 or F2 values across the four vowel categories. The Euclidean distances in the F1-F2 plane between the short and long vowels produced by the 14 subjects were 29 (SD 16) mel for /e/, 49 (SD 24) mel for /i/, 51 (SD 44) mel for /y/, and 42 (SD 31) mel for /ø/. These distances are of the order of the combined F1 and F2 difference limens (DL) reported in the literature [5, 9], indicating that the quality differences of short and long Finnish mid-front vowels spoken in citation form words are hardly noticeable.

Table 3: The mean F1 and F2 values (mel) of produced short and long /e/, /i/, /y/, and /ø/ vowels. Standard deviations are in the parentheses. dE Pa_{est} is the Euclidean distance in the F1-F2 plane between the produced vowels and estimated absolute prototypes, and dE Pω is the Euclidean distance between the produced vowels and weighted prototypes.

Vowel	F1	F2	dE Pa_{est}	dE Pω
/e/	601 (43)	1560 (113)	131 (51)	141 (53)
/i/	461 (43)	1658 (125)	176 (64)	143 (51)
/y/	445 (34)	1445 (60)	106 (49)	80 (16)
/ø/	580 (46)	1431 (62)	95 (47)	86 (46)
/e:/	602 (47)	1583 (115)	123 (44)	138 (48)
/i:/	441 (38)	1693 (133)	162 (54)	135 (52)
/y:/	436 (37)	1442 (90)	112 (55)	93 (41)
/øː/	588 (52)	1416 (75)	99 (45)	97 (40)

4. RESULTS AND DISCUSSION

The average perceptual Euclidean distance between the Finnish /e:/, /i:/, /y:/, and /øː/ categories was 218 mel (SD 15, N=14), when calculated as the mean distances between the weighted prototypes. Correspondingly, the average distances between the category centers of produced short /e/, /i/, /y/, and /ø/ vowels were 204 mel (SD 68, N=14), and of produced long /e:/, /i:/, /y:/, and /øː/ vowels 205 mel (SD 37, N=14).

The differences between individual weighted prototypes and articulated short and long vowels are presented in Figure 2. The lengths and directions of the vectors indicate that, on the average, the individual production (vector arrow) is more central and/or lower than the relevant perceptual target (vector start point). The Euclidean distances in the F1-F2 plane between the produced and perceived short and long vowels of

the 14 subjects are shown in Table 3. The mean dE Pa_{est} is 127 mel (SD 36) for short vowels and 125 mel (SD 29) for long vowels, and the mean dE Pω is 113 mel (SD 34) for short vowels and 116 mel (SD 24) for long vowels. At the group level (N=14), the produced vowels were always closest to the weighted prototypes of the category in question (Table 4).

Figure 2: Individual Euclidean distances (dE Pω) for each vowel category plotted in the F1-F2 space (mel). The upper panel represents the short and lower panel the long vowels.

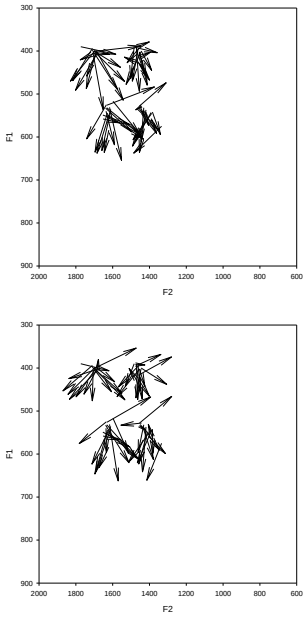


Table 4: The mean Euclidean distances (mel) of the produced short and long /e/, /i/, /y/, and /ø/ vowels from the weighted category prototypes. Standard deviations are in the parentheses.

Vowel	Pω /e/	Pω /i/	Pω /y/	Pω /ø/
/e/	**141 (53)**	269 (46)	242 (71)	178 (106)
/i/	146 (65)	**143 (51)**	213 (119)	273 (120)
/y/	205 (64)	262 (63)	**80 (16)**	124(54)
/ø/	216 (61)	329 (44)	192 (34)	**86 (46)**
/e:/	**138 (48)**	261 (47)	253 (77)	199 (109)
/i:/	213 (43)	**135 (52)**	240 (127)	311 (129)
/y:/	219 (91)	264 (88)	**93 (41)**	141 (57)
/ø:/	223 (74)	347 (61)	208 (34)	**97 (40)**

The relationship between the weighted prototypes and produced vowels was tested by using Pearson correlation. The F1 and F2 values of the weighted individual perceptual prototypes correlated significantly with the F1 and F2 values of the articulated short and long vowels: between Pω and short vowels for F1 ($r=0.860$; $p<0.01$; $df=55$) and for F2 ($r=0.666$; $p<0.01$; $df=55$), and between Pω and long vowels for F1 ($r=0.882$; $p<0.01$; $df=55$) and F2 ($r=0.708$; $p<0.01$; $df=55$).

5. REFERENCES

[1] Aaltonen, O., Suonpää, J. 1983. Computerized two-dimensional model for finnish vowel identifications. *Audiology* 22, 410-415.

[2] Benders, T., Boersma, P. 2009. Comparing methods to find a best exemplar in a multidimensional space. *Proc.10th Ann. Conf. of Int. Speech Com. Ass.*, 396-399.

[3] de Boer, B. 2000. Self-organization in vowel systems. *Journal of Phonetics* 28, 441-465.

[4] Guenther, F. 2006. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders* 39, 350-365.

[5] Hawks, J. 1994. Difference limens for formant patterns of vowel sounds. *J. Acoust. Soc. Am.* 95, 1074-1084.

[6] Iverson, P., Evans, B. 2003. A goodness optimization method for investigating phonetic categorization. *Proc. 15th ICPhS* Barcelona, 2217-2220.

[7] Klatt, D. 1980. Software for Cascade/Parallel Formant Synthesizer. *J. Acoust. Soc. Am.* 53, 8-16.

[8] Kuhl, P. 1991. Human adults and human infants show a "perceptual magnet effect" for prototypes of speech categories, monkeys do not. *P&P* 50, 93-107.

[9] Mermelstein, P. 1978. Difference limens for formant frequencies of steady-state and consonant-bound vowels. *J. Acoust. Soc. Am*. 63, 572-580.

[10] Miller, J. 1997. Internal structure of phonetic categories. *Language and Cognitive Processes* 12, 865-869.

[11] Miller, J., Connine, C., Schermer, T., Kluender, K. 1983. A possible auditory basis for internal structure of phonetic categories. *J. Acoust. Soc. Am.* 73, 2124-2133.

[12] Nábelek, A., Czyzewski, Z., Crowley, H. 1993. Vowel boundaries for steady-state and linear formant trajectories. *J. Acoust. Soc. Am.* 94, 675-687.

[13] Nearey, T. 1989. Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am*. 85, 2088-2113.

[14] Repp, B., Crowder, R. 1990. Stimulus order effects in vowel discrimination. *J. Acoust. Soc. Am.* 88, 2080-2090.

[15] Rosch, E. 1975. Cognitive reference points. *Cognitive Psychology* 7, 532-547.

[16] Strange. W. 1989. Evolving theories of vowel perception. *J. Acoust. Soc. Am.* 85, 2081-2087.

Publication II

Eerola, O., Savela, J., Laaksonen, J-P., Aaltonen, O. 2012.

The effect of duration on vowel categorization and perceptual prototypes in a quantity language.

Journal of Phonetics, Vol. 40 (2). Pages 315–328.

The effect of duration on vowel categorization and perceptual prototypes in a quantity language

Osmo Eerola[a,b,*], Janne Savela[c], Juha-Pertti Laaksonen[d], Olli Aaltonen[e,b]

[a]*Department of Biomedical Engineering, Tampere University of Technology, FI-33101 Tampere, Finland*
[b]*Centre for Cognitive Neuroscience, University of Turku, FI-20014 Turku, Finland*
[c]*Department of Information Technology, University of Turku, FI-20014 Turku, Finland*
[d]*Department of Oral & Maxillofacial Surgery, University of Turku, FI-20520 Turku, Finland*
[e]*Institution of Behavioural Sciences, University of Helsinki, FI-00014 Helsinki, Finland*

[*]Corresponding author.
Tel.: + 358 50 5016 305; fax: +358 2 2557 546; mailing address: Urheilutie 8b, FI-21620 Kuusisto, Finland.

E-mail addresses:
osmo.eerola@tut.fi (Osmo Eerola),
jansav@utu.fi (Janne Savela),
juhlaa@utu.fi (Juha-Pertti Laaksonen),
olli.aaltonen@helsinki.fi (Olli Aaltonen).

Abstract

According to the identity group interpretation of the quantity opposition in Finnish, long vowels are perceived as two successive short vowels of the same spectral quality. Some recent studies, however, challenge this general view. To investigate this, 16 listeners were first asked to categorize four sets of 19 synthesized stimuli, each set representing the Finnish vowel continuum /y/-/i/ at one of the following stimulus durations: 50 ms, 100 ms, 250 ms, and 500 ms, which cover the reported durational variations of short and long Finnish vowels. The stimuli on the /y/-/i/ continuum varied for the second formant (F2) in steps of 30 mel. Large individual variation was found in the categorization, but the category boundary F2 value and the boundary width were independent of duration in the group level, suggesting that quantity does not affect the category formation between /y/ and /i/. Normalized reaction times showed that the categorization was most difficult at 100 ms, that is, a duration that falls between a typical short and long Finnish vowel. Following the categorization task, in order to find the prototypical /i/, the same listeners were asked to evaluate the goodness of those vowels they had individually identified as /i/. The goodness rating scores and F2 frequencies of the /i/ prototypes thus found were essentially the same at all durations, suggesting that phoneme prototypes are not demonstrably dependent on the phonological quantity opposition. In conclusion, the results of this study are in accordance with the identity group interpretation of Finnish quantity opposition.

Keywords: vowel perception, phoneme prototypes, phonological quantity

1. Introduction

In *quantity* languages, such as Finnish, Czech, Estonian, Hungarian, Japanese, Mongolian, Swedish or Thai, not only the spectral quality of phones but also their duration is of importance in making judgments of phonological categories and thereby perceiving the meaning of words correctly. Finnish is an example of a contrastive quantity language where both vowels and consonants may occur independently of each other in short or long oppositions, without the quantity being bound to the word stress. For vowels, this holds for any position within a word, whereas there are certain exceptions for consonants (Suomi, 2007). The following minimal series of Finnish words demonstrates the possible occurrences of vowels and consonants in short and long oppositions: *tule-tuule-tulle-tuulle-tuullee-tuulee-tulee-tullee* [(1] (Karlsson, 1983). Native Finnish speakers normally comprehend these differences in segmental lengths easily, and therefore, one might expect that there are additional secondary cues (based on, e.g., f0 or formant frequencies F1-F3) that facilitate the distinction between a short and long occurrence of a phone. However, Finnish listeners in general ignore the possible quality differences between spoken short and long variants of the eight vowels of the Finnish vowel system: /a/, /e/, /i/, /o/, /u/, /y/, /æ/, and /ø/[(2] (Suomi, Toivanen, & Ylitalo, 2006).

- - - - - - - - - - - - - - - - -

Footnote [(1] about here

- - - - - - - - - - - - - - - - -

In written texts, the short vowels are denoted by the orthographic symbols <a>, <e>, <i>, <o>, <u>, <y>, < ä>, and <ö>, while two identical symbols indicate the long vowels <aa>, <ee>, <ii>, <oo>, <uu>, <yy>, < ää>, and <öö>. The Finnish orthography stabilized to its present form in the early 19th century and reflects the interpretation that the long segments of vowels or consonants of spoken Finnish consist of two *successive* and *identical* short segments. Karlsson (1983) refers to this interpretation as the *identity group interpretation,* and it is generally accepted in Finnish

phonetic textbooks (Suomi, Toivanen, & Ylitalo, 2006; Iivonen & Tella, 2009) as the *de facto* explanation of the phonological quantity opposition in Finnish.

One of the main implications of the identity group interpretation is that the spectral quality of the short and long Finnish vowels is assumed to be essentially the same – the distinctive difference between them is the acoustic duration, which in long vowels is twice the duration of short vowels. However, there is hardly any experimental evidence speaking for the identity group interpretation; rather, there are some reports to the opposite, as shown below in the more detailed review of literature. Therefore, the aim of this study is to examine the effect of different acoustic durations (50 ms, 100 ms, 250 ms, and 500 ms), representing the variability range of the short and long Finnish vowels, on the perception of vowel *quality* continua representing Finnish /y/ - /i/ vowels at the said durations.

- - - - - - - - - - - - - - - - -

Footnote (2 about here

- - - - - - - - - - - - - - - - -

1.1. Phoneme prototypes

In processing differences in phone quality, the best representatives of a phoneme category, also known as *phoneme prototypes*, are suggested to act as reference templates for individual quality categories. Generally, prototype based theories of perception assume that new sensory information is first processed, often in a non-linear fashion, into a particular form, which is then compared to the stored memory representations, i.e. the prototypes. Recognition takes place when the best match to a stored representation is achieved. A plethora of research reports has been published on phonetic prototypes, their relation to phonemic categorization, and the discrimination of phoneme variants close to a category boundary and within the category (e.g., Rosch, 1975; Miller, Connine, Schermer, & Kluender, 1983; Miller, 1997; Nearey, 1989; Nábelek, Czyzewski, & Crowley, 1993; Repp & Crowder, 1990; Strange, 1989). In the literature, two separate effects related to phoneme prototypes have been presented: the phoneme boundary effect, in which the sensitivity to phone

differences peaks at category borders, as shown in phone identification experiments, and the perceptual magnet effect (PME), in which the least sensitivity occurs in the vicinity of perceptual prototypes, as shown in phone discrimination experiments (Guenther & Gjaja, 1996; Iverson & Kuhl, 2000). The PME actually suggests that prototypes shrink the perceptual space around them and thereby generalize sensations to preset categories. The existence of internal structure to phonetic categories and prototypical category representatives has been shown in many reports (Miller, 1997), whereas the existence of the PME as an independent phenomenon that is not related to general perceptual contrast effects has been challenged in some articles (Lively & Pisoni, 1997; Lotto, Kluender, & Holt, 1998; Lotto, 2000); for counter-arguments, see Guenther, (2000). In a quantity language, such as Finnish, an interesting question is whether there exist spectrally different prototypes for short and long vowels, and if not, whether there is a common prototype that acts as a perceptual magnet generalizing possible spectral differences between produced short and long vowels.

1.2. The initial auditory theory of vowel perception

An important prerequisite for testing and using any prototype based theory is that the characteristic features of the stored prototypes and of the acoustic input stream are well defined and quantifiable. In their *initial auditory theory of vowel perception*, Rosner and Pickering suggest that it is the three local effective vowel indicators (LEVIs), E1, E2, and E3, which are based on the perceptual correlates of the first three physical formants (F1, F2, F3) of a vowel, and additional temporal information (D) on the physical duration (d) of the vowel, that together determine a point (E1, E2, E3, D) in the auditory vowel space (AVS) for a particular speaker (Rosner & Pickering, 1994). This theory is representative of strong auditory theories, since it is based on auditory loci in preference to physical formants. Rosner and Pickering do not present any closed form mathematical formulae for the transfer function of the time domain acoustic information to the LEVIs and D; however, they

describe some principles and introduce perceptual processes participating in this conversion (e.g., the auditory conversion of physical frequency to pitch, and the effect of speaking rate on duration in the form $D \sim d^R$, where R is the momentary speaking rate). For the purposes of the present study, we refer to two of such auditory conversions, the Hz to mel conversion, (Stevens, Volkmann & Newman, 1937), and the Hz to Bark conversion, (Zwicker & Terhardt, 1980; Traunmüller, 1990) as approximations for transforming the physical formant frequencies to the LEVIs. For the temporal information, we approximate D = d, i.e., we use the physical duration as such. In the initial auditory theory of vowel perception, the vowel identification rests on the nearest prototype rule: the listener first relies (and always can back-up to) on the *learnt* language-specific prototypes, against which he compares the speaker's AVS points. Identification then results as the best match of the speaker's AVS point with the set of the listener's prototypes. Whenever possible, the listener uses prototypes that reflect the speaker class (gender, age), and during the conversation, the listener also attempts to adjust the prototypes for a particular speaker's voice, a process that may temporarily move the prototypes away from their initial position.

Now, in quantity languages, a question of special interest in this framework is whether the LEVIs and the D of the *auditory vowel space* are independent of each other, that is, whether the AVS is an orthogonal space. In this study, we address this question in Finnish, which is a contrastive quantity language. We focus, in particular, on the relationship between E2 (as a function of F2) and D of the Finnish high-front vowels /y/ and /i/. This vowel pair was selected because it allows us to keep E1 (a function of F1) and E3 (a function of F3) constant while letting the E2 variation cause a gradual shift between the qualities /y/ and /i/ (Aaltonen & Suonpää, 1983; Aaltonen *et al*.,1997).

In terms of the AVS framework, the identity group interpretation of Finnish quantity opposition would mean that the LEVIs and D in the AVS are independent, i.e., that the space is orthogonal in

that sense. The conservative null hypothesis (H_0) of this study is formulated according to the identity group interpretation: short and long vowels are perceived similarly in terms of their spectral quality and they have similar prototypes. The alternative hypothesis (H_1) to be tested is that, because there are reports of minor spectral differences in the *produced* short and long Finnish /y/ and /i/ vowels, these differences may also be reflected in the perception of the short and long vowels.

In the world's languages there are reported quality differences (as expressed in F1 and F2 formant frequencies) between the produced short and long vowels. For a metadata analysis, we used Becker's vowel corpus (2010) and analyzed the results of 96 reports on different languages and their variants in which F2 frequency differences occur between the short and long /i/vowels produced either in isolation or as embedded in carrier words. On an average, the F2 frequency of long /i:/ vowels was 155 Hz (*SD* =155 Hz) higher than that of short /i/ vowels. The maximum difference was found in Punjabi, with the long /i:/ having 759 Hz higher F2 than the short /i/. In half of the languages, the F2 difference between short and long /i/ vowels was within the difference limen of frequency (< 3%). In 13 languages, short /i/ vowels had a higher F2 frequency than the long ones.

There are also known gender differences in the production of vowels (for a review, see Rosner & Pickering, 1994, pp. 49-73) based primarily on the shorter vocal tract of adult females, which results in greater between-category dispersion of female vowels in the F1 - F2 plane. When this anatomical difference is taken into account by using a scaling factor, there still remains a non-uniform spread of female and male vowel categories in the F1 - F2 plane: the female vowels show greater between-category dispersion especially in the /i/ and /a/ categories (Diehl *et al*., 1996). Some studies (Nordström, 1977; Goldstein, 1980) suggest that this remaining difference between

genders can be explained by articulatory behavior; female speakers prefer clear speech which results in a wider vowel triangle. Little is known whether these gender differences in production are reflected also in perception. Assuming that individual perceptual prototypes are used as articulatory targets to guide the vowel production, the observed differences in male and female production would manifest the existence of gender dependent perceptual prototypes. If this holds valid, vowel identification and goodness rating experiments should indicate gender differences both in the category dispersion and in the category internal structures in terms of F1 and F2 formants; for example, female listeners would emphasize higher F2 values for /i/ category border and /i/ prototypes than male listeners. Rosner and Pickering (1994), however, suggest in their initial auditory theory of vowel perception that the listeners rely on the speaker class specific prototypes whenever possible, which means that female listeners adjust to male speech and vice versa, thus resulting in similar (independent of F2) identification and goodness rating results between genders. We addressed this question in the present study by investigating whether male and female listeners behave differently in assessing the quality of vowels synthesized with a male voice.

1.3. Studies on the Finnish vowel system

Since the publishing of the grounding works by Wiik (1965) on the Finnish vowel system, and by Lehtonen (1970) on the quantity in Finnish, the article by Aaltonen and Suonpää (1983) was the first report to study the *perception* of the entire Finnish vowel system with a relatively large number of listeners. The /y/ - /i/ vowel continuum used in our current study is based on the results of the study by Aaltonen and Suonpää. Later, Peltola (2003) studied the perception of Finnish front vowels /i/, /e/, and /æ/, including also parts of /y/ and /ø/ categories. Savela (2009) presents identification results for synthesized Finnish vowels based on a substantial number of subjects. Table 1 summarizes the results of the above studies as regards the perceived /y/ and /i/ vowel space in terms of the first (F1) and second (F2) formant frequencies.

* * * * * * * * * * * * * * *

Table 1 about here

* * * * * * * * * * * * * * *

In the identity group interpretation, the long segments of Finnish vowels or consonants consist of two successive and identical short segments. This would suggest that the phonetic ratio of short and long segments is 1:2, an ideal pattern which would coincide with the phonological representation. However, the segmental length in Finnish is not fixed, but is extremely gradient and dependent on contextual parameters, word length, speaking rate, and speaker-specific factors (Harrikari, 2000). According to Lehtonen, and Wiik, the duration of short vowels is within the range of 60–100 ms, and that of long vowels within the range of 160–270 ms, when measured from words embedded in sentences (Lehtonen, 1970; Wiik, 1965). The corresponding phonetic ratio is 1:2.7. When measured from isolated words, the durations are slightly longer: 130–150 ms for short vowels and 250–310 ms for long vowels (Kukkonen, 1990). In Kukkonen's data from four native Finnish speakers, the mean ratio between the durations of produced short and long vowels was 1: 2.25 (variation between 1:1.7 and 1:2.4), and the mean durational differences (i.e., the category boundary width) between produced short and long vowels /u/, /y/, and /i/ were 80 ms, 111 ms, and 103 ms, respectively. In a more recent perception study (Ylinen, Shestakova, Huotilainen, Alku, & Näätänen, 2006) among native Finnish speakers, /u/ variants with a duration of less than 100 ms were perceived as short, both in a word and in an isolated vowel condition, while vowels with durations of more than 150 ms in a word context and of more than 175 ms in an isolated vowel condition were categorized as long. In that study, the mean durational ratio of perceived short and long /u/ vowels was 1: 2.03. Our earlier studies (Eerola, Laaksonen, Savela, & Aaltonen, 2002; Eerola, Laaksonen, Savela, & Aaltonen, 2003) on Finnish vowels produced by 26 subjects in an isolated word context (CVCCV and CVVCV), yielded the following durations for short and long vowels: 63 ms (*SD*=20 ms) for [y], 60 ms (*SD*=18 ms) for [i], 222 ms (*SD*=99 ms) for [y:], and 210 ms (*SD*=84 ms) for [i:]. In our studies, the mean durational ratio was 1:3.5 for both /y/ and /i/, and the mean durational difference

was 150–159 ms. The wide durational ratio (1:3.5) may partially be due to a different carrier word structure used for the short and long vowels. Further, according to the aforementioned reports, the duration difference is typically larger in isolated words than in continuous speech, since the careful pronunciation of isolated words easily prolongs the double initial vowel.

Suomi *et al.* have studied the influence of sentence accents and word stress on segmental durations in different word structures in Finnish (Suomi, Toivanen, & Ylitalo, 2003; Suomi & Ylitalo, 2004; Suomi, 2005; Suomi, 2006; Suomi, 2007). According to these studies, there are four statistically distinct, non-contrastive duration degrees for phonologically single vowels: extra short (48 ms), short (58 ms), longish (73 ms), and long (84 ms), and three degrees for double vowels: longish + longish (149 ms), long + extra short (142 ms), and very long (135 ms), indicating that, within the binary quantity opposition, there is a categorical fine structure of duration as well. The formant structures of these durational variants have, however, not been reported.

1.3.1. Acoustic correlates of the quality and quantity of spoken Finnish vowels /y/ and /i/

The results of some earlier studies on the *production* of Finnish /y/ and /i/ vowels are presented in Table 2. For example, Wiik (1965) reported clear differences in the variability ranges of Finnish single and double /y/ and /i/ vowels suggesting that the produced single vowels are more centralized than the double vowels. Unfortunately, Wiik only used five Finnish-speaking informants, and no associated statistics were published.

* * * * * * * * * * * * * * *

Table 2 about here

* * * * * * * * * * * * * * *

In a later study on vowel production by Kukkonen (1990), differences of a similar type but smaller magnitude were reported in a normal Finnish-speaking control group, but the differences were statistically significant for F1 only. In our earlier studies (Eerola, Laaksonen, Savela, & Aaltonen ,

2002), a non-significant difference of 109 Hz was found for F2 between the short and long /i/. In a more recent study by Eerola and Savela (2011), a significant difference (paired t-test, $p<0.01$, N=14) of 104 Hz was found for F2 between the short and long /i/ in uttered word pairs *tili/tiili* (‘account’/ ‘brick’), [tili/ti:li].

Iivonen and Laukkanen (1993) studied the qualitative variation of the eight Finnish vowels in 352 bisyllabic and trisyllabic words uttered by one male speaker. In their study, special attention was paid to the consonant context, vowel quantity, syllable number in word, feature structure, and auditive explanations, using the notion of the critical band (CB) of the ear (Zwicker & Terhardt, 1980). They found a clear tendency for the short vowels to be more centralized in the psychoacoustic F1 - F2 space compared to the long ones. However, except for the /u/ - /u:/ pair, this difference was smaller than one critical band, and thus was auditorily negligible. Interestingly, although the data come from one speaker only, the dispersion of F1 and F2 values on the F1 - F2 space was clearly larger for short vowels than for long ones; e.g., the standard deviations of different uttered short [y] and [i] vowels were 0.52 Bark and 0.42 Bark but only 0.27 Bark for [y:] and 0.32 Bark for [i:]. In a comparative study of the monophthong systems in Finnish, Mongolian, and Udmurt, Iivonen and Harnud (2005) report on minor spectral differences in the short/long vowel contrasts in stressed (e.g. [sika] / [si:ka] (‘pig’ / ‘whitefish’)) and non-stressed (e.g. [etsi] / [etsi:] (‘sought’ / ‘seeks’)) syllables in Finnish uttered by one male speaker; the biggest differences between short and long vowels are found in /u/. As in the study by Iivonen and Laukkanen, the [u] is more centralized and does not overlap with [u:]. Also for /y/ and /i/, the short vowels are more centralized than their longer counterparts, but now the short and long vowel versions are overlapping on the F1 axis. Interestingly, the /y/ and /i/ vowels, both short and long, also overlap on the F2 axis instead of being clearly separate phoneme categories.

