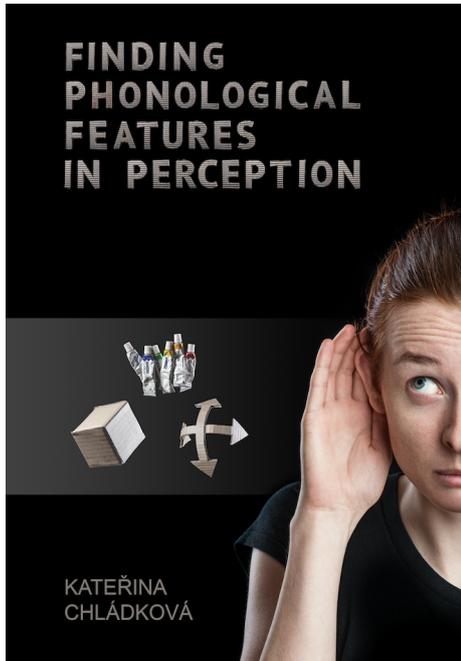


KATEŘINA CHLÁDKOVÁ

FINDING
PHONOLOGICAL FEATURES
IN PERCEPTION

Front cover



Back cover



ISBN: 978-94-6259-047-2

NUR: 616

Printed by: Ipskamp Drukkers

Layout: Typeset in L^AT_EX using the typographical look-and-feel
classicthesis developed by André Miede.

Cover photo: Tomáš Loutocký, <http://www.loutocky.com>

Copyright © 2014 Kateřina Chládková. All rights reserved.

No part of this publication may be reproduced without the prior written
permission of the author.

FINDING PHONOLOGICAL FEATURES IN PERCEPTION

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 7 maart 2014, te 12:00 uur

door

Kateřina Chládková

geboren te Šternberk, Tsjechoslowakije

PROMOTIECOMMISSIE

Promotor: prof. dr. P.P.G. Boersma

Co-promotores: dr. P.R. Escudero Neyra
dr. S.R. Hamann

Overige leden: prof. dr. J. Kingston
prof. dr. N.O. Schiller
prof. dr. A.E. Baker
dr. C.C. Levelt

Faculteit der Geesteswetenschappen

The research reported in this thesis was funded by the Netherlands Organization for Scientific Research (NWO) grant no. 277.70.008 awarded to Paul Boersma.

ACKNOWLEDGMENTS

I'm very grateful for having had Paul Boersma as my promotor. Paul has been a fabulous teacher and a devoted supervisor. Whenever I got stuck while designing an experiment, analysing data or running simulations, I just knocked on his door and he helped me solve the problem instantly. Paul, every single discussion with you made me push the limits of me as a scientist a step further. Thank you!

It's been a great experience working with my co-promotors, Paola Escudero and Silke Hamann. Paola, thank you for virtually bringing me to Amsterdam in 2007 and for encouraging me to do a master's at the UvA in the first place. Seeing your enthusiasm and determination has always been very inspiring and motivating. Silke, you introduced the sound-change perspective into my PhD project and thereby broadened my horizons quite a bit. Thank you for your supportive attitude (not only during our research trip to Kent), and for keeping reassuring me that I'm doing fine, which helped me finish this thesis (to my own surprise) rather peacefully.

I would like to thank Anne Baker, Claartje Levelt, John Kingston, and Niels Schiller for being in my doctoral committee.

This thesis has benefited from meetings and collaborations with colleagues from the ACLC and elsewhere. My experiments would not be running without Dirk Jan Vet. Dirk, you are a true tech-angel! Jan-Willem, Karin, and Margarita, thanks for making our office such a nice place to work at. A big thank you to Šárka Šimáčková and Václav Jonáš Podlipský who introduced me to the field of phonetics and phonology, and quickly transformed into colleagues and wonderful friends. Many thanks to Šárka for her comments on the English and the Czech summary. A thousand thanks to Mirjam de Jonge for translating the summary into Dutch. Mirjam and my sister Barbora will stand beside me during my defense - thank you girls!

I'm hugely indebted to the best photographer out there, Tomáš Loutocký, for designing and making the cover for this book.

I would have had an unimaginably harder time working on this thesis if it hadn't been for all my family and my dear friends, who have continuously been a great mental support and who put a big smile on my face every time I am with them. The most special thanks go to my mum Pavlína for making me who I am and for always being there for me.

David came with me to the Netherlands so that I could pursue my dream and not be alone. He has held my hand and has borne with me patiently ever since. This thesis is dedicated to him.

CONTENTS

1	INTRODUCTION	1
1.1	Phonological features and the linguist	1
1.2	Phonological features and the language user	3
1.3	Are phonological features innate or emergent?	6
1.4	Can perceptual patterns reveal feature structure?	8
1.5	Phonological features and the language learner	9
1.6	Summary	11
2	THE HUMAN LISTENER AS A PHONOLOGICAL FEATURE DETECTOR: THE PERCEPTUAL BASIS OF VOWEL HEIGHT	13
2.1	Introduction	14
2.2	Experiment 1	19
2.2.1	Method	20
2.2.2	Results and discussion	21
2.3	Experiment 2	24
2.3.1	Method	25
2.3.2	Results and discussion	27
2.4	General discussion	32
2.4.1	Main findings	32
2.4.2	The number of perceived categories	33
2.4.3	Differences in boundary location in two-peak listeners	34
2.4.4	Symmetric vs. asymmetric vowel systems	35
2.5	Conclusions	36
2.6	Acknowledgments	36
3	WHY SHE AND SHOE WON'T MERGE: REDEFINING PERCEPTUAL CUES FOR THE FRONT-BACK CONTRAST IN THE ENGLISH OF SOUTHERN ENGLAND	37
3.1	Introduction	38
3.2	Experiment 1	41
3.2.1	Method	41
3.2.2	Results	43
3.2.3	Discussion	46
3.3	Experiment 2	47
3.3.1	Method	47
3.3.2	Results	50
3.3.3	Discussion	53
3.4	General discussion and conclusions	54
3.5	Acknowledgements	56

4	PERCEPTUAL SENSITIVITY TO CHANGES IN VOWEL DURATION REVEALS THE STATUS OF THE PHONOLOGICAL LENGTH FEATURE	57
4.1	Duration in native and non-native vowel qualities	57
4.1.1	Introduction	57
4.1.2	Methods	62
4.1.3	Results	67
4.1.4	Discussion	71
4.1.5	Conclusions	76
4.2	MAAN is long but MAN is not short	77
4.2.1	Introduction	77
4.2.2	Methods	80
4.2.3	Results	82
4.2.4	Discussion	84
5	THE EMERGENCE OF PHONOLOGICAL FEATURES IN AN ARTIFICIAL NEURAL NETWORK	87
5.1	Computational models of phonology	88
5.2	A neural network version of the BiPhon model	89
5.3	First model: separate auditory dimensions for F1 and F2	92
5.4	Second model: a single auditory dimension for frequency	96
5.5	Third model: adding the knowledge of allomorphy	98
5.6	Discussion and conclusion	102
6	GENERAL DISCUSSION AND CONCLUSIONS	107
6.1	Phonological features are at the interface with phonetics	107
6.2	Perception reveals feature representations	109
6.3	Features are acquired with the help of morphophonology but have direct correlates in phonetics	110
6.4	Conclusions	111
	BIBLIOGRAPHY	113
	SUMMARY	129
	SAMENVATTING	133
	SHRNUTÍ	137
	CURRICULUM VITAE	141

LIST OF FIGURES

Figure 1.1	Three possible scenarios of the mapping between phonetics and features.	5
Figure 1.2	The mapping between phonetics and phonology (i.e., phonemes and features).	7
Figure 1.3	The model of Bidirectional Phonetics and Phonology (BiPhon, Boersma, 2009).	10
Figure 2.1	The two competing models of vowel perception.	15
Figure 2.2	Phonetic and phonological organization of a typical 5-vowel system.	18
Figure 2.3	Results of the vowel identification task pooled across the three F ₃ values.	22
Figure 2.4	The 130 stimulus pairs along each of the three continua between 6.93 Erb and 12.86 Erb (280 Hz and 725 Hz).	26
Figure 2.5	Raw and smoothed data of one listener.	28
Figure 2.6	Smoothed data, model, and smoothed model for the listener from Figure 2.5.	30
Figure 2.7	Smoothed average best model fits of one-peak and two-peak listeners.	31
Figure 2.8	Perceptual vowel spaces in Czech(-like) listeners who map F ₁ to vowel height (and F ₂ to vowel backness).	34
Figure 3.1	F ₁ -F ₂ plot of /i/ and /u/ produced by male speakers of different ages.	38
Figure 3.2	Re-association of a phonological contrast with a new phonetic cue.	39
Figure 3.3	Illustration of stimuli from Experiment 1.	42
Figure 3.4	Experiment 1: perceptual /i/-/u/ boundaries on the F ₂ dimension.	45
Figure 3.5	Experiment 2: the sampling of the F ₁ -F ₂ stimulus space.	48
Figure 3.6	Results of Experiment 2: response categories chosen for each stimulus.	50
Figure 3.7	Experiment 2: perceptual front-back phoneme boundaries in the F ₁ -F ₂ space.	53
Figure 4.1	Graphical abstract.	58
Figure 4.2	F ₁ and F ₂ plot of Dutch, Czech and Spanish vowels and of the two vowels that served as stimuli.	61

Figure 4.3	Grand-average difference waveforms at FCz in Dutch, Czech and Spanish listeners.	69
Figure 4.4	Grand-average standard and deviant waveforms at FCz.	70
Figure 4.5	Mean MMN amplitude per language and vowel quality.	71
Figure 4.6	Model of Czech, Spanish, and Dutch perception based on our findings.	73
Figure 4.7	F1 and F2 plot of Randstad Dutch vowels and of the two vowels that served as stimuli.	80
Figure 4.8	Grand-average deviant, standard, and difference waveforms at Fz, and MMN scalp distributions.	83
Figure 5.1	An example of a two-layer BiPhon neural network.	90
Figure 5.2	Architecture of the network modeled in Section 5.3 (referred to as the 2-2-1 network).	93
Figure 5.3	A 2-2-1 learner's production of the acquired 5-vowel language.	95
Figure 5.4	The results of 10 simulations of the 2-2-1 learners.	97
Figure 5.5	Architecture of the network modeled in Section 5.4 (referred to as the 1-1-1 network).	98
Figure 5.6	A 1-1-1 learner's production of the acquired 5-vowel language.	99
Figure 5.7	The results of 10 simulations of the 1-1-1 learners.	100
Figure 5.8	Architecture of the network modeled in Section 5.5 (referred to as the 1-1-2 network).	101
Figure 5.9	A 1-1-2 learner's production of the acquired 5-vowel language.	103
Figure 5.10	The results of 10 simulations of the 1-1-2 learners.	104
Figure 6.1	The mapping between phonetics and phonology based on the present findings.	108

LIST OF TABLES

Table 2.1	Czech monophthongal vowel phonemes and their height and backness features.	19
Table 2.2	Within-subject labeling uniformity in the three regions of the vowel space.	23
Table 2.3	Number of listeners with 0, 1, 2, and 3 peaks on each continuum.	30
Table 4.1	MMN amplitude averaged across 9 sites per language, vowel quality, and duration type.	68
Table 4.2	Dutch listeners: MMN amplitude averaged across 9 sites per vowel quality and duration type.	82
Table 5.1	Mean F ₁ and F ₂ values corresponding to the five meanings of our toy language.	93

LIST OF ABBREVIATIONS

ANOVA	analysis of variance
BiPhon	model for Bidirectional Phonetics and Phonology
C	consonant
EEG	electroencephalography
ERP	event-related potential
F ₀	fundamental frequency
F ₁	first formant
F ₂	second formant
F ₃	third formant
IPA	International Phonetic Association
ISI	inter-stimulus interval
MANOVA	multivariate analysis of variance
MMN	mismatch negativity
NN	neural network
OT	Optimality Theory
RM	repeated measures
SESE	the variety of Standard English spoken in Southern England
V	vowel
VOT	voice onset time

INTRODUCTION

The research presented in this thesis aims to find phonological features in perception. The search for features is approached from several angles. First, we aim to find out whether phonological features are the categories through which adult listeners process the speech signal (Chapter 2 and Chapter 3). Second, we examine listeners' perceptual patterns in order to determine which phonological features are part of their grammar (Chapter 4). Third, using simulations of perceptually driven learning, we test whether a virtual infant learns to represent the sounds of her language in terms of phonological features (Chapter 5).

In the present chapter, I first introduce the concept of phonological features, and further discuss previous literature that questioned their phonetic grounding and learnability. At the end of each section, I briefly define the questions addressed in this thesis. Finally, I describe the theoretical framework within which the present research is set.

1.1 PHONOLOGICAL FEATURES AND THE LINGUIST

The Dutch words *duin*, *tuin*, and *puin* differ in their meaning: they refer to 'dune', 'garden', and 'debris', respectively. The three Dutch words sound identical except for the consonant in their initial position: they start with [d], [t], and [p], respectively. Given the meaning and the sound contrast, one can argue that /d/, /t/ and /p/ are phonemes of Dutch. Bases for this argument can be traced back to Trubetzkoy (1939: 41) who defined the phoneme as the minimal contrastive unit of linguistic analysis. Besides *defining* them, Trubetzkoy proposed classification of phonemes in terms of distinctive oppositions. In that respect, /p/ differs from /t/ and /d/ in terms of localization (/p/ being labial, /d/ and /t/ apical); at the same time /d/ differs from /t/ and /p/ in terms of voicing (/d/ being voiced, /t/ and /p/ voiceless) (Trubetzkoy, 1939: 122–145).

In line with Trubetzkoy's distinctive oppositions, Jakobson et al. (1952: 2) inferred that in a word triplet similar to our Dutch *duin* - *tuin* - *puin* example, only *duin* vs. *tuin* and *tuin* vs. *puin* represent a minimal distinction, while *duin* vs. *puin* represent a more complex one. To formalize the difference between minimal and non-minimal distinctions, Jakobson and colleagues defined the smallest and ultimate distinctive unit of a language: the distinctive feature. Thus, *duin* vs. *tuin* and *tuin* vs. *puin* are

each contrasted by a single feature (namely, voice and gravity¹, respectively), while *duin* vs. *puin* by two features (i.e., both voice and gravity). Phonemes are then seen as concurrent combinations of features (Jakobson et al., 1952: 3, 26-27).

Similarly, in Chomsky and Halle's (1968) analysis, phonological and phonetic representations consist of matrices, in which the rows stand for individual features and the columns stand for units or segments. Thus, each speech segment is represented as a bundle of features and their values. Chomsky and Halle doubted the existence of the phoneme and seem to have considered the feature to be the only type of phonetic and phonological representation needed in the grammar (Chomsky and Halle, 1968: 390, 11).

Later phonological frameworks further refined the theory of phonological features. For instance, in Autosegmental Phonology (Goldsmith, 1976), features no longer occur in segment-sized bundles: each feature has its own tier and is (with respect to timing) relatively independent of the features on other tiers. Following up on that, Feature Geometry (Clements, 1985) posits that features are arranged in an elaborate tree-like structure and that there are thus specific hierarchical dependencies amongst the features. Within Government Phonology, or Element Theory, (Harris, 1990; Kaye et al., 1985) the ultimate unit of phonological analysis is not a feature but an 'element'. Interestingly, at least in earlier versions of the theory, the element is entirely interpretable as a matrix of features (Kaye et al., 1985). In short, since 1950's phonological features have been abundantly employed in descriptions of the world's phonological systems.

Distinctive features may indeed appear to be a particularly convenient tool for cross-linguistic analyses of sound patterns, given that they have been named after observable phonetic properties. By naming them as such, phonological theories implied that features have bases in phonetics, i.e. in the sound (e.g. Jakobson et al., 1952²) or in the articulations (e.g. Chomsky and Halle, 1968). For instance, recall that the Dutch phonemes /t/ and /d/ are phonologically differentiated by the [voice] feature (Booij, 1995: 21). This contrast is manifested acoustically by an absence versus presence of a "voice bar along the base line of the spectrogram", and articulatorily by an absence versus presence of "concomitant periodic vibrations of the vocal bands" (Jakobson et al., 1952: 26). Thus, as far as the phonologist's view is concerned, phonological features seem to have a reasonable grounding in phonetics.

¹ According to Jakobson et al., grave consonants are characterized by a lowered second formant (F₂) in an adjacent vowel, while acute consonants by a raised F₂. The gravity feature was later replaced by the coronality feature (Chomsky and Halle, 1968).

² Jakobson et al. attempted to further distinguish auditory and perceptual bases for features, although their perceptual definitions were rather sparse.

However, the fact that the linguist sees a parallel between the phonological feature and phonetic reality does not necessarily imply that the language user sees the same parallel. As Ladefoged (1980) pointed out, one should investigate whether phonological features exist at all as mental representations in the grammars of language users:

“[...] if we go on using the linguistically well-known feature sets which have been found very useful in phonological descriptions, we must do so with the realization that these feature sets – mine, Chomsky & Halle’s, or anyone else’s – have in no way been proved to be the mental representations used by people when speaking or listening to any language. [...] if they are mental representations, then I would like to know what they are mental representations of.”

(Ladefoged, 1980: 496)

In summary, the phonological feature as the ultimate distinctive category plays a crucial role in theoretical grammars. Given that features have correspondents in phonetic dimensions, the question arises whether language users form feature-like linguistic categories on those phonetic dimensions. In the next section, I review the literature that questioned the auditory and articulatory bases of distinctive features.

1.2 PHONOLOGICAL FEATURES AND THE LANGUAGE USER

Soon after phonological features were defined, psycholinguists began to ask whether features are indeed the speech categories that speakers and listeners use when producing and perceiving speech. For instance, Miller and Nicely (1955) conducted a consonant identification experiment with various degrees of acoustic masking applied to the stimuli. The authors argued that identification errors obtained in their experiment could be attributed to misperceptions of the individual features that the consonants were composed of. Miller and Nicely (1955) therefore proposed that speech is more likely to be perceived through a system with multiple independent channels each of which detects a specific feature, than through a single complex channel that would integrate all acoustic information into a single percept. Studdert-Kennedy and Shankweiler (1970) reported a dichotic listening experiment, in which participants were asked to identify two different plosive consonants that were simultaneously presented to different ears. The consonants were more likely to be correctly identified when they shared a feature, e.g. such as /p/ and /t/, than when they did not, e.g. /b/ and /t/. Furthermore, misidentifications occurred more often in a single feature (e.g. misidentification of either place or voicing) than in both features. The results were interpreted as evidence for separate extraction of distinctive features during

speech perception. Eimas and Corbit (1973) aimed to find out whether humans possess separate detectors for the features voiced and voiceless (which are in English associated with short and long VOT values respectively). Eimas and Corbit reasoned that an extensive exposure to a long-VOT stimulus would cause fatigue of the voiceless-feature detector which would in turn result in greater sensitivity of the voiced-feature detector (for ambiguous stimuli), and vice versa. The authors thus tested whether listeners shift their voice/voiceless category boundary towards one end of the VOT continuum if they are repeatedly presented with a stimulus from that end of the continuum. As predicted, listeners shifted their voiced/voiceless boundary in the expected direction. More interestingly, the boundary shift was generalized across consonantal places that were not presented during the adaptation period. Eimas and Corbit (1973) thus concluded that humans are equipped with innate feature detectors: one for short VOT and one for long VOT. An illustration of feature detectors is presented in Figure 1.1A: the figure shows that features are linked directly to the acoustic signal and that there is a separate feature detector for each phonetic dimension.

Diehl (1981) criticized the three studies reviewed above (and many others), claiming that their findings did not present unequivocal evidence for feature detectors in humans. Diehl disputed the view that feature detectors yield a phonological feature as a direct output of the signal and argued that such feature detectors would make all fine-grained acoustic information unavailable for later stages of perception. For instance, perception of the voice feature through a single detector (e.g. for VOT) would often fail as there are other acoustic and contextual cues that contribute to voicing contrasts. In that respect, Lisker and Abramson (1964) demonstrated that in American English the voice feature can have multiple acoustic correlates. A scenario in which several phonetic dimensions are used to signal a single feature is shown in Figure 1.1B. Given the existence of multiple phonetic correlates for features, Diehl (1981) suggested that, instead of being detected at the very initial stage of perception, features might be decided on at later stages of processing when all the information from the ‘neural’ spectrogram as well as contextual cues are directly available.

Potentially, features may not be detected directly from the raw acoustic signal but from some kind of a perceptual transform of the acoustics (e.g. Diehl’s ‘neural’ spectrogram). In that respect, Kingston and Diehl (1994) proposed that when implementing a phonological feature contrast, articulations are controlled in such a way that their acoustic effects mutually enhance each other. Kingston and Diehl (1995) defined such a collection of mutually enhancing acoustic properties, i.e. the stage between phonological feature representations and the raw acoustic signal, as the ‘intermediate perceptual property’ (but see Nearey, 1995 for counterar-

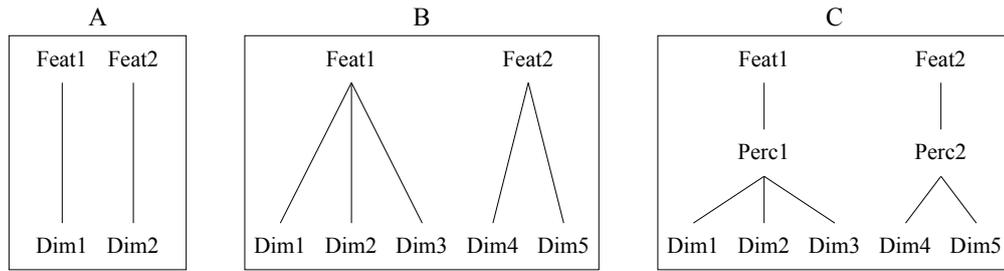


Figure 1.1: Three possible scenarios of the mapping between phonetics and features. (A) depicts a one-to-one mapping between phonetic dimensions (“Dim”) and features (“Feat”), while (B) illustrates features with multiple phonetic correlates. In (C), several phonetic dimensions are first integrated into perceptual transforms (“Perc”), which are then mapped onto phonological features.

guments). Kingston et al. (2008) proposed that the integration of acoustic dimensions is a result of a general auditory processing mechanism and is not due to listeners’ experience with these acoustic dimensions in speech. From this it follows that if there is an intermediate stage in speech processing that perceptually integrates acoustic dimensions, such a stage is not linguistic/phonological. The lowest-level linguistic/phonological representations, onto which the non-linguistic perceptual transforms are mapped, could then be phonemes or features (the latter of which was assumed by Kingston et al., 2008). Figure 1.1C illustrates detection of features from perceptual transforms of the acoustic signal.

The studies summarized above indicate that phonological features are the lowest-level linguistic representations onto which listeners map the phonetic signal, either directly or via (non-linguistic) integrated intermediate percepts. Note also that the work reviewed in the preceding paragraph advocates an auditory basis for features. Other lines of research claimed that the phonetic basis for features lies in articulatory gestures (Fowler, 1986; Liberman and Mattingly, 1985), or in the interplay between articulations and their auditory effects (Stevens, 1989). Since the experiments reported in this thesis assume an auditory-based model of speech perception (Boersma, 1997, 2009), we leave the details of articulatory-based theories outside the present review.

In sum, the above literature review suggests that over the past decades the central question relating to phonetic bases of features has shifted from “Do phonological features have direct phonetic correlates in the language users’ grammar?” (e.g. Diehl, 1981; Eimas and Corbit, 1973; Miller and Nicely, 1955; Studdert-Kennedy and Shankweiler, 1970) to “What is the nature of the features’ phonetic correlates?” As for the latter, it has been asked whether speakers link phonological features to articulatory gestures, to auditory properties of speech sounds, or to both (an articulatory basis has been advocated by e.g. Liberman and Mattingly,

1985, or Fowler, 1986; an auditory basis by e.g. Boersma, 1998; Hamann, 2003; Kingston et al., 2008; Nearey, 1995; and both types of bases by e.g. Lindau and Ladefoged, 1986, or Stevens, 1989). Also, it has been debated whether the mapping between features and phonetic dimensions is primarily one-to-one or many-to-many (a one-to-one mapping has been proposed by e.g. Stevens and Blumstein, 1981; many-to-many by e.g. Kingston and Diehl, 1995, or Kingston et al., 2008).

Evidently, in search for the specific nature of features' phonetic correlates, recent literature mostly assumes that the phonetics-to-feature mapping is a direct one (as noted by Hamann, 2011), although only some researchers formulate such an assumption explicitly (e.g. Hamann, 2011: 158–159; but see Escudero, 2005: 71–76 who claimed that phonetics is mapped onto features in infants but not in adults). Nevertheless, the prevailing surmise that phonological features lie at the interface with phonetics has not yet been verified empirically.

Regarding experimental work on the phonetics-phonology interface, Nearey (1990) demonstrated that listeners perceive the speech signal in units no larger than a segment. Since segment-sized phonological representations are phonemes (and allophones), Nearey's results can be interpreted as evidence for a mapping between acoustic signal and phonemes (as in Figure 1.2A). As Nearey pointed out, it is unclear whether speech perception employs the feature as a level of representation intermediate between the signal and the phoneme (as in Figure 1.2B), or whether the feature is only a more abstract representation which is not used in real-time phonetic perception (as in Figure 1.2C). See also Figure 1.2D illustrating yet another perception scenario, in which the phonetics is mapped onto both phonemes and features.

To find out whether listeners are tuned to features, we carried out the experiments reported in Chapter 2. Specifically, we assessed whether listeners map the F1 dimension directly onto phonological height categories or onto unanalyzed segmental phonemes. If listeners map the F1 dimension directly to the height feature, they should perceptually categorize any vowel stimulus in terms of height. Therefore, we first determined a vowel region in which our listeners do not reliably identify any phonemes, and then tested whether they perceive stimuli from this region in terms of their native height categories.

1.3 ARE PHONOLOGICAL FEATURES INNATE OR EMERGENT?

Besides addressing the question of whether phonetics is mapped onto features directly, the research reported in this thesis aims to investigate whether the mapping between phonetics and phonological features is inherent to all speakers of all languages (i.e., innate and universal) or

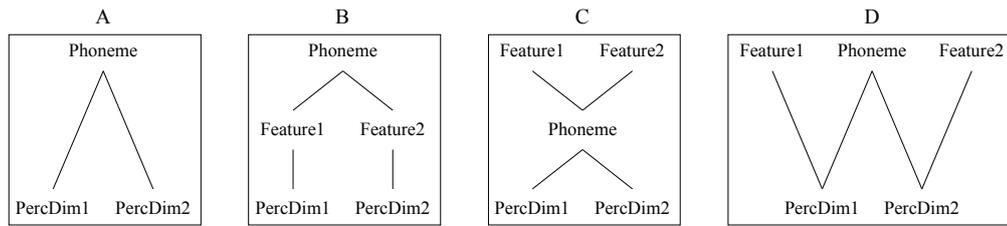


Figure 1.2: The mapping between phonetics (or, perceptual transforms of phonetics, “PercDim”) and phonology (“Phoneme” and “Feature”). (A) Phonetics is mapped onto phonemes. (B) Phonetics is mapped onto features, and features are then integrated into phonemes. (C) Phonetics is mapped onto phonemes, and phonemes are then analyzed into features. (D) Phonetics is mapped onto both features and phonemes.

acquired during exposure to one’s native language (i.e., emergent and arbitrary).

Originally, features were viewed as innate universals (Chomsky and Halle, 1968; Jakobson et al., 1952). According to Chomsky and Halle (1968: 297), for instance, the total set of features is equal to the total set of independently controllable articulatory gestures. That is, the correspondence between phonological features and phonetic dimensions is strictly one-to-one and innate (as in Figure 1.1A). The innateness view was adopted by the early studies that experimentally tested the phonetic grounding of features (e.g. Eimas and Corbit, 1973; Miller and Nicely, 1955). Features are, by definition, innate in theories that attribute them to anatomical and physiological properties of the human articulatory and auditory system (e.g. Stevens, 1989). Similarly, Stevens and Blumstein (1981) argued that the mechanism for discovering features from the acoustic signal is innate. The authors further proposed that in addition to the innate ‘primary property detectors’, speakers can rely on secondary (i.e. enhancing) acoustic cues whenever the primary cues to that feature are unavailable. It thus appears that in some innatist views (e.g. Stevens and Blumstein, 1981) the mapping between features and phonetics does not need to be a strictly one-to-one mapping.

A move away from the innatist view can be observed in studies that examined cross-linguistic differences in phonetic correlates for features. In that respect, Lisker and Abramson (1964), and later also Lindau and Ladefoged (1986), showed that a single feature can be cued by several phonetic dimensions (and vice versa) and that these mappings may differ across languages. Likewise, Kohler (1981) pointed out the between-language differences in phonetic correlates for the [voice] feature, and demonstrated that adult listeners can learn to associate a phonological feature with non-native phonetic correlates.

Most recently, various subfields of phonetics and phonology continue to provide abundant evidence for the emergent nature of features (e.g. Boersma and Hamann, 2008; Cohn, 2011; Mielke, 2008; Pulleyblank, 2006; but see Hale et al., 2006 for an opposing view) For instance, computer simulations show that sound inventories come to reflect distinctive feature patterns on the basis of the sounds' articulatory or auditory phonetic properties (Lin and Mielke, 2008). As for human learners, language acquisition studies exemplify that feature patterns develop in stages (Fikkert and Levelt, 2008; Levelt and van Oostendorp, 2007). Moreover, these stages do not follow a universal path: the feature structures emerging at various points of speech development differ across languages as well as across individuals (Menn and Vihman, 2011). With respect to adult phonologies, Morén (2003) argued that the feature systems of spoken and signed languages exhibit striking similarities: thus, since speaking and signing happen in different modalities, the mapping between the features and psychophysical reality cannot be innate. The present thesis takes on the study of feature emergence from yet another perspective. In Chapter 3, we investigate sound–feature mappings in a vowel system that has recently undergone a sound change.

In the experiments from Chapter 3, we focused on the GOOSE vowel (transcribed as /u/) of the variety of Standard English spoken in Southern England (SESE). Phonetically, /u/ has changed: along the phonetic F2 dimension that traditionally cues the phonological backness feature, /u/ seems to merge with /i/. Phonologically, however, /u/ has not changed: /i/ and /u/ still represent a backness contrast. The phonological backness distinction is manifested in phonological processes such as glide insertion: before vowel-initial words, a back glide [w] is inserted after /u/ while a front glide [j] is inserted after /i/.

Given the lack of phonetic F2 differences between /i/ and /u/, if the mapping between feature and phonetics were innate, one would have to conclude that SESE has lost the phonological backness contrast in high vowels. Such loss of contrast is however not viable given the evidence from phonological processes. On the contrary, if the mapping between feature and phonetics is emergent, one might argue that SESE speakers have learned to associate the backness feature with a phonetic cue other than F2. Chapter 3 therefore examined whether there is such a new phonetic cue that speakers associate with the phonologically back vowel /u/, and by extension – if listeners map phonetic cues directly onto features – with the backness feature in general.

1.4 CAN PERCEPTUAL PATTERNS REVEAL FEATURE STRUCTURE?

The results reported in Chapter 2 will show whether the phonological representation onto which listeners map the sound is the feature. Chap-

ter 3 will then indicate whether the mapping emerges as a result of the listeners' experience with their native language. Note that in both Chapter 2 and Chapter 3, previous phonological analyses informed us which distinctive feature is part of the listeners' language. That is, in Chapters 2 and 3, we knew beforehand what feature it is that we should see reflected in listeners' perception.

Likewise, a large body of studies compared languages with different phonological systems and found that the (a priori known) phonological differences were reflected in listeners' perception of speech sounds (e.g. Polivanov, 1931; for a review of the literature see Sebastián-Gallés, 2005). Moreover, language-specific effects have been shown to occur at early stages of neural processing (e.g. Näätänen et al., 1997). This suggests that the effect of phonology on speech sound perception is automatic and occurs without listeners' attention.

For some languages, however, phonological analyses fail to conclusively determine their feature structure. In that respect, given the well-documented effect of phonology on perception, one could examine perception in order to reveal the unknown phonology. That is, if the sound is mapped to features, listeners' perception could reveal whether their language encodes a given phonetic dimension in terms of a phonological feature.

An example of a so-far unresolved feature structure is vowel length in Dutch. Chapter 4 thus reports two experiments that aimed at uncovering whether Dutch listeners encode vowel duration in terms of the phonological length feature. The experiments assessed Dutch listeners' pre-attentive sensitivity to vowel duration and compared it across different vowels and to listeners from other languages.³ Specifically, we first tested whether Dutch listeners' processing of duration in native and non-native vowels resembles listeners who have the length feature (namely, Czech) or those who do not have it (namely, Spanish). Subsequently, we investigated whether a native Dutch vowel contrast that is realized partly by duration is represented phonologically as a length contrast.

1.5 PHONOLOGICAL FEATURES AND THE LANGUAGE LEARNER

The work reported in this thesis is done within the framework of Bidirectional Phonetics and Phonology (BiPhon; Boersma, 2007, 2009, 2011; Boersma and Hamann, 2009; Hamann, 2011; based on Boersma, 1998) Figure 1.3 shows a BiPhon model with five levels of representation.

³ Measuring pre-attentive perception enabled us to provide an assessment of listeners' speech sound processing unaffected by decision biases that can arise in behavioral tasks. Note that it is particularly desirable to eliminate the decision-bias in cross-linguistic comparisons where such biases could be specific to cultural differences.

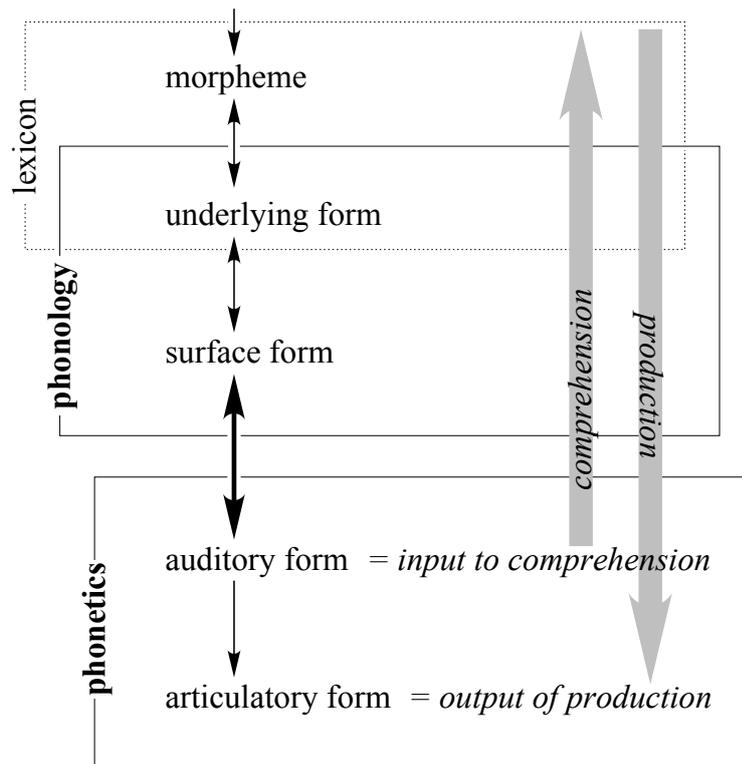


Figure 1.3: The model of Bidirectional Phonetics and Phonology (BiPhon, Boersma, 2009). The figure shows two phonetic, two phonological, and one morphological level of representation. The mappings between levels of representation are shown as thin black arrows, and the phonetics-phonology interface is marked by a thick black arrow. The thick grey arrows depict the direction of speech comprehension and speech production, and illustrate which levels are involved in these processes. Note that the phonological underlying form and the morpheme are part of the lexicon, i.e. they are stored representations; levels of representation above the morpheme are not shown here.

As is seen in Figure 1.3, the phonetics consists of two levels: the auditory and the articulatory form. The phonology also contains two levels: the underlying and the surface form. The underlying form is a collection of phonological categories of the utterance, and also contains information on morphological boundaries (which is copied from the morpheme level). The surface form consists of structured phonological units such as features, segments, syllables, and feet, which potentially form a tree-like hierarchy (Boersma, 2011). The proposed hierarchical structure suggests that the surface form could be further subdivided into several levels, each of which would contain units of the same size, e.g. a feature level separate from a segment level. While, in most theories, the feature is

the *smallest* phonological representation⁴, it is not clear whether the feature is also the *lowest*-level phonological representation, which is directly connected to the phonetics.

Importantly, note that in search for the lowest-level phonological representation, this thesis investigates perception rather than production. This is because in BiPhon, as shown in Figure 1.3, comprehension (more specifically, pre-lexical perception) is modeled as a *direct* mapping from the auditory form to the phonology, while production is a mapping from the phonology *via the auditory form* to the articulatory form. Consequently, the auditory form reflects phonological structure more straightforwardly than the articulatory form does.⁵

In BiPhon, learning, perception, and production have traditionally been modeled with algorithms and evaluation strategies of Stochastic Optimality Theory and Harmonic Grammar (e.g. Boersma, 1997; Boersma and Escudero, 2008; Boersma and Hamann, 2008). Recently, BiPhon has been implemented as a neural network (NN) model (Benders, 2013; Boersma et al., 2013a).

The BiPhon NN has been used to model phonological category emergence (Benders, 2013; Boersma et al., 2013a), and to examine whether the phonological categories that learners create are features or whether they are phonemes (Boersma and Chládková, 2013b; Boersma et al., 2013b). The outcomes of these previous simulations on feature versus phoneme emergence diverge and are summarized in Chapter 5. Furthermore, in Chapter 5, I report on follow-up simulations that aim to provide a more realistic account of vowel learning. The results of the present simulations will show whether, and under which circumstances, a virtual learner acquiring a 5-vowel system comes to represent her vowels in terms of features or in terms phonemes.

1.6 SUMMARY

To recapitulate, Chapter 2 presents an experiment that tests whether the phonological categories through which listeners perceive speech sounds are features or phonemes. In Chapter 3, we then investigate whether the mapping between the speech sound and the phonology is inherent (i.e. innate and universal) or arbitrary (i.e. emergent and based on the listeners' environment). The experiments presented in Chapter 4 assess listeners' perceptual patterns in order to uncover the as yet unclear phonological feature structure of their language. Finally, Chapter 5 re-

⁴ That the feature is the smallest unit of phonological analysis does not hold for e.g. Autosegmental Phonology where a single feature specification can stretch across several segments.

⁵ An auditory-based approach to the phonetics-phonology interface was also taken by e.g. Diehl and Kluender (1989); Kingston and Diehl (1995); Nearey (1995).

ports on computer simulations of vowel learning and perception with which we aim to determine whether a virtual infant learns to represent her native vowels in terms of features or in terms of phonemes. Chapter 6 concludes with a summary of findings from Chapters 2 through 5.

In summary, the research reported in this thesis will reveal whether phonological features are perceptually based linguistic categories. It will be shown whether listeners link perceived speech sounds directly to feature categories, and whether the link between sound and features is universal or learned from one's linguistic experience.

THE HUMAN LISTENER AS A PHONOLOGICAL FEATURE DETECTOR: THE PERCEPTUAL BASIS OF VOWEL HEIGHT

This chapter is a revised version of:

Kateřina Chládková, Titia Benders, & Paul Boersma. (in revision). The human listener as a phonological feature detector: the perceptual basis of vowel height.

ABSTRACT

For more than half a century, linguists have used distinctive features to describe speech sound inventories. Distinctive features are abstract phonological representations that have been named after actual phonetic properties of speech sounds. Thus, a direct relation has been traditionally assumed between a phonological feature and its phonetic correlate. The present study investigates whether a direct mapping between features and sound exists in the internal grammar of language users. The test case is a phonological feature that occurs in most of the world's languages, namely vowel height, and its acoustic correlate, the first formant (F₁). It was tested whether listeners map the F₁ dimension to vowel height feature values, or whether they map F₁ to phonemes. The results show that F₁ is perceived into native vowel height categories even in a vowel region that cannot be reliably identified with any phoneme of the listeners' language. This finding suggests that the phonological feature is the initial discrete representation onto which listeners map sound.

2.1 INTRODUCTION

Since the 1950's, phonological theory has described the sound patterns of the world's languages in terms of distinctive features (Jakobson et al., 1952). Distinctive features are abstract phonological representations that are supposedly directly related to observable phonetic properties of sounds: articulatory gestures, auditory cues, or both at the same time (Chomsky and Halle, 1968; Jakobson et al., 1952; Stevens, 1989). In that respect, the fact that a particular phonetic dimension is used to contrast speech sounds in a language implies that the corresponding distinctive feature is employed in that language's phonology. For instance, the feature vowel height corresponds to the first formant dimension (F₁) phonetically. Accordingly, a language that uses F₁ to contrast some of its vowels phonetically, is described as having the vowel height feature in its phonology. A contrastive speech sound, i.e., a phoneme, can then be analyzed as a bundle of features and their values. For instance, in many languages the phoneme /i/ can be analyzed as a vowel with the feature values [+high] and [-back].

Jakobson et al. (1952: 8) argued that “[a]ny distinctive feature is normally recognized by the receiver if it belongs to the code common to him and the sender, is accurately transmitted and has reached the receiver” [italics are ours]. In line with that claim, several early speech perception studies suggest that listeners extract linguistic features from the sound and that humans possess (innate) feature detectors (Eimas and Corbit, 1973; Miller and Nicely, 1955; Studdert-Kennedy and Shankweiler, 1970, illustrated in the right panel of Figure 2.1). In contrast, Pisoni and Luce (1987: 29–37) pointed out that many results that had been presented as support for the feature-detector theory could also be interpreted in favor of phonemes as the initial units of perception (as illustrated in the left panel of Figure 2.1).

The present study contributes to the long-standing debate on the nature of the units of speech perception (for a review see Pisoni and Luce, 1987) in that it investigates the initial phonological representation interfacing with the phonetics. Specifically, we test whether listeners directly perceive the speech signal in terms of features or in terms of phonemes. Figure 2.1 illustrates two possible models of low-level speech perception: one in which the sound is initially perceived in terms of phonemes (left), and one in which the sound is initially perceived in terms of features (right). The figure shows examples of mappings between F₁, the feature vowel height, and vowel phonemes. Note that we are not questioning the *existence* of features or phonemes: both features and phonemes can exist at some level of representation in the phonological grammars of language users; we test which of these two representations is *accessed first*

in perception. Below we review recent studies that relate to the question of whether distinctive features are the initial units of perception.

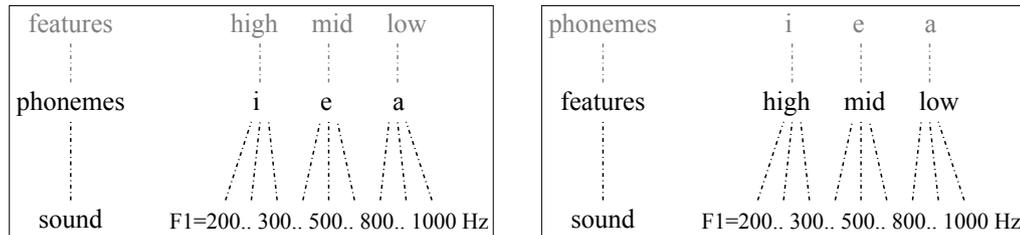


Figure 2.1: The two competing models of vowel perception. Left: the sound is initially perceived in terms of phonemes: the F_1 dimension is mapped to phoneme categories. Right: the sound is initially perceived in terms of features: the F_1 dimension is mapped to height categories.

