Universiteit
Leiden
The Netherlands

Sound of mind: electrophysiological and behavioural evidence for the role of context, variation and informativity in human speech processing

Nixon, J.S.

Citation
Nixon, J. S. (2014, October 14). *Sound of mind: electrophysiological and behavioural evidence for the role of context, variation and informativity in human speech processing*. Retrieved from https://hdl.handle.net/1887/29299

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/29299

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/29299 holds various files of this Leiden University dissertation.

Author: Nixon, Jessie Sophia
Title: Sound of mind: electrophysiological and behavioural evidence for the role of context, variation and informativity in human speech processing
Issue Date: 2014-10-14

SOUND OF MIND

Electrophysiological and behavioural evidence for the role of context, variation and informativity in human speech processing

JESSIE SOPHIA NIXON

SOUND OF MIND

Electrophysiological and behavioural evidence for the role of context, variation and informativity in human speech processing

語境、變異及信息度在人類言語加工中的作用：
來自反應時、眼動及腦電的證據

Jessie Sophia Nixon

ISBN: 978-90-8891-975-6

© 2014, Jessie Nixon

Cover design by Kamto Gary Wong
http://www.kamto.biz
Document typeset in LaTeX 2$_\varepsilon$ using the Memoir package
Printed by Proefschriftmaken.nl

SOUND OF MIND

Electrophysiological and behavioural evidence for the role of context, variation and informativity in human speech processing

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 14 oktober 2014
klokke 13:45 uur

door

Jessie Sophia Nixon
geboren te Auckland,
New Zealand

Promotor:	Prof.dr. N. O. Schiller
Co-promotor:	Dr. Y. Chen
Beoordelingscommissie:	Prof.dr. R. H. Baayen
	Dr. M. Corley
	Prof.dr. V. J. J. P. van Heuven
	Prof.dr. C. C. Levelt
	Prof.dr. R. P. Sybesma

For Margot

Acknowledgments

I consider it a great privilege to have been able to use a little piece of the world's resources in pursuit of one of life's greatest pleasures: expanding one's understanding of the world. For that, I would like to thank my supervisors, Yiya Chen and Niels Schiller for inviting me to Leiden for this research. I feel extremely fortunate to have had the opportunity to follow my curiosity in carrying out this research over the last several years. It has been highly rewarding in many ways. I have learned a lot. Not only about language and cognition and science and statistics, but about life, as well.

Many people have contributed to this journey. Perhaps one of the biggest contributions has been in the form of the warm, supportive and fun environment created by my fellow PhDs and other colleagues at LUCL. Among them were my two wonderful paranymphs, Allison and Marijn, who also helped to ease the stress of these final preparations and even make them into something fun. I have fond memories of numerous outings, extended hacky sessions and the 1166 lunches. Some of the many others who made LUCL a special place to be were Linda Badan, Yifei Bi, Enrico Boone, Martine Bruil, Edoardo Cavirani, Camelia Constantinou, Elly Dutton, Margarita Gulian, Juliette Huber, Olga Kepinska, Martin Kohlberger, Orsat Ligorio, Sara Lusini, Marieke Meelen, Khalid Mourigh, Victoria Nyst, Ongaye Orkaydo, Stanly Oomen, Leticia Pablos, Piotrek Pisarek, Christian Rapold, Bobby Ruijgrok, Alexander Schwager, Mulugeta Tsegaye, Daan van de Velde, Rebecca Voll, Josh Wilbur, Man Wang and Junru Wu.

There are also a number of people who have made practical contributions to my research. Rinus Verdonschot was a big help when I first arrived in Leiden. You seemed to spout a continuous stream of useful tips on any number of things from statistics, programming software and experiment design to medical care, economic eating places, and the current best buy in mobile phones. Not to mention your help with running participants when you were in Beijing. Many thanks also to Kalinka Timmer for hours spent working with me on all aspects of running EEG experiments, from equipment set-up to running analyses in

SPSS. I learned a huge amount and had a lot of laughs along the way. I would like to thank Hamutal Kreiner for your expertise, enthusiasm and inspiration during my Masters dissertation. By teaching me to use LME and R, you set me up with the technical skills that got me started in my PhD. I don't know what I would have done without them. Thanks to Fermín Moscoso del Prado Martín and Holger Mitterer for helpful advice on the analysis of Chapter 2. Thank you, Antoine Tremblay, for generously answering all my emails relating to the EEG pre-processing with your R packages. Thanks to Fabian Tomaschek and Denis Arnold for improving my contour graphs and to Daan van der Velde for translating my Nederlandse Samenvatting.

Special thanks to Jacolien van Rij. I consider myself incredibly lucky that you decided to come and teach the LOT course in Groningen with Martijn Wieling and became interested in my data. I have learned so much over the last year working with you on the eye tracking data in Chapter 5. I am indebted to you for the time you invested working with me and teaching me how to use Generalised Additive Modelling. You are an excellent teacher and an excellent researcher. Thanks to Martijn Wieling, not only for the excellent LOT course, but also for your enthusiasm about my data and readiness to answer questions about GAMs. Special thanks also to Harald Baayen for generously hosting me in the Quantitative Linguistics Group lab in Tübingen while I was analysing the eyetracking data. Thanks for being excited about my data and for your insights about the analysis–your enthusiasm was a huge encouragement.

I would also like to thank the people at the Chinese Academy of Science Psychology Institute who provided me with expert help, as well as making me feel extremely welcome during my two research trips to Beijing.

To my dear whanau. Those of you who have been around me the last few years will know how proud and how lucky I feel to have such a wonderful family. Dad and Fran, Jake and Tom, with your love and support, your pride, warmth, fun and laughter, I keep you close to me, even from afar. Much love.

To Kamto. You enhanced the book greatly with your cover design. You also helped a lot with many practical and technical issues and averting near disasters along the way. But most of all thank you for sharing your time with me, your humour, creativity, your sense of beauty and your insights about the world.

To Mum, you are part of this book because you are a big part of who I am. You taught me how to be happy in the world and that is the one of the most precious things I have. This book is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Communication involves transmission of a message. In the case of spoken language, transmission takes physical form in acoustic waves conveyed from the vocal folds and articulators of the speaker to the ears of the listener. This particular mode of transmission—unlike some other forms such as text—uses a continuous (non-discrete) signal. Transmission in this form is highly susceptible to the introduction of noise from motor planning and articulation, environmental noise and perceptual processes. Therefore, the acoustic signal provides only probabilistic information about a speaker's intended message. In any given utterance, various acoustic cues, such as pitch height (i.e. fundamental frequency), duration (of, for instance, phonation, frication or silence) and formant frequency all fall somewhere on a continuum. Depending on the language, certain of these acoustic cues might be relevant to determining the spoken message. Vital to the use of acoustics as a medium of communication is that within any particular language, acoustic cues pattern in language-specific ways along language-specific acoustic dimensions to create speech sound contrasts. In English, a message containing *PIN* is signalled by a longer period of aspiration (voice onset time, VOT) following the release of the bilabial stop than a message containing *BIN*. This is regardless of whether the utterance is produced in the upper or lower part of the speaker's fundamental frequency range, since pitch is not contrastive in English (i.e. it is a non-tonal language). Generally speaking, if the VOT is, say, 0 ms, the probability that the intended message was *BIN* will be higher than that it was *PIN*. The probability would also be higher than if the VOT were 50 ms. But if the word was preceded by, 'The picture was tacked to the wall with a drawing ...', this would increase the probability of *PIN* relative to *BIN*, even with a VOT of 0 ms. Of course, since VOT is a continuous measure, its values in a particular instance can fall either side of or anywhere between these prototypical values of English /b/ and /p/.

Within information theory, communication is conceptualised as (complete or incomplete) reconstruction at one point of a message issued at another point (Shannon, 1948), such as, for instance, from inside the head of a speaker to inside the head of a listener. Comprehension is considered not as decoding of messages, but rather as selection of a particular message by discriminating it from all other possible messages. Such reconstruction by the listener of a speaker's intended message is made possible by shared code. An important aspect of this conceptualisation of communication is that meaning is not in the code itself. The code is simply a means of reference, which reduces the uncertainty of the listener. Ramscar and Baayen (2013) offer the simplified example of a world in which there are only two experiences, "being hungry; being

satiated" (and no noise), for which communication would require only a one-bit signal (0 or 1). At the sentential/utterance level (and above), it seems that our choices in communication are far from binary. There are a potentially infinite number of messages we might choose to convey at any time, including situations where concepts are previously unknown to the listener. But at a lower level, speakers of a language have, to some extent, a shared code in that they each have associations with a largely overlapping set of words and phrases. However, two aspects of real-world spoken language make it somewhat more complicated than the binary choice presented above. Firstly, the degree to which speakers and listeners share code is variable, since each language user learns through associations between the code (e.g. words, phrases, speech sounds) and individual experiences. Secondly, as discussed above, the system is inherently noisy. Therefore, determining a speaker's intended message must be done by assessing the relative probabilities of the set of possible messages based on the available cues. The relationship between this continuous and inherently noisy signal and the discrete nature[1] of the underlying messages forms the basis for this thesis.

How is such highly variable, non-discrete acoustic information utilised during language processing? Early accounts assumed phonology to be processed in terms of (optimally) functional units that distinguish between lexical items: phonemes. Phonemes were conceptualized as abstract, idealized representations of sound (Foss & Swinney, 1973; Meyer, 1990, 1991; Roelofs, 1999). In most experiments investigating phonology, phonemes constitute the most fine- grained measure of phonological relatedness. In addition, some of the most influential models of language production (Dell, 1986, 1988; W. J. M. Levelt, Roelofs & Meyer, 1999; Indefrey & Levelt, 2004; W. J. M. Levelt, 2001) posit lexical access to involve activation of sequences of phonemes. However, recent evidence suggests that processing of phonetic information goes far beyond distinguishing phonemes (Clayards, Tanenhaus, Aslin & Jacobs, 2008; Goldrick & Larson, 2008; Ju & Luce, 2006; McMurray, Aslin & Toscano, 2009; Mitterer, Chen & Zhou, 2011; Newman, Clouse & Burnham, 2001).

At birth, infants know very little about what acoustic cues will be relevant in their native language(s). But infants are highly sensitive to

[1]That is not to say that there is not gradience in conceptual representations. Certainly, there are more and less prototypical instances of semantic categories, for example. However, in most cases, contrasts in continuous acoustic cues (e.g. VOT) signal discrete contrasts between lexical items. For instance, the VOT in an intended utterance of, say, 'bin' can be more /b/-like (0 ms) or somewhat more /p/-like (e.g. 30 ms). But the increased VOT does not correspond to a somewhat 'pin'-like semantic representation. The VOT is simply a more or less effective cue for discriminating between alternative possible intentions of the speaker.

fine-grained acoustic information and are able to detect subtle differences in the acoustic signal. Over time, with increased linguistic experience, listeners lose sensitivity to acoustic variation that is not contrastive in the languages they use. In a sense, it may seem surprising that learning should involve loss of sensitivity. However, prioritising sensitivity to particular cues at the expense of others is a fundamental part of becoming more finely attuned to relevant contrasts. Forming associations between linguistic events, such as (particular ranges of) acoustic cues, and particular linguistic outcomes begins in early first language acquisition, and is continually updated throughout the lifetime. Learning to ignore irrelevant cues is critical to increasing the accuracy and effectiveness of predictions (Baayen, Hendrix & Ramscar, 2013).

Phonetic variation is a fundamental property of speech. But for this statement to make any real sense with respect to language processing, it assumes the existence of or reference to speech categories within which the variation occurs. The research presented in this thesis uses a variety of experimental and statistical methods to inform our understanding of how healthy adults process phonetic information. Specifically, it investigates how native speakers of Mandarin, Cantonese and Dutch process contrastive speech sound categories and within-category variation during speech perception, production and reading aloud. Chapter 2 investigates the nature of phonological processing during production and visual processing of allophonic variants of Mandarin tones, i.e. tone sandhi. Chapter 3 investigates two types of sub-phonemic information: the first is allophonic variation during reading aloud; the second is sub-phonemic features. The fourth chapter examines how the surrounding phonetic context influences neural activation during processing of tonal variants. The fifth chapter investigates the role of statistical information in perception of speech contrasts.

1.2 Multi-level phonological processing in speech production

A long line of empirical psycholinguistic research has employed phonological priming methods, in which responses to target words are facilitated or inhibited by overlapping phonological information in congruent primes, relative to control primes which do not contain phonological overlap with the target (Costa & Caramazza, 2002; W. J. Levelt et al., 1991; Meyer & Schriefers, 1991). What is the nature of the phonological information that leads to facilitation in these studies? Is phonology processed in terms of speech categories, or are these facilitatory effects due to similarities in the actual acoustic realisation of the speech sounds? This was the question addressed in Chapter 2 (see also Nixon, Chen &

Schiller, 2014).

In order to tease apart these two types of phonological similarity, I investigated how allophonic speech variants are processed during speech production. Allophones are sounds that differ in acoustic form, but are considered to belong to the same sound category. The difference between the sounds in the words 'pin' and 'bin' can be described in different ways. They can be thought of as words that consist of different whole syllables 'pin' versus 'bin', or as words that differ only in the first sound category /p/ versus /b/ or as words that differ in various acoustic properties, including the onset time of the vowel and other acoustic properties such as formant values of the vowel. In words like 'spin', where the first sound is /s/, the second sound is considered to belong to the same sound category /p/ as in the word 'pin'. But acoustically, the voice onset time falls between the /p/ of 'pin' and /b/ of 'bin'.[2] Therefore, the phoneme /p/ is described as having (at least) two allophonic variants: a canonical, aspirated allophone [p^h] and an unaspirated allophone [p] following /s/.

Little is known about how such phonetic variation is processed. In most experiments investigating phonology, phonological relatedness is measured in terms of phoneme overlap. However, describing phonological processing simply in terms of phonemes, it is difficult to account for the kind of phonetic variation that occurs in real speech. How are speakers able to select the appropriate form? To what extent do top-down and bottom-up information shape context-dependent variation? Do speakers retain the same higher-level speech sound category, regardless of context (i.e. bottom-up processing, from the speech category to the whole word)? Does processing occur top-down, so that the surrounding context determines which variant is activated? For speech categories that have more than one variant realisation (that is, more than one peak in the distribution of acoustic cues), one possibility is that processing of that speech category always involves activation of both (or all) variants, regardless of context. Alternatively, which variant(s) are activated may be determined top-down by context, so that only the appropriate variant is activated for any given context.

This study made use of the picture-word interference paradigm (Damian & Martin, 1999; Lupker, 1982; Rosinski, Golinkoff & Kukish, 1975; Schriefers, Meyer & Levelt, 1990; Starreveld, Heij & W., 1996). In this paradigm, participants see pictures on a computer screen (e.g. of a cat) and name them as quickly and accurately as they can. Superimposed on the pictures are *distractor words* that are phonologically (or sometimes orthographically or semantically) related to the target picture

[2] It is unclear whether, cognitively, 'spin' consists of the same phonetic units as 'pin' plus an initial 's', or whether the words are simply processed as having different onsets. Here, the 'p' is assumed to belong to the same speech category in both words for the purposes of illustration of allophonic variation.

name. Participants are instructed to ignore the distractor words. However, because visual word recognition is extremely rapid—faster than picture naming—if target pictures and distractor words appear on screen at the same time, information from the distractors becomes available before or during retrieval of the picture name. Distractors that sound similar to the target facilitate production, relative to distractors that sound different, because shared phonological information is activated by the distractor word, making it easier to retrieve the target from memory. For example, overlapping phonological[3] information in the distractor word *cap* superimposed on a picture of a cat helps participants retrieve the picture name more quickly than the distractor word *book*, which has no phonological overlap.

The *phonological facilitation* effect in the picture-word interference paradigm was used to investigate how phonological information is processed when native Mandarin speakers produce words containing tonal variants. Beijing Mandarin Tone 3 usually has a low tonal contour, but when followed by another Tone 3 character, it has a rising contour, which makes it sound like Tone 2. This rising-contour variant of Tone 3 is referred to as third tone *sandhi*. Sandhi words are therefore phonologically related to both Tone 3 and Tone 2 words. This characteristic of the tonal system of Beijing Mandarin allowed for the manipulation of two types of phonological relatedness: tone contour and tone category. Sandhi words overlap with Tone 3 words in terms of the Tone 3 category (i.e. the *toneme*), but the actual realisation of the tonal contour is different (rising versus low). Sandhi words are also phonologically related to Tone 2 in that they have the same, rising contour, even though they belong to different tone categories. In Experiment 1, target pictures had names that were sandhi words. Distractors were canonical (low) tone 3 words (*toneme* condition), tone 2 words (*contour* condition) or an unrelated tone (control condition). Since words were overtly produced in this experiment, we might expect the actual realisation of the context-specific allophone (contour) to play an important role, since the speaker needs to produce this form. Therefore, in Experiment 2, the target and distractor conditions were reversed so that the phonetic variants were not overtly produced, but only processed visually in distractors that participants were instructed to ignore. In addition, in both experiments, relative timing of presentation of the target and distractor was manipulated in order to investigate the time course of processing. Results were analysed using linear mixed effects regression modelling (Baayen, 2008; Baayen, Davidson & Bates, 2008). In both experiments, there was evidence of both category-level and instantiated,

[3]In this case, there is also overlapping orthographic information, but I will not discuss that here.

context-specific processing. This indicates automatic multi-level phonological processing in both overt speech production and visual processing of written words. Interestingly, there were differences in the time course of activation of these two representational levels, depending on the mode of processing.

1.3 Does reading aloud involve sub-phonemic feature processing?

As described above, Chapter 2 shows that both speech production and visual processing of words involve multi-level phonological processing: there is activation of both the speech category and an instantiation of the particular speech sound appropriate to the phonetic context in which it occurs. Chapter 3 investigated a third way in which speakers may process acoustic information. So far, I have discussed phonetic information in terms of contrastive speech categories, such as /b/ versus /p/ and Tone 2 versus Tone 3. However, as noted above, any two contrastive sound categories are signalled by combinations of acoustic cues. Contrasting acoustic cues can be categorised into phonetic features. For example, the Dutch sounds /b/ and /p/ contrast in voicing, for which the primary acoustic cue is voice onset time. The sound pairs /p/ and /b/ are produced at the same place of articulation (bilabial), but contrast in place of articulation with /t/ and /d/ (alveolars). Chapter 3 investigates whether such phonetic features are processed during reading aloud. If it is found that feature information is activated during reading aloud, this would provide evidence of a further level of sub-phonemic processing.

In addition to the question of phonetic features, Chapter 3 also extends the question discussed in Chapter 2 regarding the nature of representations activated when one speech sound category has two or more variants. In Dutch, voiced stops (/d/ and /b/) have at least two variants. In syllable-initial position, they are voiced; however, word-finally, they are *devoiced*. That is, the voice onset time is similar to voiceless stops. For example, the word *hout* ('wood') and the word *houd* ('to hold') are homophones in Dutch. Although the acoustic realisation is similar, the question of how these sounds are processed online is not well understood. This study investigates whether the voicing distinction in word-final voiceless and devoiced is retained during reading aloud. If there is facilitation for voice-congruent primes (e.g. *HUIB – huid* compared to control primes *HUIP - huid*, this provides evidence for differential processing between voiceless and devoiced speech sound categories, despite similarities in overt production.

Very little is known about sub-phonemic processing in reading aloud.

Phonology has generally been measured at the phonemic or word level in the reading aloud literature. For example, a number of studies have shown that reading aloud is faster when targets are preceded by primes that share the same onset phonemes, compared to those whose onset phonemes differ (Kinoshita, 2000; Kinoshita & Woollams, 2002; P. Mousikou, Coltheart, Finkbeiner & Saunders, 2010; Timmer & Schiller, 2012; Schiller, 2007). Some models of speech production have proposed a featural level of representation. For example, Dell (1986) proposes that once a phoneme is selected, activation of the selected phoneme spreads to its constituent features. Other models propose that these sub-phonemic representations consist of articulatory gestures (e.g. Goldstein, Pouplier, Chen, Saltzman & Byrd, 2007). Phonetic features have been found to play a role during speech perception (see Chládková, 2014, for a full review). In reading aloud, since perceptual information is received orthographically, coded in terms of phonemes, sub-phonemic feature information may play a lesser role compared to speech perception. One recent ERP study found evidence that the voicing feature is processed in English silent reading (Ashby et al., 2009). However, because vowel duration before word-final stops differs between voiced and voiceless stops in English, it is not clear whether the effect was due to the voicing contrast in the consonant itself or due to vowel duration, or a combination of the two.

In order to investigate whether feature information is processed during reading aloud, in the study presented in Chapter 3, reaction times and EEG measures were recorded as participants read aloud real Dutch words (e.g. *huid* 'skin'). Each target word was preceded by a brief presentation of a masked non-word prime in which the final sound matched in voicing (*huib*), or place of articulation (*huit*) or mismatched in both voicing and place (control condition; *huip*). All prime conditions differed from the target equally in terms of both phonemes and letters. Only when measured at the feature level is there overlap in the matching conditions, compared to the control condition. Therefore, reduced response latencies and EEG amplitude in the matching conditions would provide evidence that processing occurs at the feature level.

Both the reaction times and the EEG measures show significant effects of feature match. Reaction times, analysed using linear mixed effects regression modelling, were significantly faster when prime and target matched in voicing, than when they did not. Consistent with the behavioural data, there was significantly less negativity in the early time window in the voice-match condition, compared to the control condition. This finding not only indicates sub-phonemic processing during reading aloud, it also has implications for processing of allophonic variation. Due to word-final devoicing, voiced stops have two variants in Dutch. Therefore, the finding that voice-congruency facilitated reading

aloud provides support for the proposal that sub-phonemic contrasts are represented as phonetic features, as proposed by, for example, Dell (1986), rather than as articulatory gestures (e.g. Goldstein et al., 2007), since the motor movement is similar for voiceless and (de)voiced final stops. Another possibility is that, as we saw for the allophonic variants in Chapter 2, both a contrastive category level (in this case phonetic features) and a context-specific instantiation (such as Goldstein-type articulatory gestures) are processed at this sub-phonemic level.

1.4 Contextual effects in sub-phonemic processing during reading aloud

When a speech category has more than one realisation, are all variants automatically activated whenever that speech category is processed? Or is activation constrained top-down by the phonetic context? As described above, Chapter 2 shows that both visual processing and overt production of allophonic variants involve activation of the context-specific realisation, as well as the speech category. Because activation in the two phonetic variants is not directly compared in that study, it is possible that the non-canonical variant is always activated, even when it is not required in the context. Chapter 4 investigates this question.

As mentioned above, previous studies of phonological processing during reading aloud have generally used homophone primes or phonemic overlap. Very little is known about processing below the phoneme level. In speech production, as we will see in Chapter 2, there is activation of the context-specific contour, as well as the speech category. During speech production and visual processing of words, acoustic properties of a prime can facilitate production of a target word, even if there is no category overlap between prime and target. A second question investigated in Chapter 4 is whether this cross-category facilitation can be found in a different task—reading aloud—when acoustic similarity in visually presented prime words is briefly presented (48 ms) and masked. Since the primes do not reach the level of articulation planning, it might be expected that context plays a lesser role and that only processing of the general speech category occurs. If context-specific differences in processing are found, this would provide strong evidence for automatic activation of context-dependent processing of speech variants.

Neural activity of native speakers of Beijing Mandarin was recorded as they read aloud Tone 2 Mandarin words, preceded by briefly presented sandhi or low Tone 3 words as masked primes. The initial character of critical primes was always Tone 3, so primes always differed from targets in terms of tone category, but either matched or mismatched the tone contour. In addition, the initial character of primes was identical

between conditions. Only the phonetic context provided by the tone of the following prime differed between conditions. Therefore, any differences found between conditions must signal different activation levels of the two variants due to context-specific processing of the tonal allophones.

An important aspect of this study is the inclusion of individual item information in the analysis. Traditional ERP analysis averages over all trials per condition, so that all information about individual word characteristics is lost. The 'language-as-fixed-effect fallacy' (Clark, 1973; Coleman, 1964) suggests that, in language research, excluding items from random effects analyses can be problematic statistically, and can lead to type 1 error. Although item analyses have been widely adopted in behavioural studies, this point has often been ignored in EEG research. This is presumably due to difficulties in coding the EEG signal and limitations of software developed for EEG analysis. Chapter 4 used an alternative method of analysis, *generalised additive mixed modelling* (GAMM; Wood, 2006). GAMMs are a type of generalised linear model, which use non-linear smooths to model linear predictors. This allows us to investigate, for example, changes in amplitude over the course of the trial. We also included trends over time as random effects for subjects and items. In addition, the model included a predictor of prime type over time, predictors of prime and target frequency and their interactions over time.

The best-fit linear mixed effects regression model found that the differences in reaction times between conditions were not significant. However, between-condition differences were found in the EEG data. This effect interacted with prime and target frequency. When, due to the tonal context, the prime and target overlap in contour, the contour no longer discriminates between prime and target. Under these conditions, the a priori probabilities of the prime and target come into play. This indicates, firstly, that the acoustic similarity in the congruent prime affects processing of the target word, even though prime and target belong to different tone categories. Secondly, it indicates that this phonetic information is context dependent. Since initial characters were identical between conditions, this suggests that the top-down processing of the surrounding phonetic context promotes activation of the appropriate allophonic variant. Further, from a methodological point of view, the interaction with prime and target frequency also highlights the importance of including individual item characteristics in EEG studies of language.

1.5 Acoustic cue variability and informativity in perception of speech contrasts

While chapters 1 to 4 investigated systematic contextual variation and sub- phonemic feature processing during speech production and reading aloud, Chapter 5 explores a different type of variation. It investigates how statistical variance or noise in the signal affects discrimination during speech perception. As discussed above, listeners have access to only highly variable, non-discrete acoustic information to extract a speaker's intended message from the speech signal. Regularities in speech allow infant and adult speakers and listeners to form contrastive speech sound categories that can be used to discriminate between word meanings, such as between the words 'pin' and 'bin'. However, the actual physical form of these speech sounds varies substantially—from speaker to speaker, from word to word and even between different instances of the same word spoken by the same speaker in a controlled setting.

The last couple of decades have seen growing interest in how statistical information is utilised in language processing, particularly in first language acquisition, but also more recently in adult native and second language processing. A continuum of speech sounds presented in a unimodal distribution is more likely to be categorised as a single sound, compared to when the same continuum is presented in a bimodal distribution (Maye & Gerken, 2000; Maye, Werker & Gerken, 2002; Maye, Weiss & Aslin, 2008). Beyond the *number* of peaks in the input distributions, few studies have investigated how the shape of the distribution—that is, the amount of variation or noise—affects processing during speech perception.

Chapter 5 investigates how the amount of variation in the acoustic signal affects certainty during perception of Cantonese speech sound contrasts. Most studies of distributional learning in adults have used offline categorisation judgments to assess learning. For example, a number of studies have used offline measures to investigate distributional effects in non-native acquisition of Dutch vowel contrasts (Escudero, Benders & Wanrooij, 2011; Gulian, Escudero & Boersma, 2007; Wanrooij, Escudero & Raijmakers, 2013) and Cantonese tone contrasts (Zhao, 2010). Categorisation measures provide information about the final outcome of a decision, but they do not directly measure the online perceptual processes. It is often implicitly or explicitly assumed that when participants categorise tokens into a single category, rather than two separate categories, this is because the tokens were not discriminated. However, this assumption may not be warranted, since the task requires a binary choice. That is not to say that offline categorisation measures are uninteresting. However, examining the early perceptual processes and the

moment-by-moment changes in processing over time up to the point of the decision could certainly inform our understanding of the effects of acoustic variation on speech perception. The data presented in Chapter 5 deal with this question. Eye movement measures were recorded as participants heard either high-variation (wide distribution) or low-variation auditory stimuli (narrow distribution) and clicked on the picture they heard. Results were analysed using generalised additive mixed modelling (Wood, 2006). This allowed the eye movement patterns over time to be analysed, rather than collapsing over the whole trial. This statistical method also made it possible to model complex interactions of time, acoustic variant, distribution condition and other predictors, such as trial and manner of articulation.

Chapter 2

Multi-level processing of tone in speech production and visual word processing

A version of this chapter is published as:
Nixon, J. S., Chen, Y. & Schiller, N. O. (2014). Multi-level processing of phonetic variants in speech production and visual word processing: evidence from Mandarin lexical tones. Language, Cognition and Neuroscience. doi:10.1080/23273798.2014.942326

Abstract

Two picture-word interference experiments provide new evidence on the nature of phonological processing in speech production and visual word processing. In both experiments, responses were significantly faster either when distractor and target matched in tone category, but had different overt realisations (toneme condition) or when target and distractor matched in overt realisation, but mismatched in tone category (contour condition). Tone 3 sandhi is an allophone of Beijing Mandarin Tone 3 (T3). Its contour is similar to another tone, Tone 2. In Experiment 1, sandhi picture naming was faster with contour (Tone 2) and toneme (low Tone 3) distractors, compared to control distractors. This indicates both category and context-specific representations are activated in sandhi word production. In Experiment 2, both contour (Tone 2) and toneme (low Tone 3) picture-naming was facilitated by visually presented sandhi distractors, compared to controls, evidence that category and context-specific instantiated representations are automatically activated during processing of visually presented words. Combined, the results point to multi-level processing of phonology, whether words are overtly produced or processed visually.