To summarize, minor spectral differences have been reported in the first (F1) and second (F2) formant frequencies of the produced short and long Finnish vowels, and this difference is largest between the high back vowels [u] and [u:].

1.3.2. Studies on perception of short and long Finnish vowels

Recent studies on the *quantity* discrimination of the single and double Finnish vowels suggest that the pitch contour may play a role in the quantity differentiation. For example, in a two-alternative forced-choice categorization experiment, Järvikivi *et al.* (2007), and Järvikivi, Vainio, and Aalto (2010) studied the perceived vowel duration in the stressed initial syllable (CV and CVV) of Finnish word pairs *sika/siika* ('pig'/ 'whitefish'), [sika/si:ka], *kisu/kiisu* ('kitten'/ 'ore'), [kisu/ki:su], *Mika/Miika* (male names), [Mika/Mi:ka], *kato/kaato* ('loss'/ 'fall'), [kato/ka:to], and *pika/piika* ('instant'/ 'maid'), [pika/pi:ka]. For the initial vowel, they used five different durations: 75 ms, 100 ms, 125 ms, 150 ms, and 175 ms, and two alternative f0 patterns: an even high pitch throughout the vowel or a dynamic fall contour. For the intermediate durations (100 ms, 125 ms, and 150 ms), the listeners were more likely to categorize the vowel of the first syllable as long [V:] in the dynamic fall condition than in the even high pitch condition. Thus, not only duration but also the tonal structure was used as a perceptual cue for the quantity opposition at the intermediate durations. However, the pitch pattern did not affect significantly the categorization for the extreme durations (75 ms and 175 ms), representing the single and double quantities most markedly. Apparently, at the extreme ends, the duration alone was a sufficiently strong cue and overran the mismatching f0 cue.

Furthermore, O'Dell (2003) questions the plain quantal nature of the duration opposition. In one experiment, O'Dell synthesized two continua of eleven stimuli, the first one using the qualitative parameters (including f0) of the short [u] vowel in the word *tuli* ('fire', [tuli]), and the second one

using those of the long [u:] in *tuuli* ('wind', [tu:li]) as the basis. Twelve listeners were requested to categorize the stimuli on the two continua as either /tuli/ or /tuuli/. If the vowel duration were the only cue for the quantity opposition, then the same durational variant should presumably form the category boundary in both series. This, however, was not the case, but the category boundaries were three duration steps apart in the two series. O'Dell also found that the formant structure between [u] and [u:] differed, with [u] being more centralized, i.e., F1 and F2 were higher than in [u:]. This is in line with the study by Iivonen and Laukkanen (1993). However, O'Dell suggests that this centralization is caused by a shorter acoustic duration, not by the phonological quantity of the vowel, an explanation that means that single and double vowels would have the same articulatory target, which is not met in articulating the single vowels.

Meister and Werner (2009) used isolated synthetic vowels in the close-open (F1) dimension to examine the micro-durational variations in perception among Finnish (N=10) and Estonian (N=10) listeners. Finnish and Estonian are phonetically closely related, and they both are quantity languages. In the experiment, the vowel duration varied between 60 ms and 140 ms in steps of 20 ms, and f0 was held constant at 100 Hz (NB: the durational range applied in the experiment does not necessarily cover the wide variation of Finnish short and long vowels in its entirety). By using a multiple forced-choice ABX setup (A and B were the category prototypes, X was an ambiguous stimulus between categories), it was found that openness correlated positively with stimulus duration in the high-mid vowel pairs (/i/-/e/, /y/-/ø/, and /u/-/o/); the longer the duration of the ambiguous stimulus (on the F1-F2 category boundary area), the more likely it was to be categorized as the more open vowel of a pair. In case of the mid-low vowel pairs (/e/-/æ/, /o/-/a/) a similar effect was found for only some Finnish subjects, while for the Estonian listeners the stimulus duration did not affect the perception of vowel categories significantly, a difference that was argued to be language specific. The results of Meister and Werner thus suggest that duration may affect the

perception of vowel quality; for example, the perception of a between category token in the /i/ - /e/ continuum is driven towards /e/ when associated with prolonged duration as a quantity cue. In other words, while the spectral quality of the stimulus remains the same, an increase in its duration widens the perceptual distance from the /i/ prototype, resulting in a better match to /e/.

On the basis of the literature discussed above one can conclude, first, that there are minor differences in the spectral properties between the *produced* short and long Finnish /y/ and /i/ phonemes suggesting that the short uttered phonemes are more centralized than the long ones, and that there are substantial differences in the F2 formant frequencies of produced short and long /i/ vowels. Second, according to most of the reports, the duration of the single Finnish /y/ and /i/ vowels is typically less than 100 ms, and the duration of the double vowels is more than 130 ms. In continuous speech, the absolute durations depend mainly on the speaking rate, but nevertheless, the duration ratio between short and long vowels is on the order of 1:1.5 to 1:3.5. Third, there are actually more than two quantity degrees in Finnish vowels, although only two form a phonological opposition. Furthermore, some recent *perception* studies question the general assumption that Finnish single and double vowels are similar in quality. The earlier studies on the Finnish vowel quality and quantity leave open such questions as to what extent the durational and qualitative properties interact in the formation of phoneme categories and their internal structures, and whether the vowel quality is statistically independent of quantity. In the following, we report on the results of two experimental trials carried out to investigate the possible impact of vowel duration on the categorization of synthetic /y/ - /i/ vowels (Experiment 1) and on the goodness rating of the categorized /i/ vowels (Experiment 2).

2. Experiment 1: Categorization

The purpose of the categorization experiment (Experiment 1) was to study the possible effect of vowel duration on the categorization of stimuli representing the Finnish /y/-/i/ continuum. To investigate this, 16 listeners were asked to categorize four sets of 19 synthesized stimuli, each set representing the Finnish vowel quality continuum /y/-/i/ at one of the following stimulus durations: 50 ms, 100 ms, 250 ms, and 500 ms, which cover the reported durational variation of short and long Finnish vowels. The vowel quality was varied by means of the second formant, while the other formants were held constant. Hence, only two acoustic variables, duration and F2 frequency, formed the independent variables in Experiment 1 (NB: for f0, see section 2.1.2.).

According to the identity group interpretation of the Finnish quantity opposition, the vowel duration does not influence the auditory perception of those spectral properties of the stimuli that form the basis for stimulus classification into the *a priori* learnt phonological quality categories of the Finnish language. However, as presented in the preceding literature review, minor spectral differences in the produced short and long Finnish /y/ and /i/ vowels have been reported, and furthermore, some perception studies indicate that quantity may affect the categorization of Finnish vowel quality. Therefore, our hypothesis (H_1) to be tested in Experiment 1 was that the category border between /y/ and /i/ is located differently for those stimulus durations that represent either the short or the long Finnish /y/ and /i/ vowels. If this is not supported by the results, the null hypothesis (H_0) will remain valid, in other words, the category border between /y/ and /i/ is located at the same place in the F2 stimulus continuum independently of the duration of the stimuli.

We further assumed that not only the category border, but also the categorization *process*[3] would be influenced by the stimulus duration. We used reaction times (RT) and the response rate as measures reflecting the categorization process. It was expected that listeners would categorize faster

and more consistently the stimuli that represent typical short and long Finnish vowels, or alternatively, those stimuli that are acoustically longer. The former case would indicate that the *quantity* prototypes of short and long vowels along the same /y/ - /i/ quality continuum affect, e.g., the *speed* of categorization to /y/ or /i/. The latter case is known as the cue-duration hypothesis: the categorization of vowel variants is presumed to be easier with longer stimuli because there is more time and more cues available for extracting the relevant features from the presented stimuli (Pisoni, 1973; Repp & Liberman, 1987).

- - - - - - - - - - - - - - - - -

Footnote [3] about here

- - - - - - - - - - - - - - - - -

2.1. Methods

2.1.1. Listeners

Sixteen adults with no reported hearing defects and all fluent speakers of modern educated Finnish of South-West Finland volunteered as listeners. Both genders were represented (9 males and 7 females), and the mean age at the time of the recordings was 27 years (range 19-44 years). Since vowels produced by female speakers show greater between-category dispersion, especially in the /i/ and /a/ categories (Diehl *et al*., 1996), gender was applied as an independent variable in order to investigate whether there are differences in categorization and goodness rating between male and female listeners for stimuli synthesized with a male voice.

2.1.2. Stimuli

Synthetic vowels presented in isolation were used in both experiments. Except for the duration and f0 contour, the synthesis parameters were the same as used in our earlier experiment (Aaltonen, Eerola, Hellström, Uusipaikka, & Lang, 1997). In order to cover the typical ranges of short and long Finnish vowels, durations of 50 ms, 100 ms, 250 ms, and 500 ms were selected for the stimuli. The ratio between the Finnish single and double vowel durations is of the order of 1:1.5 to 1:3.5. Hence,

when the stimulus duration doubles from one set to another, the steps between the stimuli are sufficiently large (> 1:1.5), and yet, the resolution over the entire durational range is appropriate for us to see possible effects suggested by the cue-duration theory.

The quality of the Finnish closed front vowels /i/ and /y/ is mainly dependent on the frequencies of two formants, F2 and F3, but variations in F2 alone are sufficient for the listeners to categorize the stimuli either as /i/ or /y/ (Aaltonen & Suonpää, 1983). Therefore, and in order to limit the number of independent acoustical variables, we used stimuli that varied only in the frequency of F2. For each duration, 19 vowel variants in the continuum of Finnish /y/-/i/ were synthesized using a parallel mode speech synthesizer (Klatt, 1980) embedded in a UNIX workstation. The F2 value varied from 1520 Hz to 2966 Hz, covering the following critical bands: 1480 Hz - 1720 Hz (Bark 11), 1720 Hz - 2000 Hz (Bark 12), 2000 Hz - 2320 Hz (Bark 13), 2320 Hz - 2700 Hz (Bark 14), and 2700 Hz - 3150 Hz (Bark 15) (Zwicker & Terhardt, 1980; Traunmüller, 1990). The 19 stimuli differed from each other in equal steps of 30 mel in the psychoacoustic F2 frequency scale (Stevens, Volkmann & Newman, 1937). This auditory frequency conversion was used as an approximation for transforming the physical formant frequency (in Hz) of F2 to LEVI E2 (in mel). A 30-mel step corresponds to 60 Hz at 1500 Hz, 75 Hz at 2000 Hz, 88 Hz at 2500 Hz, and 102 Hz at 3000 Hz, and it was considered to be a proper step size to reveal possible F2 differences between single and double Finnish [y] and [i] vowel variants. The other formants were fixed at the following frequencies: F1 = 250 Hz, F3 = 3010 Hz, F4 = 3300 Hz, F5 = 3850 Hz.

A flat f0 at 112 Hz was used for the shorter durations of 50 ms and 100 ms, whereas a rise-fall contour of f0 was used for the longer durations of 250 ms and 500 ms in order to obtain a more natural sounding synthesis result. Here, a choice had to be made between two adverse prerequisites: stimulus naturalness (fidelity) and stimulus uniformity between different durations. Because

goodness rating and finding the prototypical variants were essential in Experiment 2, the stimulus naturalness was chosen. Additionally, use of flat f0 for all durations could have jeopardized the interpretation of results because the non-normal (flat) f0 might affect the perception of the longer stimuli. Consequently, for the 250 ms stimuli, we used an f0 that rose from 112 Hz to 122 Hz during the first 50 ms and dropped to 102 Hz during the remaining 200 ms of the vowel duration. For the longest, 500 ms stimuli, f0 rose from 112 Hz to 132 Hz in 100 ms and dropped to 92 Hz during the remaining 400 ms of the vowel duration. The stimulus onsets and offsets were smoothed with linear 5 ms, 10 ms, 15 ms, and 30 ms windows (for the 50 ms, 100 ms, 250 ms, and 500 ms stimuli, respectively).

2.1.3. Procedure

Each listener participated in four randomized sessions, one for each vowel duration. The stimulus presentation order was randomized for each listener prior to the experiments. Since the aim of Experiment 1 was to examine whether different stimulus durations would affect the categorization of the /y/ - /i/ continuum, without being influenced by any prior knowledge or currently available information about the quantity differences of the vowel stimuli, only stimuli of the same duration were used in each session. The time between the sessions varied from a day to around a week. Our earlier experiments have shown that repeated categorizations vary only little from session to session (Aaltonen, Eerola, Hellström, Uusipaikka, & Lang, 1997). Therefore, repetitions with the same duration were omitted in order to keep the number of sessions reasonable and to avoid possible learning effects.

The stimuli were played with a NeuroStim PC-based stimulus presentation device at 10 kHz playback rate. A 12-bit digital-to-analogue converter with an integrated reconstruction filter fed the stimuli through the calibrated insert earphones (Ear-Tone 3A) at a sound-pressure level of 75 dB (A). The audio system was calibrated with a Brüel & Kjaer artificial ear (Type 4152) and a

precision sound level meter (Type 2230). The listeners were seated in a quiet sound-proof room (sound-pressure level of ambient noise was lower than 40 dB (A)).

The 19 vowel variants of each duration block (50 ms, 100 ms, 250 ms, and 500 ms) were played in a random order, 15 times each (i.e., 15 x 19 =285 stimuli in each of the four sessions), with a maximum inter-stimulus interval (ISI) of 2000 ms. Upon hearing the stimulus, the listeners were to categorize it by pressing one of the two response buttons (labeled as "y" or "i") of the NeuroStim response device. The next stimulus was triggered by the listener pressing the button, or alternatively, once the set ISI had elapsed. Any responses given after the 2000 ms period were marked as "non-responded" stimuli. One half of the listeners used the left thumb for "y" and the right thumb for "i", and the other half did the opposite. Reaction time was determined as the time measured from the stimulus onset to response, i.e., pressing the button (Bamber, 1969; Leibold & Werner, 2002; Reed, 1975), and the RTs were recorded with the NeuroStim device.

2.1.4. Analysis

For each listener, the category scoring percentages and reaction times versus F2 frequency were plotted in categorization graphs, separately for the different durations. The following measures characterizing the categorization were analyzed or calculated from the recorded raw data for each duration and individual: the F2 value of the category boundary (CB) in Hz, the width of the boundary area (BW) in Hz, the reaction times (RT) in seconds (for the sake of clarity, RTs are in s and the stimulus durations are in ms), and the proportion of responses given (response rate). Thus, the dependent variables used in the statistical analysis were as follows: F2 of CB, BW, RT, and response rate. The Probit non-linear curve fitting method (Bliss, 1934; Finney, 1944) available in the SPSS statistical software was applied for determining the CB and BW from the individual categorization data. Since CB is by definition the F2 value at the 50%/50% intersection for /y/ and /i/ identifications, the BW was determined, for each listener and each duration, as the mean F2

difference at the points of 75% for /y/ and /i/, and correspondingly, 25% for /i/ and /y/ identifications (see Fig. 3).

Reaction time is an established behavioral measure used in categorical perception (CP) studies. According to the CP theory, RTs are longer at the category boundary (CB) than within a category. This was first tested by comparing the RTs measured for those stimuli that fall clearly (> 90%) within the /y/ and /i/ categories against the RTs measured at the CB. The stimuli (with varying F2) and corresponding RTs, representing either the categories (> 90%) or the CB (<75%), were selected manually. The analysis was done by using Student's two-tailed t-test for two-sample sets with unequal variances (the reaction time variation at the CB differs from that within the category). Because the measured RTs could obviously be biased by the stimulus duration, which was used as the treatment in the experiment, some type of bias subtraction or normalization was necessary for the purpose of making the RTs at different stimulus durations more comparable. Subtracting the stimulus duration from the total RTs does not necessarily solve the bias problem: for longer stimuli, the listener may press the button while the stimulus is still on. Therefore, two additional measures characterizing the RTs were derived: 1) reaction time at the CB as compared to the mean RT of all presented stimuli in the continuum: $t_a = t_{CB} / t_{tot}$ and 2) reaction time at the CB as compared to the mean RT within the /y/ and /i/ categories: $t_b = t_{CB} / t_{cat}$. These two measures were also compared for their applicability regarding this kind of normalization: the former (t_a) obviously would take into account the RTs to stimuli on the entire continuum, whereas the latter (t_b) should emphasize the RT differences between stimuli at CB and within a category.

The number of non-responded stimuli is a potential measure for the consistency of categorization since it suggests either a slow general reactivity or difficulty categorizing the stimuli. In presenting

the results, we used the response rate (= 100% – non-responded stimuli %) to better indicate the percentage of stimuli for which responses of [y] and [i] were obtained.

Finally, all the measures and their derivatives were subjected to a repeated measures analysis of variance (ANOVA), with duration as the within-subjects factor and gender as the between-subjects factor. The statistical significance level $p<0.05$ was used throughout the experiments, unless otherwise mentioned. For such data sets that were not normally distributed, as tested with the Shapiro-Wilk test, non-parametric tests were used instead of an ANOVA (as explained in the relevant points in text).

2.2. Results and Discussion

2.2.1. Category boundary F2

The individual categorization results demonstrate that all the listeners were able to make the categorization, although the plot shapes of the listeners vary greatly in terms of the consistency of categorization: some listeners categorized the stimuli distinctly as /y/ and /i/, with only a few stimuli falling between categories (Fig. 1). Others were less certain in their categorization, resulting in a wider CB area between categories and in a more fluctuating categorization curves (Fig. 2). Only three listeners distinguished between [y] and [i] variants with an excellent accuracy at the CB and yielded very even categorization plots across the board for all the four durations. Four listeners had difficulties with the categorization and, in general, performed poorly with all durations. Five listeners improved clearly in their performance when the duration became longer.

* * * * * * * * * * * * * * * *

Fig. 1 and Fig. 2 about here

* * * * * * * * * * * * * * * *

We do not have a good explanation for the differences in the categorization performance. Nábelek, Czyzewski, and Crowley (1993) report a similar finding in their study with ten normal and ten hearing-impaired English-speaking listeners in an identification trial of the /I/ - /ε/ continuum. In our study, the listeners had no reported hearing impairments, so it does not explain the uncertainty observed in the poor categorizers. Similar variation in certainty was found in our earlier experiment (Aaltonen et al., 1997), in which the performance differences were also replicated in repeated runs, thus excluding a diminished concentration as a likely reason. Possible remaining reasons are that the used stimulus continuum /y/ - /i/ was not perceived as representative by all listeners, or that some of the listeners perceived the synthetic stimuli as unnatural and difficult to categorize, or that there were factual perceptual differences between the listeners, just like there are differences in musical talent. The last possibility suggests that in future research more attention should be paid to the individual differences in phoneme perception.

The averaged category scoring and reaction time curves of the four sessions (50 ms, 100 ms, 250 ms, and 500 ms) for all the 16 listeners are presented in Fig. 3a-d. At the shortest stimulus duration of 50 ms, the labeling changes over from /y/ to /i/ smoothly when F2 increases, the scoring curves are symmetric, and the RT is clearly longer at the boundary and drops to the lowest values in the middle of categories (Fig. 3a). This is in accordance with the earlier finding that categorization is consistent and precise when the stimulus duration is just long enough to trigger the recognition of the correct category (Pisoni, 1973). At the 100 ms duration, the identification of the /y/ stimuli at low F2 values is less consistent in comparison to the 50 ms duration, and the RTs are longest near the /y/ category and decrease clearly towards the center of the /i/ category (Fig. 3b). With the two longer durations (250 ms and 500 ms), the /y/ and /i/ categorization plots are similar, but with 250 ms the reaction time curve has a sharper peak at the CB (Figs. 3c and 3d).

* *

Fig. 3 about here (lay-out 2 x 2 panels: 3a and 3b top, 3c and 3d bottom)

* *

* * * * * * * * * * * * * * *

Table 3 about here

* * * * * * * * * * * * * * *

The numerical data at the group level are summarized in Table 3. When estimated with the Probit curve fitting method from individual results and then averaged for group results, the category boundary (CB) values are 2065 Hz (50 ms), 2049 Hz (100 ms), 2077 Hz (250 ms), and 2094 Hz (500 ms). These values fall below the 30 mel stimulus difference that was used in the experiment. The analysis of variance revealed that the location of the interpolated CB on the F2 axis does not depend on the duration of the stimuli at the group level ($F(3,42) = 1.490$; $p = 0.231$*; partial* η^2 *=0.096*). The results of male and female listeners did not differ significantly from each other ($F(1,15) = 0.050$; $p = 0.826$*; partial* η^2 *=0.004*), indicating that the stimulus continuum synthesized with a male voice is categorized similarly by males and females.

2.2.2. Boundary width

The mean values and standard deviations of the category boundary widths (BWs) are presented in Table 3. These BW values in Hz correspond, on the average, to a bandwidth, which is two to three times the 30 mel stimulus step used in the experiment. Because the BW values for 16 subjects were not normally distributed, the Friedman test was applied to test the dependency of BW on duration. The result was not significant (Friedman $\chi^2 = 2.553$; $p=0.466$; df=3), thus indicating that the BW does not depend on stimulus duration. Interestingly, the BW of male listeners (N=9) was narrower than the BW of female listeners (N=7) at other durations except 250 ms: at 50 ms for male 166 Hz

(*SD*=51), and for female 323 Hz (*SD*=158 Hz), at 100 ms for male 171 Hz (*SD*=50 Hz), and for female 217 Hz (*SD*=147 Hz), at 250 ms for male 194 Hz (*SD*=78 Hz), and for female 175 Hz (*SD*=87 Hz), and at 500 ms for male 142 Hz (*SD*=61), and for female 210 Hz (*SD*=170 Hz). However, the Mann-Whitney tests, which were run for each duration with gender as a group factor indicated that the result was significant only for 50 ms (for 50ms: U=12.00; p=0.042, for 100ms: U=25.0; p=0.536, for 250ms: U=23.0; p=0.408, for 500ms: U=26.50; p=0.606).

Aaltonen *et al*. (1997) found in their study, using a stimulus duration of 500 ms, that listeners were able to make a judgment between [y] and [i] with F2 differences close to the standard critical bandwidth, that is, one Bark on the F2 scale. To investigate if this is applicable to shorter stimulus durations used in the present study as well, we calculated the critical band rate (CBR) for each CB F2, and then formed the ratios of category boundary width to this critical band rate (BW/CBR). The mean values and confidence intervals (99%) for the BW/CBR ratios were 0.78 (0.60–0.95) at 50 ms, 0.71 (0.52–0.9) at 100 ms, 0.68 (0.53–0.82) at 250 ms, and 0.70 (0.35–1.02) at 500 ms. Thus, the average BW/CBR ratio was approximately 0.7 and the ratio decreased with increasing duration, although this dependency was not significant. This means that the listeners were, in general, able to make their judgment within one critical band rate (BW/CBR < 1.0) at all durations. This is in line with the findings of Aaltonen *et al*. (1997).

2.2.3. Reaction times

The averaged RTs (N=16) are presented in Table 4. Separately for each duration and individually for each listener, the RTs to stimuli at the category boundary (t_{CB}) were compared with the RTs to the stimuli within a category ($t_{/y/}$, $t_{/i/}$), and the difference was tested by *t*-test. Typically, the RTs were 0.25- 0.30 s longer at the boundary than within a category. The difference was highly significant ($p < 0.001$) for all durations and listeners, and in accordance with the earlier findings concerning categorical perception.

* * * * * * * * * * * * * * *
Table 4 about here
* * * * * * * * * * * * * * *

Because the RTs were not normally distributed, the Friedman test was performed instead of ANOVA. The duration had a significant effect (Friedman χ^2=9.150; p=0.027; df=3) on the mean RT; this result is obvious and due to the longer RTs at the 500 ms duration [4. Therefore, in order to solve the possible bias problem in comparing the measured reaction times to stimuli of varying lengths, two normalized RT ratios were formed for each listener and each duration: $t_a = t_{CB} / t_{tot,}$ and $t_b = t_{CB} / t_{cat}$. The former (t_a) is the ratio of the RT at the CB (t_{CB}) to the overall mean RT (t_{tot}), and the latter (t_b) is the ratio of the RT at the CB to the mean within-category RT of /y/ and /i/ category stimuli, respectively. The ANOVA analysis of the normalized RT ratios across the 16 listeners showed that both t_a and t_b were significantly dependent on duration: $F(3,42) = 4.037$; p = 0.013; *partial* η^2 =0.0210 for t_a, and (Huynh-Feldt corrected) $F(2.395,42) = 3.816$; p =0.026; *partial* η^2 =0.214 for t_b. The durations were further compared pair-wise: For t_a, the 100 ms stimuli were at the category boundary processed at a significantly slower rate in comparison to the 50 ms (p = 0.039), 250 ms (p = 0.014), and 500 ms stimuli (p = 0.021). Correspondingly, for t_b, the 100 ms stimuli were at the category boundary processed at a significantly slower rate in comparison to the 250 ms (p = 0.016) and 500 ms stimuli (p = 0.025).