Phonetic analyses of vowel inventories across languages provide a robust piece of evidence in favor of sound-feature mapping. Chistovich et al. (1966) noted that Swedish listeners have horizontal phoneme boundaries between high and mid vowels. Similarly, Boersma and Chládková (2011) observed horizontal boundaries between high and mid vowels in the vowel identification data of Czech, Dutch, Finnish, German, Italian, Spanish, and Polish listeners reported in Savela (2009). The horizontal boundaries between high and mid vowels in perception are remarkable given that the boundaries between high and mid vowels in *production* are diagonal. In other words, high versus mid vowels such as /i/ versus /e/ differ in both F_1 and the second formant (F_2) in production; yet, listeners seem to listen only to F_1 when classifying vowel tokens as /i/ or /e/. How does this production–perception discrepancy arise, assuming that a listener employs the same phonological grammar during both processes? In speech production, some articulatory movements require more effort than others, which may prevent the speaker from producing the vowel /e/ with the same high F_2 value as that of the corner vowel /i/. In contrast, perception is not constrained by limitations on articulatory movements and can more straightforwardly than production reflect the phonology that underlies language users’ performance. As suggested by Boersma and Chládková, the horizontal perception boundary between high and mid vowels then indicates that the F_1 dimension is mapped directly to the feature vowel height. To test whether F_1 and F_2 are mapped to vowel features or to phonemes, Boersma and Chládková ran simulations of vowel learning and subsequent vowel perception. Virtual learners were trained on input with diagonal boundaries between high and mid vowels (as produced by their virtual parents). Learners who perceived the signal in terms of features acquired horizontal (i.e., realistic) perceptual boundaries, while learners who perceived the signal in terms of phonemes acquired diagonal (i.e., unrealistic) perceptual boundaries.

Therefore, it seems plausible that human listeners map sound directly to the distinctive features of their native vowel system (and perhaps only indirectly to phonemes as shown in Figure 2.1, right).

Kingston (2003) tested whether in learning a foreign vowel system, adult human listeners extract the phonological feature structure of that system. Kingston showed that American-English listeners who had been trained with three German high–nonhigh pairs (/v-ø/, /u-œ/, and /ʏ-o/) discriminated a novel German high-nonhigh contrast /y-ɔ/ better than listeners who had been trained with only one of the three pairs. Besides these findings for vowel height, similar results were found for the vowel backness feature. Along with the outcomes of further experiments reported in that paper, Kingston's 2003 finding suggests that humans can readily learn to organize novel speech sounds in terms of features. In line with that, Lin and Mielke (2008) showed that an automated subdivision of a typical language's acoustic data (isolated segmental tokens without phoneme labels) divided up these sounds approximately into sonorants and obstruents, and that an automated subdivision of articulatory data divided up the sounds approximately in velars and non-velars. If human listeners can perform this phonetics-based induction of phonological features equally well, one could speculate that phonological features are linked directly to the acoustics (and the articulation).

Neurolinguistic research with human listeners has also claimed that phonological features affect speech sound processing. For instance, Scharinger et al. (2012) measured the neural response to the American English vowels /ɪ/, /ɛ/, and /æ/, and found that the differences in localizations of the pre-attentive response were better accounted for by a model that contained both feature differences and acoustic distance than by a model that only contained the acoustic distances between the vowels. The authors did not compare the feature-based model to a phoneme-based model, and our inspection of their data suggests that a phoneme-based model would have yielded the same results as the feature-based model has. A study more relevant for the feature vs. phoneme debate was performed by Scharinger et al. (2011a), who investigated the perception of the eight Turkish vowels. A model in terms of three phonological features (height, backness and roundedness) had a reliably better fit to the data than a model in terms of three acoustic dimensions (the first three formants). Although these authors again did not compare the feature model with a phoneme model, our inspection of their data suggests that a phoneme model would have yielded a different fit than the feature model. Scharinger et al. (2011a) therefore came close to being able to determine whether the lowest-level phonological representation is the feature or the phoneme.

Relatedly to neurophysiological studies on auditory speech sound processing, Ashby et al. (2009) assessed the neural processing of or-

thographically presented speech. Using a visual word priming experiment, Ashby et al. tested the processing of /d/- and /t/-final words that were preceded by non-word primes whose final consonant was either congruent or incongruent in voicing with the targets (e.g. /b/ or /p/). The authors demonstrated that phonological feature congruency affected written word recognition at very early stages of processing, namely by 100 ms after stimulus presentation. The early effect suggests that readers mapped the written input (i.e. letters) onto phonological features directly. Alternatively, as Ashby et al. suggested, readers might have activated an acoustic phonetic representation for the written input: under this scenario, the mapping of letters onto phonological features would pass via the reconstructed acoustic representations. In either case, the phonological feature appears to be the linguistic representation onto which Ashby et al.'s participants mapped the physical reality.

In sum, neurolinguistic literature suggests that listeners map incoming sound onto abstract phonological units. Some of the neurolinguistic studies, along with results from behavioral research and computer simulations of phonology and perception indicate that the initial phonological units in speech sound perception might be phonological features, and not phonemes. The present study addresses the feature vs. phoneme issue directly. It focuses on the phonological feature vowel height and its relation to the acoustic dimension of F₁. Starting with vowel height seems particularly useful if one aims to extend one's findings to the perceptual basis of distinctive features in general. This is because vowel height contrasts are found in all languages (Jakobson et al., 1952: 28; Halle, 1970): even the world's smallest vowel systems, namely those with 2 or 3 phonemes only, always distinguish a low vowel (e.g. /a/) and at least one non-low vowel (e.g. /ə/, /i/, or /u/) (see Crothers, 1978: 108–109; Maddieson, 1984: 125; Halle, 1970).

Whereas previous studies mostly tested feature perception in speech sounds with which listeners had (some) experience, i.e., native or newly learned sounds (Kingston, 2003; Scharinger et al., 2011a, 2012), we investigate whether listeners *generalize* the native vowel height feature to novel, *unknown*, sounds. Specifically, we test whether in the regions of the vowel space that are not used by the native vowel inventory listeners still perceive the F₁ dimension in terms of their native-language vowel height categories. The use of an unknown region was introduced by Bennett (1968) to investigate relative cue weighting in German and English in a way unbiased by the listeners' native phonemic experience. With respect to our question about the initial phonological representation interfacing with phonetics, using a novel uncolonized region allows us to collect responses that are unconfounded by the listeners' phonemic or lexical experience with the stimuli.

A suitable testing ground for the mapping between F1 and the vowel height feature is a language with a typical 5-vowel inventory of /i e a o u/. As illustrated in Figure 2.2, such a language associates low F1 values in the front and back vowel region with the high vowels /i/ and /u/ respectively, medium F1 values in the front and back vowel region with the mid vowels /e/ and /o/ respectively, and high F1 values in the central vowel region with the low vowel /a/. In the upper central part of the vowel space, i.e., in the region halfway between the non-low front and back vowels, typical 5-vowel languages do not have any phonemes. The upper central vowel region can thus be called uncolonized.

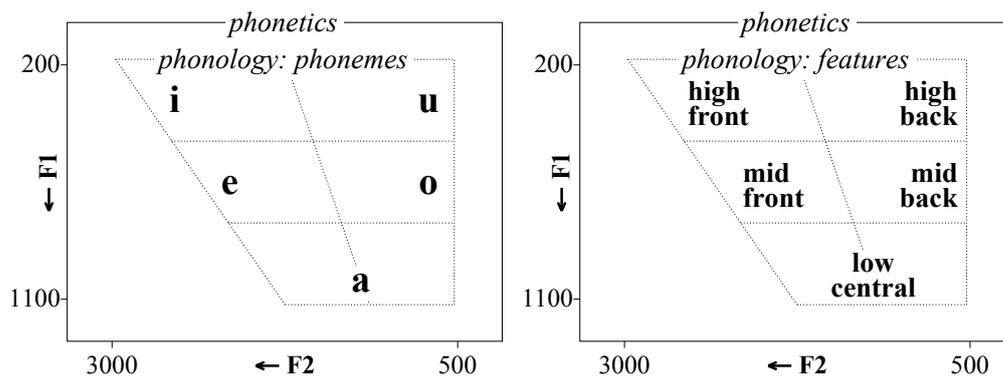


Figure 2.2: Phonetic and phonological organization of a typical 5-vowel system. Each of the five phonemes (left panel) is defined by the features vowel height and vowel backness (right panel), which correspond to the phonetic dimensions of first and second formant (F1 and F2), respectively. Note that there are no high central and mid central vowels in this 5-vowel system. The phonological quadrilateral represents the traditional IPA chart.

If the native speaker of our 5-vowel language maps the auditory signal (e.g. the F1 and F2 dimensions) directly to features (e.g. vowel height and backness), then she should generalize the high–mid distinction from the front and back vowel regions to the uncolonized central region. That is, even though the listener does not identify the F1–F2 combinations in the uncolonized region as phonemes of her language, she should still associate low F1 values with the feature high and medium F1 values with the feature mid. If, on the other hand, the native speaker of our 5-vowel language maps the auditory signal to phonemes and not to features, then she should not perceive the uncolonized continuum in terms of her native height categories.

The present study tests the perceptual basis of vowel height in native speakers of Czech, specifically, the Moravian variety of Czech. This variety has a vowel inventory with 5 monophthongal qualities (/i ε a o u/¹),

¹ As the traditionally used IPA symbols for Czech vowels suggest, the front mid vowel /ε/ is produced with slightly higher F1 values than the back mid vowel /o/

all of which occur as phonemically short and long (Šimáčková et al., 2012). The vowels are phonologically defined by three height and three backness features, as summarized in Table 2.1 (Kučera, 1961). Since / ε / and / ε :/ are phonologically mid vowels, as are / o / and / o :/, we henceforth refer to the former as / e / and / e :/ to preserve in the notation the phonological-height symmetry between the front and the back vowels.

We report on two experiments. Experiment 1 determines the location of the uncolonized region in the vowel space of Czech listeners. Experiment 2, subsequently, investigates whether Czech listeners perceive stimuli from this uncolonized region in terms of their native height categories.

	front	central	back
high	i:/i		u:/u
mid	ε :/ ε		o:/o
low		a:/a	

Table 2.1: Czech monophthongal vowel phonemes and their height and backness features.

2.2 EXPERIMENT 1

The goal of Experiment 1 was to determine the location of an uncolonized vowel region, i.e., a vowel region in which listeners are most uncertain in their identification of vowel phonemes. Therefore, it was a vowel identification task with stimuli sampled from the entire vowel space.

(Šimáčková et al., 2012). However, both / ε / and / o / have been described as mid vowels: articulatorily (Hála, 1960), acoustically (Hála, 1941), and phonologically (Kučera, 1961). Moreover, in vowel perception, the best-rated exemplars of / ε / have similar F1 values as the best-rated exemplars of / o / (Savela, 2009). This is not surprising if, as we argued above, perception but not production truly reflects the phonology (see also Boersma and Chládková, 2011). Our Experiment 1 will demonstrate that the Czech vowel system is indeed symmetrical (Figure 2.3) in that listeners associate front and back mid vowels with similar F1 values. A discussion of the cause behind the higher F1 of / ε / in speech production studies is outside the scope of the present paper. Interestingly, the Czech production data is in line with Maddieson’s (1984: 125) survey of vowel inventories from 317 languages, which shows that back vowels are universally more common than front vowels in the higher-mid range, while front vowels are more common than back vowels in the lower-mid range.

2.2.1 *Method*

2.2.1.1 *Participants*

The participants were 50 native speakers of Czech (33 female) from central and southern Moravia. They were all university students or recent graduates between 19 and 26 years of age. They were all monolingual speakers of Czech: they had been raised by native Czech-speaking parents, had never stayed in a foreign country for longer than 2 months, and self-rated their proficiency in any foreign language as poor. They reported no hearing or language problems and were each paid 7 euros for participation.

2.2.1.2 *Stimuli*

The stimuli in Experiment 1 were synthesized tokens of isolated vowels covering the whole vowel space (see e.g. Chládková and Escudero, 2012, for a similar whole-vowel-space stimulus design). F_1 , ranging from 280 to 1200 Hz, and F_2 , ranging from 800 to 3000 Hz, were both sampled in 16 steps that were auditorily equal on an Erb scale: the step size was 0.68 Erb for F_1 and 0.72 Erb for F_2 . Sixty-two F_1 – F_2 pairs were excluded: those for which F_1 would be equal to or higher than F_2 , which is by definition impossible, and those with a high F_1 and a high F_2 , which were judged to sound unnatural (frog-like). The remaining two-dimensional F_1 – F_2 vowel grid contained 194 tokens. The third formant (F_3) could have three values: 2900 Hz, 3260 Hz and 3700 Hz.² Combining three F_3 values with 194 F_1 – F_2 pairs yielded a total of 582 vowel tokens. All acoustic properties other than F_1 , F_2 , and F_3 were identical across the 582 vowel tokens. The duration of the vowels was 330 ms. The fundamental frequency rose linearly from 220 Hz at the start of the vowel to 270 Hz at one third of the total vowel duration, and then fell linearly to 180 Hz at the end of the vowel. The stimuli were modeled after a female voice and synthesized with a Klatt synthesizer (Klatt and Klatt, 1990) implemented in the program Praat (Boersma and Weenink, 1992-2013).

2.2.1.3 *Procedure*

Vowel identification was tested in a multiple forced-choice labeling task. Each trial started with a 600-ms silent interval, after which one of the 582 stimuli was presented to the participant via circumaural headphones. The participant then indicated which Czech vowel she heard by clicking

² To avoid ending up with tokens whose F_2 value would be very near to, or even higher than their F_3 value, we assigned every vowel token an actual F_3 value which was computed as the maximum of the specified F_3 value (i.e., each of 2900 Hz, 3260 Hz, and 3700 Hz) and of $F_2 + 200$ Hz.

on one of 10 buttons with orthographic labels for the 10 Czech monophthongs, /i: i e: e a: a o: o u: u/. Each stimulus was presented once, and there was no option of replaying a stimulus. After a participant's response, the next stimulus was played. Participants were allowed to take a short break after every 100th trial, and took between 35 and 45 minutes to complete the whole task. Prior to the test, the participants were not informed about the purpose of the experiment.

2.2.2 Results and discussion

Each participant labeled each stimulus once. To locate the F2 region on which Czech listeners as a group are least consistent, the results of the 50 participants were pooled. For each stimulus, we determined the winning label, i.e., the label that the stimulus received most often.

Figure 2.3 displays the winning labels for the stimulus set: the size of the symbol reflects the consistency of the winning label across listeners, which is defined as the proportion of the listeners who assigned that winning label to this stimulus. It is seen that at an F2 of about 2700 Hz and at an F2 of about 960 Hz the labeling consistency is high. This is in line with the fact that Czech has phonemes with the vowel qualities of /i/ and /e/ and phonemes with the vowel qualities of /u/ and /o/. By contrast, as the Figure also shows, at the intermediate F2 of about 1790 Hz, the between-subjects labeling consistency is low. This suggests that the F2 region at about 1790 Hz is not consistently identified with any phoneme. This is in line with the fact that Czech has no phonemes with such central qualities.

To ensure that the low labeling consistency in the central region is not due to large between-subjects variation, we tested whether a large labeling variability in this region is found within subjects as well. Around the F2 values of 2700 Hz and 960 Hz, i.e., in the front and back vowel region, we outlined areas with low F1 values, which represent the phonologically high vowels /i/ and /u/, and areas with medium F1 values, which represent the phonologically mid vowels /e/ and /o/. Around the F2 value of 1790 Hz, i.e., in the central vowel region, we outlined a low-F1 area and a medium-F1 area in a similar way (i.e., with identical F1 values as in the front and back region). Figure 2.3 illustrates these areas as shaded rectangles. Within each of these areas we then computed a first within-subject labeling uniformity (see below), which we call the "phoneme-area" labeling uniformity.

As can be seen in Figure 2.3, the parts of the vowel space between the shaded low-F1 and mid-F1 areas are likely to contain a boundary between /i/ and /e/ in the front vowels, and between /u/ and /o/ in back vowels. These boundary areas are marked with a thick dashed line. Within each of these three areas we computed a second within-subject

labeling uniformity, which we call the “boundary-area” labeling uniformity.

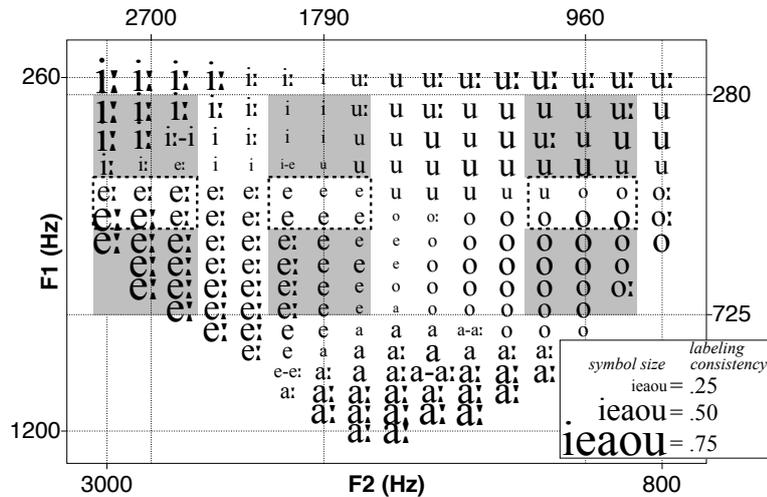


Figure 2.3: Results of the vowel identification task pooled across the three F₃ values. Symbols show the most frequently chosen label for each F₁-F₂ combination; symbol size correlates with between-subjects label consistency: the larger the label, the more subjects chose that label.

The within-subject labeling uniformity was measured in the following way. First, for every token j in a given area, we computed the proportion of tokens within the area that received the same label as j .³ The area’s labeling uniformity was then computed as the average of these proportions across all tokens in that area.⁴ For each region (front, back, central), a single measure of phoneme-area labeling uniformity was computed as the average of the uniformity of the low-F₁ area and the uniformity of the medium-F₁ area. Thus, we obtained for each participant her average phoneme-area labeling uniformity and her boundary-area labeling uniformity in the front and back vowel regions, which supposedly contain phonemes in her language, as well as in the central vowel region, which supposedly does not contain phonemes in her language.

Note that phoneme identification at phoneme boundaries is associated with uncertainty (Pisoni and Tash, 1974). Therefore, the boundary area should have a lower labeling uniformity than the phoneme area in re-

- ³ Note that we compared the vowel quality of the labels and did not consider length of the labels. That is, the labels of /o/ and /o:/, for instance, were considered the same. This is because whether a Czech vowel is phonologically short or long does not affect its phonological height feature (see Table 2.1), although long non-low vowels might be produced with a slightly lower F₁, and long low vowels with a higher F₁, than their short counterparts (Kučera, 1961).
- ⁴ Note that Figure 2.3 pools results for all three F₃ values; however, in the assessment of the within-subject labeling uniformity all three F₃ values were included separately (i.e., each of the outlined rectangles in Figure 2.3 represents a three-dimensional region).

gions of the vowel space where listeners distinguish phonemes, but not in regions where listeners do not have any phonemes.

The obtained uniformity scores were submitted to a repeated-measures analysis of variance with region (front, back, central) and area type (phoneme, boundary) as the within-subjects factors. There were main effects of region ($F[2, 98] = 64.231, p < .001$), and area-type ($F[1, 49] = 10.463, p = .002$), as well as a significant interaction between the two factors ($F[2, 98] = 12.236, p < .001$).

	front region	central region	back region	average across regions
phoneme	.875	.633	.901	.803
area	(.853-.898)	(.597-.669)	(.875-.929)	(.783-.823)
boundary	.807	.636	.730	.724
area	(.755-.859)	(.574-.698)	(.681-.779)	(.685-.763)
average	.841	.634	.815	
across areas	(.819-.864)	(.593-.676)	(.795-.835)	

Table 2.2: Within-subject labeling uniformity in the three regions of the vowel space, in the phoneme area and the boundary area. The table shows the means across 50 subjects and their 95% confidence intervals (in parentheses).

Table 2.2 lists the labeling uniformity scores in the three regions and in the two area types. The main effect of region suggests that labeling uniformity differs across the three regions of the vowel space: comparison of the means shows that the front and the back region have a larger labeling uniformity than the central region. As for the main effect of area type: labeling uniformity is larger in the phoneme areas than in the boundary areas. To further investigate the two-way interaction between region and area type we ran paired-samples t tests comparing the phoneme-area and the boundary-area uniformity within each region. The comparisons reveal that the phoneme area has a significantly larger labeling uniformity than the boundary area in both the front and the back region, while no difference between the two area types was found in the central region (front: $t[49] = 2.089, p = .021$; back: $t[49] = 5.098, p < .001$; central: $t[49] = 0.088, p = .465$).

The main effect of region shows that a listener classifies the central region more variably than the front or back region. Apparently, listeners either have a vertical phoneme boundary within the central region (separating front and back vowels), and/or they are unsure about the identity of the stimuli in the central region and therefore randomly choose labels for them. The finding that labeling was more variable in the boundary area than in the phoneme area for both front and back vowels but not for

central vowels indicates that listeners have a phoneme boundary (separating high and mid vowels) in the front and back regions but not in the central region.

The findings of Experiment 1 can be summarized as follows. The low between-subjects consistency indicates that (1) the central region is not used in native Czech speech perception and production: a vowel token from this region would often be perceived by the listeners as a different vowel category than the speaker intended, i.e., communication would fail. The large within-listener labeling variability in the central region implies that (2a) listeners are unsure about the phonemic identity of the stimuli, or that (2b) the central region contains a phoneme boundary between front and back vowels, which tends to be associated with uncertainty (Pisoni and Tash, 1974). The large labeling variability in both the phoneme area and the boundary area in the central region implies that (3) listeners do not reliably divide the central region into distinct high and mid phoneme categories. For these reasons, we interpret the result as a lack of phoneme “colonization” in the central region. We thus use the central region as an “uncolonized” region in Experiment 2, which is about the generalization of the vowel height feature.

2.3 EXPERIMENT 2

Recall that the present study investigates whether listeners map sound to features or to phonemes. If they map sound initially to features, we expect to find categorical perception of vowel height even in uncolonized regions of the vowel space, where there are no phonemes in our listeners’ language. If they map sound initially to phonemes, we do not expect to find categorical perception in these uncolonized regions. Experiment 1 has determined such an uncolonized region for Czech listeners.

Experiment 2, then, investigates whether listeners perceive F_1 differences within the uncolonized region categorically, that is, if they have perceptual boundaries along that region. We determine the presence of category boundaries by measuring discrimination along the uncolonized central continuum (denoted as $i\sim\text{ə}$) and comparing that to discrimination along the existing front and back continua (denoted as $i\sim e$ and $u\sim o$, respectively). Discrimination is tested in an AX task, in which participants have to tell whether two sounds are the same or different. This task can reveal category boundaries if listeners report to hear a difference between sounds from some parts of an auditory continuum but not between sounds from other parts (Pisoni, 1973). The data obtained in a discrimination task yield a discrimination function, which is the number of ‘different’ responses as a function of the location along the stimulus continuum. A peak in the discrimination function (i.e., a larger number of ‘different’ responses in a small part of the stimulus continuum) cor-

responds to a boundary between two categories (Liberman et al., 1957). The presence of one discrimination peak suggests that the given auditory continuum is perceived into two discrete categories; two discrimination peaks suggest that the auditory continuum is perceived into three discrete categories. The absence of discrimination peaks indicates that the auditory continuum is not perceived categorically and that listeners hear acoustic differences between sounds equally well along the whole continuum. Experiment 2 has two possible outcomes. If listeners map sound initially to phonemes, they will have discrimination peaks in the front and back regions but not in the uncolonized region. If listeners map sound initially to features they will have discrimination peaks in the uncolonized region that resemble the peaks in the front and back regions.

2.3.1 *Method*

2.3.1.1 *Participants*

A total of 81 listeners participated in the AX discrimination task: 24 participants were tested on the front *i~e* continuum (16 female), 26 on the back *u~o* continuum (17 female), and 31 on the central *i~ɘ* continuum (23 female). The criteria for the participants in this experiment were the same as in Experiment 1.⁵ Their age was between 18 and 30 years. They were each paid 5 euros for participation.

2.3.1.2 *Stimuli*

The stimuli were artificial vowels created with a synthesis procedure identical to the one in Experiment 1. Vowels were synthesized along 3 different F_1 continua: one in the front, one in the back, and one in the central region of the vowel space. F_1 always ranged from 280 to 725 Hz. On the front (*i~e*) continuum, all stimuli had $F_2 = 2700$ Hz and $F_3 = 3300$ Hz. On the back (*u~o*) continuum, all stimuli had $F_2 = 960$ Hz and $F_3 = 2900$ Hz. On the central (*i~ɘ*) continuum, all stimuli had $F_2 = 1790$ Hz and $F_3 = 3260$ Hz. The three continua thus differed in both F_2 and F_3 : the *u~o* continuum had the lowest F_3 , because back vowels in Czech are rounded. In contrast, the *i~e* continuum had the highest F_3 , because front vowels in Czech are unrounded. The F_3 of the stimuli on the *i~ɘ* continuum was relatively high, which means that the uncolonized continuum corresponded to central unrounded vowels (as

⁵ The 24 and 26 participants who discriminated the *i~e* and the *u~o* continuum respectively were the same individuals that took part in Experiment 1. To avoid any potential labeling biases during discrimination, Experiment 2 was administered before Experiment 1. Also, between Exp. 2 and Exp. 1, participants took a one-hour break outside the testing room. The 31 participants for the *i~ɘ* continuum were tested a month later and did not participate in Experiment 1.

is also implied by use of the symbols ɨ and ɘ) that have no phonemic status in the vowel inventory of Czech.

We synthesized 260 vowel tokens per continuum, which were combined into 130 stimulus pairs. The F₁ distance between the two vowels within a stimulus pair was 0.9 Erb, and the F₁ distance between two neighboring stimulus pairs (e.g. between the first vowel of pair 1 and the first vowel of pair 2) was 0.039 Erb. Figure 2.4 shows the sampling along the stimulus continua. Note that unlike most earlier speech perception studies, we used densely sampled continua of non-repeating stimuli, which should provide more ecologically valid results than stimulus sets with a small number of repeating stimuli (Boersma and Chládková, 2013a; Rogers and Davis, 2009).

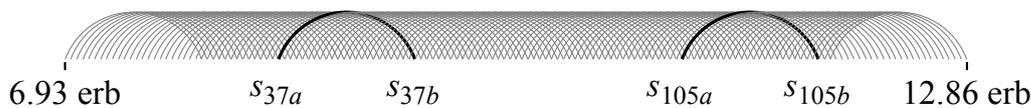


Figure 2.4: The 130 stimulus pairs along each of the three continua between 6.93 Erb and 12.86 Erb (280 Hz and 725 Hz). The members of a stimulus pair are connected by an arc. The auditory F₁ distance between the sounds within a stimulus pair is always 0.9 Erb, this is the F₁ distance between s_{37a} and s_{37b} and also the distance between s_{105a} and s_{105b} (these two pairs are shown by thick arcs). The F₁ distance between two adjacent pairs (adjacent in terms of F₁ along the F₁ continuum) is 0.039 Erb.

2.3.1.3 Procedure

On each trial, participants heard the two sounds of a stimulus pair. They indicated whether the two sounds were the same or different by clicking on one of the buttons on a computer screen that were labeled as “stejné” and “rozdílné” (‘same’ and ‘different’). There was no option of replaying the sounds. The first sound was preceded by a silence of 600 ms and the silent inter-stimulus interval was 500 ms. Each of the 130 stimulus pairs occurred twice: on one trial the sound with the lower F₁ was played first, while on the other trial the sound with the higher F₁ was played first. The complete set of the $2 \times 130 = 260$ stimulus pairs was randomized for each participant individually. Prior to testing, participants were not given any information about the language from which the stimuli were taken. Participants were allowed a short break halfway through the experiment and took about half an hour to complete the task.

Note that listeners never heard two identical stimuli within a trial in the AX task; nevertheless, we asked them to indicate whether the sounds were different or the same. The F₁ difference between the sounds was identical across all stimulus pairs, and was as small as 0.9 Erb, i.e., about

the size of a just noticeable difference for formants (Mermelstein, 1978). If the difference of 0.9 Erb is correctly perceived as different in some parts of the auditory continua but not in other parts, we will have found categorical perception.

2.3.2 Results and discussion

2.3.2.1 Determining the number of categories on a continuum

The design of the present study departs largely from that of previous studies in two respects. First, we used densely sampled non-repeating stimulus continua. Second, we tested perception of phonological features, i.e., categories for which the listeners have no labels. Since our uncolonized continuum is unidentifiable, the traditional means of assessing categorical perception, namely a comparison of the obtained discrimination scores to the discrimination scores predicted from identification data (Liberman et al., 1957; Schouten and van Hessen, 1992), were not applicable.⁶ Therefore, the present data are analyzed with the method proposed in Boersma and Chládková (2013a), which assesses categorical perception⁷ solely on the basis of peaks in the discrimination function.⁸ Moreover, the present method is suited for discrimination data on densely sampled stimulus continua.

Each listener was presented with each stimulus pair twice. Therefore, the number of times she responded ‘different’ to a stimulus pair could be 0, 1 or 2: discrimination peaks are located at those parts of the continuum where there are more 2s than in the surrounding parts. Figure 5 shows a plot of the raw data for one listener. The vertical lines indicate for every point on the continuum how many times the listener perceived that point as ‘different’. Visual inspection of peaks and valleys is possible after smoothing the raw data. Smoothing is done by convolution with a unit-area Gaussian that has a standard deviation of 10 steps along the continuum (i.e., $0.039 \times 10 = 0.39$ Erb); in Figure 2.5 this procedure produces the smooth curve, from which the peaks are easy to discern.

Visual inspection of the smoothed curves is ambiguous, though: the listener in Figure 2.5 has a clear peak around s_{58} , but does she also have

6 As also Kuhl (1981) pointed out, the traditional assessment of categorical perception with both identification and discrimination data is possible only with certain testing procedures (for instance, certain populations or stimuli).

7 Instead of “categorical perception”, the term “phoneme-boundary effect” (Wood, 1976) might be seen as more appropriate in the present study. However, we use the two terms interchangeably since the existence of a category boundary implies the existence of a different category at each side of the boundary.

8 Repp et al. (1979: 129) note that categorical perception can indeed be assessed on the basis of peaks and troughs in the discrimination function. Repp et al., however, consider this peak-based measure less important, partly because it is more difficult to quantify it than it is to quantify the fit between predicted and obtained discrimination.

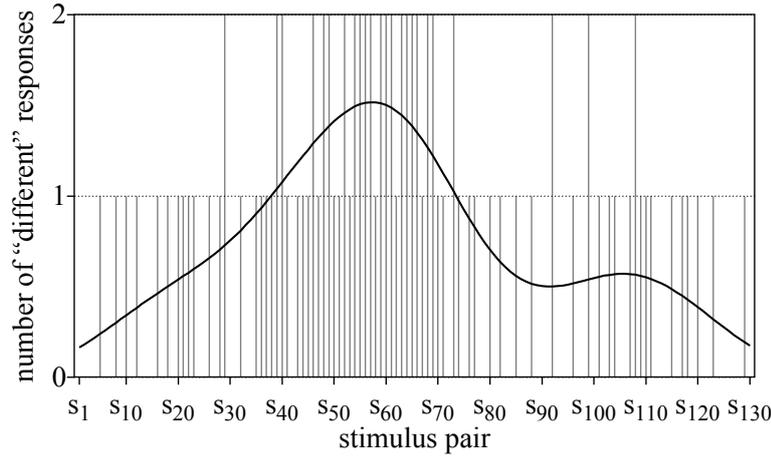


Figure 2.5: Raw (vertical lines) and smoothed data (curve) of one listener.

a peak around s_{106} ? The following mathematical method helps us to provide an answer. To quantify the number of peaks that a listener has, we submit the discrimination data to maximum-likelihood analyses. Specifically, we fit the raw data with several models that assume different numbers of discrimination peaks: each participant's discrimination function is modeled with zero, one, two, three and four peaks respectively. The model with 0 peaks corresponds to a flat discrimination function and is therefore defined by a single parameter p_- , which can be interpreted as the probability of perceiving the acoustic difference of 0.9 Erb as different. In models with 1 and more peaks, every discrimination peak is defined by 3 additional parameters: p_+ , μ , and σ , which describe the height of the peak, its location along the stimulus continuum, and its width, respectively. Thus, a model with z peaks has 3 more parameters than a model with $z-1$ peaks. For instance, the model with two discrimination peaks is defined as:

$$p_n = p_- + (p_{+1} - p_-)e^{-\frac{(n-\mu_1)^2}{2\sigma_1^2}} + (p_{+2} - p_-)e^{-\frac{(n-\mu_2)^2}{2\sigma_2^2}} \quad (2.1)$$

where n is the stimulus pair, which ranges from 1 to 130; p_- can be interpreted as the probability of judging the 0.9-Erb auditory difference *within* a category as different, a behavior that corresponds to acoustic listening; p_{+1} and p_{+2} can be interpreted as the probabilities of judging the 0.9-Erb auditory difference as different *across* a category boundary, i.e., they correspond to the heights of the first and second peak. Note that the values of p_- and p_+ range from 0 to 1, while the p_+ values are constrained to be larger than p_- . The parameters μ and σ are measured in units of 0.039 Erb, which equals the distance between neighboring stimulus pairs.

Using a maximum-likelihood method (Fisher, 1922) we then estimate which of the models best fits the participant's raw data. For every model,

we search for such values of the parameters for p_n that maximize the log-likelihood, computed as:

$$LL = \ln \prod_{n=1}^N p_n^{d_n} (1 - p_n)^{s_n} = \sum_{n=1}^N (d_n \ln p_n + s_n \ln(1 - p_n)) \quad (2.2)$$

where d_n and s_n correspond to the number of times (0, 1, or 2) that a listener judged the n^{th} stimulus pair as ‘different’ or ‘same’, respectively; and N is the total number of stimulus pairs, i.e., 130.

If the maximum likelihood of a model with $z+1$ peaks improves significantly compared to the preceding model with z peaks, then the model with $z+1$ peaks is considered a better fit to the data. When no significant improvement in maximum likelihood is seen in the model with $z+1$ peaks, then the model with z peaks is considered the best fit to the participant’s data. To test the significance of the maximum-likelihood improvement between the models with $z+1$ and z peaks, we compute ΔLL as the difference between the maximum log-likelihoods of the two models and then perform a χ^2 test on $2\Delta LL$ with 3 degrees of freedom (i.e., the 3 parameters of the $z+1^{\text{th}}$ peak), with $\alpha = .01$. This α lies in between Akaike’s Information Criterion and the Bayesian Information Criterion for significance in maximum-likelihood improvement (see Akaike, 1974; Pitt et al., 2002).

Figure 2.6 visualizes the comparison of the models with 0, 1, and 2 peaks for the listener from Figure 2.5. In Figure 2.6 it is seen that the model with 1 peak best describes this listener’s data. The improvement in maximum likelihood from a model with 0 peaks to a model with 1 peak is significant: the χ^2 test on $2\Delta LL$ yields a p -value of $3 \cdot 10^{-13}$ ($\chi^2[3] = 61.282$). Accordingly, it can be seen that the curve for the smoothed data overlaps better with the curve of the smoothed 1-peak model than with the curve of the smoothed 0-peak model. The improvement from a model with 1 peak to a model with 2 peaks is not significant ($\chi^2[3] = 3.234$; $p = .357$), and therefore the model with 2 peaks is not considered a better fit to the data than the model with 1 peak. We conclude that the peak that is visible around s_{106} in Figure 2.5 might well be spurious.

2.3.2.2 Comparing the number of categories across continua

Table 2.3 summarizes the results for all 81 listeners. It can be observed that on all three continua, most listeners had 1 or 2 discrimination peaks. In other words, they had one or two category boundaries along the continuum, which implies two and three perceived categories, respectively.

Inspection of the data in Table 2.3 suggests that the perception of vowels is similar across the three continua. In order to assess whether there were differences in perceptual strategies across the continua, we carried

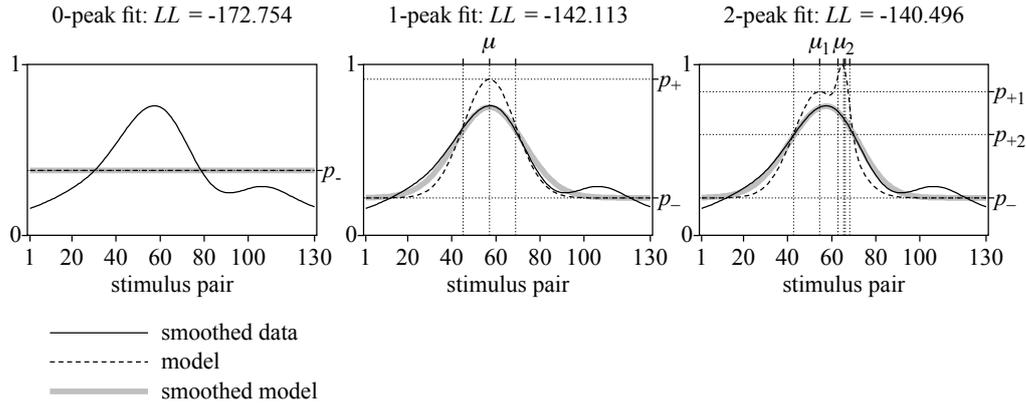


Figure 2.6: Smoothed data (black solid line), model (black dashed line), and smoothed model (thick grey line) for the listener from Figure 2.5.

n peaks↓	i~e	i~ə	u~o	total
0	3 (12.5)	7 (22.6)	3 (11.5)	13 (16.1)
1	12 (50.0)	16 (51.6)	12 (46.2)	40 (49.4)
2	9 (37.5)	7 (22.6)	11 (42.3)	27 (33.3)
3	0 (0)	1 (3.2)	0 (0)	1 (1.2)
total	24	31	26	81

Table 2.3: The number and percentage (in parentheses) of listeners with 0, 1, 2 and 3 peaks on each of the three continua.

out a χ^2 test of independence for groups. The test did not find a significant difference between the three continua with respect to the distribution of listeners with 0, 1, 2, and 3 peaks ($\chi^2[6] = 4.871, p = .560$). The absence of significant differences across the three continua suggests that the pattern of categorical perception on the i~ə continuum is similar to the pattern of categorical perception on the i~e and u~o continua.

Since most listeners had either one or two discrimination peaks, we further compared the parameters of models with 1 and 2 peaks across the three continua. That is, we compared the 12 i~e, 12 u~o, and 16 i~ə listeners with one peak, and also the 9 i~e, 11 u~o, and 7 i~ə listeners with two peaks. The averaged model fits of the one-peak and two-peak listeners are plotted in Figure 2.7.

To test whether categorical perception differs across the front, central, and back continua, the values of the three parameters p_+ , μ , and σ of the 1-peak listeners were submitted to a multivariate analysis of variance (MANOVA) with continuum (front, central, back) as fixed factor; a similar MANOVA was done for the six parameters p_{+1} , μ_1 , σ_1 , p_{+2} , μ_2 , and σ_2 of the 2-peak listeners. The MANOVA for the 1-peak listeners did not yield a significant effect of continuum (Wilk's $\lambda = 0.899, F[6, 70] =$

0.639, $p = .699$). The MANOVA for the 2-peak listeners yielded a significant effect of continuum ($\lambda = 0.260$, $F[12, 38] = 3.045$, $p = .004$). Univariate ANOVAs revealed that continuum had a significant effect on two parameters: σ_1 ($F[2, 24] = 4.614$, $p = .02$), and μ_2 ($F[2, 24] = 4.448$, $p = .023$). Pairwise comparisons showed that σ_1 was smaller on the $i\sim\vartheta$ continuum than on both the $i\sim e$ and the $u\sim o$ continuum ($i\sim e$: mean difference = 6.9, $p = .007$; $u\sim o$: mean difference = 5.0, $p = .034$). This implies that the first peak of 2-peak listeners is narrower on the $i\sim\vartheta$ continuum than it is on the $i\sim e$ and $u\sim o$ continua by about 6×0.039 Erb. Further, the pairwise comparisons showed that μ_2 is smaller on the $i\sim\vartheta$ continuum than on both the $i\sim e$ and the $u\sim o$ continuum ($i\sim e$: mean difference = 18.406, $p = .017$; $u\sim o$: mean difference = 18.959, $p = .011$). This implies that the second peak on the $i\sim\vartheta$ continuum is located at lower F1 values than on the $i\sim e$ and $u\sim o$ continua by about 19×0.039 Erb.

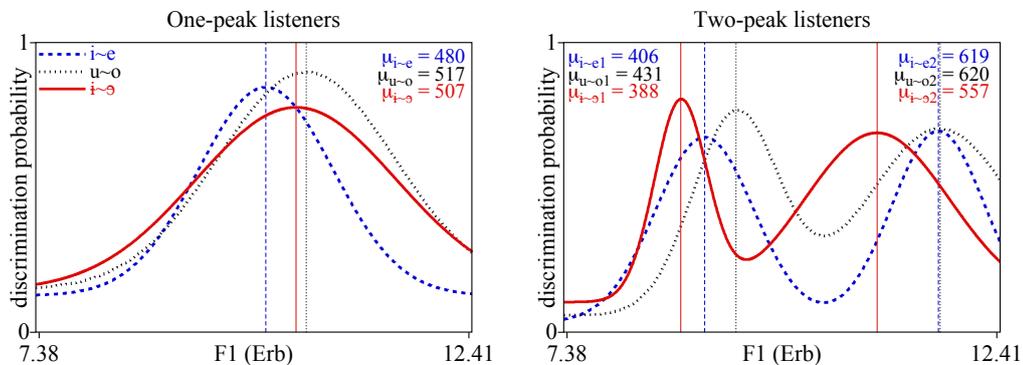


Figure 2.7: Smoothed average best model fits of one-peak (left) and two-peak listeners (right). There was no effect of continuum for one-peak listeners. In two-peak listeners, the first peak (i.e., the high-mid boundary) was narrower on the $i\sim\vartheta$ continuum than on the other two continua, and the second peak (i.e., the mid-low boundary) was located at lower F1 values on the $i\sim\vartheta$ continuum than on the other two continua. Locations of the peaks in Hz are shown in the Figure.

2.3.2.3 Summary of the results of Experiment 2

Experiment 2 was a vowel discrimination task that measured the degree of categorical perception on the uncolonized (central) continuum and on the existing (front and back) continua. We argued that if the uncolonized continuum, on which listeners have no phonemes, is perceived categorically and similarly to the existing continua, we will have found evidence for feature-based perception. We found that the number of discrimination peaks did not differ across the three continua (see Table 2.3), which suggests that perception on the central continuum is similar to perception on the front and back continua. On all continua, about half of the listeners had one discrimination peak, i.e., one category boundary,

which suggests that each continuum was perceived by most listeners into two categories. About a third of the listeners had two discrimination peaks, i.e., two category boundaries, which suggests three categories. Experiment 2 thus found categorical perception on the uncolonized continuum and the number of categories on the uncolonized continuum did not differ from the number of categories on the existing continua. We therefore conclude that listeners map sound initially to features and not to phonemes.