2.1 Introduction

How are the sounds of language stored in memory and accessed during language production? Early accounts assumed phonology to be processed in terms of (optimally) functional units that distinguish between lexical items: phonemes. Phonemes were conceptualized as abstract, idealized representations of sound (Foss & Swinney, 1973; Meyer, 1990, 1991; Roelofs, 1999). In most experiments investigating phonology, phonological relatedness is measured in terms of phoneme overlap. In addition, some of the most influential models of language production (Dell, 1986, 1988; W. J. M. Levelt et al., 1999; Indefrey & Levelt, 2004; W. J. M. Levelt, 2001) posit lexical access to involve activation of sequences of phonemes.

Phonemes (e.g. /t/ or /k/) are the smallest units of sound that distinguish between words in a particular language (e.g. 'top' versus 'cop' in English). In contrast, allophones vary with phonetic context, but do not affect word meaning. For example, word-initially, English /t/ is aspirated (has a puff of air, e.g. 'top'), but is unaspirated (no puff of air) following /s/, (e.g. 'stop'). Experimental evidence suggests that phoneme-like generalisation plays a role in online speech processing. For instance, in a perceptual learning experiment, McQueen, Cutler and Norris (2006) had Dutch participants perform a training phase of auditory lexical decisions to words in which either the final /f/ or the final /s/ was replaced by an ambiguous [f-s] fricative sound. These words created a lexical bias to interpret the ambiguous sound as a particular phoneme. For example, participants in the ambiguous /f/ condition heard [witlɔ?], where *witlof* is a real Dutch word, but *witlos* is not, thereby creating a bias to interpret the ambiguous sound as an /f/. In the following test phase, participants made lexical decisions to visually presented minimal pair words (for example, doof 'deaf'; doos 'box') preceded by auditory primes containing the ambiguous sound [doo?]. Facilitation depended on which ambiguous phoneme participants were trained with. Participants who heard the ambiguous sound in /f/- words during training were faster to identify visually presented /f/-words (e.g. *doof* 'deaf'), while participants who heard ambiguous /s/ were faster to name /s/-words (e.g. *doos* 'box'). Participants had adjusted ('re-tuned') their perceptual categories by matching the distorted sound to lexical items stored in memory. Importantly, since different sets of words were used in training and test, re-tuning was not restricted to specific words, but instead must have generalized to elements common to both training and test words; that is, to phoneme categories. Similarly, McLennan, Luce and Charles-Luce (2003) found evidence for category-level processing in production. In American English, word-medial /d/ and /t/ are often produced as a flap, making the two sounds ambiguous. In a repetition

priming experiment, McLennan et al. (2003) had participants produce words containing /d/ and /t/, preceded by auditory primes that were either carefully articulated or flapped. Results showed that flapped and carefully produced forms primed each other, evidence for processing at the category level.

On the other hand, there is mounting evidence that processing of phonetic information goes far beyond distinguishing phonemes (Clayards et al., 2008; Goldrick & Larson, 2008; Ju & Luce, 2006; McMurray et al., 2009; Mitterer et al., 2011; Newman et al., 2001; Trude & Brown-Schmidt, 2012). Exquisite perception and memory for detail have also been shown in auditory (Agus, Thorpe & Pressnitzer, 2010) and visual processing (Brady, Konkle, Alvarez & Oliva, 2008). At the extreme, it has been proposed that lexical processing can be explained without any sublexical categories. For example, the memory model MINERVA 2 (Hintzman, 1986) takes phonological representations to be built up from episodic memory traces of whole lexical items. Goldinger (1998) found that MINERVA 2 correctly predicted both reaction times and speakers' spontaneous mimicking of voice onset time in perceived speech, which cannot be explained by purely abstractionist models. This has been taken as evidence that there are no abstract categories below the word level (but *cf* Fowler, 2010; Mitterer, 2006).

Taken together, the above findings suggest that speech processing involves both phonemic and sub-phonemic representations. This conclusion is further supported by recent evidence for both abstraction and detailed information obtained within the same experiment (Mitterer et al., 2011; Nielsen, 2011). For instance, Mitterer et al. (2011) tested the extent to which abstract and detailed acoustic information influence perceptual learning of tones in Mandarin. Analogous to the McQueen et al. (2006) perceptual learning study, listeners heard ambiguous tonal contours (a synthesised continuum between Tone 1 and Tone 2) in phrases that biased interpretation to either Tone 1 or Tone 2. Results showed that participants who received the ambiguous contours in contexts that biased interpretation to Tone 1 in the exposure phase were more likely to perceive ambiguous tones as Tone 1 during test than those who received the ambiguous Tone 2 context. This generalised to words not in the exposure phase, suggesting sublexical abstraction of tone category. There was also a specific-word effect: perceptual learning was greater for exposure-phase words than new words, evidence that detailed acoustic information was retained in representations of individual words. In this paper, we extend the investigation of specificity in lexical prosodic representations to the realm of speech production. In addition, the study makes an important distinction between the level and the nature of phonological representations.

The present study

The aim of the present study is to determine two inter-related aspects of lexical tone processing in Beijing Mandarin. The first concerns whether the level of processing corresponds to the tone category or a context-specific sub-phonemic level. The second question examines whether the nature of the representations is purely abstract or involves an internal instantiation of an actual sound, i.e. the tonal contour. In addition, Experiment 2 investigates whether sub-phonemic processing occurs with visually presented words.

Beijing Mandarin has four lexical tones (*tonemes*) represented schematically in Figure 2.1. Characters[1]. that have the same *segmental syllable* (i.e the non-tonal syllable) can be distinguished by this inherent pitch contour, such as bi2[2] (鼻, 'nose') versus bi3 (笔, 'pen'). In connected speech, Tone 3 (T3) has at least two variants (*allotones*)[3], shown in Figure 2.2. The canonical realisation is the low contour, but preceding another T3 syllable, T3 is realized with a rising contour. This allophonic variant of T3 is known as *third tone sandhi* (hereinafter, 'T3 sandhi'). Tone sandhi refers to the phenomenon whereby the acoustic realization of a tone is influenced by a neighbouring tone in a particular environment. Importantly for the present study, the contour of T3 sandhi is very similar to another tone, Tone 2. Figure 2.3 shows the tonal contours for Tone 2, the canonical, low Tone 3 and T3 sandhi. Detailed acoustic analyses have been able to detect subtle differences between the pitch contours of Tone 2 and the T3 sandhi (Yuan & Chen, 2014). However, listeners generally cannot consciously distinguish between them (Peng, 2000; W. S.-Y. Wang & Li, 1967). Peng (2000) had native Mandarin speakers produce minimal pairs of bisyllabic words that were sequences of either Tone 3 + Tone 3 (sandhi) or Tone 2 + Tone 3. Although subtle acoustic differences were detected, in a following identification task, a different group of native speakers performed only at chance level in distinguishing the two word types.

These aspects of the tonal system of Beijing Mandarin allowed us to manipulate two types of phonological relatedness: tone contour and tone category. In two picture-word interference (PWI) experiments (Damian & Martin, 1999; Lupker, 1982; Rosinski et al., 1975; Schriefers et al.,

[1]The term 'character' is used here in the sense that Chinese speakers would use *zi* (?). It is used to denote the linguistic unit that includes the morpheme, the tonal syllable, the orthographic character and the associations between them.

[2]Mandarin tones are referred to using a number system (Tones 1 to 4). Here, the numeral following the syllable represents the tone number, in this case, Tone 2.

[3]The third tone is sometimes described as 'falling-rising'. However, the rising part of the contour is optional and does not usually occur when there is a following syllable. In addition, the gradient of the fall is very shallow. For these reasons and for simplicity, we refer to the contour as 'low'.

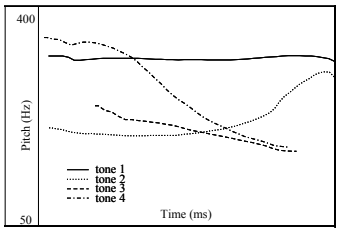


Figure 2.1: Pitch contours of the four tones of Beijing Mandarin

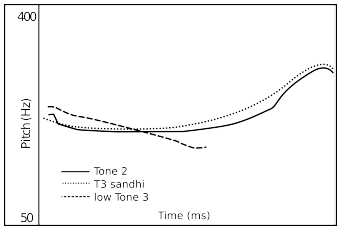


Figure 2.2: Pitch contours of Tone 2, Tone 3 sandhi and canonical low Tone 3

1990; Starreveld et al., 1996), T3 sandhi and Tone 2 words share the same (rising) realisation contour, but belong to different tone categories (*contour* condition), or T3 sandhi and low Tone 3 words share the tone category (T3), but have different overt realisations (*toneme* condition). Experiment 1 investigates phonemic and sub- phonemic processing during overtly produced T3 sandhi words. Facilitation in the toneme condition indicates processing at the tone category level; facilitation in the contour condition indicates sub-phonemic processing of tone. Experiment 2 tests whether these two levels of processing occur during visual processing of T3 sandhi words that are not overtly produced. In addition, both experiments used two stimulus onset asynchronies (SOAs) to investigate differences in the time course of processing between the contour and the tone category. Target picture and superimposed distractor word were presented either simultaneously (SOA = 0 ms) or with the distractor word delayed by 83 milliseconds (SOA = 83 ms). Varying of SOAs has been used in this paradigm to investigate the time course of processing in speech production. Although the details of the stages and their time course are disputed, it is generally agreed across speech production models that producing speech involves access to at least two levels of information: a conceptual level and a form (phonological and/or orthographic) level. The distractor word (and the phonological, orthographic or semantic information it contains) can be made available be-

fore, at the same time as or after the target picture is presented, with differential effects. For example, with visual presentation of distractor words, phonological facilitation from overlap of segmental phonemes has been found from 200 ms preceding up to 100 ms following target presentation, while semantic interference has been found only at simultaneous and positive SOAs, 0 ms to 200 ms (Damian & Martin, 1999). The decrease or disappearance of phonological facilitation between 100 ms and 200 ms presumably occurs because phonological processing has by 200 ms already reached a stage at which the speaker no longer benefits from the segmental overlap.

To date, very little is known about tone processing in speech production in general, or its time course in particular. However, Zhou and Zhuang (2000) found in a PWI experiment that tone processing is faster than segmental processing. While facilitation from segmental overlap occurred at both short and long SOAs, facilitation for tone was found only at the short SOA. We therefore selected a relatively short positive SOA in the present experiments to maximise the chances of obtaining facilitation effects.

If differences in the time course are found between the tonal category and the tonal contour, we see a number of possibilities for how this could manifest. Firstly, it is possible that each word is initially processed holistically, so that each morpheme is processed in its context-specific form. This would then be followed by inductive activation of the context-general tone category. That is, in this scenario, the initial syllable of T3 sandhi words is processed as a rising tone first, followed by activation of the Tone 3 toneme. The second possibility is that activation begins at the general category level, followed by processing of the context-specific variant. If this is the case, we would expect an early effect in the toneme congruent conditions (i.e. at SOA = 0 ms), and late effects of the contour congruent conditions (at SOA = 83 ms). A third possibility is that both of these processes occur simultaneously, leading to simultaneous activation of both levels of processing. Finally, it is also possible that there are differences in the time course of activation of the two levels of processing, depending on the task. We might expect the actual contour of the context-specific variant to play a greater role in overt speech production than in silent processing of written words, while the reverse might be true for the tone category.

2.2 Experiment 1 Tone 3 sandhi picture naming with contour and toneme distractors

Method

In Experiment 1, participants named pictures of objects with T3 sandhi names. T3 sandhi words are made up of two Tone 3 characters. When both characters are Tone 3, the first character has a rising contour, instead of the canonical low contour. Crucially, this rising contour is similar to the contour of Tone 2. Therefore, Tone 2 distractors share the overt realisation—the rising contour—with T3 sandhi target pictures (contour condition). Low T3 distractors differ in overt contour realisation, but match in tone category (toneme condition).

If T3 sandhi picture naming is facilitated by toneme distractors, this demonstrates that there is activation at the tone category level, despite differences in the overt pronunciation. If contour distractors facilitate T3 sandhi picture naming, this indicates two things. Firstly, it is evidence for context-specific processing of the T3 sandhi allotone. Since, in most contexts, the contour of T3 is unrelated to T2, shorter latencies in the contour condition indicate a context-specific representation of the T3 sandhi allotone (rather than the general Tone 3 category). Secondly, even though T3 sandhi and T2 have similar realisations, if they are represented in a purely abstract form, they could still be processed as separate categories. Only through similarities in the actual pitch contour can facilitation from contour distractors occur. This suggests activation of an *instantiated* representation of the tonal contour.

Participants Thirty native speakers of Beijing Mandarin (24 female; mean age 21.5), students at universities within Haidian district in Beijing, were paid for their participation. All participants and their parents were born and raised in Beijing, except three participants who had one parent from the nearby Northern Mandarin-speaking province of Hebei, two participants for whom both parents were from Hebei, and one participant whose parents were from Shanghai.

Stimuli The experimental conditions and sample stimuli are shown in Table 2.1. Critical targets were 27 pictures with two-character T3 sandhi names. Pictures were black-on-white line drawings selected from the MPI (12 pictures, two with modifications) and the Alario and Ferrand (1999) picture databases (three pictures), supplemented with pictures from the internet (12 pictures). Distractors were contour (T2 characters), toneme (T3 characters) and control one-character words (T1 or T4 characters) with the same segmental syllable as the target initial syllable. Contour, toneme and control distractors were matched for word

frequency and stroke number. Targets and distractors were semantically and orthographically unrelated. An additional 27 picture-distractor pairs were used as fillers to add variety and make the design less obvious to participants. None of the characters or initial syllables used in critical trials appeared in filler trials. Word and character frequencies were obtained from Subtlex-CH, a large (46.8 million characters, 33.5 million words) Chinese database based on film subtitles (Cai & Brysbaert, 2010).

Before going on to the experiment design, we make a brief note about the notion of 'word' in Chinese. The distinction between words and phrases is less clear-cut in Chinese than it is in alphabetic languages. Although lexicality could be said to be a gradient property in any language, for alphabetic language speakers, intuitions about what constitutes a word may be so deep-seated that we do not usually define it. Generally speaking, word boundaries are indicated by white spaces in the script. In Chinese script, spaces are instead inserted between characters. Characters correspond not to words, but to single syllables and (almost always) single morphemes. A word can consist of one or more characters. However, native Chinese speakers do not always agree on what constitutes a word versus a phrase. Therefore, in this paper, we have used bigram frequencies as a measure of lexicality. Sandhi stimuli had medium-to-high bigram frequencies, so they are expected to be processed more like 'words' than multi-word phrases. In Experiment 1, mean bigram frequency (measured in mutual information; MI) of 6.3 (SD=3.7). MI is a measure of how likely two characters are to co-occur (see Da, 2004, for an explanation of the method).

With medium-to-high bigram frequencies, one might expect the surface level to play a greater role. However, as described above, McLennan et al. (2003) failed to find evidence for surface-level processing in within-word American English flap production. This has yet to be investigated in tone processing.

Design Experiment 1 consisted of 324 trials, divided into six blocks of 54 trials, with breaks between the blocks. The experiment followed a 3 x 2 within-participant factorial design, with the factors Distractor type (Contour, Toneme, and Control) and SOA (0 or 83). At SOA = 0 ms, the target picture and distractor word appeared simultaneously, while at SOA = 83 ms the distractor word was presented 83 milliseconds after target picture onset. A relatively short delay was selected for the positive SOA because tonal effects have been found to be short-lived relative to segmental effects (Zhou & Zhuang, 2000). The SOA of 83 ms was calculated to match the screen refresh rate (60 Hz). There were 27 trials per condition. Three distractor word lists were constructed for each

SOA, with distractor words divided equally between conditions. Each target word was presented six times (once in each distractor condition for each SOA). The script was programmed to counterbalance the order of presentation of the distractor word lists across participants. All lists were pseudo-randomised for each participant. Each block was preceded by three warm-up trials, which were excluded from analysis.

Table 2.1: Experiment design and sample stimuli Experiment 1

	Target picture	Distractor condition		
		Toneme	**Contour**	**Control**
Tone category	Tone3+Tone3	Tone3+ToneX	Tone2+ToneX	Tone1/4+ToneX
Tonal contour	rising	low	rising	other (high or falling)
Example	fu3dao3 辅导	fu3 斧	fu2 服	fu4 付

Procedure Participants were tested individually in a quiet room at the Psychology Institute of the Chinese Academy of Sciences in Beijing. Stimulus presentation and data acquisition were conducted using the E-Prime 2.0 software package with the addition of a voice key. After being familiarised with target pictures and picture names, participants were seated approximately 60 cm from a 17-inch cathode ray tube computer monitor and given a practice session prior to the actual experiment.

Each experimental trial began with a fixation cross for 500 milliseconds, followed by the target picture for a maximum of 2,000 milliseconds, or until the participant responded. Distractor words appeared superimposed on target pictures either simultaneously (SOA = 0 ms) or 83 milliseconds after picture onset (SOA = 83 ms). An inter-stimulus interval of 500 milliseconds preceded the next trial. Participants were instructed to ignore the words and name the pictures as quickly and accurately as possible. The experimenter coded response accuracy during the experiment. Response time was calculated from the time of target picture presentation until the voice key was triggered by the participant response.

Results

Data were analysed using linear mixed effects modelling, using the lmer function of the lme4 package (Bates, Maechler & Bolker, 2013) see also (Baayen, 2008; Baayen et al., 2008) in R (R Core Team, 2013). Analysis was conducted on the 4,667 data points remaining after stutters, errors, false starts (3%) and null responses (0.8%) were removed. Since error

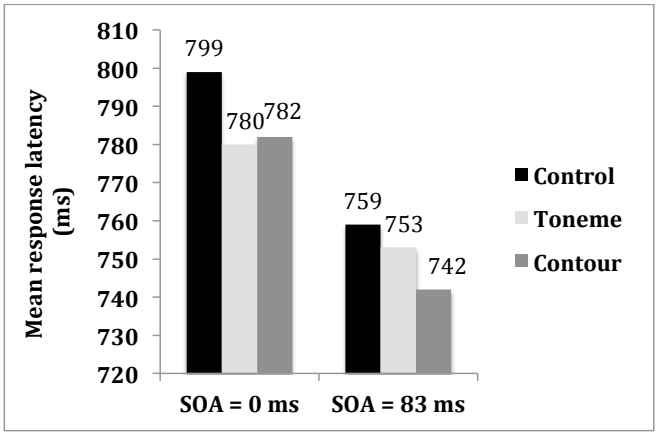


Figure 2.3: Mean reaction times (ms) per distractor type (toneme versus contour versus control) and stimulus onset asynchrony (SOA; 0 ms versus 83 ms) in Experiment 1

rates were low, no further analyses were conducted on the errors. Inspection of response latency distributions revealed a skewed distribution, which was normalized by logarithmic transformation. Mean response times per distractor condition and SOA are shown in Figure 2.3.

There is currently debate in the literature concerning the appropriate method for constructing statistical models. Therefore, in this paper, both forward and backward algorithms were used for model comparison (Baayen, 2008; Barr, Levy, Scheepers & Tily, 2013). The two types of algorithm converged on the same final model. Firstly, a forward algorithm was conducted, which gradually built up complexity in the model. The baseline model was a regression line of log reaction times (log RT), with random intercepts for subjects and target pictures. Each fixed effect and interaction was individually added to the model and tested by comparing the log likelihood ratio to that of the simpler model. Trial was included as a control variable to investigate effects of learning or fatigue over the course of the experiment (Baayen, 2008). Only effects that significantly improved the fit were retained in the final model. Once the fixed effects were established, random effects structure was tested. A random slope was individually added and tested for each of the significant fixed effects. Only random slopes that improved model fit were retained.

It has been argued that random effects structure should be kept maximal, in that model testing should be conducted by starting first with maximal fixed effects, eliminating non-significant predictors, then entering maximal random slopes for all significant predictors of interest

and eliminating only those that do not improve model fit (Barr et al., 2013). Therefore, in addition to the forward algorithm, a backward elimination algorithm was also implemented here. After fixed effects were established, random intercepts and slopes for all significant fixed effects (Trial, SOA and Distractor type) were added to the model.

However, the maximum likelihood estimation of this model failed to reach convergence. This is a common problem with complex maximal models, particularly those with complex random effects structure (Barr et al., 2013). Barr and colleagues suggest that by removing correlation parameters, this problem can be solved while still meeting the objective of maximising the model. Therefore, a model was constructed containing separate by-subject random slopes for Trial, SOA and Distractor type without correlation parameters. With three random slopes, the likelihood estimation still failed to converge. With two random slopes, convergence was reached and model comparisons could be completed. No significant difference was found when Trial or Distractor random slopes were removed, but removing SOA significantly reduced log likelihood ratio. The backward algorithm converged on the same final model as the forward algorithm.

The best-fit model (Table 2.2) included main effects of Trial, Distractor type and SOA, but no interactions, random intercepts for subjects and target pictures, and a by- subject random slope for SOA. Bigram frequency was tested, but did not improve the model as a main effect or interaction with other fixed effects, so was removed. In the model summary in Table 2.2, the control condition at SOA = 0 ms lies on the intercept (the baseline condition) and the estimates show the coefficients for each of the predictors. The Trial coefficient indicates there was a small but significant increase in reaction times across participants over the course of the experiment. The main effect of SOA indicates that responses were faster when distractors were delayed (SOA = 83 ms) than with simultaneous presentation of stimuli (SOA = 0 ms). More importantly, naming latencies were significantly[4][5] shorter for both Contour and Toneme distractors, compared to controls. The effect appears to be slightly stronger in the Contour condition than in the Toneme condition.

Although the log likelihood ratio showed no significant improvement in model fit by adding an interaction between distractor type and SOA ($p > .23$), the mean reaction times (Figure 2.4) suggest differences in

[4]Significance is reported at the 95% confidence level, unless otherwise specified.

[5]In LME, it is unclear what the appropriate degrees of freedom should be. Therefore, in the lme4 package, p-values are not provided in the output. However, the t-value provides confidence intervals for sufficiently large data sets (1,000 data points or more). T-values below -2 or above 2 can be taken as significant at the 95% confidence level (see, for example, Baayen, 2008; Baayen et al., 2008; Baayen & Milin, 2010, for full discussion).

Table 2.2: Results summary Experiment 1: coefficient estimates, standard errors (SE) and t- values for all significant predictors in the log-transformed naming latencies for pictures with Tone 3 sandhi names

Predictor	Coefficient estimate	SE	t
Intercept	6.6108	0.0282	234.75
Trial	0.0014	0.0002	8.15
Distractor Type Contour	-0.0207	0.0065	-3.20
Distractor Type Toneme	-0.0164	0.0065	-2.53
SOA 83	-0.0479	0.0191	-2.51

effects between the SOAs. Since our primary interest was to investigate the effects of different distractors on target picture naming, we split the data set by SOA and ran separate models for each. The model summary for SOA = 0 ms (Table 2.3) shows that the predictors for the SOA = 0 ms model are similar to that of the full data set. Model fit was improved by main effects of Trial ($p = 0$) and Distractor ($p < .02$), but not their interaction ($p = .8$). Random slopes did not improve the model. The model summary shows that for both Contour and Toneme distractors, response times are significantly faster than with the Control distractor.

Table 2.3: Results summary Experiment 1 SOA = 0 ms: coefficient estimates, standard errors (SE) and t-values for all significant predictors in the log-transformed naming latencies for Tone 3 sandhi pictures

Predictor	Coefficient estimate	SE	t
Intercept	6.6099	0.0282	234.80
Trial	0.0015	0.0002	6.53
Distractor Type Contour	-0.0200	0.0089	-2.25
Distractor Type Toneme	-0.0232	0.0089	-2.61

The model for the SOA = 83 ms data set is shown in Table 2.4. There was a main effect of Trial ($p = 0$), but Distractor type only approached significance ($p > .07$). It may be that there was insufficient power in the experiment to yield a significant improvement in model fit for the 3-level factor at the later SOA. However, since the predictor approached significance, we include it in the model here for comparison with SOA = 0 ms. The interaction with Trial was not significant ($p > .3$), but there was a significant random slope for Trial ($p < .01$). With delayed presentation of the distractor (SOA = 83 ms), the t-values of this model

suggest that while T3 sandhi naming seems to be faster in the Contour condition than the Control condition, there is no longer facilitation from Toneme distractors.

Table 2.4: Results summary Experiment 1 SOA = 83 ms: coefficient estimates, tandard errors (SE) and t-values for all significant predictors in the log-transformed naming latencies forTone 3 sandhi pictures

Predictor	Coefficient estimate	SE	t
Intercept	6.5651	0.0275	238.66
Trial	0.0013	0.0004	3.69
Distractor Type Contour	-0.0233	0.0955	-2.44
Distractor Type Toneme	-0.0076	0.0952	-0.80

Discussion

The purpose of Experiment 1 was to investigate whether production of T3 sandhi words activates the Tone 3 toneme, the context-specific rising contour, or both. The results show that both levels of activation occur. Main effects of SOA and Trial indicate that responses were faster when distractors were delayed and that there was a slight increase in reaction times over the course of the experiment. More importantly, there was also a main effect of Distractor type, such that responses were faster when the distractor and target matched in contour, but mismatched in toneme (contour distractors), or when they matched in toneme, but mismatched in contour (toneme distractors), compared to unrelated controls.

Although there were no significant interactions, the numerical means (Figure 2.3) indicate a difference in facilitation effects between SOAs. This was investigated in a separate model for each SOA. Similar to the full data set model, with simultaneous presentation of target picture and distractor word (SOA = 0 ms), faster naming latencies were found when distractors matched either the realisation (contour distractors) or the Tone 3 category (toneme distractors).

At the later SOA, including distractor type in the model resulted in only a marginal improvement of model fit. However, the model summary suggested faster naming with contour distractors, but not toneme distractors, compared to controls. This suggests that the congruent context-specific rising contour continues to facilitate production even with delayed presentation. This would suggest that while activation of the tone category may be fleeting, similarity in the actual acoustic-

phonetic contour continues to facilitate production at later stages during overt production. Presumably this reflects activation of an acoustic and/or articulatory target in preparation for speech.

In summary, the finding of both contour and toneme priming effects in Experiment 1 indicates activation of multiple levels of representation during T3 sandhi word production. This raises the question of whether the results are due to automatic, lexical processes or to articulation preparation. It is possible that lexical processing of T3 sandhi words involves only their abstract form, but that the context-specific instantiation is only generated for overt speech. If this is the case, visually presented T3 sandhi words that are not overtly produced should lead to activation of the Tone 3 category only. Experiment 2 addresses this question.

2.3 Experiment 2 Naming of contour and toneme pictures with visually presented Tone 3 sandhi distractors

Method

Experiment 2 reversed the distractor and target conditions of Experiment 1, such that distractors were T3 sandhi words or controls and targets were pictures with contour (Tone 2 initial syllable) and toneme (Tone 3 initial syllable) bisyllabic names. If context-specific representations are activated only during speech preparation, then toneme targets, but not contour targets, should see facilitation from T3 sandhi compared to control distractors. If contour picture naming is quicker with T3 sandhi compared to control distractors, this suggests automatic activation of an instantiated representation of the context-specific T3 sandhi allotone, even when it is not overtly produced.

Participants Thirty native Beijing Mandarin speakers (24 female; mean age 22.7), students at universities within Haidian district of Beijing, were paid for their participation. None of them had participated in Experiment 1. All participants and their parents were born and raised in Beijing, except for four participants who had one parent from another Northern Mandarin-speaking province, and two participants whose parents were each from (different) Northern Mandarin speaking provinces.

Stimuli The experiment design and sample stimuli are shown in Table 2.5. Targets were 48 bisyllabic pictures; 24 with initial Tone 2 syllable (contour condition) and 24 with initial Tone 3 names (toneme condition). Distractors were bisyllabic T3 sandhi or control (Tone 1 or 4)

words that shared the same initial segmental syllable. T3 sandhi and control distractors were matched for word frequency, first character frequency, second character frequency, whole word stroke number, first character stroke number and second character stroke number. Mean bigram frequency of sandhi stimuli was 6.44 MI (SD = 3.7). There was no orthographic overlap or semantic relatedness between prime and target. An additional 48 picture-distractor pairs were used as filler trials.

Table 2.5: Experiment design and sample stimuli for Experiment 2

	Target condition		Distractor condition	
	Contour	**Toneme**	**Sandhi**	**Control**
Tone category	Tone2+ToneX	Tone3+ToneX	Tone3+Tone3	Tone1/4+ToneX
Tonal contour	rising	low	rising	other (high or falling)
Example	bi2kong3 鼻孔	bi3ji4 笔记	bi3shou3 比首bi4zhi4	币值

Design Experiment 2 consisted of a factorial 2 x 2 x 2 within-participants design. Experimental factors were Target type (Contour versus Toneme), Distractor type (T3 sandhi versus Control) and SOA (0 ms versus 83 ms). The experiment consisted of 384 trials, divided into six blocks of 64 trials, with breaks between the blocks. Each target word was presented four times (once in each distractor condition for each SOA). Other aspects of the design were the same as Experiment 1.