- - - - - - - - - - - - - - - - -
Footnote [4 about here
- - - - - - - - - - - - - - - - -

The effect of RT normalization is interesting; it appears that, among the 50 ms, 100 ms, 250 ms, and 500 ms stimulus durations, the 100 ms stimuli are the most difficult to categorize either as /i/ or /y/ although the results of the categorization process (i.e., the CB and BW values) remain the same. In other words, at the 100 ms stimulus duration, the time used by the listener to make the

categorization at the (quality) category boundary increases to a higher extent in relation to the overall RT or to the within-category RT than at the other durations of 50 ms, 250 ms, and 500 ms. The result suggests that vowels with duration of 100 ms, which according to earlier reports (see section 1.3.) represent the borderline duration between the short and long Finnish vowels, may be perceived differently and processed at a slower rate than the vowels representing more clearly either the short or the long Finnish vowels.

2.2.4. Non-responded stimuli

As described above in section 2.1.4, there was a limited time window of 2000 ms for responding to the stimuli. If no response was detected by the recording system within that time, the stimulus in question was marked as "non-responded". The response rate (given as percentage, 100% = all responded) was afterwards calculated by subtracting the number of non-responded stimuli from all presented stimuli (N = 15 for each stimulus variant). The average response rates were 93% for 50 ms, 92.5% for 100 ms, 96.0% for 250 ms, and 97.5% for 500 ms. Because the response rates were not normally distributed, the Friedman test was performed instead of ANOVA. The test showed significantly (Friedman χ^2=15.382; p=0.002; df=3) higher response rates at longer durations. This result is in accordance with the cue-duration hypothesis. The Mann-Whitney tests were used for each duration, with gender as a group factor: none of the values was significant (for 50ms: U=27.0; p=0.633, for 100ms: U=20.5; p=0.244, for 250ms: U=26.0; p=0.559, for 500 ms: U=26.0; p=0.559), thus indicating that there were no differences between the genders.

In summary of Experiment 1, large individual variation was found in the categorization, but the category boundary F2 value and the boundary width were independent of duration in the group level, suggesting that quantity does not affect the category formation between /y/ and /i/. Further, the listeners were, in general, able to make their judgment within one critical band rate (BW/CBR <

1.0) at all durations. Male listeners showed significantly narrower BWs at 50 ms durations compared to female listeners, however, no other significant differences were found between the genders. Normalized reaction times showed that the (quality) categorization was most difficult at 100 ms, that is, a duration that falls between a typical short and long Finnish vowel.

3. Experiment 2: Goodness rating

The purpose of the goodness rating experiment (Experiment 2) was, first, to find the prototypical [i] variants within each listener's individual /i/ category, as determined in Experiment 1, at the four durations of 50 ms, 100 ms, 250 ms, and 500 ms, and, second, to study the possible effect of duration on the perceptual quality differences and on the F2 values of these prototypes.

According to hypothesis H_1, the experiment was expected to reveal significant F2 differences in the prototypical [i] phonemes at different durations. Assuming that there are 63 Hz–200 Hz differences (see Table 2) in the F2 values of the produced single and double Finnish /i/ vowels, similar F2 differences should be found in the perception of these vowels, as well; in other words, the prototypical [i] variants should differ from the prototypical [i:] variants in terms of F2. We also hypothesized that the goodness ratings would vary at different durations so as to reflect the cue-duration hypothesis, i.e., that the longer durations achieve higher ratings. The conservative null hypothesis (H_0) of Experiment 2 was, in compliance with the identity group interpretation, that duration does not influence the goodness ratings and the F2 values of the prototypical variants, but rather that the short and long vowels are perceived similarly.

3.1. Methods

The same sixteen adults as in Experiment 1 volunteered as listeners, with the exception that in Experiment 2 one listener did not participate in the 250 ms session, and was excluded from the analysis (N=15, 8 males, 7 females). As the purpose of the goodness rating experiment was to find the best ranked stimulus variants (prototypes) within each listener's individual /i/ category, and to investigate whether these prototypes vary with duration, only those synthesized stimuli of Experiment 1 were used that the listeners had consistently categorized as /i/ in more than 75% of

cases. Thus, in Experiment 2, the number of stimuli representing the /i/ category varied between the listeners, and also between the durations in some individual listeners.

The variants representing consistently the [i] phonemes of the individual /i/ categories were presented in a random order, 15 times each, in four separate sessions, one for each duration. The listeners were asked to rate the stimuli using the scale from 1 to 7 (1 = a poor category exemplar, 7 = a good category exemplar) and mark the score on a form sheet. The stimulus presentation was self-paced, with the minimum ISI set at 2000 ms (i.e., it was not possible to trigger the next stimulus until 2000 ms had elapsed). The goodness ratings (1–7) were first saved in a computer database, and the mean rating scores versus the F2 frequency were calculated. For each listener and each duration, the stimulus with the highest rating was labeled as the candidate prototype (P) and the one with the lowest rating as the non-prototype (NP). The significance of the difference in the mean ratings between the P and NP stimulus variants (N=15) was then t-tested for each listener and each duration. A significant difference ($p<0.05$) was required between P and NP ratings for regarding P as a representative category prototype (Kuhl, 1991). The mean goodness scores and the F2 frequencies (in Hz) of the prototype stimuli were subjected to a repeated measures analysis of variance (ANOVA), with duration as the within-subjects factor and gender as the between-subjects factor.

3.2. Results and discussion

Examples of goodness ratings within the individually scored /i/ category are presented in Figs. 4, 5, and 6. Three different types of curves emerged for goodness ratings (scoring value versus F2 frequency). The most common curve type (see Table 5) across all durations was a “hill” curve, where the highest scoring stimuli occur in the middle of the individual F2 continuum of [i] vowels (Fig. 4). This curve type represents a category structure similar to that obtained by Kuhl (1991). The second most frequent curve type was a “down” curve with the most prototypical [i] vowels

occurring close to the category boundary against /y/ (Fig. 5). The least frequent curve type was the "up" curve with the prototypes occurring at the other extreme, i.e., at the highest F2 values in the continuum (Fig. 6). This curve type represents a category structure similar to that reported by Lively (1993). The differences in the /i/ category internal structures are similar to those found in our earlier studies (Aaltonen *et al*., 1997) with long /i/ vowels (500 ms). For the "up" type listeners, the hyper-space effect offers another possible explanation: in the goodness evaluation, they may prefer stimuli with higher F2, resembling hyper-articulated vowels rather than vowels of normal effortless speech (Johnson, Flemming, & Wright, 1993).

* *
Fig. 4, Fig. 5, and Fig. 6 about here
* *

The mean goodness ratings of the 15 listeners for all stimuli, and separately for the prototype (P) and non-prototype (NP) stimuli, at the durations of 50 ms, 100 ms, 250 ms, and 500 ms are presented in Table 5. All the listeners were able to give a consistent quality evaluation of the vowel variants that they had earlier in the categorization task labeled as members of the /i/ category in the sense that in all cases the mean ratings were significantly higher for prototypes than for non-prototypes ($p < 0.01$).

* * * * * * * * * * * * * * *
Table 5 about here
* * * * * * * * * * * * * * *

At the group level, the averaged score value for all vowel samples was 4.1on the scale 1–7, the prototypical [i] was scored as 5.68 and the non-prototypical [i] as 1.80, on the average. The individual scores of the prototypical [i] were subjected to a repeated measures analysis of variance (ANOVA), with duration being the within-subjects factor and gender the between-subjects factor. No duration-dependent main effect on stimulus ratings was found ($F(3,39) = 2.073$; $p = 0.120$; *partial* $\eta^2 = 0.138$). Nor did the listener's gender affect the ratings ($F(1,13) = 0.224$; $p = 0.976$;

partial η^2 =0.017). However, pair-wise comparisons showed that there was a significant difference (p = 0.041) between the goodness ratings at the durations of 50 ms and 100 ms, indicating that while the shortest stimulus duration of 50 ms is long enough for a listener to identify the best vowel exemplar from a set of stimuli representing the same phoneme category, a significant increase in the goodness rating is achieved by doubling the duration from 50 ms to 100 ms, but not any more for prolonging from 100 ms to 250 ms or from 250 ms to 500 ms.

As can be seen from Table 5, the mean F2 values of the prototypical [i] vowels at different durations ranged from 2493 Hz (50 ms) to 2561 Hz (500 ms). The biggest F2 frequency difference thus was obtained between the shortest and longest duration, and was 68 Hz (non significant). This is of the order of F2 differences in produced short and long /i/ vowels reported by Kukkonen (F2 is 63 Hz higher in long /i/), but much less than the values reported by, e.g.,Wiik (140 Hz), and about half of the average (118 Hz) of the earlier reported F2 differences between short and long Finnish /i/ (for details, see Table 2). The individual F2 values of the prototypical [i] vowels were subjected to a repeated measures analysis of variance (ANOVA), with the duration being the within-subjects factor and gender the between-subjects factor. Neither the duration of the stimulus nor the listener's gender had any significant main effect on F2: $F(3,42) = 0.931$; $p = 0.435$; *partial* η^2 =0.067 for duration, and $F(1,13) = 1.386$; $p = 0.260$; *partial* η^2 =0.096) for gender. To summarize, the F2 frequencies of the highest scoring (prototypical) stimuli are not statistically dependent on duration, suggesting that the phonological quantity categories do not influence significantly the perception of quality differences within a particular vowel category.

Another interesting question is whether the perceptual prototype has an inherent minimum RT within a category. If there were a clear minimum RT for the prototype stimulus, the RTs could be

used to disclose the category prototypes directly from the categorization data and the subsequent goodness rating experiment could be omitted. In Experiment 1, within the /i/ category, the shortest RTs were recorded to stimulus 16 (F2 = 2672 Hz) at the duration of 50 ms, 100 ms, and 500 ms, and to stimulus 17 (F2 = 2767 Hz) at the duration of 250 ms (see Table 4). However, in Experiment 2, stimuli 16 and 17 were not among the prototype stimuli, while they were 30 mel – 60 mel higher in F2 than the best rated [i] variants (see Table 5). The results indicate that even if there are differences between the within-category stimuli, as measured by reaction times in a categorization task, the stimuli showing the shortest reaction times are not necessarily identical with the prototypical stimuli emerging in a dedicated goodness rating setting.

4. General discussion and conclusion

The conservative null hypothesis (H_0) of this study was that, in spoken Finnish, the perceived vowel quality is independent of vowel quantity, as formulated in the identity group interpretation of Finnish quantity opposition by Karlsson (1983). The main results of this study leave the null hypothesis valid: In Experiment 1, duration had no significant effect on the location and width of the /y/-/i/ category boundary (on the F2 axis), and in Experiment 2, duration had no significant effect on either the F2 value or the goodness rating value of the prototypical /i/ within the individually determined /i/ categories (however, for the difference between 50 ms and 100 ms, see section 3.2). In other words, the listeners' category boundaries between /y/ and /i/, and the /i/ prototypes (in terms of F2 frequency) were not demonstrably dependent on the stimulus duration. This result is noteworthy also from the perspective that different f0 contours were used for the longer durations of 250 ms and 500 ms for the purpose of achieving better stimulus naturalness (see section 2.1.2). In spite of this additional f0 cue (Järvikivi, Vainio, and Aalto, 2010; see section 1.3.2), no difference was observed in the categorization or goodness rating of the stimuli. In the experiments, the formants varied only in one dimension (F2), and therefore, the results cannot be generalized to apply to the entire formant space of /y/ and /i/ vowels in the Finnish vowel system; rather they represent one cross-section along the F2 axis while the F1 was held constant. Keeping this limitation in mind, the results do not challenge the general view that the single and double Finnish vowels are perceived essentially identically in terms of quality.

Another important finding in Experiment 1 was that the listener's gender had no effect on the location (F2 frequency) of the category border between /y/ and /i/, although statistical analysis revealed that the category boundary area (BW) was narrower in male listeners at 50 ms. In Experiment 2, neither the F2 frequency nor the goodness rating values of the prototypical /i/ differed between genders. The stimuli were synthesized using f0 values that are typical for male

speakers. Thus, if the listeners were using speaker class (gender) specific prototypes in their assessments, both the male and female listeners behaved similarly and apparently used their prototypes for a male speaker. This is in line what Rosner & Pickering (1994) propose in their initial auditory theory of vowel perception.

One goal of the present study was to find possible duration-dependent effects on the categorization process itself. In Finnish, the vowel quantity determines the meaning of a word in certain minimal word pairs, so one may hypothesize that the consistency of *quality* categorization and the measured reaction times would differ at durations that represent the typical *quantity* categories of Finnish vowels. We expected either a better labeling performance with less variability when the stimuli are close to the durations of the typical Finnish short and long vowels, or an overall poor performance with the shorter durations, which would emphasize the role of auditory cue processing instead of stimulus typicality. According to the main part of research published on the duration of Finnish vowels, the short vowels are within the range of 40 ms - 80 ms, long vowels within the range of 130 ms - 350 ms, and the category border area is within the range of 90 ms - 130 ms. The stimulus durations used in the present study covered the typical short and long Finnish vowels: 50 ms represented short vowels, 100 ms category border area, 250 ms long vowels, and 500 ms "prolonged" vowels in carefully uttered speech. Interestingly, the normalized reaction times to the stimuli with the duration of 100 ms showed a significant difference in comparison to the other durations. This could be interpreted so as to indicate that the 100 ms stimuli do not represent properly either the short or the long Finnish vowels, and consequently, the normalized reaction times at the boundary of *quality categories* are slightly longer. These results thus suggests that stimulus typicality (quantity) affects the categorization process but not its end result. The response rate might be feasible as a potential categorization performance indicator since the number of recorded responses increased significantly at longer stimulus durations, which may be explained by

the cue-duration hypothesis: there is more time and more cues available for extracting the relevant features from the longer stimuli (Pisoni, 1973; Repp & Liberman, 1987).

The results of this study indicate that two key characteristics of the initial auditory theory of vowel perception (Rosner & Pickering, 1994), namely, the local effective vowel indicator E2 (approximated by the auditory Hz to mel frequency conversion of F2) and the factor D (representing here directly the physical duration d), are not seemingly dependent on each other, thus suggesting that the AVS is orthogonal for these two variables in the Finnish vowel space of /y/ and /i/. A possible explanation for this comes from studies measuring more directly the neural processing of vowel quality and quantity. On the basis of fMRI studies, Jacquemot et al. (2003) suggest that the spectral cues of vowels are represented through the tonotopic organization of the auditory cortex, whereas the quantity is processed separately through temporal integration in the auditory pathway. Ylinen *et al*. give further support for this in their studies on Finnish vowel quantity (Ylinen, Huotilainen, & Näätänen, 2005; Ylinen, 2006). They used a component of the event-related brain potential, the mismatch negativity (MMN), to investigate the processing of phoneme quality and quantity in the human brain. Upon finding that the MMN responses to changes in phoneme quality and quantity are additive, they concluded that these features are processed independently of each other, thus representing separate neural processes that can be seen as different levels in the phonological system.

The duration-independent F2 values of the CB obtained in this study suggest that individual quality categories are determined by the psychoacoustic processing of *spectral cues,* and even the shortest (50 ms) stimulus duration of an isolated vowel is long enough for a listener to consistently judge between quality oppositions. The observation that perceptual /i/ prototypes did not depend on duration further supports the notion that the quality of the single and double vowels is perceived as

the same. This result may also be interpreted as giving indirect support to the perceptual magnet effect (Kuhl, 1991): regardless of the minor F2 differences reported between the produced Finnish short and long /i/ vowels, they are perceived equally due to the perceptual /i/ prototypes that generalize the minor differences in vowel quality. If perceptual prototypes form the basis for articulatory targets used in speech production, the results of this study support O'Dell's notion (see section 1.3.2.) that the reported centralization of short vowels is caused by a shorter acoustic duration, not by the phonological quantity of the vowel, an explanation that means that single and double vowels would have the same articulatory target, which is not met in articulating the single vowels.

The results of the present study seem to differ from the results obtained by Meister and Werner (2009) for the high-mid vowel pairs /i/-/e/, /y/-/ö/ and /u/-/o/ of Finnish and Estonian listeners (see section 1.3.2.). They found that openness correlates positively with the stimulus duration in an ABX setup, where A and B represent the prototypical vowels of the pair (e.g., /i/ and /e/) and X represents a vowel variant on the continuum between the pair. The conclusion was that the longer the duration of the ambiguous stimulus on the category boundary area, the more likely it is categorized as the more open vowel of the pair. The main differences between the study design of these two studies are that, first, in the present study we varied only the F2 of the stimuli (front-back), whereas Meister and Werner varied primarily the F1 formant (high-low), and second, Meister and Werner used the ABX setup, which differed from the categorization setup used in our study by offering two prototypical references at the opposite ends of the continuum for the comparison. They also used shorter vowel durations (covering only the 50 ms and 100 ms durations of our study), and the formant frequencies for the prototypical /i/ reference were 250 Hz (F1) and 2205 Hz (F2). With F1 fixed at 250 Hz, our rating experiment, however, resulted in F2 values of about 2500 Hz for a prototypical /i/ regardless of duration. These differences may offer an explanation for the seemingly

discrepant results between the two studies. Essentially, the ABX setup gives physical references to which the subject is asked to compare the ambiguous stimulus, whereas in our study design there is only a mental reference available. Given that the F2 value of the reference /i/ used by Meister and Werner is typical to a produced short /i/ (Table 2), prolongation of the ambiguous X stimulus may thus cause a growing mismatch to the typical produced long /i:/.

In the face of recent challenges that suggest that quality co-vary with quantity, the main results of this study support the identity group hypothesis: the location of the category boundary between /y/ and /i/ on the F2 formant frequency axis, the width of the category boundary on the F2 formant frequency axis, the goodness rating value of the prototypical /i/, and the location of the prototypical /i/ on the F2 formant frequency axis were all independent of the stimulus duration.

Acknowledgments

The study was partially supported by a grant from the Finnish Cultural Foundation. We wish to thank Professor Heikki Lyytinen, University of Jyväskylä, and Professor emeritus Åke Hellström, Stockholm University, for their valuable comments on the manuscript, and Lea Heinonen-Eerola, M.A. for revising the English language of the manuscript.

Textual footnotes

[1)] *tule* (‘come!’) - *tuule* (‘blow!’) - ei *tulle* (‘it may not come’) - ei *tuulle* (‘it may not blow’) - *tuullee* (‘it may blow’) - *tuulee* (‘it blows’) - *tulee* (‘it comes’) - *tullee*, (‘it may come’); phonetically with IPA symbols: [tule] - [tu:le] - [tul:e] - [tu:l:e] - [tu:l:e:] - [tu:le:] - [tule:] - [tul:e:].

[2)] The following terms and notations are used in relation to *quantity*: The term *duration* refers to the acoustic length (in seconds or milliseconds) of a phone or a word. The words *single* and *double* refer to phonological or linguistic *quantity categories*, denoted as /V/ and /VV/ for vowels and /C/ and /CC/ for consonants. The notation [phone] denotes the *short* duration and [phone:] the *long* duration of an uttered phone. The following notations and terms are used in relation to *quality*: [phone] (for example, [i]) denotes a phone as an acoustic variant (allophone) of a phoneme, and /phoneme/ (for example, /i/) denotes a phoneme as a representative of a linguistic quality category.

When orthography is emphasized the following notation is used: <V> for vowel V and <C> for consonant C (for example, the Finnish vowels are: <a>, <e>, <i>, <o>, <u>, <y>, < ä>, and <ö>).

[3)] Categorization process refers here to the psychological functions or steps needed for identifying the vowel and deciding on its *quality* category. The end result of the categorization process may be the same (identical CB and BW), but e.g. the process timing may depend on stimulus duration.

[4)] In Experiment 1, the subjects were instructed to listen to the stimuli and make their choice, but it was not especially emphasized that the stimuli should be listened to the end. Since the 500 ms stimulus duration represents a prolonged vowel, listeners may have responded occasionally while the stimulus was still on. However, considering the longer mean RT and the distribution of responses to the longest 500 ms stimulus set (mean RT= 0.73 s, *SD*= 0.11 s), it is evident that major part of the responses (>95.45%) took place after the stimulus offset (mean - 2 x *SD* = 0.51 s).

Abbreviations

AVS: auditory vowel space; BW: (category) boundary width; CB: category boundary; CBR: critical band rate; CV: coefficient of variation; d: physical duration; D: auditory temporary information; ISI: inter-stimulus interval; LEVI: local effective vowel indicator (E1, E2, E3); N: sample size; NP: non-prototype; P: prototype; PME: perceptual magnet effect; RT: reaction time; *SD*: standard deviation.

References

Aaltonen, O., Eerola, O., Hellström, Å., Uusipaikka, E., & Lang, H., A. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *Journal of the Acoustical Society of America, 101*(2), 1090-1103.

Aaltonen, O., & Suonpää, J. (1983). Computerized two-dimensional model for Finnish vowel identifications. *Audiology, 22*, 410-415.

Bamber, D. (1969). Reaction times and error rates for 'same' - 'different' judgements of multidimensional stimuli. *Perception & Psychophsyics 6*(3), 169-174.

Becker-Kristal, R. (2010). Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus. (Ph.D. thesis), Department of Linguistics, UCLA, Los Angeles, U.S.A. (http://www.linguistics.ucla.edu/faciliti/research/research.html#Dissertations)

Bliss, C. I. (1934). The method of probits. *Science, 79*, 38-39.

Diehl, R., Lindblom, B., Hoemke, K., & Fahey, R. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics, 24*, 187-208.

Eerola, O., Laaksonen, J., Savela, J., & Aaltonen, O. (2002). Suomen [y] / [i] ja [y:] / [i:] -vokaalien tuotto havaintokokeiden tulosten valossa. *Fonetiikan Päivät 2002 - Phonetics Symposium 2002,* Espoo, Finland. *, 67,* 109-113.

Eerola, O., Laaksonen, J., Savela, J., & Aaltonen, O. (2003). Perception and production of the short and long Finnish [i] vowels: Individuals seem to have different perceptual and articulatory templates. *Proceedings of the 15th International Congress of Phonetics Sciences,* University of Barcelona, Barcelona, Spain.

Eerola, O., Savela, J. (2011). Differences in Finnish front vowel production and weighted perceptual prototypes in the F1-F2 space. *Proceedings of the 17th International Congress of Phonetics Sciences,* University of Hong Kong, Hong Kong, China.

Finney, D.J. (1944). The application of Probit analysis to the results of mental tests. *Psychometrica, 9*(1).

Goldstein, U. (1980). An articulatory model for the vocal tracts of growing children. Doctoral dissertation, M.I.T. (http://mit.dspace.org/handle/1721.1/22386).

Guenther, F. H. (2000). An analytical error invalidates the "depolarization" of the perceptual magnet effect. *Journal of the Acoustical Society of America, 107*, 3576-3577.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100*(2), 1111-1121.

Harrikari, H. (2000). Segmental length in Finnish - studies within constraint-based approach. (Ph.D. thesis), *Publications of the Department of General Linguistics, University of Helsinki, 33*, 1-151.

Iivonen, A., & Harnud, H. (2005). Acoustical comparison of the monophtong systems in Finnish, Mongolian, and Udmurt. *Journal of the International Phonetic Association, 35*(1), 59-71.

Iivonen, A., & Laukkanen, A. (1993). Explanations for the qualitative variation of Finnish vowels. *Studies in Logopedics and Phonetics, 4*, 29-55.

Iivonen, A., & Tella, S. (2009). Vieraan kielen ääntämisen ja kuulemisen opetus ja harjoittelu. In O. Aaltonen, R. Aulanko, A. Iivonen, A. Klippi, & M. Vainio (Eds.), *Puhuva ihminen - puhetieteiden perusteet* (1st ed., pp. 269-281). Helsinki: Kustannusosakeyhtiö Otava.

Iverson, P., & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from common mechanism? *Perception & Psychophysics, 62*(4), 874-886.

Järvikivi, J., Aalto, D., Aulanko, R., & Vainio, M. (2007). Perception of vowel length: Tonality cues categorization even in a quantity language. In J. Trouvain, & W.J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetics* Sciences, Universität des Saarlandes, Saarbrücken, Germany (pp. 693-696).

Järvikivi, J., Vainio, M., Aalto, D. (2010). Real-time correlates of phonological quantity reveal unity of tonal and non-tonal languages. *PLoS ONE, 5*(9), p. e12603. 10 p.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *The Journal of Neuroscience, 23*(29), 9541-9546.

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

Karlsson, F. (1983). *Suomen kielen äänne- ja muotorakenne* [Sound and Form Structures in Finnish]. Porvoo: Werner Södesrstöm Oy.

Klatt, D. H. (1980). Software for Cascade/Parallel formant synthesizer. *Journal of the Acoustical Society of America, 53*, 8-16.

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

Kukkonen, P. (1990). Patterns of phonological disturbances in adult aphasia. Faculty of Arts, University of Helsinki. *Suomalaisen Kirjallisuuden Seuran Toimituksia,* (529), 1-231.

Lehtonen, J. (1970). Aspects of quantity in standard Finnish. University of Jyväskylä. *Studia Philologica Jyväskyläensia, IV* .

Leibold, L.J., Werner, L.A. (2002). Relationship between intensity and reaction time in normal-hearing infants and adults. *Ear and Hearing, 23*(2), 92-97

Lively, S. E. (1993). An examination of the perceptual magnet effect. *Journal of the Acoustical Society of America, 93*(4), 2423.

Lively, S. E., & Pisoni, D. B. (1997). On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology, 23*(6), 1665-1679.

Lotto, A. J. (2000). Reply to "an analytical error invalidates the 'depolarization' of the perceptual magnet effect" [J.acoust.soc.am. 107, 3576-3577 (2000)]. *Journal of the Acoustical Society of America, 107*(6), 3578-3580.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1998). Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America, 103*(6), 3648-3655.

Meister, E., & Werner, S. (2009). Duration affects vowel perception in Estonian and Finnish. *Linguistica Uralica, 3*, 161-177.