A comparison of the locations, widths and heights of the peaks in the one-peak listeners did not show differences across the continua. However, a comparison of the two-peak listeners showed that the width of the first peak and the location of the second peak on the central continuum differ from those on the front and back continua. The absence of between-continua differences for one-peak listeners is in line with feature-based perception. For two-peak listeners, however, there were slight differences between the uncolonized and the existing continua; in Section 2.4.3, we explain how these further support feature-based perception.

2.4 GENERAL DISCUSSION

2.4.1 *Main findings*

The present study investigated whether listeners perceive speech sounds in terms of distinctive features or in terms of phonemes. Experiment 1 was a vowel identification task and aimed to determine an uncolonized vowel region, on which listeners have low categorization certainty and differ from each other in their phoneme identification. Such a region was found in the central part of the vowel space, which does not contain phonemes in the listeners' language.

To test whether listeners map sound to features or to phonemes, we carried out Experiment 2, which was a vowel discrimination task on a vowel continuum in the uncolonized (central) region and in the existing (front and back) regions. This task assessed whether listeners perceive these vowel continua categorically. We predicted that if listeners map sound to features, perception on both the uncolonized and the existing continua would be equally categorical. In contrast, if listeners map sound to phonemes, perception would be categorical on the existing continua but not on the uncolonized continuum. The results showed that listeners perceive both the uncolonized central continuum and the existing front and back continua categorically. Moreover, the number of perceived categories did not differ across continua. On the basis of these findings, we conclude that the auditory F1 dimension is mapped to

vowel height categories, i.e., that listeners perceive F₁ values in terms of distinctive height feature categories, rather than in terms of phonemes.

The early speech perception studies that argued for feature-based perception assumed that feature detectors in human listeners were innate (e.g., Eimas and Corbit, 1973). We argue instead that feature categories could be acquired during native language development and therefore be language-specific. For instance, in a language that uses F₁ to distinguish vowels, the learner will realize, after a sufficient amount of input, that this dimension is relevant for phonological contrasts in her language and will start to create discrete categories along that dimension. Czech or Spanish infants, whose language contrasts 3 vowel heights, will form 3 categories along this dimension, while French or Portuguese infants will learn that there are 4 categories. Thus, in our vowel discrimination task on an uncolonized central continuum, Spanish listeners should have the same number of discrimination peaks as Czech listeners, whereas French and Portuguese listeners should have an extra discrimination peak.

2.4.2 *The number of perceived categories*

We now discuss why some listeners perceived two categories while others perceived three. Previous vowel identification experiments have shown that the location of the category boundaries along auditory continua may be influenced by the number of available response categories (Benders et al., 2012; Sawusch and Nusbaum, 1979). It appears that the category boundaries are distributed along a given stimulus continuum so as to allow for a sufficient auditory space for each of the available categories (cf. Benders et al., 2012). Unlike identification tasks, in which a listener chooses her responses from a predetermined set of categories, a discrimination task does not specify which categories she “should” perceive. Therefore, some listeners in the present study may have attempted to fit all their three height categories into the F₁ range 280–725 Hz, while others did not consider the third (low) category while discriminating the continua; see Figure 2.8 for an illustration of these two listening strategies.

Recall that among the front and back vowel phonemes, Czech contrasts two heights, high and mid (see Table 2.1). Still, there were listeners who discriminated 3 height categories along the front or the back vowel continuum. This finding further supports feature-based perception: high F₁ values are mapped to the feature low even if the presented F₁–F₂ *combination* is untypical of any phoneme in the listener’s phoneme inventory. In other words, Czech listeners can perceptually differentiate between a low and a mid height category in front or back vowels, even though this distinction does not contribute to phonemic contrasts in Czech.

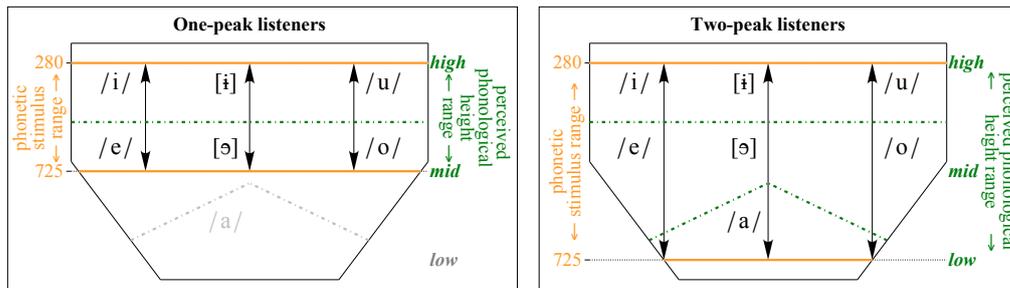


Figure 2.8: Perceptual vowel spaces in Czech(-like) listeners who map F1 to vowel height (and F2 to vowel backness). Left: one-peak listeners, right: two-peak listeners. The black arrows indicate the *phonological* height range that listeners perceived within the *phonetic* stimulus F1 range between 280 and 725 Hz. The phonetic stimulus range is marked by orange solid lines; the perceived phonological boundaries in that range are illustrated by green dot-and-dash lines.

2.4.3 Differences in boundary location in two-peak listeners

We argued that the two-peak (i.e., two-boundary) listeners perceive three height categories (high, mid, and low) on all the three continua. Thus, the first peak in these listeners corresponds to the high–mid boundary and their second peak corresponds to the mid–low boundary. We further found that the mid–low boundary on the central continuum was located at lower F1 values than on the front and back continua. This finding is in line with what has been reported earlier for vowel perception in virtual listeners who perceive vowel sounds in terms of features (Boersma and Chládková, 2011). The two graphs in Figure 2.8 show perceptual vowel boundaries in listeners who map F1 to the feature vowel height (and F2 to vowel backness).

Boersma and Chládková demonstrated that virtual feature-based listeners separate neighboring vowels that differ only in the height feature by a horizontal boundary, vowels that differ only in the backness feature by a vertical boundary, and vowels that differ in both height and backness by a diagonal boundary. Since the Czech high vowels /i/ and /u/ differ from the mid vowels /e/ and /o/ in height only, the high–mid boundary is horizontal. At the same time, as shown in Table 2.1, the Czech mid vowels /e/ and /o/ differ from the low vowel /a/ in both height and backness, which implies diagonal mid–low boundaries. See also Figure 2.3, where the boundaries separating the 5 vowel areas roughly correspond to the visualization in Figure 2.8.

Figure 2.8 (right) shows that the diagonal shape of the /e/-/a/ and /o/-/a/ boundaries affects the location of the mid–low boundary: in the front and back regions the mid–low boundary is at higher F1 values than in the central region. As seen in the Figure, the diagonal shape of the mid–low boundary explains why on the central continuum we

found a relatively low F₁ value of the mid–low boundary for the two-peak listeners in Experiment 2.

2.4.4 *Symmetric vs. asymmetric vowel systems*

Our claim that the phonological feature is the lowest-level phonological representation interfacing with the phonetics is further supported by the structure of vowel inventories across languages. Many languages have symmetrical vowel systems with respect to vowel height, that is, they have the same number of height distinctions across front and back vowels, e.g. Arabic, Spanish, Czech, Slovak, Portuguese, and Catalan (see, respectively, Carbonell and Llisterra, 1992; Cruz-Ferreira, 1995; Hanulíková and Hamann, 2010; Martínez-Celdrán et al., 2003; Thelwall and Akram Sa'Adeddin, 1990; Šimáčková et al., 2012). In addition, when vowel systems change diachronically, front and back vowels often shift in parallel to maintain the front-back symmetry in vowel height; see for instance Alkire and Rosen (2010) for the diachronic vowel changes in Romance languages.

However, there are also languages with asymmetric vowel inventories, e.g. Australian English and Dutch (Cox and Palethorpe, 2007; Gussenhoven, 1992).⁹ Data from such languages appear to run contrary to our present finding that listeners map the F₁ dimension directly to the feature vowel height. However, the front-back asymmetry in the number of apparent vowel height categories can occur even if listeners perceive F₁ in terms of vowel height. For instance, speakers of a language with three height contrasts in the front vowels and two height contrasts in the back vowels have three height categories in their phonology onto which they map any incoming F₁ value, even though they do not use one of these three heights phonemically in the back region. Our prediction is that speakers of such an asymmetric language *discriminate* the same number of height categories along both the back and front dimension (namely, three), even though they *identify* or *recognize* a different number of phonemes in the front than in the back of the vowel space (namely, three and two, respectively).

⁹ Universally, it is slightly more common that asymmetric vowel inventories have more phonemes in the front than in the back of the vowel space, than vice versa (Maddieson, 1984: 124), which suggests that the asymmetrical languages usually distinguish more heights among front than among back vowels. The disfavoring of phonemic contrasts among back vowels may be due to the lower acoustic saliency of back vowels, as noted by Maddieson (1984: 125) to explain the universal preference of /i/ over /u/ in 3-vowel systems.

2.5 CONCLUSIONS

The present study investigated whether listeners perceptually map speech sounds to distinctive feature categories. Specifically, we tested whether the F₁ dimension is perceived in terms of the vowel height feature, or whether it is perceived in terms of unanalyzed phonemes. We found that in an uncolonized vowel region that cannot be reliably identified with any of the phonemes of one's language, the acoustic F₁ dimension is perceived categorically. Moreover, the pattern of categorical perception in the uncolonized vowel region resembles categorical perception in regions in which the listeners' language does have phonemes. The present results thus show that listeners map the F₁ dimension initially to the vowel height feature rather than to phonemes. Therefore, we argue that the phonological feature is the *initial* discrete representation onto which listeners map the incoming speech signal.

2.6 ACKNOWLEDGMENTS

This research has been funded by the Netherlands Organization for Scientific Research grant no. 277.70.008 awarded to the third author. We would like to thank Barbora Chládková for help with participant recruitment, and to the Department of English and American Studies at Palacký University in Olomouc for providing testing facilities.

WHY SHE AND SHOE WON'T MERGE: REDEFINING
PERCEPTUAL CUES FOR THE FRONT-BACK
CONTRAST IN THE ENGLISH OF SOUTHERN
ENGLAND

This chapter is an adapted version of:

Kateřina Chládková, Silke Hamann, & Daniel Williams. (under revision). Why SHE and SHOE won't merge: Redefining perceptual cues for the front-back contrast in the English of Southern England.

ABSTRACT

The vowel /u/ (GOOSE lexical set) of the Standard English variety spoken in Southern England (SESE) has shifted from the back to the front area of the vowel space, so that it comes to be realized with high midpoint second formant (F₂) values similar to those of the vowel /i/ (FLEECE lexical set). Yet, there is no evidence of merger: recent production data suggest that /i/ and /u/ are differentiated by diphthongization of F₂ (and F₃): /i/ is realized with a rising and /u/ with a falling formant contour. Therefore, the present study tested whether diphthongization serves as a cue to the SESE /i/-/u/ contrast also in perception. The present findings show that both young and older SESE listeners rely on diphthongization to distinguish /i/ from /u/: an otherwise ambiguous token is identified as /i/ if it has a rising F₂ contour and as /u/ if it has a falling F₂ contour. Furthermore, the results indicate that listeners generalize their reliance on diphthongization to other contrasts, namely /ε/-/ʊ/ and /æ/-/ʊ/. This suggests that in SESE, a rising F₂ seems to be perceptually associated with the feature [+front] while a falling F₂ with the feature [-front].

3.1 INTRODUCTION

It has been well documented in the literature that the GOOSE vowel (i.e., /u/) of the variety of Standard English spoken in Southern England (SESE) has shifted from the back region of the vowel space, i.e. from low values of the second formant (F2), towards the front, i.e. to high F2 values (e.g. Bauer, 1985; Harrington et al., 2008; Hawkins and Midgley, 2005; Henton, 1983). Figure 3.1 illustrates this fronting of /u/ with data from old and young generations of speakers reported in the literature. It can be seen that due to its considerable phonetic fronting, /u/¹ comes to be realized with F2 values that are close to those of the FLEECE vowel (i.e., /i/). In the study by Harrington et al. (2008: 2829, their Figure 2), the realizations of /u/ both by the young female and young male speakers show considerable overlap with that of /i/. For this reason, the process of /u/-fronting in SESE is occasionally referred to as a phonetic merger (e.g. Uffmann, 2010). Perceptual support for a merger in progress comes from observations like the one by Collins and Mees (2008: 102) that “older-generation speakers sometimes interpret this new GOOSE vowel as FLEECE, and may even confuse pairs such as *two* – *tea*, *through* – *three*, etc.”.

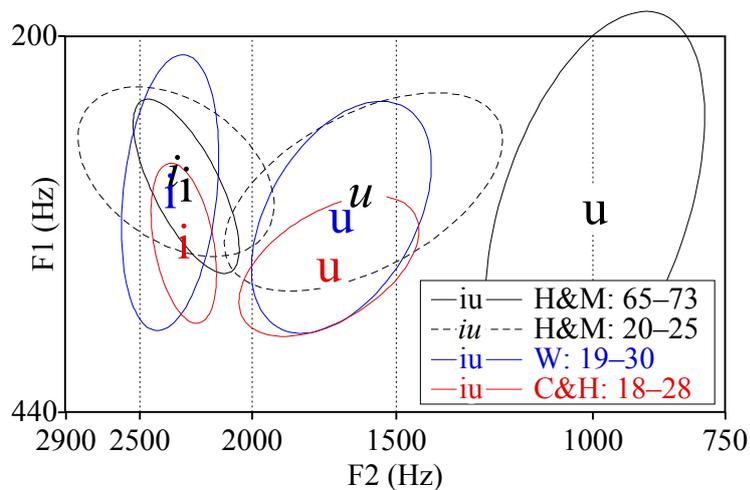


Figure 3.1: F1-F2 plot of /i/ and /u/ produced by male speakers of different ages. Symbols indicate means in different age groups and ellipses show 2 standard deviations. The figure shows data of the oldest and the youngest group from Hawkins and Midgley (2005, H&M, black), and the young male speakers from Chládková and Hamann (2011, C&H, red) and Williams (2013, W, blue). The figure also lists the age range of the speakers in each study.

¹ Despite its considerable phonetic fronting in younger speakers, we transcribe the GOOSE vowel as /u/ throughout this article.

Impressionistic phonetic descriptions of several varieties of British English over the last 50 years have been mentioning a slight diphthongization of the two tense high vowels, with /i/ sounding like [ii], [iɪ], or [əi], and /u/ like [ʊu], [ʊu:], [ʊʊ], or [ʊʊ]; see Wells (1962), Collins and Mees (2008), and Roach (2009: 20) for Southern-England English/RP; for varieties of British English other than SESE, see Stoddart et al. (1999) on the Sheffield dialect; Trudgill (1999) on Norwich; and Docherty and Foulkes (1999) on Derby and Newcastle. The first acoustic study supporting these observations was performed by Chládková and Hamann (2011), who analyzed young SESE speakers' productions of /i/ and /u/. They found that speakers acoustically differentiate /i/ and /u/ not only by the vowels' midpoint F2 and F3 values but also by the direction of F2 (and F3) diphthongization: the formants had a rising contour in /i/, but a falling contour in /u/, irrespective of the consonantal context in which the vowels were embedded. Given Chládková and Hamann's findings of diphthongization differences between /i/ and /u/ in production, it is plausible that diphthongization might also be an important cue to the /i/-/u/ difference in perception. The present study therefore tests whether diphthongization serves as a cue to the /i/-/u/ contrast in SESE listeners' perception. A consistent use of this cue would predict that /u/ is not going to merge with the front vowel /i/ because these two tense high vowels can be reliably distinguished by F2 diphthongization. See Figure 3.2A and B, which illustrates a possible re-definition of diphthongization as a new (or, additional) phonetic cue to the /i/-/u/ contrast.

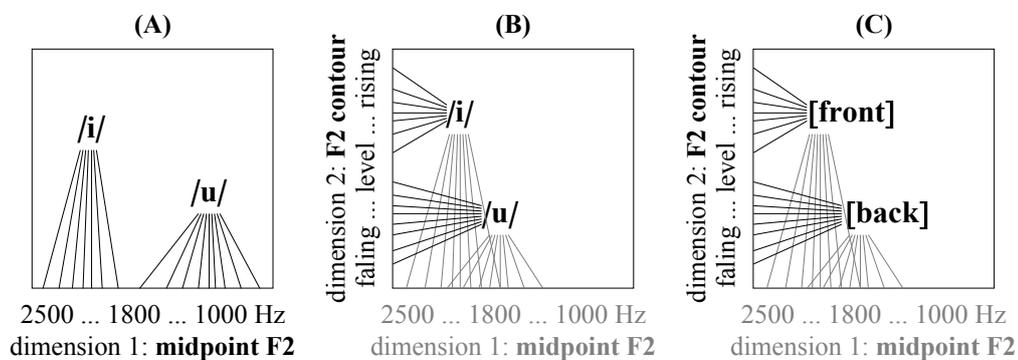


Figure 3.2: Re-association of a phonological contrast with a new phonetic cue. (A) shows a mapping between a single phonetic dimension (namely, midpoint F2) and a phoneme contrast (namely, /i/-/u/). In (B) the two phonemes are no longer reliably distinguished by midpoint F2, therefore, a new phonetic dimension (namely, F2 contour) becomes used as a cue to the phonological /i/-/u/ contrast. (C) visualizes a scenario in which phonetic cues are directly mapped onto (and thus also re-associated with) phonological feature categories.

As the process of /u/-fronting is a very recent change (and possibly still ongoing), Harrington et al. (2008) found a difference between young

and older listeners in the use of the midpoint F2 as perceptual cue: the /i/-/u/ boundary along this dimension was more fronted in young than in older listeners. On the basis of their findings one can expect a similar age-dependent difference for the use of diphthongization, with young listeners relying on diphthongization as a cue to distinguish the two vowels more heavily than older listeners. Support for this hypothesis comes from reports that only older listeners seem to confuse /u/ with /i/ in the speech of younger speakers (Collins and Mees, 2008). The present study therefore compares the use of diphthongization as a cue to the /i/-/u/ contrast in young and older listeners.

If the /i/-/u/ contrast is (at least partially) cued by diphthongization, it is plausible that diphthongization is employed as a perceptual cue to other front-back contrast as well. In that respect, results of various speech perception experiments suggest that listeners map the heard speech signal directly to phonological features (e.g. Chládková et al., ms; Kraljic and Samuel, 2006; Scharinger et al., 2011a). This suggests that cue re-association (e.g. after a sound change) might not be phoneme specific but might occur as a re-association of a new phonetic cue to a phonological feature, as illustrated in Figure 3.2C. The present study therefore also tests the follow-up hypothesis that if SESE listeners use diphthongization as a perceptual cue to the /i/-/u/ contrast, they might employ the same cue for other front-back contrasts, such as KIT vs. FOOT (i.e., /ɪ/-/ʊ/) or DRESS vs. THOUGHT (i.e., /ɛ/-/ɔ/).

The final point of interest in the present study is the influence of phonetic context on perceptual cues. Recent vowel production data of young speakers show that both /i/ and /u/ have a higher midpoint F2 in coronal than in non-coronal contexts, and that this effect is more pronounced for /u/ than for /i/ (Chládková and Hamann, 2011). Furthermore, the fronting effect of the coronal context on /u/ is larger in older than in young speakers (Harrington et al., 2008). This is because in non-coronal contexts /u/ is more retracted in older than in younger speakers, which means that in older speakers it can undergo a larger fronting shift triggered by coronal context. A coarticulatory effect like this is usually perceptually compensated for (see Harrington et al., 2008; Lindblom and Studdert-Kennedy, 1967; Mann and Repp, 1980). A study by Ohala and Feder (1994) tested the identification of /i/ and /u/ in VCə stimuli where the C was either /b/, /d/, or a fully masked /b/ or /d/ (white noise masking; no place cues remained). Their results showed a perceptual compensation for coarticulation, as the boundary between /i/ and /u/ was at higher F2 values for the /d/- than for the /b/-stimuli. Interestingly, stimuli with masked /b/ that were presented in the same block with the unmasked /d/-stimuli (and vice versa) triggered the same boundary shifting. This is because listeners seemed to have interpreted the context of the masked stimuli as being the same as those of the non-

masked stimuli in the same block, i.e. they interpreted masked /b/ as coronal. Ohala and Feder's findings suggest that an imagined context can trigger similar compensatory effects than a context that is acoustically present. Based on this suggestion, the present study tests whether a context that is only given in the orthography of the answer categories can trigger compensation for coarticulation, i.e. a shift of the /i/-/u/ boundary to higher F2 values when the orthography indicates a coronal context (compared to labial and dorsal contexts).

In summary, the present study tests the following hypotheses.

- (1) Diphthongization is used as a perceptual cue for the /i/-/u/ contrast in SESE.
- (2) There are age-specific differences in the use of diphthongization as a perceptual cue: older speakers show less or no use.
- (3) Diphthongization is also used as perceptual cue for other front-back contrasts in SESE.
- (4) Orthographically presented consonantal context triggers compensatory effects on the /i/-/u/ boundary (with coronals causing boundary fronting).

The present study consists of two experiments. Experiment 1 tests whether the direction of F2 diphthongization affects the location of the perceptual /i/-/u/ boundary (hypothesis 1), whether there is a difference between young and older listeners in their reliance on diphthongization (hypothesis 2), and whether orthographic information about consonantal context affects the /i/-/u/ boundary in perception (hypothesis 4). The follow-up Experiment 2 examines whether diphthongization serves as a cue to front-back phoneme contrasts other than /i/-/u/ (hypothesis 3).

3.2 EXPERIMENT 1

3.2.1 *Method*

3.2.1.1 *Stimuli*

The stimuli were synthetic vowels made with a Klatt synthesizer (Klatt and Klatt, 1990) built into the program Praat (Boersma and Weenink, 1992-2013). A single F2 continuum ranging from 1800 Hz to 3200 Hz was divided into 12 values equidistant on an Erb scale (step size = 0.43 Erb). Each of the 12 F2 values was synthesized with two durations: 181 and 200 ms; this was to render the stimulus set more variable and thus more realistic. All stimuli had a mid-point F1 of 330 Hz and a mid-point F3 of 2700 Hz. The stimuli were synthesized with three diphthongization types: rising, level, and falling. For 'level' stimuli, all formants were

stable throughout the duration of the vowel. For ‘rising’ stimuli, F2 and F3 rose linearly by 0.5 Erb from the beginning to the end of the vowel, while for ‘falling’ stimuli, F2 and F3 fell linearly by 0.5 Erb. Both ‘rising’ and ‘falling’ stimuli contained a linear 0.5-Erb fall in F1. The fundamental frequency (Fo) rose linearly from 230 Hz at the beginning of the vowel up to 275 Hz at 15% of the vowel’s duration and then decreased linearly to 175 Hz at the end of the vowel. There were in total 72 different stimuli: 12 F2 values \times 2 durations \times 3 diphthongization types. Figure 3.3 illustrates the three diphthongization types as well as the pitch contour of the stimuli.

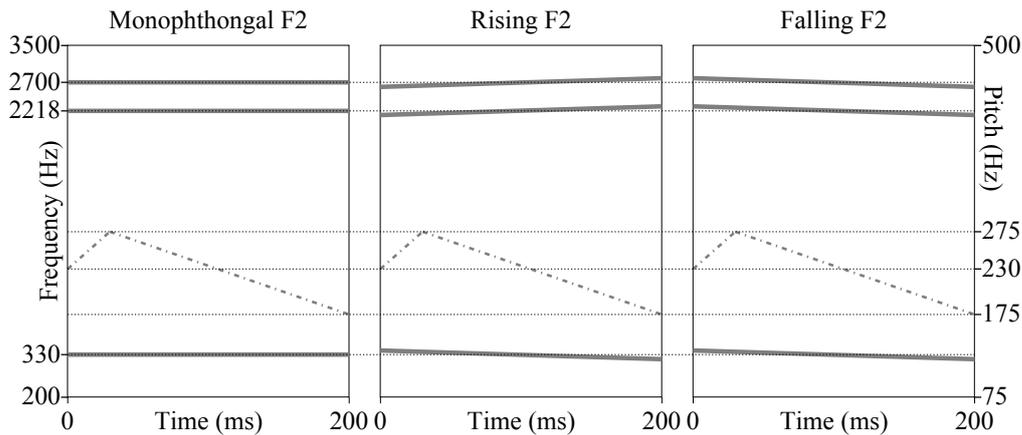


Figure 3.3: Illustration of stimuli from Experiment 1. The figure shows the three different diphthongization types for a stimulus with mid-point F2 value of 2218 Hz and duration of 200 ms: the grey solid lines represent the first three formants (left axis), and the dotted-dashed line shows the pitch contour (right axis).

3.2.1.2 Participants

Forty-two young speakers and twelve older speakers of SESE took part. The young speakers were university students between 18 and 33 years of age (mean age = 21.8; 16 male). They were tested at the University of Sheffield. Before coming to study in Sheffield, they had lived all their lives in the south of England and considered their dialect to be representative of that area. The young participants were randomly assigned to one of three groups according to which consonantal context they were tested with: labial ($n = 16$, mean age = 21.1, 7 male), coronal ($n = 14$, mean age = 22.4, 6 male), and dorsal ($n = 12$, mean age = 22.2, 3 male).

The older listeners were aged between 57 and 67 years (mean age = 63.2; 2 male). They were tested at their homes or work place: ten in London, and two in Royal Tunbridge Wells. All participants were healthy and reported normal hearing. Due to a limited number of recruited participants, older listeners were only tested with the coronal consonantal

context. The choice of coronal context was motivated by the following. In Harrington et al.'s (2008) experiment, the young and the older listeners' /i/-/u/ boundaries at midpoint F2 were shown to differ least in coronal context, i.e. both older and young listeners had fronted /u/ in coronal context. If diphthongization serves as a perceptual cue, it should do so especially when midpoint F2 becomes uninformative, i.e. in the coronal context for both young and older listeners.

3.2.1.3 Procedure

The experiment was a two-alternative forced-choice identification task. Participants were instructed that they would hear vowels cut from recordings of an English speaker, and they would have to identify which of two words the vowel came from. Depending on whether they were assigned to the coronal, labial, or dorsal context-group, participants' response options were *teed* and *tood*, *feeb* and *foob*, or *keeg* and *koog*, respectively. To ensure that participants were familiar with how the nonce words would sound in English, they were given written instructions that the words rhyme with *leap* and *loop*, respectively.²

The stimuli were presented in random order and there was no option of replaying the sound; if unsure, participants were asked to give their best guess. The experiment was preceded by a short practice round with 7 stimuli to ensure that participants understood the task.

Each trial started with a 400-ms silent interval, after which the stimulus was played. Participants were asked to listen to the whole sound, and then indicate their response by clicking on one of the two buttons on the computer screen (labeled as e.g. *teed* and *tood*). The whole randomized set of 72 stimuli was presented once to the older listeners, and twice to the young listeners. During the experiment, young participants could take two short breaks (after every 50th trial), and the older participants could take three breaks (after every 20th trial).

3.2.2 Results

For each of the 42 young and 12 old listeners, we ran binomial logistic regression models with vowel midpoint F2 as the regression factor and proportion /i/-responses as the dependent variable. The /i/-/u/ boundary is located at such a midpoint F2 value x that would receive the label

² In fact, *teed* (past tense; 'to place on a tee') and *feeb* (slang; 'a stupid person') do exist in English but are quite rare. To further ensure that participants in all groups regarded both of their response options as nonce words, the exact instructions were as follows: "We would like you to learn two new English words: *teed* and *tood* [or *feeb* and *foob*, or *keeg* and *koog*, depending on the group in which they were assigned]. Although they don't have a meaning in English, they could be English words because the *sound* English. (They rhyme with *leap* and *loop*.)"

/i/ with the probability of 0.5 (and, analogously, the label /u/ with the probability $1-0.5$):

$$\ln \frac{0.5}{1-0.5} = \beta_0 + \beta_1 x \quad (3.1)$$

where β_0 and β_1 are the logistic regression coefficients. Since $\ln \frac{0.5}{1-0.5} = 0$,

$$x = -\frac{\beta_0}{\beta_1} \quad (3.2)$$

The boundaries of the 42 young listeners were submitted to a repeated-measures analysis of variance (RM-ANOVA)³ with diphthongization type as the within-subjects factor (rising, level, falling) and orthographic context as the between-subjects factor (labial, coronal, dorsal). The analysis revealed a main effect of diphthongization type ($F[2,78] = 37.847, p < .001$). No significant main or interaction effects involving context were found.

Pairwise comparisons (Fisher's LSD) of the mean boundary locations across the three diphthongization types showed that the /i-/u/ boundary for stimuli with rising F2 contour was at lower F2 values than the boundary for stimuli with level F2, which in turn was at lower F2 values than the boundary for stimuli with falling F2 contour (rising-level: mean difference = -0.159 Erb, 95% confidence interval [c.i.] = $-0.228.. -0.089, p < .001$; level-falling: mean diff. = -0.205 Erb, c.i. = $-0.292.. -0.118, p < .001$; rising-falling: mean diff. = -0.364 Erb, c.i. = $-0.460... -0.268, p < .001$). Figure 3.4 (top graph) plots the logistic regression fit averaged across the 42 young listeners.

In line with previous studies that only compared the coronal context to one type of non-coronal (i.e., labial) context, we ran a second RM-ANOVA where the three consonantal contexts were re-coded into two levels of coronality: either coronal or non-coronal (= dorsal and labial collapsed). The analysis again yielded a main effect of diphthongization type ($F[2,80] = 33.729, p < .001$). This time, the effect of context also approached significance ($F[1,40] = 3.150, p = .084$): the /i-/u/ boundary for participants who had labels with coronal context was at higher F2 values than for those who had labels with non-coronal context (mean difference = 0.188 Erb, c.i. = $-0.026..0.402$).

To test for the effect of age, the boundaries of the 12 old and the 14 young listeners (who were tested with coronal context) were submitted to a third RM-ANOVA with diphthongization type as the within-subjects factor and age group as the between-subjects factor. The analysis revealed a main effect of diphthongization type ($F[2\varepsilon, 48\varepsilon, \varepsilon = .882] =$

³ In this and all subsequent analyses, if Mauchly's test of sphericity is not passed, we employ Huynh-Feldt's correction, which reduces the number of degrees of freedom by a factor ε .

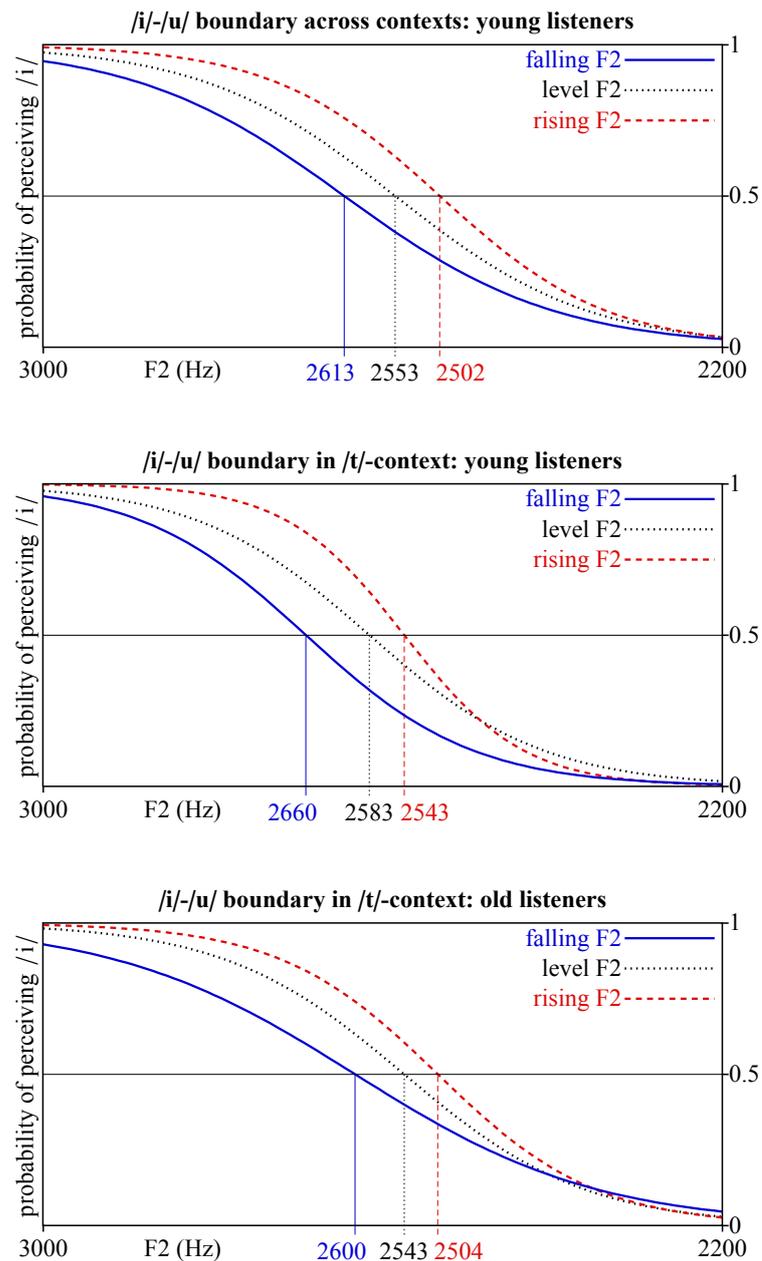


Figure 3.4: Experiment 1: perceptual /i-/u/ boundaries on the F2 dimension. Top graph: average over all three contexts across 42 young listeners; middle graph: /t/-context across 14 young listeners; bottom graph: /t/-context across 12 older listeners. Note that the graphs zoom in on an F2 range between 2200 and 3000 Hz but the stimulus continuum ranged from 1800 to 3200 Hz.

9.974, $p < .001$). There were no significant main or interaction effects involving age. Pairwise comparisons (Fisher's LSD) of the mean boundary locations across the three diphthongization types showed that the /i-/u/ boundary for stimuli with rising F2 contour was at lower F2 values than the boundary for stimuli with level F2, which in turn was at lower

F2 values than the boundary for stimuli with falling F2 contour (rising-level: mean difference = -0.123 Erb, c.i. = $-0.230.. - 0.016$, $p = .026$; level-falling: mean diff. = -0.203 Erb, c.i. = $-0.373.. - 0.033$, $p = .021$; rising-falling: mean diff. = -0.326 Erb, c.i. = $-0.496.. - 0.155$, $p = .001$). Figure 3.4 (middle and bottom graphs) plots the logistic regression fits of the 14 young and 12 old listeners in coronal context.

3.2.3 Discussion

Experiment 1 provides support for hypothesis 1 on the use of diphthongization: we found that the /i-/u/ boundary was at lower F2 values for stimuli with rising F2 contour than for stimuli with falling F2 contour, which implies that native speakers of SESE use diphthongization as perceptual cue to the /i-/u/ contrast. This finding is in line with the acoustic data by Chládková and Hamann (2011), in which young SESE speakers produced /i/ with a rising F2 contour and /u/ with a falling F2 contour.

With respect to hypothesis 2 on age-specific differences in the use of diphthongization, we did not find any difference between young and older listeners: both groups showed a similar influence of diphthongization on the perceptual boundary between /i/ and /u/. Two factors might be responsible for this similar behavior. First, the older listeners were only tested with the orthographically imposed coronal context, and their use of diphthongization might be context specific: In Harrington et al. (2008), the context influence on the perceptual boundary was smallest for coronals, and older listeners had a very front boundary in coronal context. Second, the stimuli modeled the voice of a young female speaker (i.e., they had a rather high F_0 with a pronounced rise-fall contour), therefore it is possible that the old listeners, on the basis of their linguistic experience, may have adapted their perception to the speech of a young speaker with overlap on midterm F2-values and employed diphthongization as secondary cue, which they would not employ when expecting the speech of an older speaker (see e.g. Drager, 2010, on the effect of expected speaker age on vowel perception). This proposal is, however, difficult to reconcile with the observations mentioned in the introduction that older speakers sometimes confuse /u/ with /i/ when listening to younger speakers.

Both of these factors remain to be tested in future work. If they can be excluded as possible explanations for the lack of differences in the use of the diphthongization cue by older and young listeners, we have to deduct that diphthongization has been available as a perceptual cue for a longer time than initially assumed, and that it did not emerge as a consequence of /u/-fronting.

An effect of orthographic consonantal context (hypothesis 4), which was tested for young listeners only, was not detected when we compared labial, coronal and velar place of articulation. However, following most previous studies that compared /u/-perception between a coronal and a single non-coronal context, we ran a second analysis in which we tested for effects of coronal versus non-coronal (labial and velar) context in the orthographic labels. In that comparison, we found a nearly significant effect of coronality: the /i/-/u/ boundary appeared to be more fronted in listeners who were presented with coronal labels than in listeners who were presented with non-coronal labels. This indicates that not only a context that is acoustically present (as e.g. in Harrington et al., 2008) but also a context that is only orthographically present might affect the /i/-/u/ perceptual boundary.

3.3 EXPERIMENT 2

To assess whether diphthongization is used as a cue to front-back contrasts in general (hypothesis 3) we carried out Experiment 2. Additionally, the design of Experiment 2 improved several aspects of Experiment 1. It was a vowel identification task with a more realistic design: stimuli were sampled from the whole vowel space (not just a single continuum), and the response labels consisted of the eleven possible English monophthongs (not just two vowels). Experiment 2 was run with young SESE speakers who have always lived in Kent, and were slightly younger than the young participants in Experiment 1. Experiment 2 thus investigated whether front-back contrasts other than /i/-/u/ are cued by diphthongization, and whether we can replicate the findings of Experiment 1 with a larger stimulus set, a larger number of response options, and a group of participants who are more homogenous with respect to linguistic experience and age.

3.3.1 *Method*

3.3.1.1 *Stimuli*

The stimuli were synthetic vowels sampled from the whole possible vowel space, with relatively more stimuli from the upper region of the vowel space. Figure 3.5 shows the F1-F2 stimulus grid. F1 and F2 were both sampled into 11 values equidistant on an Erb scale. F1 ranged from 300 to 1000 Hz (7.28 to 15.29 Erb, step size was 0.80 Erb), F2 ranged from 800 to 3300 Hz (13.59 to 25.07 Erb, step size was 1.15 Erb). We excluded F1-F2 combinations that are by definition impossible (when F1 would be above F2, i.e. the lower right corner of the vowel grid) or highly unlikely, frog-like sounding, speech sounds (high F1 values combined with high

F2 values, i.e. the lower left corner of the vowel grid). This procedure yielded 93 unique F1-F2 pairs. 55 F1-F2 pairs from the upper part of the vowel grid (outlined by the rectangle in Figure 3.5) were synthesized with two F3 values: 2200 Hz and 2800 Hz (21.72 and 23.72 Erb)⁴, and two durations: 245 ms and 181 ms. The remaining 38 F1-F2 pairs had an F3 of 2566 Hz (23 Erb) and duration of 211 ms. All stimuli contained the same pattern of F0 contour as the stimuli in Experiment 1.

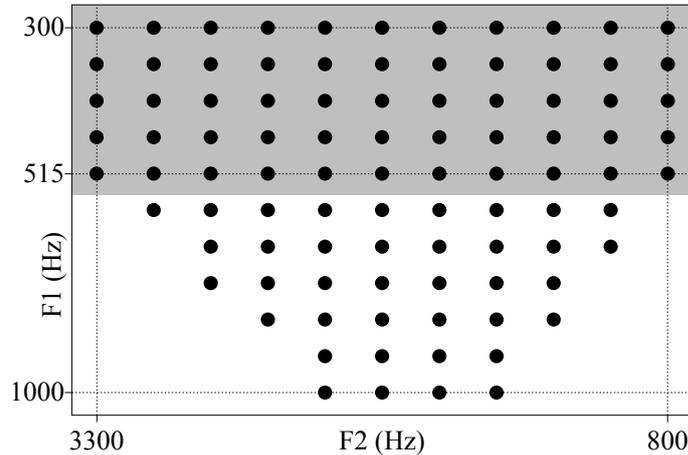


Figure 3.5: Experiment 2: the sampling of the F1-F2 stimulus space. The 55 F1-F2 pairs in the upper grey region were synthesized with two F3 values, two durations, and three diphthongization types (monophthongal, rising, falling). The remaining F1-F2 pairs from the lower region were synthesized with one F3 value, one duration, and one diphthongization type (monophthongal).

To test whether listeners rely on diphthongization as a cue to the front-back contrast among non-low vowels, we also varied the diphthongization of the 55 tokens in the upper part of the vowel grid. The upper 55 tokens were synthesized with three possible diphthongization values: monophthongal, rising and falling (similarly to Experiment 1). The 38 tokens from the lower part of the vowel grid were monophthongal.

Combining 55 F1-F2 values from the upper part of the vowel space with 2 F3 values, 2 durations, and 3 diphthongization types, and adding the 38 tokens from the lower part of the vowel space yielded 698 stimuli in total.

⁴ Note that the F2 ranged up to 3300 Hz. This means that for stimuli with high F2 values, the F2 in fact became an F3. Previous research has shown that when F2 and F3 are close, listeners perceptually integrate the acoustic F2 and F3 into the ‘effective F2’ or ‘F2 prime’ (Bladon, 1983; Delattre et al., 1952). Therefore, we analyze and plot the results as if F2 ranged from 800 to 3300 Hz. This F2 is meant to represent the perceptual F2 and not its actual acoustic value.

3.3.1.2 *Participants*

The participants were 49 young monolingual native speakers of SESE (different individuals from the subjects in Experiment 1). They were sixth-form high-school students between 17 and 19 years of age. At the time of testing, they had lived all their lives in Kent, UK. All but five participants (who were excluded) had been raised by monolingual SESE speakers. Two further participants had to be excluded because they did not complete the perception task. All participants were paid for taking part in the experiment.

3.3.1.3 *Procedure*

The experiment was a multiple forced-choice identification task. Participants had to identify every stimulus with one of 11 labels corresponding to nonce⁵ monosyllabic words each containing one of the 11 SESE monophthongal vowels /i ɪ ε æ ɜ ʌ ʊ ɒ ɔ ʊ u/. The words were presented orthographically on a computer screen as CeeC, CiC, CeC, CaC, CerC, CuC, CarC, CoC, CawC, CuCC, and CooC (the order corresponding to the 11 vowels listed above, with C = consonant). As in Experiment 1, the consonantal frames were fVb, tVd, and kVg (V = vowel). Participants were randomly assigned to one of three groups depending on the consonantal context in which the vowels were embedded (i.e., labial, coronal, or velar).

The 698 stimuli were presented one at a time in random order over headphones. Each trial started with a 1000-ms silence, after which a stimulus was played. Participants were asked to wait until the entire stimulus was played and then give their answer by clicking on one of the 11 buttons on the computer screen containing the 11 English nonce words. There was a 5-second break after every 88th stimulus; the fourth out of a total of 7 breaks was somewhat longer and participants could decide themselves when to resume the experiment. Participants were tested in small groups in a quiet computer room at the Charles Darwin School in Kent, UK.