Procedure The procedure for Experiment 2 was identical to Experiment 1.

Results

Analysis was conducted on the 5,589 data points remaining after removal of stutters, errors, false starts (2%) and voice key errors (0.9%). Mean reaction times are shown in Figure 2.4. Forward and backward algorithms were used to establish fixed and random effects structure, and converged on the same final model. In the backward algorithm, the maximal model with random slopes for Distractor Type, SOA and Trial reached convergence, but the log likelihood ratio revealed that the random slope for Distractor Type was not significant, so it was removed.

The summary of results for the LME model for Experiment 2 is shown in Table 2.6. The control distractor condition at SOA = 0 ms lies on the intercept. The model was improved by a main effect of Trial, reflecting an overall increase in response time over the course of the experiment.

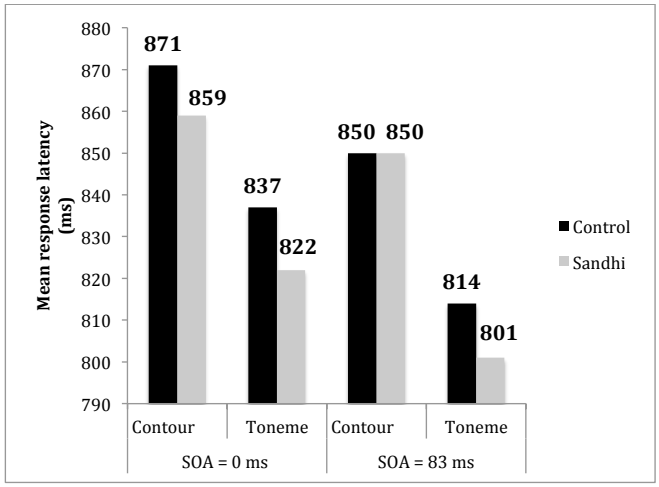


Figure 2.4: Mean reaction times (ms) per target type (contour versus toneme), distractor type (sandhi versus control) and stimulus onset asynchrony (SOA; 0 ms versus 83 ms) in Experiment 2

Responses were faster with a delayed distractor (SOA = 83 ms) than with simultaneous presentation, replicating the findings in Experiment 1. More interestingly for the present study, the model reveals a main effect of Distractor Type, with log naming latencies significantly shorter for T3 sandhi distractors, compared to control distractors. No improvement of the model was found with either a main effect of Target Type ($p = .12$), nor its interaction with Distractor Type ($p = .28$). The main effect of bigram frequency was not significant, and there were no significant interactions, so it was not included in the model. With a random slope for SOA, the fixed effect for SOA was no longer significant, but since the random effect term was significant, the fixed effect was retained in the model.

The main effect of Distractor Type in absence of an interaction between Distractor type and Target type or SOA suggests that the T3 sandhi distractors facilitated both target types at both SOAs. However, as with Experiment 1, between- condition differences in the numerical means suggest differential effects for the two target types, particularly at the late SOA. In order to further investigate these numerical differences, we split the data and modelled each target type separately. The model summary for Toneme targets is shown in Table 2.7. The model contained very similar predictors as in the combined data set. Model fit was improved by main effects of Trial ($p = 0$), SOA ($p = 0$) and

Table 2.6: Results summary Experiment 2: coefficient estimates, standard errors (SE) and t-values for all significant predictors in the log-transformed picture naming latencies with Tone 3 sandhi versus control distractors

Predictor	Coefficient estimate	SE	t
(Intercept)	6.6951	0.0230	291.49
Distractor Type Sandhi	-0.0127	0.0048	-2.67
SOA 83	-0.0221	0.0151	-1.47
Trial	0.0006	0.0001	5.17

Distractor type ($p < .01$), but no interactions. With a random slope for SOA, the t-value for the fixed effect was only marginally significant (see Table 2.7), but since SOA significantly improved model fit and because the random slope term was significant, the fixed effect was retained in the model. The model confirms for Toneme targets the findings of the full model, namely that presentation of T3 sandhi distractor words facilitates production of picture names with the same tone category.

Table 2.7: Results summary Experiment 2 Toneme targets: coefficient estimates, standard errors (SE) and t-values for all significant predictors in the log-transformed picture naming latencies with Tone 3 sandhi versus control distractors

Predictor	Coefficient estimate	SE	t
(Intercept)	6.6854	0.0286	233.99
Distractor Type Sandhi	-0.0224	0.0068	-3.29
SOA 83	-0.0293	0.0154	-1.90
Trial	0.0005	0.0001	4.41

The model summary for Contour targets is shown in Table 2.8. The only significant fixed effects were Trial ($p < .001$) and SOA ($p < .02$), which did not interact ($p = .42$). Distractor Type did not significantly improve the model. Model fit was further improved by a random slope for SOA ($p = 0$).

Although the model shows no effect of Distractor Type for the contour target only data with both SOAs included, there was a substantial difference in mean response times between T3 sandhi and control distractors at SOA = 0 ms (Figure 2.4). Because neither the interaction with SOA nor the interaction with Target Type was significant, it is

unclear whether this numerical difference is due to facilitation in processing from contour overlap or simply due to random variation. In order to further investigate this issue we ran a Bayesian analysis on the SOA = 0 ms data. If it is found that both target types are facilitated by T3 sandhi, compared to controls (that is, that there is no interaction between Distractor type and Target type) this would provide evidence that the contour is activated in visual word processing. Each of the predictors included in the LME model were individually tested in the Bayesian model. Substantial support was found for including a predictor of Trial, compared the baseline intercept (BF = 10.5)[6]. Adding a predictor for Distractor Type further improved the model (BF = 42.6), but neither Target Type (BF = .08) nor a Target Type:Distractor Type interaction (BF = .08) were supported. The absence of an interaction indicates that both target types were facilitated by T3 sandhi, compared to control distractors, providing further evidence for activation of the context-specific rising contour in visual processing of T3 sandhi words.

Table 2.8: Results summary Experiment 2 Contour targets: coefficient estimates, standard errors (SE) and t-values for all significant predictors in the log-transformed picture naming latencies with Tone 3 sandhi versus control distractors

Predictor	Coefficient estimate	SE	t
(Intercept)	6.6073	0.0276	242.86
SOA 83	-0.0163	0.0160	-1.02
Trial	0.0006	0.0001	5.30

Discussion

Experiment 2 investigated whether the findings from Experiment 1, namely that overt production of speech variants involves multilevel processing, can be extended to visual processing of written speech variants. There was a main effect of Trial, reflecting an increase in reaction times over the course of the experiment. More importantly, there was a main effect of Distractor type, indicating faster picture naming with T3 sandhi distractors, compared to control distractors. Although the interaction between Distractor Type and Target type was not significant, mean reaction times (Figure 2.4) suggested differences in the amount of fa-

[6] A Bayes Factor (BF) of 3 or more shows a substantial preference for the model over the alternative. A BF of less than 1 indicates lack of evidence or a preference for the alternative model.

cilitation for Toneme and Contour targets. We therefore split the data and analysed each Target Type separately. A robust effect of Distractor Type remained for the Toneme targets, indicating activation of the tone category during visual presentation of T3 sandhi words.

For the Contour targets, Distractor Type did not improve model fit with both SOAs in the model. However, a number of factors pointed to a facilitatory effect of contour. Firstly, there was a substantial difference in mean response times at SOA = 0 ms (Figure 2.4). In addition, in the full model containing both target types, the effect of Distractor type was significant, and there was no statistical support for an interaction with Target Type, suggesting that Distractor Type plays a role for both Toneme and Contour targets. Therefore, a Bayesian analysis was run in order to investigate whether the main effect of Distractor Type in the full model and the shorter reaction times at SOA = 0 ms can indeed be attributed to contour facilitation. The model showed a preference for including Distractor Type, but not Target Type or the interaction, confirming that the context-specific contour is activated during visual processing of T3 sandhi words.

2.4 General discussion

Two picture-word interference experiments provide new evidence on the nature of the phonological processing during speech production and visual processing of words that are not overtly produced. In particular, they address the question of whether allophonic variation is processed at the higher level of the phonemic category or at the lower, subphonemic level of the context-specific variant. In Experiment 1, during overt production of T3 sandhi picture names, significantly shorter naming latencies were found when distractor and target picture matched in tone category, but had different overt realisations (toneme condition), and when target and distractor matched in overt realisation, but mismatched in tone category (contour condition). This demonstrates that production of allophonic variants of Mandarin tones involves multilevel phonological processing: both the tone category and the context-specific variant are activated. The time course of activation was further investigated by splitting the data by SOA. When target and distractor were presented simultaneously (SOA = 0 ms), there was facilitation from both contour and toneme distractors, compared to controls, indicating early activation of both the tone category and the context-specific contour. With delayed presentation of the distractor (SOA = 83 ms), only the contour distractor showed significant effects; the toneme distractor no longer facilitated naming latencies. This can be explained if the overt realisation contour remains activated for longer than the tone category.

An alternative explanation is that, while both the contour and the category remain activated, as the task shifts from lexical retrieval to articulation preparation, only the articulatory/acoustic congruency benefits production. The present results do not allow us to tease apart these two possibilities.

In Experiment 2, target and distractor conditions were reversed to investigate whether the multilevel processing of allophonic variants found in Experiment 1 could be extended to visual processing of ignored distractor words. The model revealed a significant effect of Distractor Type that did not interact with Target Type, suggesting that both the tone category and the context-specific variant were activated. However, mean response times indicated differential effects for the two target types. Therefore, the data were split and analysed separately for the two target types. The model for toneme targets confirmed the results from the full model: T3 sandhi distractor words facilitated naming of toneme (low Tone 3) pictures, compared to control distractors. This demonstrates that automatic visual processing of allophonic variants in ignored distractor words involves processing of the tone category.

The tone contour was also found to be activated during visual processing of T3 sandhi words. Interestingly, there seem to be differences in the time course compared to toneme targets and compared to overt speech production. When run with data from only the contour targets, but with both SOAs included, the model did not show a significant effect of Distractor Type. In addition, at SOA = 83 ms, mean reaction times (Figure 2.4) were identical in the two distractor conditions. However, with simultaneous presentation of target and distractor (SOA = 0 ms), naming latencies were substantially shorter with congruent T3 sandhi distractors, compared to control distractors. We speculated that there may be an early facilitation effect from contour overlap, but that the present experiment had insufficient power to capture it in the reduced data set, due to the absence of facilitation at SOA = 83 ms. We investigated this with a Bayesian model, which showed support for facilitation of T3 sandhi distractors on both target types, indicating activation of both the tone category and the tone contour at SOA = 0 ms during visual processing of ignored distractor words.

Overall, the present results point to a different pattern of effects for overt production compared to ignored distractor words. When target and distractor were presented simultaneously, there was facilitation in both the contour and toneme conditions, regardless of whether the allophonic variants were overtly produced or processed visually. However, when presentation of the distractor was delayed, there was an effect of contour congruency only with overt production (Experiment 1) and an effect of the toneme only with visual processing (Experiment 2).

Taken together, the present results provide evidence for automatic

activation of both categorical and context-specific, instantiated representations during both overt production and visual processing of T3 sandhi words. This is consistent with previous studies that have found both abstract and fine-grained processing of segments (McLennan et al., 2003; McLennan, Luce & Charles-Luce, 2005) and tone (Mitterer et al., 2011) in speech perception, as well as segmental processing in speech production (McLennan et al., 2003, 2005; Nielsen, 2011). The present study extends the evidence for multilevel processing of speech variants to lexical tone production (see also Y. Chen, Shen & Schiller, 2011) and visual processing of Chinese characters. The results also provide evidence of differences in the time course of processing of the two levels during overt production, compared to when words are visually presented and not overtly produced.

The toneme and contour priming effects seem to reflect two separate processes corresponding to different levels of representation. The toneme effect addresses the question of whether processing of allophonic variants activates a category-level representation. Facilitation in the toneme conditions must have occurred at the tone category level because the actual realisation of the T3 sandhi targets and distractors (rising contour) is unrelated to the toneme targets and distractors (low contour). During overt production of the T3 sandhi variants (Experiment 1), t-values for the SOA = 83 ms model suggest continued facilitation from contour distractors, but not toneme distractors. In Experiment 2, the reverse pattern seems to emerge. The t-values for SOA = 83 ms indicate that the toneme targets are still facilitated by T3 sandhi distractors, while the contour targets are not.

The present results have interesting implications for the debate about the role of statistical distributions in speech category acquisition and processing. Although a growing body of research demonstrates that statistical information about acoustic cues plays an important role in first and second language acquisition and speech perception (Escudero et al., 2011; Gulian et al., 2007; Maye & Gerken, 2000; Maye et al., 2002; Wanrooij et al., 2013), there is also substantial between-category overlap in distributions. In the case of T3 sandhi and Tone 2, the overlap is almost total. Accounts based purely on statistical distributions of acoustic cues would have difficulty explaining the forming of separate categories in such cases where acoustic information from two speech categories is very similar. Recent evidence from computational models also suggests that acoustic distributional information alone may not be sufficient for phonetic category acquisition (Feldman, Myers, White, Griffiths & Morgan, 2011; McMurray et al., 2009). Feldman et al. (2011) suggest that phonetic category development occurs as part of extracting meaning from language, through association of phonetic distributions with lexical items. Association with the meanings (and orthography)

of the respective characters can explain how different speech categories, such as T2 and T3 sandhi, can form separate representations at the category level, despite very similar acoustic distributions.

The contour priming effect, on the other hand, seems to reflect a different sort of representation. Contour facilitation must have occurred due to acoustic and/or articulatory similarity. This entails representations that are both context-specific and instantiated. They are context-specific because the variant occurs only in the particular phonetic environment when two or more Tone 3 characters occur directly one after the other. They are instantiated because, since target and distractor are unrelated at the category level, the effect must occur due to similar physical properties, that is, acoustic and/or motor-movement similarity. The idea of an instantiated internal representation is consistent with models that posit involvement of the sensori-motor system in speech production and studies showing that auditory and somatosensory feedback are utilised in guiding and adjusting speech production (Davis & Johnsrude, 2007; Guenther, Ghosh & Tourville, 2006; Guenther & Vladusich, 2009; Houde & Jordan, 1998; Jones & Munhall, 2002; Liberman & Whalen, 2000; Purcell & Munhall, 2006).

The contour effect in Experiment 1 opens up new questions about the processing of the contour itself. For example, is it stored lexically? In fact, T3 sandhi occurs not only in words, but also across word boundaries. In the present study, there was no effect of bigram frequency. However, since we were interested in tone processing in sandhi *words*, the bigram frequency range in the selected stimuli was restricted. Recall that bigram frequency is a measure of how often two characters occur together, and is used here as a measure of lexicality, since word boundaries are not explicit in Chinese. If similar effects were found even with very low bigram frequency, this would rule out the possibility that the context-specific contour is stored only in lexical items. Two further possibilities are that the contour is stored as part of the morpheme and, alternatively, that it is stored as part of the tone category. Future research could disentangle these possibilities. If the sandhi contour is processed as part of a purely abstract Tone 3 category, effects should be equal for all morphemes. However, if the contour is processed by exemplar, morphemes which rarely occur in sandhi contexts should see significant attenuation of the contour effect relative to morphemes that frequently occur in sandhi contexts.

Conclusion

In conclusion, the present study provides new insights into phonological processing during speech production and visual word processing. Experiments 1 and 2 showed toneme and contour priming effects, indicat-

ing multiple levels of representation in production and visual processing of Mandarin tones. The toneme effect indicates activation of the tone category representation, which may be formed through processing of regularities in input data distributions. The contour effect suggests a context-specific and instantiated representation of the actual pitch contour. This can be explained in terms of a somato-motor/auditory target by which speakers gauge production accuracy.

Chapter 3

Early negativity reveals rapid sub-phonemic processing during reading aloud

A version of this chapter is submitted for publication as:
Nixon, J. S., Timmer, K., Linke, K., Schiller, N. O. & Chen, Y. (submitted). Early negativity reveals rapid sub-phonemic processing during reading aloud.

Abstract Little is known about whether or when sub-phonemic features, such as voicing and place of articulation, are processed during reading aloud. Event-related potentials were recorded while participants named Dutch words preceded by non-word primes. Primes were identical to targets, except for the final letter, which matched in voicing (voice-match condition) or place of articulation (place-match condition), or mismatched in both voicing and place (controls). Responses were faster in the voice-match condition, but not place-match condition, compared to controls. Consistent with behavioural results, EEG measures revealed reduced negativity in the voice-match condition, but not place-match condition, compared to controls. The effect occurred in the early, 25-75 ms time window. These combined electrophysiological and behavioural results indicate that sub-phonemic information is processed early in reading aloud. In addition, the results also have implications for the processing of allophonic speech variants. In Dutch, voiced stops are 'devoiced' in final position. That is, when voiced stops occur at the end of a word, the actual overt realisation is similar to voiceless stops. The present voice-congruency effect shows that despite similarities in the realisation, (de)voiced and voiceless stops are processed as separate categories.

Keywords sub-phonemic features, allophonic variation, ERPs, reading aloud

3.1 Introduction

Sub-phonemic features While traditional views of language processing took phonemes to be the basic units of sound, recent evidence shows that speech production involves multi-level phonological processing, with both category-level and sub-phonemic information playing a role (McLennan et al., 2005; Nixon et al., 2014). Very little is known about sub-phonemic processing in reading aloud. Whether sub-phonemic features, such as voicing and place of articulation, are processed is not yet well understood. There is abundant evidence that a range of lexical processes are facilitated by activation of sub-*lexical* phonological components that make up a word. These phonological components have generally been measured in terms of phonemes. For example, the masked onset priming effect (MOPE) shows that reading aloud is faster when targets are preceded by primes that share the same onset phonemes, compared to those whose onset phonemes differ (Kinoshita, 2000; Kinoshita & Woollams, 2002; B. Mousikou, Roon & Rastle, 2014; Timmer & Schiller, 2012; Timmer, Vahid-Gharavi & Schiller, 2012; Schiller, 2004) (see also Ferrand & Grainger, 1992, 1993, 1994)

However, recent evidence suggests that viewing sub-lexical phonology as consisting of permanent, unchanging, abstract phoneme categories may be too simplistic. Language processing has been shown to be influenced by a broad range of types of sub-*phonemic* detail (Clayards et al., 2008; Ju & Luce, 2006; Maye et al., 2008; McMurray et al., 2009; Mitterer et al., 2011; Newman et al., 2001; Nixon et al., 2014; Trude & Brown-Schmidt, 2012). In addition, some models of speech production include representations of features. For example, Dell (1986) proposes that following phoneme selection, activation spreads from the selected phoneme to its constituent features. An alternative suggestion is that these sub-phonemic representations consist of articulatory gestures (e.g. Goldstein et al., 2007) (e.g. Goldstein, Pouplier, Chen, Saltzman & Byrd, 2007).

While Roelofs (1999) failed to find evidence of facilitation from feature overlap during speech production, such phonetic features (or articulatory gestures) have more recently been shown to play a role in speech perception and silent reading (Ashby, Sanders & Kingston, 2009; Chládková, 2014). Using event-related potentials (ERPs), Ashby et al. (2009) conducted a silent reading task, in which targets ended in a voiced or voiceless consonant coda (e.g. *fat* or *fad*). Targets were preceded by masked non-word primes that were identical to targets, except for the coda, which either matched or mismatched in voicing (e.g. *faz* - FAD; *faz* - FAT). ERPs showed less negativity in the 80-120 ms time window when prime and target matched in voicing, compared to the mismatch

control condition.

As Ashby and colleagues point out, there are two possible reasons for their voice-congruency effects. In English, there are duration differences in vowels preceding voiced and voiceless final stops. For example, the 'a' in 'fad' is longer than that in 'fat'. Therefore, although their voice-congruency effect indicates sub-phonemic processing of some sort, it may reflect processing of voicing in the final consonant, or processing of vowel duration. In addition, some of the prime-target stimuli pairs differed in both voicing and manner of articulation (e.g. *faz* is a voiced fricative, while *FAT* is a voiceless stop). The question of whether phonological processing involves processing of voicing and place of articulation information could be further elucidated by teasing apart consonant voicing from vowel duration, as well as controlling for manner of articulation.

In addition, one recent study has investigated processing of feature information in English non-word reading aloud. B. Mousikou et al. (2014) found that non-word naming latencies were shorter when the onset of primes overlapped with target onset phonemes (e.g. bez-BAF) or target onset place and manner of articulation (e.g. piz-BAF), compared to the unrelated condition (suz-BAF).

Allophonic variation A second area of sub-phonemic processing that is not well understood concerns allophonic variation. In certain cases, the realisation of phonemes may vary reliably with phonetic context, without affecting word meaning. These variants are called *allophones*. For example, word-initial English voiceless stops /t/, /p/ and /k/ are normally aspirated — that is, there is a delay between the release of the stop closure and the onset of the vowel. But following /s/ (e.g. the 't' in 'stop') they are unaspirated. When a phoneme category has more than one output pattern (i.e. target distribution, or allophone), are the two or more distinct outputs processed as a single category or as separate categories?

Sometimes, allophonic variation can lead to ambiguity between two otherwise distinct categories. In Dutch word-initial position, the voiceless stop /t/ is distinguished from its voiced counterpart /d/ primarily by *voice onset time* (VOT: 20 ms for voiceless; -70 ms for voiced; Lisker & Abramson, 1964; Slis & Cohen, 1969; Meijers, 1971). In word-final position, however, the VOT values of voiced stops are comparable with voiceless stops, a phenomenon known as *final devoicing*. This leads to ambiguity in speech. For example, the words *hout* ('wood') and *houd* ('to hold') are homophones in Dutch. Baumann (1995)found that Dutch listeners performed at chance level in distinguishing devoiced-voiceless minimal pairs.

Processing of allophonic variants has been investigated with respect

to Mandarin tone production (Nixon et al., 2014). In Beijing Mandarin, the third tone is usually produced with a low contour, but preceding another third tone, it is realised with a rising contour, similar to the second tone. Participants named pictures that were superimposed with distractor words that matched the actual realisation of the tonal contour, but mismatched the tone category (the contour condition), matched the tone category, but mismatched in surface realisation (category condition) or mismatched both. Results showed reduced latencies in both the contour and category conditions, compared to controls. This finding suggests simultaneous activation of multiple levels of phonological representation, at least in Mandarin tone production.

The present study In the present study, the masked priming paradigm (Forster & Davis, 1991; Kinoshita, 2000; Kinoshita & Woollams, 2002; Timmer & Schiller, 2012) was employed to address two questions relating to sub-phonemic processing. Firstly, the study examines whether and when sub-phonemic features are processed in Dutch reading aloud. Secondly, it asks what kind of representation is activated when one phoneme category has two possible variants. ERPs were recorded as participants read aloud Dutch words from a computer screen. Each word was preceded by a brief presentation of a masked non-word prime. Primes were identical to targets, except for the final letter, which either matched in voicing (voice condition) or place of articulation (place condition) or mismatched both voicing and place (control condition). While most previous studies have investigated onsets or whole syllables, in the present study we chose to manipulate the word-final consonants in order to replicate as closely as possible the Ashby et al. study described above.

With respect to the first question, if features are processed, we expect this to be reflected in the reaction times as faster responses in feature-congruent (i.e. voice and place) conditions, compared to controls. In ERP measures, we expect reduced negativity early in processing in feature-congruent conditions, compared to controls (Ashby et al., 2009).

The second question investigates whether voiceless and 'devoiced' final stops are processed as separate categories. Recall that, at the articulatory level, due to final devoicing, both voiceless and devoiced stops are 'voiceless'. If devoiced variants are processed at a purely articulatory level (e.g. Goldstein et al., 2007), we would not expect differences between match and mismatch conditions for voice. However, at the category level they differ in voicing. If a voicing congruency effect is found, this suggests processing of a contrastive feature category level (e.g. Dell, 1986), despite similar articulation, in Dutch reading aloud.

3.2 Method

Participants Twenty-seven native Dutch speakers were paid for their participation (19 female). Mean age was 23.4 years (s.d. 4.99). All participants signed an informed consent form, had normal or corrected-to-normal vision and reported no reading difficulties. Participants were excluded from analysis if fewer than 75% of trials were left due to noisy EEG data (3). A further participant was removed due to failure of the voice key trigger. Analysis was conducted on the 23 remaining participants (17 female; mean age 23.4 years; s.d. 4.94).

Materials Target words were selected from the CELEX database (Baayen, Piepenbrock & van Rijn, 1993). Critical targets consisted of 39 three- to four-letter Dutch nouns ending in a voiced ('d' or 'b') or voiceless stop consonant ('t' or 'p'). Sample stimuli are shown in Table 3.1. Word structure was either CVC or CCVC. No targets contained word-final consonant clusters. A further 39 words were used as fillers to add variety and make the design less obvious to participants. Each target was preceded by a non-word prime that was identical to the target, except for the final letter (see Table 3.1). The final letter either matched in voicing (voice condition) or place of articulation (place condition), or mismatched in both voicing and place (controls). Primes were matched for phonological neighbourhood frequency across voice (3.29), place (3.35) and control conditions (3.24) using Clearpond (Marian, Bartolotti, Chabal & Shook, 2012).

Table 3.1: Experiment design and sample stimuli

	Prime condition		
Target word	**place**	**voice**	**control**
HUID	huit	huib	huip

Design The experiment consisted of 234 trials, divided into three blocks of 78 trials, with breaks between the blocks. There were 39 trials per condition. Each target word was presented three times (once in each prime condition). Three prime word lists were constructed, with primes divided equally between conditions so as to control for order effects of particular items. All participants received all lists, and the order of presentation of the prime lists was counterbalanced across participants. All lists were pseudo-randomised for each participant. Each block was preceded by ten warm-up trials, which were excluded from analysis.

Procedure Participants were tested individually in a dimly lit, sound-proof room. They were instructed to read aloud the target words that appeared on the screen as quickly and accurately as possible. Participants were not aware of the masked primes. The trial procedure is shown in Figure 3.1. All stimuli were presented in black letters on a white background in the centre of the screen. Screen refresh rate was 60 Hz. Each trial began with a fixation cross with jittered presentation time (400-700 ms) to reduce time-induced expectancy waves. A forward-mask of five hash symbols ('#') followed for 500 ms, before presentation of the prime in lower case for 48 ms. A backward mask was presented for 17 ms to avoid continuation of visual processing due to imprinting on the retina. Finally, the target word was presented in upper case for a maximum of 2,000 ms or until the participant response, which triggered the voice key and caused the word to disappear. Prime and target were presented in lower and upper case, respectively, to avoid lower-level visual effects. The experimenter coded incorrect responses and voice key errors in a 1,400 ms interval before the beginning of the next trial.

3.3 Analysis and Results

Reaction time analysis Behavioural data were analysed with linear mixed effects (LME) modelling, using the lmer function of the lme4 package (Bates et al., 2013) see also (Baayen, 2008; Baayen et al., 2008) in R (R Development Core Team, 2013). Analysis was conducted on the 2,546 data points remaining after errors and voice key errors (<4%) were removed.

Electrophysiological recording and analysis The electroencephalogram (EEG) was recorded using 32 Ag/AgCl electrodes at the standard scalp sites of the extended international 10/20 system. A further six flat reference electrodes were placed above and below the left eye to record blinks, at the external canthi of each eye to record horizontal eye movements, and at the left and right mastoid, for offline re-referencing. The EEG signal was sampled at 512 Hz and off-line band filtered from 0.01 to 40 Hz. Epochs were computed from target onset to +500 ms with a -300 to -100 ms baseline. Ocular artefacts were (Gratton, Coles & Donchin, 1983) corrected using the algorithm. For non-ocular artefacts, trials with amplitudes below -200 μV, above +200 μV or trials which contained a voltage step of 100 μV or more within 200 ms were removed from analysis. A recent review of EEG research

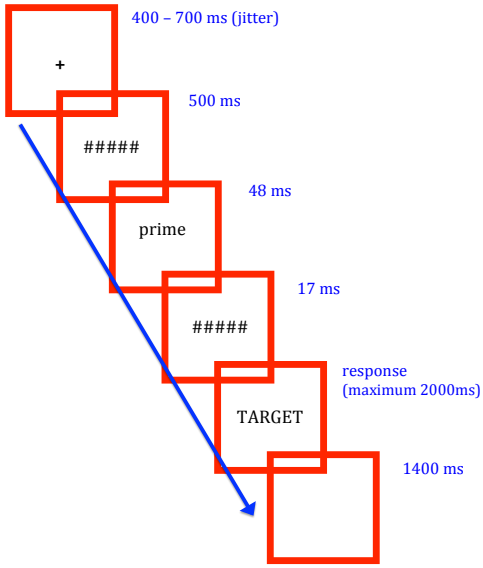


Figure 3.1: Trial procedure: a fixation cross is presented with a jittered duration, followed by a forward mask (500 ms), the non-word prime (48 ms), a backward mask (17 ms) and finally the target (until participant response or a maximum of 2000 ms). Errors are coded during a blank screen preceding the next trial.

with overt speech production shows that artefact-free brain responses can be measured up to at least 400 ms following presentation of the target stimulus (Ganushchak, Christoffels & Schiller, 2011). ERP grand averages were time-locked to the onset of the target-word and averaged across participants for the three conditions (voice, place and controls).