Miller, J. L. (1997). Internal structure of phonetic categories. *Language and Cognitive Processes, 12*(5/6), 865-869.

Miller, J. L., Connine, C. M., Schermer, T. M., & Kluender, K. R. (1983). A possible auditory basis for internal structure of phonetic categories. *Journal of the Acoustical Society of America, 73*(6), 2124-2133.

Nábelek, A. K., Czyzewski, Z., & Crowley, H. J. (1993). Vowel boundaries for steady-state and linear formant trajectories. *Journal of the Acoustical Society of America, 94*(2), 675-687.

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

Nordström, P-E. (1977). Female and infant vocal tracts simulated from male area functions. *Journal of Phonetics, 5*, 81-92.

O'Dell, M. (2003). Intrinsic timing and quantity in Finnish. (Doctoral Dissertation. *Acta Universitatis Tamperensis, 979*, 1-128.

Peltola, M. S. (2003). *The attentive and preattentive perception of native and non-native vowels.* Unpublished Doctoral Thesis, University of Turku, Department of Phonetics, Turku.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13*(2), 253-260.

Reed, C. (1975). Reaction times for a same-different discrimination of vowel-consonant syllables. *Perception & Psychophysics, 18*(2), 65-70.

Repp, B. H., & Crowder, R. G. (1990). Stimulus order effects in vowel discrimination. *Journal of the Acoustical Society of America, 88*(5), 2080-2090.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. Harnad (Ed.), *Categorical perception, the groundwork of cognition* (1 st ed., pp. 89-112). New York: Press Syndicate of the University of Cambridge.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*, 532-547.

Rosner, B. S., & Pickering, J. B. (1994). *Vowel perception and production*. New York: Oxford University Press.

Savela, J. (2009). Role of selected spectral attributes in the perception of synthetic vowels. (PhD thesis, Turku Centre for Computer Science, University of Turku). *TUCS Dissertations, 119*, 1-82.

Stevens, S. S., & Volkmann, J., Newman, E.B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8*, 185-190.

Strange, W. (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America, 85*(5), 2081-2087.

Suomi, K. (2005). Temporal conspiracies for a tonal end: Segmental durations and accentual f0 movement in a quantity language. *Journal of Phonetics, 33*, 291-309.

Suomi, K. (2006). *Stress, accent and vowel durations in Finnish* No. Working Papers 52). Lund: Department of Linguistics & Phonetics, Lund University.

Suomi, K. (2007). On the tonal and temporal domains of accent in Finnish. *Journal of Phonetics, 35*, 40-55.

Suomi, K., Toivanen, J., & Ylitalo, R. (2003). Durational and tonal correlates of accent in Finnish. *Journal of Phonetics, 31*, 113-138.

Suomi, K., Toivanen, J., & Ylitalo, R. (2006). *Fonetiikan ja suomen äänneopin perusteet*. Helsinki: Gaudeamus Kirja.

Suomi, K., & Ylitalo, R. (2004). On durational correlates of word stress in Finnish. *Journal of Phonetics, 32*, 35-63.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America, 88*, 97-100.

Wiik, K. (1965). Finnish and English vowels. (Doctoral Thesis, University of Turku). *Annales Universitatis Turkuensis, Series B* (94)

Ylinen, S. (2006). Cortical representation for phonological quantity. (Doctotal Thesis, Cognitive Brain Research Unit, Department of Psychology, University of Helsinki).

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., Huotilainen, M., Alku, P., & Näätänen, R. (2006). Mismatch negativity (MMN) elicited by changes in phoneme length: A cross-linguistic study. *Brain Research, 1072*, 175-185.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America, 68*(5), 1523-1525.

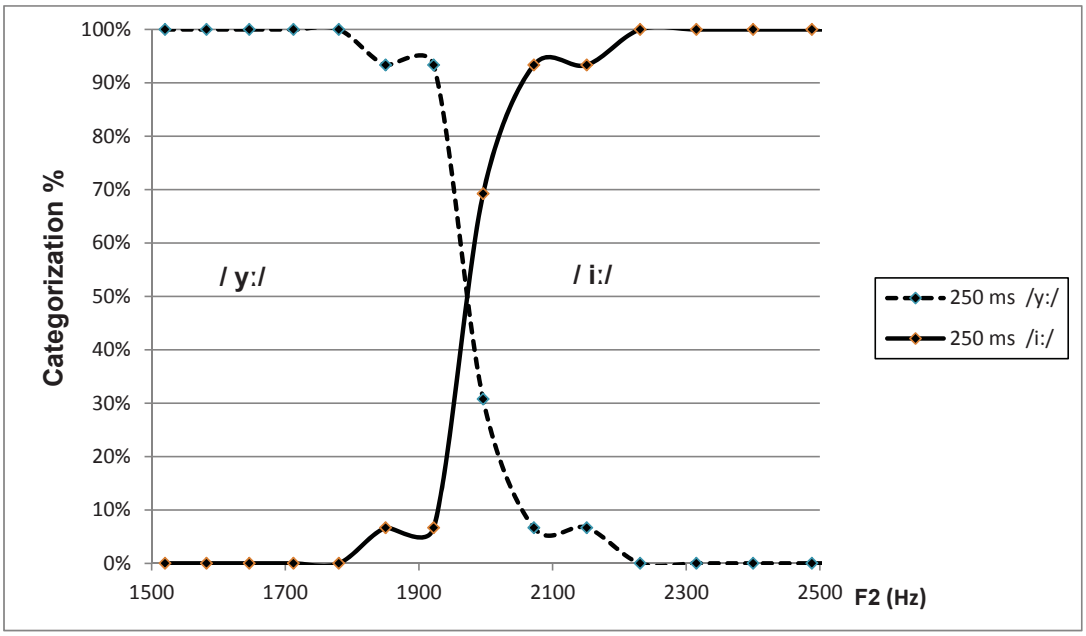


Fig. 1. Example of a consistent /y:/-/i:/ categorization (Listener 2) as a function of formant F2 frequency at a stimulus duration of 250 ms. Stimulus step size is 30 mel.

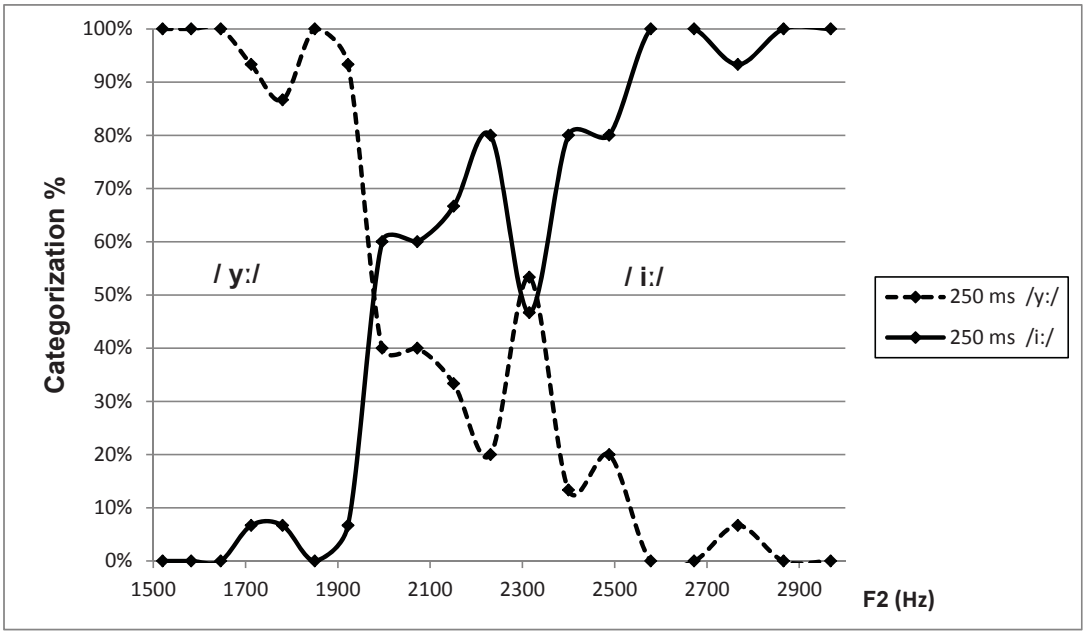
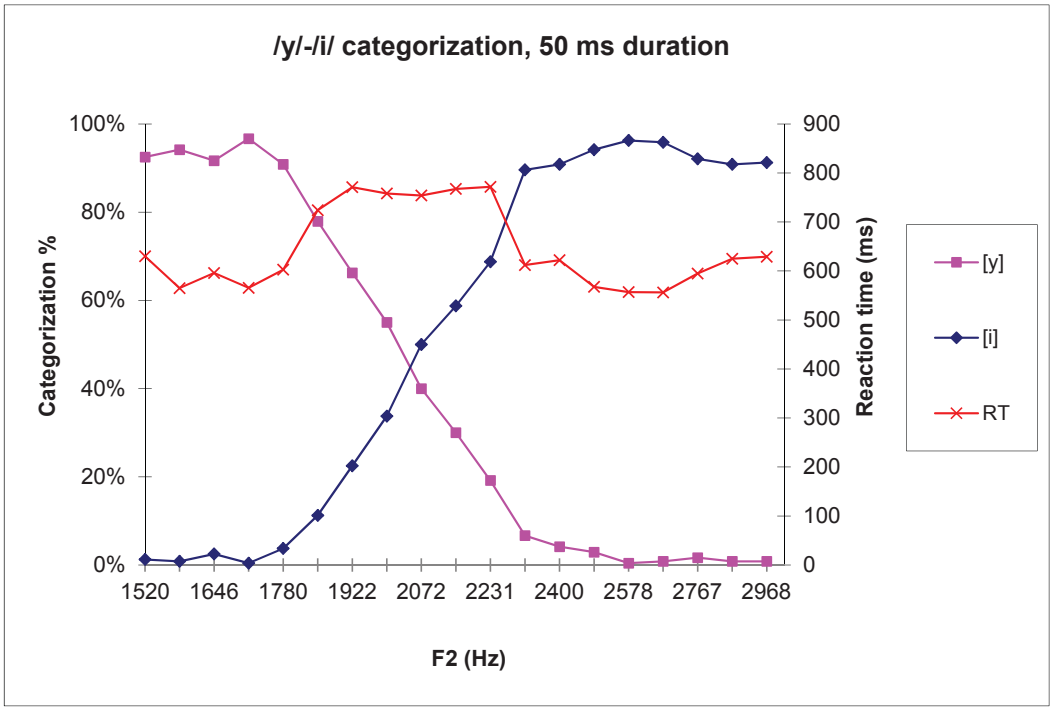
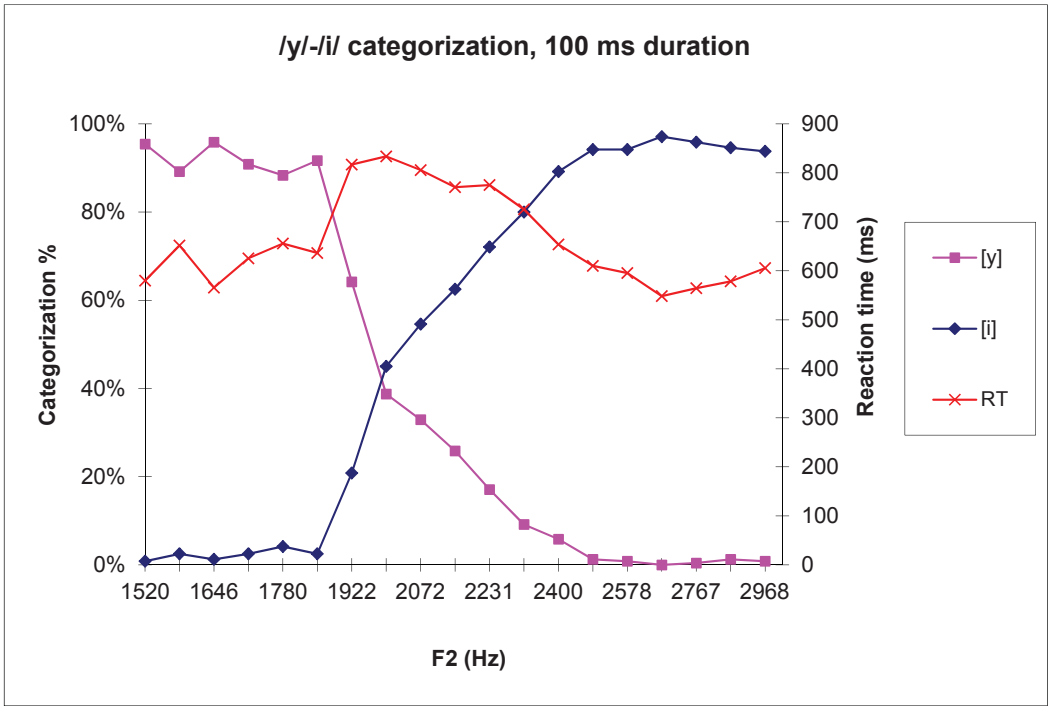
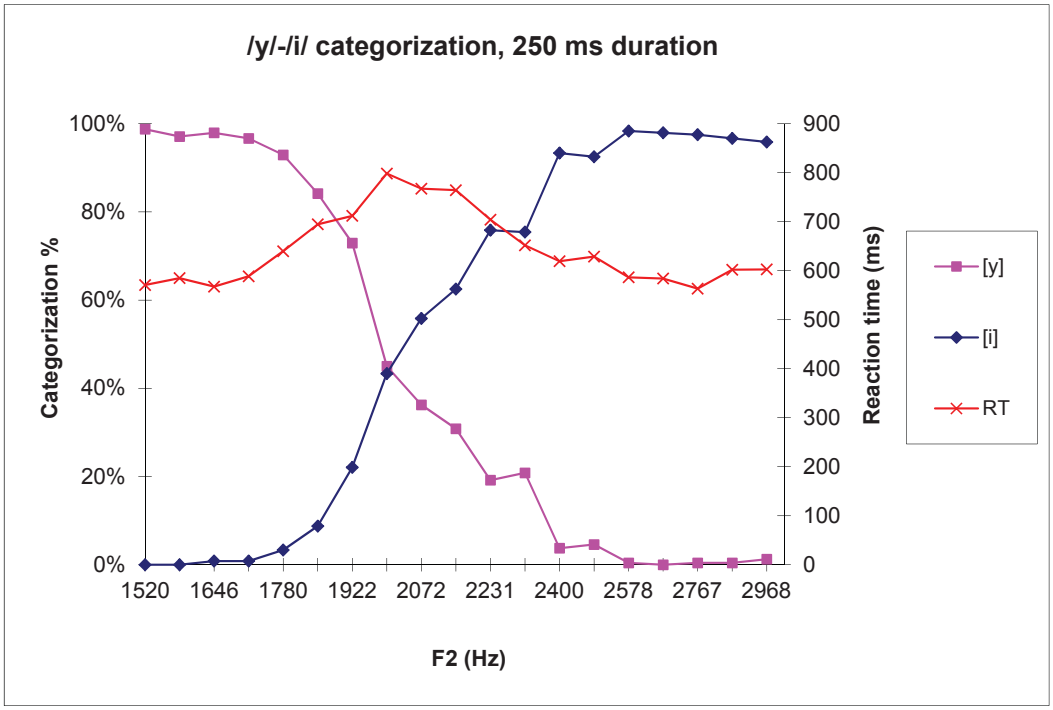


Fig. 2. Example of an inconsistent /y:/-/i:/ categorization (Listener 17) as a function of formant F2 frequency at a stimulus duration of 250 ms. Stimulus step size is 30 mel.


/y/-/i/ categorization, 50 ms duration
Categorization %
100%
80%
60%
40%
20%
0%
Reaction time (ms)
900
800
700
600
500
400
300
200
100
0
1520 1646 1780 1922 2072 2231 2400 2578 2767 2968
F2 (Hz)
[y]
[i]
RT


/y/-/i/ categorization, 100 ms duration
Categorization %
100%
80%
60%
40%
20%
0%
Reaction time (ms)
900
800
700
600
500
400
300
200
100
0
1520 1646 1780 1922 2072 2231 2400 2578 2767 2968
F2 (Hz)
[y]
[i]
RT



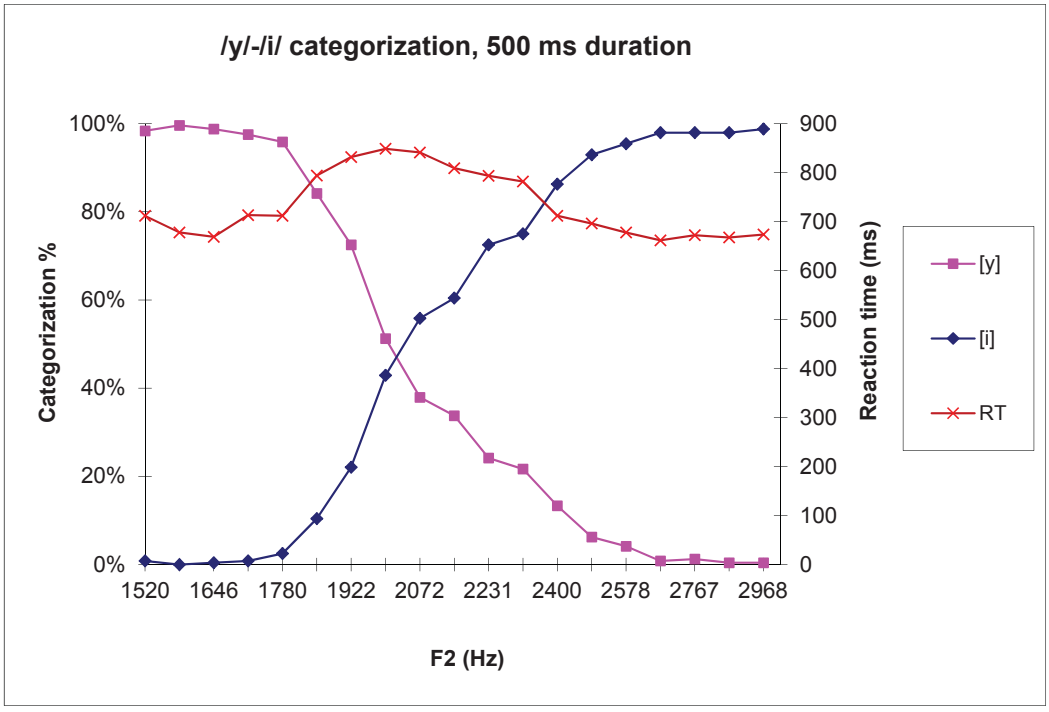


Fig. 3. a-d. The effect of duration on vowel categorization. Categorization of 19 synthesized vowel stimuli to [y] and [i] phones (Categorization %), and categorization reaction times (RT, in ms) as a

function of the second formant (F2, in Hz) at stimulus durations of 50 ms (Fig. 3a), 100 ms (Fig. 3b), 250 ms (Fig. 3c), and 500 ms (Fig. 3d). The F2 continuum spans from 1520 Hz (1290 mel) to 2968 Hz (1830 mel) in steps of 30 mel (meaning that, e.g., four stimulus increments correspond to 260 Hz at 1520 Hz but to 367 Hz at 2400 Hz).

= = = = = = = = = = = = = =

Note for publisher: Fig. 3 in colors online (web), BW when printed.

The suggested layout for the four panels is 2x2, with the 50 ms and 100 ms panels on top, and the 250 ms and 500 ms panels in bottom.

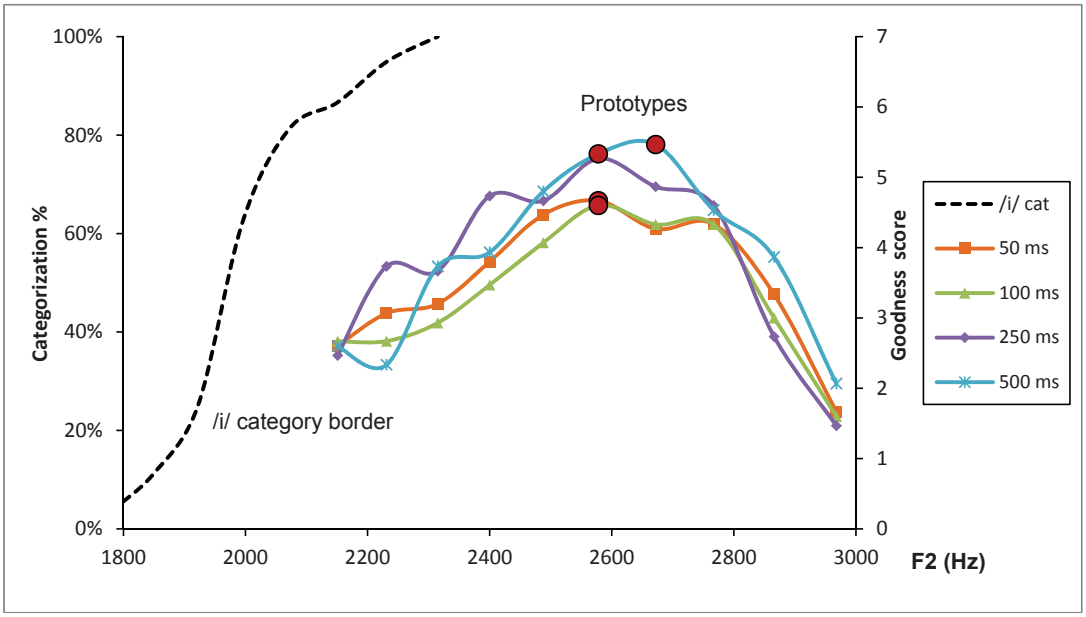


Fig. 4. Example of "hill" type goodness ratings (scale 1-7) of stimuli within the individual /i/ category of Listener 2 at stimulus durations of 50 ms, 100 ms, 250 ms, and 500 ms. The /i/ category border is shown as a dotted line. The highest scoring stimuli (perceptual prototypes, marked as circles) are at 2578 Hz (50 ms), 2578 Hz (100 ms), 2578 Hz (250 ms), and 2672 Hz (500 ms). Stimulus step size is 30 mel.

= = = = = = = = = = = = = =

Note for publisher: Fig. 4 in colors online (web), BW when printed.

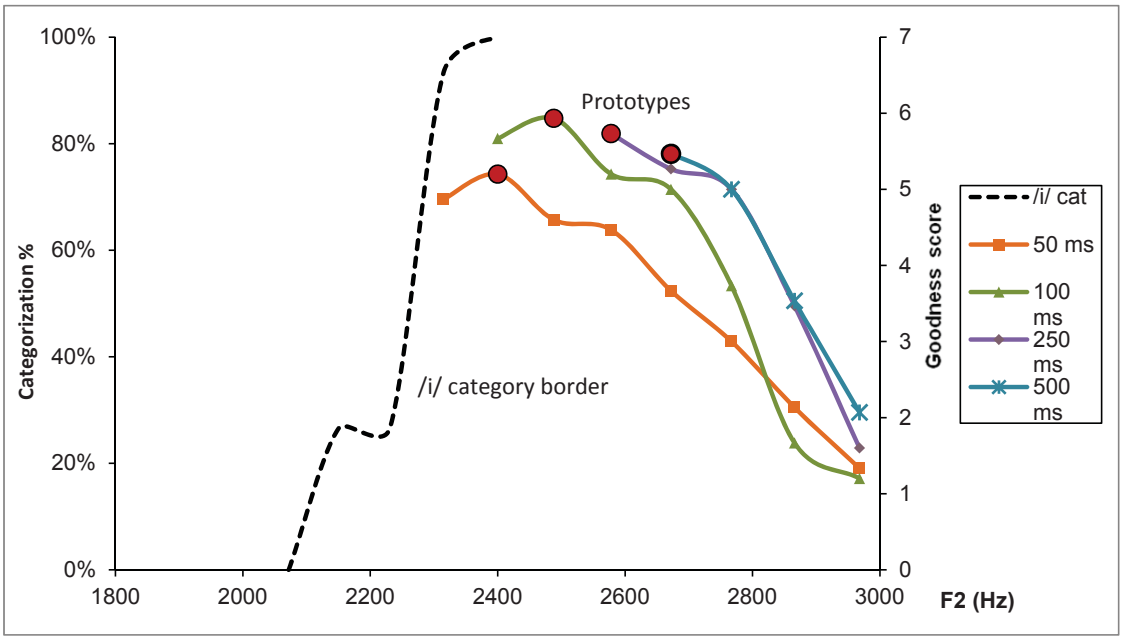


Fig. 5. Example of "down" type goodness ratings (scale 1-7) of stimuli within the individual /i/ category of Listener 14 at the stimulus durations of 50 ms, 100 ms, 250 ms, and 500 ms. The /i/ category border is shown as a dotted line. The highest scoring stimuli (perceptual prototypes, marked as circles) are at 2400 Hz (50 ms), 2488 Hz (100 ms), 2578 Hz (250 ms), and 2672 Hz (500 ms). Both the prototypes and category borders of Listener 14 sift towards higher frequencies with longer durations. Stimulus step size is 30 mel.

= = = = = = = = = = = = = =

Note for publisher: Fig. 5 in colors online (web), BW when printed.