Prior to the perception experiment, participants were presented with a printed list of their 11 answer categories together with a set of rhyming words embedded in a sentence. For instance, the text relevant for the /i/-word in the coronal-context group were: “**Teed** rhymes with feed and seek. In **teed** we have an ‘ee’. **Teed**.” The participants were asked to try to quietly learn the pronunciation of the 11 new words and were given

⁵ Some of the monosyllables do in fact represent words that exist in English. Since these are names, abbreviations, or rather infrequent words, we untruthfully told the participants that the words they were going to learn do not exist in English. We supposed this would further draw participants’ attention away from the possible existent meaning of these words. Therefore, we did not expect the possible-word status of some response labels to affect participants’ identification of the stimuli.

approximately 5 minutes for this task. They were told that the purpose of the subsequent listening experiment was to test how well they had learnt the pronunciation of these eleven new words.

3.3.2 Results

Figure 3.6 shows the labeling results pooled across the 42 participants. For each stimulus, the figure plots the vowel category that was chosen by the majority of participants (in case of a tie, both response categories are plotted).

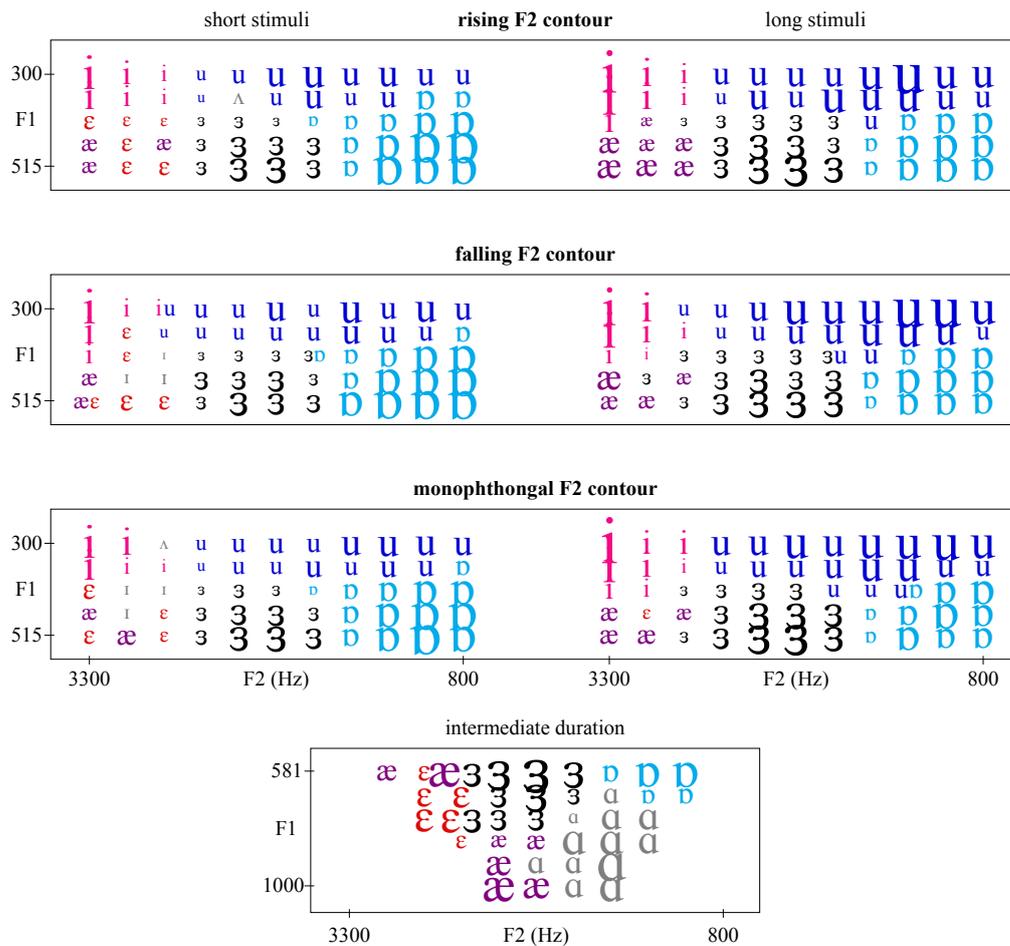


Figure 3.6: Experiment 2: response categories chosen for each stimulus (pooled across two different F₃ values). For each stimulus, the label that was given by the majority of participants is plotted: the larger the symbol the more participants chose that label (in case of a tie both labels are plotted). The F₁ and F₂ axes indicate formant values measured at the mid-point of the stimulus. Recall that for F₁ greater than 515 Hz, i.e. the lower part of the vowel space, the stimuli were monophthongal, all had one (intermediate) duration value, and one F₃ value.

As can be seen from Figure 3.6, three response categories were hardly ever used: /ɔ/, /ʊ/, and /ʌ/. Apparently, the labeling task with 11 labels was rather difficult and the young participants were not fully able to learn the spelling-sound mapping for *tawd*, *tudd*, and *tud*. The labeling patterns also show that subjects used the latter two labels interchangeably. Due to the lack of reliable /ʊ/ responses, we could not include the /ɪ/-/ʊ/ contrast in our analysis.

For stimuli from the upper vowel region (i.e., stimuli with a F1 between 300 and 515 Hz), we ran binomial logistic regression with midpoint F1 and F2 as the regression factors and proportion /i/-responses as the dependent variable. The /i/-/u/ boundary in the two-dimensional F1-F2 space runs through such F1-F2 value pairs, i.e. y and x values, that would receive the label /i/ with the probability of 0.5:

$$\ln \frac{0.5}{1-0.5} = \beta_0 + \beta_1 y + \beta_2 x \quad (3.3)$$

where β_0 , β_1 , and β_2 are the logistic regression coefficients, y is the value of F1 and x is the value of F2. We are further interested in the boundary location on the F2 axis for an intermediate F1 value (i.e., for the value of y halfway between 300 and 515 Hz along an Erb scale). Therefore, since $\ln \frac{0.5}{1-0.5} = 0$,

$$x = -\frac{\beta_0 + \beta_1 y}{\beta_2} = -\frac{\beta_0 + \beta_1 \cdot 8.88}{\beta_2} \quad (3.4)$$

The F2 locations of the boundaries were submitted to a RM-ANOVA with diphthongization type as the within-subjects factor with three levels (rising, falling, level), and context as the between-subjects factor with two levels (coronal, non-coronal)⁶. Boundaries that were found to lie below 0 Erb or above 30 Erb were excluded from the statistical analysis: this happened for one participant's boundary for the monophthongal stimuli, thus leaving us with /i/-/u/ boundary data from 41 participants (out of whom 18 had the coronal and 23 the non-coronal context). The ANOVA yielded a significant main effect of diphthongization ($F[2\varepsilon, 78\varepsilon, \varepsilon = 0.985] = 4.484, p = .015$). Pairwise comparisons showed that the F2 boundary was at significantly lower F2 values for stimuli with rising F2 contour than for stimuli with falling F2 contour (mean difference = 0.624 Erb, $p = .012$, c.i. = 0.144..1.105). The analysis did not detect any main or interaction effects involving context.

Although we were not able to assess boundary locations for the /ɪ/-/ʊ/ contrast (possibly due to the confusion of the /ʊ/ and /ʌ/ labels),

⁶ Since Experiment 1 found a nearly-significant difference between coronal vs. non-coronal context, the analyses of Experiment 2 tested for the effect of context with two levels: coronal vs. non-coronal.

the data provide us with other front-back contrasts for which the boundary can be reliably determined. Figure 3.6 suggests that, apart from /i/ and /u/, stimuli from the upper region of the vowel space (i.e. they grey area of Figure 3.5) were often labeled as /æ/, /ɛ/, /ɜ/, and /ɒ/. In SESE, the vowels /æ/ and /ɛ/ are front, /ɒ/ back, and /ɜ/ central (see e.g. Roach, 2009). Thus, to further examine whether diphthongization serves as a cue to a front-back contrast in general, we ran a binomial logistic regression for the two remaining front-back contrasts in our data: /æ/-/ɒ/ and /ɛ/-/ɒ/. Note that for /ɛ/-/ɒ/ in one subject and for /æ/-/ɒ/ in nine subjects there were not enough of the respective vowel responses to fit the logistic regression. From the regression coefficients we again computed, per participant, the location of the /æ/-/ɒ/ and /ɛ/-/ɒ/ boundaries for each diphthongization type. As with /i/-/u/, boundaries below 0 Erb or above 30 Erb were excluded from further analyses. We thus had boundary data for all three contrasts from 32 subjects, out of whom 14 had coronal and 18 had non-coronal labels. We submitted the /æ/-/ɒ/ and /ɛ/-/ɒ/ boundaries together with the /i/-/u/ boundaries to a second RM-ANOVA with diphthongization type and vowel contrast as the within-subjects factors with three levels each (i.e. diphthongization: rising, falling, level; vowel contrast: /i/-/u/, /æ/-/ɒ/, and /ɛ/-/ɒ/), and context as the between-subjects factor with two levels (coronal, non-coronal).

The ANOVA yielded a main effect of vowel contrast ($F[2, 60] = 15.884, p < .001$) and a main effect of diphthongization type ($F[2, 60] = 5.325, p = .007$). The analysis did not detect a significant interaction between vowel contrast and diphthongization type, nor any effects involving the between-subjects factor context. The main effect of vowel contrast indicates that, unsurprisingly, the F2 boundary differed across the 3 vowel pairs. Pairwise comparisons of the means showed that the /ɛ/-/ɒ/ boundary was at lower F2 values than the /æ/-/ɒ/ boundary, which was in turn at lower F2 values than the /i/-/u/ boundary (/ɛ/-/ɒ/ vs. /æ/-/ɒ/: mean difference = -0.870 Erb, c.i. = $-1.355.. -0.384, p = .001$; /æ/-/ɒ/ vs. /i/-/u/: mean diff. = -0.642 Erb, c.i. = $-1.236.. -0.047, p = .035$; /ɛ/-/ɒ/ vs. /i/-/u/: mean diff. = -1.511 Erb, c.i. = $-2.075.. -0.948, p < .001$). As for the main effect of diphthongization type, pairwise comparisons showed that the boundary for stimuli with rising F2 contour was at significantly lower F2 values than the boundary for stimuli with falling F2 contour (mean diff. = -0.535 Erb, c.i. = $-0.846.. -0.224, p = .001$). Figure 3.7 plots the /i/-/u/, /æ/-/ɒ/, and /ɛ/-/ɒ/ boundaries for each diphthongization type.

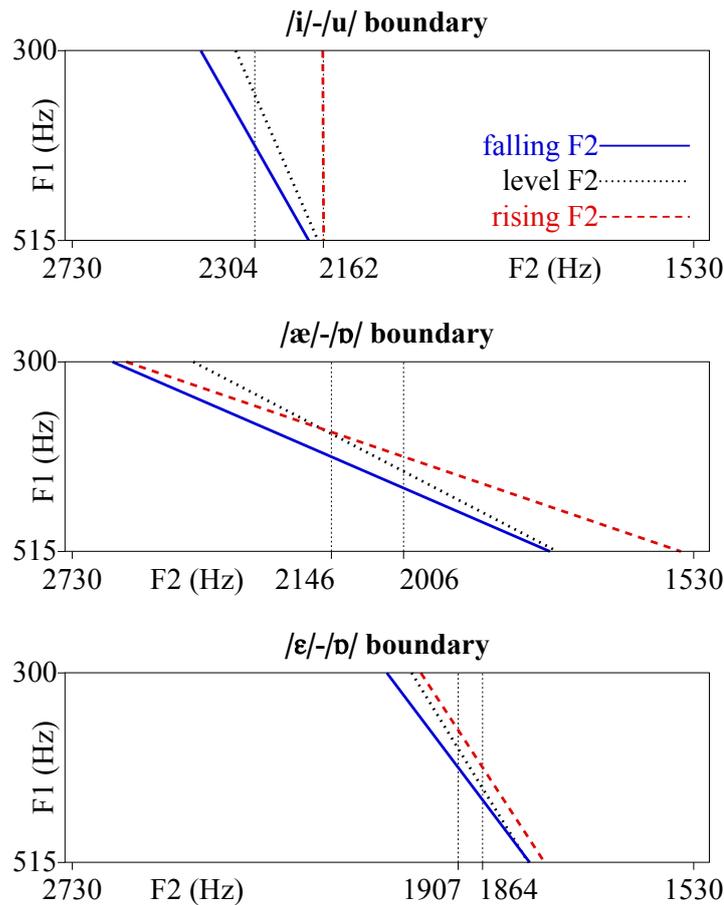


Figure 3.7: Experiment 2: perceptual front-back phoneme boundaries in the F1-F2 space; collapsed over consonantal contexts. Top graph: /i-/u/ boundary; middle graph: /æ-/ɒ/ boundary; bottom graph: /ɛ-/ɒ/ boundary.

3.3.3 Discussion

The findings of Experiment 2 replicated those of Experiment 1 in that the /i-/u/ boundary was affected by the F2 diphthongization of the stimuli (hypothesis 1): the boundary was at lower F2 values for stimuli with rising F2 than for stimuli with falling F2. The results of Experiment 2 further suggest that diphthongization affects boundary location in two other contrasts, namely /æ-/ɒ/ and /ɛ-/ɒ/, in a similar way as it does in the /i-/u/ contrast. This indicates that (at least) in young SESE listeners, front-back contrasts other than /i-/u/ are also perceptually cued by diphthongization (hypothesis 3).

Experiment 2 did not detect a main effect of context on the perceived /i-/u/ boundary (hypothesis 4). This suggests that the nearly significant effect in Experiment 1 might either have been due to chance, or failed to be manifested in Experiment 2 due to the demanding nature of the task. Specifically, the requirement to memorize spelling-sound mappings

for 11 new words in a short time may have interfered with the listeners' ability to perceptually compensate according to context. This speculation needs further research that would compare the effect of orthographically presented context in tasks with varying degrees of complexity.

Though unrelated to our research questions, we would like to report on the unexpected finding that stimuli with high F2 values and rather low F1 values (i.e. the space that is occupied by the vowel / ϵ /) were labeled / \ae /, as can be seen in Figure 3.6. This effect is even stronger for the long stimuli. Given the fact that the vowel / \ae / has been reported to shift towards an [a]-like quality in the production of young SESE speakers (de Jong et al., 2007; Gimson, 2001, see also Harrington, 2007, for one older speaker), it is rather surprising that our listeners labeled stimuli with very low F1 values as / \ae /. We speculate that the unexpected labeling happened because participants consider these stimuli as being too long for an / ϵ /, and / \ae / is the only front vowel that is slightly longer in duration. This speculation seems to be supported by recent studies on SESE vowel production and perception: / \ae / is produced with 1.2 times longer duration than / ϵ / in male speakers (Williams, 2013: Table 4.3; although no such difference has been found for female speakers), and listeners' perceptual judgments show that the best perceptual exemplar of / \ae / is 1.33 times longer than that of / ϵ / (Evans and Iverson, 2004: Table II).

Finally, Experiment 2 demonstrated that a vowel identification task with stimuli from the whole vowel space, and with labels for 11 'new' English words is rather demanding, which may be the reason why our participants failed to reliably use some of the response labels.

3.4 GENERAL DISCUSSION AND CONCLUSIONS

The two experiments reported in this paper demonstrate that diphthongization is a perceptual cue to the /i/-/u/ distinction in SESE. When classifying vowels modeled after a young female voice, adolescents, young adults, and older adults use diphthongization in a similar way: a vowel with ambiguous midpoint F2 is perceived as /i/ if it has a rising F2 contour and as /u/ if it has a falling F2 contour. In other words, in the traditional F1-F2 vowel space, the /i/-/u/ boundary for stimuli with falling F2 contour is more front than for stimuli with rising F2 contour.

Experiment 2 indicated that diphthongization may not be specific to the /i/-/u/ contrast but may be a perceptual cue to a more general front-back contrast: no difference was found across /i/-/u/, / \ae /-/ɒ/, and / ϵ /-/ɒ/ with respect to the effect of diphthongization on boundary location. Therefore, we propose that SESE speakers might have learned to associate a rising F2 contour with front vowels or a feature such as [+front] and a falling F2 contour with non-front vowels or a feature such as [-

front]. Note that we do not claim that listeners no longer use midpoint F₂; both F₂ contour and midpoint F₂ may serve as perceptual cues to vowel frontness in SESE.

If in SESE the direction of diphthongization signals whether a vowel is [+front] or [-front], one would expect any front vowel to be realized with rising F₂ contour and any back vowel with falling F₂ contour. In the vowel production data collected by Williams (2013), we do not observe such a clear pattern of rising F₂ in front vowels and falling F₂ in back vowels: only /i/ has a rising F₂ contour whereas all other monophthongs seem to have falling F₂.⁷ We speculate that diphthongization could still be a perceptual cue to vowel frontness even if it is not manifested in production yet. That is, the mapping between the phonetic cue of diphthongization and the phonological /i/-/u/ contrast is easily generalized to other non-low vowel contrasts in perception, even if it is not consistently used in the production (yet).

Experiment 1 demonstrated that diphthongization affected the /i/-/u/ boundary in both age groups, i.e. older listeners employed this perceptual cue to the /i/-/u/ contrast in a similar way as young listeners (though this might be restricted to the tested coronal context and the voice of a young speaker). This indicates that diphthongization has been systematically present in the production of the high tense vowels in SESE for some time and prior to the emergence of /u/-fronting. Evidence for a longer and consistent presence of diphthongization in SESE can also be seen in the fact that young listeners readily generalized this cue to vowels other than /i/ and /u/ in Experiment 2.

The existence of the secondary perceptual cue of diphthongization might even have triggered /u/-fronting: the presence of a distinguishing secondary cue could have allowed an allophonic split of /u/ with an allophone with high midpoint F₂ values in coronal context (as documented in the data by Harrington et al., 2008, for the older generation) and a subsequent shift of all /u/ realizations to a high midpoint F₂, without the danger of perceptual confusion or merger. This proposed diachronic development therefore provides a supplement to previous phonetic proposals on the emergence of SESE /u/-fronting that refer to factors such as articulatory ease (Harrington et al., 2011a,b), a prevalence for /u/ to occur post-coronally (Harrington, 2007; Harrington et al., 2008), and a failure of the younger generation to compensate for coarticulation (Harrington et al., 2008, based on Ohala's, 1981, hypocorrection account).

The addition of diphthongization as perceptual cue to vowel contrasts is not unique to the case described here, as it can be seen in the develop-

⁷ Interestingly, however, the front vs. back distinction realized by rising vs. falling F₂ contour appears to be valid in Sheffield English vowel production (Williams, 2013): the front vowels /i/, /ɪ/, and /ɛ/ have rising F₂ values whereas the back vowels /u/, /ʊ/, /ɔ/, /ɒ/, and /ɑ/ have falling F₂ contour; /æ/ does not fit the pattern as it is a front vowel but has a falling F₂.

ment from high tense vowels to diphthongized vowels (and eventually to diphthongs) in Middle English as part of the Great Vowel Shift (Jespersen, 1909; Stockwell, 2002).

3.5 ACKNOWLEDGEMENTS

The research reported in this study was funded by the Netherlands Organization for Scientific Research (NWO) grant 277-70-008 awarded to Paul Boersma. We would like to thank Paul Boersma for comments on experiment design and previous versions of the manuscript, and for providing the funding. We are grateful to the Charles Darwin School in Biggin Hill, Kent, and particularly to Jill Green, for kindly hosting us to run our Experiment 2, for allowing us to use their multimedia facilities and the opportunity to recruit participants. We would like to thank Šárka Šimáčková for comments on experiment design, Sam Hellmuth and Mary Pearce for testing some of the older participants, Clara Martín Sánchez for assistance in testing, and Carmen Lie-La Huerta for providing us with testing equipment. Part of the results were presented at the 2nd *Workshop on Sound Change* in Kloster Seeon (May 2, 2012) and at the 13th *Conference on Laboratory Phonology* in Stuttgart (July 27, 2012), and we would like to thank the audience, especially Mary Beckman and James Kirby, for comments.

PERCEPTUAL SENSITIVITY TO CHANGES IN VOWEL DURATION REVEALS THE STATUS OF THE PHONOLOGICAL LENGTH FEATURE

4.1 PRE-ATTENTIVE SENSITIVITY TO VOWEL DURATION REVEALS NATIVE PHONOLOGY AND PREDICTS LEARNING OF SECOND-LANGUAGE SOUNDS

This section has been published as:

Kateřina Chládková, Paola Escudero, & Silvia Lipski (2013). Pre-attentive sensitivity to vowel duration reveals native phonology and predicts learning of second-language sounds. Brain and Language, 126 (3): 243-252.

Abstract

In some languages (e.g. Czech), changes in vowel duration affect word meaning, while in others (e.g. Spanish) they do not. Yet for other languages (e.g. Dutch), the linguistic role of vowel duration remains unclear. To reveal whether Dutch represents vowel length in its phonology, we compared auditory pre-attentive duration processing in native and non-native vowels across Dutch, Czech, and Spanish. Dutch duration sensitivity patterned with Czech but was larger than Spanish in the native vowel, while it was smaller than Czech and Spanish in the non-native vowel. An interpretation of these findings suggests that in Dutch, duration is used phonemically but it might be relevant for the identity of certain native vowels only. Furthermore, the finding that Spanish listeners are more sensitive to duration in non-native than in native vowels indicates that a lack of duration differences in one's native language could be beneficial for second-language learning.

4.1.1 Introduction

Languages differ in their phonemic inventories, that is, in the number of speech sounds that can distinguish word meaning. For instance, the English phonemic inventory includes the two vowels of "sheep" and "ship", namely /i/ and /ɪ/, while Spanish only has /i/. All languages have vowel phonemes that are distinguished in terms of their quality (Crothers, 1978; Maddieson, 1984), as measured by the position of the tongue and jaw or by the acoustic spectral properties of the vowel. How-

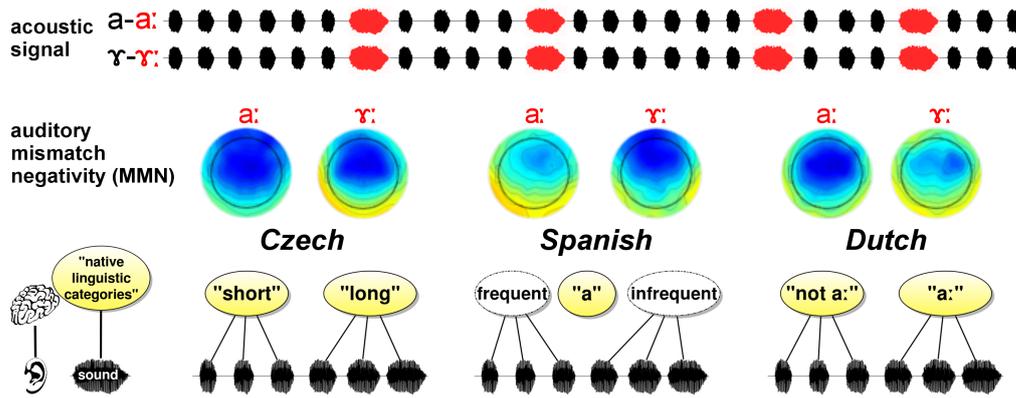


Figure 4.1: Graphical abstract.

ever, not all languages distinguish vowel quantity (also known as phonological vowel length), as measured by the duration of the vowel. In quantity languages such as Czech, vowel length is encoded in the phonology so that replacing a long vowel with a short one leads to a change in word meaning such as in the Czech words /sa:t/ ‘to suck’ and /sat/ ‘orchard’, which are only distinguished by the duration of the vowel. Spanish, on the other hand, is not a quantity language and its phonology does not encode vowel length so that whether the first vowel in the Spanish word /kasa/ ‘house’ is long or short does not change its meaning.

It is unclear whether the Dutch language encodes the acoustic dimension of vowel duration as phonological vowel length¹, in other words, whether this language has discrete long and short vowel categories. In his analysis of the Dutch vowel inventory, Moulton (1962) differentiates phonologically short and phonologically long vowels. Similarly, Zonneveld (1993) posits that vowel length is part of the Dutch native phonology. Booij (1995) argues that the vowel /a:/ equals to two units of the vowel /a/ within a syllable. Although Booij claims that vowel length as such is not a phonemic property of Dutch vowels, the proposal of a ‘doubling’ of an otherwise phonologically identical unit within a syllable implies that language users should have some representation of quantity in their grammar. van Oostendorp (1995) presents several arguments against phonological vowel length in Dutch and argues that the phonological property of vowel tenseness better accounts for Dutch vowel phonology. Most recently, Botma and van Oostendorp (2012) argue that the Dutch phonology does not at all distinguish between tense and lax (or, long and short) vowel segments, but that the phonetic (i.e. durational) differences between Dutch vowels are due to the structure of the syllable in which a vowel occurs. Botma and van Oostendorp dis-

¹ Here we use the term *duration* when we refer to the acoustic dimension, i.e. the phonetic property of the sound, while we use the term *length* when we refer to the abstract linguistic category, i.e. the phonological and contrastive mental representation.

cuss a large number of phonological studies with opposing views, which suggests that the long-lasting phonological debate has not yet led to a consensus on whether or not Dutch has vowel length.

Phonetic studies on Dutch do not clarify the issue of Dutch vowel length either. In that respect, recent speech production studies (Adank et al., 2004, 2007; van Leussen et al., 2011) show that Dutch speakers only use differences in vowel duration to distinguish a limited number of vowels, that is, the use of duration in speech production is inconsistent across vowels. Interestingly, however, speech perception studies suggest that vowel length in Dutch, as compared to English, does have a contrastive role. For instance, two recent studies (Dietrich et al., 2007; van der Feest and Swingley, 2011) have found that Dutch 18-month-olds and adults are more sensitive to differences in vowel duration than their English counterparts, which the authors attributed to the contrastive role of vowel length in Dutch as opposed to English. Importantly, some studies show that Dutch listeners use vowel duration to distinguish the vowels in the words /man/ 'man' and /ma:n/ 'moon' (Nooteboom and Doodeman, 1980), while others demonstrate that Dutch listeners predominantly use vowel spectral properties to distinguish these vowels (Escudero et al., 2009).

The present study aims to resolve the controversy around the abstract phonological representation of vowel length in the Dutch language. We examined Dutch listeners' pre-attentive processing of vowel duration changes, and compared it to that of Czech listeners, who clearly have short and long vowel phonemes, and to Spanish listeners, whose native phonology treats all vowel durations as equal. Listeners were presented with duration changes in both native and non-native vowels, which enabled the investigation of whether vowel duration processing depends on the listeners' phonemic inventory.

We recorded behavior-independent responses of the auditory system to vowel duration in a categorical oddball-paradigm using electroencephalography (EEG), examining the mismatch negativity (MMN). The MMN is elicited at about 100-250 ms latency when infrequent deviations occur among frequently repeated sound patterns. The MMN is widely accounted as a marker of pre-attentive change detection and is obtained for simple and complex patterns of auditory changes (Näätänen et al., 2007, 2001). What makes the MMN ideally suited for the present investigation of phonological representations is its sensitivity to listeners' linguistic experience: native phonemic contrasts elicit a stronger and often earlier MMN than speech sound contrasts without relation to the listeners' phonology. Crucially, many studies have demonstrated that listeners' native phonology modulates the pre-attentive processing of acoustic information (Hisagi et al., 2010; Kazanina et al., 2006; Kirmse et al., 2008; Lipski and Mathiak, 2007; Menning et al., 2002; Näätänen

et al., 1997; Nenonen et al., 2003; Sharma and Dorman, 2000; Tervaniemi et al., 2006; Ylinen et al., 2006). As for vowel duration, it has been shown that speakers of quantity languages such as Czech or Finnish, that is, languages which represent vowel duration in terms of abstract phonological categories, have stronger mismatch responses to vowel duration changes than speakers of other languages, including non-quantity languages such as Spanish or Russian (Hisagi et al., 2010; Kirmse et al., 2008; Menning et al., 2002; Nenonen et al., 2003, 2005; Tervaniemi et al., 2006; Ylinen et al., 2006). Unlike these previous studies, our three-way comparison of pre-attentive processing of vowel duration in Dutch, Czech and Spanish listeners will unravel the phonology underlying Dutch listeners' perception. Specifically, we will be able to show whether Dutch encodes vowel duration in terms of discrete short and long categories, in other words, whether Dutch is like quantity languages such as Czech, or whether it is like non-quantity languages such as Spanish.

Incidentally, even if vowel duration is not encoded in the phonology of a certain language, native speakers of that language tend to rely on duration to distinguish novel vowels that are present in a foreign language. Using behavioral tasks, a number of studies have shown that Spanish, Catalan, Portuguese, Mandarin, Polish, and Russian learners distinguish English or Dutch vowels through their duration differences, while native listeners predominantly use the vowels' spectral differences (Bogacka, 2004; Cebrian, 2006; Escudero et al., 2009; Escudero and Boersma, 2004; Flege et al., 1997; Kondaurova and Francis, 2008; Rauber et al., 2005). One explanation for second language learners' reliance on duration states that duration is acoustically highly salient and, therefore, universally accessible to learners regardless of its status in their native phonology (Bohn, 1995). An alternative explanation holds that the processing of duration is always transferred from the learner's native phonology (Escudero and Boersma, 2004). That is, listeners whose language does not employ vowel duration transfer a blank slate for this dimension, which allows them to readily form new length categories in a novel language (Escudero and Boersma, 2004).

Neurophysiological studies have also shown pre-attentive reliance on duration despite its irrelevance in the listeners' native phonology. In Lipski et al. (2012) Spanish learners of Dutch and Dutch natives had similar MMN responses to vowel duration changes for the Dutch vowels /a:/ and /ɑ/. Interestingly, however, speakers of non-quantity languages such as Spanish or Russian seem to process duration changes depending on how close a novel vowel is to their native vowel inventory. Nenonen et al.'s (2005) Russian learners of Finnish had smaller MMNs for vowel duration differences than Finnish natives when they were presented with stimuli that resembled a Russian vowel, while the two groups had similar MMNs for stimuli that did not resemble any Rus-

sian vowel. The authors attributed the Russian learners' strong MMN for duration differences in non-native vowels to the fact that they had successfully acquired second-language length categories after considerable exposure to Finnish. Since most previous studies have considered second language learners, it is unclear whether speakers of non-quantity languages such as Russian or Spanish have pre-attentive sensitivity to vowel duration when first exposed to the non-native length contrasts.

Given Nenonen et al.'s (2005) surprising results, the present study aimed at demonstrating whether pre-attentive processing of non-native vowel duration differences is *universal*, that is, independent of how vowel duration is encoded in the listener's native phonology, or *language-specific*, that is, dependent on its encoding within the listener's native phonology. To this end, we presented native and non-native vowels with different durations to Czech, Dutch, and Spanish listeners whose native phonologies are likely to differ on how they encode vowel length. Figure 4.2 shows the quality properties (first and second formant frequencies) of the native and non-native vowel stimuli used in the present study together with those of the Czech, Spanish and Dutch vowel inventories. It can be observed that [a] (the native vowel quality) falls within /a/ in all three languages, while [ɤ] (the non-native vowel quality) is far from any of the listeners' native vowels.

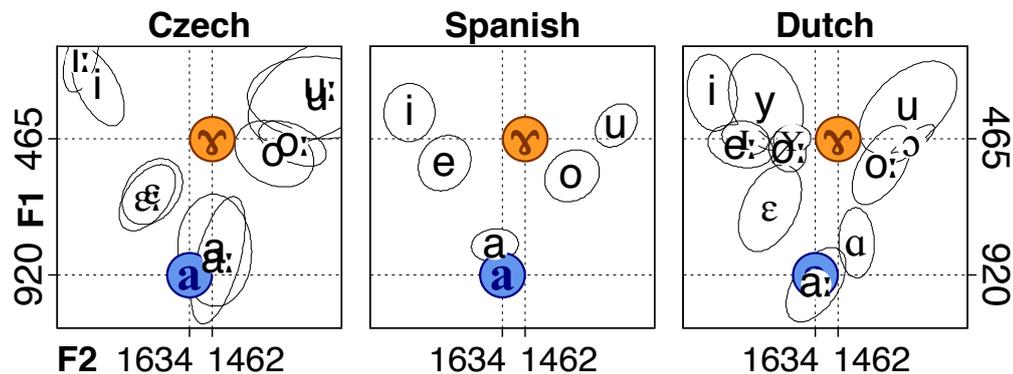


Figure 4.2: F1 and F2 plot of the two vowels produced by a female Estonian speaker that served as stimuli (native vowel quality = blue filled circle; non-native vowel quality = orange filled circle), and the female vowel inventories of the participants' native languages, specifically their native dialects: Moravian Czech (Šimáčková et al., 2012), Iberian Spanish (Chládková et al., 2011), and Randstad Dutch (van Leussen et al., 2011). Symbols represent the mean value of the population, ellipses show 2 standard deviations. Marks are in Hz, axes are scaled in Erb. The quality of the native stimulus resembles the native (long or short) phoneme /a(:)/ in all three participant languages, while the non-native stimulus does not resemble any phoneme in any of the three participant languages.

We tested Czech, Dutch, and Spanish monolingual young adults with very little experience in foreign languages. EEG was recorded in two sessions that took place on different days. In one session, participants listened passively to short and long tokens of the native vowel [a], and in the other, they listened passively to short and long tokens of the non-native vowel [ɤ]. As with Finnish and Russian listeners (Nenonen et al., 2005), if the auditory cortical processing of vowel duration is modulated by whether or not vowel duration can contrast native vowel phonemes, Czech and Spanish listeners should show opposite MMN responses for the native vowel quality. Specifically, Czech listeners should exhibit the largest MMN, while Spanish listeners will have no or the smallest MMN response for [a]. Dutch listeners, for whom the phonological role of vowel duration is unclear, may behave similarly to either the Czech or the Spanish listeners, or, alternatively, show an MMN response that is intermediate between the Czech and Spanish responses.

If non-native perception of vowel duration is phonology-specific and not universal, duration differences in the non-native vowel [ɤ] will elicit the largest MMN in Czech listeners. If the Spanish listeners transfer their native disregarding of duration differences to non-native perception, they will, again, have the smallest MMN in the non-native vowel. Alternatively, based on the findings of numerous L2 perception studies discussed above, duration differences in the non-native vowel [ɤ] may well elicit a large MMN in the Spanish listeners, one comparable to that of quantity language listeners. The latter would demonstrate that not only advanced Russian learners of Finnish (Nenonen et al., 2005), but also non-native listeners with little exposure to a novel language exhibit rapid pre-attentive sensitivity to vowel duration. Crucially, if Dutch phonology encodes vowel duration in terms of abstract length categories as it is in quantity-languages, Dutch listeners will resemble Czechs, and will thus have a large MMN for duration changes in the non-native vowel. If Dutch does not encode vowel duration in its phonology at all, Dutch listeners will resemble the Spanish in the non-native vowel.

Alternatively, if non-native perception of vowel duration is modulated by the universal salience of this acoustic dimension, all three groups should have an equally large MMN for the non-native vowel quality.

4.1.2 *Methods*

4.1.2.1 *Participants*

24 Czech, 24 Spanish, and 26 Dutch right-handed listeners, all university students or recent graduates aged 19 to 31 years, took part in the study. They were all monolinguals, who were raised in a monolingual family, had never spent more than 2 months in a foreign country, and had not had exposure to foreign languages above the level of high-school

classroom instruction. They rated their knowledge of foreign languages below 4 on a scale from 0 to 7 (where 0 means none, and 7 native-like), and were not linguistics students. None of the participants reported to have had a history of neurological, hearing, or language-related disorders. The Czech participants were from central and southern Moravia in the Czech Republic. The Dutch participants were from the Randstad area in the Netherlands. The Spanish participants were from various regions in Spain. 23 Czech (13 female, mean age = 22.4 years), 22 Spanish (10 female, mean age = 23.0 years) and 24 Dutch (13 female, mean age = 22.6 years) were included in the ERP analysis. Five participants were excluded due to a large number of artifacts (one Spanish and one Dutch), technical errors during data acquisition (one Czech and one Spanish) and ambidexterity revealed after the experiment (one Dutch). Participants gave a written informed consent and were paid for participation. The study was approved by the ethical committee of the Faculty of Humanities, University of Amsterdam and conforms to the guidelines of the Declaration of Helsinki (2008).

4.1.2.2 *Stimuli*

VOWEL QUALITIES AND DURATION STEPS The stimuli were natural tokens of the Estonian vowels /æ/ and /ɤ/ (henceforth transcribed as [a] and [ɤ], respectively), spoken by a 26-year old native female speaker of standard Estonian, a trained phonetician. The values of the first three formants were 920 Hz, 1634 Hz and 2707 Hz for [a], and 465 Hz, 1462 Hz, and 2920 Hz for [ɤ]. As shown in Figure 4.2, the quality of [a] is acoustically close to that of the participants' native vowel category (/a/ in Spanish and Czech, and /a:/ in Dutch and Czech), while the quality of [ɤ] is not close to any vowel in the participants' native languages. Since cross-language acoustic similarity of vowels is a good predictor of their *perceived* similarity (e.g. Chládková and Podlipský, 2012; Escudero and Chládková, 2010; Escudero et al., 2012; Escudero and Vasiliev, 2011; Escudero and Williams, 2011, 2012), [a] was used as the native vowel quality, and [ɤ] was used as the non-native vowel quality. The two vowels were produced with a flat pitch contour. Formant on-glides and off-glides were discarded so that the duration of the middle vowel portion with stable formants and pitch was 351 ms for [a] and 349 ms for [ɤ]. These tokens were subsequently manipulated using the time-domain pitch-synchronous overlap-and-add algorithm (Moulines and Charpentier, 1990) implemented in the software Praat (Boersma and Weenink, 1992-2013) to yield 6 different durations in psychoacoustically equal steps: 118, 136, 157, 181, 208, and 239 ms; that is, the six stimuli of each of the two vowel qualities differed only in their duration. These 6 duration values were selected on the basis of a pilot behavioral experiment with Czech listeners and the literature on Dutch vowels (described in

the paragraph below) so that the three short tokens in the present study are in the short category and the three long tokens are in the long category. The stimuli were presented in the categorical oddball paradigm (i.e. many-to-many oddball paradigm; Hisagi et al., 2010; Lipski et al., 2012; Scharinger et al., 2011c) described in Section 4.1.2.3. The three shortest items served as the short stimulus category while the three longest items served as the long stimulus category. This paradigm elicits an MMN if listeners perceive the acoustically varied standard stimuli as different from the acoustically varied deviant stimuli.

BEHAVIORAL PILOT EXPERIMENT: DETERMINING THE SHORT-LONG STIMULUS BOUNDARY The pilot experiment to determine the short-long boundary was a two-alternative forced-choice identification task. Only Czech listeners participated in this experiment because they are the only group who has explicit labels for the short and long vowel categories. We tested twenty-five young native speakers of Moravian Czech, who did not participate in the EEG experiment. Three stimulus duration continua were created: one with the quality of a high-mid back unrounded vowel [ɤ], one of a low front unrounded [a] and one of a low-mid back rounded [ɑ]. The first and the second vowel quality served as the non-native and the native stimulus quality, respectively, in the EEG experiment. The three different qualities from distinct vowel space regions were used to determine a general short-long boundary that applies across the vowel space. Each continuum ranged from 95 to 245 ms and consisted of 13 duration values equidistantly spaced along the logarithmic scale.

Testing was conducted in a quiet room and the stimuli were presented via circumaural headphones. Participants were instructed to label each stimulus as either a short or a long vowel by clicking on “short” or “long” written in Czech orthography on the computer screen. Each stimulus was repeated 5 times, resulting in a total of 195 stimuli (13 duration values * 3 continua * 5 repetitions) which were randomly shuffled before the experiment. We used logistic regression to obtain an identification function for each participant. Per participant, from the regression function we then computed the location of the short-long boundary. The short-long boundary is located at such a stimulus that would receive each of the two labels “short” and “long” with probability of 0.5. Therefore, the boundary x was computed from the formula:

$$\ln \frac{0.5}{1-0.5} = \beta_0 + \beta_1 x \quad (4.1)$$

where β_0 and β_1 are the logistic regression coefficients. Since $\ln \frac{0.5}{1-0.5} = 0$,

$$x = -\frac{\beta_0}{\beta_1} \quad (4.2)$$

The pilot experiment detected the average short-long boundary in Czech to lie at 168 ms. Therefore, the durations of the stimuli in the EEG experiment were manipulated so that 168 ms is the boundary between the 3 short tokens and the 3 long tokens.

The short-long boundary used to separate short and long stimuli in the present EEG study reliably separates phonetically short and long Dutch vowels. The phonetically long Dutch vowels (e.g. /a:/, diphthongized vowels, and true diphthongs) are longer than the present short-long boundary of 168 ms, while phonetically short vowels (e.g. /ɑ/, /ɪ/, /i/) are shorter (Adank et al., 2004, 2007). The average Spanish vowel is (slightly) shorter than the present short-long boundary (Chládková et al., 2011; Zimmerman and Sapon, 1958).

4.1.2.3 Procedure

EEG was recorded in two sessions, one for the native and one for the non-native vowel stimuli, in two different days within a week. Native and non-native vowel stimuli were presented in separate sessions to avoid the influence of their differential status within the listeners' phonemic inventory. The order of the two sessions was counterbalanced across subjects so that in the first session (first day) half of the subjects of each language group listened to native vowels, while the other half listened to non-native vowels. Each session consisted of two 30-minute blocks of EEG-recording (block 1, block 2), with a 15-minute break between blocks.

In one block, short vowels were the standard stimuli and long vowels were the deviants, while in the other long vowels were standards and short vowels deviants. The order of blocks was counterbalanced across subjects but was kept identical across a participant's two sessions. Within a block, the deviant category occurred with a probability of 15.2%. All three deviants and standards were evenly represented in the deviant and the standard category, respectively. Each block started with 20 standards, followed by the oddball sequence which contained 300 deviants (100 deviants of each type), for a total of 2022 stimuli per block. A deviant was always followed by 3 to 8 standards. The inter-stimulus interval was varied randomly in 5 steps between 800 and 932 ms. Stimuli were presented at 60 dB SPL via a single loudspeaker placed in front of the participant at a distance of 1 m at chin level.

Testing took place in sound-attenuated speech laboratories at the University of Amsterdam and at the Palacký University in Olomouc. Eighteen Czech participants were tested in Olomouc, while the Dutch, Spanish and six Czech participants were tested in Amsterdam (the Spanish and the 6 Czechs were exchange university students who had arrived in Amsterdam less than 2 weeks prior to the time of their second session to ensure they had as little foreign-language exposure as possible). All

participants were tested with the same equipment. During stimulus presentation, participants watched a muted movie of their choice (originally spoken in their native language) with subtitles in their native language. At the beginning of each session, participants were given information about the sounds to be played (either sounds from their native language or sounds from a foreign unknown language, depending on the session) and were instructed to disregard the sounds and just watch the movie.

4.1.2.4 *EEG recording and pre-processing*

EEG was recorded from 64 active Ag-AgCl electrodes placed according to the International 10/20 placement in a cap (BioSemi) fitted to participant's head size. Seven external electrodes were used: placed on the nose (offline reference), below and above the right eye, on the left and right temple (ocular activity), and on the right and left mastoid. The input/output gain was 31.25 nV/bit, the EEG signal was recorded at 8kHz and later downsampled to 512 Hz.