Reaction time results Table 3.2 shows the mean response times for each condition. The fixed effects for the LME model of response latencies is shown in Table 3.3. The model was built up from a baseline model with random intercepts for participants and target words (Baayen, 2008). Main effects of trial and condition and their interaction were added individually to the model and tested by comparing the log likelihood ratio to the simpler model. Only effects that

significantly improved the model fit were retained in the model. Next, random effects structure was tested. A random by-participants slope for Trial, but not Condition, was found to improve the model. The best-fit model included main effects of Trial and Condition, random intercepts for participants and target items and a by-participants random slope for Trial. The summary of the model reveals significantly[1] shorter response latencies for the voice condition, compared to controls. There was no difference in response times between the place and control conditions.

Table 3.2: Mean reaction times per prime condition

Prime condition	**control**	**place**	**voice**
Mean reaction time	529	530	520

Table 3.3: Results summary: coefficient estimates, standard errors (SE) and t-values for all significant predictors in the best-fit model of response latencies

	Coefficient estimate	**SE**	**t**
Intercept (Condition: control)	521.59	17.02	30.65
Condition:place	0.87	2.92	0.30
Condition:voicing	-8.35	2.91	-2.9
Trial	0.21	0.11	2.00

ERP measures ERPs were analysed with a repeated measures ANOVA, conducted in SPSS, with Localization (anterior: AF3, AF4, F3, F4, F7, F8, Fz vs. central: C3, C4, Cz, FC1, FC2, CP1, CP2 vs. posterior: P3, P4, P7, P8, PO3, PO4, Pz) and Condition (voice vs. place vs. control) as within-participants factors.

Time windows were selected based on visual inspection. Figure 3.2 shows the average amplitude over time for each condition in six electrodes. The 25-75 ms time window revealed a main effect of Condition ($F(2,44) = 3.30$, $MSe = 52.85$, $p < .05$) that did not interact with

[1] T-values below -2 or above 2 can be taken as significant at the 95% confidence level for sufficiently large (1,000 or more data points) data sets (see, for example, Baayen, 2008; Baayen et al., 2008; Baayen & Milin, 2010).

Localization ($F(4,88) = 1.20$, $MSe = 5.86$, ns). Planned comparisons of Condition revealed smaller negative amplitudes for the voice-match condition (3.01 μV; SE = 0.75) compared to the control condition (1.84 μV; SE = 0.69; $F(1,22) = 6.18$, $MSe = 655.07$, $p < .05$). In contrast, the place-match condition (2.29 μV; SE = 0.66) did not differ from the control condition (1.84 μV; SE = 0.69; $F < 1$).
The 100-150 and 175-250 ms time windows did not reveal a main effect of Condition ($F(2,44) = 1.84$, $MSe = 93.52$, ns; $F(2,44) = 1.52$, $MSe = 88.59$, ns, respectively) or an interaction with Localization ($F(4,88) = 1.02$, $MSe = 8.78$, ns; $F<1$, ns, respectively).

3.4 Discussion

The present study provides new behavioural and electrophysiological evidence on the nature of phonological processing in reading aloud. In the reaction times, match between prime and target in the sub-phonemic feature voice led to reduced response latencies, compared to the control condition. There was no effect of place-match, compared to the control condition. Consistent with the behavioural results, ERP measurements revealed decreased negativity across the entire scalp in the 25-75 ms time window for the voice-match condition, but not the place-match condition, compared the control condition.

Although this effect occurs early, by manipulating mask duration between experiments, Ashby et al. (2009) demonstrated that the voice-congruency effect is not tied to the N1. In two experiments, they used two different mask durations (100 ms in Experiment 1; 22 ms in Experiment 2). The timing of the N1 at FCz was delayed (148 ms after target onset) with the long mask duration, compared to with the short mask duration (100 ms after target onset). However, mask duration did not appear to affect the timing of the congruency effect, which appeared at around 80 ms in both experiments.

The present findings in Dutch are consistent with studies that have shown feature-level processing of voicing information in English. Ashby et al. (2009) found that voice congruency during silent reading in English led to reduced negativity beginning at 80 ms, compared to controls. Since voicing in English final consonants is confounded with preceding vowel length, it was not clear whether their voice-congruency effect resulted from sub-phonemic differences in the vowel or the consonant. In the present study, the use of Dutch final stops allowed us to tease apart (underlying) consonant voicing from duration of the preceding vowel.

These results provide further support for early processing of sub-phonemic feature information, and extend the results to Dutch reading

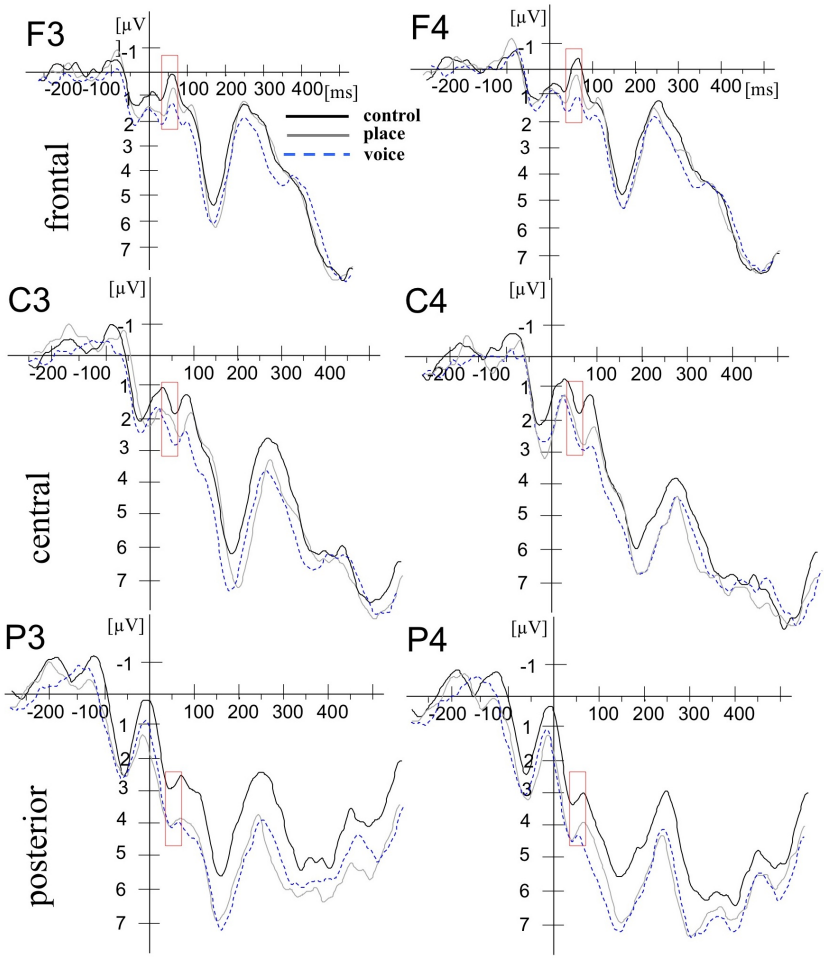


Figure 3.2: Average ERP responses for voice-match, place-match and control conditions.

aloud. This cannot be due to orthographic or phonemic processes, as the manipulation was identical for the critical and control conditions in terms of both letters and phonemes: each prime-target pair differed by exactly one phoneme and letter. Only when measured at the sub-phonemic feature level was there greater overlap in congruent prime-target pairs (voice and place conditions), compared to controls.[2]

Processing of voicing information has also been found in speech production studies. Nielsen (2011) found that imitation of extended VOT in production of words containing /p/ generalized not only to new words containing the same phoneme /p/, but also to other voiceless stops (Nielsen, 2011). There is also evidence that infants are able to learn and generalise voicing distinctions (Maye et al., 2002). When infants were presented with a novel category distinction in the form of a bimodal distribution of VOT for one place of articulation (e.g. alveolar), they were able to generalise the same VOT category distinction to a new place of articulation (e.g. velar).

The present results extend the evidence for voicing feature processing found in English (Ashby et al., 2009; Maye et al., 2002; B. Mousikou et al., 2014; Nielsen, 2011) to a new language and a new task. Moreover, the reduced early negativity found in English (Ashby et al., 2009) was replicated in Dutch, which does not confound voicing and vowel length, providing further evidence for the rapid processing of sub-phonemic feature information in consonants.

In addition to sub-phonemic feature processing, the voice-congruency effect sheds light on the processing of allophonic variation. Despite similarities in overt production, voiceless and (de)voiced stops seem to be processed as distinct categories. Previous studies have shown that production of allophonic variants of lexical tone involves processing of both the sound category and the context-specific realization (Nixon et al., 2014). The present results cannot speak to the processing of the context-specific allophone (i.e. the devoiced variant), as this question was not part of the experimental design. (Although, presumably, it is processed at some level in order to produce the actual devoiced realization found in speech). However, the present voice-congruency effect provides evidence that voiced and voiceless stops are processed as distinct categories. For example, for target *HUID* (voiced category, surface devoiced), the prime *loeb* (voiced category, surface devoiced) facilitated responses compared to *loet* (voiceless category). If devoiced and voiceless stops were

[2] Although one study (Warner, Jongman, Sereno & Kemps, 2004) was able to detect very minor acoustic differences (3.5 ms) in preceding vowel duration, they suggest this may be an orthographic effect that occurs as a result of careful articulation during list reading in production experiments. It is unlikely that such an effect occurs during subliminal processing. Moreover, even if such an effect did occur, it would not account for the size of the effect found in the response times reported here.

not processed as separate categories, then all conditions would match in voicing. For all critical stimuli, the overt realisation was voiceless ([t], [p]). Only at the category level was there a distinction between voiced (/d/, /b/) and voiceless stimuli (/t/, /p/). The reduced negativity and shorter reaction times in the voice-match condition suggest that the distinct voicing categories are processed, and are not collapsed to one (voiceless) category. This seems to provide support for a fairly abstract, contrastive level of representation for features (e.g. Chomsky & Halle, 1968; Dell, 1986). However, the results do not exclude the possibility that features are also processed at the articulatory level. (Nixon et al., 2014) showed that speech production and processing of visual words involve multi-level processing. The present results could also be explained if multi-level processing also occurs at the feature level. That is, reading aloud may involve processing of both a contrastive feature category (voiced-voiceless) and the context-specific articulatory gesture. More work is needed to verify this possibility.

In contrast to the voice-match condition, the place-match condition was not significantly different to controls. Although this is a null effect, in light of the positive result for voice-match, it is worth considering why there was no significant effect here. One possible explanation is that sub-phonemic features are not a fundamental unit of processing and that the present results reflect a language-specific effect, particular to Dutch. For example, it might be argued that the voice congruency effect may actually occur *due to* final devoicing, rather than in spite of it. However, this seems unlikely given previous findings of processing of voicing information in English, in which devoicing does not occur (Ashby et al., 2009; B. Mousikou et al., 2014; Nielsen, 2011)and of infants' ability to learn and generalise new VOT category distinctions (Maye et al., 2002). A better explanation is that VOT is processed because it is highly informative for distinguishing between phonetic categories. The finding of an effect for voice congruency but not place congruency may be because place is less informative.

One of the factors that might affect the degree of informativeness of a particular phonetic cue is the number of categories for which it is informative (Pajak, 2012). If a phonetic dimension (such as VOT) distinguishes several (pairs of) categories, that dimension may be more informative than a dimension that distinguishes only one or relatively few category pairs. In Dutch, VOT distinguishes between four pairs of native stops and fricatives (d-t, b-p, z-s, v-f), plus two more pairs when borrowed words are included (g-k, sh-zh). In the case of place of articulation, several place categories are used to distinguish only a relatively limited number of phonetic categories. There are a total of eight place categories (bilabial, labiodental, alveolar, post-alveolar (loanwords only), palatal, velar, uvular and glottal)) that produce ten contrasts.

In sum, the present study challenges previous assumptions in reading and speech production research that phonological processing simply involves activation of strings of phonemic units. Consistent effects in the ERP measures and response latencies in the voice-congruent condition, compared to controls, reveal processing of sub-phonemic voicing information in reading aloud. Moreover, this effect was found despite similarities in overt production of voiceless and (de)voiced stops in final position in Dutch. This suggests category-level, rather than purely surface-level phonetic processing, for the voicing dimension in Dutch.

An interesting possibility for future work would be to reexamine the results with a method that allows for analysis of individual experimental items, such as linear mixed regression or generalised additive modelling. This would allow us to include more fine-grained information about the individual items, such as target word frequency, bigram frequency of targets and primes, and so on. This may offer a deeper understanding of the underlying processes in the perception of the non-words in this study. Recent work on Dutch voiceless versus (de)voiced final stops (Ernestus & Baayen, 2003, 2004) suggests that processing of phonological information is sensitive to analogical relationships with other words in the language. In Dutch, formation of affixes, such as the past tense, are either voiceless (e.g. -*te*) or voiced (e.g. -*de*) depending on the underlying voicing in the stem-final consonant. Ernestus and Baayen (2004) showed that during perception of spoken non-words, where orthographic information is not available, participants' production voiced versus voiceless past tense forms for a particular non-word depended on the voicing in similar real Dutch words. Therefore there may be neighbourhood effects in the current data that we are unable to determine with the current analysis and which could be elucidated with a more accurate statistical method.

Chapter 4

Context constrains neural activity during speech variant processing: a non-linear model of ERP data

A version of this chapter is in preparation as:
Nixon, J. S., van Rij, Li, X. Q. & Chen, Y. (in preparation). Context constrains neural activity during speech variant processing: a non-linear model of ERP data

Abstract

The phonetic realisation of speech sounds depends on their context, yet most psycholinguistic theories do not account for such phonetic variation. The present study investigated how such phonetic variation is processed by measuring ERP amplitude during a reading aloud task with masked priming. In Beijing Mandarin, Tone 3 usually has a low contour, but preceding another Tone 3 syllable, it has a rising contour. All critical targets had a Tone 2 initial character, which also has a rising contour. In the Contour match condition, primes consisted of two Tone 3 characters (T3 + T3), so the (rising) contour matched the targets, even though the tone category was different. In the mismatch condition, the second character of the prime was another tone (T3 + TX), so prime and target differed in both contour and tone category. ERPs were analysed using Generalised Additive Mixed Modelling, a non-linear model with random effects for subjects and items. Models revealed a complex interaction between prime type and prime and target frequencies. In the mismatch condition, there was relatively little effect of the item frequencies. In the contour match condition, in contrast, there seemed to be a cross-over effect in the prime and target frequencies. When target frequency is relatively high and prime frequency is low, there is reduced negativity. However, when prime and target frequency are both high or both low, there is increased negativity, which suggests competition between prime and target. This suggests that when tonal contour no longer discriminates between prime and target (i.e. in the contour match condition), the a priori probabilities of the prime and target come into play. The conflict that arises between prime and target requires increased processing effort as reflected in greater ERP amplitudes. This difference in the pattern of effects between the Contour match and mismatch primes provides evidence for top-down effects of context on phonological processing of masked primes during reading aloud.

4.1 Introduction

Sub-phonemic processing

Phonetic variation is a fundamental property of speech. Regularities in speech make it possible for speakers to form speech sound categories that distinguish between word meanings, such as between the words 'pin' and 'bin'. In alphabetic languages, sub-lexical processing of speech sounds has been posited to involve activation of strings of phonemes (Dell, 1986, 1988; Foss & Swinney, 1973; Indefrey & Levelt, 2004; W. J. M. Levelt, 2001; W. J. M. Levelt et al., 1999; McClelland & Elman, 1986; Meyer, 1990, 1991; Roelofs, 1999). However, the actual acoustic form of these phoneme categories is far from uniform. It is likely that speakers' categorisation of sounds reflects the way they are represented in the orthography. For example, in words like 'spin', where the first sound is /s/, the second sound is considered to belong to the same sound category /p/ as in the word 'pin'. But acoustically, the voice onset time falls between the /p/ of 'pin' and /b/ of 'bin'. How this kind of 'within-category' variation is processed is not yet well understood. Many current psycholinguistic models, particularly models of speech production and reading aloud, fail to account for processing of context-dependent phonetic variation.

Context effects are well attested during speech perception. Perception of an incoming speech signal is influenced by the semantic context. For instance, processing of meaningless syllables is more similar to real-word processing when semantic information is available from a surrounding sentence (Bonte, Parviainen, Hytönen & Salmelin, 2006). In speech perception, recalibration is a process where the acoustic cues used to discriminate established, native phonemes are shifted after training with contextual or indexical cues. For example, ambiguous acoustic cues are biased towards one or other interpretation by a lexical or visual context or a specific speaker (Dahan, Drucker & Scarborough, 2008; Kraljic & Samuel, 2005, 2007, 2011; Kraljic, Samuel & Brennan, 2008; McQueen et al., 2006; Norris, McQueen & Cutler, 2003). Reinisch, Wozny, Mitterer and Holt (2014) tested whether generalisation of visually guided recalibration of cues contained in meaningless bisyllables depended on phonetic context. Interestingly, they found that recalibration did not occur when the surrounding vowels differed between training and test, suggesting a context-specific utilisation of acoustic cues to distinguish speech contrasts during perception.

Some recent research has begun to address the question of context effects in speech production. Goldrick and Larson (2008) found that speech errors were sensitive to statistical frequencies of syllable positions of features. Errors were less likely to result in fricatives being produced in syllable-final position when 75% of the fricatives in

the training were syllable-initial. This shows phoneme processing includes information about the surrounding phonetic context. There is also evidence that overt production and visual processing of speech variants involves multilevel phonological processing. Using the picture-word interference paradigm, Nixon et al. (2014) investigated whether processing of allophonic tonal variants (sandhi words) in Beijing Mandarin involved activation of the tone category (toneme), the context-specific variant (allotone) or both. Experiment 1 revealed that during overt production of sandhi words, both the tone category and the context-specific variant were activated. In Experiment 2, sandhi words were not overtly produced, but were instead visually presented as distractor words superimposed on target pictures. This led to a different time course of effects, compared to overt production. While facilitation from the context-specific variant was stable across both simultaneous and delayed presentation conditions during overt production (Experiment 1), when processed visually as ignored distractor words (Experiment 2) the context-specific variant no longer affected reaction times when presentation of the distractor was delayed. The tone category, in contrast, remained stable for both presentation conditions during visual processing (Experiment 2) but was less robust with delayed presentation during overt production (Experiment 1). This suggests the possibility that there may be differences in processing of sub-phonemic information depending on the task and modality. In particular, during overt production, both the tone category and the context-specific variant are activated early. During visual processing of tonal variants, on the other hand, there seems to initially be activation of the tone category, while it takes time to activate the context-specific variant.

One of the assumptions of priming studies is that activation of overlapping phonological units leads to facilitation of words containing those same units. Few studies have investigated whether priming occurs across speech categories. That is, do similarities in the physical acoustic properties of speech facilitate processing, even if there is no category overlap? Most investigations of sub-lexical phonology have taken the phoneme to be the basic unit of sound. For example, similarity between prime and target is usually measured in terms of phoneme overlap. Recently, a number of studies have provided evidence for sub-phonemic processing of speech sounds (Clayards et al., 2008; Ju & Luce, 2006; McMurray et al., 2009; Mitterer et al., 2011; Newman et al., 2001; Nixon et al., 2014; Nixon, van Rij, Mok, Baayen & Chen, submitted; Nixon, Timmer, Linke, Schiller & Chen, submitted). The contour effect found in Nixon et al. (2014) suggests that acoustic similarity in visual words presented simultaneously with targets is sufficient to facilitate tone processing during speech production, even when there is no category overlap. Whether acoustic similarity in masked primes also facilitates tone processing dur-

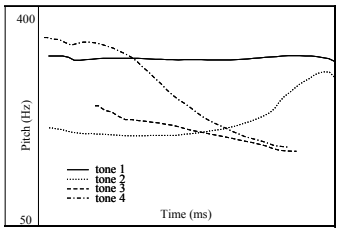


Figure 4.1: Pitch contours of the four tones of Beijing Mandarin

ing reading aloud is not yet known.

Lexical tone processing

Despite the fact that well over half of the world's languages use tone to distinguish between word meanings, a survey of the literature reveals a mere handful of studies on the processing of tone in speech production. As far as we are aware, the present study is the only study to investigate tone processing in reading aloud. Beijing Mandarin has four lexical tones (tonemes). The ptich contours of each of the four tones were shown in Chaper 2, Figure 2.2. They are reproduced here for convenience (Figure 4.1) Characters that have the same segmental syllable can be distinguished by this inherent pitch contour, such as yu2[1] (鱼, 'fish') versus yu3 (雨, 'rain'). In connected speech, Tone 3 (T3) has at least two variants (allotones)[2]. The canonical realisation is the low contour, but preceding another T3 syllable, T3 is realized with a rising contour. This allophonic variant of T3 is known as third tone sandhi (hereinafter, 'T3 sandhi'). Tone sandhi refers to the phenomenon whereby the acoustic realization of a tone is influenced by a neighbouring tone in a particular environment. Importantly for the present study, the contour of T3 sandhi is very similar to another tone, Tone 2. Figure 4.2, reproduced here from Chapter 2, shows the tonal contours of Tone 2, T3 sandhi and the canonical low Tone 3.

The present study has two main objectives. Firstly, it investigates the effects of acoustic contour on reading aloud when prime and target belong to different speech categories. Secondly, it investigates whether the tonal context in which a character occurs affects the relative activation levels of the two alternative tonal variants. One possibility is

[1] Mandarin tones are referred to using a number system (Tones 1 to 4). Here, the numeral following the syllable represents the tone number, in this case, Tone 2.

[2] The third tone is sometimes described as 'falling-rising'. However, the rising part of the contour is optional and does not usually occur when there is a following syllable. In addition, the gradient of the fall is very shallow. For these reasons and for simplicity, we refer to the contour as 'low'

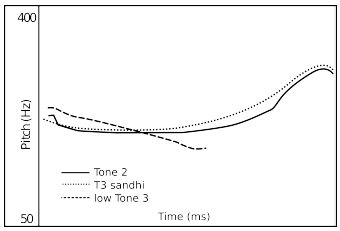


Figure 4.2: Pitch contours of Tone 2, Tone 3 sandhi and canonical low Tone 3

that processing occurs at the speech category level, until the speech preparation stage, when the context-specific articulatory information is activated. A second possibility is that for words containing allophonic variants all variants are activated, regardless of the context in which they occur. Finally, a third possibility is that top-down information available from the surrounding phonetic context contributes to the relative activation levels of the variants, boosting activation of the appropriate context specific-variant. In the former case, since the conditions contain identical initial syllables, we would not expect to see differences between conditions. Only if the actual, context-specific contour is more highly activated in the sandhi (contour) primes than the low-tone (mismatch) primes will we see differences between conditions.

4.2 Method

Participants

Twenty-four native speakers of Beijing Mandarin were paid for participation in the experiment. All participants signed an informed consent form and had normal or corrected-to normal vision.

Materials

Critical targets consisted of 25 two-character Chinese words, of which the initial character was Tone 2 (e.g. 鱼缸, yu2gang1, 'fish tank'). Each target was preceded by a two-character prime. The initial syllable of each prime was a Tone 3 character, which had the same segmental syllable as the first character of the target. The second character was either Tone 3 (sandhi prime, e.g. 雨水, yu3shui3, 'rain') or a different tone (low-tone prime, e.g. 雨衣, yu3yi1 'raincoat'). For each target word, the initial character of the prime was identical between prime conditions. When two Tone 3 characters occur together, the first is realised with

a rising contour, similar to that of Tone 2. Therefore all prime target pairs differed in terms of the tone category. They either matched or mismatched in the actual contour realisation.

Word frequencies were obtained from Subtlex-CH, a large (46.8 million characters, 33.5 million words) Chinese database based on film subtitles (Cai & Brysbaert, 2010). There was no orthographic overlap (i.e. no shared radicals) between primes and targets. Visual inspection of word frequencies revealed that word frequency was not normally distributed. Therefore, frequencies were transformed using the Johnson transform for normality (jtrans Package version 0.1 in R; Y. C. Wang, 2013) (jtrans Package version 0.1 in R; Wang, 2013).

Table 4.1: Experiment design and sample stimuli

	Contour prime Sandhi (Tone 3 + Tone 3)	**Mismatch prime** Low (Tone 3 + Tone X)
Target yu2gang1 鱼缸	yu3shui3 雨水	yu3yi1 雨衣

Design

The experiment consisted of 200 trials, divided into two blocks of 100 trials, with breaks between the blocks. Each target word was presented twice (once in each prime condition). Two lists were constructed, the order of which was counterbalanced across participants. Lists were pseudo-randomised for each participant. Each block was preceded by three warm-up trials, which were excluded from analysis.

Procedure

Participants were tested individually in a dimly lit, soundproof room, seated approximately 60cm from a 17-inch cathode ray tube computer monitor. A practice session preceded the actual experiment to familiarise participants with the procedure and test the equipment. Stimulus presentation and reaction time data acquisition were conducted using the E-Prime 2.0 software package with a voice key trigger. Participants were instructed to read aloud the words that appeared on the screen as quickly and accurately as possible. All stimuli were presented in black characters on a white background. Each trial began with a fixation cross with jittered presentation time (400-700 ms) to reduce time-induced expectancy waves. A forward-mask of five hash symbols ('#') followed for

100 ms, before presentation of the prime for 48 ms. A backward mask (row of hash symbols) was presented for 17 ms to avoid images of the prime remaining on the retina. Finally, the target word was presented for a maximum of 2,000 ms or until the participant response, which triggered the voice key and caused the word to disappear. The experimenter coded incorrect responses and voice key errors in a 1,400 ms interval before the beginning of the next trial. Response time was calculated from the time of target word presentation until the voice key was triggered by the participant response.

4.3 Analysis and Results

Reaction time data: analysis and results

Reaction time data were analysed using linear mixed effects modelling, using the lmer function of the lme4 package (see also Bates et al., 2013; Baayen, 2008; Baayen et al., 2008) in R (R Development Core Team, 2013). Analysis was conducted on the 1,178 data points remaining after stutters, errors, false starts (<1%) and null responses (<1%) were removed. Since error rates were low, no further analyses were conducted on the errors. Inspection of response latency distributions revealed a skewed distribution, which was normalized by logarithmic transformation. Mean response times were numerically longer in the contour match condition (609 ms) than in the mismatch condition (603 ms).

The baseline model was a regression line of log reaction times (log RT), with random intercepts for subjects and target pictures. Prime type was included to investigate the effect of phonetic context on allophonic variants processing and to determine whether congruency in the acoustic contour affects processing of tone during word reading aloud. Johnson transformed prime and target frequencies were also included to investigate whether item frequencies influenced reaction times or interacted with prime type. None of the predictors or their interactions significantly improved model fit. Only target frequency approached significance ($p>.06$)

Electrophysiological data: recording and pre-processing

The 64-channel electroencephalogram (EEG) was recorded using Neuroscan with electrodes secured in a nylon Electrocap International electrode cap. Electrodes were located at the midline (Fpz, Fz, FCz, Cz, CPz, Pz, POz, Oz) and left and right hemisphere (Fp1, Fp2, AF3, AF4, AF7, AF8, F1, F2, F3, F4, F5, F6, F7, F8, FT7, FT8, FC1, FC2, FC3, FC4, FC5, FC6, C1, C2, C3, C4, C5, C6, T7, T8, CP1, CP2, CP3, CP4, CP5, CP6, TP7, TP8, P1, P2, P3, P4, P5, P6, P7, P8, PO3, PO4, PO5,

PO6, PO7, PO8, O1, O2). Eye movements were monitored using additional electrodes placed above and below the left eye and at the external canthi of the left and right eye. These were offline bipolarized to obtain vertical (VEOG) and horizontal electro-oculograms (HEOG). Electrodes placed on the left and right mastoids served as reference points and the GND electrode served as ground. Electrode impedances were kept below 5ω.

The analogue electrophysiological signal was amplified with a band-pass filter between 0.05 and 100 Hz and digitized at a rate of 500 Hz. The digitized EEG was partially processed off-line using Brain Vision Analyzer 2.0. The signal was DC detrended 400 ms before stimulus markers and band-pass filtered from 0.01 to 40 Hz using an inverse discrete wavelets transform. The raw data was then exported and all other pre-processing was conducted in R (R Development Core Team, 2013). The signal was segmented into epochs of 660 ms (160 ms before and 500 ms after stimulus presentation). Each epoch was baseline corrected on the 160 ms before target onset using the baseline function of the eRp package (version 0.9.8.11; Tremblay, 2013a). Blinks and other ocular artefacts were corrected based on vertical and horizontal EOGs using independent component analysis (ICA) with the icaOcularCorrection package in R (Tremblay, 2013b). Figure 4.3 shows the grand average wave form for the mismatch and contour prime conditions at nine electrodes. In all figures, positive is plotted up.

Electrophysiological data: analysis and results

EEG data were analysed using Generalized Additive Mixed Modeling (GAMM; Wood, 2006) see also Tremblay & Baayen, 2010) using the mgcv package 1.7-28 (Wood, 2011) conducted in R (version 3.0.1; R core team, 2013; www.r-project.org). GAMM is a type of Generalised Linear Model (GLM) that uses non-linear smooth functions to model linear predictors. The method used for this is penalized iteratively reweighted least squares (PIRLS; see Wood, 2006, for details). PIRLS determines the optimal linear and/or non-linear equation for avoiding both over-fitting and over-generalizing of the model.