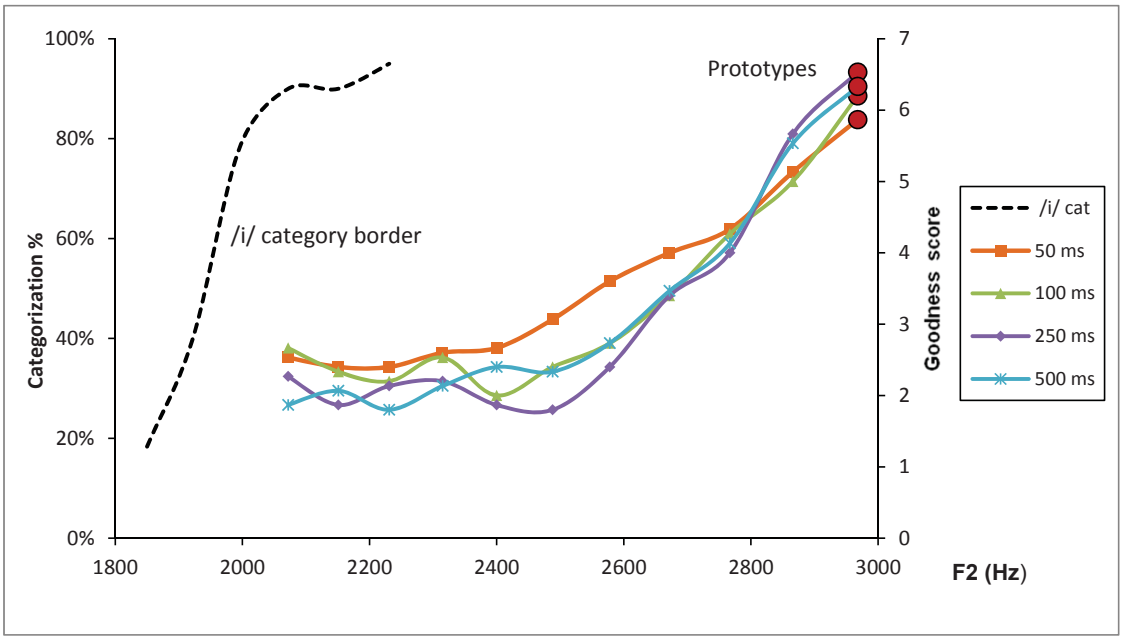


Fig. 6. Example of "up" type goodness ratings (scale 1-7) of stimuli within the individual /i/ category of Listener 13 at the stimulus durations 50 ms, 100 ms, 250 ms, and 500 ms. The /i/ category border is shown as a dotted line. The highest scoring stimuli (perceptual prototypes, marked as circles) are at 2968 Hz at all durations. Stimulus step size is 30 mel.

= = = = = = = = = = = = = = =

Note for publisher: Fig. 6 in colors online (web), BW when printed.

Table 1. Formant F1 and F2 values in Hz of long Finnish /i/ and /y/ vowel categories obtained in different identification studies using synthesized long vowels.

	F1 /i:/ (Hz)	F2 /i:/ (Hz)	F1 /y:/ (Hz)	F2 /y:/ (Hz)	duration (ms)	n	Source
1	250-310	> 2100	250-325	1500-1900	300	32	Aaltonen & Suonpää, 1983
2	250-330	<2880	250-330	<1644	350	9	Peltola, 2003
3	248-326	2200-2800	248-354	1460-1900	350	68	Savela, 2009

Table 2. Observed values and differences in Hz for formants F1 and F2 in produced Finnish short and long /i/ and /y/ vowels obtained in different studies.

	F1 /i/	F1 /i:/	ΔF1	F2 /i/	F2 /i:/	ΔF2	F1 /y/	F1 /y:/	ΔF1	F2 /y/	F2 /y:/	ΔF2	n	Source
	Hz	Hz	Hz	Hz	Hz	Hz	Hz	Hz	Hz	Hz	Hz	Hz		
1	340	275	65	2355	2495	-140	340	300	40	1920	1995	-75	5	Wiik, 1965
2	333	317	16	2326	2389	-63	340	320	20	1774	1849	-75	4	Kukkonen, 1990
3	300	295	5	2262	2380	-118	335	292	43	1751	1805	-54	1	Iivonen et al., 1993
4	355	319	36	2064	2155	-91	365	326	39	1620	1633	-13	4	Kuronen, 2000
5	n.a.	n.a.	-	2391	2500	-109	n.a.	n.a.	-	1860	1841	19	26	Eerola et al., 2002
6	300	240	60	1900	2100	-200	300	260	40	1600	1680	-80	1	Iivonen et al., 2005
7	346	328	18	2422	2525	-104	331	323	8	1861	1854	7	14	Eerola & Savela, 2011
	329	296	33	2246	2363	-118	335	304	32	1769	1808	-39		Mean value

Table 3. Categorization as a function of stimulus duration (Experiment 1).

1. Formant F2 frequencies (Hz) of the category boundary (CB) between /y/ and /i/ as determined by Probit non-linear estimation (n=16). 2. Boundary width (BW) values: F2 frequency differences (Hz) at the 25%/75% identification points. 3. Categorization consistency: the response rates of the 16 listeners participating in the categorization experiment. *SD*=standard deviation, CV=coefficient of variation.

	50 ms *n*=16	100 ms *n*=16	250 ms *n*=16	500 ms *n*=16	Unit
1. Category boundary					
Mean of F2	2065	2049	2077	2094	Hz
SD of F2	144	158	171	196	Hz
Max of F2	2305	2304	2423	2546	Hz
Min of F2	1852	1769	1909	1823	Hz
Median of F2	2054	2032	1990	2061	Hz
2. Boundary width					
Mean of BW	235	191	186	172	Hz
SD of BW	134	102	80	122	Hz
CV of BW	57,0	53,6	42,9	71,0	%
BW/CBW	0.77	0.71	0.67	0.68	
3. Response rate	93.0	92.5	96.0	97.5	%

Table 4. Reaction times as a function of stimulus duration (Experiment 1). Mean reaction times (t) and standard deviations (*SD*) of 16 listeners categorizing 19 stimuli, each repeated 15 times, on the Finnish /y/-/i/ continuum (with stimulus F2 ranging from 1520 Hz to 2968 Hz in steps of 30 mel) at four different vowel durations 50 ms, 100 ms, 250 ms, and 500 ms. $t_{/y/}$ = mean reaction time within the /y/ category, $t_{/i/}$ = mean reaction time within the /i/ category, t_{CB}= mean reaction time at the category boundary area, $t_{/i/min}$ = the shortest mean reaction time recorded for a stimulus within the /i/ category (stimulus F2 given in the Table), t_{tot} =mean reaction time to all stimuli, t_{cat}= ($t_{/y/}$ + $t_{/i/}$) / 2.

Reaction times 50 ms duration	Mean (s)	*SD* (s)	F2 (Hz)	Reaction times 100 ms duration	Mean (s)	*SD* (s)	F2 (Hz)
t/y/	0.59	0.24		t/y/	0.61	0.23	
t_{CB}	0.84	0.22	1852-2305	t_{CB}	0.96	0.27	1909-2412
t/i/	0.55	0.14		t/i/	0.58	0.18	
t/i/min	0.41	0.07	2672	t/i/min	0.40	0.07	2672
t_{tot}, overall mean	0.65	0.19	1520-2968	t_{tot}, overall mean	0.66	0.18	1520-2968
$t_a = t_{CB} / t_{tot}$	1.31			$t_a = t_{CB} / t_{tot}$	1.44		
$t_b = t_{CB} / t_{cat}$	1.51			$t_b = t_{CB} / t_{cat}$	1.67		

Reaction times 250 ms duration	Mean (s)	*SD* (s)	F2 (Hz)	Reaction times 500 ms duration	Mean (s)	*SD* (s)	F2 (Hz)
t/y/	0.58	0.14		t/y/	0.68	0.13	
t_{CB}	0.85	0.21	1909-2423	t_{CB}	0.96	0.20	1823-2546
t/i/	0.58	0.11		t/i/	0.66	0.15	
t/i/min	0.38	0.08	2767	t/i/min	0.45	0.08	2672
t_{tot}, overall mean	0.64	0.13	1520-2968	t_{tot}, overall mean	0.73	0.12	1520-2968
$t_a = t_{CB} / t_{tot}$	1.32			$t_a = t_{CB} / t_{tot}$	1.28		
$t_b = t_{CB} / t_{cat}$	1.48			$t_b = t_{CB} / t_{cat}$	1.42		

Table 5. Goodness rating of vowels categorized as /i/ at varying stimulus durations (Experiment 2). The mean rating scores and standard deviations (*SD*) of prototypes (P), non-prototypes (NP), and of all stimuli on the scale 1-7 (1 = a poor category exemplar, 7 = a good category exemplar), the formant F2 frequencies (Hz) of the prototype vowels, and the number (#) of response types ("hill", "down" , "up") for 15 listeners at the stimulus durations of 50 ms, 100 ms, 250 ms, and 500 ms.

	50 ms	**100 ms**	**250 ms**	**500 ms**
P, mean score	5.53	5.88	5.71	5.60
P, *SD* of scores	0.90	0.83	0.75	0.71
NP, mean score	1.72	1.89	1.59	1.99
NP, *SD* of scores	0.80	0.90	0.48	0.94
All, mean score	4.04	4.27	4.06	4.05
All, *SD* of scores	0.82	0.83	0.73	0.65
P F2 (Hz), mean	2493	2533	2511	2561
P F2 (Hz), *SD*	184	258	191	219
# "hill" type	10	8	11	10
# "down" type	4	4	3	3
# "up" type	1	3	1	2

Publication III

Eerola, O., Savela, J. 2012.

Production of short and long Finnish vowels with and without noise masking.

Linguistica Uralica, Vol. XLVIII (3). Pages 200-208.

 doi:10.3176/lu.2012.3.06

OSMO EEROLA, JANNE SAVELA (Turku)

PRODUCTION OF SHORT AND LONG FINNISH VOWELS WITH AND WITHOUT NOISE MASKING

Abstract. In order to further examine the possible quality differences between produced short and long Finnish vowels, we studied the formant frequencies F1–F4 and duration of the eight Finnish vowels /a/, /e/, /i/, /o/, /u/, /y/, /æ/ and /ø/[1] when uttered in carrier words (e.g., /tili/ — /tiili/) in two different masking conditions and without a noise mask. Babble noise at 92dB SPL was used to simulate a loud, crowded cocktail party, and pink noise at 83dB SPL an environment with the maximum noise level allowed for continuous working. Minor quality differences were found between the short and long vowels. Noise masking caused a significant prolongation of produced short vowels, and a significant increase in the F1 frequency.

Keywords: Finnish, vowel production, vowel quality and quantity, noise masking.

1. Introduction

The Finnish vowel system includes eight vowels: /a/, /e/, /i/, /o/, /u/, /y/, /æ/ and /ø/, which all can occur as short (single) or long (double) in any position of a word (Suomi, Toivanen, Ylitalo 2006). The modern orthography of Finnish reflects the interpretation that the long vowel segments of spoken Finnish consist of two similar shorter segments (Karlsson 1983).

Karlsson (1983) presents three possible phonological interpretations for the Finnish quantity opposition. According to the monophonematic interpretation, the short and long vowels and consonants represent different phonemes: e.g. /tule/ — /tUle/, or /tule/ — /tuLe/ (here, a capital letter stands for a long phoneme). This interpretation has not been widely accepted since it would almost double the number of Finnish phonemes from the 8 vowels and 22 core consonants, which is undesirable for the economy of linguistic description. Karlsson further argues that this interpretation is against the Finnish orthography and also against the intuition of Finnish speakers. According to the second interpretation, the long phonemes are short phonemes followed by a chrome /:/ (originally proposed by Jones (1944)), which extends the duration of a short phoneme. This interpretation can partially be justified on the basis of the fact that the phonetic quality differences between short and long vowels in Finnish are small as compared, for example, to English or Swedish. However, it would complicate the analysis of certain morphological categories in Finnish. According to the third interpretation, the long segments of vowels or consonants consist of two successive and identical short segments. Karlsson refers to this

[1] The symbols used in this paper are those of the International Phonetic Alphabet (IPA). Equivalents in the Finno-Ugric transcription system are as follows: ɑ = a, æ = ä, ø = ö, y = ü.

interpretation as the identity group interpretation, and it is generally accepted in Finnish phonetic textbooks (Suomi, Toivanen, Ylitalo 2006; Iivonen, Tella 2009) as the de facto explanation of the phonological quantity opposition in Finnish. We refer to this interpretation in the following also as the general view. Harrikari (2000) has presented a complementary and partially opposing view on identity group interpretation, using the optimality theory of generative phonology as the framework and considering dialectic epenthesis, gemination, and language games as examples. However, Harrikari approaches the segmental length in Finnish from the viewpoint of theoretical phonology and morphology, not from that of experimental phonetics.

Generally, the two durational variants of the eight Finnish vowels are regarded as being similar in p e r c e i v e d q u a l i t y. Eerola, Savela, Laaksonen and Aaltonen (2012) investigated the perception of short and long Finnish /y/ and /i/ vowels, and found that the location of the category boundary between /y/ and /i/ on the F2 formant frequency axis, the width of the category boundary on the F2 formant frequency axis, the goodness rating value of the prototypical /i/, and the location of the prototypical /i/ on the F2 formant frequency axis were all independent of the stimulus duration. The main results of the study by Eerola, Savela, Laaksonen and Aaltonen (2012) thus did not challenge the general view that the perceived Finnish short and long vowels are of equal quality.

However, the results of some earlier studies on the p r o d u c t i o n of Finnish vowels suggest that there exist minor spectral dissimilarities in the formant frequencies F1—F3 of the produced short and long vowels. For example, based on five informants, Wiik (1965) reported clear differences in the variability ranges of Finnish single and double /y/ and /i/ vowels, as measured in terms of F1, F2 and F3, stating that F1 is 40 Hz higher and F2 is 75 Hz lower in [y] than in [y:], and, correspondingly, F1 is 65 Hz higher, F2 is 140 Hz lower, and F3 is 265 Hz lower in [i] than in [i:]. The results indicate that the produced single vowels are more centralized than the double vowels are. In a later study on vowel production by Kukkonen (1990), differences of similar type but smaller magnitude were reported in a normal Finnish-speaking control group (N = 4): F1 was 16 Hz higher, and F2 and F3 were 63 Hz and 32 Hz lower in single than in double /i/ vowel. Correspondingly for single and double /y/ vowels, the differences were as follows: F1 was 19 Hz higher, F2 was 75Hz lower, and F3 was 20 Hz lower in the single vowel. However, only differences in F1 were statistically significant. In our earlier studies (Eerola, Laaksonen, Savela, Aaltonen 2003), a non-significant difference of 108 Hz was found for F2 between the short /i/ (F2 = 2391 Hz, SD = 194 Hz) and long /i:/ (F2 = 2500 Hz, SD = 212 Hz) produced by 26 informants in the first syllables of the words *tikki* and *tiili*. In a more recent study by Eerola and Savela (2011), a significant difference (paired t-test, $p < 0.01$, N = 14) of 104 Hz was found for F2 between the short /i/ and long /i:/ in an uttered word pair *tili*/*tiili*.

Iivonen and Laukkanen (1993) studied the qualitative variation of the eight Finnish vowels in 352 bisyllabic and trisyllabic words uttered by a single male speaker. They found a clear tendency for the short vowels to be more centralized in the psychoacoustic F1—F2 space, as compared to the long ones. However, except for the /u/—/u:/ pair, this difference was smaller than one critical band, and thus auditorily negligible. In a comparative study of the monophthong systems in the Finnish, Mongolian and Udmurt languages, Iivonen and Harnud (2005) report on minor spectral differences in the short/long vowel contrasts in stressed (e.g., [sika] / [si:ka]) and non-stressed (e.g., [etsi] / [etsi:]) syllables in Finnish words uttered by a single male speaker. The biggest differences between short and long vowels were found in /u/. As in the study by Iivonen and Laukkanen (1993), [u] is more centralized and does not overlap with [u:]. Also for /y/ and /i/, the short vowels are more centralized than their longer counterparts, but the short and long vowel versions overlap on the F1 axis. Interestingly, the /y/ and /i/ vowels, both short and long, also overlap on the F2 axis instead of being clearly separate phoneme

categories. To summarize, minor spectral differences have been reported in the F1 and F2 formant frequencies of the produced short and long Finnish vowels, and the biggest difference occurs between the high back vowels [u] and [u:].

In this study, we further examine the reported quality differences between produced short and long variants across the entire Finnish vowel system in two different noise masking conditions and without any noise mask. It was assumed that noise masking may cause hyperarticulation, and possibly accentuate the reported minor quality differences between short and long Finnish vowels. Since speakers are known to alter their vocal production in noisy environments (Lane, Tranel 1971, the Lombard effect), such as a loud restaurant or a noisy factory, we included two different types of masking noise to simulate these conditions. Multi-talker babble noise at 92 dB SPL (sound pressure level) was used to simulate a loud, crowded cocktail party, and pink noise at 83 dB SPL an environment with the maximum noise level allowed for continuous working. The Lombard effect has been reported to cause measureable differences in vowel intensity and duration, and also in formant frequencies: ambient noise elevates the speech amplitude by 5—10 dB, increases word durations by 10—20%, and increases significantly the F1 and F2 frequencies, thus causing a shift in the vowel space (van Summers, Pisoni, Bernacki, Pedlow, Stokes 1988; Castellanos, Benedi, Casacuberta 1996; Beckford Wassink, Wright, Franklin 2007).

2. Materials and methods

2.1. Subjects

Ten normally hearing young adults speaking the modern educated Finnish of South-West Finland volunteered as subjects. All subjects were screened for hearing impairments by means of an audiometer (Amplivox 116). For different vowels, the number of recorded subjects varied: 10 subjects for /i/, /e/, /y/, and /ø/. 9 subjects for /u/, and 4 subjects for /a/, /æ/, and /o/.

2.2. Procedure and analysis

The articulation of the eight Finnish vowels /a/, /e/, /i/, /o/, /u/, /y/, /æ/ and /ø/ when uttered in different carrier words and non-words (e.g., /tili/ — /tiili/, see Table 1) was recorded in two different masking conditions and without a noise mask. The subjects were asked to utter each word five times successively using their normal speech style, first without the noise mask, and then in the masking conditions. The recordings were carried out in an acoustically dampened room (27 dB_A SPL) by using a high quality microphone (AKG D660S) that was connected via an amplifier to a PC. The recordings were made at a sampling rate of 44.1 kHz, and saved as sound files for later analysis. Praat software was used for both the recordings and analysis.

Table 1

Carrier utterances used in the experiments

Short		**Long**	
IPA, Finnish	**Meaning**	**IPA, Finnish**	**Meaning**
[tɑli], *tali*	'tran'	[tɑ:li], *taali*	non-word
[teli], *teli*	'twin axle'	[te:li], *teeli*	non-word
[tili], *tili*	'account'	[ti:li], *tiili*	'brick'
[toli], *toli*	non-word/NA	[to:li], *tooli*	non-word
[tuli], *tuli*	'fire'	[tu:li], *tuuli*	'wind'
[tyli], *tyli*	non-word	[ty:li], *tyyli*	'style'
[tæli], *täli*	non-word	[tæ:li], *tääli*	non-word
[tøli], *töli*	non-word	[tø:li], *töölі*	non-word

The sound samples were automatically analyzed using a text grid in which the steady-state part of each target vowel was windowed varying between utterances. The f0, formants F1—F4, and vowel durations were analyzed by using the Burg method in which short-term LPC coefficients are averaged for the length of an entire sound. The Praat formant analysis settings were 0.025 s for Window length, and 5000 Hz (male) and 5500 Hz (female) for Maximum formant. The analysis results of the five repetitions were averaged for individual results.

2.3. Noise masks

Multi-talker babble noise at 92 dB SPL was used to simulate a loud, crowded cocktail party, and pink noise at 83 dB SPL an environment with the maximum noise level allowed for continuous working. Being difficult to synthesize, recorded babble noise was used. Pink noise was selected because of its good speech masking properties (Rao, Letowski 2006). Its spectral envelope follows the spectral properties of speech signals: the peak intensity in the f0—F1 range and an even roll-out of 6 dB per octave at the higher frequencies of F2—F5 formants. Masking was on throughout the recording of each utterance, and the noise masks were presented via Sennheiser PC161 headphones, which were calibrated in the beginning of each session by Brüel and Kjaer Type 2235 SPL meter to deliver 83 +/– 0.5 dB_A SPL at the pink noise mask.

3. Results

3.1. Short versus long vowels

The individual results of articulated Finnish vowels in the F1—F2 space are illustrated in Figure 1. As can be seen from the figure the /y/ and /i/, and correspondingly, /ø/ and /e/ categories overlap clearly with each other. The short and long vowels differ in terms of F1 and F2 between the categories with the differences being largest between /u/ and /u:/. Except for /y/ and /ø/, the other vowel categories show a pattern where short vowels are more centralized than long vowels. This is in accordance with the results of Iivonen, Laukkanen 1993.

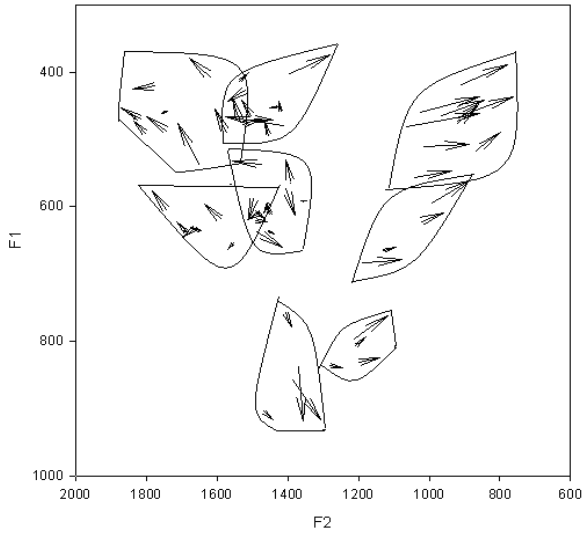


Figure 1. **Individual articulations of the short and long Finnish vowels in the F1—F2 space (in mel). Vector starting points represent the short vowels and end points the long vowels. Note that the number of subjects varies in different categories. The categories are from top left to right down: /i/, /y/, /u/, /e/, /ø/, /o/, /æ/ and /ɑ/.**

The mean values of the five repetitions of all subjects for the short and long Finnish vowels are shown in Table 2 and Table 3, respectively, and illustrated in Figure 2. The grand average of mean durations of all vowel categories was 125 ms (and for standard deviations SD 34 ms) for the short vowels and 345 ms (SD 75 ms) for the long vowels, resulting the durational ratio 1 : 2.8 between the short and long vowels. The coefficient of variation (CV = SD/mean) was slightly higher for the short vowels (0.27) than for the long vowels (0.21). These results are in line with the earlier reports on the durational variation of the Finnish short and long vowel quantities (for a review, see Eerola, Savela, Laaksonen, Aaltonen 2012).

Table 2

Mean values (and standard deviations) of the durations (in ms) and formants F1—F4 (in mel) for the produced short Finnish vowels

Vowel	Duration	F1	F2	F3	F4
i	103 (25)	471 (37)	1708 (107)	1902 (63)	2135 (104)
e	120 (32)	617 (27)	1608 (93)	1862 (62)	2129 (111)
æ	140 (42)	840 (61)	1408 (43)	1786 (14)	2010 (52)
y	118 (36)	452 (33)	1452 (60)	1748 (88)	2037 (48)
ø	125 (24)	599 (33)	1448 (46)	1805 (69)	2093 (88)
u	113 (28)	483 (40)	968 (71)	1791 (104)	2037 (92)
o	139 (43)	642 (41)	1083 (92)	1803 (62)	2032 (85)
ɑ	140 (43)	818 (19)	1225 (37)	1801 (33)	2054 (45)

The averaged results confirm the earlier findings that there are minor quality differences of 29—128 mel between short and long vowels in Finnish (Table 4, column S—L). The mean individual distance in the F1—F2 plane between the long and short vowels without noise masking was 62 mel over all vowel categories. Variation was found between vowel categories: /e/ and /ø/ had distances of 29—39 mel and no centralization tendency was observed, whereas /o/, /u/ and /æ/ showed clearly larger distances, up to 128 mel. Noticeable centralization of the short vowels was found especially in /i/, /u/, /o/, /a/, and /æ/ (Figure 2). The individual differences in F1 and F2 values were tested using Wilcoxon signed rank test. Differences between short and long vowels were significant for /i/ in F1 ($Z = -2.497$, $p = 0.013$) and F2 ($Z = -2.807$, $p = 0.005$), for /e/ in F2 ($Z = -2.499$, $p = 0.012$), for /y/ in F1 ($Z = -2.499$, $p = 0.012$), and for /u/ in F1 ($Z = -2.524$, $p = 0.012$).