The EEG was offline referenced to the nose channel. Slow drifts were removed by subtracting from each channel a line so that the first and the last sample become zero. The data were band-pass filtered in the frequency domain with a low cut-off of 1 Hz (0.5 Hz bandwidth) and a high cut-off of 30 Hz (15 Hz bandwidth). The data were epoched from -100 ms to 700 ms relative to stimulus onset. For subsequent baseline correction the mean voltage in the 100-ms pre-stimulus interval was subtracted from each sample in the epoch. Artifact correction was done automatically (rejection of epochs with $\pm 75 \mu\text{V}$ at any channel) and by subsequent visual inspection. Participants (one Spanish and one Dutch) with more than 50% of artifact-contaminated epochs were excluded from further analysis.

Per participant per block, the epochs of the three short stimuli and the epochs of the three long stimuli were averaged. Per participant, two difference waves were derived by subtracting (1) the average waveform of short standards (from one block) from the average waveform of short deviants (from the other block), and (2) the average waveform of long standards from the average waveform of long deviants. There was thus a within-subject factor "duration-type" with two levels, namely short and long, referring to the comparison of short standards with short deviants from reversed blocks, and long standards with long deviants from reversed blocks, respectively. Previous studies have reported an asymmetry in MMN to duration decrements versus duration increments: duration decrements (equal to the duration-type short in the present study) often yield smaller MMN than duration increments (i.e. duration-type long) (Hisagi et al., 2010; Kirmse et al., 2008; Lipski et al., 2012). Therefore, the factor duration-type was included to test whether asymmetries

between MMN to short and MMN to long stimuli were also present in this study.

In the first block of EEG recording, half of the participants per language were presented with short deviants among long standards, while the other half of participants were presented with long deviants among short standards. It has been suggested that MMN to speech stimuli may be reduced over time due to habituation (McGee et al., 2001). To control for any habituation effects, our analyses also included the between-subjects factor “first-deviant-duration” with two levels: short and long, which refers to the duration-type of deviants from the first block.

We searched for a negative peak (“group-peak”) between 200 and 360 ms post stimulus-onset for each channel in the grand-average difference waveforms per language, first-deviant-duration, vowel-quality, and duration-type. Subsequently, per participant, we computed the mean amplitude over a 40-ms time window centered at the group-peak, which was our measure of MMN amplitude. Statistical tests were done with the alpha level of 0.05.

4.1.3 Results

We first ran an exploratory repeated-measures ANOVA on the MMN amplitude measured at Fz with language and first-deviant-duration as the between-subjects factors, and vowel-quality and duration-type as the within-subject factors. This analysis yielded a significant two-way interaction of first-deviant-duration and duration-type ($F[1, 63] = 67.027; p < 0.001$). Inspection of this two-way interaction revealed that the average MMN in participants who were first presented with *long* deviants was $-1.653 \mu\text{V}$ for *long* stimuli and $-0.357 \mu\text{V}$ for *short* stimuli. The average MMN in participants who were first presented with *short* deviants was $-1.117 \mu\text{V}$ for *short* stimuli and $-0.169 \mu\text{V}$ for *long* stimuli, with the latter not being significantly different from 0. Thus, the MMN was approximately five times larger for deviants from the first block than for deviants from the second block. Importantly, the considerable attenuation of MMN to deviants presented in the second block is independent of duration-type and might be a result of habituation to the frequently repeated standards in the first block (McGee et al., 2001). This means that the small MMN from the second block likely represents a less reliable measure of the subjects’ sensitivity to duration changes. Therefore, all further cross-language and cross-vowel comparisons were conducted on the MMN elicited by the deviants from the first block. This resulted in a single difference wave, or duration-type, per participant, which is the duration-type of the deviant presented in the first block. That is, duration-type was treated as a between-subjects factor in the subsequent analysis. Ac-

cordingly, Table 4.1 and Figures 4.3 tot 4.5 show the data that were compared across groups, namely the MMN for deviants from the first block.

The subsequent repeated-measures ANOVA thus had language (Czech vs. Spanish vs. Dutch) and duration-type (short vs. long) as the between-subjects factors, and vowel-quality (native vs. non-native) as the within-subject factor. This analysis was run on the MMN amplitude measured at 9 channels (Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4), and also included the within-subject factors anteriority (frontal: Fz, F3, F4; fronto-central: FCz, FC3, FC4, central: Cz, C3, C4) and laterality (midline: Fz, FCz, Cz; left: F3, FC3, C3; right: F4, FC4, C4). The mean MMN amplitudes averaged across the 9 sites (Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4) are listed in Table 4.1. Figure 4.3 shows the grand-average difference waveform at FCz and the topographical MMN distributions for each language, vowel-quality and duration-type; Figure 4.4 then shows the standard and deviant waveforms at FCz.

language (n)	duration type (n)	native vowel		non-native vowel	
		mean	95% c.i.	mean	95% c.i.
Czech (23)	long (11)	-1.782	-2.465..-1.099	-1.771	-2.284..-1.258
	short (12)	-0.884	-1.378..-0.389	-1.344	-1.824..-0.863
	average long/short	-1.313	-1.742..-0.885	-1.548	-1.883..-1.213
Spanish (22)	long (11)	-0.795	-1.502..-0.088	-1.783	-2.408..-1.159
	short (11)	-0.706	-1.173..-0.238	-1.278	-2.085..-0.470
	average long/short	-0.750	-1.137..-0.364	-1.531	-2.010..-1.051
Dutch (24)	long (12)	-1.830	-2.715..-0.944	-0.972	-1.541..-0.404
	short (12)	-0.655	-1.121..-0.188	-0.926	-1.333..-0.518
	average long/short	-1.242	-1.768..-0.717	-0.949	-1.271..-0.627

Table 4.1: MMN amplitude (in μV) averaged across 9 sites (Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4). The table shows means and 95% confidence intervals (c.i.) per vowel-quality, language, and duration-type, and the number of subjects (n) in each group.

The ANOVA revealed significant main effects of duration-type ($F[1, 63] = 8.618, p = 0.005$), anteriority ($F[2, 126] = 12.916, p < 0.001$) and laterality ($F[2, 126] = 14.018, p < 0.001$), as well as a two-way interaction of language and vowel-quality ($F[2, 63] = 5.247, p = 0.008$).²

² A similar ANOVA that was run on MMN amplitude elicited by deviants in block 2 did not yield any main effects of vowel-quality, language, or duration-type, neither any interactions involving at least two of these factors (all p 's > 0.1). This provides further evidence for the fact that all listeners, irrespective of their language, vowel-quality or duration-type, habituated to the standards form block 1, which did not yield an MMN when they were presented as deviants in block 2.

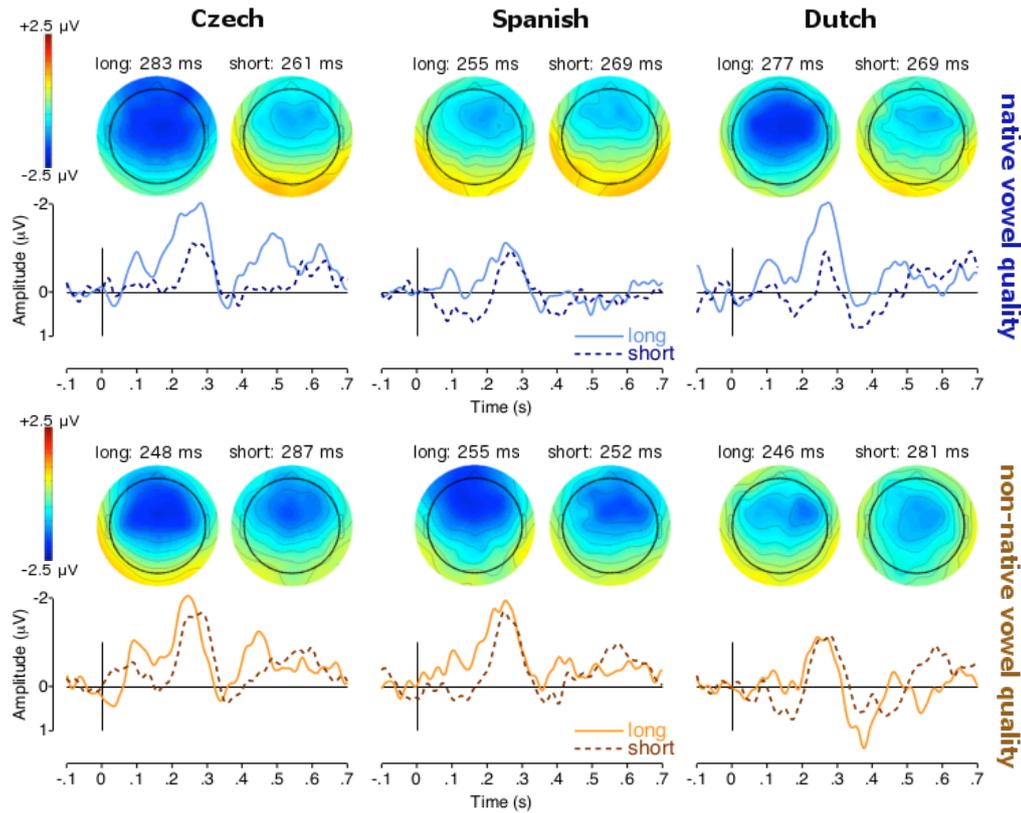


Figure 4.3: Grand-average difference waveforms at FCz of the three language groups in the native vowel quality (blue scale; upper graphs) and the non-native vowel quality (orange scale; bottom graphs), for the short stimuli (dashed dark line) and the long stimuli (solid light line). Topographic MMN distributions at average MMN peak latencies at FCz per language, duration-type, and vowel quality (upper plots = native vowel, lower plots = non-native vowel). Peak latencies at FCz are given above the respective scalps.

As for the main effect of duration-type, long deviants elicited a larger MMN than short deviants by on average $0.524 \mu\text{V}$. Regarding the main effects of laterality and anteriority, MMN amplitude was largest at frontal and fronto-central sites (by, on average, 0.173 and $0.164 \mu\text{V}$, respectively), and it was more prominent at the midline and the right hemisphere than at the left hemisphere for all groups (by, on average, 0.207 and $0.188 \mu\text{V}$, respectively); see Figure 4.3. Below we inspect the significant two-way interaction of language and vowel-quality.

Independent-sample *t*-tests were carried out to compare the MMN amplitude (averaged across the 9 sites) across languages separately for the native and for the non-native vowel. The comparisons showed that for duration changes in the native vowel, Spanish listeners had significantly smaller MMN amplitude than Czech listeners by, on average, $0.563 \mu\text{V}$ ($t[43] = 2.019$, $p = 0.025$, 95% confidence interval [c.i.] of the difference = $0.001..1.126$), and smaller MMN than Dutch listeners by,

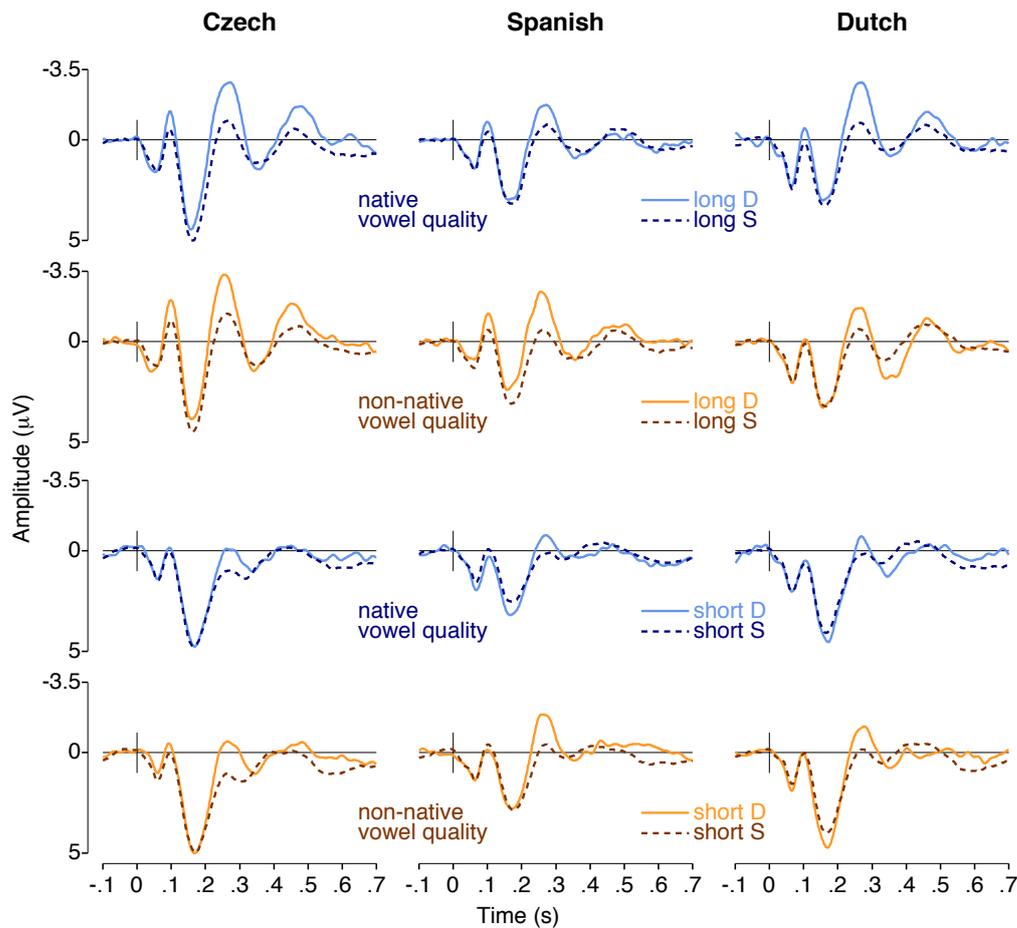


Figure 4.4: Grand-average standard (dashed darker lines) and deviant (solid lighter lines) waveforms for long stimuli (top 2 rows) and short stimuli (bottom 2 rows) at FCz of the three language groups (per column) in the native condition (blue scale) and the non-native condition (orange scale).

on average, $0.492 \mu\text{V}$ (although only nearly-significant with $\alpha = 0.05$; $t[44] = 1.540, p = 0.065, \text{c.i.} = -0.152..1.136$); no significant differences were found between Dutch and Czech. In contrast, in the non-native vowel quality, Dutch listeners had a significantly smaller MMN amplitude than Czech listeners by, on average, $0.599 \mu\text{V}$ ($t[45] = 2.674, p = 0.005, \text{c.i.} = 0.148..1.050$) and smaller MMN than Spanish listeners by, on average, $0.582 \mu\text{V}$ ($t[44] = 2.674, p = 0.020, \text{c.i.} = 0.030..1.133$), while no significant difference was found between Spanish and Czech.

Subsequently, paired-samples t -tests were run to compare the MMN amplitude (averaged across the 9 sites) between the native and the non-native vowel within each language. These comparisons showed that, in Spanish listeners, the MMN amplitude in the non-native vowel was larger than in the native vowel by, on average, $0.780 \mu\text{V}$ ($t[21] = 2.756, p = 0.006, \text{c.i.} = 0.191..1.369$). Although in Czech and Dutch, the difference does not reach significance, it can be observed that Czech listeners fol-

low a similar trend to that of Spanish listeners in that their MMN in the non-native vowel also appears to be larger than in the native vowel (by, on average, $-0.235 \mu\text{V}$; $p = 0.095$). Conversely, the Dutch MMN follows the opposite trend in that their MMN tends to be larger in the native than in the non-native vowel (with the average difference being $+0.293 \mu\text{V}$; $p = 0.134$). Figure 4.5 shows the mean MMN amplitude per language and vowel-quality averaged across the two duration-types and across the 9 sites.

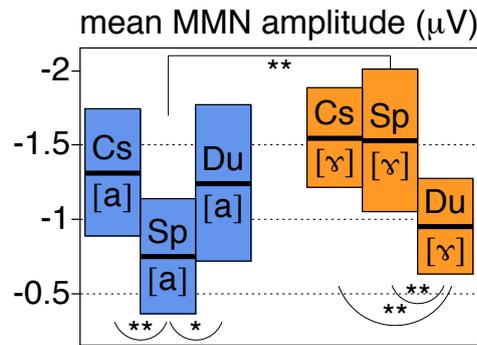


Figure 4.5: Mean MMN amplitude per language (Cs = Czech, Sp = Spanish, Du = Dutch) and vowel quality (native = [a], non-native = [ɣ]) pooled across 9 sites (Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4). The bars indicate 95% confidence intervals, the thick lines mark the mean. Asterisks mark (nearly-)significant differences between groups; $**p < 0.05$, $*p = 0.065$.

In sum, our results show that in the native vowel [a], Czech and Dutch listeners have a larger MMN to duration changes than Spanish listeners. In contrast, in the non-native vowel [ɣ], Czech and Spanish listeners have a larger MMN than Dutch listeners. Spanish listeners have a reliably larger MMN in the non-native than in the native vowel. The same, although non-significant, trend is observed in Czech listeners, while the opposite trend is seen in Dutch listeners.

4.1.4 Discussion

Neurophysiological research demonstrated that experience with native phonology shapes the early pre-attentive speech sound processing. In that respect, both the phonemic status of the sound within one's native language or dialect (Kazanina et al., 2006; Näätänen et al., 1997; Nenonen et al., 2005; Scharinger et al., 2011c) and the specific phonological structure of the sound (the contrastive role of its various acoustic dimensions) (Hisagi et al., 2010; Lipski et al., 2007; Obleser et al., 2004; Scharinger et al., 2011a,b; Ylinen et al., 2006) can modulate listeners' pre-attentive response to speech sounds. To this date, however, the phonological structure of some languages remains unclear (e.g. status of vowel

length in languages such as Dutch). Therefore, we used cross-linguistic and cross-stimulus comparisons of the mismatch response – a measure of pre-attentive processing of speech sounds – to reveal phonological structure. With this approach, our study represents a first step towards resolving the unclear phonological status of vowel length in Dutch, for which speech production and perception studies gave conflicting evidence.

We compared Dutch listeners' pre-attentive receptiveness for vowel duration to that of Czech listeners whose native phonology unequivocally encodes vowel length, and to that of Spanish listeners whose native phonology unequivocally does not encode vowel length. We assessed duration processing in two types of stimuli: a native and a non-native vowel quality, because if a language encodes a particular phonetic dimension (i.e. vowel duration) in terms of discrete phonological categories (i.e. vowel length categories) in its phonology, then this encoding should be generalized across vowel phonemes.

4.1.4.1 *Phonological role of vowel duration in the native phonology*

Our results demonstrate that Dutch listeners' MMN response to vowel duration in the native vowel quality [a] does not differ from that of quantity-language listeners (Czechs) and is larger than that of non-quantity language listeners (Spanish). This indicates that duration differences between phonetically short and long tokens of [a] may signal a category boundary in Dutch: that is, Dutch listeners do not perceive short [a] as the same category as long [a:]. However, our results also show that Dutch listeners' receptiveness to duration in the non-native vowel quality [ɤ] was smaller than that of Czech listeners. This indicates that the change between a short [ɤ] and a long [ɤ:] is more likely to represent a phonemic, i.e. linguistically relevant, change for Czech than for Dutch listeners, despite the fact that the spectral quality of [ɤ] is not phonemic in either language.

We thus find that Dutch does not pattern with either Czech or Spanish, which suggests that next to quantity and non-quantity phonologies, there is (at least) one other type of phonological system with respect to phonetic vowel duration and phonological length. Specifically, the Dutch quantity-like large MMN to duration in [a] indicates that the phonetic dimension of duration is used phonemically in some way in Dutch. However, the Dutch MMN to duration in [ɤ], which was smaller than that of quantity-language listeners, indicates that the phonological encoding of vowel duration in Dutch differs from that of quantity languages.

Here we put forward that in Dutch, phonetic vowel duration is not encoded in terms of abstract phonological categories for vowel length, but it is used to define the identity of certain vowel qualities, such as the vowel /a:/ which has to have a long duration. This is illustrated in

Figure 4.6 where it can be seen that in Dutch, an [a]-like stimulus with short duration is not perceived as the /a:/ category. Note that Dutch also has diphthongs such as /e:/ (realized as [e¹]) or ϵI , which are phonetically longer than most monophthongs (Adank et al., 2004). If duration is a cue to diphthongs, perception of an [e]-like vowel may then work similarly to an [a]-like vowel: a short tokens of [e] will not be perceived as /e:/, but possibly as a different vowel category.

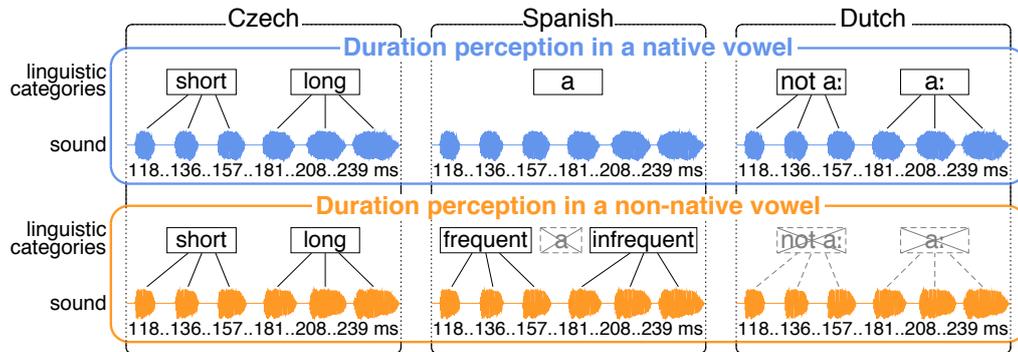


Figure 4.6: Model of Czech, Spanish, and Dutch perception based on our findings. Top graph: Czech and Dutch have some linguistic categories defined by duration. In Czech they are abstract length categories independent of vowel quality, while in Dutch these categories are specific vowels (the vowel /a:/ in this case). Bottom graph: Czech but not Dutch can apply their native linguistic categories for duration in perception of non-native vowel qualities. In Spanish, vowel duration does not contribute to native linguistic categories (top graph), and therefore, Spanish listeners can use this dimension to create new non-native categories in a second language (bottom graph).

Recall however that we did not detect a significant difference within Dutch listeners between their MMN to the native versus the non-native vowel quality. However, we believe that this within-language finding for Dutch should only be interpreted with caution. This is because it is well known that in speech production, different vowel qualities have different intrinsic durations: high or mid vowels such as /i/ or /e/ are intrinsically shorter than low vowels such as /a/ (Lehiste, 1970). In line with that, a recent experiment by Meister et al. (2011) demonstrated that vowel height affects the perceived short-long boundary for listeners of quantity languages: high vowels require less duration difference to be perceived as long than low vowels. This implies that a physically identical duration difference is likely to be perceived as relatively large in a mid (or high) vowel, and as relatively small in a low vowel. Since our stimuli included a mid (non-native) and a low (native) vowel, a quantity-language listener should exhibit a (slightly) different MMN for each of these vowels, i.e. the MMN for the non-native mid vowel [ɤ] should be larger than that of

the native low vowel [a]. And, in fact, the Czechs appear to follow this trend, while Dutch listeners seem to follow the opposite trend.

It thus seems that Dutch listeners' MMN for the native versus the non-native conditions may have been influenced by their differential sensitivity for duration across different vowel qualities (i.e. different vowel heights). However, this explanation is only preliminary since further research should compare Dutch /a:/ to another low vowel, e.g. /ɑ/, in order to conclusively determine Dutch duration sensitivity across vowels.

4.1.4.2 *Non-native use of vowel duration*

As can be seen in Figure 4.2, Dutch differentiates twelve spectral qualities, while Czech and Spanish only five. In line with that, Escudero et al. (2009) have shown that Dutch listeners weigh spectral properties heavier than durational properties (while Spanish learners of Dutch weighted duration heavier than spectrum). One might therefore argue that Dutch listeners will only show high sensitivity to spectral and less so to durational changes for all vowels. This may explain why they had the smallest MMN to duration differences in the non-native condition, but it cannot explain why they had a similarly large MMN to that of Czechs in the native condition, since neither condition contained spectral differences that could be used to distinguish the vowels.

Regarding Spanish listeners, their duration sensitivity in the non-native condition was surprisingly similar to that of the Czech, who are quantity-language speakers, and larger than that of the Dutch. In that respect, recall that Nenonen et al. (2005) also showed that pre-attentive sensitivity to duration is larger in non-quantity listeners for non-native than for native vowel qualities. Unlike Nenonen et al. whose participants were advanced learners of a quantity language, our Spanish monolinguals' strong duration sensitivity in non-native vowels cannot be due to experience with second-language length categories. Since the system underlying MMN generation is affected by linguistic experience (e.g. Näätänen et al., 1997), it is plausible that the complete lack of meaningful duration differences in Spanish signifies that duration is phonologically a blank slate dimension and that therefore, processing of non-native vowel duration is not affected by native phonemic representations.

Importantly, these findings for Dutch and Spanish listeners' duration processing in non-native vowels speak to theories of second-language speech development. Bohn (1995) proposed that in areas of vowel space where spectral properties do not contrast any native phonemes, listeners would attend to duration to differentiate novel vowel contrasts, which explains the larger MMN to duration in the non-native versus the native condition for Spanish listeners, but does not explain why Dutch listeners do not use duration to the same extent as Czech and Spanish listeners

in the non-native condition. Additionally, one could interpret Bohn's hypothesis as suggesting that listeners with few spectral contrasts (Spanish) in their L1 would be more sensitive to duration than listeners with many spectral contrasts (Dutch), which also explains why Spanish listeners resort to duration in the non-native condition. However, such an interpretation would predict that Dutch listeners are not sensitive to duration in any vowel quality, which runs contrary to the results for the native condition.

Alternatively, our data seems to be in line with Escudero et al. (2009); Escudero and Boersma (2004), who proposed that sensitivity to vowel duration in foreign vowels depends on the learners' native phonology. Specifically, our results show that Czech listeners, who encode phonetic vowel duration in terms of abstract phonological length categories (i.e. 'short' and 'long') in their phonology, apply the duration cue equally in native and non-native perception. That is, they generalize their use of the *phonological length contrast* to non-native vowel inventories. Spanish listeners either have a blank slate for duration (i.e. they do not encode phonetic duration at all in their phonology) and are able to learn different categories along this dimension in the same way infants learn the sound categories of their first language (Escudero and Boersma, 2004)³, or automatically use duration when confronted with any non-native vowel quality (Bohn, 1995).

In Dutch, vowel duration is neither a blank slate dimension (as it is in Spanish) nor is it divided in phonological length contrasts such as 'short' vs. 'long' (as it is in Czech), but it is an acoustic dimension that contributes to the identity of only certain vowel phonemes, e.g. /a:/ (Adank et al., 2004; Nootboom and Doodeman, 1980). This explains why Dutch listeners have a lower sensitivity to duration for the non-native condition than Spanish and Czech listeners. This low sensitivity to non-native duration-based contrasts might also help to explain Dutch listeners' substantial difficulty in distinguishing English words such as *buzz* and *bus*, whose difference is primarily cued by vowel duration in native English listeners (Broersma, 2005, 2010; Elsendoorn, 1985).

Beside the between-language differences, we found that long deviants elicited larger mismatch responses than short deviants, independently of vowel type or listeners' native language. This asymmetry has been reported previously for various languages (Hisagi et al., 2010; Kirmse et al., 2008; Lipski et al., 2012). The asymmetry seems to result from

³ Infants learn to discriminate novel speech sounds through a learning mechanism known as statistical or distributional learning (Maye et al., 2008, 2002) and adults seem to acquire L2 phonemes via the same mechanism (Escudero et al., 2011; Gulian et al., 2007; Hayes-Harb, 2007; Maye and Gerken, 2001). If this learning mechanism remains active throughout the life span, it is not surprising that after brief exposure to systematic durational variation in non-native vowels, Spanish listeners become very sensitive to short and long vowels.

general processing mechanisms in that an unexpected extra portion of acoustic signal (long deviant among short standards) evokes stronger responses than an unexpected absence of acoustic signal (short deviant among long standards).

4.1.5 *Conclusions*

The present study attempted to uncover the thus far uncertain status of vowel length within Dutch phonology. A cross-linguistic comparison of listeners' pre-attentive processing of duration in a native and in a non-native vowel quality suggests that Dutch is neither like Spanish (a non-quantity language) nor like Czech (a quantity language). Unlike Spanish listeners, Dutch listeners exploit the phonetic dimension of vowel duration in their native language, as reflected by their large MMN in the native vowel condition. And unlike Czech who use the phonetic dimension of vowel duration to cue phonological length categories 'short' and 'long,' Dutch appears to use phonetic vowel duration as a cue to the phonological identity of specific vowel phonemes, as reflected by their smallest MMN in the non-native condition.

Further, Spanish listeners demonstrated large sensitivity to duration changes when exposed to a novel vowel quality. It is plausible that the widely reported Spanish speakers' reliance on vowel duration in a second language is caused by the complete lack of this dimension in their native language. However, further research should show whether this sensitivity is the result of automatic psychoacoustic salience, as proposed by Bohn (1995), or is the result of the learning of frequency distributions, as proposed by Escudero and Boersma (2004).

Acknowledgements

We thank Paul Boersma for valuable comments on experiment design, data and statistical analyses, and previous versions of this paper, Karin Wanrooij for comments on experiment design and for sharing her testing and data analysis experience, Silke Hamann for comments on a previous version of this paper, Dirk Jan Vet for technical assistance, Nele Salveste for recording the stimuli, the Department of Czech Studies at Palacký University and Václav Jonáš Podlipský for providing and customizing their speech lab to our needs and Barbora Chládková, Clara Martín Sánchez, Vanina Grippo, Ana Díaz Dasi, Michelle van Bokhorst, Sascha Couvee and Gisela Govaart for participant recruitment and assistance in data collection. This study was funded by NWO (The Netherlands Organization for Scientific Research) grant 277-70-008 awarded to Paul Boersma.

4.2 MAAN IS LONG BUT MAN IS NOT SHORT: NEUROPHYSIOLOGICAL EVIDENCE FOR THE STATUS OF VOWEL LENGTH IN DUTCH

This section is an adapted version of:

Kateřina Chládková, Paola Escudero, & Silvia Lipski (submitted). MAAN is long but MAN is not short: Neurophysiological evidence for the status of vowel length in Dutch.

Abstract

The phonological role of vowel length in Dutch is under debate. Speech production and perception studies give conflicting evidence on the relevance of vowel duration in Dutch phonology. The present study assessed Dutch listeners' pre-attentive processing of duration in two vowel qualities: [a] and [ɑ] (as in *maan* 'moon' and *man* 'man'). If a language encodes phonetic duration into phonological length categories, listeners should be equally sensitive to duration across all vowels. Thus, duration changes in both [a] and [ɑ] should elicit similar neural mismatch responses (MMN). However, we found that duration changes evoked larger MMN amplitude for [a] than for [ɑ]. We propose that duration is phonemically more relevant for the *maan*-vowel, which has to be long, while duration is not phonemically specified for the *man*-vowel. Thus, our findings suggest that in Dutch, vowel duration is a phoneme-specific property and is not represented in terms of phonological length categories.

4.2.1 *Introduction*

The phonological status of vowel length in Dutch, i.e. whether Dutch speakers have abstract phonological representations⁴ of short and long vowels, has been debated for decades and remains a question (for a review see Botma and van Oostendorp, 2012). In languages with phonological vowel length (i.e., quantity languages such as Finnish, Estonian, Czech, or Japanese), phonologically short vowels are produced, i.e. phonetically realized, with short duration, while long vowels are phonetically realized with long duration. One of the reasons why the relation between phonetic duration and mental phonological representations for length is not clear in Dutch is because speech perception and production studies provide conflicting evidence.

⁴ Phonological representations are the stored functional entities of speech sounds, that is, they can be described as abstract correspondents of speech sounds that function at a discrete linguistic level free from the actual physical (i.e. auditory or articulatory) properties of speech signals.

Northern Standard Dutch has 15 vowels, all produced with different spectral properties: 9 monophthongs /i ɪ y ʏ ε a: ɑ ɔ u/, three diphthongs /ɛɪ œy ɔu/, and three ‘potential’ diphthongs /e: ø: o:/ (realized as [ei œy ou], respectively). The six diphthongs and /a:/ are usually produced with a long duration, while the remaining eight monophthongs are produced with a short duration Adank et al. (2004). Even though /i/, /y/, and /u/ are phonetically short, phonologists who argue that Dutch has phonological vowel length (e.g. Moulton, 1962) describe these vowels as phonologically long (e.g. because they occupy the same syllabic positions as phonetically long vowels). This means that in production, Dutch speakers do not seem to use duration consistently across all phonologically short-long contrasts.

Chládková et al. (2013) assessed pre-attentive processing of vowel duration in Dutch listeners and compared them to listeners who unequivocally have abstract length and those who do not (Czech and Spanish, respectively). The Dutch differed from either group depending on the spectral properties of the vowel they heard. For [a], which has a native vowel quality (of e.g. Dutch *maan* ‘moon’), Dutch listeners exhibited large sensitivity to duration changes comparable to that of Czech and larger than that of Spanish listeners. In contrast, for [ɤ], which has a non-native vowel quality, Dutch listeners had a smaller sensitivity to duration than both Czech and Spanish listeners. The authors proposed that Dutch listeners might not have abstract representations for vowel length across their vowel system. However, they did not find a significant difference within Dutch listeners for duration changes in native versus non-native vowel quality: therefore, no reliable conclusion could be drawn about the phonological role of vowel length in Dutch. The authors suggested that the height difference between native [a] and non-native [ɤ] (i.e., a low versus a mid vowel) might have obscured a between-vowel difference in duration sensitivity. That is, differences in intrinsic vowel length between mid and low vowels may cause differences in relative perception of duration changes (see Meister et al., 2011). Specifically, listeners may be universally more sensitive to a specific absolute duration change in an (intrinsically short) mid vowel than in an (intrinsically long) low vowel. If, however, one’s phonology uses duration contrastively in low but not in mid vowels (as could be the case for Dutch), then this language-specific phonological effect may clash with the universal psychoacoustic effect, thus cancelling each other out.

The present study investigates whether Dutch listeners generalize their duration processing across native vowel qualities that do not differ in height, namely [a] and [ɑ]. The aim is to investigate whether Dutch listeners perceptually rely on duration to the same extent for all native vowels, including /a:/ and /ɑ/, as should be the case if length was a phonological feature in Dutch.

Length-based phonological descriptions of Dutch consider /a:/-/a/ a length contrast (Moulton, 1962), partly because these vowels are usually produced with a long and short duration, respectively (Adank et al., 2004). Dutch listeners' perception of /a:/ and /a/ should also reflect the status of vowel length for this contrast. In that respect, in an overt vowel classification task, Escudero et al. (2009) found that Dutch listeners almost neglected the duration differences between /a:/ and /a/ and instead relied on spectral properties to distinguish these two vowels (for a similar finding see also van Heuven et al., 1986). Auditorily, however, this duration contrast is clearly processed by Dutch listeners, as shown by Lipski et al. (2012) where duration and spectral changes between the vowels evoked similar MMN responses. A difference between Escudero et al.; Lipski et al.'s set-up of tested vowel contrasts should be noted here: the former tested durational reliance for both /a:/ and /a/, while the latter did so only for /a:/.

Interestingly, earlier behavioral perception studies have shown that Dutch listeners identify a token of /a:/ with a short duration as /a/, but do *not* identify a token of /a/ with a long duration as /a:/ (Nootboom and Doodeman, 1980; van der Feest and Swingley, 2011), which may suggest that duration is perceptually relevant only for /a:/ and not for /a/. However, listeners' responses in behavioral tasks could be frequency-driven. Specifically, Dutch listeners may be able to discriminate [a]-[a:] equally well as [a]-[a:], but identify only the former two as a single phoneme. That is, they may less likely classify [a] and [a:] as a single phoneme, possibly because their experience tells them that [a] can occur as a realization of /a/ in some Dutch dialects and consonantal contexts (see Benders, 2013: 91). In order to demonstrate whether vowel duration is an equally strong perceptual cue across Dutch vowels, we carried out a direct comparison of Dutch listeners' pre-attentive detection of duration changes, as reflected by the MMN, for the two vowel categories /a:/ and /a/.

The MMN is elicited when infrequent deviations occur among frequently repeated sounds, and is modulated by linguistic experience: acoustic deviations that represent a phonemic change can elicit a stronger MMN response than those that do not represent a phonemic change (Näätänen et al., 1997; Nenonen et al., 2005; Sharma and Dorman, 2000; Ylinen et al., 2006).

Dutch listeners were presented with duration changes in [a] and [a:], which resemble the quality of their native phonemes /a:/ and /a/, respectively, as shown in Figure 4.7. If duration is phonemically relevant for /a:/ and not for /a/, the change between [a] and [a:] should elicit a stronger MMN response than the change between [a] and [a:]. If, on the other hand, duration is phonemically relevant for both /a:/ and /a/, the

change between [a] and [a:] and the change between [ɑ] and [ɑ:] should elicit equally large MMN responses.

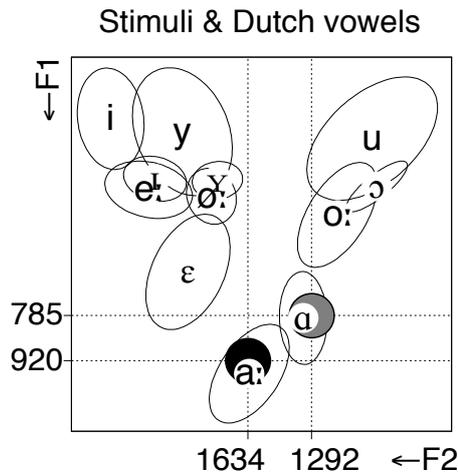


Figure 4.7: F1 and F2 values of Randstad Dutch vowels produced by female speakers (van Leussen et al., 2011) and the two vowels produced by a female Estonian speaker that served as stimuli in the present study: [a] = black filled circle, [ɑ] = grey filled circle. Phonetic symbols indicate the mean values of the Dutch vowels, ellipses show two standard deviations. Axes are scaled in Erb, marks are in Hz.

The present results will thus provide strong evidence for whether or not Dutch listeners have equal sensitivity to vowel duration across all native vowels, as would be the case if phonological length was part of the Dutch vowel system.

4.2.2 Methods

4.2.2.1 Participants

Eighteen young healthy right-handed listeners took part. We measured their MMN to duration changes in two separate sessions: in one session they listened to duration changes in [a], while in the other session they listened to duration changes in [ɑ]; the order of the two sessions was counterbalanced across subjects. Nine participants' data for [a] comes from the data reported in Chládková et al. (2013), i.e. 9 participants who were presented with [a] in that study's first session (mean age at first session = 22.8, range = 19–26; 3 male). The nine participants listening to [ɑ] in their first session were newly recruited participants for the present study (mean age at first session = 22, range = 19–24; 5 male). For all participants, the second session was administered 10 to 11 months after the first session.

The participants were all monolingual Dutch native speakers from the Randstad area in the Netherlands. Seven additional participants were

recruited for the first session: two of them had a large number of artifacts (> 50%) in the first session and were thus further excluded from the study, and five participants chose not to take part in the second session. Participants gave a written informed consent and were paid for participation. The study was approved by the ethical committee of the Faculty of Humanities, University of Amsterdam and conforms to the guidelines of the Declaration of Helsinki (2008).

4.2.2.2 *Stimuli*

The stimuli were natural tokens of the Estonian vowels /æ/ and /ɑ/ (henceforth transcribed as [a] and [ɑ], respectively). The values of the first three formants were 920 Hz, 1634 Hz and 2707 Hz for [a], and 785 Hz, 1292 Hz, and 2675 Hz for [ɑ]. As shown in Figure 4.7, [a] is acoustically similar to Dutch /a:/, and [ɑ] is acoustically similar to Dutch /ɑ/. The procedure for creating 6 different durations in psychoacoustically equal steps: 118, 136, 157, 181, 208, and 239 ms, is described in Chládková et al. (2013). The stimuli were presented in the categorical oddball paradigm, in which the 118-, 136-, and 157-ms items served as the short stimulus category while the 181-, 208-, and 239-ms items served as the long stimulus category.

4.2.2.3 *Procedure*

As noted above, participants were presented with duration changes in [a] in one session, and with duration changes in [ɑ] in the other session. A testing session consisted of two 30-minute blocks of EEG-recording (block 1, block 2), with a 15-minute break between blocks.

In one block, short vowels were the standard stimuli and long vowels were the deviants, while in the other long vowels were standards and short vowels deviants. The order of blocks was counterbalanced across subjects. Within a block, the deviant category occurred with a probability of 15.2%. All three deviants and standards were evenly represented in both the deviant and the standard category. Each block started with 20 standards, followed by the oddball sequence which contained 300 deviants (100 deviants of each type), for a total of 2022 stimuli per block. A deviant was always followed by 3 to 8 standards. The ISI was varied randomly in 5 steps between 800 and 932 ms. Stimuli were presented at 60 dB SPL via a single loudspeaker placed in front of the participant at a distance of 1 m at chin level.

Testing took place in a sound-attenuated speech laboratory at the University of Amsterdam. During stimulus presentation, participants watched a muted movie of their choice (originally spoken in Dutch) with subtitles in Dutch. Before the session started, participants were told they

would hear Dutch vowels and were instructed to disregard them and just watch the movie.

4.2.2.4 EEG recording and pre-processing

The EEG recording, pre-processing, and MMN quantification methods are identical to those described in Chládková et al. (2013).

4.2.3 Results

For each condition (i.e., vowel quality and duration type), we searched for a negative peak of the grand mean difference between standard and deviant response⁵ within the time window 200 to 360 ms after stimulus onset. Subsequently, in a 40-ms window centered at the detected grand-peak, we measured the mean MMN amplitude for each individual subject.

Table 4.2 lists the mean MMN amplitudes averaged across 9 sites: Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4. Figure 4.8 shows the grand-average standard, deviant, and difference waveforms at Fz, as well as the topographical MMN distributions for each vowel-quality and duration-type.

duration type (n)	MMN amplitude: mean and c.i.			
	[a]		[ɑ]	
long (9)	-1.722	(-2.348...-1.096)	-1.034	(-1.662...-0.406)
short (9)	-0.972	(-1.598...-0.347)	-0.672	(-1.300...-0.044)
average (18)	-1.347	(-1.790...-0.905)	-0.853	(-1.297...-0.409)

Table 4.2: MMN amplitude (in μV) averaged across 9 sites (Fz, FCz, Cz, F3, F4, FC3, FC4, C3, C4). The table shows means and 95% confidence intervals (c.i.) per vowel-quality and duration-type, and the number of subjects (n) in each group.

The MMN amplitudes were compared in statistical analyses similar to those reported in Chládková et al. (2013). An exploratory repeated-measures analysis of variance (ANOVA) on the MMN amplitude measured at Fz was run first. It had vowel-quality and duration-type as the within-subject factors, and first-deviant-duration as the between-subjects factor. There was a significant two-way interaction of duration-type and first-deviant-duration ($F[1, 16] = 12.293, p = .003, r = .66$). Pairwise comparisons of the means revealed that the average MMN in participants who were first presented with *long* deviants was $-1.486 \mu\text{V}$ for *long* stimuli and $-0.069 \mu\text{V}$ for *short* stimuli (with the 95% confidence interval [c.i.]