Due to the high computational cost of modelling each electrode, analysis was conducted on a reduced set of 24 electrodes with an even distribution over the scalp (Figure 4.4). Decompositional models were initially run to determine which individual predictors and interactions influenced EEG amplitude. Auto-correlation (i.e. rho) parameters were determined for each model. The full model included trends over time as random effects for participants and items. A smooth of time was included in the model to examine the changes in amplitude over the course of each trial. Our main predictor of interest, prime type, was included in order to

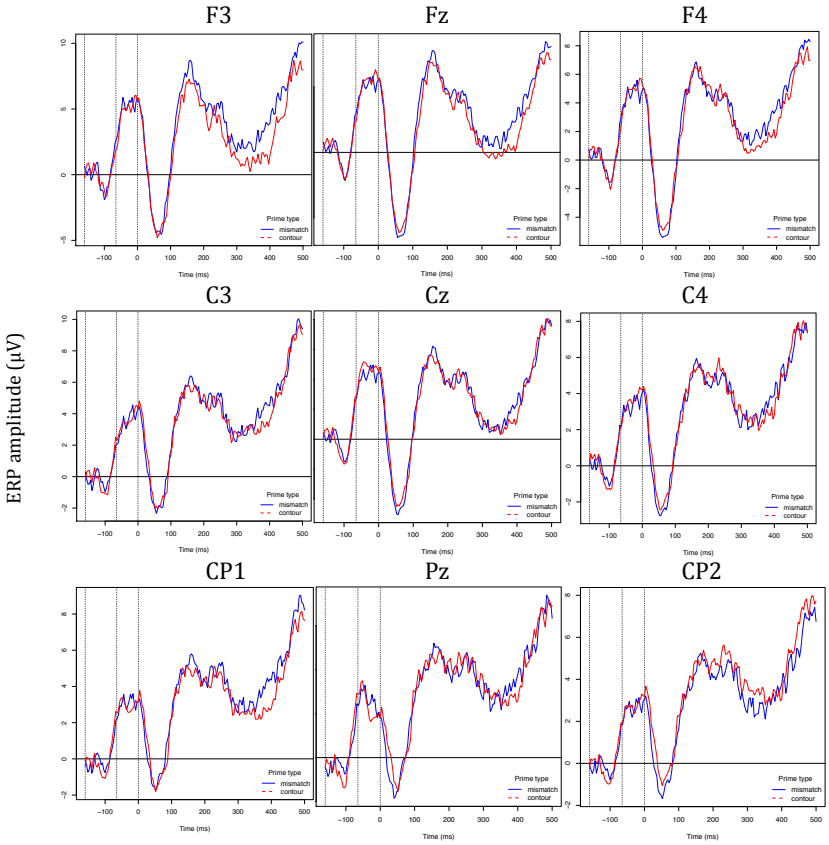


Figure 4.3: Average ERP signal for Contour and Mismatch primes at nine electrodes

determine whether the phonetic context of the following tone affects the processing of Mandarin tone variants and whether cross-category acoustic similarity affects processing during reading aloud. Prime condition was coded as a binary variable, so that in each sample, prime was either low-T3 (0) or sandhi (1). Johnson-transformed prime and target frequencies were also included in the model to investigate whether individual word frequencies affected amplitude of the signal or interacted with the effect of prime type. All two- and three-way interactions were included for each prime type. Model summaries showed significant higher-order interactions between prime type, prime frequency and target frequency over time.

These interactions were then investigated by running separate tensor

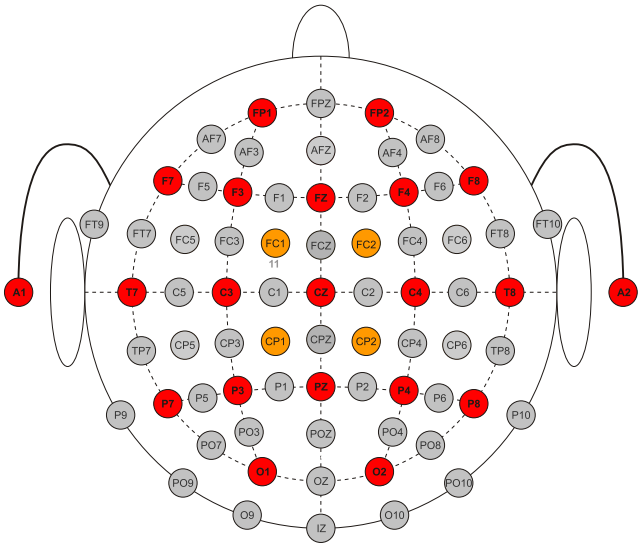


Figure 4.4: Electrode map: models were run for channels marked in red and orange

models for each individual electrode. Rho parameters were determined for and included in each model for each channel. Model comparisons were run separately for each channel using the compareML function (version 2.0; Van Rij, 2014) to determine which predictors contributed to model fit. The compareML function performs a chi-square test on the fREML scores of each model, taking into account the degrees of freedom. Results of the model comparisons are shown in Appendix A. Appendix B shows the model summaries for the best-fit model for each individual electrode.

Random effects

The best-fit model includes trends over time as random effects for participants and target items. The random wiggly curves are shown in Figure 4.5 for participants (left panel) and items (right panel) for the Fz elec-

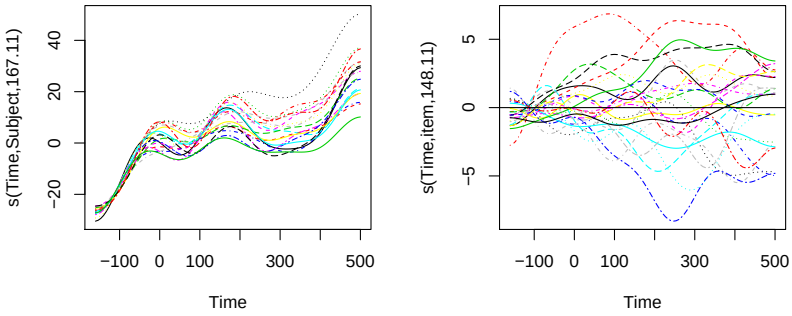


Figure 4.5: Random wiggly curves over time for subjects (left panel) and items (right panel) for the Fz electrode

trode. The figures show the by-subject and by-item adjustments to the amplitude over time.

Fixed effects

Model comparisons (Appendix A) show that model fit was significantly improved for almost all channels by inclusion of a predictor of prime frequency over time (model 2 versus model 1). The models were further improved by an interaction between prime type, prime frequency and target frequency over time (model 3 versus model 2). This 3-way interaction effect over time was consistent across almost the entire scalp (Appendix B).

Figure 4.6 shows the interaction between prime frequency and target frequency over a series of time points (-100, -50ms, 0 ms, 100 ms, 200 ms, 300 ms, 400 ms) for trials with the control prime (left panels) and contour match prime (right panels) for electrode Fz. Yellow indicates relatively more positive amplitude; blue indicates more negative amplitude, green is at the intercept. On the x-axis is prime frequency and on the y-axis is target frequency. The panel rows show the different time points from -100 ms in the top row to 400 ms in the bottom row. The grey dots indicate the individual items.

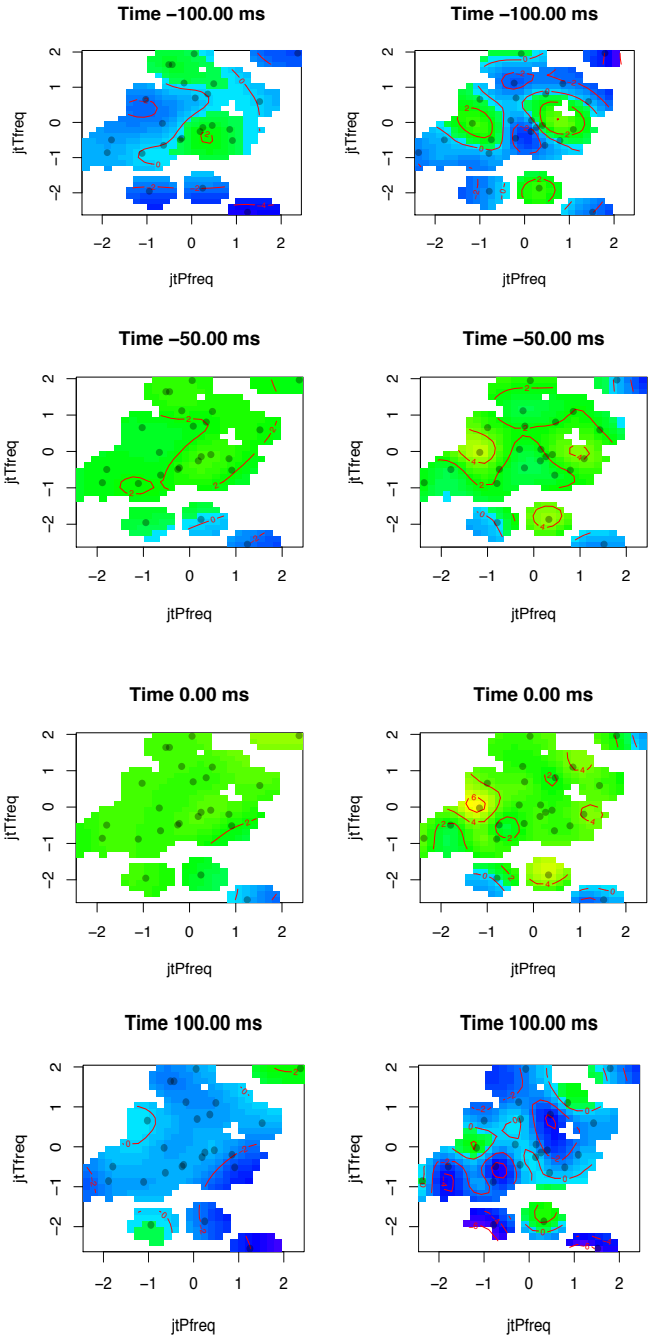
Consistent with the model comparisons and model summaries, the plots show that the two prime types elicit a different pattern of effects. In the mismatch prime condition (left panels), amplitude is relatively flat across prime and target frequencies over time. In the match condition, in contrast, there is a much greater effect of prime and target

frequency over time. At -50 ms and 0 ms, the distribution is quite flat in both conditions; however, at 100 ms, there is greater negativity in the match condition, indicated by the dark blue pools. This is near the peak of the first negative-going waveform. By 200 ms, following the peak of first positive-going waveform, the plots show that the amplitude for both prime types is becoming more positive. In the mismatch prime condition, the amplitude is still relatively flat across frequencies, while in the match condition, positive and negative peaks begin to emerge. These positive and negative peaks continue to increase at 300 ms and 400 ms. Generally speaking, there seems to be more negativity with higher-frequency primes and lower-frequency targets. However, this does not account for all the variance. This point will be returned to in the Discussion.

4.4 Discussion

In a reading aloud task with masked primes, electrophysiological measurements (EEG) and reaction times were used to investigate how phonetic context affects processing of allophonic speech variants. Participants read aloud two-character Mandarin words preceded by briefly presented (48 ms) two-character masked primes whose initial character always differed from the target in terms of tone category, but either matched or mismatched the tone contour (that is, the actual acoustic realisation). The lexical phonetic context Ðspecifically, the tone of the following character - determines whether the tone of the initial character is realized in its canonical (low-tone; mismatch condition) form or in its *sandhi* (rising; contour match condition) form.

No significant differences were found in reaction times between conditions. However, the EEG analysis revealed significant between-condition effects. In the mismatch condition, amplitude of the EEG signal was relatively flat across prime and target word frequencies over time. However, in the contour match condition, there was more negativity on the first negative-going waveform at 100 ms and on the second negative-going waveform, particularly from 300 ms to 400 ms. This effect was modulated by prime and target frequency. There was generally more negativity for lower-frequency targets and higher-frequency primes, but this pattern was not consistent for all items. There were also peaks of positivity at central values. Note that in traditional EEG analysis using ANOVA, only the main effect of prime over time would be analysed, since data are averaged over conditions and analysis of by-item characteristics is not possible. In fact, with the present data, if prime and target frequency information is excluded from the models, prime type is still significant. However, since prime and target frequency contributed to model fit, they were retained in the models. Therefore, it is clear


Time −100.00 ms
Time −100.00 ms
Time −50.00 ms
Time −50.00 ms
Time 0.00 ms
Time 0.00 ms
Time 100.00 ms
Time 100.00 ms
jtTfreq
jtPfreq
−2
−1
0
1
2

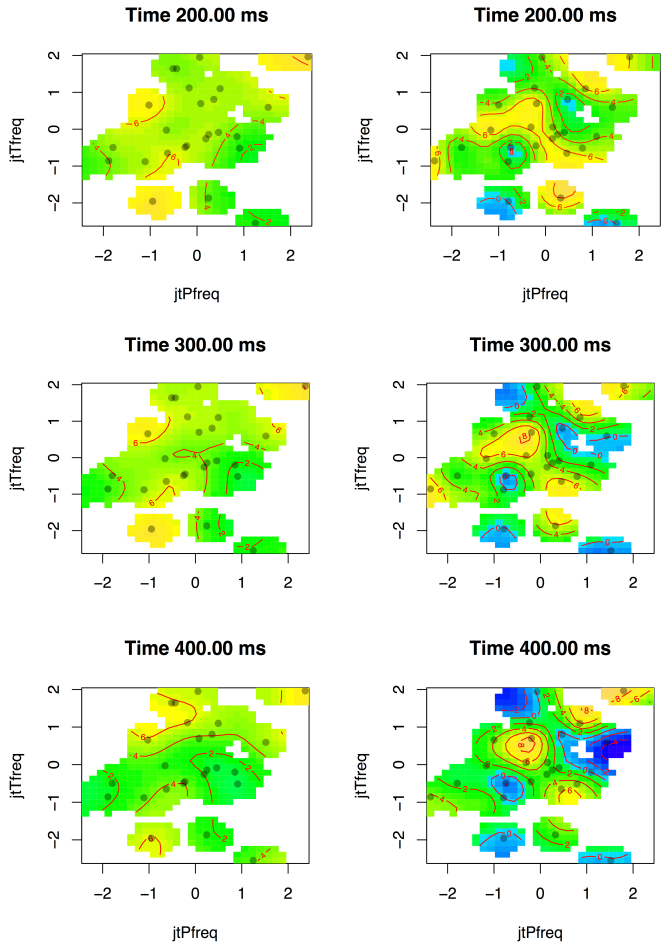


Figure 4.6: Interaction between prime frequency and target frequency at time points -100 ms, -50ms, 0 ms, 100 ms, 200 ms, 300 ms, 400 ms for trials with the control prime (left panels) and contour match prime (right panels) at electrode Fz. Yellow indicates relatively more positive amplitude; blue indicates more negative amplitude; green is at the intercept. Prime frequency is on the x-axis; target frequency is on the y-axis. The panel rows show the different time points from -100 ms in the top row to 400 ms in the bottom row. The grey dots indicate the individual items.

that processing is significantly modulated by prime type, indicating that there is greater activation of the context-specific contour in the contour match condition, compared to the mismatch condition. This supports the hypothesis that processing of phonetic variants is context-specific, i.e. that the surrounding phonetic context constrains activation of the alternative variants. Effects were seen on both the first and second negative-going waveforms. The early time window is associated with sub-phonemic processing of masked primes in silent reading and reading aloud (Ashby et al., 2009). The later time window is associated with sub-lexical phonological processing (e.g. Carreiras, Perea, Vergara & Pollatsek, 2009; Grainger, Kiyonaga & Holcomb, 2006).

What is not yet entirely clear is the role that prime and target frequency play here. It may be that the current analysis does not include the correct measures. For example, it may be that the results are better predicted by bigram frequencies or other co-occurrence measures, rather than simple word frequency measures. There is one other important point to be made here. Although amplitude is comparable between conditions at -50 ms and 0 ms, the plot for -100 ms shows that the model is showing differences between conditions at this early time point. In fact, this should not be possible, because neither prime nor target have been seen yet at this time. In the present study, there were 25 items per prime condition. It may be that this number of items was too small relative to the effect size and that effects may be driven by individual item characteristics. Further work is need to disentangle the precise role that item frequency plays.

Nonetheless, the importance of these frequency measures in processing the target depends on whether or not there is contour overlap between prime and target. This result is reminiscent of previous findings in the tone processing literature. In a Mandarin implicit priming study, J. Y. Chen, Chen and Dell (2002) showed that there was significant priming from overlap of the segmental syllable + prime, and less but still significant priming from the segmental syllable only (when tones differed), but there was no facilitation from tone-only overlap. Similar effects have been found in Cantonese word production (Wong & Chen, 2008). Matching tones did not facilitate processing if the segmental syllable differed. Similarly, in the present results, frequency measures seem to only come into play when there is contour overlap.

The present results extend findings from a previous study, which found evidence for multi-level processing of speech variants (Nixon et al., 2014). This picture-word interference study showed that processing occurs both at the tone category-level and at the level of the context-specific tone variant during overt speech production. Naming of target pictures with sandhi names was facilitated both when target and distractor matched in tone category, but mismatched in tone contour

(toneme condition), and when target and distractor matched in tone contour, but mismatched in tone category (contour condition), compared to controls, which mismatched in both tone category and contour. A second experiment suggested that multi-level processing also occurs when tone variants (sandhi words) are not overtly produced targets, but instead visually presented, ignored distractor words superimposed on targets. The present study shows that automatic context-specific processing can also occur during reading aloud.

The direction of effects was counter to predictions. Based on the earlier study (Chapter 2 of this thesis, Nixon et al., 2014), we expected a facilitatory effect when prime and target matched in contour. Here we see mostly a larger ERP amplitude in the match condition. There are a number of factors which may explain the different direction of effects between the two studies: inhibitory effects in the present experiment, compared to facilitatory effects in the previous study. Firstly, the present study used a different manipulation of the tonal conditions compared to the picture-word interference (PWI) study. The PWI study compared contour match conditions with an unrelated tone (Tone 1 or Tone 4). Although the present results show that context constrains the degree to which the allophonic variants are activated, it is likely that the context-inappropriate allophonic variant receives at least partial activation. This may lead to a reduction in the relative facilitation between conditions. Secondly, the presentation time of the prime relative to the target differed between studies. In the PWI study, the contour-congruency effect was sensitive to the manipulation of stimulus onset asynchrony (SOA). The contour was beneficial only when presented simultaneously with the target. When presentation was delayed (83 milliseconds after the target), there was no facilitation effect from the contour, although there was still facilitation from the tone category. This suggests that, with visually presented stimuli, the sandhi contour may take longer to generate than the tone category. In the present experiment, the prime was presented 65 ms prior to the target word, so lexical retrieval of the prime should have progressed further at the time of target onset, compared to the PWI study. It may be that at early stages, similarity in the acoustic realisation is beneficial to lexical retrieval. However, as lexical access progresses, activation of the matching contour leads to suppression (inhibition) of the competing Tone 2 category.

In summary, the present results provide evidence for automatic retrieval of sub-phonemic information in visually processed masked prime words. Although prime and target belonged to different tone categories, a match in the actual physical contour led to differences in the amplitude of the encephalogram, compared to when there was no contour match between prime and target. Interestingly, counter to predictions,

contour congruency led to greater negativity over time, depending on prime and target frequency. This may reflect an interference effect due to competition between the mismatching tone categories, following the matching contour. Finally, the finding that target word frequency influences neural activity and interacts with other predictors suggests that it is useful to include individual item characteristics in ERP studies of language processing.

Appendix A Model Comparisons

Predictor / Channel	Fp1	Fp2	Fpz	F7	F3	Fz	F4	F8	FC1	FC2	CP1	CP2	T7	C3	Cz	C4	T8	P7	P3	Pz	P4	P8	O1	O2
Interaction of target and prime versus target only	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Interaction of target and prime versus prime only	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Predictor of target frequency	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Predictor of prime frequency	***	***	***	***	*	***	***	***	***	***	ns	***	***	***	***	***	ns	ns	***	*	***	ns	**	***

Appendix B Model Summaries

Predictor \ Channel	Fp1	Fp2	Fpz	F7	F3	Fz	F4	F8	FC1	FC2	CP2	T7	C3	Cz	C4	P3	Pz	P4	O1	O2
Time	***	***	***	***	***	***	***	***	***	***	***	*	***	***	***	***	***	***		***
jtPfreq				.	***	**	*			**	**	.	**		*			*	*	***
jtTfreq									*	*				.	*			***		.
Time:prime	**				*		*	***				**							***	
jtPfreq):prime		***	***	***		***	**	***	***	***	***	***	***	***	***	***	***	***	***	***
jtTfreq):prime		***	***	***			***	***	***	**	***	***	**	***	*	***		***		***
Time,jtPfreq		**			*						*		.						*	***
Time,jtTfreq	**	***		*		*				.	***		**		*	***	***	***		
jtPfreq,jtTfreq					*	.		*				.	.			**		.		
Time,jtPfreq):prime	***	*	**	.	***	***	***		***	***	***		**	**	***				*	
Time,jtTfreq):prime	.	***	***	***		**	*			*	*		*		*	**	.	*	.	
jtPfreq,jtTfreq):prime	**	.	*	***	***	***	***	**	**	***	***	**	***	***	***	**	**	***		
Time,jtPfreq,jtTfreq	*	***		.	**	*	.	***	***	***	*	**		**			*	*	***	***
Time,jtPfreq,jtTfreq):prime	***	***	***	***	***	***	***	***	***	***	***	***	*	***	***	**	.	***	***	**

Chapter 5

Eye movements reflect acoustic cue informativity and statistical noise

A version of this chapter is submitted for publication as:
Nixon, J. S., van Rij, J. Mok, P. Baayen, R. H. & Chen, Y. (submitted). Eye movements reflect acoustic cue informativity and statistical noise.

Abstract

Determining a speaker's message requires discrimination between discrete alternatives based on inherently noisy, non-discrete acoustic cues. Despite growing interest in the role of statistical information in language processing, very little is yet known about how statistical variation affects speech perception. Two visual world experiments investigated how the degree of statistical noise in the input affects perceptual certainty during perception of Cantonese speech sound contrasts. Participants saw four pictures on screen and heard an auditory stimulus. Critical pictures were of word pairs that were identical except for initial consonants (Experiment 1), which were unaspirated (bou2, 'treasure') or aspirated (pou2 'shop'); or tones (Experiment 2), which were high (jin1, 'carpet') or mid (jin3, 'arrow'). Auditory stimuli consisted of a VOT continuum of 12 tokens of increasing VOT (Experiment 1) or pitch (Experiment 2). The number of times participants heard each token followed a bimodal distribution. The width of distribution varied between conditions: wide (high variability) versus narrow (low variability). Eye movements were monitored until participants selected a picture by clicking on it. Fixations were analysed using Generalised Additive Mixed Modelling. Results showed that fixations on the clicked object over the course of the trial varied as a function of VOT value (Experiment 1) or pitch (Experiment 2). In both experiments, this effect significantly interacted with distribution condition. In the narrow conditions, a clear shape of the distribution emerged, with differential eye movement patterns for category means (prototypical values), category boundaries and peripheries. In contrast, in the wide condition, the eye movement behaviour was much flatter. That is, there was a smaller effect of the VOT/pitch value. This indicates that the shape of the acoustic cue distribution plays an essential role in perceptual processing. Increased statistical noise immediately leads to decreased effectiveness of the cue for discrimination during perception.

5.1 Introduction

Human listeners rely on highly variable, non-discrete acoustic information to extract a speaker's intended message from the speech signal. How listeners deal with this variability has been the focus of a torrent of recent research in both speech perception and language acquisition. An initial tack was to attempt to establish invariant cues in the signal. However, such invariance has remained elusive, leading most researchers to seek other explanations. McMurray and Jongman (2011) conducted an extensive investigation of a set of 24 speech cues that have been posited to contribute to distinguishing place, voicing or sibilance, as well as a number of other cues not previously considered, with the aim of determining whether any of the cues had properties of invariance that could be relied on for categorisation. None of the 24 cues investigated were invariant, all being affected to some extent by vowel context. It is clear that acoustic signals, like other physical signals such as visual cues, provide only probabilistic information about a speaker's intended message and that listeners must rely on other mechanisms to comprehend such messages. One of the mechanisms that has been proposed to account for this ability is statistical learning.

Infants start life with very little linguistic knowledge, but with high sensitivity to fine-grained acoustic information. As linguistic knowledge grows, sensitivity to low-level acoustic variation declines and the ability to discriminate between speech categories is enhanced. Though it may seem puzzling that development should involve loss of sensitivity, this phenomenon makes sense from the point of view of enhancing the effectiveness of prediction. When humans learn that the likelihood of a particular outcome or event is associated with a particular cue, they can use that cue to predict future outcomes (Baayen et al., 2013). During early first language acquisition, infants learn how various acoustic cues predict the likelihood of various linguistic outcomes. They can then use this learned process of prediction to facilitate subsequent associations between particular acoustic cues and new linguistic events.

Importantly, this process involves not only tracking of when cues accurately predict a certain outcome, but also when they fail to predict an outcome. Through repeated association between acoustic cues and events in the world, infants learn to discriminate between the various experiences in the world that might be conveyed in speakers' messages. However, contrary to the intuition of learning as acquisition of knowledge, learning may instead essentially amount to learning to ignore uninformative cues. Through prediction error, infants learn which cues are most informative in discriminating between the set of possible speaker messages and which cues do not facilitate this process. At the lexical level, it has been shown that children favour informativity over logic for

assignment of labels to objects (Ramscar, Dye & Klein, 2013).

Assuming that phonological acquisition is an error-driven process, as long as acoustic cues are organised within language in systematically informative ways, they should become associated with lexical-semantic items during speech acquisition. There is abundant evidence that acoustic-phonetic information is organised in such a way. For an acoustic dimension to be informative, particular (ranges of) cue values (e.g short VOT) should have both a high co-occurrence (positive evidence) with one set of outcomes (such as sounds or words, e.g /b/ or *bear*) and a low co-occurrence (negative evidence) with the contrasting set of sounds or words (e.g. /p/ or *pear*). The informativity of a particular cue value can be calculated as a function of both its positive and negative evidence (Ramscar et al., 2013). An important consequence is that the representations of associations between form and meaning are subject to continuous change through experience, and that these changes lead to loss of information. In speech, linguistic events act as both cues and outcomes. This means that experience with language leads to changes in the way cues are used to predict language over time. One implication for adult phonetic processing is that there can be an immediate influence of the informativity of perceptual cues on the processing of speech sounds.

Tracking of transitional probabilities by infants In a seminal article, Saffran, Aslin and Newport (1996) showed that 8-month-old infants were sensitive to statistical information about transitional probabilities in an artificial language. In a training phase, infants were exposed to a two-minute, continuous stream of four different trisyllabic CVCVCV sequences (i.e. 'words'), presented in random order without intonation or prosodic cues. In the test phase, the syllables were identical, but the syllable sequences were either the same or new. Results showed that infants listened longer to syllable strings that spanned 'word' boundaries (i.e. unfamiliar sequences) than they listened to the trisyllabic sequences presented during the training phase, indicating that they had learnt the trisyllabic words. This was arguably the strongest evidence to date of the power of infants' statistical learning ability. If infants were able to process such statistical regularities, this hugely expanded the information available in the environment that could potentially be utilized in language acquisition and processing. Numerous studies have since provided further evidence for infants' ability to use various types of statistical regularities for processing of a wide range of perceptual and linguistic information, including conditional probabilities between syllables and backward transitional probabilities (Aslin, Saffran & Newport, 1998; Pelucchi, Hay & Saffran, 2009b) non-adjacent dependencies between segments (Newport & Aslin, 2004) and native (L1) English in-

fants were able to track transitional probabilities between syllables in natural Italian speech (Pelucchi, Hay & Saffran, 2009a).

Phonetic category learning in infants In natural language, acoustic information is continuous and values of acoustic cues tend to cluster along phonetic dimensions. For a given acoustic cue, speech contrasts correspond to two or more distributional peaks along a continuum. For example, in English, the voicing distinction has a bimodal distribution of voice onset time with peaks around 0ms and 60ms (Lisker & Abramson, 1964). Although the extent to which infants are able to use distributional information in phonetic category learning is not yet clear (Cristià, McGuire, Seidl & Francis, 2011), there is some evidence that infants may use information about statistical clustering of acoustic cues to determine the number of categories along a phonetic dimension. Using the preferential looking paradigm, Maye et al. (2002) trained 6- and 8-month-old infants on speech sounds resynthesized into an 8-point continuum between voiced [d] and unaspirated [t] (spliced from words beginning with /s/). The infants heard the same sounds, but the presentation frequency varied between conditions. Theyeither heard a distribution of stimuli with a single peak in the centre of the continuum (unimodal distribution) or with two peaks towards the endpoints of the continuum (bimodal distribution). In the test phase, infants heard stimuli tokens from the endpoints of the continuum either in alternating or non-alternating (i.e. blocked) trials. Results showed that only infants in the bimodal distribution looked longer in the alternating trials, suggesting that the infants in the unimodal condition had lost sensitivity to the contrast. A later study found that distributional information can also enhance discrimination of difficult contrasts (Maye et al., 2008) Recent evidence suggests that sensitivity to statistical probabilities in acoustic input may change over time. Liu and Kager (2011) found that 11-12-month-old, but not 5-6-month-old Dutch-learning infants benefitted from a bimodal, compared to the unimodal distribution during Mandarin tone discrimination. However, using the mismatch negativity paradigm, Wanrooij, Boersma and van Zuijen (2014) showed that fast phonetic learning can occur in infants as young as 2-3 months.