Table 3

Mean values (and standard deviations) of the durations (in ms) and formants F1—F4 (in mel) for the produced long Finnish vowels

Vowel	Duration	F1	F2	F3	F4
i	301 (59)	449 (29)	1749 (108)	1946 (66)	2147 (113)
e	316 (51)	617 (30)	1630 (99)	1872 (59)	2142 (109)
æ	387 (96)	883 (69)	1374 (57)	1797 (39)	2078 (67)
y	329 (58)	436 (41)	1449 (88)	1732 (92)	2044 (77)
ø	326 (71)	603 (42)	1444 (62)	1791 (85)	2110 (101)
u	336 (74)	461 (45)	842 (57)	1799 (113)	2071 (107)
o	396 (94)	628 (53)	1004 (98)	1818 (51)	2032 (74)
ɑ	366 (95)	805 (34)	1170 (56)	1801 (37)	2055 (69)

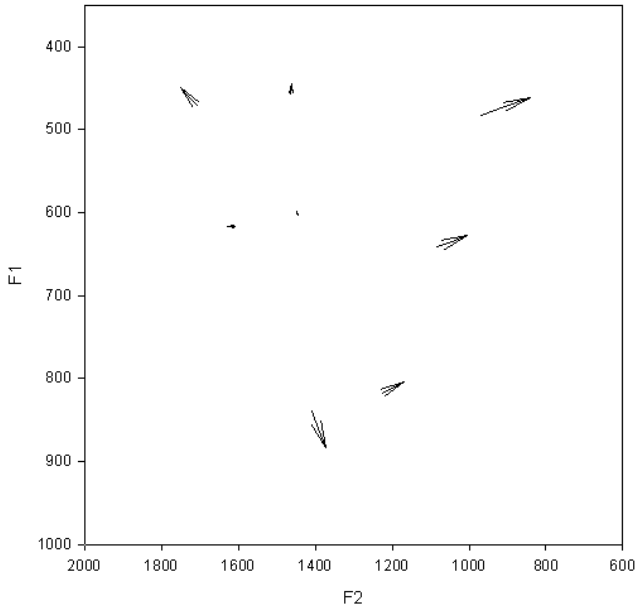


Figure 2. **The grand averages of short and long Finnish vowels in the F1—F2 space (in mel). Vector starting points represent the short vowels and end points the long vowels. The number of subjects varies in different categories. The categories are from top left to right down: /i/, /y/, /u/, /e/, /ø/, /o/, /æ/ and /ɑ/.**

3.2. The effect of a masking noise

Interestingly, both types of noise masking caused a highly significant prolongation in the duration of the short vowels, but not of the long vowels. With babble noise, the mean durations over 61 subjects were 143, ms (SD 37 ms) and 349 ms (SD 76 ms), and correspondingly with pink noise, 130 ms (SD 32 ms) and 341 ms (SD 79 ms). By using Wilcoxon signed rank test, the differences in duration between the quiet (Q) and noise (B = Babble, P = Pink) conditions were significant for short vowels in Q versus P ($Z = -3.040$, $p = 0.002$), and in Q versus B ($Z = -6.037$, $p = 0.000$). In case of long vowels the differences between the two noise conditions were significant; in B versus P ($Z = 2.069$, $p = 0.039$).

Table 4

Mean values of individual Euclidean distances (and standard deviations) in mels between the produced short (S) and long (L) Finnish vowels without noise masking (column S—L), and between the short vowels without and with babble (SBN) and pink noise (SPN) masking, and between the long vowels without and with the babble (LBN) and pink noise (LPN)

Vowel	S—L	SBN	SPN	LBN	LPN
i	49 (22)	59 (33)	58 (35)	60 (33)	46 (25)
e	29 (16)	53 (34)	53 (34)	63 (40)	59 (38)
æ	59 (36)	50 (11)	33 (11)	92 (116)	40 (17)
y	56 (48)	59 (34)	56 (37)	80 (57)	86 (68)
ø	39 (23)	72 (52)	53 (23)	77 (73)	76 (58)
u	128 (44)	51 (26)	51 (27)	86 (42)	85 (43)
o	80 (37)	55 (24)	38 (16)	51 (36)	43 (17)
ɑ	57 (32)	46 (20)	44 (14)	28 (7)	33 (7)
Mean	62	56	48	67	58

Since the durations increased along with increasing sound pressure level, the phenomenon may rather be explained by the amplitude of the mask than its type. However, when using a low pass filtered white masking noise, Summers, Pisoni, Bernacki, Pedlow and Stokes (1988) did not find any significant differences between the effects of the 80 dB and 90 dB SPL masks on durations, but instead, they found a highly significant ($p < 0.0001$) difference between non-masking and masking conditions. On the other hand, Beckford Wassink, Wright and Franklin (2007) did not find significant differences in segment durations between Lombard speech and (non-mask) citation speech. Our finding that the short vowels are prolonged with Lombard speech is interesting and motivates further investigation.

The effect of noise on the produced vowel quality was similar in both two masking conditions, and no major differences between babble and pink noise were found (Figure 3). Both noise types seem to cause higher F1 frequencies in the production of the mid-high vowels: On the average, the F1 values of the short and long vowels produced in the masking conditions are about 34 mel higher than without masking. No similar effect was found for the low vowels /a/ and /æ/. The results indicate that noise masking causes a systematic shift of F1—F2 values in the production of mid-high Finnish vowels, as illustrated in Figure 3. By using Wilcoxon signed rank test, the differences in F1 between the quiet (Q) and noise conditions (B = Babble, P = Pink) were significant for short vowels in Q versus P ($Z = –5.872$, $p = 0.000$), and in Q versus B ($Z = –5{,}983$, $p = 0.000$), and for long vowels in Q versus P ($Z = –5.732$, $p = 0.000$), and in Q versus B ($Z = –5.671$, $p = 0.000$).

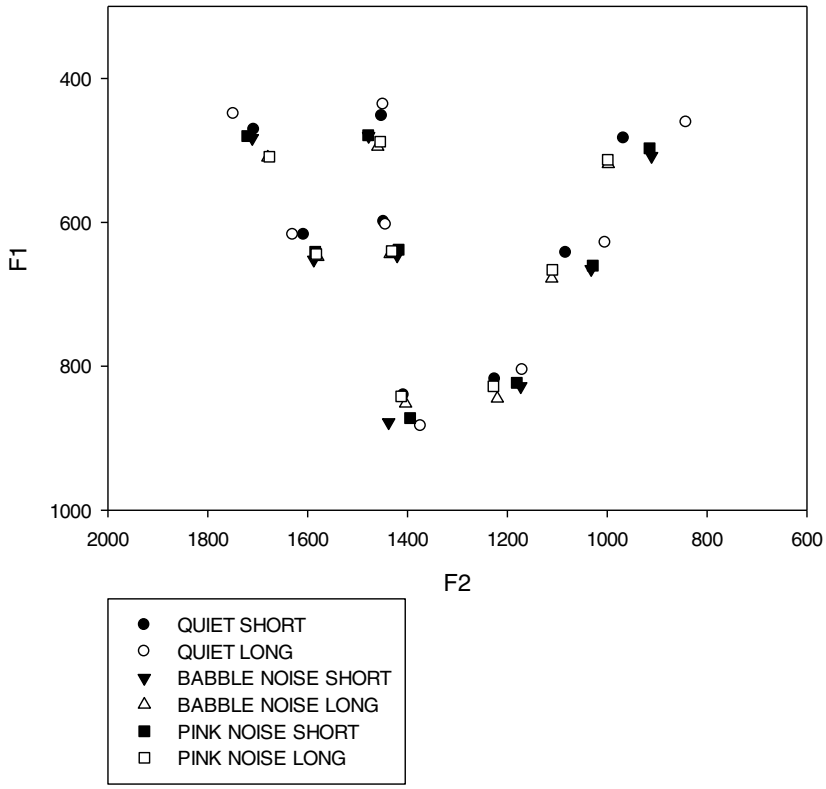


Figure 3. **The grand averages of short and long Finnish vowels in the F1—F2 space (in mel) in the two different masking conditions and without noise masking. The number of subjects varies in different categories. The categories are from top left to right down: /i/, /y/, /u/, /e/, /ø/, /o/, /æ/ and /ɑ/.**

4. Discussion and conclusions

The results of this study on the production of the short and long Finnish vowels confirmed, first, the earlier findings that the short vowels /i/, /u/, /o/, /a/ and /æ/ are more centralized in the F1—F2 space than their longer counterparts. Second, the Lombard effect induced by the two different noise masks caused the duration of the short vowels, but not the long ones, to increase significantly. The increase was larger with the louder babble noise than with the pink noise. Whether this difference was due to the higher amplitude of the babble noise or due to the noise type itself is a subject for further studies.

Third, the Lombard effect resulted in an increase in the F1 of the mid-high vowels, but had no effect on the Euclidean distances of the short and long vowels. These results in terms of the F1 value and the Euclidean distances are in line with the findings of Summers, Pisoni, Bernacki, Pedlow, Stoke (1988), and Beckford Wassink, Wright, Franklin (2007). The latter study among Jamaican speakers is particularly interesting, since Jamaican Creole utilizes the phonemic vowel length in a similar manner as Finnish, which, however, is a distinctive quantity language. The vowel quality (in terms of F1 and F2) was affected similarly by the Lombard speech in both these languages, but a clear durational prolongation of short vowels was only found in Finnish

Acknowledgments

The study was supported by a grant from the Finnish Cultural Foundation. We wish to thank Lea Heinonen-Eerola, M.A., for revising the English language of the manuscript.

Addresses

Osmo Eerola
Faculty of Telecommunication and e-Business, Turku University of Applied Sciences, and Department of Biomedical Engineering, Tampere University of Technology, and Centre for Cognitive Neuroscience, University of Turku
E-mail: osmo.eerola@tut.fi

Janne Savela
Department of Information Technology, University of Turku
E-mail: jansav@utu.fi

REFERENCES

Beckford Wassink, A., Wright, R., Franklin, A. 2007, Intraspeaker Variability in Vowel Production. An Investigation of Motherese, Hyperspeech, and Lomabrd Speech in Jamaican Speakers. — Journal of Phonetics 35, 363—379.

Castellanos, A., Benedi, J.-M., Casacuberta, F. 1996, An Analysis of General Acoustic-Phonetic Features for Spanish Speech Produced with the Lombard Effect. — Speech Communication 20, 23—35.

Eerola, O., Laaksonen, J. P., Savela, J., Aaltonen, O. 2003, Perception and Production of the Short and Long Finnish [i] Vowels. Individuals Seem to Have Different Perceptual and Articulatory Templates. — Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, 989—992.

Eerola, O., Savela, J. 2011, Differences in Finnish Front Vowel Production and Weighted Perceptual Prototypes in the F1—F2 Space. — Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong,

631—634.
Eerola, O., Savela, J, Laaksonen, J. P., Aaltonen, O. 2012, The Effect of Duration on Vowel Categorization and Perceptual Prototypes in a Quantity Language. — Journal of Phonetics 40, 315—328.
Harrikari, H. 2000, Segmental Length in Finnish. Studies within Constraint-Based Approach, Helsinki (Publications of the Department of General Linguistics, University of Helsinki 33).
Iivonen, A., Harnud, H. 2005, Acoustical Comparison of the Monophthong Systems in Finnish, Mongolian, and Udmurt. — Journal of the International Phonetic Association 35, 59—71.
Iivonen, A., Laukkanen, A. 1993, Explanations for the Qualitative Variation of Finnish Vowels. — Studies in Logopedics and Phonetics 4, 29—55.
Iivonen, A., Tella, S. 2009, Vieraan kielen ääntämisen ja kuulemisen opetus ja harjoittelu. — Puhuva ihminen — puhetieteiden perusteet, Helsinki, 269—281.
Jones, D. 1944, Chronemes and Tonemes. A Contribution to the Theory of Phonemes. — Acta Linguistica IV, 1—10.
Karlsson, F. 1983, Suomen kielen äänne- ja muotorakenne, Porvoo.
Kukkonen, P. 1990, Patterns of Phonological Disturbances in Adult Aphasia, Helsinki (SKST 529).
Lane, H. L., Tranel, B. 1971, The Lombard Sign and the Role of Hearing in Speech. — Journal of Speech and Hearing Research 14, 677—709.
Rao, M., Letowski, T. 2006, Callsign Acquisition Test (CAT). Speech Intelligibility in Noise. — Ear & Hearing 27, 120—128.
Summers, W. Van, Pisoni, D., Bernacki, R., Pedlow, R., Stokes, M. 1988, Effects of Noise on Speech Production. Acoustic and Perceptual Analyses. — Journal of the Acoustical Society of America 84, 917—928.
Suomi, K., Toivanen, J., Ylitalo, R. 2006, Fonetiikan ja suomen äänneopin perusteet, Helsinki.
Wiik, K. 1965, Finnish and English Vowels, Turku (Annales Universitatis Turkuensis. Series B 94).

ОСМО ЭЭРОЛА, *ЯННЕ САВЕЛА* (Турку)

ПРОИЗНОШЕНИЕ КОРОТКИХ И ДЛИННЫХ ФИНСКИХ ГЛАСНЫХ ПРИ ШУМОВОЙ МАСКИРОВКЕ И БЕЗ ШУМА

Для дальнейшего изучения возможных различий по качеству коротких и длинных финские гласных, мы исследовали формантные частоты F1—F4 и длительности восьми финских гласных /a/, /e/, /i/, /o/, /u/, /y/, /æ/ and /ø/, произнесенных в контексте слова (например, /tili/ — /tiili/) в двух различных условиях шумовой маскировки и без шума. Для шумовой маскировки использовали речевой шум на уровне 92 дБ и розовый шум на уровне 83 дБ. Установлено, что различия по качеству между короткими и длинными гласными незначительны, но шум маскировки привел к значительному удлинению длительности кратких гласных и, кроме того, к значительному повышению частоты F1.

Publication IV

Savela, J., Eerola, O., Aaltonen, O. 2014.

Weighted vowel prototypes in Finnish and German.

Journal of the Acoustical Society of America, Vol. 135 (3). Pages 1530-1540.

Weighted vowel prototypes in Finnish and German

Janne Savela[a)]
Department of Information Technology, University of Turku, FI-20014 Turku, Finland

Osmo Eerola
Department of Electronics and Communications Engineering, Tampere University of Technology, FI-33101 Tampere, Finland

Olli Aaltonen
Institution of Behavioural Sciences, University of Helsinki, FI-00014 Helsinki, Finland

(Received 9 May 2012; revised 31 December 2013; accepted 21 January 2014)

This study explores the perceptual vowel space of the Finnish and German languages, which have a similar vowel system with eight vowels, /ɑ/ /e/ /i/ /o/ /u/ /y/ /æ∼ɛ/ /ø/. Three different prototypicality measures are used for describing the internal structuring of the vowel categories in terms of the F1 and F2 formant frequencies: The arithmetic mean (centroid) of the F1–F2 space of the category (Pc), the absolute prototype of the category (Pa), and the weighted prototype of the category (Pω), in which the stimulus formant values are weighted by their goodness rating values. The study gave the following main results: (1) in both languages, the inter-subject differences were the smallest in Pω, and on the order of Difference Limen (DL) of F1–F2 frequencies for all of the three measures, (2) the Pa and Pω differed significantly from the centroid, with the absolute prototypes being the most peripheric, (3) the vowel systems of the two languages were similar (Euclidean distances in Pω of Finnish and German 7–34 mels) although minor differences were found in /e/, / ø/, and /u/, and (4) the mean difference of the prototypes from some earlier published production data was 100–150 mels.
 [http://dx.doi.org/10.1121/1.4864305]

PACS number(s): 43.71.Es, 43.71.An, 43.71.Hw [BRM] Pages: 1530–1540

I. INTRODUCTION

Prototype theories form a major framework for understanding how people interpret the phonetic percepts as belonging to particular categories (Samuel, 1982; Kuhl, 1993; Rosner and Pickering, 1994; Iverson and Kuhl, 2000; Boersma, 2006). Within this framework, perception is built on the emergence of auditory-phonetic prototypes that serve as the basis for categorization. In contrast to the classical boundary-based approaches (e.g., Nábelek *et al.*, 1993), the prototype theories emphasize the role of memory reference points in perception processes. These reference points are represented in the vowel space as areas of high prototypicality. In the present study, it is assumed that, within a particular phonetic category, there are subsets of category members that are more representative to the category than the category members in general. This study explores, by means of three different prototypicality measures, the areas of high prototypicality in the vowel systems of Finnish and German. The comparison of these two linguistically unrelated languages is motivated by the fact that their vowel systems are phonologically comparable.

A. Prototypes

The prototype view stems from the fundamental work of Rosch (1975), which suggests that perceptual categories are gradual and based on prototypes. By definition, prototypes are more representative of their categories than the other exemplars of the same category are. A prototype can be defined as the centroid of the category (e.g., Rendell, 1986), which means that the most representative category member is the one which is closest to the majority of stimuli identified as belonging to the same category. In exemplar theories, the identity of the stimulus is based on the highest relative familiarity with a certain category, as compared to other categories in the given perceptual space (Nosofsky, 1988; Lacerda, 1995). The idea that all category members identified as belonging to a particular category are not equally important representatives of the category was presented several decades ago by, for example, Fairbanks and Grub (1961), but the understanding of this phenomenon has been accumulated in recent years (e.g., Kuhl, 1991; Thyer *et al.*, 2000; Iverson and Kuhl, 2000; Evans and Iverson, 2004, 2007).

The vowel system of an individual listener is not invariable but can evolve over time (Evans and Iverson, 2004). In English, the dialects differ drastically in the numbers and loci of the individual vowel categories. According to Evans and Iverson, the listeners use the prototypes of their own dialect but adjust them depending on the dialect of the carrier sentence. The listeners changed their prototypes in the direction of the dialect of the carrier sentence although their basic models were based on their native area. In a later study (Evans and Iverson, 2007), they showed that the accent can change during the later age (during university studies). However, the prototypes are more robust in the case of native speakers.

The concept of adaptive dispersion (e.g., Johnson *et al.*, 1993; Johnson, 2000) suggests that the listeners use more peripheral formant values for the prototype vowels, as compared

a)Author to whom correspondence should be addressed. Electronic mail: jansav@utu.fi

to the same vowels in produced sentences. However, several prototype studies have also shown that individual vowel prototypes vary, and this variation can be based on individual differences in phonetic knowledge or experience (Aaltonen *et al.*, 1997; Morais and Kolinsky, 1994; Harinen *et al.*, 2011). Multiple studies support the idea that on-going sound processing utilizes the vowel representations in very primitive levels of the vowel identification.

In the present study, a goodness rating is used as the basic tool to examine the prototypicality and its relationship with identification constancy in the perceptual vowel space of the Finnish and German languages, which have a similar vowel system with eight vowels, /ɑ/ /e/ /i/ /o/ /u/ /y/ /æ∼ɛ/ /ø/. Three different prototypicality measures are used for describing the internal structuring of the vowel categories: The centroid of the category Pc(F1,F2), the absolute prototype Pa(F1,F2) of the category, and, as a novel approach, the weighted prototype Pω(F1,F2) of the category, in which the stimulus formant values are weighted by their goodness rating values (Eerola and Savela, 2011). Through these methods, data will be presented on the internal structure of the vowel categories, and the spread of prototypical vowels within the category. It will be examined how the differences in the measures of prototypicality affect the pattern and interpretation of the general role of the prototypes in the identification of categories.

There are obvious reasons motivating the use of weighted prototypes. Theoretically, there may be a true absolute prototype Pa(F1, F2) in the sense that it represents the highest ranking exemplar of all distinguishable category members for an individual listener. This value, however, is difficult to measure experimentally, since there are numerous amounts of distinguishable phones within each category, and it is practically impossible to attempt to synthesize and present them all in a listening experiment. Therefore, some new optimization methods have recently been presented for investigating the phonetic categorization without the need to play huge numbers of stimuli in the experiments (Iverson and Evans, 2003; Oglesbee and de Jong, 2007; Benders and Boersma, 2009). The weighted prototype Pω(F1,F2) approach enables us to avoid some of these experimental problems. The Pω(F1,F2) is robust in the sense that it represents the center of gravity of the category: The absolute prototype can most likely be found within the area of the vowel space where the majority of the stimuli with high goodness values are located.

B. Experiments on vowel perception

The experimental basis of this vowel study is the Turku Vowel Test (Raimo *et al.*, 2002b; Raimo *et al.*, 2002a). In contrast to earlier studies using large amounts of synthesized vowel stimuli for identification (e.g., Aaltonen and Suonpää, 1983; Määttä, 1983; Hose *et al.*, 1983), the Turku Vowel Test also systematically collects the goodness ratings for the identified vowels (from a set of 386 synthesized vowels). The Turku Vowel Test is a method for structured observational studies, and it has been used for exploring the vowel categories in multiple languages; for example, Savela (2009) showed data for 13 languages, and Raimo *et al.* (2002a) used it for the determination of subsets of the most solid vowel categories in 10 languages.

The present study focuses on two languages: Finnish and German. These two languages have similar vowel categories, at least, as far as the long vowels are concerned, and they have a relatively simple relationship between a sound and its orthographic symbol. They belong to two distinct language families (Finno-Ugric-Finnic and Indo-European-Germanic, respectively), which, however, are geographically close. Both Finnish and German have eight vowel categories (/ɑ/,/e/,/i/,/o/,/u/,/y/,/æ∼ɛ/,/ø/). In these two languages, there is a distinction between rounded and unrounded front vowels. All Finnish dialects have a similar number of vowel categories and distinctions. The possible differences between dialects are only studied cursorily in the present study. The Finnish vowel system is extremely quantitative: Each vowel can occur as a short or a long version in any position in a word. According to the identity group interpretation of the Finnish vowel system (Karlsson, 1983), the long vowel has the same quality but two times the duration of the short vowel. The studies on the quality differences between *produced* long and short vowels have shown minor differences of 40–120 mels in formants F1–F3 (e.g., Wiik, 1965; Kuronen, 2000; Eerola *et al.*, 2003), whereas the *perceived* quality differences are minimal (Eerola *et al.*, 2012). In Finnish, the duration difference between short (<100 ms) and long (>150 ms) vowels is large (e.g., Eerola and Savela, 2011).

In German, there are larger differences in the loci of F1 and F2 in the vowel space between long and short vowels (for an extensive review, see Becker, 1998). Bohn and Flege (1992) studied how German listeners perceived the (English) vowels with high F1 and F2 values. They found that inexperienced German subjects (without this distinction in their own dialect) used the duration in making a difference between the two English vowel categories (/ɛ/ and /æ/). Although not all German speakers make a difference between /e:/ and /æ:/ when pronouncing words, this distinction exists between the names of the symbols <e> and <ä> (if pronounced separately). The differences in the phonological quality of these variants are mainly based on duration, but also on the location in the vowel space (e.g., Heid *et al.*, 1995). According to a study on the duration of vowels in the PhonDat database of spoken German, a difference between long and short vowels exists, although it is not very large (67 vs 98 ms). However, the stimuli of the Turku Vowel Test were always of the same duration, 350 ms, which is clearly longer than the duration of the vowels reported by Heid and co-workers. German speakers are known to be sensitive for the quality difference between short and long vowels (e.g., Bennett, 1968; Sendlmeier, 1981; Jessen *et al.*, 1995). They are able to distinguish between the first vowels in word pairs such as <offen> [**o**f:en] and <Ofen> [o:fen] or <Mitte> [m**I**tte] and <Miete> [mi:te]. In those words, the Euclidean (quality) distance between a short and long vowel is larger than it would be in corresponding cases in Finnish. In produced vowels, the distance is about 40–120 mels in Finnish (e.g., Kuronen, 2000) and 120–140 mels in German (Sendlmeier and Seebode, 2006). There is a classical view that German would have 15 distinctive vowel categories (e.g., the database study

of Heid *et al.*, 1995). However, the quality difference between the German vowels of different quantities is dependent on the syllable structure of the word and results from different centralization mechanisms (Becker, 1998). Becker argues that there are only eight vowel categories in German, and they are in correspondence with the eight orthographic vowel qualities.

C. Aims of the present study

1. The individual differences

The first aim of the present study is to explore the expected differences in vowel categories between native speakers of a given language. The question is whether a similar variation in the loci of prototypes can be found for a larger set of vowels and for two languages as has been obtained for the Finnish /i/ in some earlier studies. In the data for Aaltonen *et al.* (1997) and Eerola *et al.* (2012), the inter-subject differences of /i/ prototypes were over 8% in mel scale in the F1-F2 space. In another study, by Lively and Pisoni (1997), it was found, by using an orbit of 33 vowels, that 21 of these vowels were perceived as the best exemplar by at least one of the subjects (N = 78), whereas only four of the stimuli were labeled as the best representatives of the category by the majority of the subjects. These findings suggest that there are individual differences in the goodness ratings of vowels.

Furthermore, since there is no clear evidence in the literature how the individual prototypes are distributed within a category (e.g., either evenly throughout the category or concentrated in the central or in the more peripheral subareas), the normality of their distribution is examined for all vowel types. Kuhl (1991) showed that, among the stimuli used in goodness rating, the best stimulus was systematically in the center of the goodness rating and the goodness rating decreased when the distance from the orbit center increased. The data presented by Lively and Pisoni (1997) suggest that the differences between subjects are presumably larger than those reported by Kuhl.

2. Differences between prototype measures

The measures subject to comparison are the centroid, the weighted prototype (ratings 3–7 on scale 1–7 are included, see Sec. II), and the absolute prototype (ratings 6–7 are included). The comparisons were assumed to make it possible to show the difference between the prototypes based on goodness rating and the category centroid. Assuming that the general hyperspace effect (Johnson *et al.*, 1993; Johnson, 2000) affects the vowel perception, the most peripheral stimuli should receive the best ratings, and consequently, the loci of the absolute and weighted prototypes should differ essentially from those of the centroids. On the other hand, if the centroid (obtained as the arithmetic category center without weighting by goodness) forms the prototype, the loci of all three measures should be the same.

3. The differences between Finnish and German vowels

In the present study, the clustering of vowels belonging roughly to the same category is compared between two languages, namely, Finnish and German. Mean prototypes may differ between languages, and it is tested how the differences between sounds are manifested in Finnish and German. There are few studies on multilingual comparisons in this respect. For example, Willerman and Kuhl (1996) compared vowels in Sweden-Swedish, Texan Spanish and American English. The study showed language-specific differences in the discrimination of [y]. Swedish listeners performed poorer in discriminating the vowel pairs typical of their native /y/ type of sounds than did the speakers of the other two languages, which do not have any /y/ vowels.

On the basis of the results obtained in some recent vowel production tests, the greatest difference between German and Finnish vowels is in the high F1–F2 area (e.g., Kuronen, 2000; Sendlmeier and Seebode, 2006). The Finnish /æ/ should have lower F1 than the German /ɛ/. On the other hand, German /e/ should have lower F1 than Finnish /e/. If the vowel prototypes reflect the idea of adaptive dispersion (Johnson *et al.*, 1993; Johnson, 2000), and dispersion in general (e.g., Becker-Kristal, 2010), the languages should be similar. If the vowel prototypes reflect the differences in production, there should be larger differences.