⁵ Difference waves were computed by subtracting responses to physically identical standard from deviant stimuli.

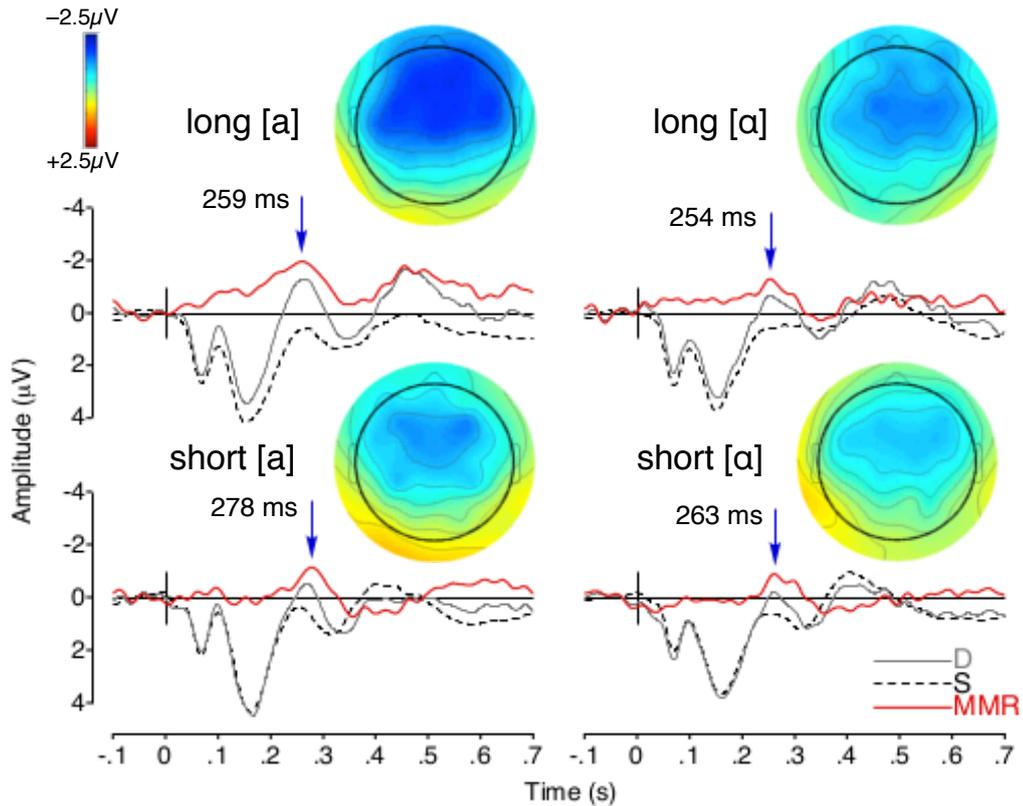


Figure 4.8: Grand-average deviant (grey line), standard (black dashed line), and difference waveforms (red line) at Fz, and scalp distribution at MMN peaks in the two vowel qualities for long (top) and short stimuli (bottom). The MMN peaks are marked by blue arrows with an indication of their latencies

of the latter not significantly different from o). The average MMN in participants who were first presented with *short* deviants was $-0.841 \mu\text{V}$ for *short* stimuli and $-0.333 \mu\text{V}$ for *long* stimuli (with the c.i. of the latter not significantly different from o). That is, the MMN was considerably larger for deviants from the first block than for deviants from the second block.

This finding replicates the block-effect reported in Chládková et al. (2013), where the attenuation of MMN to deviants presented in the second block was interpreted as a result of habituation to the frequently repeated standards in the first (McGee et al., 2001). Since the declined MMN responses from block 2 may not reliably represent the listeners' true sensitivity to duration, we follow Chládková et al. (2013), and further compare the MMNs elicited by deviants from the first block only (Table 4.2 and Figure 4.8, accordingly, show MMN to deviants from the first block).

A second repeated-measures ANOVA was carried out with vowel-quality (native vs. non-native) as the within-subjects factor and with duration-type (short vs. long) as the between-subjects factor. The MMN amplitudes measured at 9 channels (Fz, FCz, Cz, F3, F4, FC3, FC4, C3,

C₄) were included in the analysis, and therefore anteriority (frontal: F_z, F₃, F₄; fronto-central: FC_z, FC₃, FC₄, central: Cz, C₃, C₄) and laterality (midline: F_z, FC_z, Cz; left: F₃, FC₃, C₃; right: F₄, FC₄, C₄) were also within-subject factors.

The ANOVA revealed a main effect of vowel quality ($F[1, 16] = 4.976, p = .040, r = .49$). Pairwise comparisons revealed that duration changes in [a] yielded a larger MMN than duration changes in [ɑ] by on average 0.494 μV (95% c.i. of the difference = 0.025..0.964 μV). The analysis did not detect any other significant main effects or interactions.⁶

4.2.4 Discussion

This study investigated whether Dutch listeners are equally sensitive to duration changes in the vowels /a:/ and /ɑ/. These vowels are produced with long and short durations, respectively, and distinguish Dutch words such as *maan* and *man* ('moon' and 'man'). If duration were an equally important phonetic cue to both members of the Dutch /a:/-/ɑ/ contrast, one would expect to find an equally strong mismatch response to duration changes in both these vowels.

We measured the amplitude of the mismatch response to duration changes for [a] and [ɑ]. The results showed that duration changes elicited a larger MMN response in [a] than in [ɑ], which indicates that Dutch listeners do not rely on duration to the same extent across all native vowels. Crucially, our Dutch listeners' MMN amplitude for duration changes in [a] was comparable to those of Dutch and Czech listeners from Chládková et al. (2013) for the same vowel, while their MMN amplitude for duration changes in [ɑ] was comparable to that of Chládková et al.'s Spanish listeners for [a]. This further indicates that Dutch listeners may represent duration differently for /a:/ than for /ɑ/: for the former, they have a strong, *quantity-language-like* reliance on duration, while for the latter they have a weak, *non-quantity-language-like* reliance on duration.

Chládková et al. (2013) compared Dutch listeners' duration sensitivity in [a] and [ɤ], and although the listeners tended to be less receptive to duration in [ɤ] than in [a], that difference was not significant. The lack of a significant difference was possibly due to an additional vowel-height confound between [ɤ] and [a]: specifically, in listeners who are equally sensitive to duration across all vowel qualities, one expects the MMN to physically identical duration changes to be larger in an intrinsically shorter mid vowel [ɤ] than in an intrinsically longer low vowel [a]. In

⁶ A similar ANOVA run on MMN amplitude elicited by deviants in block 2 did not yield any main effects or interactions involving vowel-quality and duration-type. This further supports the fact that all listeners, irrespective of vowel-quality or duration-type, habituated to the standards from block 1, which did not yield an MMN when they were presented as deviants in block 2. This finding replicates that of Chládková et al. (2013).

the present study, both stimuli had the same height (i.e. they were low vowels [a] and [ɑ]), and we detected a reliable difference in duration receptiveness between these two vowels. Therefore, the present results demonstrate that Dutch listeners do not generalize their duration processing across different vowel qualities, which further supports the proposal made in Chládková et al. that Dutch phonology may not encode phonetic duration into phonological length categories ‘short’ and ‘long’, but that duration is a vowel-specific cue.

Our findings thus indicate that duration is a reliably less relevant phonetic property for Dutch /ɑ/ than for Dutch /a:/. We propose that this differential sensitivity may be phonemic: /a:/ is stored as a long vowel, while /ɑ/ does not have a specification for vowel duration. The differential phonemic status could explain why these vowels are produced with distinct durations. That is, since /a:/ is represented as ‘long’, it is produced with a long duration, while /ɑ/, which has no length representation, can be produced with any duration, but its short version is most common because it involves less articulatory effort (Boersma, 1998: 149–151).⁷

The proposed differential phonemic relevance of duration for these two vowels also explains the finding of previous behavioral studies where the perceived identity of the stimulus was more likely to be affected by duration changes in [a] than in [ɑ] (e.g. Nooteboom and Doodeman, 1980; van der Feest and Swingley, 2011). Specifically, since duration is relevant for /a:/ but not for /ɑ/, listeners perceive a phonemic difference between [a] and [a:] but not between [ɑ] and [ɑ:].

In sum, the present study found that Dutch listeners have a reliably larger MMN amplitude to duration changes in [a] than in [ɑ], which indicates that duration is not an equally important perceptual cue for all vowels in Dutch. This finding suggests that Dutch uses duration as a vowel-specific property and may thus not contain the feature vowel length in its phonology.

Acknowledgements

This study was funded by the Netherlands Organization for Scientific Research (NWO) grant 277-70-008 awarded to Paul Boersma. We are grateful to Paul Boersma for providing the funds, and for comments on data analysis and previous versions of this paper. We thank Dirk Jan Vet for technical assistance, Nele Salveste for recording the stimuli, Clara Martín Sánchez, Michelle van Bokhorst, Sascha Couvee, Gisela Govaart,

⁷ Speakers of non-quantity languages that do not use duration phonemically (e.g. Spanish) may commonly realize all vowels with short duration because it involves smaller articulatory effort (Chládková et al., 2011; Zimmerman and Sapon, 1958).

Marieke van den Heuvel and Brechje van Osch for participant recruitment and assistance in data collection.

THE EMERGENCE OF PHONOLOGICAL FEATURES IN AN ARTIFICIAL NEURAL NETWORK

This chapter will be incorporated in:

Paul Boersma, Kateřina Chládková, & Titia Benders. (in progress). Learning phonological structures from auditory input and phonological alternations [working title].

*Section 5.3 has been presented as Boersma & Chládková (2013b),
Section 5.4 as Boersma, Chládková & Benders (2013b),
and Section 5.5 as Boersma & Chládková (2013c).*

ABSTRACT

This study aims to determine whether language learners create phonological feature representations for the sounds of their language. We model an artificial neural network with three layers: sound, phonology and lexicon. We implement three versions of the network and simulate lexicon-driven learning of a typical five-vowel language (the three network versions differ in their sound-layer architecture and in the type of information available in the lexicon). The results of the simulations show that learners who have separate auditory layers for the first and the second formant (network 1) create mostly feature-like representations for their vowels, while those who have a single auditory layer for formant frequency (network 2) create mostly phoneme-like representations. Finally, learners who have a single auditory layer for formant frequency but who are also able to employ morphological knowledge at some point during vowel learning (network 3) create mostly feature-like representations for their vowels. Since network 3 represents formant frequency on a single auditory dimension (cf. basilar membrane), and allows the virtual infant to also use her knowledge of morphemes at later stages of learning (cf. Berko, 1958; Fikkert and Freitas, 2006), we argue that it models human language acquisition more realistically than networks 1 and 2. Thus, we conclude that phonological feature categories emerge from learners' exposure to the phonetics and morphophonology of the ambient language.

5.1 COMPUTATIONAL MODELS OF PHONOLOGY

The literature indicates that language users represent the sounds of their language in terms of phonological features (e.g. Kingston, 2003; Miller and Nicely, 1955; Scharinger et al., 2011a; see also Chapter 2 and Chapter 3 in this thesis). Psycholinguistic and neurolinguistic experiments traditionally test whether and how the potential phonological representations are *reflected* in the participants' speech production and perception (either overtly or pre-attentively). Ideally, however, a laboratory phonologist seeks not only to observe the reflections of speakers' mental representations for phonology, but also to directly and in real time observe whether and how these representations are learned, and how they are employed in speech production and comprehension.

A straightforward assessment of phonological representations and their learnability is viable with computational modeling. Perhaps most notably, Optimality Theoretic (OT) models have been widely employed to explain various aspects of phonetics and phonology such as perceptual warping (Boersma et al., 2003) or auditory dispersion (Boersma and Hamann, 2008). Although OT can account for a wide range of phonetic and phonological phenomena, biologically it is rather implausible. For instance, OT posits an infinite candidate set from which a language user selects her perception or production output, so if OT is to represent the language user's processing it will require the language user to have an infinite mental storage capacity. In that respect, the properties of the human brain, i.e. of a biological neural network, are more closely approximated by artificial neural network (NN) models.

To date only a few studies have employed artificial neural networks to model phonetics and phonology. For instance, Guenther and Gjaja (1996) trained a two-layer neural network with language-specific auditory distributions of sounds, and showed that the network comes to exhibit language-specific perceptual warping of the auditory space. Guenther and Gjaja thus simulated the perceptual magnet effect found in earlier experiments with human listeners (e.g. Kuhl, 1991). Recently, Boersma et al. (2013a) implemented a two-layer NN with which they successfully modeled two aspects of human language development: the emergence of phonological categories and auditory dispersion. Here we adopt Boersma et al.'s model in order to further examine whether the phonological categories that it creates for the sounds in its environment are features or phonemes. Below, we first briefly describe Boersma et al.'s model, and subsequently present three different implementations of the model that we have used to investigate feature emergence.

5.2 A NEURAL NETWORK VERSION OF BIDIRECTIONAL PHONETICS AND PHONOLOGY

In this section, we describe Boersma et al.'s (2013a) NN model for phonetics and phonology (BiPhon NN). In BiPhon NN, each phonetic and phonological level of representation corresponds to a set of nodes (a layer). A node can be either active or inactive: a specific phonetic or phonological representation then corresponds to a specific activation pattern in the respective layer. Figure 5.1 shows an example of a two-layer BiPhon NN. As illustrated in the Figure, the bottom layer can be interpreted as an auditory-phonetic continuum, e.g. F₁ ranging from 200 to 1000 Hz. Different F₁ values (of an incoming stimulus or of a produced sound) correspond to different activity patterns in the auditory layer. For instance, Figure 5.1A shows that for a sound with F₁ of about 200 Hz, auditory node 3 (counted from left) is activated most, nodes 2 and 4 are activated a bit less strongly, and nodes 1 and 5 even less; other nodes in the sound layer have activities near zero.¹

The mapping between levels of representation is modeled by excitatory connections: every node in each layer is connected to every node in the neighboring layer. The connections between nodes have specific weights² that determine the extent to which one node's activity will excite the activity of the other node. For instance, in Figure 5.1, there are strong excitatory connections between sound nodes of about 200 Hz and phonology nodes 1, 3, 4 and 9. In line with that, as shown in Figure 5.1A, activity at F₁ nodes near 200 Hz strongly excites phonological nodes 1, 3, 4 and 9. Besides the excitatory connections, the network also contains inhibitory connections between nodes within layers. The weight of an inhibitory connection determines to what extent one node's activity will inhibit the activity of the other node.

The process of passing information across levels of representation is implemented as activity spreading. During activity spreading, node activities can change depending on whether the nodes are clamped (i.e. have fixed activities) or unclamped (i.e. are free to change their activities). For instance, to model speech perception, i.e. the mapping from

¹ This Gaussian-shaped pattern of activity in the auditory F₁ layer is biologically inspired. The nodes along auditory layer could be seen as the human auditory nerve fibers that fire at a high rate to their characteristic frequency, but they also fire – at a lower rate – to frequencies similar to their characteristic frequency (see e.g. Delgutte, 1997: Fig 2). In the present simulations, the amount of activity at an auditory node i is defined as: $a_i = e^{-\frac{0.5(i-F_1)^2}{0.6084}}$, where i is the index of the current auditory node ranging from 1 to n , and F₁ is the F₁ value of the stimulus measured on a scale from 1 to n ; with n being the number of auditory nodes.

² Here we only refer to an adult-like network, i.e. to a network that has already acquired specific connection weights. In the network's initial state, i.e. before learning, all connections have random low weights. For description of learning, see Section 5.3

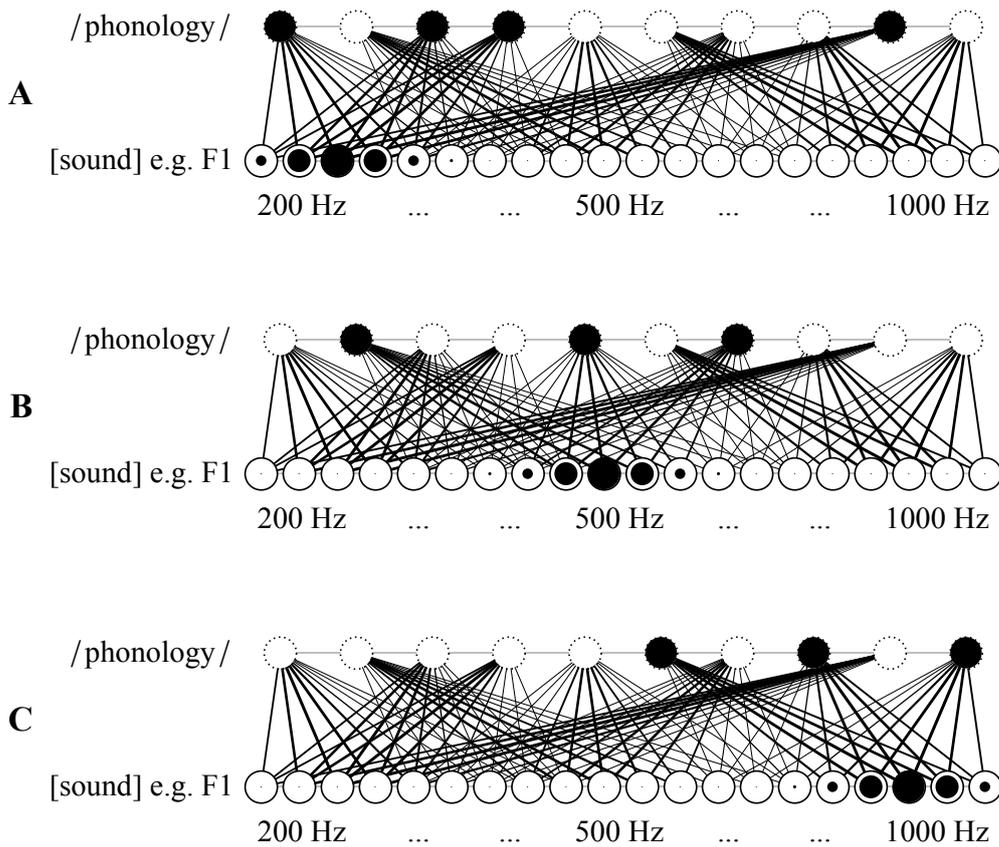


Figure 5.1: An example of a two-layer BiPhon neural network (see Boersma et al., 2013a). The bottom layer corresponds to the auditory F1 dimension and consists of 20 nodes (circles), and the top layer corresponds to phonology and consists of 10 nodes. Excitatory connections between layers are drawn as black lines, inhibitory connections within layers as grey lines. Strong connections are drawn by thick lines, weak ones by thin lines. Clamped nodes are drawn as solid-line circles, unclamped ones as dotted-line circles. An active node is shown as a black filled circle, an inactive node as a white circle. The figure models perception of three different F1 values.

the auditory to the phonological layer, we clamp the nodes at the auditory layer (i.e. the input layer) and leave the nodes at the phonology layer (i.e. the output layer) unclamped. The auditory nodes are clamped because their activity is determined by the physical stimulus properties and not by the activity of other nodes in the network. By contrast, the activities of the nodes in the phonological layer will be affected by the activities of other nodes in the network. To compute how the network would perceive an incoming F1 value, we switch on the nodes for the desired F1 value and let activity spread towards the unclamped nodes in the phonology layer.

Initially, the activity of the unclamped nodes starts at zero and is subsequently updated in several hundreds of small steps, during which ac-

tivity spreads towards each unclamped node j from its neighbors i . The amount by which the activity of the unclamped node j will be updated at every step is computed as follows:

$$\Delta e_j = \eta_a \left(\sum_{\text{connected nodes } i} w_{ij} a_i - e_j \right) \quad (5.1)$$

where η_a is the spreading rate, w_{ij} is the weight of connection between nodes i and j , a_i is the activity of node i , and e_j is the current excitation of node j .

After activity spreading, we can examine the pattern of activity in the output layer, which represents the phonological category onto which the network maps the auditory F1 value. Figure 5.1 shows three examples of mapping different F1 values to different phonological categories. As has been noted above, Figure 5.1A shows that a 200 Hz-input activates phonological nodes 1, 3, 4 and 9; similarly, an F1 of about 500 Hz activates phonological nodes 2, 5, and 7 (Figure 5.1B), and an F1 of about 1000 Hz activates phonological nodes 6, 8 and 10 (Figure 5.1C). The three different activity patterns in the top layer shown in Figure 1A–C can be interpreted as three different phonological representations, e.g. three phonemes or three vowel height categories.

Boersma et al. (2013a) demonstrated that a BiPhon NN with two layers, an auditory and a phonological layer, can handle category creation.³ Specifically, Boersma et al. (2013a) showed that from a continuous distribution with three peaks at the auditory level, the network learned to represent three discrete categories at the surface level.⁴ Thus, the network exhibited distributional category learning, a mechanism that appears to be employed in language acquisition by human infants and adults (Escudero et al., 2011; Maye et al., 2008).

Since the BiPhon NN can handle category creation, we have employed it here to investigate the emergence of phonological features. Specifically, we examined whether the network can learn to represent a 5-vowel system in terms of the phonological features height and backness. The following sections describe three different implementations of the BiPhon NN with which we addressed feature emergence (Boersma and Chládková, 2013b,c; Boersma et al., 2013b).

³ Boersma et al. (2013a) further showed that a BiPhon NN with an additional articulatory layer successfully models auditory dispersion.

⁴ See also Benders (2013) who modeled category emergence in a BiPhon NN with two auditory layers representing two different auditory continua.

5.3 FIRST MODEL: SEPARATE AUDITORY DIMENSIONS FOR F1 AND F2

Boersma and Chládková (2013b) implemented a BiPhon NN with three levels: sound, phonology, and meaning. The authors trained the three-layer network with sound-meaning pairs from a 5-vowel language and showed that in the hidden phonological layer the virtual learner created discrete feature-like representations. In this section, we first describe the architecture of Boersma and Chládková's model, and subsequently report on a replication of their simulations.

Figure 5.2 shows the architecture of Boersma and Chládková's three-layer network. The bottom layer corresponds to the auditory level of representation, the top layer to the lexicon, and the middle layer to the phonology. The bottom layer is split into halves that thereby represent two separate phonetic dimensions, namely, F1 and F2. This split into halves is also reflected in the phonology layer. In the lexicon layer, there are five quadruplets of nodes that represent the five meanings 'I', 'E', 'A', 'O', and 'U', respectively. There are excitatory connections between neighboring layers. The left and the right part of the sound layer are connected to the left and to the right part of the phonology layer, respectively. Both parts of the phonology are connected to the whole meaning layer. Besides the excitatory connections between layers, there are also inhibitory connections within layers.⁵ Since there are two sets of nodes in the bottom and middle layer and one set of nodes in the top layer, we further refer to this network architecture as 2-2-1.

This network becomes a learner of a toy language with 5 meanings, 'I', 'E', 'A', 'O', and 'U', which are pronounced as [i], [e], [a], [o], and [u], respectively. In this language, the 5-way meaning contrast is thus realized through systematic variation along two separate phonetic dimensions: F1 and F2. Specifically, the meaning 'I' is paired with a low value on the F1 dimension and a high value on the F2 dimension, 'E' with medium F1 and high F2, 'A' with high F1 and medium F2, 'O' with medium F1 and low F2, and 'U' with low F1 and low F2. The F1 and F2 values for the five meanings are listed in Table 5.1.

Following Boersma et al. (2013a), in the learner's initial state all connections have random low weights. To simulate supervised learning, the network is fed with sound-meaning pairs. At each learning step, one random meaning out of 'I', 'E', 'A', 'O', and 'U' is selected and the respective nodes in the meaning and the sound layer are switched on,⁶

⁵ In the learning simulations reported here and in the next two sections, weights of the excitatory connections could range from 0 to +1. Weights of the inhibitory connections were fixed at -0.1 (lexicon and sound layer) and -0.25 (phonology layer).

⁶ At each learning step, the input F1 and F2 values were drawn from a Gaussian distribution defined by the mean F1 and F2 values for that meaning and their standard deviations as shown in Table 5.1.

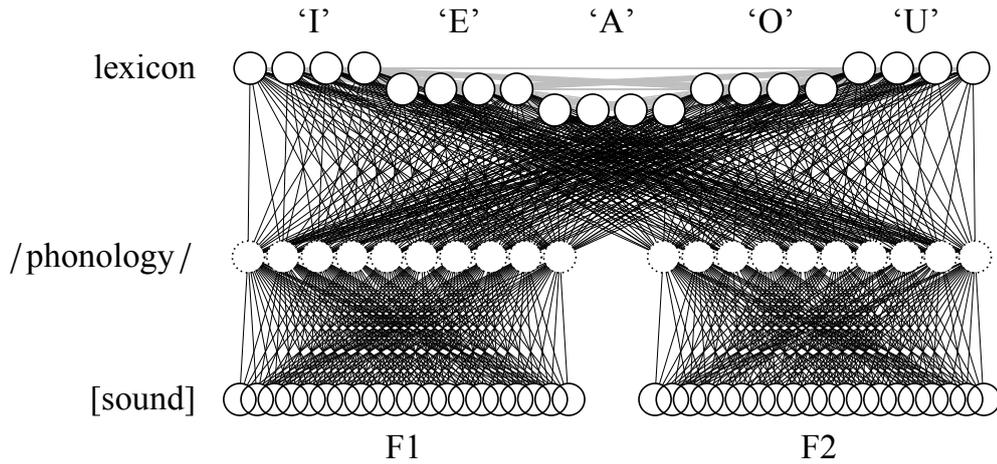


Figure 5.2: Architecture of the network modeled in Section 5.3 (referred to as the 2-2-1 network). The figure shows the network in an initial stage before learning: all the excitatory connections have random low weights, and none of the nodes is active.

	'I'	'E'	'A'	'O'	'U'
F1	4	10	16	10	4
F2	36	33.5	30	26.5	24

Table 5.1: Mean F1 and F2 values corresponding to the five meanings of our toy language. The formant values are defined on a scale from 1 to 40. In the network from Section 5.3, the F1 dimension ranged from 1 to 20, and F2 ranged from 21 to 40. In the networks from Section 5.4 and Section 5.5, a single frequency dimension for F1 and F2 ranged from 1 to 40. The standard deviation of all the means is 1.95.

while all the other meaning and sound nodes are off. Subsequently, the activity is allowed to spread from the meaning and sound layers to the phonological layer according to Equation 5.1, in 500 small steps. In the present stimulations, the spreading rate is kept constant at 0.01.

After activity spreading, each connection updates its weight according to the Hebbian-inspired (Hebb, 1949) inoutstar learning algorithm of Boersma et al. (2013a): a connection is strengthened if both its nodes are on, unchanged if both nodes are off, and weakened if one node is on and the other off. The formula defining the amount by which the connection weight will change is as follows:

$$\Delta w_{ij} = \eta_w \left(a_i a_j - \frac{a_j w_{ij} - a_i w_{ij} - w_{ij}}{2} \right) \quad (5.2)$$

where η_w is the learning rate (which was set to 0.001 in the present stimulations), a_i is the activity of the input node, a_j is the activity of the

output node, and w_{ij} is the current weight of the connection between the input and output node.

After the update of the weights, the connection weights are normalized to ensure that the sum of weights incoming to node i from all nodes j at a neighboring lower layer be maintained at a fixed value (see Rumelhart and Zipser, 1985). Weight normalization is thus formalized as:

$$w_{ij \text{ normalized}} = k \frac{w_{ij}}{\sum_{j=1}^n w_{ij}} \quad (5.3)$$

where w_{ij} is the current weight of the connection between node i and node j from a neighboring lower layer, and k is the value at which the sum of connection weights incoming to node i from the neighboring lower layer j is fixed. Here, $k = 0.1 \cdot n$, where n is the number of nodes in layer j .

Each learning step thus consists of feeding the network with a random sound–meaning pair, spreading of activity, updating and subsequently normalizing connection weights. After a sufficient number of learning steps (40,000 in the present simulation), the connection weights come to exhibit a stable pattern that does not change with further learning.

After learning, we can examine how the learner would produce the 5 meanings of the language that she was trained on. To simulate production, for each meaning the relevant nodes in the meaning layer are switched on and the activity is allowed to spread through the network, i.e. to the phonology and the sound layer. Figure 5.3 shows how the learner produces each of the 5 words. It is seen that the 5 meanings are produced correctly, i.e. in a ‘parent-like’ way, e.g. the meaning ‘I’ is produced with low F1 and high F2, and ‘A’ with high F1 and medium F2.

More interestingly, the activity patterns at the middle layer reveal whether the representations that emerge in the hidden phonology layer are phoneme- or feature-like. As is seen in Figure 5.3, some nodes in the phonological layer display an activity pattern that resembles features. Specifically, in the phonology layer above the auditory F1 dimension, nodes 5 and 7 are shared by ‘I’ and ‘U’⁷, and nodes 2, 3 and 4 are shared by ‘E’ and ‘O’. Similarly, in the phonology layer above F2, nodes 11, 14 and 17 are shared by ‘I’ and ‘E’ and nodes 16 and 20 by ‘U’ and ‘O’. These activation patterns indicate that the learner has created height and backness features in her phonology: nodes 5 and 7 correspond to

⁷ In the present model, node activity ranged from 0 to 1. In our evaluations of phonological patterns we consider as ‘active’ only nodes with the maximum activity of 1. In Figure 5.3, notice the partially activated phonology node 1 for ‘I’ and the partially activated node 10 for ‘U’. Despite the lack of strong connections to the respective meanings, these two phonological-layer nodes exhibit some activity, which is caused by a bottom-up spreading of activity from the sound layer.

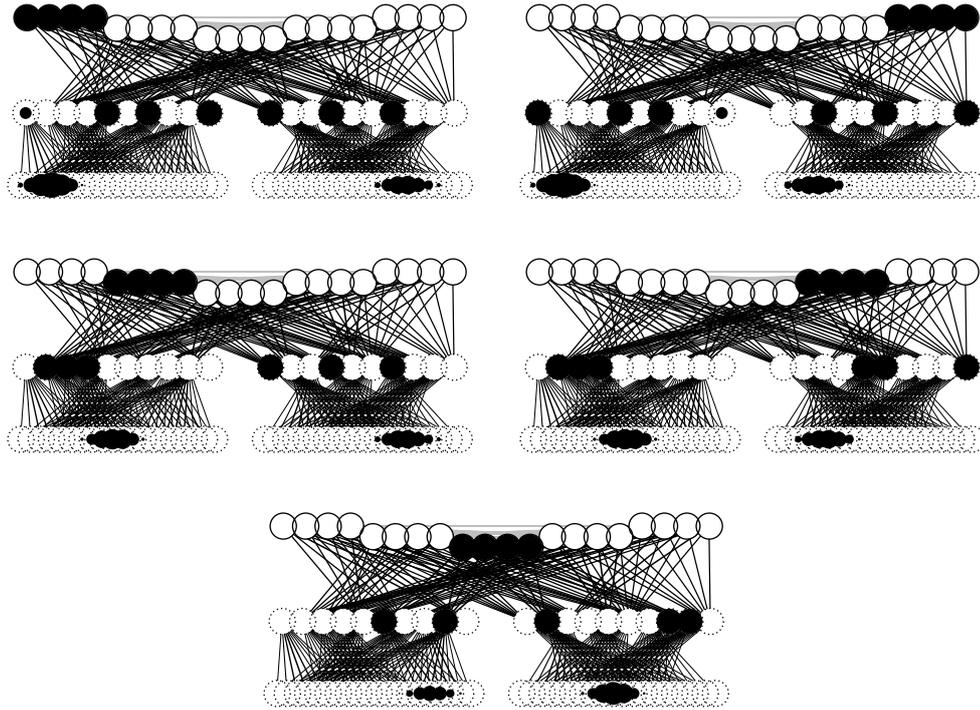


Figure 5.3: A 2-2-1 learner's production of the acquired 5-vowel language: top left = production of 'I', middle left = 'E', bottom = 'A', middle right = 'O', top right = 'U'. The top layer is clamped, relevant meaning nodes for one of the 5 meanings are switched on and activity spreads to the phonology and the sound layer. Each meaning is produced with parent-like formant values. The phonology layer exhibits mostly feature-specific activity patterns.

the feature high, nodes 2, 3 and 4 to the feature mid, nodes 11, 14 and 17 to the feature front, and nodes 16 and 20 to the feature back. Interestingly, note also that some phonology nodes seem to be phoneme-specific: for instance, node 13 is activated by the meaning 'U' only, and node 15 only by 'O'. The meaning 'A' activates 5 phonological nodes that are not shared with any of the remaining four meanings: nodes 6, 9, 12, 18 and 19, which can be interpreted as either feature-specific nodes for low and/or central, or as phoneme-specific nodes.

Note that in terms of phonological features, the 5-vowel system of our toy language can be fully specified by four feature categories: high, mid, front, and back, and by at least one other category for 'A' (i.e. low and/or central). To that end, we can conclude that the learner in Figure 5.3 has created feature representations with which she sufficiently represents her whole vowel system.⁸

⁸ Our present result is slightly different from that of Boersma and Chládková (2013b). The phonological representations that emerged in Boersma and Chládková were *exclusively* feature-like, that is, there were three distinct activity patterns in each part of the

We simulated a total of ten 2-2-1 learners, whose acquired phonological patterns are summarized in Figure 5.4. It can be seen that nine learners created representations for all four features: high, mid, front and back; the remaining learner (number 3) created representations for three of these features. Pooling across learners shows that the average number of feature representations emerging in the 2-2-1 model is 3.9. The phonologies of most learners also exhibit phoneme-specific activity patterns for at least four of the five meanings 'I', 'E', 'A', 'O', and 'U'; an exception is learner 9 who has phoneme-specific representations for only 3 meanings. We can thus conclude that in the 2-2-1 model, features as well as phonemes have emerged.⁹

5.4 SECOND MODEL: A SINGLE AUDITORY DIMENSION FOR FORMANT FREQUENCY

In the 2-2-1 model described in the previous section, vowel F1 and F2 were each represented at a separate set of nodes at the bottom sound layer, each of which was paired with a separate set of nodes at the middle phonology layer. The split of the sound layer into two would be plausible for two physically different dimensions such as duration and pitch. However, F1 and F2 are in fact values along a single phonetic continuum, i.e. frequency. Since the human auditory system contains a single basilar membrane along which it represents the whole frequency range, a network with single layer for frequency is biologically more plausible than a network with separate layers for each formant frequency. Therefore, Boersma et al. (2013b) proposed a potentially more realistic version of the three-level BiPhon NN in which both F1 and F2 are represented at the same phonetic dimension, i.e. at a single sound layer. The authors showed that in such a model with a single sound and a single phonology layer, the phonological representations that emerge are phoneme-like. Here we replicated the architecture of the Boersma et al.'s (2013b) network and used it to simulate vowel learning.

As has been noted above, the architecture of the second network is identical to the first one except that the bottom and the middle layer are

phonology layer. Each of the 5 meanings was then represented as a combination of one height and one backness feature and there were no phoneme-specific nodes. The difference between Boersma and Chládková's result and the one reported here is due to a difference in parameter settings of the two simulations: most notably, here we applied weight normalization, which was not used in the previous simulation. Weight normalization was included here in order to make the 2-2-1 simulation comparable to the 1-1-1 and 1-1-2 simulations (presented in the following sections), which both applied weight normalization.

⁹ Note that when the number of learning steps is reduced to 20.000, the simulations yield a similar result: both features and phonemes emerge and the average number of features is 3.7. Since some learners did not arrive at a complete equilibrium at 20.000 steps, we report here the results after 40.000 steps.

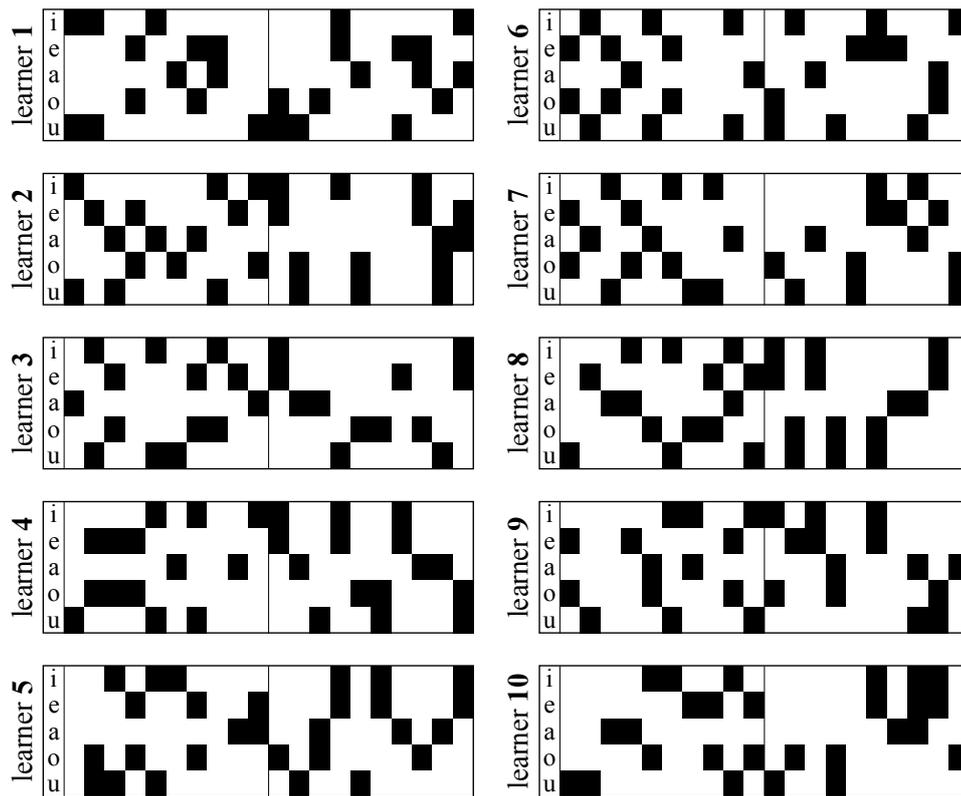


Figure 5.4: The results of 10 simulations of the 2-2-1 learners (learner 4 is the learner from Figure 5.3). Each box shows the activity pattern at the phonology layer during vowel production. Within each box, rows represent the 5 meanings. Columns represent the 20 phonology-layer nodes; vertical lines mark the split between the left and the right half of the phonology. Black squares mark fully activated nodes, i.e. with the maximum possible activity of 1. As can be deduced by averaging over learners, out of the four features front, back, high and, mid, an average 2-2-1 learner comes to represent 3.9 features in her phonology.

not split into two halves; we thus refer to this network architecture as a 1-1-1 model. As is seen in Figure 5.5, every node in the sound layer is connected to every node in the middle layer (which was not the case in the previous 2-2-1 model). Learning is implemented identically to the first simulation, that is, the virtual infant again learns the same 5-vowel language from sound-meaning pairs (in this simulation, an equilibrium was reached at 20.000 learning steps). To compare across models we kept all parameter settings of the present 1-1-1 simulations for activity spreading, weight updates, and weight normalization identical to the parameter settings of the 2-2-1 simulations.

Figure 5.6 shows that the 1-1-1 learner has successfully acquired her language: she produces the 5 meanings correctly, i.e. in a parent-like

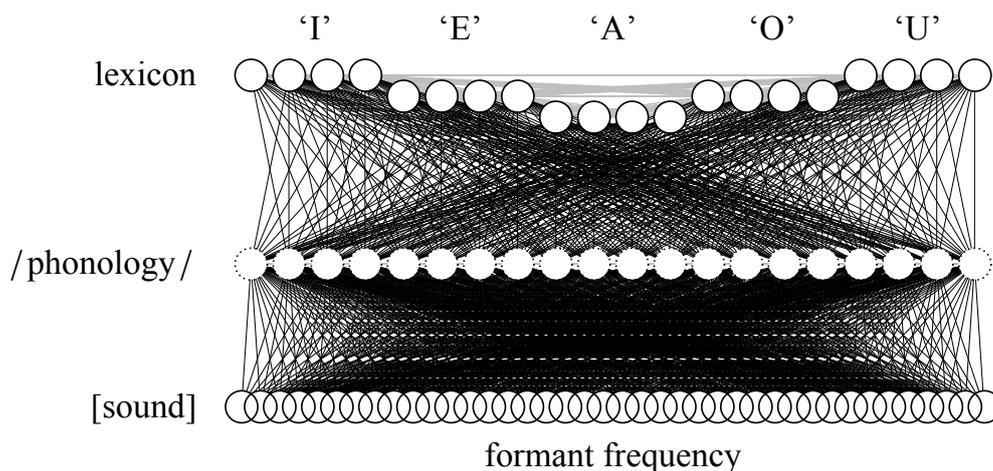


Figure 5.5: Architecture of the network modeled in Section 5.4 (referred to as the 1-1-1 network). The figure shows the network in an initial stage before learning: all excitatory connections have random low weights, and none of the nodes is active.

way. Unlike in the previous 2-2-1 model, most activity patterns in the phonology are phoneme-like; exceptions are nodes 6 and 10, which are reminiscent of the features front and back respectively.

We simulated a total of 10 learners whose resulting phonological structures are shown in Figure 5.7. It is seen that the acquired phonological representations are mostly phonemes and only to a limited extent features. Pooling across learners shows that the average number of feature representations is 1.8. As discussed in the previous section, at least four distinct feature-like activity patterns are needed to fully specify the 5-vowel system. Since none of the learners acquired representations for 4 features, we conclude that the phonological categories emerging in 1-1-1 learners are phoneme-specific.

5.5 THIRD MODEL: ADDING THE KNOWLEDGE OF ALLOMORPHY

From the results of the 1-1-1 simulations, it appears that phonological features typically associated with vowel F_1 and F_2 , i.e. height and backness, do not emerge from phonetic information alone. Nevertheless, the grammars of adult 5-vowel language speakers do reflect feature-based phonological structures through phenomena such as morphophonological alternations. If features do not emerge on the basis of phonetic information alone, it is possible that they emerge once the learner has acquired some morphological knowledge. The model presented in this section thus addresses the question of whether experience with morphological alternations helps the learner create feature-like representations

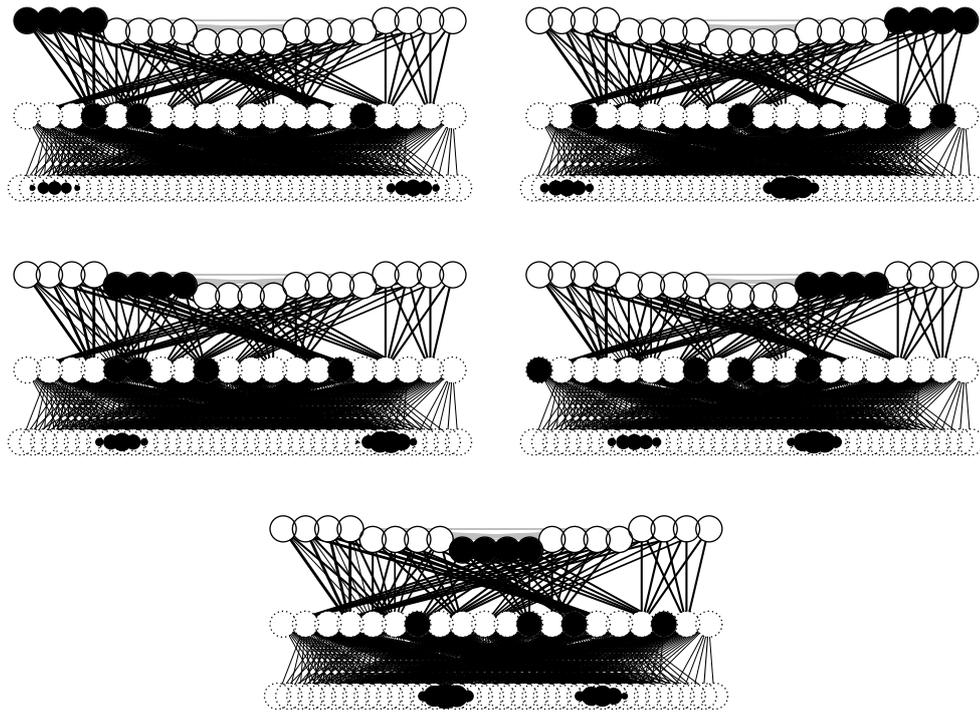


Figure 5.6: A 1-1-1 learner's production of the acquired 5-vowel language: top left = production of 'I', middle left = 'E', bottom = 'A', middle right = 'O', top right = 'U'. The top layer is clamped, relevant meaning nodes for one of the 5 meanings are switched on and activity spreads to the phonology and the sound layer. Each meaning is produced with parent-like formant values. The phonology layer mostly shows phoneme-specific activity patterns.