Distributional learning in adults Although the majority of distributional learning studies have aimed at investigating development of phonetic categories during first language acquisition, a number of studies have also examined the role of distributional learning in adults. Maye and Gerken (2000) used unimodal versus bimodal distributions to demonstrate that adults can use statistical information to determine the number of phonetic categories along an acoustic continuum. In

these experiments, participants were exposed to a familiar acoustic cue dimension, voice onset time (VOT), contrastive in their native language (English). Although VOT is contrastive in English, the particular values selected for the stimuli do not correspond to the usual English contrast. One half of the stimuli corresponded to a context-specific allophone (the /t/ stimuli were spliced from words beginning with /s/), which does not normally occur word-initially. Therefore, the stimuli pairs fall somewhere between contrastive and non-contrastive. Although English speakers can discriminate these sounds, results showed that after exposure to the unimodal distribution, they were more likely to categorise the sounds as the same, compared to participants exposed to the bimodal condition.

Much of the research in adult distributional learning work has focused on the acquisition and development of non-native contrasts. For example, a series of recent studies has investigated the effects of statistical distributions on non-native perception of Dutch vowel contrasts (Escudero et al., 2011; Gulian et al., 2007; Wanrooij et al., 2013) Gulian et al. (2007) presented adult native Bulgarian naive listeners with Dutch vowel contrasts which were non-contrastive in Bulgarian. After a 5-minute exposure period, participants in the bimodal group distinguished the Dutch vowels better than those in the unimodal group.

Acoustic distance between categories in cue distributions The studies described up to this point have generally used bimodal versus unimodal distributions of stimuli. In these studies, the same tokens were presented to all participants, but at different presentation frequency. Escudero et al. (2011) instead used two types of bimodal distributions to examine the affects of acoustic distance between vowel categories in second language acquisition. This was motivated by the observation that infant and foreigner directed speech has a stretched vowel space. They used *natural bimodal* versus *enhanced bimodal* distributions to train Spanish learners to distinguish a Dutch vowel contrast not present in the native language and which Spanish learners of Dutch typically find difficult to acquire. In both conditions, participants heard a bimodal distribution. In the natural bimodal condition, the acoustic distance between vowel categories was reduced (i.e. the two vowel sounds were more similar to each other) compared to natural speech; in the enhanced distribution condition, acoustic distance was increased (i.e. the vowel sounds were more different from each other) relative to natural speech. They exposed Spanish learners of Dutch to two minutes of natural bimodal or enhanced distributions or to music (as a control group). Compared to the music group, there was an increase in 'correct' categorisation in the enhanced distribution group, and a small, but al-

most significant increase in the natural bimodal group. The increase in the enhanced group was significantly different to 0.

In a follow-up study, Wanrooij et al. (2013) replicated and extended the study to investigate individual differences in the use of acoustic cues (*'listening strategies'*) using *latent class regression modelling*. They split the participants into two groups, low- and high-performers, based on pre-test performance: namely whether participants made use of the two most important cues to the Dutch vowel contrast, the first and second formants. Both training distributions led to increased use of F1 and/or F2 as cues to vowel discrimination and there was no difference between conditions. However, although the enhanced distribution did not have the predicted effect on the trained cues, F1 and F2, there were interesting effects in the *untrained* cues. There was a significant effect of distribution (natural bimodal vs enhanced) on bootstrapping the use of vowel duration, modulated by pretest performance (listening strategy). All groups increased their use of the duration cue, but for the pre-test low performers (participants who tended not to use the most important cues in the pre-test) the increase in use of duration was greatest in the enhanced group, followed by the bimodal group, then the music group. For the pre-test high performers, the enhanced distribution seemed to bootstrap the use of a further cue, in most cases F3, relative to the natural bimodal distribution.

Online measures of perceptual uncertainty during processing of speech contrasts Most studies of distributional learning in adults have used offline categorisation responses as the measure of learning. Categorisation measures provide information about the final outcome of the decision process; however, they do not provide information about the online processing of the perception itself. In discussions of effects on categorisation, it is often implicitly or explicitly assumed that assigning tokens to one category rather than two occurs because the two tokens were not discriminated. This assumption may not necessarily be justified. In a forced-choice categorisation task, regardless of the degree of uncertainty, or any gradient degree of goodness of fit with one category or another, the participant must make a binary choice. While it is interesting that factors such as cue distribution can affect even the final outcome of the decision process, examining the moment by moment online processing can tell us about how subtle differences in statistical distributions can affect the development of perceptual processes over time, prior to the decision process.

One interesting recent study, to the best of our knowledge the only study that has used online measures to investigate statistical processing of acoustic cues during perception of native speech contrasts, is an in-

novative eyetracking study by Clayards and colleagues. Clayards et al. (2008) found that when English listeners heard wide distributions of acoustic cues, it led to a shallower slope in the categorisation function, and less certainty in eye movements, compared to the narrow distribution.

The present study The present study makes several important new contributions to the investigation of acoustic cue distributions in speech perception. Firstly, in the Clayards et al. study, eye movement measures were collapsed across the whole trial, giving only a single measurement of the proportion of fixations per VOT value per condition. Secondly, Clayards and colleagues made specific predictions about the relative size of distribution effects according to the location within the VOT distribution. Due to the relatively small number of trials at particular points in the distribution, they were unable to test the predictions for all VOT values. In the present study, we increased the number of participants and used a new method of analysis (described below), which allowed us to investigate the full range of the cue distribution. We were also able to analyse the changes in eye movement patterns over the course of the whole trial, rather than collapsing over the trial. This gives a substantially more in-depth picture of the complex interaction of effects over time. As will be shown, while some of Clayards and colleagues' predictions are upheld, our results also show other effects not predicted in their model, suggesting a possible difference in the underlying mechanism of the use of acoustic cues.

The present study investigates how the degree of variation in acoustic cues affects perception of Cantonese speech contrasts. Experiment 1 investigates whether Clayards and colleagues' findings can be generalised to a new language, which uses a different range of VOT values for the aspirated-unaspirated distinction in stops. The experiment also included affricates to investigate whether the effects on VOT perception also apply in a different phonetic environment. This is important because, as Rost and McMurray (2009) point out, the effects of distributional learning are likely to be dependent on the specific acoustic properties in question. For example, Jongman, Wayland and Wong (2000) found a relatively large overlap in contrastive and non-contrastive cues in English fricatives, which is greater than that of stops. Experiment 2 extended the investigation to a suprasegmental cue, lexical tone.

5.2 Experiment 1 Voice onset time

Method

Participants Thirty-seven native Cantonese speaking undergraduate students from the Chinese University of Hong Kong were paid for participating in the experiment. Participants were tested individually in a quiet room.

Experiment design and stimuli Visual stimuli were picture pairs whose names were word pairs that began with either bilabial stops ('b', 'p') or alveolar affricates ('j', 'ch'). The two members of each word pair were identical except for the initial consonants, which were either unaspirated (bou3, 'cloth'; jun1 'brick') or aspirated (pou3, 'shop'; chun1, 'village'). Pictures were black-on-white line drawings selected from the MPI and Els Severens (Severens, Lommel, Ratinckx & Hartsuiker, 2005) picture databases, supplemented with pictures from the internet.

All auditory stimuli were recorded by a male native speaker of Hong Kong Cantonese. Stimuli were then resynthesised into a 12-step VOT continuum using the Pitch-Synchronous-Overlap-and-Add (PSOLA) method in PRAAT (Boersma & Weenink, 2012). For the stops, the duration of the consonant ranged from 0 ms to 88 ms. For the affricates, the frication duration ranged from 60 ms to 280 ms. Table 5.1 shows the presentation frequency of each step on the continuum. The number of times participants heard each step followed a bimodal distribution, with the two peaks of the distributions corresponding to the prototypical mean VOT for the unaspirated and aspirated stimuli, respectively. Each condition contained 456 tokens, 76 for each word pair. All participants heard the same number of tokens; only the number of times they heard each token (i.e. the width of distribution) varied between conditions: wide (high variability) versus narrow (low variability).

Table 5.1: Presentation frequency per variant per condition: each variant represents one step on the VOT continuum

		Number of iterations											
	Variant	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Distribution	**Narrow**	0	6	54	108	54	6	6	54	108	54	6	0
condition	**Wide**	6	24	54	60	54	30	30	54	60	54	54	6

The experiment consisted of 456 experimental trials, divided into six

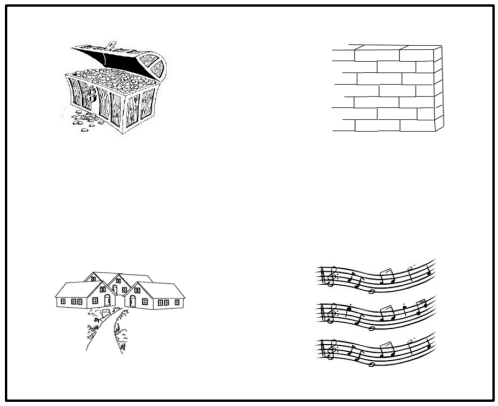


Figure 5.1: Sample screen display during stimulus presentation

blocks of 76 trials, with breaks between blocks. The order of presentation was randomised for each participant.

Procedure Participants sat at a comfortable viewing distance from the computer screen and wore an SR Eyelink II head mounted eye-tracker with a sampling rate of 500 Hz. Stimulus presentation and data acquisition were conducted using SR Research Experiment Builder software. The session began with 12 familiarization trials in which participants saw the pictures and their corresponding written labels once each. This was followed by a practice block to familiarize participants with the experimental procedure. None of the experimental pictures or words were presented during the practice phase.

Each experimental trial began with drift correction to ensure accurate calibration of the equipment, followed by brief presentation (500ms) of four pictures, one in each quadrant of the screen (see Figure 5.1). The purpose of giving an advance preview of the stimuli was to reduce the time and likelihood of participants scanning the pictures at the beginning of the trial, and hence to reduce noise in the eye movement data. The display always contained two test items and two filler items. The location of the picture conditions on screen, as well as their relative location, was randomised to avoid strategic effects. The pictures disappeared, replaced with a gaze-contingent fixation cross, which ensured participants were looking at the centre of the screen at the beginning of the critical trial period. One of the auditory stimuli was presented and participants chose the picture they thought most appropriate by clicking on it with the mouse. Eye movements were monitored from the onset of the auditory stimulus until participants made a response.

5.3 Analysis

Eye movement data for Experiments 1 and 2 were analysed using *Generalized Additive Mixed Modeling* (GAMM; Wood, 2006) using the mgcv package 1.7-28 conducted in R (R core team, 2013; www.r-project.org). GAMM is a type of Generalised Linear Model (GLM) that uses non-linear smooth functions to model linear predictors. The method used for this is penalized iteratively re-weighted least squares (PIRLS; see Wood, 2006, for details). PIRLS determines the optimal linear or non-linear equation for avoiding both over-fitting and over-generalizing of the model.

All predictors of interest were entered into a GAMM model using an iterative backward fitting model updating procedure, and predictors that did not contribute to model fit were eliminated. Model comparison was conducted using the statistics provided by the model summaries and the Somers' Dxy Rank Correlation in the Hmisc package [version] in R. Because the auditory stimuli were an acoustic continuum rather than two distinct categories, the target picture was defined as the picture selected and clicked on at the end of each trial (henceforth *clicked target*). This was done in order to allow participant-defined category boundaries, rather than imposing a pre-defined boundary. This led to similar proportions as if we had analysed the target as belonging to the pre-defined boundaries (i.e. VOT/pitch steps 1-6 versus 7-12). The dependent variable fixation on the clicked target was coded as a binary variable, so that in each sample, fixation was on clicked target (1) or competitor (0). VOT (Experiment 1) and pitch (Experiment 2) variants were modelled as a continuous variable, centred around 0. Because we were interested in processing over the course of the whole trial, from early perceptual processing to later decision processes, the predictor *time* was included. Since generation of a saccade takes about 150-200 ms and inspection of standard errors revealed that the time period between 0 ms (i.e. the presentation of the auditory stimulus) and 200 ms contained too few data points to yield reliable analysis, an 800 ms time window from 200 ms to 1000 ms was selected for analysis. Manner of articulation (hereafter *manner*) was included to investigate whether there were differences in effects between stops and affricates. Log transformed frequency was included in the model as a control variable. Frequency information for each target item was obtained by entering the character in the Google search engine and recording the number of hits[1]. In order to ensure

[1]Hong Kong uses traditional script. In Mainland China, some characters have been simplified, while others remain the same as those in traditional script. In addition, the population and, therefore, the volume of internet traffic is much greater in China than in Hong Kong. Therefore, the overlap in scripts means that searches involving unsimplified characters (i.e. characters that are the same in traditional and

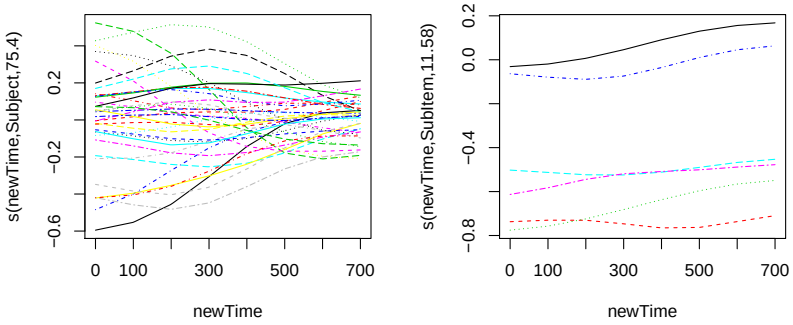


Figure 5.2: Random wiggly curves per subject (left panel) and item (right panel) over time (ms).

frequencies reflected frequency in Cantonese and not other Chinese languages, searches were restricted to Hong Kong websites. A predictor for *trial* was entered in the model in order to investigate learning over the course of the experiment, however this did not contribute substantially to model fit, so was removed. Inspection of the residuals of the first statistical model revealed a high degree of correlation between subsequent measurements. Therefore the data were downsampled to 10 Hz, which eliminated the autocorrelation. The best-fit model is presented in Appendix A. The estimated effects are on logit scale. A logit of 0 is equivalent to 0.5, negative logit values represent values lower than 0.5 and positive logit values represent values over 0.5.

5.4 Results

Random effects

The best-fit model (Appendix A) includes trends over time as random effects for participants and target picture items. The random wiggly curves are shown in Figure 5.2 for participants (left panel) and items (right panel). The figures show the by-subject and by-item adjustments to the proportion of looks to the clicked target picture over time.

simplified script) are likely to yield much a higher number of hits than traditional characters. The word frequency search was restricted to Hong Kong websites to avoid this issue.

Voice onset time

The summary of the best-fit model (Appendix A) revealed a significant effect of VOT over time ($\chi^2(16.29)=325.65$, $p<.001$). Estimated effects of VOT over time (i.e. over the course of each single trial) are shown for the baseline (narrow) condition in the left panel of Figure 5.3. Yellow (positive values) indicates relatively more looks to the clicked target; blue (negative values) indicates relatively more looks to the competitor. Category means are at -2.5 (for the unaspirated stimuli) and 2.5 (for the aspirated stimuli).

The plot indicates that changes in eye movements over the course of the trial occur differently at different points on the VOT continuum. The early part of the trial reveals a bimodal distribution of increased fixations on target (yellow areas) around the two means. Over the course of the trial, this high proportion of fixations on the clicked target gives way to an increasing number of fixations on the competitor. Around the category means, fixation proportions to clicked target and competitor even out (green areas) and remain stable until the end of the trial period. However, quite a different pattern of eye movements emerges near the category boundary. There are fewer looks to the clicked target in the early part of the trial, and fixations are increasingly likely to land on the competitor picture over time.

Effects of distribution condition

In addition to the interaction between Time and VOT, there was also a significant interaction between distribution condition and VOT over time ($\chi^2(13.75)=172.91$, $p<.001$). This is further supported by the differential patterns of looking behaviour over time between distributions, shown in the right panel of Figure 5.3 (wide condition), compared to the left panel (narrow condition). The pattern of fixation behaviour in the wide condition lacks the distinct boundaries seen in the later time period in the narrow condition. Until approximately 400 ms, looking behaviour is similar between conditions. After this time, the wide condition becomes quite flat, while the narrow condition displays divergent patterns across VOTs. The pattern of eye movement seems to reflect different stages of processing: an initial, perceptual stage and a later process of verification of whether the initial perceived sound matches the identified picture. While VOT appears to have an early effect on the likelihood of the first fixations landing on target, distribution effects come into play at a relatively late stage of processing. During this later stage of processing, participants seem to be looking at the competitor object as part of the process of rejecting it. When there is more noise in the signal (in the high-variability wide condition) more fixations on the

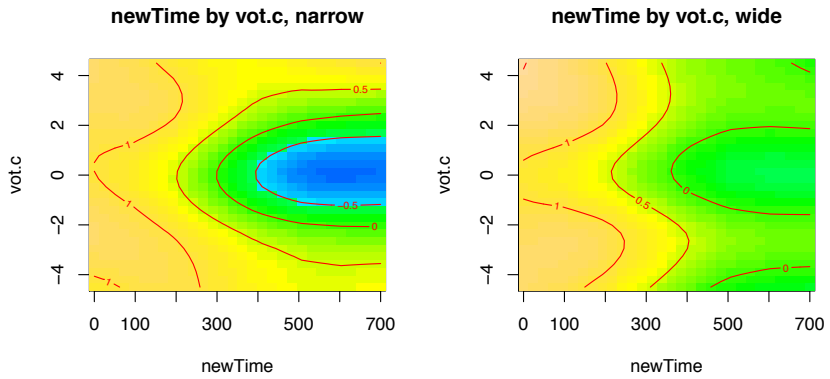


Figure 5.3: Topographical maps of the proportion of fixations on clicked target versus competitor for VOT over time for the narrow (left panel) and wide conditions (right panel) in the GAMM model. Estimated effects are on logit scale. Yellow (positive values) indicates relatively more looks to clicked target; blue (negative values) indicates relatively more looks to the competitor. Voice onset time (VOT) is on the vertical axis. VOT is centred around 0, the category boundary. The negative vot.c values correspond to unaspirated stimuli, the positive values to aspirated stimuli. Category means are at vot.c -2.5 and 2.5, respectively. Time (ms) is on the horizontal axis. These plots are plotted with MannerBin (binary variable of manner of articulation) equal to zero. They include the intercept and subject and item effects.

competitor are necessary in order to reject it. This is perhaps analogous to the increased latencies rejecting non-words in lexical decision tasks when with increased similarity to real words (e.g. Coltheart, 1977; Perea & Lupker, 2003).

The low noise (i.e. high cue predictiveness) in the narrow condition led to greater overall certainty for cues within the expected range, compared to the wide condition. However, there seems to also be a trade-off of this relative certainty in the narrow condition: cues near the category boundary incurred a greater cost of rejecting the competitor, compared to the wide condition.

Frequency effects

Model fit was significantly improved by adding an interaction between log frequency of clicked target items and time ($\chi^2(10.23)=248.83$, $p<.001$). Estimated effects of log frequency over time are shown in Figure 5.4. It should be noted that these are partial effects, not absolute

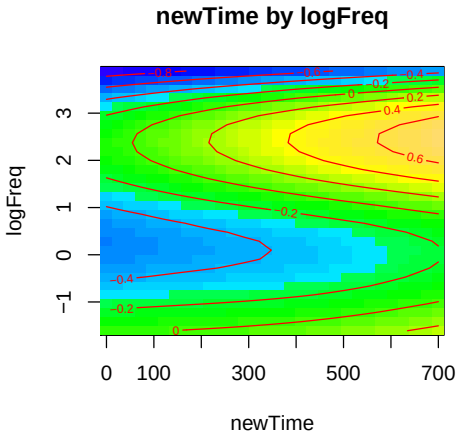


Figure 5.4: Topographical map of log frequency over time in the GAMM model for Experiment 1: proportion of fixations on the clicked target versus competitor over the course of the trial as a function of target log frequency. Estimated effects are on logit scale. Yellow (positive values) indicates relatively more looks to clicked target; blue (negative values) indicates relatively more looks to the competitor. Log frequency is on the vertical axis.

values. In order to obtain the estimated values, these plots need to be added to the other effects in the model. In the early part of the trial, for the lower frequency items, there are more of fixations on the competitor, while for the higher frequency items, there are more of fixations on the target. For both low- and high-frequency items, the proportion of fixations on the target increases over the course of the trial. This suggests that frequency has a very early effect, influencing the likelihood of the first fixations landing on target and affecting the proportion of fixations throughout the whole trial period. Essentially, the first fixations are drawn towards the higher frequency item, even when it is not selected as the target picture. This is likely to reflect a greater number of verification fixations on the high-frequency competitor due to difficulty in rejecting the incorrect picture when it is high frequency, compared to low-frequency competitors.

In order to further investigate the effect of frequency in the very early fixations, we created a new data set containing only the fixations at 200ms after target presentation. The model (see Appendix B for the model summary) showed no effect of VOT, but main effects of Trial and

frequency ratio - the frequency of the target relative to the competitor. A main effect of Trial (β=0.001, z=3.01; p=.003) indicates that the likelihood of these early fixations landing on the clicked target increased over the course of the experiment. However, a negative main effect of frequency ratio (β=-0.053, z=-2.21; p=.03) indicates that the higher the relative frequency of the clicked target, the more likely the fixation is to land on the low frequency competitor. This early onset of frequency effects is consistent with previous findings that high-frequency competitors will have an advantage in the early stage of the trial (Tanenhaus, Magnuson, Dahan & Chambers, 2000).

Manner of articulation

The model summary shows a significant interaction of manner by VOT (χ^2(2.81)=86.85 p<.001). This interaction is visualised in Figure 5.5, which shows the adjustments to the VOT effect (shown in the left panel of Figure 5.3) for the different manner conditions. Positive values of manner correspond to the affricates, and negative values to stops. Around the central VOT values, a slight positive adjustment needs to be made for the affricates and a slight negative adjustment for stops. Note that this effect was not modulated by condition, so applies to both the narrow and wide distributions.

Discussion

The GAMM model showed that the pattern of eye movements over time was predicted by VOT value. The baseline (narrow distribution) condition (left panel of Figure 5.3), shows that the pattern of eye movements reflects the shape of the input distribution. The early part of the trial has a bimodal distribution of slightly increased proportion of fixations on the clicked target (darker yellow) around the category means. In the later part of the trial, after about 200 ms, differential fixation patterns occur for the category means (green areas), peripheries (yellow areas) and category boundaries (blue area).

More interestingly, in addition to VOT over time, there was also an interaction between VOT and distribution condition over time. In contrast to the narrow condition, fixation behaviour in the wide distribution condition (the right panel of Figure 5.3) is very similar across VOT values in the later part of the trial. In other words, when the acoustic cue VOT was less informative, it had less influence on eye movements as participants attempted to discriminate between words. The greater the acoustic variation in the input, the less effective the cue was for discrimination between candidate pictures.

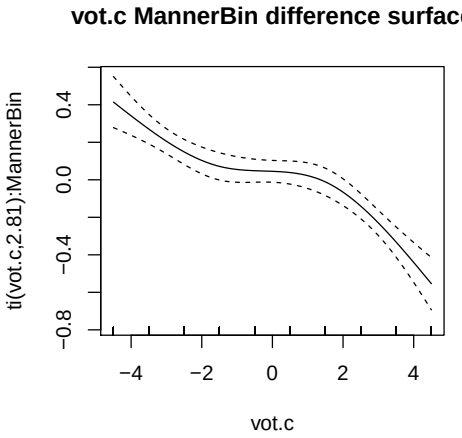


Figure 5.5: Plot of the interaction between manner and VOT. Estimated effects are on logit scale. The difference logit is on the vertical axis. Positive values correspond to affricates and negative values to stops. Voice onset time (VOT) is on the horizontal axis. VOT is centred around 0, the category boundary. The negative vot.c values correspond to unaspirated stimuli, the positive values to aspirated stimuli. Category means are at vot.c -2.5 and 2.5, respectively.

5.5 Experiment 2 Tones

Method

Participants Forty native Cantonese speaking undergraduate students from the Chinese University of Hong Kong were paid for participating in the experiment. Participants were tested individually in a quiet room.

Experiment design and stimuli The experiment design was the same as Experiment 1, except that different target items were used. Visual stimuli were picture pairs with whose names were word pairs that were either high (e.g. jin1 'carpet'; gun1 'crown') or mid tone (jin3 'arrow'; gun3 'can'). The two members of each word pair had the same segmental syllable. Initial consonants were either velar stops ('g') or alveolar affricates ('j'). Auditory stimuli were created in the same manner as in Experiment 1, except that the continuum was of f0, instead of VOT values.

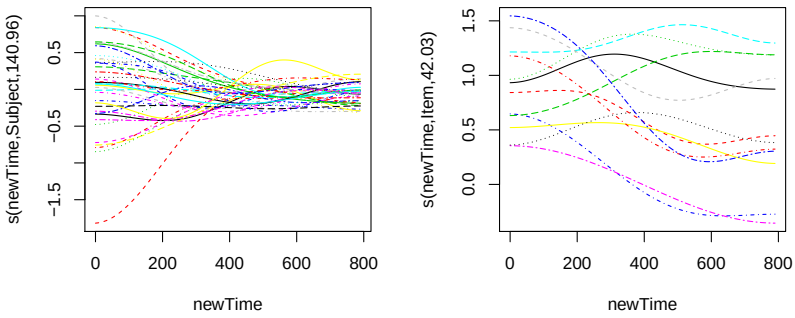


Figure 5.6: Random wiggly curves over time for participants (left panel) and items (right panel) in Experiment 2.

Procedure The procedure was identical to Experiment 1.

5.6 Analysis

Analysis was conducted using the same variables as Experiment 1, except that the acoustic cue was a continuum pitch (f0), instead of VOT values.

5.7 Results

Random effects

Figure 5.6 shows the random wiggly curves for participants and target picture items. By-subject and by-item adjustments to the proportion of looks to the clicked target picture over time are shown for participants (left panel) and items (right panel).

Pitch

The model summary (Appendix C) shows a significant effect of pitch over time ($\chi^2(18.132)=389.20$, $p<.001$). Estimated effects of pitch over time are shown for the baseline (narrow) condition in the left panel of Figure 5.7. Yellow (positive values) indicates relatively more looks to the clicked target; blue (negative values) indicates relatively more looks to the competitor. Category means are at -2.5 (for the mid-tone stimuli) and 2.5 (for the high-tone stimuli).

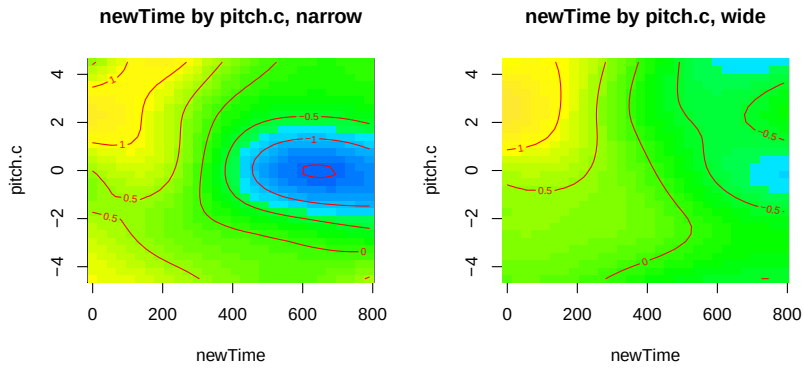


Figure 5.7: Topographical maps of the proportion of fixations on clicked target versus competitor for pitch over time for the narrow (left panel) and wide conditions (right panel) in the GAMM model. Estimated effects are on logit scale. Yellow (positive values) indicates relatively more looks to clicked target; blue (negative values) indicates relatively more looks to the competitor. Pitch is on the vertical axis. Pitch is centred around 0, the category boundary. The negative pitch.c values correspond to mid tone stimuli, the positive values to high tone stimuli. Category means are at pitch.c -2.5 and 2.5, respectively. Time (ms) is on the horizontal axis. These plots are plotted with MannerBin (binary variable of manner of articulation) equal to zero. They include the intercept and subject and item effects.

The plot indicates that changes in eye movements over the course of the trial occur differently at different points on the pitch continuum. In the early part of the trial, there are relatively more fixations on the target (yellow areas), particularly around the mean of the high tone and just below the mean for the mid tone. This distance from the mean in the mid tone may suggest that our predefined category boundaries were slightly higher than the listeners' initial boundaries. Over the course of the trial, this initial high proportion of fixations on the clicked target gives way to an increasing number of fixations on the competitor. After a few hundred milliseconds, fixation proportions to the clicked target and competitor even out around the category means (green areas), and remain stable until the end of the trial period. In contrast, eye movement behaviour is quite different near the category boundary. There are fewer looks to the clicked target in the early part of the trial, and fixations are increasingly likely to land on the competitor picture over time.