The question of language-specific and general phonetic prototypicality (Gottfried and Beddor, 1988; Strange *et al.*, 2007) can be examined in this context by comparing the differences between languages in terms of prototypicality. It can be asked how these two types of prototypicality co-occur in the present data representing two languages with a similar vowel pattern, but with different prototypicality areas. It can be further asked whether non-prototypical vowels are perceived similarly on the basis of the phonetic prototypes or the abstract phonology system.

4. The differences between production and perception

Finally, the perceptual data obtained in this study are compared to some earlier published data for produced vowels (Kuronen, 2000; Sendlmeier and Seebode, 2006). The idea is to test whether the observations presented by Eerola and Savela (2011) would be replicated when the perceptual and production data originate from different subject groups: In the production data, the formants of the weighted prototypes were closer to the produced vowels (in Euclidean vowel space) than the absolute prototypes.

II. METHOD

A. Participants

The perceptual data collected for this study comes from altogether 86 Finnish and German speaking listeners who volunteered as subjects in the Turku Vowel Test (for these two languages). The subjects were students, exchange students, and staff members at the University of Turku. There were 68 Finnish and 18 German speaking listeners. The mean age of Finnish listeners was 25.5 yr, and there were 33 female and 35 male listeners. The mean age of German listeners was 26.2 yr, and 12 were female and 6 male. Of the Finnish listeners, 36 represented South Western dialects, 19 Tavastian dialects, 7 Savonian dialects, and 6 other dialectal

areas. Of the German listeners, 4 were North German, 8 Middle German, and 6 Upper German (including South German, Swiss, Austrian, and South Tyrolean) listeners. In both language groups, the listeners' self-reported knowledge of other languages varied: On an average, 2.5 languages for Finnish listeners (English and Swedish), and 2.1 for German listeners (English and French). The Finnish subjects were university students with high fluency in Finnish. The German subjects were mainly exchange students. Three of them reported some knowledge of Finnish, and 16 Finnish students reported knowledge of German. None of the subjects reported hearing impairments.

B. Stimuli

The stimuli (n = 386) consisted of synthetic vowels covering the entire vowel space. Diphthongs and nasal vowels were not included. The stimuli were synthesized with the Klatt parallel synthesizer (Klatt, 1980). The vowel space was created by varying F1 from 250 to 800 Hz in steps of 30 mels and F2 from 600 to 2800 Hz in steps of 40 mels. F3 was 2517 Hz, when F2 was 1500 mels (1995 Hz) or below, and increased by 200 mels, when F2 was above 1995 Hz (i.e., F3 = F2 + 200 mels, when F2 > 1995 Hz). F4 was 3500, when F3 was 3300 Hz or below, and 4000 Hz, when F3 was over 3300 Hz. The duration of the vowel stimuli was fixed at 350 ms, with the pitch initially (0–120 ms) rising from 100 Hz to 120 Hz and then falling to 80 Hz during the rest of the stimulus. The duration of the stimuli was chosen to represent the long vowels in both languages.

C. Procedure

The experiments were run in the Language Centre of the University of Turku, which has a noise-free environment designed for spoken language exercises. The instructor was present during the entire session, which lasted 45 min. The number of listeners participating in each session varied.

The stimuli were presented in a pseudorandom order through head phones (PMB K 800) that were connected to a PC by using a special-purpose JAVA applet. The listeners were asked to categorize the stimuli to one of their native categories that were displayed on the screen. The list of possible answers is based on the vowel systems of the two languages. For Finnish they were as follows: /ɑ/ /e/ /i/ /o/ /u/ /y/ /æ/ /ø/ (in orthography: <a> <e> <i> <o> <u> <y> <ä> <ö>). In German they were as follows: /ɑ/ /e/ /i/ /o/ /u/ /y/ /ɛ/ /ø/ (in orthography: <a> <e> <i> <o> <u> <ü> <ä> <ö>).

A matrix was presented for each stimulus. In this matrix, there was a column for each vowel category and a row for each of the different ratings. The scale of the goodness rating was 1–7 (with 7 corresponding to a good and 1 to a poor exemplar). There was no option for a listener to give a zero (0) rating. Listeners responded to the stimuli by clicking the appropriate onscreen button with the mouse and then asking for the next stimulus. It was possible to repeat each stimulus as many times as the listener wanted. The last answer was saved for a particular stimulus. The subjects had the possibility to adjust the loudness level of stimuli to a comfortable level.

D. Analysis method

Three different prototypicality measures were used for describing the internal structuring of the vowel categories: The centroid of the category Pc(F1,F2), the estimated absolute prototype Pa(F1,F2) of the category, and the weighted prototype Pω(F1,F2) of the category. The centroid of the category Pc(F1,F2) was calculated as the mean value of the F1 and F2 values of all the stimuli identified as category members, whereas the estimate of the absolute prototype Pa(F1,F2) was formed by including only those stimuli that were rated as 6 or 7 (on scale 1–7).

The weighted prototype Pω(F1,F2) of each category was formed by applying Eq. (1) presented by Eerola and Savela (2011),

$$\mathbf{Fi} = (a_1 r_1 Fi_1 + a_j r_j Fi_j + \cdots + a_n r_n Fi_n)/ (a_1 r_1 + a_j r_j + \cdots + a_n r_n), \quad (1)$$

where

Fi = weighted formant frequency, i = 1,2;
Fi_j = formant i of stimulus j, j = 1,2,…,n;
a_j = evaluation score (1–7), j = 1,2,…,n;
r_j = identification consistency (0.7–1.0), j = 1,2,…,n;
n = number of stimuli identified as category members.

Pω(F1,F2) thus represents a point in the F1–F2 space (in mel scale) that is obtained by weighting the F1 and F2 values of each stimulus identified as a category member by the goodness rating value and unanimousity. In the present study, the stimulus identification and goodness ratings were done within the same session. The goodness rating scale was 1–7 without the possibility to rate 0 (i.e., a stimulus would not belong to any category); consequently, the identification consistency (r_j) is always 1 (in contrast to Eerola and Savela, 2011), and the evaluation score (a_j) represents the goodness rating value given to each stimulus. Stimuli with rating values 3, 4, 5, 6, and 7 were included in the calculation of Pω(F1,F2), and stimuli with rating values 6 and 7 were included for Pa(F1,F2). In both cases, the prototypes were calculated by using the weighting power of the ratings. If there were no answers meeting the criteria described above (6–7) for a particular stimulus category, the centroid (arithmetical mean) was used for all measures (4% of all cases in absolute prototypes). The subjects differed slightly in their manner of scoring the goodness of stimuli on scale 1–7 (Savela *et al.*, 2005). Some listeners utilized the entire scale, whereas some listeners had fairly even goodness ratings, resulting in less variation between different stimulus tokens. However, the weighting formula for Pω(F1,F2) rectifies possible bias to some extent, since the same rating value is used both in the nominator and denominator of the formula.

The F1 and F2 formant values (in mels) of the obtained prototype measures were subjected to standard statistical analyses, as explained in more detail in Sec. III. The Turku Vowel Test is a method for structured observational studies, and it has been used for exploring the vowel categories in multiple languages. This study reports the results obtained in collecting data with the Turku Vowel Test for two phonologically

similar languages, Finnish and German. The number of subjects differs for Finnish (N = 68) and German (N = 18), but unequal sample sizes do not prevent statistical comparisons between the languages; rather, the larger number of Finnish listeners improves the power of the statistical comparison tests of these two populations.

III. RESULTS

A. Weighted prototypes—Individual variation

For each subject, the weighted category prototypes were computed by using Eq. (1) and the obtained stimulus goodness ratings (Table I).

The individual weighted prototypes for Finnish and German are presented in Fig. 1. The individual vowel categories can be seen as clusters in the psychoacoustic F1–F2 space. Figure 1 further shows that the weighted prototypes are not evenly distributed within a category, but rather, they are strongly concentrated around the central areas of each category. However, in both languages there are a few outliers outside the 90% identification curve: For /i/, /y/, and /e/ in Finnish, and for /e/ and /ɛ/ in German.

Maps of goodness ratings were computed for the Finnish and German vowel systems (Fig. 2). Visual analysis indicates that the weighted prototypes form fairly similar clusters in the two languages and always falls (by definition) within the domain of the areas of ≥3 ratings.

The areas of the highest goodness ratings (6 and 7), representing the absolute prototypes, are in some cases more peripheral than the weighted prototype clusters, especially in the case of German /e/ and /ɛ/. This difference will be discussed later in this article. The weighted vowel prototypes were, in general, normally distributed in both languages. The question is discussed in more detail in Sec. III B below.

TABLE I. The mean values for formant F1 and F2 frequencies (in mel scale), with standard deviations (SD), coefficients of variation (CV in %), median, maximum, and minimum, of the weighted prototypes (Pω) for the vowels used in the Turku Vowel Test in Finnish (FI, N = 68) and German (DE, N = 18).

		Mean		SD		CV		Median		Max		Min	
		F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
FI	/i/	376	1689	9	16	2.4	0.9	377	1688	399	1726	350	1632
FI	/y/	385	1383	13	43	3.4	3.1	385	1387	424	1496	353	1285
FI	/e/	547	1595	26	19	4.8	1.2	550	1596	601	1653	451	1551
FI	/æ/	777	1415	17	16	2.1	1.1	779	1414	821	1450	728	1372
FI	/ø/	565	1315	24	32	4.2	2.4	565	1314	625	1396	514	1238
FI	/ɑ/	792	1045	17	16	2.1	1.5	790	1045	824	1115	754	1000
FI	/o/	576	864	20	21	3.5	2.4	579	861	615	911	510	802
FI	/u/	406	830	13	46	3.2	5.5	406	826	443	926	378	744
DE	/i/	376	1681	6	15	1.6	0.9	376	1683	388	1708	368	1652
DE	/y/	383	1375	7	40	1.8	2.9	384	1374	395	1463	372	1284
DE	/e/	517	1604	20	23	3.9	1.4	518	1604	551	1652	464	1551
DE	/ɛ/	761	1441	21	24	2.7	1.7	762	1437	792	1494	721	1405
DE	/ø/	556	1294	16	29	2.9	2.2	554	1291	586	1363	537	1247
DE	/ɑ/	806	1064	16	18	2.0	1.7	804	1060	845	1112	783	1046
DE	/o/	558	861	11	19	2.0	2.2	556	859	587	921	538	834
DE	/u/	398	848	8	40	2.0	4.7	399	841	411	958	385	790

In order to test the significance of individual differences within a language, the coefficients of variation (CV = σ/μ) were computed for F1 and F2 of each vowel prototype (Table I). CV was interpreted as an indicator of the Difference Limen (DL = ΔF/F in AX discrimination experiment, where F = formant frequency, and ΔF = the just noticeable difference of F) for the Pω(F1,F2) of a particular vowel category. The DL of frequency is a measure that describes the size of the step within the vowel space that is still distinguishable. This measure is affected by the different experimental designs that facilitate or non-facilitate the discrimination task (for a review, see Hawks, 1994). In Flanagan's test, the compared stimulus pairs are presented in random order and the subject cannot know what kind of difference to expect (Flanagan, 1955).

The results (see Table I) showed that, on the average (over all listeners and all vowels), the inter-individual variation (CV) was 2.8% for F1, and 2.3% for F2, and was slightly larger for Finnish vowels (3.2% for F1, and 2.3% for F2)

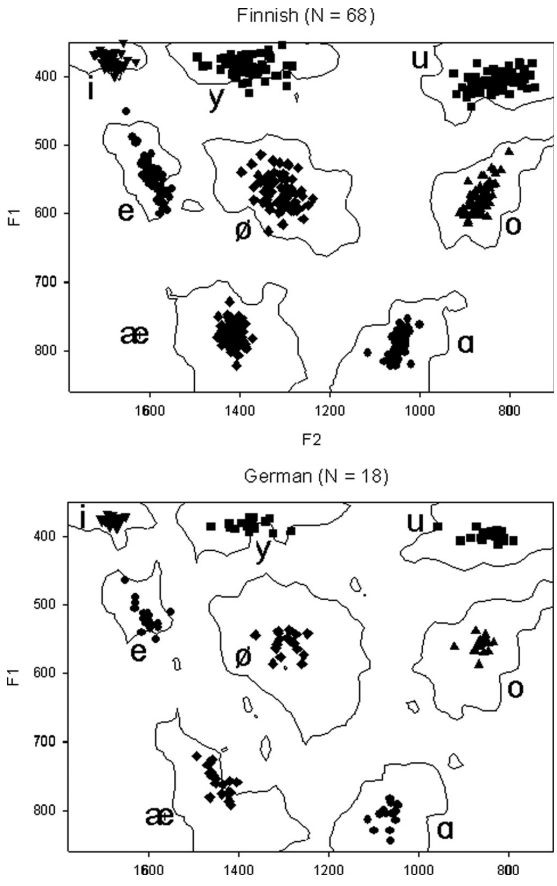


FIG. 1. Clusters of individual weighted vowel prototypes Pω(F1,F2) in the F1–F2 formant space (mel scale) obtained in the identification experiments for Finnish and German. Different symbols are used for prototypes representing different categories within a language. The clusters of prototypes indicate the spread of vowel categories in each language. The solid lines indicate areas where 90% of listeners categorized the stimuli similarly.

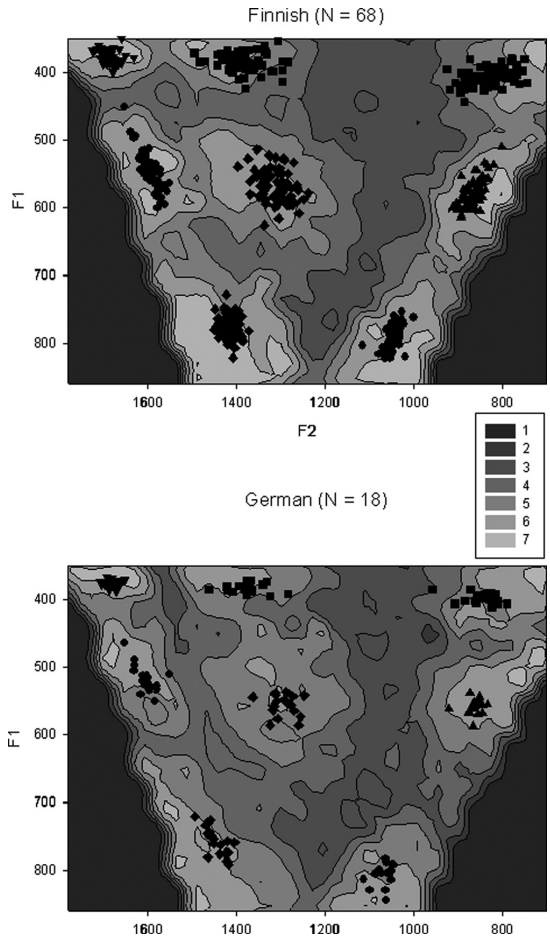


FIG. 2. Goodness ratings (on scale 1–7) and weighted vowel prototypes Pω(F1,F2) in the F1–F2 formant space (mel scale) obtained in the identification and rating experiments in Finnish and German. The average goodness rating of the majority category is represented in shades of gray color as indicated by the legend.

than for German vowels (2.4% for F1, and 2.2% for F2). This difference may be due to the larger sample size of Finnish listeners, and it is mainly reflected in F1. For Finnish listeners, the inter-individual variation was largest in /e/ (CV = 4.8%) and /æ/ (CV = 4.2%) for F1, and in /u/ (CV = 5.5%) and /y/ (CV = 3.1%) for F2. Correspondingly for German listeners, the variation was largest in /e/ (CV = 3.9%), /ø/ (CV = 2.9%) and /ɛ/ (CV = 2.8%) for F1, and /u/ (CV = 4.7%) and /y/ (CV = 2.9%) for F2. Interestingly, the largest variation manifests in vowels that are similar in both languages. In all categories (except perhaps F2 of the Finnish /u/), the CVs were at the level of the DL for F1 and F2 (3–5%) (Flanagan, 1955; Hawks, 1994), thus suggesting that the observed differences between individual weighted prototypes and the mean of individual prototypes are hardly noticeable and similar across the listeners of the same language. The possible effects of the listener's dialect background were calculated for Finnish subjects. When comparing the two largest dialect groups (South Western, N = 36, and Tavastian, N = 19), the differences between the grand averages of similar vowel categories in these dialects were small: $\Delta F/F_T = 0.48\%$ (ΔF is the difference in F1 or F2 formant frequency of the two dialect groups, and F_T is the corresponding formant frequency of the Tavastian dialect group), and the independent sample *t*-tests for all vowel types showed that the differences were insignificant.

B. Weighted prototypes, absolute prototypes and the centroid of the category

The absolute prototypes were calculated for all vowel categories in the two languages (Table II and Fig. 3). The stimuli that received ratings of 6 and 7 were included in the absolute prototype. As indicated in Fig. 3, there are differences between the prototype measures of Finnish and German for the non-rounded open front vowels (/e/ and /æ - ɛ/) and rounded vowel /o/. This is reflected in the differences between Pa(F1, F2) and Pω(F1,F2), as illustrated in Fig. 3.

Second, the centroid of the category (Pc) was calculated. Mathematically, the centroid is the mean of F1 and F2 of all those stimuli which were categorized as belonging to the same category. In this approach, the goodness evaluation of the stimuli is not considered for weighting the centroid. Since the grid used in the test covered the entire vowel space, the category centroid measure is supposed to show how the listeners identify the poor vowels that contribute more to this measure than they do in the weighted prototype, because all of the stimuli were categorized into one of the given categories. Furthermore, the shift in category boundary also shifts the category centroid, whereas in the prototype measures, the boundary shift does not necessarily shift the measure.

The absolute distances to the median value of the entire vowel grid used in the experiment (F1, 605 mels, F2, 1240 mels) were computed for all vowel categories. This was done to show the possible centralization-hyperspace effects for the category prototypes. In order to test the differences between prototype measures and vowel types, analysis of variance (ANOVA) was performed by using *vowel type* and *prototype measure* as the independent variables.

In Finnish, for F1, both the vowel type and the prototype measure were significant in explaining the absolute distance: Vowel type $F(7,469) = 1833.962$, $p < 0.001$, and prototype measure (Pa 146 mels, Pω 142 mels, Pc 134 mels) $F(2,134) = 150.847$, $p < 0.001$. The interaction between vowel type and prototype measure showed a significant effect [$F(14\,938) = 35.638$, $p < 0.001$] on the F1 distance, indicating that the prototype measures differed in different vowels. In Finnish, for F2, the effect of vowel type was significant in explaining the absolute distance [$F(7,469) = 2401.263$, $p < 0.001$]. The prototype measure (Pa 292 mels, Pω 272 mels, Pc 242 mels) had a significant effect [$F(2,136) = 544.611$, $p < 0.001$], and the interaction between vowel type and prototype measure showed a significant effect [$F(14\,938) = 104.618$, $p < 0.001$] on the acoustic distances. To further analyze the interactions, Wilcoxon's signed ranks test was first conducted for the comparison of the different prototype measures (with all vowels combined

TABLE II. The formant F1 and F2 frequencies (in mel scale) of the vowel prototypes (Pa = absolute prototype, Pω = weighted prototype) and category centroids (Pc) for the Finnish (N = 68) and German listeners (N = 18). The columns represent the mean, standard deviation and coefficient of variation (CV in %). Unrounded values were used in calculating CV. The different Euclidean distances are measured in F1–F2 formant space (in mel scale). The normally distributed formant samples are marked with #.

							Euclidean distance		
Finnish	F1 Pa	F2 Pa	F1 Pω	F2 Pω	F1 Pc	F2 Pc	Pa - Pω	Pc -Pω	Pc -Pa
/i/	372 (8) (2.2)	1693 (21) (1.2)	376 (9) (2.4)#	1689 (16) (0.9)#	383 (11) (2.9)#	1677 (21) (1.3)	6	14	19
/y/	379 (16) (4.2)	1425 (52) (3.6)	385 (13) (3.4)#	1383 (43) (3.1)#	398 (13) (3.3)#	1322 (40) (3.0)#	42	62	105
/e/	547 (26) (4.8)#	1593 (23) (1.4)	547 (26) (4.8)#	1595 (19) (1.2)#	546 (28) (5.1)	1591 (22) (1.4)#	2	4	2
/ø/	562 (21) (3.7)#	1344 (40) (3.0)#	565 (24) (4.2)#	1315 (32) (2.4)#	570 (27) (4.7)#	1280 (29) (2.3)#	29	35	64
/æ/	788 (16) (2.0)#	1421 (23) (1.6)	777 (17) (2.2)#	1415 (16) (1.1)#	762 (20) (2.6)#	1394 (14) (1.0)#	13	26	37
/ɑ/	801 (20) (2.5)#	1041 (20) (1.9)	792 (17) (2.1)#	1045 (16) (1.5)	777 (18) (2.3)#	1050 (16) (1.5)#	10	16	26
/o/	579 (20) (3.5)#	850 (23) (2.7)#	576 (20) (3.5)	864 (21) (2.4)#	577 (23) (4.0)#	895 (24) (2.7)	14	31	45
/u/	401 (16) (4.0)#	772 (45) (5.8)	406 (13) (3.2)#	830 (46) (5.5)#	415 (13) (3.1)#	907 (38) (4.2)#	58	78	136
German									
/i/	368 (8) (2.2)#	1687 (18) (1.1)#	376 (6) (1.6)#	1681 (15) (0.9)#	382 (8) (2.1)#	1659 (39) (2.4)	10	23	31
/y/	379 (10) (2.6)#	1374 (53) (3.9)#	383 (7) (1.8)#	1375 (40) (2.9)#	390 (9) (2.3)#	1339 (40) (3.0)#	4	37	37
/e/	497 (21) (4.2)#	1628 (24) (1.5)#	517 (20) (3.9)#	1604 (23) (1.4)#	527 (27) (5.1)#	1579 (37) (2.3)#	31	27	57
/ø/	530 (34) (6.4)#	1292 (29) (2.2)#	556 (16) (2.9)#	1294 (29) (2.2)#	570 (17) (3.0)#	1268 (25) (2.0)#	26	30	47
/ɛ/	772 (24) (3.1)#	1456 (41) (2.8)#	761 (21) (2.8)#	1441 (24) (1.7)#	752 (17) (2.3)#	1415 (31) (2.2)#	19	28	46
/ɑ/	822 (28) (3.4)#	1068 (26) (2.4)#	806 (16) (2.0)#	1064 (18) (1.7)	780 (14) (1.8)#	1055 (14) (1.3)#	16	28	44
/o/	542 (22) (4.1)	836 (35) (4.2)	558 (11) (2.0)#	861 (19) (2.2)	570 (14) (2.5)	900 (18) (2.0)#	30	41	70
/u/	393 (10) (2.5)#	823 (53) (6.4)	398 (8) (2.0)#	848 (40) (4.7)#	405 (9) (2.2)#	905 (31) (3.4)#	25	57	83

and Boniferroni correction used for multiple comparisons; a non-parametric test was used since distributions of measures were not normal). All pairwise comparisons (Pa - Pω, Pa - Pc, and Pω - Pc) reached significance ($p < 0.016$). Then, ANOVAs were performed individually for each vowel type, with prototype measure as the independent variable. The differences in all vowel types were significant ($p < 0.005$), except for /e/ (see Fig. 3) where the differences between prototype measures were insignificant [for F1, $F(2,134) = 0.717$, $p = 0.495$, and for F2, $F(2,134) = 2.219$, $p = 0.131$]. The pairwise comparisons showed non-significant differences between Pω and Pa in F1 for /ø/ and /o/ and in F2 for /ɑ/, and between Pω and Pc in F1 for /o/, and between Pc and Pa in F1 for /o/.

In German, for F1, both the vowel type and prototype measure were significant in explaining the absolute distance: Vowel type $F(7,119) = 475.467$, $p < 0.001$, and prototype measure (Pa 163 mels, Pω 150 mels, Pc 139 mels) $F(2,34) = 94.963$, $p < 0.001$. The interaction between vowel type and prototype showed a significant effect [$F(14\,238) = 6.172$, $p < 0.001$] on the F1 distance. In German, for F2, the effect of vowel type was significant in explaining the absolute distance [$F(7,119) = 575.843$, $p < 0.001$]. The prototype measure (Pa 278 mels, Pω 268 mels, Pc 240 mels) had a significant effect [$F(2,34) = 72.541$, $p < 0.001$], and the interaction between vowel type and prototype measure showed a significant effect [$F(14\,238) = 8.947$, $p < 0.001$] on acoustic distances. Using Wilcoxon's signed ranks test for prototype measures as explained above, all pairwise comparisons reached significance ($p < 0.016$). The ANOVAs were performed individually for each vowel type, with prototype measure as the independent variable. The differences in all vowel types were significant ($p < 0.005$), except for /ɑ/ where the differences between prototype measures were insignificant for F2, $F(2,34) = 3.734$, $p = 0.060$ (Greenhouse corrected). In pairwise comparisons, the difference was insignificant for F1 in /u/, and for F2 in /i/, y/, /ɛ/, /ø/ between Pω and Pa.

The ANOVA results thus showed that, in both languages, all the vowel types were distinct from each other in terms of F1 and F2, and the different prototype measures deviated from the median value of the entire vowel grid (F1, 605 mels; F2, 1240 mels) in the following order Pa > Pω > Pc.