(see Boersma and Chládková, 2013c, for a slightly different version of the model).

In typical 5-vowel languages, one finds morphological alternations between phonologically high and mid vowels that share backness. For instance, in Spanish, the mid vowels /e/ and /o/ alternate with the high vowels /i/ and /u/ respectively in stems of conjugated *-ir* verbs as in examples 1a and 1b below. Likewise, in some noun stems in Czech we observe alternations between high front /i/ and mid front /ɛ/, and between high back /u/ and mid back /o/, which are due to e.g. adjectivization, verbalization, pluralization or declension; see examples 2a and 2b.

- (1) a. /sentir/ 'feel, inf.' > /sinti'o/ 'felt, 3 sg.'
 b. /dormir/ 'sleep, inf.' > /durmi'o/ 'slept, 3 sg.'
- (2) a. /ji:ra/ 'hole' > /jɛravi:/ 'holey'
 b. /du:m/ 'residence' > /domovji:/ 'residential'

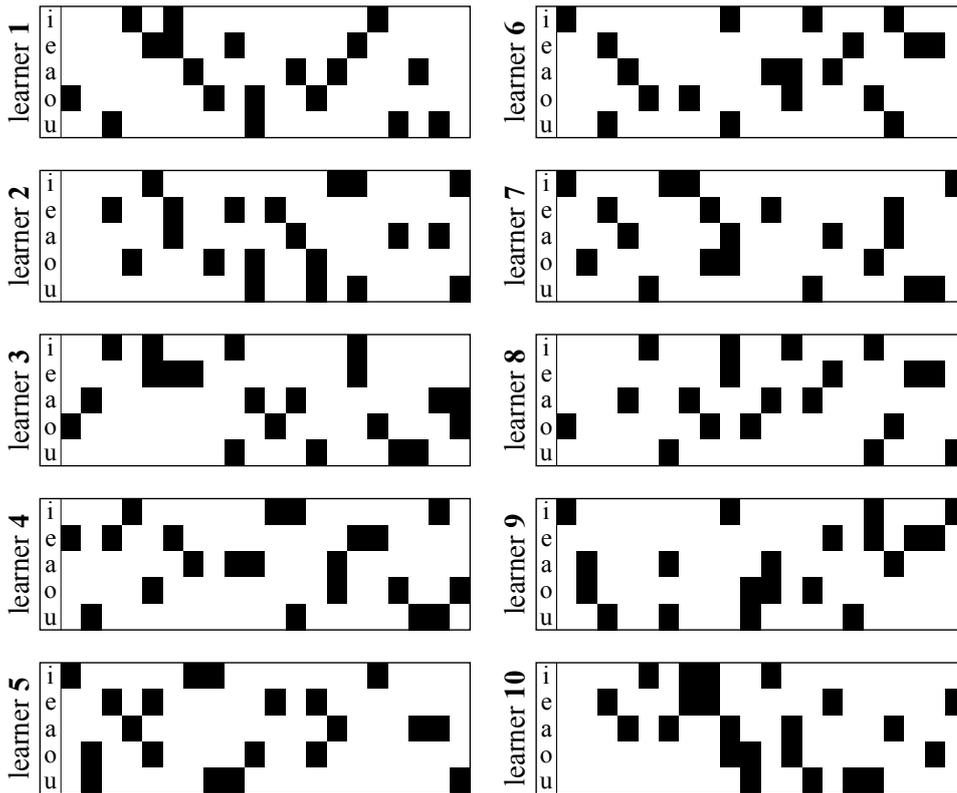


Figure 5.7: The results of 10 simulations of the 1-1-1 learners (learner 1 is the learner from Figure 5.6). Each box shows the activity pattern at the phonology layer during vowel production. Within each box, rows represent the 5 meanings. Columns represent the 20 nodes in the phonology layer. Black squares mark fully activated nodes, i.e. with the maximum possible activity of 1. Out of the four features front, back, high and, mid, an average 1-1-1 learner comes to represent only 1.8 features in her phonology.

Our toy language thus resembles typical 5-vowel languages like Spanish or Czech not only in that it contrasts the 5 word meanings 'I' 'E' 'A' 'O' 'U' that are realized as [i], [e], [a], [o], and [u], but also in that it contains morphological contexts in which [i] and [u] change into [e] and [o] respectively (or vice versa). Such knowledge of morphological alternations was added to the 1-1-1 model described in the preceding section. The morphological alternations were added to the top layer. As shown in Figure 5.8, the top layer, i.e. the lexicon, is now split into two parts, the left part representing the unanalyzed meanings (i.e. lexicon of words), and the right part representing all the component meanings (i.e. lexicon of morphemes). Within the meaning layer and within the morphology layer, there are inhibitory connections between nodes, but the two layers are not connected to each other. We refer to the present network architecture with a split top layer as the 1-1-2 model.

The lexicon thus contains the 5 unanalyzed meanings 'I' 'E' 'A' 'O' 'U' in the word layer, and 3 stem morphemes: <ie>, <uo>, <a> and 2 suffixes <sg> and <pl> in the morpheme layer. In our toy language, the word 'I' is composed of the stem morpheme <ie> and the suffix <sg>, 'E' is composed of <ie> and <pl>, 'U' is composed of <uo> and <sg>, 'O' is composed of <uo> and <pl>, and 'A' is composed of <a> and both <sg> and <pl> since there are no morphologically conditioned alternations affecting 'A'.

As shown in Figure 5.8, the word layer is connected to the phonology by excitatory connections from the initial stage of learning. The morphology layer becomes connected to the phonology by excitatory connections as well, but only at a later stage of learning.

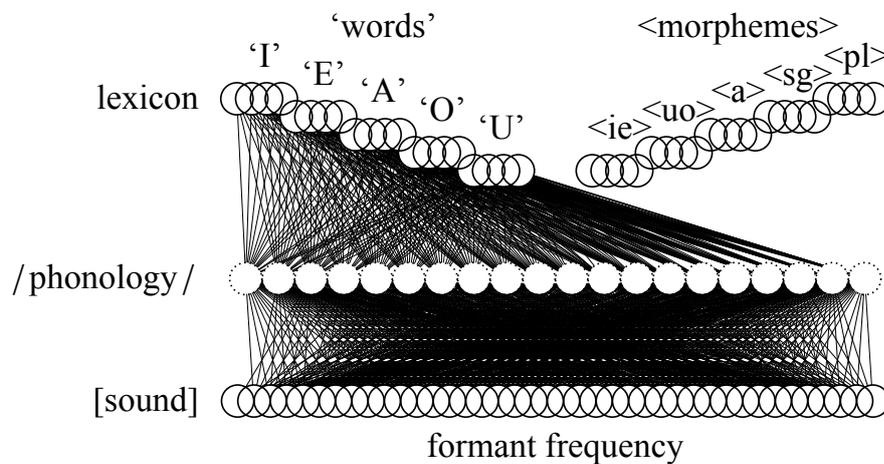


Figure 5.8: Architecture of the network modeled in Section 5.5 (referred to as the 1-1-2 network). The figure shows the network in an initial stage before learning: the connections have random low weights, none of the nodes is active, and morphology is not yet connected to the phonology.

Learning is thus implemented in two stages. The first stage consists of 20,000 learning steps and is identical to the previous two simulations: that is, the virtual infant starts with learning from sound-(word)meaning pairs. Importantly, we assume that after step 20,000 the learner will have acquired the ability to analyze words into morphemes. Thus, during the second stage, which begins at step 20,001 and consists of another 40,000 learning steps,¹⁰ the network learns from sound-word-morpheme

¹⁰ The number of learning steps assigned to each learning stage was based on the two previous models. Stage 1 contains 20,000 steps because that was the number of steps at which learners in the 1-1-1 model reached equilibrium and acquired their phonemes. Stage 2 contains 40,000 steps because that was the number of steps at which learners in the 2-2-1 model reached equilibrium and acquired their features. Thus, if the ac-

triplets. A learning step from a sound-word-morpheme triplet proceeds as follows: one random meaning out of 'I', 'E', 'A', 'O', and 'U' is selected and the respective nodes in the whole lexicon, i.e. in both the word and the morpheme layer, and the respective nodes in the sound layer are switched on. Subsequently, activity is allowed to spread from the clamped word, morpheme, and sound layers. The parameter settings for activity spreading, weight updates and weight normalization are identical to the parameter settings from the previous two models.

Figure 5.9 shows that a 1-1-2 learner successfully acquires her language: she produces the 5 words correctly. It is seen that the learner has created feature-like representations in her phonology. Specifically, nodes 9 and 11 are shared (exclusively) by 'I' and 'U' and thus correspond to the feature high, node 1 is shared by 'E' and 'O' and thus corresponds to the feature mid, nodes 7 and 13 are shared by 'I' and 'E' and thus correspond to the feature front, and node 15 is shared by 'U' and 'O' and thus corresponds to the feature back. Note that four of the five meanings also have their own phoneme-specific nodes: node 3 is specific for 'E', nodes 4, 6, 12 and 17 for 'A', nodes 2 and 16 for 'O', and nodes 5 and 8 for 'U'.

As with the previous two models, we simulated ten 1-1-2 learners; Figure 5.10 shows the activity patterns in their phonology layers. The Figure shows that eight learners formed feature representations for all four features (i.e. high, mid, front and back), and two learners formed feature representations for three features. The average number of emerged feature representations in the present 1-1-2 model is thus 3.8. It can be seen that besides feature representations, the learners created also phoneme representations for, on average, 3.5 phonemes. We conclude that the phonological representations that emerge in a 1-1-2 network are both features and phonemes.¹¹

5.6 DISCUSSION AND CONCLUSION

In the simulations described in this chapter, we investigated whether the phonological representations that language learners acquire for a 5-

quisition of phonological representations consists of a phoneme- and a feature-stage, 20.000 and 40.000 may be about the right amount of steps required at the two stages respectively.

¹¹ Note that when the total number of learning steps is reduced to 20.000 (i.e. 10.000 at either stage), the simulations yield a similar result: both phoneme and feature representations emerge and the average number of features is 3.8. As it is not clear when learners start analyzing words into morphemes, we model here a case in which morphological knowledge comes in as soon as some phonological representations have been created without the involvement of morphology. Thus, since the 1-1-1 learners needed 20.000 steps to create stable phonological representations (namely phonemes), we define the onset of morphological knowledge as the 20.001th step. Subsequently, we let the network learn 40.000 more times to allow at least as much time for potential feature creation as was needed by the 2-2-1 learners.

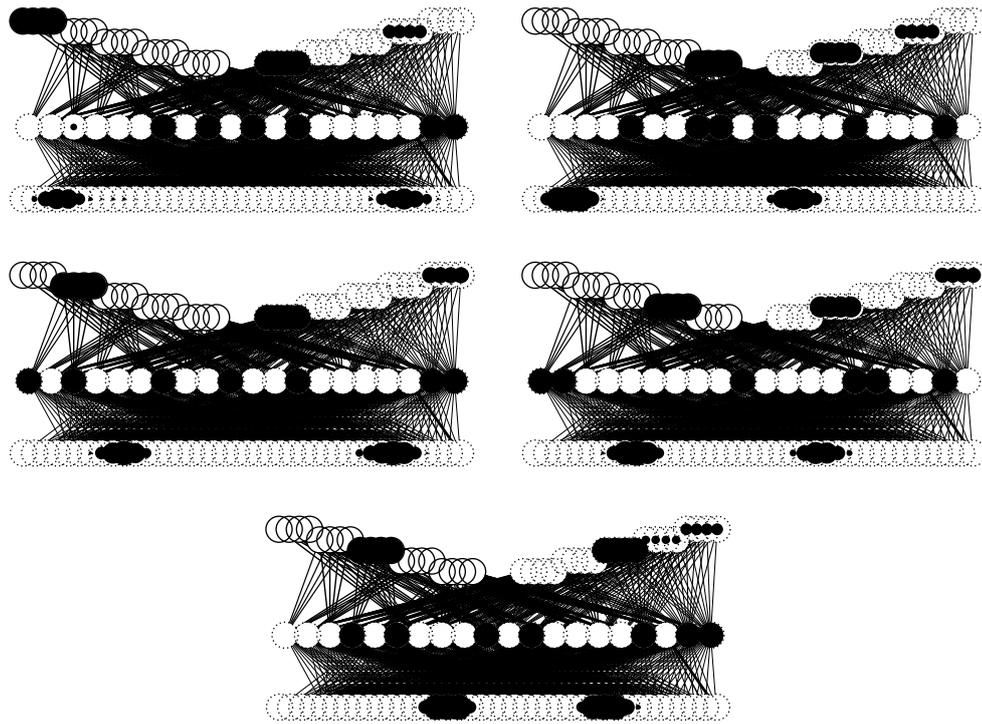


Figure 5.9: A 1-1-2 learner's production of the acquired 5-vowel language: top left = production of 'I', middle left = 'E', bottom = 'A', middle right = 'O', top right = 'U'. The word layer (top left) is clamped, relevant meaning nodes for one of the 5 words are switched on and activity spreads to through the network. The activities in the sound layer show that each word is produced with parent-like formant values. The activities in the phonology layer show that the learner created mostly feature-specific but also phoneme-specific representations. Note also that the morphology layer (top right) is unclamped: the activities of the moprheme nodes show that the learner can analyze the words she produces into the correct morphemes.

vowel system are features or phonemes. To that end, we employed a neural network with three layers of nodes corresponding to three levels of representation: sound, phonology and lexicon. We implemented three different versions of the model and simulated supervised (or, lexicon-driven) learning.

In the first model, F1 and F2 were represented on separate auditory layers that were in turn mapped onto separate sets of nodes in the phonology layer. We showed that with such a network architecture, the representations that emerged in the phonology were both features and phonemes. The average number of features that emerged in the first model was 3.9 (out of 4, which was considered the number of features necessary to fully specify the 5-vowel system). The second model that we implemented was biologically more realistic than the first model in

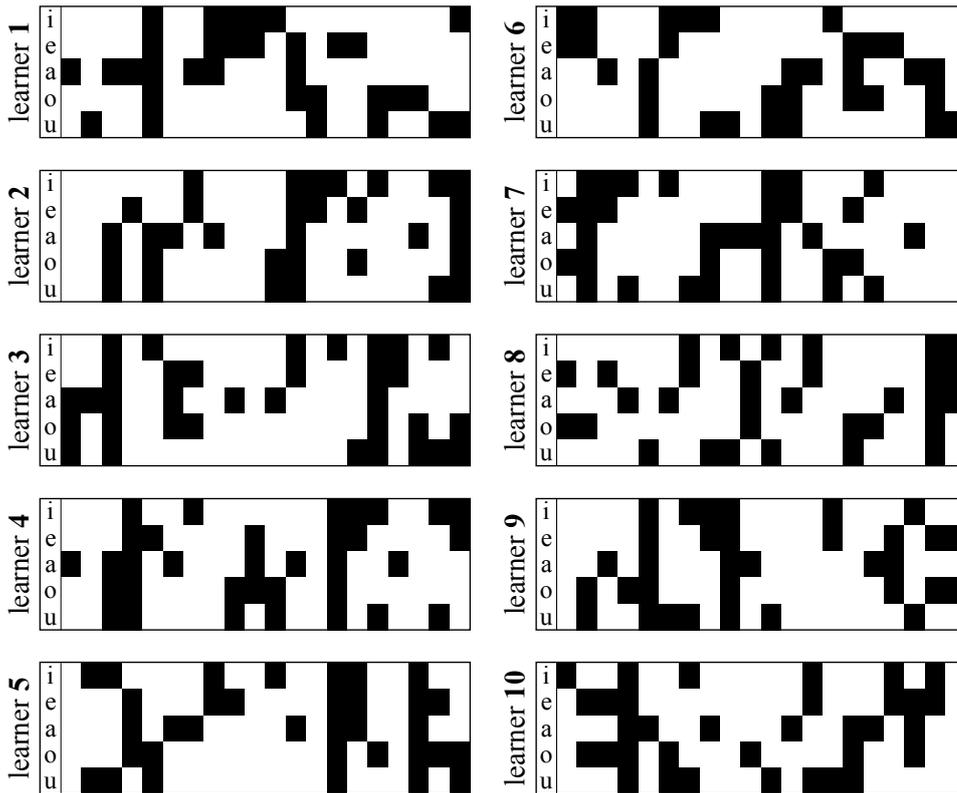


Figure 5.10: The results of 10 simulations of the 1-1-2 learners (learner 8 is the learner from Figure 5.9). Each box shows the activity pattern at the phonology layer during vowel production. Within each box, rows represent the 5 meanings. Columns represent the 20 nodes in the phonology layer. Black squares mark fully activated nodes, i.e. with the maximum possible activity of 1. Out of the four features front, back, high and, mid, an average 1-1-2 learner comes to represent 3.8 features in her phonology.

that it did not enforce a separation of F1 and F2, which are in fact values along a single auditory dimension. F1 and F2 were thus represented at a single auditory layer. The second model yielded mostly phoneme-like representations; the average number of feature-like representations was only 1.8. In the third model, we kept the structure of the sound and phonology layer identical to that of the second model. This time, we added morphological knowledge to the lexicon. The phonological representations that emerged in the third model were both phonemes and features. The average number of features that emerged in the third model was 3.8, which (as in the first model) can be considered a sufficient number of features to fully represent the 5-vowel system of our learners' language.

A comparison of the second and the third model indicates that phonetic information alone may not be enough for feature representations

to emerge in the phonology. Once the knowledge of morphologically conditioned vowel alternations, i.e. allomorphy, is added to the model, features start to emerge.¹²

The implementation of an initial ‘meaning-driven’ and a subsequent ‘meaning-and-morphology-driven’ learning is inspired by human language development: the literature shows that children first acquire the items of their language as unanalyzed words, and only later learn to analyze the lexical items into their component parts, i.e. morphemes (Berko, 1958). Relatedly, for phonological acquisition, Hayes (2004) proposed that the child first learns to differentiate native phonemic contrasts and later also acquires the knowledge of morphophonological alternations. Therefore, a model that also captures the morphological knowledge is potentially a more realistic implementation of human language acquisition than a model that only assumes knowledge of unanalyzed meanings and vowels’ phonetic properties. We thus consider the third model to be the closest approximation of the human language acquisition system.

The finding that phonetic information alone may not yield feature representations but that adding morphological knowledge boosts feature emergence is in line with phonological acquisition studies with humans. For instance, Fikkert and Freitas (2006) showed that the knowledge of vowel alternations helps European Portuguese children acquire the phonological features that characterize their native vowels. Similarly, Drescher (2004) argued that Manchu learners can acquire phonological feature specifications for their native vowels if they consider not only the vowels’ phonetic properties but also the phonological processes in which the vowels participate. It seems plausible that at the initial stages of acquisition, learners form phonetically motivated phonological representations which are phonemes (as in our second model). Later, with the development of a more abstract morphological knowledge, learners start representing features (as in our third model). Such gradual emergence of feature representations has also been observed in the productions of human language learners (see e.g. Menn and Vihman, 2011).

The present results show that simulated learners create phonological feature representations for their vowels. Importantly, feature representations emerge when the learners have access not only to the phonetic but also to the morphophonological evidence for feature patterns. Our findings also suggest that language users come to represent both features and phonemes in their phonology.¹³ To sum up, the present findings indicate that phonological features are emergent categories that learners

¹² Interestingly, even with a markedly increased number of feature representations (from 1.8 to 3.8), phoneme-specific representations were still preserved.

¹³ With the present simulations we can only conclude that both features and phonemes emerge in language learners’ phonology. Since we only modeled a single phonological level, we cannot determine whether features and phonemes would be represented at

acquire from experience with their phonetic and morphological environment.

different levels if the model contained more than one hidden phonological layer. This remains to be addressed in future work.

GENERAL DISCUSSION AND CONCLUSIONS

In this thesis we investigated the perceptual bases of phonological features from different perspectives: from that of an adult listener who has a fully acquired language system in place, from that of a linguist who aims to uncover feature structures in languages, and from the perspective of a learner who acquires the phonological representations for her native speech sounds. Specifically, we employed discrimination and identification experiments to determine whether phonological features are the categories through which adult listeners process the speech signal (Chapter 2 and Chapter 3). Subsequently, we assessed listeners' pre-attentive sensitivity to a particular phonetic dimension in order to find out whether they encode that dimension in terms of a phonological feature (Chapter 4). Finally, we carried out simulations of perceptually driven learning to reveal whether a virtual learner comes to represent the sounds of her language in terms of phonological features (Chapter 5).

Chapter 2 and Chapter 3 showed that adult listeners map the auditory properties of speech sounds onto phonological feature categories and Chapter 3 further indicated that within a single language the perceptual mappings between sound and phonological features can be redefined when sound change occurs. Chapter 4 assessed listeners' perceptual patterns and revealed that a phonetic dimension that is used contrastively in a particular language is not necessarily encoded in terms of a phonological feature, but can be associated with a specific phoneme category instead. The simulations of vowel learning presented in Chapter 5 seem to have confirmed the combined findings of Chapters 2 through 4 by showing that virtual infants can learn to represent the sounds of their language in terms of both features and phonemes. Importantly, Chapter 5 demonstrated that feature representations are emergent categories: they are created on the basis of the phonetic and morphophonological input that learners are exposed to. The present findings are further discussed in the following sections and visualized in Figure 6.1.

6.1 PHONOLOGICAL FEATURES ARE AT THE INTERFACE WITH PHONETICS

In Chapter 2, we tested whether listeners perceive speech sounds in terms of features or phonemes. The results indicated that in regions of the vowel space where they do not reliably identify any phonemes, listeners still categorize vowels in terms of their native height categories.

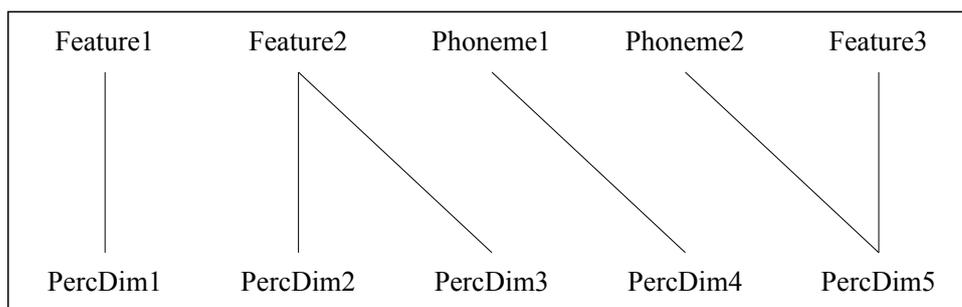


Figure 6.1: The mapping between phonetics (bottom row) and phonology (top row) based on the present findings. The figure shows that phonetics is linked directly to features, and that multiple dimensions can be linked to a single feature (as demonstrated in Chapters 2 and 3 respectively). Also, all the connections are acquired during one's language development and not innate to the learner (as demonstrated in Chapters 3 and 5). Besides features, the phonological level also contains phonemes: that is, there are both sound-feature and sound-phoneme mappings (as suggested by Chapters 4 and 5).

This finding can be interpreted as evidence for a direct mapping between sound and phonological features.

Note that in Chapter 2, in which we investigated the perceptual basis of vowel height, we focused on only one possible phonetic cue to the height feature, namely, the auditory F₁ dimension. In that respect, the literature suggests that the phonetic correlate of vowel height may not be solely the F₁, but for instance the difference between F₁ and F₀ (Diehl and Kluender, 1989; Syrdal and Gopal, 1986). Since in our design, we only varied the F₁ dimension and kept F₀ constant, we cannot rule out the possibility that it is the F₁-F₀ difference that listeners map onto vowel height categories. Relatedly, our finding that F₁ is mapped directly onto phonological height, does not rule out the possibility that there are multiple cues for vowel height (such as, F₁, F₀, F₁-F₀ difference, or even duration) that are either integrated into a single percept for height (e.g. Kingston, 1991) or mapped onto the height feature separately.¹

While Chapter 2 showed that the phonological height feature has at least one direct phonetic correlate, i.e. F₁, the results of Chapter 3 provide evidence for multiple phonetic correlates for phonological features. In Chapter 3, we primarily investigated whether the mapping between auditory dimensions and phonological features is inherent or learned. To that end, we showed that speakers can learn to re-associate a phonological feature with a new phonetic cue and thus link multiple auditory dimensions (e.g. an old and a new one) to a single feature.

¹ Conversely, it is plausible that F₁ is a cue to features other than vowel height (e.g. consonant voicing, see Kingston and Diehl, 1995).

Specifically, in Chapter 3 we focused on the variety of Standard English spoken in Southern England (SESE) in which the phonologically back vowel /u/ has undergone a sound change and is nowadays produced with F2 values similar to those of a phonologically front vowel /i/. In other words, F2, i.e. the dimension traditionally associated with phonological backness, no longer appears to be a reliable cue to at least one SESE front-back contrast, namely /i/-/u/. In Chapter 3 we showed that SESE listeners also rely on another cue, namely diphthongization, to perceptually distinguish their front and back vowels. Importantly, the perceptual use of diphthongization was found even for front-back vowel contrasts that are still sufficiently distinguished by F2, which suggests that the cue re-association has occurred for the backness feature in general rather than for specific phonemes. This generalization effect can be interpreted as further evidence for a direct mapping between phonetic dimensions and phonological features. Importantly, since the backness contrast was still differentiated by F2 as well, it seems that both F2 and diphthongization are mapped onto the backness feature in SESE.

6.2 PERCEPTUAL PATTERNS AND LEARNING SIMULATIONS REVEAL THE NATURE OF PHONOLOGICAL REPRESENTATIONS

On the basis of our finding that listeners map phonetic cues directly onto phonological features, the experiments in Chapter 4 attempted to determine whether Dutch has a phonological feature associated with vowel duration, i.e. the phonological length feature. Using measures of pre-attentive speech sound processing we found that Dutch listeners had a large sensitivity to duration changes when the stimulus had the spectral quality of their native vowel /a:/, which indicated that short and long instances of an [a]-like vowel are perceived as different categories. However, this perceptual categorization of duration in the spectral quality of /a:/ was not generalized to other, native or non-native, spectral qualities. Therefore, we concluded that vowel duration in Dutch is a phoneme-specific cue.

Note that such a proposal of phoneme-specific phonetic cues implies that phonological representations for phonemes exist at the interface with phonetics, i.e. that listeners map the sound directly onto phonemes. Thus, while Chapters 2 and 3 found that listeners map the auditory information onto features, the results of Chapter 4 indicate that listeners map some auditory information onto phonemes. Importantly, these seemingly contradictory results do not disprove one another: that is, Chapters 2 and 3 do not disprove the existence of a direct mapping between sound and phonemes, whereas Chapter 4 does not disprove a direct mapping between sound and features. Therefore, one might ar-

gue that the interface between phonetics and phonology contains *both* features and phonemes.

In order to reveal whether phonetic dimensions that contrast vowels are phonologically encoded in terms of feature or phoneme categories, we implemented in Chapter 5 an artificial neural network model, with which we simulated the phonological acquisition of a hypothetical five-vowel language. Interestingly, the phonological representations that the virtual learner created for her vowels were features as well as phonemes. We thus found both types of representation at the phonological level despite the fact that with the features only, the learner would be able to sufficiently represent all her vowels.

It is plausible that human language users are like the virtual ones from Chapter 5 in that they represent both features and phonemes at the phonological level that interfaces with the phonetics. In line with that, one can then observe evidence for sound-feature mapping (Chapter 2 and Chapter 3) as well as for sound-phoneme mapping (Chapter 4).

6.3 FEATURES ARE ACQUIRED WITH THE HELP OF MORPHOPHONOLOGY BUT HAVE DIRECT CORRELATES IN PHONETICS

In Chapter 5, we demonstrated that phonological features emerge on the basis of learners' phonetic and morphophonological input. It was shown that initially, when the learner has access only to unanalyzed word meanings and phonetic properties of the vowels, she learns to represent the contrastive sounds in her language in terms of phonemes. Later, as the learner acquires the knowledge of morphological structure, she redefines her phonological representations and comes to represent the contrastive sounds of her language in terms of both phonemes and features.

Chapter 5 thus showed that features are emergent but possibly only if higher-level linguistic knowledge is employed. That is, besides phonetic information, the language-learning infant also needs more abstract linguistic knowledge (e.g. morphological) to create phonological representations for features. However, Chapter 2 and Chapter 3 indicate that once feature representations are in place, the adult language user determines features directly on the basis of the phonetic properties of the sounds.

Previous phonetic training and imitation studies with adults provide further evidence for our argument that adult language users indeed have an established direct mapping between phonological features and phonetic signal. For instance, Kraljic and Samuel (2006) trained American English speakers to perceptually identify a stop consonant with a realization ambiguous between that of /d/ and /t/ as either of the two phonemes. Interestingly, listeners exhibited perceptual adaptation

not only for the /d/-/t/ contrast,² but also for a /b/-/p/ contrast, with which they were not presented during training. As for phonetic imitation, Nielsen (2011) exposed American English listeners to speech with extended VOT values in /p/ and showed that they subsequently produced extended VOTs in /p/ but also in /k/. These findings thus indicate that, throughout their language development, language users have learned to link phonetic information directly to phonological features, and that they can adjust these established sound-feature mapping when adapting to a new speaker or dialect.

6.4 CONCLUSIONS

The findings reported in this thesis indicate that adult listeners map phonetic dimensions directly onto phonological features. To acquire their feature representations, however, language learners use both the phonetic as well as the more abstract linguistic information available in their language environment. Our results further suggest that language users map phonetic information not only to features but also to phonemes, and that both features and phonemes might potentially lie at the interface with the phonetics.

Note, however, that even if both features and phonemes are connected to the phonetics, they do not necessarily have to exist within the same level of representation. For instance, one could speculate that the phoneme is at a higher level of representation than the feature and is thus connected to the phonetics only indirectly through feature representations. At the same time, phonetic dimensions that as such are not mapped onto any phonological features, might be mapped directly onto phonemes, as a result of e.g. speakers' articulatory experience.³ An alternative scenario, i.e. phonemes being at a lower level of representation than features, runs contrary to the present findings, which indicated that features that are part of one's phonological system seem to be connected to the phonetics directly. In sum, the findings of the present experiments are least compatible with a model in which the feature is represented above the phoneme. However, we cannot conclude that the reverse is

² See Norris et al. (2003) who developed the lexical adaptation task and demonstrated the effect with fricatives in Dutch.

³ We can illustrate this speculation on the findings of Chapter 4. We found that phonetic duration does not cue a phonological feature in Dutch but it still has a phonological function: it signals the identity of the phoneme /a:/. Duration may have become a phoneme-specific cue possibly because Dutch speakers have 'phonologized' the long duration that was inherent to the open articulatory realization of the low [a]-quality. Supposedly thus, the Dutch phoneme /a:/ is represented at a higher level than the features [low] and [central] (which are directly linked to the relevant phonetic dimensions of F1 and F2), but this high-level phonemic representation itself is also connected to the phonetic dimension of duration.

true, i.e. that the phoneme is represented above the feature, since features and phonemes may both exist within the same level of representation. Whether there is any hierarchy at all between speakers' feature and phoneme representations remains a question open for future research.

BIBLIOGRAPHY

- Adank, P., van Hout, R., and Smits, R. (2004). An acoustic description of the vowels of Northern and Southern standard Dutch. *Journal of the Acoustical Society of America*, 116:1729–1738.
- Adank, P., van Hout, R., and van de Velde, H. (2007). An acoustic description of the vowels of Northern and Southern standard Dutch II: Regional varieties. *Journal of the Acoustical Society of America*, 121:1130–1141.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- Alkire, T. and Rosen, C. (2010). *Romance languages: a historical introduction*. Cambridge: Cambridge University Press.
- Ashby, J., Sanders, L. D., and Kingston, J. (2009). Skilled readers begin processing sub-phonemic features by 80 ms during visual word recognition: Evidence from ERPs. *Biological Psychology*, 80:84–94.
- Bauer, L. (1985). Tracing phonetic change in the received pronunciation of British English. *Journal of Phonetics*, 13:61–81.
- Benders, T. (2013). *Nature's distributional learning experiment: infants' input, infants' perception, computer simulations*. PhD thesis, University of Amsterdam.
- Benders, T., Escudero, P., and Sjerps, M. (2012). The interrelation between acoustic context effects and available response categories in speech sound categorization. *Journal of the Acoustical Society of America*, 131(4):3079–3087.
- Bennett, D. C. (1968). Spectral form and duration as cues in the recognition of English and German vowels. *Language and Speech*, 11:65–85.
- Berko, J. (1958). The child's learning of English morphology. *Word*, 14:150–177.
- Bladon, A. (1983). Two-formant models of vowel perception: shortcomings and enhancements. *Speech Communication*, 2:305–313.
- Boersma, P. (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences Amsterdam*, 21:43–58.

- Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.
- Boersma, P. (2007). Some listener-oriented accounts of h-aspiré in French. *Lingua*, 117:1989–2054.
- Boersma, P. (2009). Cue constraints and their interactions in phonological perception and production. In Boersma, P. and Hamann, S., editors, *Phonology in Perception*, pages 55–110. Berlin: Mouton De Gruyter.
- Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In Benz, A. and Mattausch, J., editors, *Bidirectional Optimality Theory*, pages 33–72. Amsterdam: John Benjamins.
- Boersma, P., Benders, T., and Seinhorst, K. (2013a). Neural network models for phonology and phonetics. Manuscript in preparation.
- Boersma, P. and Chládková, K. (2011). Asymmetries between speech perception and production reveal phonological structure. In *Proceedings of 17th ICPHS, Hong Kong*, pages 328–331.
- Boersma, P. and Chládková, K. (2013a). Detecting categorical perception in continuous discrimination data. *Speech Communication*, 55:33–39.
- Boersma, P. and Chládková, K. (2013b). Emergence of vowel features in a neural network. Presentation at the CUNY conference on the feature in phonetics and phonology, New York, January 16–18, 2013.
- Boersma, P. and Chládková, K. (2013c). Neural networks learn features more easily if there are phonological alternations. Presentation at Phonology, University of Massachusetts Amherst, November 8–10, 2013, upcoming.
- Boersma, P., Chládková, K., and Benders, T. (2013b). Learning phonological structures from sound-meaning pairs. Presentation at the 21st Manchester Phonology Meeting, Manchester, May 23–25, 2013.
- Boersma, P. and Escudero, P. (2008). Learning to perceive a smaller L2 vowel inventory: an Optimality Theory account. In Avery, P., Dresher, E., and Rice, K., editors, *Contrast in Phonology: Theory, Perception, Acquisition*, pages 271–301. Berlin: Mouton De Gruyter.
- Boersma, P., Escudero, P., and Hayes-Harb, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. In *Proceedings of 15th ICPHS*, pages 1013–1016, Barcelona.

- Boersma, P. and Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology*, 25:217–270.
- Boersma, P. and Hamann, S. (2009). Loanword adaptation as first-language phonological perception. In Calabrese, A. and Wetzels, W. L., editors, *Loanword phonology*, pages 11–58. Amsterdam: John Benjamins.
- Boersma, P. and Weenink, D. (1992-2013). Praat: doing phonetics by computer (versions 5.1.3 [ch2], 5.2.26 [ch3], 5.2.40 [ch4]). [Computer program], available from <http://www.praat.org/>.
- Bogacka, A. (2004). On the perception of english high vowels by Polish learners of English. In *CamLing 2004: Proceedings of the University of Cambridge second postgraduate conference in language research*, pages 43–50. Cambridge.
- Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In Strange, W., editor, *Speech perception and linguistic experience: Issues in cross-language research*, pages 279–300. Timonium, MD: York Press.
- Booij, G. (1995). *The Phonology of Dutch*. Oxford: Oxford University Press.
- Botma, B. and van Oostendorp, M. (2012). A propos of the dutch vowel system 21 years on, 22 years on. In Botma, B. and Noske, R., editors, *Phonological Explorations: Empirical, Theoretical and Diachronic Issues*, pages 1–16. Berlin: Mouton de Gruyter.
- Broersma, M. (2005). Perception of familiar contrasts in unfamiliar positions. *Journal of the Acoustical Society of America*, 117:3890–3901.
- Broersma, M. (2010). Perception of final fricative voicing: Native and non-native listeners' use of vowel duration. *Journal of the Acoustical Society of America*, 127:1636–1644.
- Carbonell, J. F. and Llisterri, J. (1992). Catalan. *Journal of the International Phonetic Association*, 22(1–2):53–56.
- Cebrian, J. (2006). Experience and the use of duration in the categorization of l2 vowels. *Journal of Phonetics*, 34:372–387.
- Chistovich, L., Fant, G., and de Serpa Leitao, A. (1966). Mimicking and perception of synthetic vowels, part II. *STL-QPSR*, 7(3):1–3.
- Chládková, K., Benders, T., and Boersma, P. (ms). The human listener as a phonological feature detector: the perceptual basis of vowel height.

- Chládková, K. and Escudero, P. (2012). Comparing vowel perception and production in Spanish and Portuguese: European versus Latin American dialects. *Journal of the Acoustical Society of America*, 131(2):EL119–EL125.
- Chládková, K., Escudero, P., and Boersma, P. (2011). Context-specific acoustic differences between Peruvian and Iberian Spanish vowels. *Journal of the Acoustical Society of America*, 130:416–428.
- Chládková, K., Escudero, P., and Lipski, S. C. (2013). Pre-attentive sensitivity to vowel duration reveals native phonology and predicts learning of second-language sounds. *Brain and Language*, 126:243–252.
- Chládková, K. and Hamann, S. (2011). High vowels in Standard British English: /u/-fronting does not result in merger. In *Proceedings of XVII ICPhS 2011, Hong Kong*, pages 476–479.
- Chládková, K. and Podlipský, V. J. (2012). Native dialect matters: Perceptual assimilation of Dutch vowels by Czech listeners. *Journal of the Acoustical Society of America*, 130:EL186–EL192.
- Chomsky, N. and Halle, M. (1968). *Sound Pattern of English*. Cambridge, MA: MIT Press.
- Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, 2:225–252.
- Cohn, A. (2011). Features, segments, and the sources of phonological primitives. In Clements, G. N. and Ridouane, R., editors, *Where Do Phonological Features Come From? Cognitive, physical and developmental bases of distinctive speech categories*, pages 15–41. Amsterdam: John Benjamins.
- Collins, B. and Mees, I. (2008). *Practical Phonetics and Phonology*. Abingdon: Routledge, second edition.
- Cox, F. and Palethorpe, S. (2007). Australian English. *Journal of the International Phonetic Association*, 37(3):341–350.
- Crothers, J. (1978). Typology and universals of vowel systems. In Greenberg, J., editor, *Universals of Human Language: Phonology*, volume 2, pages 93–152. Stanford, CA: Stanford University Press.
- Cruz-Ferreira, M. (1995). Portuguese (European). *Journal of the International Phonetic Association*, 25(2):90–94.
- de Jong, G., McDougall, K., Hudson, T., and Nolan, F. (2007). The speaker discriminating power of sounds undergoing historical change: A formant-based study. In *Proceedings of XVI ICPhS*, pages 1813–1816, Saarbrücken.

- Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color: observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, 8:195–210.
- Delgutte, B. (1997). Auditory neural processing of speech. In Hardcastle, W. J. and Laver, J., editors, *The Handbook of Phonetic Sciences*, pages 507–538. Cambridge, MA and Oxford: Blackwell.
- Diehl, R. L. (1981). Feature detectors for speech: a critical reappraisal. *Psychological Bulletin*, 89:1–18.
- Diehl, R. L. and Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, 1(2):121–144.
- Dietrich, C., Swingle, D., and Werker, J. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Science USA*, 104:16027–16031.
- Docherty, G. and Foulkes, P. (1999). Derby and newcastle: instrumental phonetics and variationist studies. In Foulkes, P. and Docherty, G., editors, *Urban Voices*, pages 47–71. London: Arnold.
- Drager, K. (2010). Speaker age and vowel perception. *Language and Speech*, 54:99–121.
- Dresher, B. E. (2004). On the acquisition of phonological representations. In *Proceedings of the first workshop on psychocomputational models of human Language Acquisition*, pages 41–49. <http://www.colag.cs.hunter.cuny.edu/psychocomp>.
- Eimas, P. D. and Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4:99–109.
- Elsendoorn, B. A. G. (1985). Production and perception of dutch foreign vowel duration in english monosyllabic words. *Language and Speech*, 28:231–254.
- Escudero, P. (2005). *Linguistic Perception and Second Language Acquisition. Explaining the attainment of optimal phonological categorization*. PhD thesis, Utrecht University.
- Escudero, P., Benders, T., and Lipski, S. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, 37:452–465.
- Escudero, P., Benders, T., and Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second-language vowels. *Journal of the Acoustical Society of America*, 130:EL206–EL212.

- Escudero, P. and Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, 26:551–585.
- Escudero, P. and Chládková, K. (2010). Spanish listeners' perception of American and Southern British English vowels. *Journal of the Acoustical Society of America*, 128:EL254–EL260.
- Escudero, P., Simon, E., and Mitterer, H. (2012). The perception of English front vowels by North Holland and Flemish listeners: Acoustic similarity predicts and explains cross-linguistic and L2 perception. *Journal of Phonetics*, 40:280–288.
- Escudero, P. and Vasiliev, P. (2011). Cross-language acoustic similarity predicts perceptual assimilation of Canadian English and Canadian French vowels. *Journal of the Acoustical Society of America*, 130:EL277–EL283.
- Escudero, P. and Williams, D. (2011). Spanish listeners' perception of dutch vowels. *Journal of the Acoustical Society of America*, 129:EL1–EL7.
- Escudero, P. and Williams, D. (2012). Native dialect influences second-language vowel perception: Peruvian versus Iberian Spanish learners of Dutch. *Journal of the Acoustical Society of America*, 131:EL406–EL412.
- Evans, B. G. and Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *Journal of the Acoustical Society of America*, 115(1):352–361.
- Fikkert, P. and Freitas, M. J. (2006). Allophony and allomorphy cue phonological acquisition: evidence from the European Portuguese vowel system. *Catalan Journal of Linguistics*, 5:83–108.
- Fikkert, P. and Levelt, C. (2008). How does Place fall into Place? The lexicon and emergent constraints in children's developing phonological grammar. In Avery, P., Dresher, E., and Rice, K., editors, *Contrast in Phonology: Theory, Perception, Acquisition*, pages 231–270. Berlin: Mouton De Gruyter.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. A*, 222:309–368.
- Flege, J. E., Bohn, O.-S., and Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25:437–470.
- Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14:3–28.