As shown in the model summary (Appendix C) there was a signi-

ficant interaction between distribution condition and pitch over time ($\chi^2(18.00)=219.84$, $p<.001$). Consistent with this, the pattern of eye movements over time differs between distribution conditions, shown in the right panel of Figure 5.7 (wide condition), compared to the left panel (narrow condition). The effects are similar to Experiment 1. In the later part of the trial, the pattern of fixation behaviour in the wide condition is overall much flatter than in the narrow condition. Until approximately 400 ms, the pattern of fixations is similar between conditions, except that the narrow condition shows more looks to the mid tone clicked target. Later in the trial, the wide condition becomes flat. Fixation proportions in the narrow condition, in contrast, depend on the pitch value, with many more looks to the competitor near the category boundary.

Frequency effects

The model summary (Appendix C) shows a significant effect of the log frequency of clicked target items over time ($\chi^2(12.79)=214.6$, $p<.001$). Estimated effects of log frequency over time are shown in Figure 5.8. The plot shows partial effects, which are added to the other effects in the model. In the early part of the trial, there are more fixations on the clicked target for the higher frequency items. For both low- and high-frequency items, the proportion of fixations on the competitor increases over the course of the trial. As in Experiment 1, frequency has a very early effect and continues to influence fixations throughout the whole trial period.

Manner of articulation

The model summary shows a significant interaction of manner by pitch ($\chi^2(2.96)=57.81$ $p<.001$). Figure 5.9 shows this interaction. Note that this is a partial effect. The figure shows the adjustments to the pitch effect (shown in the left panel of Figure 5.7) for the different manner conditions. Positive values of manner correspond to the affricates, and negative values to stops. Around the central pitch values, a slight positive adjustment needs to be made for the affricates and a slight negative adjustment for stops.

Discussion

The results for Experiment 2 were similar to Experiment 1, except that the acoustic cue *pitch* in Experiment 2 replaced *VOT* in Experiment 1. The GAMM model showed that the proportion of fixations on the clicked target over the course of the trial period was predicted by pitch

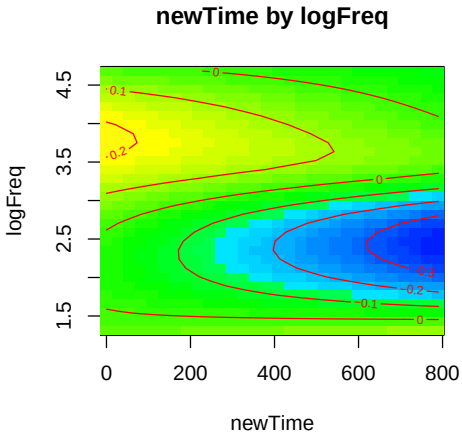


Figure 5.8: Topographical map of log frequency over time in the GAMM model for Experiment 2: proportion of fixations on the clicked target versus competitor over the course of the trial as a function of target log frequency. Estimated effects are on logit scale. Yellow (positive values) indicates relatively more looks to clicked target; blue (negative values) indicates relatively more looks to the competitor. Log frequency is on the vertical axis.

value. Fixations on the clicked target in the early part of the trial were greatest around the mean of the high tone and just below the mean of the mid tone (Figure 5.7). That the effect was below the category mean for the mid tone suggests that listeners' initial boundaries may have been slightly lower than our pre-defined boundaries. As in Experiment 1, the effect of the acoustic cue value depended on the width of distribution condition that participants heard. The plots show that the effect of the pitch value was much flatter in the wide condition (right panel) compared to the narrow condition (left panel). This indicates that, like VOT, when the acoustic cue pitch was informative for discriminating between words (narrow condition), it influenced the proportion of eye movements to the clicked target more than when it was less informative (wide condition).

5.8 General Discussion

The present study investigated how the shape of acoustic cue distributions affects perceptual certainty during perception of speech contrasts.

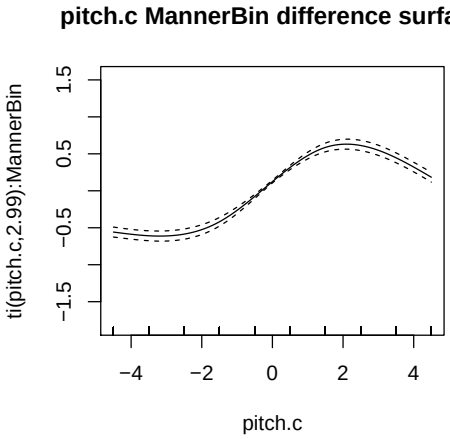


Figure 5.9: Plot of the interaction between manner and pitch. Estimated effects are on logit scale. The difference logit is on the vertical axis. Positive values correspond to affricates and negative values to stops. Pitch is on the horizontal axis. Pitch is centred around 0, the category boundary. The negative pitch.c values correspond to mid-tone stimuli, the positive values to high-tone stimuli. Category means are at pitch.c -2.5 and 2.5, respectively.

Participants heard either a relatively large amount of variation (the *wide* distribution condition) or relatively little variation in acoustic stimuli (the *narrow* distribution condition) and saw pictures of word pairs consisting of aspirated and unaspirated counterparts (Experiment 1) or mid and high tone counterparts (Experiment 2). Eye movements to these pictures were monitored until participants selected a picture by clicking on it. We hypothesised that greater variation in the signal would lead to greater uncertainty in processing of the speech contrasts.

Analysis of eye movement data using generalised additive mixed modelling revealed that the proportion of fixations on the clicked target over the course of the trial varied as a function of VOT value (Experiment 1) or pitch (Experiment 2), and that this significantly interacted with distribution condition. The left and right panels of Figure 5.3 (Experiment 1) show that in both conditions, there were relatively many fixations on the clicked object compared to the competitor in the early part of the trial period. This early tendency to fixate on the clicked target was modulated by VOT. The effect was greater and continued for longer for prototypical VOT values compared to values near the category

boundary. The pattern was very similar in Experiment 2 (Figure 5.7), except that the high proportion of fixations for the mid-tone was at lower pitch values in these early fixations, suggesting that the pre-defined category boundaries may have been slightly higher than listeners' initial category boundaries. Following this initial few hundred milliseconds, around the prototypical pitch and VOT values in both the narrow and wide conditions, looks to the competitor increased around 500-600 ms into the trial and then levelled off. As seen in the left panel of Figure 5.3 (Experiment 1), a clear shape of the distribution emerges in the narrow (low-variability) condition, with differential looking behaviour at category means, boundaries and peripheries. For VOT values near the category boundary, there is a relatively steep and steady increase in looks to the competitor object over the whole trial period. For VOT values at the distribution perimeters, there is a continued target advantage, with few looks to the competitor throughout the whole trial period. In contrast, in the wide (high-variability) condition, the distribution is flatter; in the latter part of the trial, there is a weaker effect of VOT, such that after 600 ms the distribution appears quite flat across all VOTs. This suggests that at later stages of processing, VOT is relied on less for verification of the decision in the wide condition, when VOT is a less informative cue, than in the narrow condition when it is more informative. Interestingly, while the robustness of the category may be stronger in the narrow condition, the greater number of looks to the competitor at the category boundaries in the narrow condition (compared to the wide condition) suggest that there may also be a greater cost when VOT values fall outside the expected range and compete with neighbouring categories. The pattern of effects was similar in Experiment 2, except that the overall proportion of looks to clicked target was lower. There are not the large areas of yellow indicating high proportions of fixations on the clicked target. This may reflect a generally higher level of difficulty distinguishing tone contrasts, compared to segmental contrasts.

What do the present results imply for the role of statistical information about acoustic variation in speech perception? While traditional speech perception models posit phonetic representation to be abstract and invariant (e.g. Stevens, 2002), the present results show that subtle differences in acoustic cue distributions can affect the way a particular acoustic cue is perceived. A number of recent studies have demonstrated that increased variation in the acoustic signal can actually be beneficial in discriminating speech contrasts, particularly in acquisition of cue dimensions. For example, the number of distributions (bimodal versus unimodal) influences offline categorisation responses in adults (Gulian et al., 2007; Maye & Gerken, 2000) and looking times (Liu & Kager, 2011; Maye et al., 2002, 2008) and ERPs in infants (Wanrooij et al., 2014). Beyond number of distributions, Rost and McMurray (2010)

demonstrated a crucial role for acoustic phonetic variation in infant language acquisition. In a series of experiments in which phonetic cues were either varied or held constant, 14-month-olds were able to acquire the voicing contrast only when indexical speaker cues were varied. Statistical information in VOT values themselves within the same speaker was not sufficient for learning, but variance in *non-contrastive indexical dimensions* in the multi-speaker condition enabled infants to extract the relative invariance in the contrastive VOT dimension. These findings differ with respect to the present results in that the variation in the Rost and McMurray studies was along non-contrastive dimensions. In fact, whether a cue is contrastive or non-contrastive, variation within the cue dimension seems to have the same effect: it lowers the cue weighting of the cue. This can be beneficial for the non-contrastive cues, or detrimental for contrastive cues. This is consistent with the assumption in learning models that learning involves not only acquisition of knowledge, but also learning to ignore cues that are not effective discriminators (Baayen et al., 2013). Both in the Rost and McMurray study and in the present study, experience with less informative cues led to decreased use of the cue for discrimination/identification. The present results are also consistent with effects of the degree of within-speaker variation in natural speech. When within-speaker variability in natural English fricative production is high, identification response times by native listeners are longer than responses to low-variability speakers (Newman et al., 2001).

A further piece of evidence suggesting a beneficial role of a wide acoustic distribution comes from Wanrooij, Escudero and colleagues. In two recent studies (Escudero et al., 2011; Wanrooij et al., 2013), they demonstrated that a wide distribution (which they call *enhanced*) can aid in the acquisition of L2 sound categories. This seems to pose an interesting question with respect to the present results. In both their studies, learners benefitted from the wider distribution, while in the present study, the wider distribution led to greater uncertainty. This seeming discrepancy in the results is easily explained by looking at the specific manipulation of the acoustic distributions. Although both their study and the present study used a narrow and a wide distribution, the manipulation was fundamentally different between studies. The focus of their studies was the acoustic distance between the two vowel categories, while keeping presentation frequency constant. At all points in the continuum (means, category boundaries and peripheries (which they call *endpoints*) the distance between categories was greater in the enhanced condition compared to the natural bimodal condition. That is, the vowel categories were either similar (natural bimodal condition) or very different (enhanced bimodal condition). In the present study, in contrast, acoustic distance was held constant between conditions. Category means had the same acoustic (VOT or pitch) value and were equally distant in the

two conditions, both from each other and from the category boundary. Only the presentation frequency of each token was manipulated. Further studies that disentangle the contribution of the degree of overlap, the distance between means, the degree of variance and the remoteness of the peripheries would contribute to our understanding of how variation in acoustic cues affects perception of speech sounds.

The differential patterns of looking behaviour seen at the central VOTs (i.e. the distribution peaks) in the present results (Figure 5.3) are precisely the pattern predicted by Clayards and colleagues. Clayards et al. (2008) investigated the effect of degree of VOT variation on perception of English voiced and voiceless stops. They hypothesised that the largest difference in looks to the competitor object between the narrow and wide conditions would be at the VOT values closest to the category boundaries. However, due to a smaller number of participants in their experiment and a different method of analysis the relatively few trials at the most central VOT values meant that there was insufficient power to test this prediction for all VOTs. In the present experiment, with increased power using GAMM, we were able to evaluate the fixations at these VOT values. In line with their predictions, Figure 5.3 shows the strongest effects of condition at the central VOTs with effects decreasing towards the mean. On the other hand, the effects at the periphery of the distributions are at variance with their predictions. Based on their calculations of the posterior probability functions using (1), they predicted no differences between distributions at these peripheral VOT values.

P(categoryA | stimX) = P(stimX | categoryA) / P(stimX | categoryA) + P(stimX | categoryB)
(1)

Eq (1) calculates the slope of the categorization function as a function of the degree of overlap in the distributions. However, if the assumptions of the ideal observer models are correct, participants should be making use of all available cues. This suggests that not only the degree of overlap, but also the other parts of the distribution play a role. This point will be returned to later in the discussion.

Clayards et al. (2008) also raised the question of the level of processing at which participants were tracking probabilities. In their study, participants might have been tracking, for example at the word level ('beach'-'peach'), the phoneme level (/b/-/p/), the syllable level (/bi/-/pi/) or the feature level (voiced-voiceless). The present experiment was not designed to, nor is it able to tease apart these possibilities, but the absence of an interaction between word frequency and condition suggests that the effects of the probability distributions may be operating at a

lower level than the word.

Two ways in which statistical information has been shown to affect categorisation processes are *phonetic recalibration*, in which repeated exposure to a prototypical stimulus leads to fewer items on an acoustic continuum being classified as the test category (Eimas & Corbit, 1973; Samuel, 1986; Vroomen, van Linden, Keetels, de Gelder & Bertelson, 2004; Vroomen, van Linden, de Gelder & Bertelson, 2007) and *selective adaptation*, where exposure to acoustically ambiguous stimuli combined with a disambiguating visual or lexical cue leads to *more* of the ambiguous stimuli being classified as the test category (Norris et al., 2003; Vroomen et al., 2007). These two effects have traditionally been interpreted as occurring as the result of two fundamentally different mechanisms (Vroomen et al., 2007) . However, Kleinschmidt and Jaeger (2012) have demonstrated that these effects can be modelled in terms of a Bayesian belief-updating model, in which categorisation behaviour is continually updated in response to statistical information available in the acoustic stimuli during the course of an experiment. They created conditions in which the ambiguity of the stimuli varied between unambiguous, more like those that led to recalibration (e.g. Vroomen et al., 2007) and fully ambiguous, which led to adaptation (Norris et al., 2003) with steps in between. Their results show that the degree of ambiguity can account for the resulting affects on categorisation behaviour. With presentation of ambiguous, inter-ambiguous, prototypical and inter-prototypical auditory stimuli, they modelled the crossover point at which /b/ categorisation with /d/ visual cues surpassed /b/ categorisation with /b/ visual cues. The less ambiguous the auditory stimuli, the earlier participants begin to categorise in an adaption-like way. That is, with less auditory ambiguous stimuli, participants used the relied more on the visual cues earlier.

Most models of speech perception assume that the recognition of spoken words involves the two processes of activation and competition (Luce & Pisoni, 1998; W. Marslen-Wilson, 1989; McClelland & Elman, 1986; Norris, 1994). There has been some debate over which lexical items become activated and participate in the competition process. In the Cohort model (W. D. Marslen-Wilson, 1987; W. Marslen-Wilson, 1989, 1990), word-initial information takes precedence in the activation of lexical candidates. As the signal unfolds, only lexical candidates with overlapping information become and remain activated. The minimal discrepancy hypothesis states that minimal mismatch in the incoming speech signal is sufficient to exclude candidates from the competition process.

Evidence against this account comes from eyetracking data using the visual world paradigm. Allopenna, Magnuson and Tanenhaus (1998) showed that rhyming competitors (e.g. *speaker* for target *beaker*) influ-

enced eye movements early in the recognition process, at least in English. However, this effect has not been found in Chinese. While initial overlap creates competition in Mandarin spoken word recognition, attempts to replicate the rhyme effect have previously been unsuccessful. The high rate of competitor fixations in the present study provide evidence for the rhyme effect in another Chinese language, Cantonese.

An important methodological finding of the present study is that there is not a simple one-to-one relation between where participants look and what they believe they are hearing. Although accuracy was high in the actual responses, and the early fixations were likely to fall on the target, later fixations were increasingly drawn to the competitor object over time. Tanenhaus and colleagues developed and made explicit the link hypothesis as an assumption underlying the use of the visual world paradigm to investigate spoken language processing (e.g. Tanenhaus, Magnuson, Dahan & Chambers, 2000).

> The assumption providing the link between lexical activation and eye movements is that the activation of the name of a picture determines the probability that a subject will shift attention to that picture and thus make a saccadic eye movement to fixate it... Thus the predicted fixation probability is determined both by the amount of evidence for an alternative and the amount of evidence for that alternative compared to the other possible alternatives (p. 567-569).

The present study provides further support for the second part of this statement, namely that activation levels of both the target and the competitor contribute to the likelihood of fixating the target. However, the results also suggest that we need to be careful about our interpretation regarding the exact nature of link between activation levels and likelihood of fixating the target. When acoustic evidence for the target over the competitor was low, the number of fixations on the competitor actually surpassed the number of fixations on the target, even though the target was eventually selected at the end of the trial, and even though early fixations were more likely to land on target. This suggests differential effects at different stages of processing. While accuracy is relatively high in the very early perceptual processes, later processing seems to involve a verification process, in which fixations on the competitor object are increasingly likely with increasing perceptual uncertainty. These verification fixations seem to parallel increases in response latencies seen in rejecting non-words in lexical decision tasks with increased similarity to real words. Attention is focused on the point of difficulty. It is unlikely that the competitor is actually being considered for selection at this point. Instead, these verification fixations seem to reflect a checking process before taking action to reject it.

Chapter 6

Discussion

6.1 Introduction

This thesis investigated native speakers' processing of contrastive and non-contrastive phonetic variation during speech perception, production and reading aloud. The research presented here provides evidence that several types of sub-phonemic information are processed during presentation of both auditory and visual stimuli, as well as during speech production. While most studies of phonological processing during speech production and reading aloud have taken the phoneme to be the basic processing unit, the present results show that speech production (Chapter 2) and reading aloud (Chapters 3) involve multi-level processing. That is, activation of both the speech category and the actual realisation of the context-specific variant occur. This is true whether the allophonic variants are overtly produced (Chapter 2, Experiment 1) or whether they are processed visually as ignored distractor words (Chapter 2, Experiment 2). Chapter 3 demonstrated that reading aloud involves processing of sub-phonemic feature information. Reaction times and electrophysiological measurements showed that overlap in the sub-phonemic feature voicing facilitates reading aloud. Chapter 4 investigated how phonetic context influences processing of phonetic variants. Context-specific representation of speech sounds are activated even with briefly presented masked primes (Chapter 4). As in speech production (Chapter 2), reading aloud latencies can be facilitated by cross-category primes (primes which mismatch in terms of speech category) if the context-specific realisation matches the target word. Chapter 4 also addressed how the appropriate form is selected when a speech category has more than one variant. Rapidly processed top-down information available in the surrounding phonetic context affects the relative activation of the two variants as evidenced by amplitude of the EEG signal (Chapter 4). Finally, Chapter 5 investigated one of the fundamental mechanisms by which these continuous acoustic signals become contrastive in human speech to begin with: informativity of acoustic cues. Results showed that high-noise (i.e. low-informativity, wide distribution) input leads to less reliance on acoustic cues relative to low-noise (high-informativity, narrow distribution) input.

6.2 Multi-level processing of phonology during Mandarin tone production

In this thesis, Chapter 2 addressed the question of whether processing of allophonic variants during speech production occurs at the higher level of the phonemic category or at the lower, subphonemic level of the context-specific variant. This study made use of the picture-word interference paradigm (Damian & Martin, 1999; Lupker, 1982; Rosinski et al., 1975; Schriefers et al., 1990; Starreveld et al., 1996). Most speech production models involve activation of sequences of phonemes (e.g. Dell, 1986, 1988; Indefrey & Levelt, 2004; W. J. M. Levelt et al., 1999; W. J. M. Levelt, 2001; W. J. M. Levelt et al., 1999). However, it is not clear whether phonological effects are due to abstract phoneme-level representations, or similarity in actual, instantiated (acoustic and/or motor) representation of the speech sound. A number of recent studies have investigated processing of non-canonical variants in speech processing (Bürki, Ernestus & Frauenfelder, 2010; Connine, 2004; Gaskell & Marslen-Wilson, 1996). McLennan et al. (2003) investigated processing of word-medial alveolar stops /t/ and /d/ which, in casual American English speech are free-varying allophones often produced as flaps. In a shadowing study, they found that carefully articulated variants primed production of flapped variants, and vice-versa, indicating activation of the higher-level speech category. However, the study specifically used words in which the flapped variant did not make the phoneme ambiguous between /t/ and /d/ (i.e. it did not contain word pairs such as rater and raider). Therefore it did not test whether there was cross-category facilitation from the flapped variant of /t/ to /d/ or vice versa.

In two picture-word interference experiments the phonological facilitation effect was used to investigate native Mandarin speakers' processing of phonological information in tonal variants. Recall that the tonal contour of Beijing Mandarin Tone 3 is usually low, but when followed by another Tone 3 character, it is rising, like the contour of Tone 2. This is known as third tone sandhi. Sandhi words are therefore phonologically related to both Tone 3 and Tone 2 words. They overlap with Tone 3 words in terms of the Tone 3 category (i.e. the toneme), but the actual realisation

of the tonal contour is different (rising versus low). Sandhi words are also phonologically related to Tone 2 in that they have the same, rising contour, even though they belong to different tone categories. The question addressed in Chapter 2 was which of these two types of phonological relatedness is important during speech production? Does speech production involve retrieval of speech categories? Or is it activation of the actual acoustic realisation (such as the tonal contour) that is important in speech production?

The best-fit LME (linear mixed effects) model (Baayen, 2008; Baayen et al., 2008) revealed that production of T3 sandhi picture names was significantly faster when distractor and target picture matched in tone category, but had different overt realisations (i.e. Tone 3 distractors, the toneme condition), and when target and distractor matched in overt realisation, but mismatched in tone category (Tone 2 distractors, the contour condition), compared to control distractors, which mismatched the target in both the toneme and the contour. These two types of facilitatory effects indicate that speech production involves multilevel phonological processing. More specifically, production of allophonic tone variants activates both the tone category and the actual context-specific tonal contour.

The finding of any speech category effect is particularly interesting in Chinese, since phonology is not directly represented in the script. Even more so in the case of tone: although many Chinese characters contain hints about the pronunciation of the segmental syllable (the non-tonal part), there is no representation of tonal information in the orthography, at all. Therefore, the finding that activation of the tone category from one word facilitates speech production of an otherwise completely unrelated other word provides strong evidence for generalisation of tone categories in Mandarin. Although many studies have assumed or investigated the importance of speech categories (such as phonemes) in language processing, a question that is often overlooked is whether such speech categories are language-general or whether they occur as a result of the specific orthography and education system within which they are acquired. Alphabetic languages use an inherently phonetic system of representing words. Therefore, since

phonology is confounded with orthography in these languages, it is difficult to generalise findings to other languages. The findings presented here provide one piece of evidence for processing of speech categories that are not represented orthographically.

On the other hand, the cross-category phonological facilitation of sandhi picture naming from Tone 2 distractor words demonstrates that phonology is not simply processed in terms of abstract phoneme categories. Similarities in the tonal contour only were sufficient to reduce naming latencies, even though target and distractor belonged to separate speech categories. This finding poses a challenge to theories of speech production that view phonological representations as abstract, phoneme-sized units. In terms of phonemes, there was no overlap between the Tone 2 distractors and the sandhi targets. The facilitation must have occurred due to similarities in either the acoustic-phonetic properties or the motor commands used to produce the sounds, or both.

6.3 Processing of phonological and tonal information in visually presented words

A second question addressed in Chapter 2 concerned processing of Mandarin tones in visually presented words containing allophonic variants. As mentioned above, visual processing of tone in Chinese is an interesting subject, since tone is not represented in the script. In fact, whether phonological processing is necessary at all in Chinese reading has been a matter of debate in the literature. Unlike in alphabetic languages, which use a phonetic system to encode the sounds of words, semantic information is encoded directly in the characters in Chinese. Therefore, many early accounts of Chinese reading suggested that semantic processing proceeds directly from the orthography, bypassing phonological information altogether (Barren, 1978; Biederman & Tsao, 1979; Coltheart, 1978; Smith, 1985; W. S.-Y. Wang, 1973). More recent research has established that early automatic activation of phonology does occur in Chinese character processing (Perfetti & Zhang, 1995; Spinks, Liu, Perfetti & Tan, 2000). However, most research on phonological processing in Chinese reading and visual word processing has focused on segmental information. Little is known about how tones are processed. To the best of our know-

ledge, this is the first study to investigate phonological processing in visually presented words containing tonal allophones.

The experimental set up was similar to that described above for the investigation of tone sandhi processing during speech production. However, target and distractor conditions were reversed: target pictures were of objects with Tone 2 or Tone 3 names and distractors were sandhi words. Data were analysed using linear mixed effects regression modelling, in combination with Bayesian modelling. Results of the models showed that, although participants were instructed to ignore the distractor words, visual processing of sandhi words (allophonic variants of Tone 3) facilitated production of Tone 3 words. This was even though the actual realisation of the tonal contour is different between sandhi distractor words (rising contour) and Tone 3 targets (low contour). Since the tonal contours were different, the facilitation effect found here must have occurred at the level of the tone category. It is interesting that, even though tones are not represented in the orthography, they are still processed as speech categories. Since there is no orthographic representation, the speech category must have been formed through regular association between the pitch contour and the meanings (and orthography) of particular characters. The question of how speech categories are formed and, in particular, the role of statistical acoustic information was addressed in Chapter 5 and is returned to below.

In addition to activation of the tone categories, the study also provided evidence for sub-phonemic processing of phonological information in visually presented Chinese words. When sandhi distractors were superimposed on Tone 2 target pictures, naming was faster than with unrelated distractors. Note that the target and distractor belong to different tone categories (Tone 3 versus Tone 2). Therefore, the facilitation effect must be due to acoustic-phonetic and/or motor movement similarity in the actual (rising) tonal contour itself. This poses a challenge to theories that posit speech production to involve activation of series of abstract phonemic units. Activation of an internal instantiated representation during visual processing of words is consistent with models that posit involvement of the sensori-motor system in phonological processing and studies showing that auditory

and somatosensory feedback are utilised in guiding and adjusting speech production (Davis & Johnsrude, 2007; Guenther et al., 2006; Guenther & Vladusich, 2009; Houde & Jordan, 1998; Jones & Munhall, 2002; Liberman & Whalen, 2000; Purcell & Munhall, 2006). In summary, as we have already seen for overt production of allophonic variants, visual processing of allophonic variants also involves multi-level phonological processing: both the speech category and the context-specific variant are activated.

6.4 Phonological processing during reading aloud

As we have seen, overt speech production and visual processing of Mandarin tones involve both category-level and sub-phonemic processing. Chapters 3 and 4 investigated sub-phonemic processing of tonal and segmental information in a different task. Very few studies have investigated sub-phonemic processing during reading aloud. Facilitation has been found in reading aloud in alphabetic languages, such as Dutch and English, when targets and primes have the same onset phonemes, compared to those whose onset phonemes differ (Kinoshita, 2000; Kinoshita & Woollams, 2002; Timmer & Schiller, 2012; Schiller, 2004, 2007). This may not be surprising, given that these languages use a phonemic system to represent phonology. However, as shown in Chapter 2 and described above, language processing also involves processing of sub-phonemic detail, at least during speech production. In two EEG studies with masked priming, Chapters 3 and 4 investigated two types of sub-phonemic processing during reading aloud. The first question addressed whether reading aloud involves processing of sub-phonemic features in a typologically different language, Dutch.

Sub-phonemic feature processing

The first question addressed in Chapter 3 was whether and when sub-phonemic features are processed in Dutch reading aloud. Evidence for featural representations comes from a variety of sources. As early as the 1950s in a consonant identification study, Miller and Nicely (1955) suggested that speech perception may involve multiple features. Speech error studies show that substi-

tution of phonemes that differ in only one feature is more likely than phonemes that differ in more than one feature (Goldrick & Blumstein, 2006; McMillan & Corley, 2010). Some models of speech production include a feature level. For example, in Dell (1986) model, features are activated after word retrieval prior to articulation. Phonetic features also have been found to play a role during speech perception and acquisition (Chládková, 2014) and silent reading (Ashby et al., 2009). So, far the question of whether features are processed during reading aloud has not been investigated. Further to the question of whether or not features play a role in reading aloud, a matter of debate in the literature concerns the type of information features represent. One possibility is that features are relatively abstract contrastive representations (Chomsky & Halle, 1968; Dell, 1986). Alternatively, they may consist of articulatory gestures (e.g., Goldstein et al., 2007).

In Dutch, the sound pairs t-d and p-b are produced at the same place of articulation (alveolar and bilabial, respectively), while the pairs t-p and d-b match in voicing (voiceless and voiced, respectively). In this ERP study, participants read aloud real Dutch words (e.g. huid 'skin') from a computer screen. Each target word was preceded by a brief presentation of a masked non-word prime in which the final sound matched in voicing (huib), place of articulation (huit) or mismatched in both voicing and place (control condition, huip). The best-fit linear mixed effects regression model revealed that reaction times were significantly faster when prime and target matched in voicing, than when they did not. Consistent with this, there was also reduced negativity in the voice-match condition, compared to the control condition in the early time window 25-75 ms after presentation of the target word.

These results indicate rapid processing of sub-phonemic voicing information in Dutch reading aloud. This cannot be due to orthographic or phonemic processes, since there was no difference between the critical and control conditions in terms of either letters or phonemes: each prime-target pair differed by exactly one phoneme and one letter. Only when measured at the sub-phonemic feature level was there greater overlap in congruent prime-target pairs (voice and place conditions), compared to controls. Both the ERP measures and reaction times provide evid-

ence for processing of sub-phonemic voicing information in reading aloud. This finding challenges previous assumptions in models of reading aloud that phonological processing simply involves activation of strings of phonemic units.