In Finnish, the Kolmogorov-Smirnov test (more than 50 subjects) was used to test the normality of distributions. Only /e/, /ø/ and /æ/ were normally distributed for all prototypicality measures. On the basis of the normality tests, there were 8 normally distributed formant values out of the 16 formant values for the absolute prototypes, 14 out of 16 for the weighted prototypes, and 13 out of 16 for the centroids (see Table II).

In German, the Shapiro-Wilk test (fewer than 50 subjects) was used to test the normality of distributions. The vowels /y/, /e/, /ø/, /ɛ/ and /u/ were normally distributed for all prototypicality measures. On the basis of the normality tests, there were 13 normally distributed formant values out of the 16 formant values for the absolute prototypes, while the corresponding proportion was 14 out of 16 for the weighted prototypes and the centroids (see Table II).

The normality test results thus showed that the formant distributions of the Pω and Pc measures were normal in the vast majority of the cases in both languages. This was also the case for Pa in German, whereas only one half of the Pa formant values were normally distributed in Finnish. This may reflect the earlier finding for Finnish /i/ where three distinct absolute prototype classes were found (Aaltonen *et al.*, 1997; Eerola *et al.*, 2012).

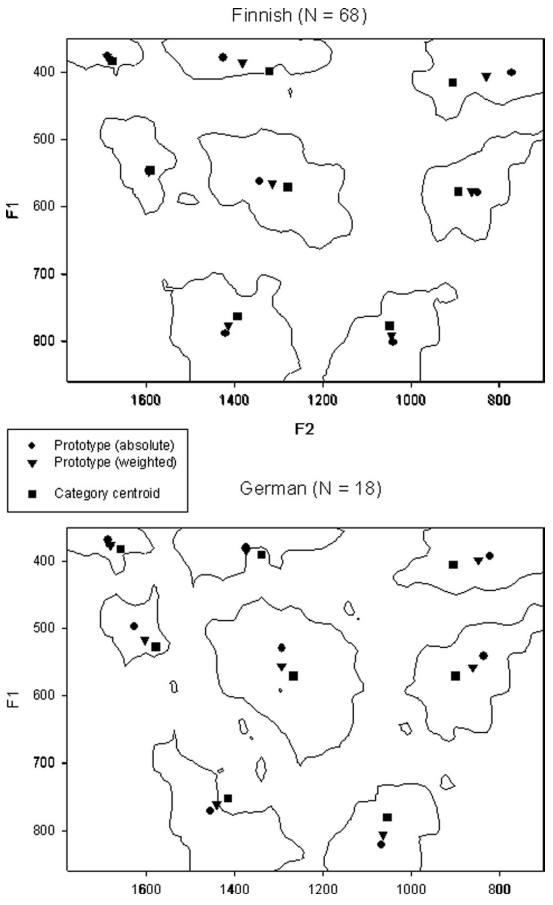


FIG. 3. The loci of the absolute prototypes, weighted prototypes, and category centroids in the F1–F2 formant space (mel scale) of the Finnish and German vowels. Areas of >90% identification consistency between listeners are shown by a solid line.

In order to test whether the absolute prototypes vary to a larger extent between the subjects than the weighted prototypes do in the F1-F2 space, the CV values of each prototype measure were examined. The mean CV values for different prototype measures [16 measures (8 categories * 2 formants) per language] were 3.1 (SD 1.4) % for absolute, 2.5 (SD 1.1) % for weighted, and 2.7 (SD 1.0) % for centroid type prototypes. Since the measures were non-normally distributed, the Bonferroni corrected Wilcoxon's signed ranks tests were used to test the differences of CVs. The measures were corrected using the Sokal-Braumann correction (Sokal and Braumann, 1980). The difference between the absolute and weighted prototype was significant ($p < 0.016$), and the difference between the absolute prototype and centroid was insignificant ($p > 0.016$). The difference between the weighted prototype and centroid was insignificant ($p > 0.016$). The results indicate that the absolute prototypes had significantly larger variation than the weighted prototypes.

TABLE III. Comparison of Finnish and German vowels expressed as the arithmetic differences of formants F1 and F2 (mel scale) of Finnish and German prototype measures (Pc, Pω, and Pa). The ratio of the Finnish - German difference to the German prototype is given in the parentheses (%). Significant differences are denoted by *.

	Pc		Pω		Pa	
	F1	F2	F1	F2	F1	F2
/i/	−1 (0.3)	−18 (1.1)	0 (0)	−8 (0.5)	−4 (1.1)	−6 (0.4)
/y/	−8 (2)*	−17 (1.3)	−2 (0.5)	−8 (0.6)	0 (0)	−51 (3.6)*
/e/	−19 (5.5)*	−9 (0.6)	−30 (5.4)*	9 (0.6)	50 (9.1)*	−35 (2.1)*
/ø/	0 (0)	12 (0.9)	9 (1.6)	21 (1.6)*	31 (5.8)*	52 (4.0)*
/ɛ-æ/	−10 (1.3)	21 (1.5)*	−16 (2.1)*	26 (1.9)*	−16 (2.1)*	35 (2.5)*
/ɑ/	3 (0.4)	5 (0.5)	14 (1.7)*	19 (1.8)*	21 (2.6)*	27 (2.5)*
/o/	−10 (2.5)	−2 (2)	8 (2.2)*	−18 (2.1)	−8 (2.0)*	51 (6.6)
/u/	−7 (1.2)*	5 (0.6)	−18 (3.2)*	−3 (0.3)	−37 (6.4)*	−14 (1.7)*

C. Differences between languages on the basis of prototype measures

The differences between the two languages were examined (Table III). For normally distributed prototype formant frequencies, independent sample *t*-tests were used. For non-normally distributed comparisons, the Mann-Whitley test for independent samples was used.

There were several significant differences in the Pc, Pω, and Pa measures of vowel categories between Finnish and German (Table III). However, such differences that were over DL (found only for Pa measure), and hence hearable, were found in /y/ (F2 51 mels), /e/ (F1 50 mels, F2 35 mels), and /ø/ (F1 31 mels, F2 52 mels). These observed differences in the vowel categories were expected on basis of production data. However, for the open front vowels /ɛ/ and /æ/ that were essentially similar in both languages, the differences were smaller than expected, a finding that contradicts the earlier results on production (e.g., Kuronen, 2000; Sendlmeier and Seebode, 2006).

D. Differences between production and perception

The comparisons between the (three) different prototype measures and the F1 and F2 values of produced Finnish (Kuronen, 2000) and German (Sendlmeier and Seebode, 2006) vowels showed differences across the languages (Table IV), ranging between 19 mels for Pa of /ø/ and 236 mels for Pa of /o/ in Finnish, and between 38 mels for Pw of /y/ and 174 mels for Pa of /æ/ in German. In the study of Eerola and Savela (2011), the same 14 subjects participated both in a listening and a production experiment using four vowels (/i/, /e/, /y/, /ø/). In that study, the calculated mean Euclidean distance in the F1–F2 space between the produced Finnish vowels and their weighted category prototypes was 113 mels for short and 116 mels for long vowels. For the same vowels (/i/, /e/, /y/, /ø/) as were used in Eerola and Savela (2011), the distances between the produced vowels and the Turku Vowel Test weighted prototypes were, on an average, 64 mels. Correspondingly, the Euclidean distances between the produced vowels (Kuronen, 2000) and the weighted prototypes for those four vowels (/i/, /e/, /y/, /ø/) in

TABLE IV. The Euclidean distances between produced vowels (Kuronen, 2000; Sendlmeier and Seebode, 2006) and the different types of prototypes (Pc=centroid, Pω=weighted prototype, and Pa=absolute prototype) in Finnish and German. The results of produced vowels were transformed to the mel scale by using the formula presented in Lindsay and Norman (1977).

	Produced		Euclidean distance		
Finnish	F1	F2	Pc(F1,F2)	Pω(F1,F2)	Pa(F1,F2)
/i/	431	1562	125	138	143
/y/	439	1344	47	67	101
/e/	585	1435	161	164	163
/ø/	576	1332	52	20	18
/æ/	685	1273	143	169	180
/ɑ/	700	1222	189	199	207
/o/	598	1085	191	222	236
/u/	460	824	94	54	79
Mean			125	129	140
German					
/i/	367	1591	70	90	96
/y/	413	1399	64	38	42
/e/	463	1551	70	76	84
/ø/	488	1282	83	69	43
/ɛ/	598	1462	161	164	174
/ɑ/	791	1164	110	101	101
/o/	500	892	70	66	70
/u/	421	901	16	58	82
Mean	81	82	86		
Grand Average both languages	102	106	114		

the present experiment were, on an average, 97 mels. For German vowels, the mean Euclidean distance between produced vowels (Sendlmeier and Seebode, 2006) and the weighted prototypes of this study was 68 mels. However, the observed differences may be explained by the fact that Kuronen (2000) used carrier sentences, whereas Eerola and Savela (2011) used isolated words. In an imitation study by Repp and Williams (1985), the differences were 20–50 mels for /i/ and /y/.

IV. GENERAL DISCUSSION

The present study examines the prototypicality in the context of a large vowel grid in two languages that have a similar vowel system exhibiting the same number of vowel categories and basic distinctions. The adjacent stimuli of the grid differed by 30 mel in F1 and 40 mel in F2, which corresponds roughly with the DL of frequency. This made it possible to accurately compare the different prototypicality measures and their differences between the languages. The results gave four major findings.

First, the inter-individual variation (in terms of CV) from the mean value of each category was within the DL of the F1 and F2 frequencies. Of the different prototype measures, Pω had the smallest variation, whereas Pc and the Pa showed larger individual variation. The within-category formant distributions of the individual weighted prototypes and centroids were normal for most vowel types in both languages, whereas the distributions of absolute prototypes differed; in German they were normal in most cases, but in Finnish they were normal in only one half of the cases. Thus, for normally distributed prototypes, especially for Pω, 68.3% of subjects ($\mu \pm \sigma$) evaluated the best category representatives from a subset of stimuli that lie within the limits of just noticeable frequency difference from each other in the F1–F2 space. Differences in variation were also found between vowel types: In both languages, the largest differences were in /e/ and /ø/ for F1, and in /u/ for F2.

All prototype measures showed less variation than has been reported in earlier literature. This might be related to the larger grid used in the present study. For example, in the study by Aaltonen *et al.* (1997), the larger individual variation (CV 8%) may have been related to differences in the subjects' strategies in identification experiment leading to hyper-articulation-like behavior in the identifications. In that study, only the F2 values were varied in the identification test, and therefore the possibility that subjects did not recognize familiar sounds was greater than in the present study, in which the F1 was varied as well. This may have resulted to a smaller variation between subjects in the present experiment. The weighted prototypes had significantly smaller variation than the absolute prototypes in the experiment of Aaltonen and coworkers.

Second, the absolute prototypes were, in general, more peripheric than the weighted prototypes. Figure 3 illustrates that weighting moves the prototype to a more peripheral location from the centroid. In peripheral vowels, the absolute prototypes are in the most peripheral positions of the vowel space. Interestingly, in German, the absolute prototype of /e/ has a lower F1 in comparison to the other measures, whereas for the Finnish /e/, all these measures were similar.

The present study gave some support to the theory on adaptive dispersion effect in perception by Johnson (2000), since there was a main effect indicating that the absolute prototypes were the most peripherical. The weighted prototypes also differed from the category centroids. This finding suggests that the "gravity center" of a category differs from the arithmetic mean of the category. In German, the absolute prototypes had lower F1 values in mid-vowels as compared to the other prototype measures, whereas in Finnish no such shift was observed. This result may be related to the tense-lax relationship in German that decreases F1, a phenomenon that does not exist in Finnish.

Third, there were some minor differences in the vowel systems of Finnish and German as indicated by the prototype measures: The absolute prototypes showed the largest differences between the languages in /e/, /ø/ and /u/. This is in line with the earlier investigations on produced vowels in Finnish and German. However, in general, the vowel systems of the two languages were similar, as suggested by dispersion theories (e.g., Becker-Kristal, 2010). In terms of weighted prototypes, the Euclidean distances of corresponding categories in the two languages varied between 8–30 mels. This result indicates that the acoustical differences of the vowel systems in these two different languages is strikingly small. The largest differences were observed in the non-closed front vowels, while the other types of vowels showed minor differences, as expected on the basis of their production.

Fourthly, the differences observed between the various prototype measures and produced vowels (obtained in some

earlier studies) were similar to earlier findings, with the mean difference being approximately 110 mels across all categories in both languages (131 for Finnish and 83 for German). In the study by Eerola and Savela (2011), the difference was approximately 110 mels for four vowels (/i/, /e/, /y/, /ø/), whereas in the present study, the difference for the same four vowels was 99 mels. Nevertheless, there was a significant difference between the studies in the formant range of F2 of the used stimuli (1780 vs 1830 mels, respectively), which makes direct comparisons difficult. There were differences between vowel categories in terms of differences between production data and perception data. The non-peripheral vowels /y/ and /ø/ had the smallest differences between production and perception, whereas for other vowels, adaptive dispersion in terms of production-perception was found (e.g., Johnson *et al*., 1993). In contrast to Eerola and Savela (2011), the produced vowels were always less peripheral than the perceived stimuli. The smallest differences were found in /ø/. In that study, the listener's own production was compared with his/her goodness ratings of synthetic vowels (male voice). The absolute distance was similar, although the direction of the difference depended on the subject. The differences between vowel categories may be related to the vowel type (rounded/unrounded, back/front, open/close). In Johnson *et al*. (1993) data, the largest difference between spoken vowels and prototypes (although the method was different) was found in /u/.

In this study, Finnish and German listeners identified and rated a large number of synthetic vowels according to their native vowel systems. Based on the results, these two languages with the same number of vowels, openness levels, and secondary rounded vowels appeared strikingly similar, especially when the weighted prototype is used. When the absolute prototype measure is used, the prototypicality appears to be more language specific (e.g., Strange *et al*., 2007). This finding may reflect, e.g., the dispersion principles (e.g., Becker-Kristal, 2010), some general phonetic features (Strange *et al*., 2007), or color of the timbre (Savela, 2009). However, the results obtained for German and Finnish are not necessarily applicable to languages in general.

V. CONCLUSION

In general, there were differences between the different vowel prototype measures in terms of their location (peripherality) and normality of distribution in the F1–F2 space. In the case of Finnish and German, the absolute prototype method seems to be more sensitive to language differences, but it suffers from larger individual variation in the loci of vowel prototypes, and to some extent, from non-normal density distribution within the category. This was also found in the earlier studies concerning absolute prototypes (e.g., Aaltonen *et al*., 1997). What actually causes this type of distribution of absolute prototypes is not evident on the basis of this study. The weighted prototype approach provides a new method for defining the loci of perceptual sound spaces, even with a smaller grid of stimuli than the one used in this study (Eerola and Savela, 2011). Furthermore, it seems to provide a robust way for approximating an area within a category where individual results differ from the group mean to a lesser extent than or equally to the difference limens of F1 and F2 frequencies. This can be interpreted to show that the formation of individual prototypes is similar among the speakers of a particular language, and even between languages that have similar sound systems. Further studies are needed to investigate the discrimination capability around the weighted prototypes for the purposes of comparing the results to those obtained by using absolute prototypes. Another potential topic for future research is the comparison of the different prototypicality measures in other languages in addition to Finnish and German.

Aaltonen, O., Eerola, O., Hellström, Å., Uusipaikka, E., and Lang, H. A. (**1997**). "Perceptual magnet effect in the light of behavioral and psychophysiological data," J. Acoust. Soc. Am. **101**, 1090–1103.

Aaltonen, O., and Suonpää, J. (**1983**). "Computerized two-dimensional model for Finnish vowel identifications," Audiology **22**, 410–415.

Becker, T. (**1998**). *Das Vokalsystem der deutschen Standardsprache (Vowel system of Standard German)* (Peter Lang, Frankfurt am Main, German), pp. 1–200.

Becker-Kristal, R. (**2010**). "Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus," Ph.D. thesis (Department of Linguistics, University of California, Los Angeles, CA). Available at http://www.linguistics.ucla.edu/faciliti/research/research.html#Dissertations (Last viewed 9/4/2012).

Benders, T., and Boersma, P. (**2009**). "Comparing methods to find a best exemplar in a multidimensional space," in *Proceedings of Interspeech 2009*, September 6–10, Brighton, UK, pp. 396–399.

Bennett, D. C. (**1968**). "Spectral form and duration as cues in the recognition of English and German vowels," Lang. Speech **11**, 65–85.

Boersma, P. (**2006**). "Prototypicality judgments as inverted perception," in *Gradedness in Grammar*, edited by G. Fanselow, C. Féry, M. Schlesewsky, and R. Vogel (Oxford University Press, Oxford, UK), pp. 167–184.

Bohn, O.-S., and Flege, J. (**1992**). "The production of new and similar vowels by adult German learners of English," Stud. Second Lang. Acquis. **14**, 131–158.

Eerola, O., Laaksonen, J., Savela, J., and Aaltonen, O. (**2003**). "Perception and production of the short and long Finnish [i] vowels: Individuals seem to have different perceptual and articulatory templates," in *Proceedings of the 15th International Congress of Phonetics Sciences*, Barcelona, Spain, pp. 989–992.

Eerola, O., and Savela, J. (**2011**). "Differences in productions of Finnish front vowels and weighted prototypes vary in the F1-F2 space," in *Proceedings of 17th International Conference of Phonetics 2011*, Hong Kong, pp. 631–634.

Eerola, O., Savela, J., Laaksonen, J., and Aaltonen, O. (**2012**). "The effect of duration on vowel categorization and perceptual prototypes in a quantity language," J. Phonetics **40**, 315–328.

Evans, B. G., and Iverson, P. (**2004**). "Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences," J. Acoust. Soc. Am. **115**(1), 352–361.

Evans, B. G., and Iverson, P. (**2007**). "Plasticity in vowel perception and production: A study of accent change in young adults," J. Acoust. Soc. Am. **121**(6), 3814–3826.

Fairbanks, G., and Grubb, P. (**1961**). "A psychophysical investigation of vowel formants," J. Speech Lang. Hearing Res. **4**, 203–219.

Flanagan, J. L. (**1955**). "A difference limen for vowel formant frequency," J. Acoust. Soc. Am. **27**, 613–617.

Gottfried, T. L., and Beddor, P. S. (**1988**). "Perception of temporal and spectral information in French Vowels," Language Speech **31**(1), 57–75.

Harinen, K., Aaltonen, O., Salo, E., Salonen, O., and Rinne, T. (**2013**). "Task-dependent activations of human auditory cortex to prototypical and nonprototypical vowels," Human Brain Mapp. **34**(6), 1272–1281.

Hawks, J. W. (**1994**). "Difference limens for formant patterns of vowel sounds," J. Acoust. Soc. Am. **95**, 1074–1084.

Heid, S., M.-B. Wesenick, and C. Draxler (**1995**). "Phonetic analysis of vowel segments in the PhonDat database of spoken German," in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, pp. 416–419.

Hose, B., Langer, G., and Schleich, H. (**1983**). "Linear phoneme boundaries for German synthetic two-formant vowels," Hearing Res. **9**, 13–25.
Iverson, P., and Evans, B. (**2003**). "A goodness optimization method for investigating phonetic categorization," in *Proceedings of the 15th Congress of Phonetic Sciences*, Barcelona, Spain, pp. 2217–2220.
Iverson, P., and Kuhl, P. K. (**2000**). "Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism?," Percept. Psychophys. **62**, 874–886.
Jessen, M., Marasek, K., Schneider, K., and Claßen, K. (**1995**). "Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German," in *Proceedings of the International Congress of Phonetic Sciences* (Stockholm University, Stockholm), pp. 428–431.
Johnson, K. (**2000**). "Adaptive dispersion in vowel perception," Phonetica **57**, 181–188.
Johnson, K., Flemming, E., and Wright, R. (**1993**). "The hyperspace effect: Phonetic targets are hyperarticulated," Language **69**, 505–528.
Karlsson, F. (**1983**). *Suomen kielen äänne- ja muotorakenne (Sound and Form Structures in Finnish)* (Werner Söderström Oy, Porvoo, Finland), pp. 1–444.
Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**(3), 971–995.
Kuhl, P. K. (**1991**). "Human adults and human infants show a 'Perceptual Magnet Effect' for the prototypes of speech categories, monkeys do not," Percept. Psychophys. **50**(2), 93–107.
Kuhl, P. K. (**1993**). "Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception," J. Phonetics **21**, 125–139.
Kuronen, M. (**2000**). "Vokaluttalets akustik i sverigesvenska, finlandssvenska och finska" ("The acoustic character af vowel pronunciation in Sweden-Swedish, Finland-Swedish and Finnish"), Ph.D. thesis, Studia Philologica Jyväskylensia, University of Jyväskylä, Jyväskylä, Finland, Vol. 49, pp. 1–234.
Lacerda, F. (**1995**). "The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory," in *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Stockholm University, Stockholm), Vol. 2, pp. 140–147.
Lindsay, P. H., and Norman, D. A. (**1977**). *Human Information Processing: An Introduction to Psychology*, 2nd ed. (Academic Press, New York), 777 pp.
Lively, S. E., and Pisoni, D. B. (**1997**). "On prototypes and phonetic categories: A critical magnet effect in speech perception," J. Exp. Psychol. Hum. Percept. Perform. **23**, 1665–1679.
Määttä, T. (**1983**). "Hur finskspråkiga uppfattar svenskans vokaler" ("Contrastive Studies in the Perception of the Vowel Sounds of Swedish by Speakers of Finnish"), Ph.D. thesis, University of Umeå, Acta Universitatis Umensis (Almqvist and Wiksell International, Stockholm, Sweden), Vol. 55, 211 pp.
Morais, J., and Kolinsky, R. (**1994**). "Perception and awareness in phonological processing: The case of the phoneme," Cognition **50**, 287–297.
Nábelek, A. K., Czyzewski, Z., and Crowley, H. J. (**1993**). "Vowel boundaries for steady-state and linear formant trajectories," J. Acoust. Soc. Am. **94**, 675–687.
Nosofsky, R. M. (**1988**). "Exemplar-based accounts of relations between classification, recognition, and typicality," J. Exp. Psychol. Learn. Mem. Cogn. **14**, 700–770.
Oglesbee, E., and De Jong, K. (**2007**). "Searching for best exemplars in multidimensional stimulus spaces," J. Acoust. Soc. Am. **122**(4), EL101–EL106.
Raimo, I., Savela, J., and Aaltonen, O. (**2002a**). "Turku vowel test," in *Fonetiikan Päivien Paperit*, Espoo, Finland (Helsinki University of Technology, Helsinki), pp. 45–52.
Raimo, I., Savela, J., Launonen, A., Kärki, T., Mattila, M., Uusipaikka, E., and Aaltonen, O. (**2002b**). "Multilingual vowel perception," in *Abstracts of Temporal Integration in the Perception of Speech*, April 8–10, Aix-en-Provence, France (Cambridge University Press, London, UK), pp. 86.
Rendell, L. A. (**1986**). "General Framework for *Induction* and a Study of Selective *Induction*," Mach. Learn. **1**, 177–226.
Repp, B. H., and Williams, D. R. (**1985**). "Categorical trends in vowel imitation: Preliminary observations from a replication experiment," Speech Comm. **4**, 105–120.
Rosch, E. (**1975**). "Cognitive reference points," Cognit. Psychol. **7**, 532–547.
Rosner, B. S., and Pickering, J. B. (**1994**). *Vowel Perception and Production* (Oxford University Press, Oxford, UK), 413 pp.
Samuel, A. G. (**1982**). "Phonetic prototypes," Percept. Psychophys. **31**, 307–314.
Savela, J. (**2009**). "Role of selected spectral attributes in the perception of synthetic vowels," Ph.D. thesis, University of Turku, TUCS Dissertations 119, pp. 1–92.
Savela, J., Pikkanen, O., Raimo, I., Uusipaikka, E., and Aaltonen, O. (**2005**). "Stability of vowel perception: results from the Turku vowel test," in *Fonetiikan päivät 2004 – The phonetics symposium 2004 Oulussa 27.–28.8.2004*, edited by. T. Seppänen, K. Suomi, and J. Toivanen (MediaTeam Oulu ja Suomen kielen, informaatiotutkimuksen ja logopedian laitos, Oulun yliopisto, Oulu: Oulun yliopistopaino), pp. 30–31.
Sendlmeier, W. F. (**1981**). "Der Einfluss von Qualität und Quantität auf die Perzeption betonter Vokale des Deutschen" ("The Influence of Quality and Quantity on the Perception of Stressed Vowels in German"), Phonetica **38**, 291–308.
Sendlmeier, W. F., and Seebode J. (**2006**). "Formantkarten des deutschen Vokalsystems" ("The formant charts of German vowel system"), TU Berlin, http://www.kw.tu-berlin.de/fileadmin/a01311100/ Formantkarten_des_deutschen_Vokalsystems_01.pdf (Last viewed 9/4/2011).
Sokal, R. R., and Braumann, C. A. (**1980**). "Significance tests for coefficients of variation and variability profiles," Syst. Zool. **29**, 50–66.
Strange, W., Levy, E. S., and Law, F. F. (**2009**). "Cross-language categorisation of French and German vowels by naive American listeners," J. Acoust. Soc. Am. **126**(3), 1461–1476.
Thyer, N., Hickson, L., and Dodd, B. (**2000**). "The perceptual magnet effect in Australian English vowels," Percept. Psychophys. **62**, 1–20.
Wiik, K. (**1965**). "Finnish and English vowels," Ph.D. thesis, University of Turku, Ann. Univ. Turkuensis, Ser. B 94, 1–192.
Willerman, R., and Kuhl, P. (**1996**). "Cross-language speech perception: Swedish, English, and Spanish speakers' perception of front rounded vowels," in *Proceedings of the ICSLP-1996*, pp. 442–445.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3333-4
ISSN 1459-2045