- Gimson, A. C. (2001). *Gimson's pronunciation of English*. London: Arnold, sixth edition.
- Goldsmith, J. (1976). *Autosegmental phonology*. PhD thesis, Cambridge: MIT.
- Guenther, F. H. and Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, 100:1111–1121.
- Gulian, M., Escudero, P., and Boersma, P. (2007). Supervision hampers distributional learning of vowel contrasts. In *Proceedings of the International Congress of Phonetic Sciences*, pages 1893–1896, Saarbrücken.
- Gussenhoven, C. (1992). Dutch. *Journal of the International Phonetic Association*, 22(1–2):45–47.
- Hála, B. (1941). *Akustická podstata samohlásek*. [The acoustic basis of vowels]. Praha: Czech Academy of Sciences and Arts.
- Hála, B. (1960). *Fonetické obrazy hlásek*. [Phonetic projections of speech sounds]. Praha: SPN.
- Hale, M., Kissock, M., and Reiss, C. (2006). Microvariation, variation, and the features of universal grammar. *Lingua*, 32:402–420.
- Halle, M. (1970). Is Kabardian a vowelless language? *Foundations of Language*, 6(1):95–103.
- Hamann, S. (2003). *The phonetics and phonology of retroflexes*. PhD thesis, Utrecht University.
- Hamann, S. (2011). The Phonetics-Phonology Interface. In Kula, N., Botma, B., and Nasukawa, K., editors, *Continuum Companion to Phonology*, pages 202–224. London: Continuum.
- Hanulíková, A. and Hamann, S. (2010). Slovak. *Journal of the International Phonetic Association*, 40(3):373–378.
- Harrington, J. (2007). Evidence for a relationship between synchronic variability and diachronic change in the Queen's annual Christmas broadcasts. In Cole, J. and Hualde, J., editors, *Laboratory Phonology 9*, pages 125–143. Mouton: Berlin.
- Harrington, J., Hoole, P., Kleber, F., and Reubold, U. (2011a). The physiological, acoustic, and perceptual basis of high back vowel fronting: Evidence from German tense and lax vowels. *Journal of Phonetics*, 39:121–131.

- Harrington, J., Kleber, F., and Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *Journal of the Acoustical Society of America*, 123:2825–2835.
- Harrington, J., Kleber, F., and Reubold, U. (2011b). The contributions of the lips and the tongue to the diachronic fronting of high back vowels in Standard Southern British English. *Journal of the International Phonetic Association*, 41:137–156.
- Harris, J. (1990). Segmental complexity and phonological government. *Phonology*, 7:255–300.
- Hawkins, S. and Midgley, J. (2005). Formant frequencies in RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association*, 35:183–195.
- Hayes, B. (2004). Phonological acquisition in Optimality Theory: the early stages. In Kager, R., Pater, J., and Zonneveld, W., editors, *Constraints in phonological acquisition*, pages 158–203. Cambridge, MA: CUP.
- Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, 23:1–31.
- Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley.
- Henton, C. G. (1983). Changes in the vowels of Received Pronunciation. *Journal of Phonetics*, 123:353–371.
- Hisagi, M., Shafer, V. L., Strange, W., and Sussman, E. S. (2010). Perception of a Japanese vowel length contrast by Japanese and American English listeners. *Brain Research*, 1360:89–105.
- Jakobson, R., Fant, G., and Halle, M. (1952). *Preliminaries to Speech Analysis: the Distinctive Features and their Correlates*. Cambridge, MA: MIT Press.
- Jespersen, O. (1909). *A Modern English Grammar on Historical Principles*, volume 1. London: Allen and Unwin.
- Kaye, J. D., Lowenstamm, J., and Vergnaud, J.-R. (1985). The internal structure of phonological representations: a theory of charm and government. *Phonology Yearbook*, 2:305–328.
- Kazanina, N., Phillips, C., and Idsardi, W. (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences USA*, 103:11381–11386.

- Kingston, J. (1991). Integrating articulations in the perception of vowel height. *Phonetica*, 48:149–179.
- Kingston, J. (2003). Learning foreign vowels. *Language and Speech*, 46:295–349.
- Kingston, J. and Diehl, R. L. (1994). Phonetic knowledge. *Language*, 70(3):419–454.
- Kingston, J. and Diehl, R. L. (1995). Intermediate properties in the perception of distinctive feature values. In Connell, A. and Arvaniti, A., editors, *Papers in Laboratory phonology 4: Phonology and phonetic evidence*, pages 7–27.
- Kingston, J., Diehl, R. L., Kirk, C. J., and Castleman, W. A. (2008). On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics*, 36:28–54.
- Kirmse, U., Ylinen, S., Tervaniemi, M., Vainio, M., Schröger, E., and Jacobsen, T. (2008). Modulation of the mismatch negativity (MMN) to vowel duration changes in native speakers of Finnish and German as a result of language experience. *International Journal of Psychophysiology*, 67:131–143.
- Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America*, 87(2):820–856.
- Kohler, K. J. (1981). Contrastive phonology and the acquisition of phonetic skills. *Phonetica*, 38:213–226.
- Kondaurova, M. V. and Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *Journal of the Acoustical Society of America*, 124:3959–3971.
- Kraljic, T. and Samuel, A. G. (2006). Generalization in perceptual learning of speech. *Psychonomic Bulletin & Review*, 13(2):262–268.
- Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *Journal of the Acoustical Society of America*, 70(2):340–349.
- Kuhl, P. K. (1991). Human adults and human infants show a ‘perceptual magnet effect’ for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, 50:93–107.
- Kučera, H. (1961). *The Phonology of Czech*. s’ Gravenhage: Mouton & Co.

- Ladefoged, P. (1980). What are linguistic sounds made of? *Language*, 56(3):485–502.
- Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
- Levelt, C. and van Oostendorp, M. (2007). Feature co-occurrence constraints in L1 acquisition. In Los, B. and van Koppen, M., editors, *Linguistics in the Netherlands*, pages 162–172. Amsterdam: John Benjamins.
- Lieberman, A. and Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, 21:1–36.
- Lieberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, 54(5):358–368.
- Lin, Y. and Mielke, J. (2008). Discovering place and manner features: what can be learned from acoustic and articulatory data. *University of Pennsylvania Working Papers in Linguistics*, 14:241–254.
- Lindau, M. and Ladefoged, P. (1986). Variability of speech processes. In Perkell, J. and Klatt, D., editors, *Invariance and variability of speech processes*, pages 464–478. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lindblom, B. E. F. and Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, 42(4):830–843.
- Lipski, S., Escudero, P., and Benders, T. (2012). Language experience modulates weighting of acoustic cues for vowel perception: an event-related potential study. *Psychophysiology*, 49:638–650.
- Lipski, S. C., Lahiri, A., and Eulitz, C. (2007). Differential height specification in front vowels for German speakers and Turkish-German bilinguals: an electroencephalographic study. In *Proceedings of the International Congress of Phonetic Sciences*, pages 809–812, Saarbrücken.
- Lipski, S. C. and Mathiak, K. (2007). A magnetoencephalographic study on auditory processing of native and nonnative fricative contrasts in Polish and German. *Neuroscience Letters*, 415:90–95.
- Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word*, 20(3):384–422.
- Maddieson, I. (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.

- Mann, V. A. and Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, 28(3):213–228.
- Martínez-Celdrán, E., Fernández-Planas, A. M., and Carrera-Sabaté, J. (2003). Castilian spanish. *Journal of the International Phonetic Association*, 33(2):255–259.
- Maye, J. and Gerken, L. A. (2001). Learning phonemes: how far can input take us? In *BUCLD 25 Proceedings*, pages 480–490, Somerville, MA.
- Maye, J., Weiss, D., and Aslin, R. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, 11:122–134.
- Maye, J., Werker, J. F., and Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82:B101–B111.
- McGee, T. J., King, C., Tremblay, K., Nicol, T. G., Cunningham, J., and Kraus, N. (2001). Long-term habituation of the speech-elicited mismatch negativity. *Psychophysiology*, 38:653–658.
- Meister, E., Werner, S., and Meister, L. (2011). Short vs. long category perception affected by vowel quality. In *Proceedings of 17th ICPhS*, pages 1362–1365, Hong Kong.
- Menn, L. and Vihman, M. (2011). Features in child phonology: Inherent, emergent, or artefacts of analysis? In Clements, G. N. and Ridouane, R., editors, *Where Do Phonological Features Come From? Cognitive, physical and developmental bases of distinctive speech categories*, pages 261–301. Amsterdam: John Benjamins.
- Menning, H., Imaizumi, S., Zwitserlood, P., and Pantev, C. (2002). Plasticity of the human auditory cortex induced by discrimination learning of non-native, mora-timed contrasts of the Japanese language. *Learning and Memory*, 9:253–267.
- Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound formants. *Journal of the Acoustical Society of America*, 63(2):572–580.
- Mielke, J. (2008). *The emergence of distinctive features*. Oxford: OUP.
- Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *Journal of the Acoustical Society of America*, 27(2):338–352.

- Morén, B. (2003). The parallel structures model of feature geometry. In *Working Papers of the Cornell Phonetics Laboratory*, volume 15, pages 194–270.
- Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467.
- Moulton, W. (1962). The vowels of Dutch: Phonetic and distributional classes. *Lingua*, 11:294–312.
- Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R., Luuk, A., Allik, J., Sinkkonen, J., and Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385:432–434.
- Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology*, 118:2544–2590.
- Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., and Winkler, I. (2001). Primitive intelligence in auditory cortex. *Trends in Neuroscience*, 25:283–288.
- Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, 18:347–373.
- Nearey, T. M. (1995). A double-weak view of trading relations: comments on Kingston and Diehl. In Connell, A. and Arvaniti, A., editors, *Papers in Laboratory phonology 4: Phonology and phonetic evidence*, pages 28–40.
- Nenonen, S., Shestakova, A., Huotilainen, M., and Näätänen, R. (2003). Linguistic relevance of duration within the native language determines the accuracy of speech-duration processing. *Cognitive Brain Research*, 16:492–495.
- Nenonen, S., Shestakova, A., Huotilainen, M., and Näätänen, R. (2005). Speech-sound duration processing in a second language is specific to phonetic categories. *Brain and Language*, 92:26–32.
- Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 39:132–142.
- Nooteboom, S. G. and Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America*, 67:276–287.

- Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47:204–238.
- Obleser, J., Lahiri, A., and Eulitz, C. (2004). Magnetic brain response mirrors extraction of phonological features from spoken vowels. *Journal of Cognitive Neuroscience*, 16:31–39.
- Ohala, J. J. (1981). The listener as a source of sound change. In Masek, C. and Hendrik, R. A. M. M. F., editors, *Proceedings of the Chicago Linguistic Society 17*, pages 178–203, Chicago.
- Ohala, J. J. and Feder, D. (1994). Listeners' normalization of vowel quality is influenced by 'restored' consonantal context. *Phonetica*, 51:111–118.
- Pisoni, D. B. (1973). Auditory short-term memory and vowel perception. *Memory Cognition*, 3(1):7–18.
- Pisoni, D. B. and Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, 25:21–52.
- Pisoni, D. B. and Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, 15(2):285–290.
- Pitt, M., Myung, I., and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109(3):472–491.
- Polivanov, E. D. (1931). La perception des sons d'une langue étrangère [The perception of the sounds of a foreign language]. *Travaux du Cercle Linguistique de Prague*, 4:79–96.
- Pulleyblank, D. (2006). Minimizing UG: Constraints upon Constraints. In Baumer, D., Montero, D., and Scanlon, M., editors, *Proceedings of the 25th West Coast Conference on Formal Linguistics*, pages 15–39.
- Rauber, A. S., Escudero, P., Bion, R., and Baptista, B. O. (2005). The interrelation between the perception and production of English vowels by native speakers of Brazilian Portuguese. In *Proceedings of Interspeech*, pages 2913–2916, Lisbon.
- Repp, B. H., Healy, A. F., and Crowder, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 5(1):129–145.
- Roach, P. (2009). *English Phonetics and Phonology*. Cambridge: Cambridge University Press, fourth edition.

- Rogers, J. C. and Davis, M. H. (2009). Categorical perception of speech without stimulus repetition. In *Proceedings of Interspeech 2009*, pages 376–379.
- Rumelhart, D. and Zipser, D. E. (1985). Feature discovery by competitive learning. *Cognitive Science*, 9:75–112.
- Savela, J. (2009). *Role of selected spectral attributes in the perception of synthetic vowels*. PhD thesis, University of Turku.
- Sawusch, J. R. and Nusbaum, H. C. (1979). Contextual effects in vowel perception i: Anchor- induced contrast effects. *Perception and Psychophysics*, 25(4):292–302.
- Scharinger, M., Idsardi, W. J., and Poe, S. (2011a). A comprehensive three-dimensional cortical map of vowel space. *Journal of Cognitive Neuroscience*, 23(12):3972–3982.
- Scharinger, M., Merickel, J., Riley, J., and Idsardi, W. J. (2011b). Neuromagnetic evidence for a featural distinction of English consonants: Sensor- and source-space data. *Brain and Language*, 116:71–82.
- Scharinger, M., Monahan, P. J., and Idsardi, W. J. (2011c). You had me at “Hello”: Rapid extraction of dialect information from spoken words. *NeuroImage*, 56:2329–2338.
- Scharinger, M., Monahan, P. J., and Idsardi, W. J. (2012). Assymetries in the processing of vowel height. *Journal of Speech, Language and Hearing Research*, 55:903–918.
- Schouten, M. E. H. and van Hessen, A. J. (1992). Modelling phoneme perception. I: Categorical perception. *Journal of the Acoustical Society of America*, 92(4):1841–1855.
- Sebastián-Gallés, N. (2005). Cross-language speech perception.
- Sharma, A. and Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *Journal of the Acoustical Society of America*, 107:2697–2703.
- Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17:3–46.
- Stevens, K. N. and Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In Eimas, P. D. and Miller, J. L., editors, *Perspectives on the study of speech*, pages 1–38. Hillsdale, NJ: Lawrence Erlbaum Associates.

- Stockwell, R. (2002). How much shifting actually occurred in the historical English vowel shift? In Minkova, D. and Stockwell, R., editors, *Studies in the History of the English Language: A Millennial Perspective*, pages 267–281. Berlin: Mouton de Gruyter.
- Stoddart, J., Upton, C., and Widdowson, J. (1999). Sheffield dialect in the 1990s: revisiting the concept of NORMs. In Foulkes, P. and Docherty, G., editors, *Urban Voices*, pages 72–89. London: Arnold.
- Studdert-Kennedy, M. and Shankweiler, D. (1970). Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America*, 48(2B):579–594.
- Syrdal, A. K. and Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, 79(4):1086–1100.
- Tervaniemi, M., Jacobsen, T., Röttger, S., Kujala, T., Widmann, A., Vainio, M., Näätänen, R., and Schröger, E. (2006). Selective tuning of cortical sound-feature processing by language experience. *European Journal of Neuroscience*, 23:2538–2541.
- Thelwall, R. and Akram Sa'Adeddin, M. (1990). Arabic. *Journal of the International Phonetic Association*, 20(2):37–41.
- Trubetzkoy, N. (1939). *Grundzüge der Phonologie [Principles of Phonology]*. Los Angeles: University of California Press. Translated by C. Baltaxe, 1969.
- Trudgill, P. (1999). Norwich: endogenous and exogenous linguistic change. In Foulkes, P. and Docherty, G., editors, *Urban Voices*, pages 124–140. London: Arnold.
- Uffmann, C. (2010). The non-trivialness of segmental representations. Talk presented at the Old World Conference in Phonology 7, Nice, France. [abstract available at: <http://www.unice.fr/dsl/ocp7/abstracts/uffmann.pdf>, accessed May 7, 2013].
- van der Feest, S. and Swingle, D. (2011). Dutch and English listeners' interpretation of vowel duration. *Journal of the Acoustical Society of America*, 129:EL57–EL63.
- van Heuven, V. J. E., van Houten, J. E., and de Vries, J. W. (1986). De perceptie van Nederlandske klinkers door Turken [The perception of Dutch vowels by Turkish listeners]. *Spectator*, 15-4:225–238.

- van Leussen, J. W., Escudero, P., and Williams, D. (2011). Acoustic properties of Dutch steady-state vowels: Contextual effects and a comparison with previous studies. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 1194–1197, Hong Kong.
- van Oostendorp, M. (1995). *Vowel quality and phonological projection*. PhD thesis, Tilburg University.
- Šimáčková, Š., Podlipský, V. J., and Chládková, K. (2012). Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association*, 42(2):225–232.
- Wells, J. (1962). A study of the formants of the pure vowels of british english. Master's thesis, University of London. available at: <http://www.phon.ucl.ac.uk/home/wells/formants/index.htm>, accessed May 28, 2013.
- Williams, D. P. (2013). *Cross-language acoustic and perceptual similarity of vowels: The role of listeners' native accents*. PhD thesis, University of Sheffield.
- Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustical Society of America*, 60(6):1381–1389.
- Ylinen, S., Shestakova, A., Huotilainen, M., Alku, P., and Näätänen, R. (2006). Mismatch negativity (MMN) elicited by changes in phoneme length: A cross-linguistic study. *Brain Research*, 1072:175–185.
- Zimmerman, S. A. and Sapon, S. M. (1958). Note on vowel duration seen cross-linguistically. *Journal of the Acoustical Society of America*, 30:152–153.
- Zonneveld, W. (1993). Schwa, superheavies, stress and syllables in Dutch. *The Linguistic Review*, 10:61–110.

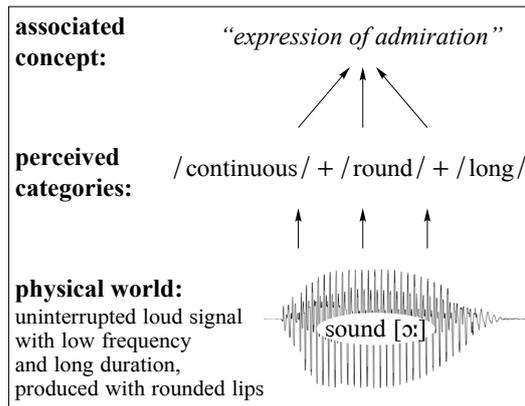
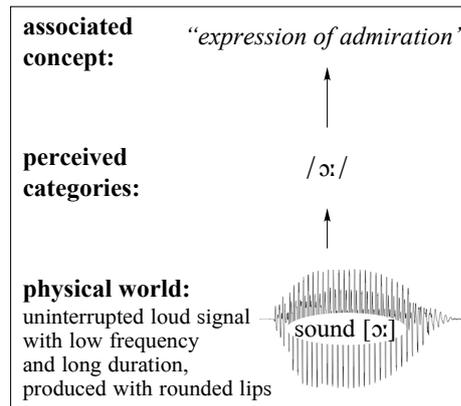
SUMMARY

Most objects in the world around us are associated with abstract concepts. A real-world object is thus a physical realization of an abstract category. For instance, the object pictured on the back cover of this book is associated with the concept “Rubik’s cube”. Also, every physical object has a number of individual properties some of which are contrastive, which means that they differentiate the given object from other objects associated with other concepts. For instance, some of the properties of the object  are: shape of a cube, six different colours, rotating parts.

Speech sounds are like objects. Firstly, speech sounds, or strings of speech sounds, are associated with abstract concepts. A speech sound is thus a physical realization of an abstract category. For instance, the speech sound that is produced by the speaker pictured on the back cover of this book is a single vowel ‘o’ (which phoneticians transcribe as [ɔ:]), and in some languages, it is associated with the concept of “expression of admiration”. Similarly, the vowel [ɔ:] is also found in the string of sounds [dɔ:g], which is, in American English, associated with the concept “a companion animal that barks”. Secondly, every speech sound has a number of individual properties some of which are contrastive, which means that they differentiate the given sound from other sounds associated with other concepts. For instance, the sound [ɔ:] consists of an uninterrupted loud acoustic signal, it is relatively long, it is produced with rounded lips, and it has a considerable amount of energy in low frequencies. The properties of speech sounds that are contrastive are called *phonological features*.

Linguists know what properties are contained in every speech sound because people can be recorded as they speak and their speech sounds can be acoustically analyzed with a computer. However, linguists do not entirely know yet how people actually *listen* to speech sounds. Specifically, it has not yet been shown whether during speech comprehension, listeners recognize each of the contrastive properties of the sound (i.e. the phonological features) individually, or whether listeners immediately recognize the whole sound segment without the need to recognize each of its contrastive features on its own. In this thesis, we aimed to resolve that puzzle and reveal whether phonological features are found in perception. The Figure on the following page illustrates what speech comprehension may look like with and without phonological feature categories.

In order to investigate whether listeners perceive speech sounds through phonological feature categories, and whether they also learn speech sounds

Speech comprehension **with** phonological features... and **without** phonological features

as sets of phonological features, we carried out a number of experiments with human and virtual listeners. Importantly, note that when you play the sound [ɔ:] and ask a listener to tell you what she hears, she will most certainly report hearing an /ɔ:/, that is, she will name the whole speech segment. This is because speech segments have explicit labels, or names, which are known to all speakers of a given language (think of the alphabet, which contains names for most sounds of a given language). On the other hand, phonological features do not have any labels or names that would be known to an ordinary language user. For that reason, it is impossible to ask a listener whether she perceives [ɔ:] in terms of its individual features (i.e. as a category /continuous/ plus a category /round/ plus a category /long/) or whether she perceives it as an unanalyzed segment (i.e. as a category /ɔ:/).

Therefore, in our experiments, instead of straightforwardly asking participants whether they hear phonological features, we used a variety of behavioral, electrophysiological and computational methods that allowed us to uncover how humans process speech. For instance, we tested the perception of sound segments that have no abstract associations and no labels in a given language, but that contain some of the features that are present in the listeners’ native-language sounds. We reasoned that if humans are able to perceptually categorize the unknown sounds in the same way as they categorize their native-language sounds, we have evidence for feature categories in perception.

The results of our experiments indicate that adult listeners indeed perceive speech sounds in terms of phonological feature categories. For instance, a sound like [ɔ:] is perceived in terms of individual feature categories /continuous/ + /round/ + /long/ (as shown in the left panel of the Figure above). An illustration of this result is also provided on the cover of this thesis, where the physical object 🧊 is perceived as a set of contrastive features /cube/ + /colours/ + /rotation/.

Our findings further suggest that during speech comprehension, listeners recognize those feature categories that are used contrastively in their own language. For instance, if – in an imaginary language – the sound [ɔ:] were always produced with the tip of the tongue stuck out of the mouth, the native speaker of this language would probably perceive [ɔ:] as /continuous/ + /round/ + /long/ + /tongue out/. Let's illustrate this finding using our favorite object. If – in an imaginary world – every single 🎲 were not only rotatable but also squeezable, it would probably be perceived as /cube/ + /colours/ + /rotation/ + /squeeze/.

Finally, the results of our computational simulations show that when virtual infants acquire their native language, they initially learn to represent speech sounds as whole segments, but after enough experience with their language, they also create feature categories for the sounds. For instance, a very young baby may first perceive the sound [ɔ:] as an unanalyzed segment /ɔ:/. Subsequently, as the baby encounters many different instances of [ɔ:] and of all the other sounds and words of her language, she comes to figure out that [ɔ:] has the phonological features /continuous/ + /round/ + /long/ that differentiate it from a sound like [e:] which is not produced with rounded lips, or from a sound like [p] which is neither continuous nor long. We again exemplify this result using the Rubik's cube. At first, a baby may just perceive 🎲 as one unanalyzed whole, but as she gains experience with the world, she comes to realize that this object has the contrastive features /cube/ + /colours/ + /rotation/ that differentiate it from an object like 🎲 which is not rotatable, or from an object like 🍷 which is neither cube-shaped nor rotatable.

To sum up, the research reported in this thesis aimed to uncover the role of phonological features in speech comprehension. Our results indicate that phonological features are likely to be the categories through which listeners perceive speech. Furthermore, our findings suggest that language users learn to recognize those phonological features that are relevant in their own language environment

Note that all references to object perception are only meant to illustrate the present findings about *speech* perception using a domain that is familiar to the general audience. No claims are made here about object perception in general.

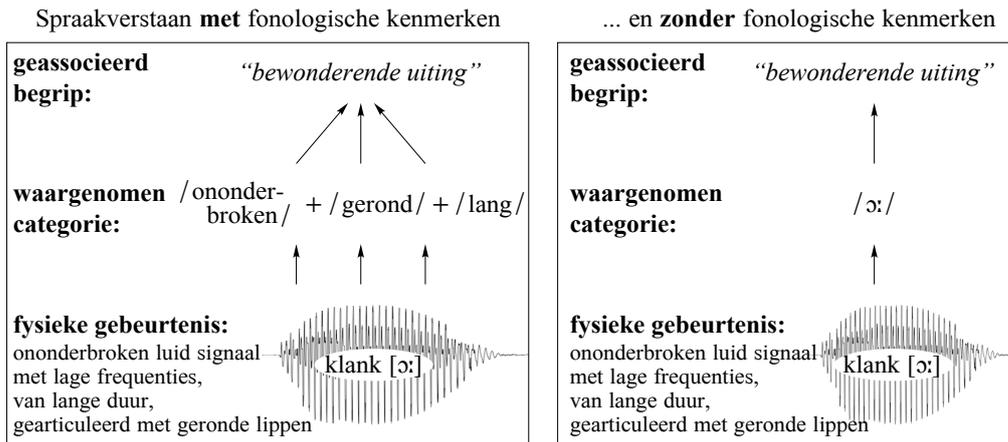
SAMENVATTING

De meeste objecten in de wereld om ons heen worden geassocieerd met abstracte begrippen. Het object is dan de fysieke realisatie van een abstracte categorie. Zo wordt het object op de omslag van dit boek geassocieerd met het begrip “Rubikskubus”. Naast een associatie met een abstract begrip heeft een fysiek object eigenschappen waarmee het object van andere objecten - die met andere begrippen geassocieerd zijn - kan worden onderscheiden. Enkele eigenschappen van het object  zijn: kubusvormig, zeskleurig, beweegbaar.

Spraakklanken zijn ook een soort objecten. Allereerst worden (reeksen van) spraakklanken geassocieerd met abstracte begrippen. De spraakklank is dan de fysieke realisatie van een abstracte categorie. De spreker op de achterflap van dit boek spreekt bijvoorbeeld de klinker ‘o’ uit (in fonetische notatie [ɔ:]), die in sommige talen wordt geassocieerd met het concept “bewonderende uiting”. De klinker [ɔ:] komt ook voor in de klankreeks [rɔ:zə] die in het Nederlands wordt geassocieerd met het begrip “licht rood- of magenta-achtige kleur”. Daarnaast hebben ook spraakklanken eigenschappen op basis waarvan ze van andere spraakklanken – die met andere begrippen geassocieerd worden – kunnen worden onderscheiden. Het geluid [ɔ:] bestaat uit een ononderbroken akoestisch signaal van een relatief lange duur dat wordt geproduceerd met geronde lippen, waarin de energie is geconcentreerd bij lage frequenties. De onderscheidende eigenschappen van spraakklanken worden *fonologische kenmerken* genoemd.

Taalwetenschappers weten van spraakklanken vrij precies welke eigenschappen ze hebben, omdat spraakgeluid kan worden opgenomen en met de computer akoestisch kan worden geanalyseerd. Wat taalwetenschappers echter niet weten is hoe mensen precies naar spraakklanken *luisteren*. Een open vraag is of luisteraars de onderscheidende kenmerken van klanken allemaal apart herkennen, of dat ze een klank in zijn geheel in één keer herkennen. De vraag die we in dit proefschrift proberen te beantwoorden is dus of fonologische kenmerken een rol spelen bij het waarnemen van spraak. De afbeelding op de volgende pagina illustreert hoe spraakbegrip tot stand zou kunnen komen als de categorieën al dan niet in termen van fonologische kenmerken worden gedefinieerd.

Of luisteraars spraakgeluid inderdaad waarnemen in termen van fonologische kenmerken, en of ze klankcategorieën ook leren als verzamelingen van fonologische kenmerken, is onderzocht in experimenten met gewone mensen en virtuele luisteraars. Een hindernis hierbij is de neiging van mensen om hen bekende klanken met een conventionele naam



of label aan te duiden, zoals bijvoorbeeld de namen van letters in het alfabet. Dit zorgt ervoor dat het antwoord op de expliciete vraag “Wat heb je gehoord?” niet in termen van fonologische kenmerken zal worden gegeven, ongeacht de mogelijke rol ervan in het luisterproces, omdat voor deze kenmerken geen conventionele namen bestaan waar een doorsnee luisteraar mee bekend is. Het is dus niet zinvol om te vragen of een luisteraar [ɔ:] waarneemt als verzameling van individuele kenmerken (dus als /ononderbroken/, /gerond/, en /lang/) of als ongeanalyseerd geheel (dus als de categorie /ɔ:/).

Om dit probleem te omzeilen is gebruik gemaakt van verschillende gedragsmaten, electrofysiologische methoden, en computationele methoden die het mogelijk maken onbewuste processen van de menselijke spraakwaarneming te onderzoeken. Zo is bijvoorbeeld onderzocht hoe klanken worden waargenomen wanneer ze voor een luisteraar geen link met een abstract begrip hebben en geen conventioneel label, maar wel kenmerken hebben die in de moedertaal van de luisteraar voorkomen. Als de toewijzing aan klankcategorieën in dit geval toch hetzelfde verloopt als bij klanken uit de moedertaal is dat een aanwijzing dat de waarneming van spraakklanken door fonologische kenmerken wordt gemedieerd.

De experimentele resultaten wijzen erop dat spraakwaarneming bij volwassen luisteraars inderdaad tot stand komt op basis van fonologische kenmerken. De klank [ɔ:] wordt dus waargenomen als verzameling van de individuele kenmerken /ononderbroken/ + /gerond/ + /lang/ (zie bovenstaande afbeelding, links). Dit resultaat wordt ook geïllustreerd op de kaft van deze dissertatie, waar het fysieke object  wordt waargenomen als verzameling van de onderscheidende kenmerken /kubusvormig/ + /zeskleurig/ + /beweegbaar/.

De huidige bevindingen duiden er verder op dat luisteraars om spraak te verstaan precies die kenmerkcategoryen aanwenden die in hun moedertaal gebruikt worden om onderscheid te maken. Als er dus een – be-

dachte – taal zou zijn waarin de klank [ɔ:] zou worden gearticuleerd met uitgestoken tongpunt, en dit zou niet gelden voor een andere klank die met een ander begrip was geassocieerd, dan zou een moedertaalspreker van deze taal een [ɔ:] waarschijnlijk waarnemen als /ononderbroken/ + /gerond/ + /lang/ + /uitgestoken tong/. In het geval van ons favoriete object  stellen we ons een wereld voor waarin iedere  niet alleen beweegbaar maar ook indrukbaar is, terwijl er in deze wereld ook objecten bestaan die niet indrukbaar zijn. Op basis van de bevindingen van ons onderzoek voorspellen we dat dit object dan waarschijnlijk zou worden waargenomen als /kubusvormig/ + /zeskleurig/ + /beweegbaar/ + /indrukbaar/.

Tenslotte laten simulaties met computermodellen zien dat virtuele baby's die hun moedertaal leren spraakklanken in eerste instantie representeren als gehele klanksegmenten, maar naarmate ze meer ervaring krijgen met hun moedertaal gaan ze kenmerkcategoryën opbouwen. Voor echte mensen betekent dit bijvoorbeeld dat een jonge baby de klank [ɔ:] in eerste instantie waarneemt als het ongeanalyseerde geheel /ɔ:/. Vervolgens komt de baby regelmatig opnieuw realisaties van de klank [ɔ:] en van andere klanken en woorden uit haar moedertaal tegen, en zal dan ontdekken dat [ɔ:] de fonologische kenmerken /ononderbroken/ + /gerond/ + /lang/ heeft, in tegenstelling tot de klank [e:] die niet gerond is (maar wel lang), en de klank [p] die niet ononderbroken of lang is. In een vergelijkbare situatie met de Rubikskubus zou een jonge baby  waarnemen als een ongeanalyseerd geheel, maar naarmate haar ervaring met de wereld vordert zou ze zich realiseren dat dit object de onderscheidende kenmerken /kubusvormig/ + /zeskleurig/ + /beweegbaar/ heeft, in tegenstelling tot een object als  dat niet beweegbaar is (maar wel kubusvormig), en  dat niet kubusvormig of beweegbaar is.

Samenvattend was het doel van het hier gerapporteerde onderzoek om te ontdekken welke rol fonologische kenmerken spelen bij het verstaan van spraak. De resultaten laten zien dat fonologische kenmerken de eenheden zijn waarin spraakgeluid door luisteraars wordt waargenomen. Bovendien suggereren de bevindingen dat taalgebruikers precies die fonologische kenmerken leren herkennen die relevant zijn in hun eigen taalomgeving.

Verwijzingen naar fysieke objecten vormen een illustratie van de bevindingen aangaande *spraak*waarneming met voorbeelden die voor een breed publiek herkenbaar zijn. Er worden geen claims gedaan over waarneming van fysieke objecten in het algemeen.

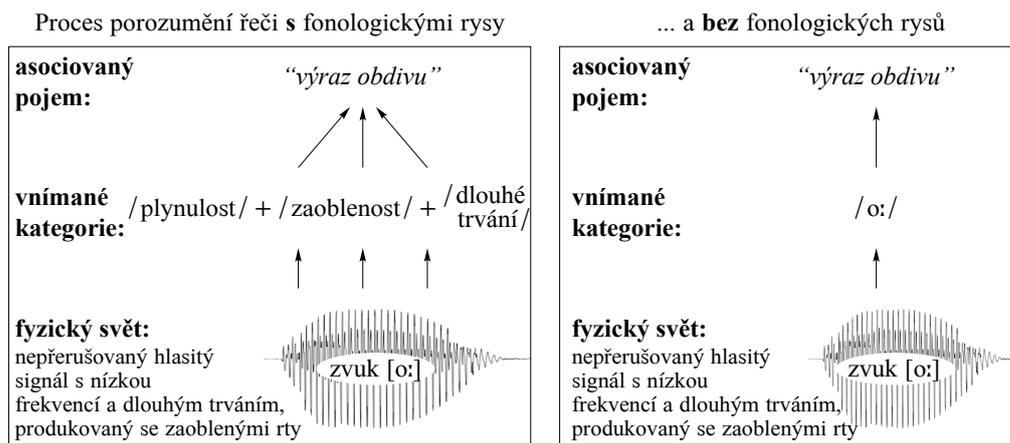
SHRNUTÍ

Většina objektů ve světě okolo nás je spojena s abstraktními pojmy. Konkrétní objekt je tedy fyzickou realizací abstraktní kategorie. Například, objekt vyobrazený na zadní straně přebalu této dizertace je asociován s pojmem “Rubikova kostka”. Dále, každý fyzický objekt má řadu vlastností, z nichž některé jsou kontrastivní, což znamená, že daný objekt odlišují od ostatních objektů spojených s jinými pojmy. Mezi vlastnosti objektu  patří: tvar krychle, šest různých barev, rotující části.

Hlásky jsou jako objekty. Hlávky, nebo seskupení hlásek, jsou spojeny s abstraktními pojmy. Hlávka je tedy fyzickou realizací abstraktní kategorie. Například, zvuk, který vyslovuje mluvčí na zadní straně přebalu této dizertace, je samohlávka ‘o’ (fonetici ji píšou jako [o:]) a tato je v některých jazycích asociována s pojmem “výraz obdivu”. Podobně, samohlávku [o:] najdeme také v seskupení hlásek [fo:r], které je v češtině asociováno s pojmem “krátké vyprávění, jehož účelem je pobavit”. Každá hlávka má řadu vlastností, z nichž některé jsou kontrastivní, což znamená, že odlišují danou hlávku od ostatních hlásek spojených s jinými pojmy. Hlávka [o:] je charakterizována nepřerušovaným, hlasitým akustickým signálem, je relativně dlouhá, produkována se zaoblenými rty, a má značné množství energie v nízkých frekvencích. Vlastnosti hlásek, které jsou kontrastivní, se nazývají *fonologické rysy*.

Je dobře známo, jaké vlastnosti každá hlávka má, protože mluvčí můžeme nahrát a jejich hlávky potom na počítači akusticky zanalyzovat. Nicméně, doposud není zcela jasné, jak jsou hlávky vlastně *vnímány*. Ještě nebylo prokázáno, jestli lidé při vnímání řeči rozeznávají každý kontrastivní rys hlásek jednotlivě, nebo jestli okamžitě rozeznají hlávku jako jeden celistvý segment bez toho, aniž by museli poznat každý z jejích kontrastivních rysů zvlášť. V této dizertační práci, nazvané *Pátrání po fonologických rysech v percepci*, jsme se tuto záhadu pokusili vyřešit – zkoumali jsme, jestli fonologické rysy existují jako abstraktní kategorie během procesu vnímání řeči. Obrázek na následující straně znázorňuje, jak by mohlo vnímání řeči vypadat s fonologickými rysy a bez nich.

Abychom zjistili, jestli lidé vnímají hlávky prostřednictvím fonologických rysů, a jestli si hlávky jako skupiny fonologických rysů i osvojují, provedli jsme řadu experimentů s reálnými a s virtuálními posluchači. Zde je důležité si uvědomit, že když přehrajeme zvuk [o:] a zeptáme se posluchače, co slyšel, téměř jistě odpoví, že slyšel /o:/, tedy, pojmenuje danou hlávku jako jeden celistvý segment. Děje se tak proto, že hlávky jako segmenty jsou explicitně označeny, čili mají jména, která znají všichni mluvčí daného jazyka (uvědomme si, že abeceda obsahuje



jména pro téměř všechny hlásky daného jazyka). Na druhou stranu, fonologické rysy nemají žádná označení či jména, která by znal každý běžný uživatel jazyka. Z tohoto důvodu je tedy nemožné, zeptat se posluchače, jestli vnímá hlásku [o:] prostřednictvím jejích fonologických rysů (t.j. jako kategorii /plynulost/ plus kategorii /zaoblenost/ plus kategorii /dlouhé trvání/), nebo jestli tuto hlásku vnímá jako jeden celistvý segment (t.j. jako kategorii /o:/).

V našich experimentech jsme se tedy účastníků přímočaře neptali, jestli slyší fonologické rysy. Namísto toho jsme použili řadu behaviorálních, elektrofyziologických a výpočetních metod, které nám umožnily odhalit, jak lidé vnímají řeč. Například jsme testovali percepci zvukových segmentů, které nemají v daném jazyce žádné jméno a nejsou spojeny s žádným pojmem, ale které obsahují některé z rysů přítomných v hláskách mateřského jazyka našich posluchačů. Usuzovali jsme, že pokud budou lidé schopni poslechově kategorizovat takové neznámé zvuky stejně, jako kategorizují zvuky svého mateřského jazyka, bude to znamenat, že fonologické rysy existují jako percepční kategorie.

Výsledky našich experimentů naznačují, že dospělí posluchači vskutku vnímají hlásky prostřednictvím fonologických rysů. Například, zvuk jako je [o:] je vnímán prostřednictvím svých fonologických rysů /plynulost/ + /zaoblenost/ + /dlouhé trvání/ (viz. levá část obrázku výše). Tento výsledek je také znázorněn na přebalu této knihy, kde je fyzický objekt  vnímán jako skupina kontrastivních rysů /krychle/ + /barvy/ + /rotace/.

Naše zjištění dále ukazují, že během porozumění řeči rozeznávají posluchači ty fonologické rysy, které mají kontrastivní funkci v jejich vlastním jazyce. Například, pokud by – v imaginárním jazyce – byl zvuk [o:] vždycky vyslovován se špičkou jazyka vystrčenou z úst, rodilý mluvčí takového jazyka by pravděpodobně vnímal [o:] jako /plynulost/ + /zaoblenost/ + /dlouhé trvání/ + /jazyk venku/. Pro znázornění tohoto výsledku použijeme opět náš oblíbený objekt. Pokud by – v imaginár-

ním světě – měla úplně každá  nejen rotující části, ale bylo by možné ji i celou zmačkat, byla by pravděpodobně vnímána jako /krychle/ + /barvy/ + /rotace/ + /stlačitelnost/.

A nakonec, výsledky našich výpočetních simulací ukazují, že když si virtuální nemluvnata osvojují jazyk, tak se nejprve učí interpretovat hlásky jako celistvé segmenty, ale po dostatečné zkušenosti se svým mateřským jazykem si vytvoří abstraktní kategorie i pro fonologické rysy těchto hlásek. Velmi malé dítě tedy možná nejprve vnímá zvuk [o:] jako jeden celistvý segment /o:/. Jak se ale setkává s mnoha různými instancemi hlásky [o:] a se všemi dalšími hláskami a slovy svého jazyka, tak přijde na to, že hláska [o:] má fonologické rysy /plynulost/ + /zaoblenost/ + /dlouhé trvání/, které ji odlišují od hlásky jako je [e:], jež není vyslovována se zaoblenými rty, nebo od hlásky jako je [p], jež není plynulá ani dlouhá. I tento výsledek znázorníme na Rubikově kostce. Zpočátku může dítě vnímat  jako jeden celek, ale jak získává zkušenosti s okolním světem, přijde na to, že tento objekt má kontrastivní rysy /krychle/ + /barvy/ + /rotace/, které jej odlišují od objektu jako je , jež nemá rotující části, nebo od objektu jako je , jež nemá rotující části ani tvar krychle.

Pro zrekapitulování, výzkum popsany v této dizertační práci měl za cíl objasnit roli fonologických rysů v procesu porozumění řeči. Naše výsledky naznačují, že fonologické rysy jsou percepčními kategoriemi, skrze něž posluchači vnímají řeč. Naše zjištění dále svědčí o tom, že uživatelé jazyka se naučí rozpoznávat ty fonologické rysy, které jsou relevantní v jejich vlastním jazykovém prostředí.

Všechny odkazy na percepci objektů mají za účel pouze ilustrovat naše zjištění o percepci *řeči* na příkladech, jež jsou bližší běžnému čtenáři. Naše závěry nejsou míněny jako závěry o percepci objektů obecně.

CURRICULUM VITAE

Kateřina Chládková was born on 1 June 1984 in Šternberk, Czechoslovakia. She obtained her BA degree (cum laude) in English and Dutch philology from Palacký University in Olomouc, Czech Republic. During her undergraduate studies, Kateřina spent a semester at Utrecht University, the Netherlands, where she later also worked as a research assistant. In 2009, Kateřina obtained her MA degree (cum laude) in General Linguistics from the University of Amsterdam, the Netherlands. Between 2009 and 2013, she carried out PhD research at the Amsterdam Center for Language and Communication (ACLIC), University of Amsterdam, which resulted in the present dissertation. Kateřina currently works as a post-doctoral researcher at the ACLIC. She is a recipient of the Endeavour Research Fellowship, funded by the Australian Government, which gives her the opportunity to spend six months of 2014/2015 at the University of Western Sydney, Australia.