Processing of allophonic variants

In addition to sub-phonemic feature processing, the voice-congruency effect presented in Chapter 3 and described above also sheds light on the processing of allophonic variation. In Dutch, voiced stops have two realisations: a voiced and a voiceless allophonic variant. In word-initial position, voiced stops (e.g. /d/ and /b/) are distinguished from their voiceless counterparts (/t/ and /p/) primarily by voice onset time (VOT). But in word-final position, the VOT values of voiced and voiceless stops are very similar. For example, the words hout ('garden') and houd ('to hold') are homophones in Dutch. The voiced sounds are described as devoiced (e.g. [t], [p]). When Dutch listeners were asked to distinguish between voiceless-devoiced minimal pair words they performed at chance level (Baumann, 1995). This study investigated the question of whether, when a sound category has more than one output pattern (i.e. target distribution, or allophone), the two or more distinct outputs are processed as a single category or as separate categories.

As described above, response latencies were shorter and the amplitude of the EEG was reduced with voice-congruent primes, compared to mismatching control primes. This is a particularly interesting result, given that final stops are devoiced in Dutch. Articulatorily, due to final devoicing, both prime types are voiceless (and therefore 'match' in voicing). However, the voice-congruency effect indicates that the voicing distinction is retained and processed during reading aloud. This suggests that, although the overt realisation is similar, voiceless and devoiced stops are processed as separate categories. This is consistent with the data presented in Chapter 2 that processing of speech variants during speech production and processing of visual words activates both the speech category and the context-specific allophonic variant. In the present study, the experiment was not designed to test for activation of the context-specific allophonic variant, but

it does provide evidence that the voicing distinction is processed during reading aloud, even for devoiced variants. This seems to provide support for a fairly abstract representation for features (e.g., Chomsky & Halle, 1968; Dell, 1986). However, the results do not rule out processing at the articulatory level. The present results could also be explained if multi-level processing of the type seen in Chapter 2 occurs. There may be processing of both a contrastive feature category (voiced-voiceless) and the context-specific articulatory gesture. More work is needed to verify this possibility.

6.5 Context effects on processing of speech variants

In the previous section, we saw that reading aloud Dutch segmental allophones was facilitated by congruency at the feature level, despite similarity at the articulatory level in both match and mismatch conditions. In other words, distinctive feature categories are retained for voiced-voiceless pairs despite final devoicing. In Chapter 2 we saw that production and visual processing of allophonic variants involves multi-level processing. Although these studies inform the question of what type of information is activated, they did not directly test how phonetic context affects processing of allophonic variants. Chapter 4 examined how the tonal context of a following character affects neural processing of tonal variants during Mandarin reading aloud.

Effects of phonetic context are well attested in speech perception. For example, Mann and Repp (1980) showed that an ambiguous target syllable /da/-/ga/ is perceived differently depending on the preceding context, /ar/ versus /al/. Numerous studies have demonstrated that various contextual cues affect perception of speech categories (Creel, Aslin & Tanenhaus, 2012; Kraljic et al., 2008; Toscano & McMurray, 2012). Evidence from laboratory-induced speech errors also provides evidence for contextual effects in the form of phonotactic constraints in speech production (Goldrick & Larson, 2008). In addition, Chapter 2 showed that production of Beijing Mandarin tone sandhi (tonal allophones) involves activation of an instantiation of the actual, context-specific speech sound. The PWI study showed that speech production can be facilitated by activation of another speech category that (due

to context) has a similar realisation. That is, similarities in the acoustic properties of a prime can facilitate production of a target word, even if there is no category overlap between prime and target. Chapter 4 investigated whether this is also true for reading aloud. It also extended the study in two ways. Firstly, it used briefly presented masked primes, so that processing would occur below the level of awareness. Secondly, it examined the question of whether, when a speech category has two (or more) phonetic variants, top-down contextual information constrains the relative activation of the alternative variants.

An electroencephalogram and reaction times were recorded as participants read aloud two-character Mandarin words, preceded by masked primes. The initial character of critical primes was always Tone 3, so primes always differed from targets in terms of tone category, but either matched or mismatched the tone contour. In addition, the initial character of primes was identical between conditions. Only the phonetic context provided by the tone of the following prime differed between conditions. Therefore, any differences found between conditions must be due to the context-specific processing of the tonal allophone.

Although there were numerical differences in reaction times, the best-fit linear mixed effects regression model found that the differences were not significant. However, in the EEG data, modelled using GAMMs (Wood, 2006), significant differences were found depending on the phonetic context provided by the tone of the following character. The effect was modulated by prime and target frequency. This indicates, firstly, that the acoustic similarity in the congruent prime affects processing of the target word, even though prime and target belong to different tone categories. Secondly, it indicates that this phonetic information is context dependent. Since initial characters were identical between conditions, this suggests that the top-down processing of the surrounding phonetic context promotes activation of the appropriate allophonic variant.

Traditional methods of ERP analysis often average over trials for each experimental condition. In particular, item information is collapsed, so that only by-subject (F1) and no by-item (F2, nor F' or min F') analysis is possible. Between-item variation is an

important consideration. Just as participants are sampled from the wider population (e.g. of speakers of a particular language), linguistic items are typically sampled from a larger population of possible items relating to the experimental question (e.g. English nouns in an English picture-naming experiment). That is the experimental items do not exhaust all examples available in the language. Therefore, just as there are faster and slower participants, particular characteristics of linguistic items (such as frequency) may make them easier or more difficult in a particular task. This is demonstrated by the 'language-as-fixed-effect-fallacy' (Clark, 1973; Coleman, 1964) and testing for item random effects is now widely adopted in behavioural psycholinguistic research. However, within EEG research, this problem has largely been ignored, and analysis is typically done with no examination of item variation.

The present study addressed this problem by analysing data using Generalised Additive Mixed Modelling (GAMM) in R. Full random effects structure for subjects and items, as well as word-specific frequency properties were included in the model. The finding that prime and target word frequency influence neural activity and interact with other effects suggests that it is useful to include items as random effects in ERP studies of language processing.

6.6 Phonetic variation and acoustic cue informativity

So far we have seen evidence for multi-level processing of acoustic regularities across several domains of language processing. Although speakers seem able to process these regularities, acoustics provide an inherently noisy medium for communication. Each acoustic dimension recruited in human languages for contrasting (word) meanings is continuous. There are no particular defined acoustic cue values associated with any given speech sound, but rather values occur relative to contrasting sounds. The voice onset time of /b/, for instance, is short relative to /p/, the pitch (fundamental frequency) of a Cantonese high tone is high relative to the mid tone, which is high relative to the low tone. But the actual value of any given speech sound depends on many factors, such as speech rate, phonetic context, the voice of the speaker and so on. Considering that phonetic variation (i.e. noise) is a fundamental

property of the speech signal, relatively little is known about how it affects online speech processing. A number of recent studies have investigated various aspects of the statistical distributions of the input. The majority of these studies have investigated whether the *number of distributions* (unimodal versus bimodal) affects categorisation judgments (Gulian et al., 2007; Maye & Gerken, 2000; Maye et al., 2008), infant looking times (Liu & Kager, 2011; Maye et al., 2002) or ERPs (Wanrooij et al., 2014). Other studies have investigated effects of training with increased or reduced acoustic distance in second-language acquisition(e.g. Escudero et al., 2011; Wanrooij et al., 2013). Very little is known about how the *shape of statistical distributions* in the input influences perception of speech contrasts.

Chapter 5 investigated the effects of variability of acoustic cues on native Cantonese listeners' processing of speech sound contrasts. Results showed that the informativity of acoustic cues has immediate consequences on the extent that that cue is relied on during speech perception.

Eye movements were recorded as participants heard acoustic stimuli that contained either a relatively large amount variation (the wide distribution condition) or relatively little variation (the narrow distribution condition) and saw pictures of word pairs consisting of aspirated and unaspirated counterparts (Experiment 1) or mid- and high-tone counterparts (Experiment 2). We hypothesised that greater variation in the signal would lead to greater uncertainty in processing of the speech contrasts. The best-fit generalised additive mixed model (GAMM) revealed that the proportion of fixations on the clicked object over the course of the trial varied as a function of distribution condition (narrow versus wide) and VOT or pitch value (location on the 12-step continuum), and that VOT/pitch value significantly interacted with distribution condition. In the narrow (low-variability) condition, a clear shape of the distribution emerged, with differential looking behaviour at category means, boundaries and peripheries. In contrast, in the wide (high-variability) condition, the distribution was flatter, particularly in the latter part of the trial. In the wide condition, the effect of VOT/pitch was weak, so that after 600 ms the distribution appeared quite flat across all cue values. The pattern

of looking behaviour suggests that there is a change in the stage of processing over the course of the trial. Interestingly, the early fixations were very likely to fall on the clicked object in both distribution conditions. However, the distribution condition seemed to come into play most strongly in later stages of processing. This suggests that variability in the signal has the strongest influence on the process of verification. With high variability, more looks to the competitor object are necessary in order to reject it in favour of the clicked target object. At later stages of processing, the VOT or pitch cue is relied on less for verification of the decision in the wide condition, when it is a less informative cue, than in the narrow condition when it is more informative.

These results show that subtle differences in acoustic cue distributions can affect the way a particular acoustic cue is perceived and utilised in processing of speech contrasts. It is well documented that individual listeners attend to different acoustic cues. For example, adult second-language (L2) learners often have trouble distinguishing certain L2 speech contrasts. Yet, the question of how listeners come to utilise certain cues and not others is not yet well understood. The finding that acoustic cue informativity influences the degree to which a cue can be utilised for discrimination can inform our understanding of this process. These effects were found with adult participants in the short period of a laboratory experiment. This demonstrates that learning is rapid and on-going throughout the lifetime. The degree to which an acoustic cue is utilised during speech perception is updated depend on its effectiveness in discriminating between alternative messages.

References

Agus, T. R., Thorpe, S. J. & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron*, *66*(4), 610–618.

Allopenna, P. D., Magnuson, J. S. & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38*(4), 419–439.

Ashby, J., Sanders, L. D. & Kingston, J. (2009). Skilled readers begin processing sub-phonemic features by 80 ms during visual word recognition: Evidence from ERPs. *Biological psychology*, *80*(1), 84–94.

Aslin, R. N., Saffran, J. R. & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, *9*(4), 321–324.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., Hendrix, P. & Ramscar, M. (2013). *Sidestepping the Combinatorial Explosion: An Explanation of n-gram Frequency Effects Based on Naive Discriminative Learning*. Language and Speech.

Baayen, R. H. & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Baayen, R. H., Piepenbrock, R. & van Rijn, H. (1993). *The CELEX lexical database*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. (2013). Random effects

structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Barren, R. W. (1978). Access to the meaning of printed words: Some implications for reading and learning to read. In F. B. Murray (Ed.), *The recognition of words: IRA series on the development of the reading process* (pp. 34–56). Newark, DE: International Reading Association.

Bates, D., Maechler, M. & Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes*. R package version 0. Retrieved from `http://CRAN.R-project.org/package=lme4`

Baumann, M. (1995). *The production of syllables in connected speech*. Unpublished Ph. D.

Biederman, I. & Tsao, Y. C. (1979). On processing Chinese ideo- graphs and English words: Some implications from Stroop-test results. *Cognitive Psychology*, *11*, 125–132.

Boersma, P. & Weenink, D. (2012). *Praat* (Vol. 5.).

Bonte, M., Parviainen, T., Hytönen, K. & Salmelin, R. (2006). Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex*, *16*(1), 115–123.

Brady, T. F., Konkle, T., Alvarez, G. A. & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325.

Bürki, A., Ernestus, M. & Frauenfelder, U. H. (2010). Is there only one "fenêtre"in the production lexicon? On-line evidence on the nature of phonological representations of pronunciation variants for French schwa words. *Journal of Memory and Language*, *62*(4), 421–437.

Cai, Q. & Brysbaert, M. (2010). Subtlex-ch: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*(6), e10729.

Carreiras, M., Perea, M., Vergara, M. & Pollatsek, A. (2009). The time course of orthography and phonology: ERP correlates of masked priming effects in Spanish. *Psychophysiology*, *46*(5), 1113–1122.

Chen, J. Y., Chen, T. M. & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language*, *46*.

Chen, Y., Shen, R. & Schiller, N. O. (2011). *Representation of allophonic tone sandhi variants*. Proceedings of Psycholinguistics Representation of Tone.

Chládková, K. (2014). *Finding phonological features in perception*.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, *12*(4), 335–359.

Clayards, M., Tanenhaus, M. K., Aslin, R. N. & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, *14*(1), 219–226.

Coltheart, M. (1977). *Access to the ignternal lexicon.* The psychology of reading.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151–216). New York: Academic Press.

Connine, C. M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin and Review*, *11*.

Costa, A. & Caramazza, A. (2002). The production of noun phrases in English and Spanish: Implications for the scope of phonological encoding in speech production. *Journal of Memory and Language*, *46*(1), 178–198.

Creel, S. C., Aslin, R. N. & Tanenhaus, M. K. (2012). Word learning under adverse listening conditions: Context-specific recognition. *Language and Cognitive Processes*, 1021–1038.

Cristià, A., McGuire, G. L., Seidl, A. & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of phonetics*, *39*(3), 388–402.

Da, J. (2004). A corpus-based study of?character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction: Studies on the theory and methodology of digitalized Chinese teaching to foreigners. In P. Zhang, T. Xie & J. Xu (Eds.), *Proceedings of the fourth international conference on new technologies in teaching and learning chinese* (pp. 501–511). Beijing: Tsinghua University Press.

Dahan, D., Drucker, S. J. & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? Cognition, 108(3). *710-718. doi:*, *10.*.

Damian, M. F. & Martin, R. C. (1999). Semantic and phonological codes interact in single word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 345–361.

Davis, M. H. & Johnsrude, I. S. (2007). *Hearing speech sounds: Top-down influences on the interface between audition and speech perception.* Hearing Research, 229(12), 132-147. doi:10.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3).

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2).

Eimas, P. D. & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*(1), 99–109.

Ernestus, M. & Baayen, R. H. (2003). *Predicting the unpredictable: Interpreting neutralized segments in Dutch.* Language, 79:5–38.

Ernestus, M. & Baayen, R. H. (2004). *Analogical eects in regular past tense production in Dutch.* Linguistics, 42:873–903.

Escudero, P., Benders, T. & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America*, *130*(4).

Feldman, N., Myers, E., White, K., Griffiths, T. & Morgan, J. (2011). Learners use word-level statistics in phonetic category acquisition. In N. Danis et al. (Eds.), *Proceedings of the 35th annual Boston University Conference on Language Development* (pp. 197–209).

Ferrand, L. & Grainger, J. (1992). Phonology and orthography in visual word recognition: Evidence from masked non-word priming. *Quarterly Journal of Experimental Psychology: Section A*, *45*(3), 353–372.

Ferrand, L. & Grainger, J. (1993). The time course of orthographic and phonological code activation in the early phases of visual word recognition. *Bulletin of the psychonomic society*, *31*(2), 119–122.

Ferrand, L. & Grainger, J. (1994). Effects of orthography are independent of phonology in masked form priming. *The Quarterly Journal of Experimental Psychology*, *47*(2), 365–382.

Forster, K. I. & Davis, C. (1991). The density constraint on form- priming in a naming task: interference effects from a masked prime. *Journal of memory and language*, *30*.

Foss, D. J. & Swinney, D. A. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, *12*(3).

Fowler, C. A. (2010). The reality of phonological forms: a reply to Port. *Language sciences*, *32*(1), 56–59.

Ganushchak, L. Y., Christoffels, I. K. & Schiller, N. O. (2011). The use of electroencephalography in language production research: a review. *Frontiers in psychology*, *2.*.

Gaskell, M. G. & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human perception and performance*, *22*(1), 144–158.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Goldrick, M. & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*(6), 649–683.

Goldrick, M. & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition*, *107*(3), 1155–1164.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E. & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, *103*(3), 386–412.

Grainger, J., Kiyonaga, K. & Holcomb, P. J. (2006). The time course of orthographic and phonological code activation. *Psychological Science*, *17*(12), 1021–1026.

Gratton, G., Coles, M. G. H. & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*.

Guenther, F. H., Ghosh, S. S. & Tourville, J. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, *96*(3), 280-301.

Guenther, F. H. & Vladusich, T. (2009). A neural theory of speech acquisition and production. *Journal of neurolinguistics*, *25*(5).

Gulian, M., Escudero, P. & Boersma, P. (2007). *Supervision hampers distributional learning of vowel contrasts.* Proceedings of the international congress of phonetic sciences.

Hintzman, D. L. (1986). ÒschemaabstractionÓin a multiple-trace memory model. *Psychological Review*, *93*(4).

Houde, J. F. & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354).

Indefrey, P. & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92(1-2)*, 101-144.

Jones, J. A. & Munhall, K. G. (2002). The role of auditory feedback during phonation: studies of Mandarin tone production. *Journal of Phonetics*, *30*(3).

Jongman, A., Wayland, R. & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263.

Ju, M. & Luce, P. A. (2006). Representational specificity of within-category phonetic variation in the long-term mental lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 120–138.

Kinoshita, S. (2000). The left-to-right nature of the masked onset priming effect in naming. *Psychonomic Bulletin & Review*, *7*(1), 133–141.

Kinoshita, S. & Woollams, A. (2002). The masked onset priming effect in naming: Computation of phonology or speech planning? *Memory & Cognition*, *30*(2), 237–245.

Kleinschmidt, D. & Jaeger, T. F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. In *Proceedings of the 34th*

annual meeting of the Cognitive Science Society (CogSci12) (pp. 107–115).

Kraljic, T. & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive psychology*, *51*(2), 141–178.

Kraljic, T. & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Kraljic, T. & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, *121*(3), 459–465.

Kraljic, T., Samuel, A. G. & Brennan, S. E. (2008). First impressions and last resorts how listeners adjust to speaker variability. *Psychological science*, *19*(4), 332–338.

Levelt, W. J., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T. & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological review*, *98*(1), 122–142.

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, *98*(23), 13464.

Levelt, W. J. M., Roelofs, A. & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*. Retrieved from `http://journals.cambridge.org/article_S0140525X99001776`

Liberman, A. & Whalen, D. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5).

Lisker, L. & Abramson, A. S. (1964). *A cross-language study of voicing in initial stops: acoustical measurements.* Word 20.

Liu, L. & Kager, R. (2011). How do statistical learning and perceptual reorganization alter dutch infant's perception to lexical tones? In *Icphs* (Vol. 17, pp. 1270–1273).

Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, *19*(1).

Lupker, S. J. (1982). The role of phonetic and orthographic similarity in picture-word interference. *Canadian Journal of Psychology*, *36*, 349–367.

Mann, V. A. & Repp, B. H. (1980). *Influence of vocalic context on perception of the [S]-[s] distinction.* Perception and Psychophysics, 28(3):213–228.

Marian, V., Bartolotti, J., Chabal, S. & Shook, A. (2012). Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS ONE*.

Marslen-Wilson, W. (1989). Lexical representation and process. In W. Marslen-Wilson (Ed.), (pp. 3–24).

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In *Cognitive models of speech processing* (pp.

148–172).
Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1), 71–102.
Maye, J. & Gerken, L. (2000). *Learning phonemes without minimal pairs*. Proceedings of the 24th Annual Boston University Conference on Language Development.
Maye, J., Weiss, D. & Aslin, R. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1).
Maye, J., Werker, J. F. & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3).
McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1–86.
McLennan, C. T., Luce, P. A. & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4).
McLennan, C. T., Luce, P. A. & Charles-Luce, J. (2005). Representation of lexical form: Evidence from studies of sublexical ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1308–1314.
McMillan, C. T. & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition*, *117*(3), 243–260.
McMurray, B., Aslin, R. N. & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*(3), 369–378.
McMurray, B. & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological review*, *118*(2), 219-246.
McQueen, J. M., Cutler, A. & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.
Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language*, *29*(5).
Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language*, *30*(1), 69–89.
Meyer, A. S. & Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(6), 1146-1160.

Miller, G. A. & Nicely, P. E. (1955). *An analysis of perceptual confusions among some english consonants.* Journal of the Acoustical Society of America, 27(2):338–352.

Mitterer, H. (2006). Is vowel normalization independent of lexical processing? Phonetica, 63(4). *209-229. doi:10*, 1159/000097306.

Mitterer, H., Chen, Y. & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science*, *35*(1).

Mousikou, B., Roon, K. & Rastle, K. (2014). Masked primes activate feature representations in reading aloud. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Mousikou, P., Coltheart, M., Finkbeiner, M. & Saunders, S. (2010). Can the dual-route cascaded computational model of reading offer a valid account of the masked onset priming effect? *The Quarterly Journal of Experimental Psychology*, *63*(5), 984–1003.

Newman, R. S., Clouse, S. A. & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3).

Newport, E. L. & Aslin, R. N. (2004). Learning at a distance I. *Statistical learning of non-adjacent dependencies. Cognitive psychology*, *48*(2), 127–162.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142.

Nixon, J. S., Chen, Y. & Schiller, N. O. (2014). Multi-level processing of phonetic variants in speech production and visual word processing: evidence from mandarin lexical tones. *Language, Cognition and Neuroscience.*, *10.*.

Nixon, J. S., Timmer, K., Linke, K., Schiller, N. O. & Chen, Y. (submitted). Early negativity reveals rapid sub-phonemic processing during reading aloud.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H. & Chen, Y. (submitted). Eye movements reflect acoustic cue informativity and statistical noise.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Norris, D., McQueen, J. M. & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, *47*(2), 204–238.

Pajak, B. (2012). *Inductive Inference in Non-Native Speech Processing and Learning (Doctoral dissertation, University of California.* San Diego).

Pelucchi, B., Hay, J. F. & Saffran, J. R. (2009a). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244–247.

Pelucchi, B., Hay, J. F. & Saffran, J. R. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child development*, *80*(3), 674–685.

Peng, S. H. (2000). *(2000)* (L. versus 'phonological' representations of Mandarin sandhi tones. In M. B. Broe & J. B. Pierrehumbert, Eds.). Acquisition and the lexicon: Papers in Laboratory Phonology V.

Perea, M. & Lupker, S. J. (2003). Transposed-letter confusability effects in masked form priming. In S. Kinoshita & S. J. Lupker (Eds.), *Masked priming: State of the art* (pp. 97–120). Hove, U. K.: Psychology Press.

Perfetti, C. A. & Zhang, S. (1995). Very early phonological activation in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 24–33.

Purcell, D. W. & Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America*, *119*.

R Core Team. (2013). *R: A language and environment for statistical computing*. Retrieved from `http://www.R-project.org/`

Ramscar, M. & Baayen, R. H. (2013). Production, comprehension, and synthesis: a communicative perspective on language. *Frontiers in psychology*, *4*..

Ramscar, M., Dye, M. & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological science*, *24*(6), 1017–1023.

Reinisch, E., Wozny, D. R., Mitterer, H. & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics*, *45*, 91–105.

Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes*, *14*(2), 1080/016909699386338.

Rosinski, R. R., Golinkoff, R. M. & Kukish, K. S. (1975). *Automatic semantic processing in a picture-word interference task*. Child Development.

Rost, G. C. & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

Rost, G. C. & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635.

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294).

Samuel, A. G. (1986). The role of the lexicon in speech perception. *Pattern recognition by humans and machines: Speech perception*,

1, 89–112.

Schiller, N. O. (2004). The onset effect in word naming. *Journal of memory and language*, *50*.

Schiller, N. O. (2007). Phonology and orthography in reading aloud. *Psychonomic bulletin and review*, *14*(3), 460–465.

Schriefers, H., Meyer, A. S. & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of memory and language*, *29*(1), 86–102.

Severens, E., Lommel, S. V., Ratinckx, E. & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta psychologica*, *119*(2), 159–187.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*.

Smith, F. (1985). *Reading without nonsense (2nd ed.)*. New York: Teachers College Press.

Spinks, J. A., Liu, Y., Perfetti, C. A. & Tan, L. H. (2000). Reading Chinese characters for meaning: The role of phonological information. *Cognition*, *76*(1).

Starreveld, P. A., Heij, L. & W. (1996). Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 896–918.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872-1891.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D. & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, *29*(6), 557–580.

Timmer, K. & Schiller, N. O. (2012). The role of orthography and phonology in English: An ERP study on first and second language reading aloud. *Brain research*, *1483*, 39–53.

Timmer, K., Vahid-Gharavi, N. & Schiller, N. O. (2012). *Reading aloud in Persian: ERP evidence for an early locus of the masked onset priming effect*. Brain and language.

Toscano, J. C. & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception and Psychophysics*, *74*(6), 1284–1301.

Tremblay, A. (2013a). erp: Pre-processing and visualization of event-related brain potential and field (ERP/erf) data. *R package version 0.9. 8*, *8.*(11).

Tremblay, A. (2013b). *icaOcularCorrection: Independent Components Analysis (ICA) based artifact correction.* Retrieved from `http://CRAN.R-project.org/package=icaOcularCorrection`

Trude, A. M. & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, 979–1001.

Van Rij, J. (2014). compareML. *R package version 2.0*.

Vroomen, J., van Linden, S., de Gelder, B. & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–577.

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B. & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, *44*(1), 55–61.

Wang, W. S.-Y. (1973). (1973). *The Chinese language. Scientific American*, *228*, 50–60.

Wang, W. S.-Y. & Li, K. P. (1967). Tone 3 in pekinese. *Journal of Speech and Hearing Research*, *10*, 629–636.

Wang, Y. C. (2013). jtrans: Johnson transformation for normality. *R package version 1.0*. Retrieved from `http://CRAN.R-project.org/package=jtrans`

Wanrooij, K., Boersma, P. & van Zuijen, T. L. (2014). Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. *Frontiers in psychology*, *5*..

Wanrooij, K., Escudero, P. & Raijmakers, M. E. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. *Journal of Phonetics*, *41*(5), 307–319.

Warner, N., Jongman, A., Sereno, J. & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of Phonetics*, *32*(2), 251–276.

Wong, A. W. K. & Chen, H. C. (2008). Processing segmental and prosodic information in Cantonese word production. Journal of Experimental Psychology: Learning, Memory and Cognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*, 1172–1190.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Yuan, J. H. & Chen, Y. (2014). 3rd tone sandhi in Standard Chinese: A corpus approach. *Journal of Chinese Linguistics*, *42*.

Zhao, Y. (2010). *Statistical inference in the learning of novel phonetic categories*. (Unpublished doctoral dissertation).

Zhou, X. L. & Zhuang, J. (2000). Lexical tone in the speech production of Chinese words. *Stroke 9. 8. 8*, *8.*(8), 9–5. Retrieved from `http://www.isca-speech.org/archive/icslp_2000/i00_`

Summary

This thesis presents evidence obtained through a variety of empirical and statistical methods to investigate how healthy adult native speakers process the inherently noisy acoustic information that spoken communication consists of. While psycholinguistic models (especially of speech production and reading aloud) have generally measured phonological processing in terms of sequences of phonemes, this thesis has demonstrated several types of subphonemic processing in a variety of tasks. One of the main focuses of this dissertation is how speakers and listeners process phonetic variation. Importantly, a distinction is made between informative, regular variation, as in allophonic variants, and, in contrast, random noise, which reduces the informativity of acoustic cues. In addition, we also show that, even in a language where no information about speech contrasts is represented in the orthographic script, both overt production and visual processing of written words involve processing of contrastive speech sound categories.

Chapter 2 made use of the picture-word interference paradigm to investigate how tonal variants are processed during Mandarin speech production. In Beijing Mandarin, when Tone 3 is followed by another Tone 3 character (third tone sandhi), it sounds like Tone 2. This aspect of the tonal system means that sandhi words phonologically related to both Tone 3 and Tone 2 words, which allowed us to manipulate two kinds of phonological information: the phonological tone category and the acoustic tonal contour. Chapter 2 investigated which of these two types of phonological relatedness is important during speech production. While there

is a wealth of psycholinguistic research showing that processing is facilitated by pre-activation of congruent phonological information, the exact nature of this phonological information is not yet clear. Does speech production involve retrieval of the speech sound category? Or is it activation of the actual acoustic realisation that is important? Chapter 2 (see also Nixon et al., 2014) showed that speech production involves multilevel phonological processing. Production of sandhi picture names was significantly faster when distractor and target picture matched in tone category, but had different overt realisations and when target and distractor matched in overt realisation, but mismatched in tone category (compared to control distractors, which mismatched the target in both the toneme and the contour). Interestingly, there were differences in the time course of effects. The tonal contour facilitated production both when presented simultaneously with the target picture and when it was delayed by 83 milliseconds. The tone category, in contrast, only facilitated production with simultaneous presentation. This suggests that the speech category is activated early (perhaps during word retrieval), but that in later stages of processing (during speech preparation, for instance), the actual acoustic realisation becomes more important. There are two possible explanations for this pattern of results. It may be that the realisation contour remains activated for longer than the tone category during overt production. Another possibility is that, while both the contour and the category remain activated, as the task shifts from lexical retrieval to articulation preparation, only the articulatory/acoustic congruency benefits production. Further research is required to tease these two possibilities apart. A second question raised by the results of Experiment 1, is whether processing of the context-specific tonal contour is automatic, or whether it occurs only with overt production of the speech variants, in preparation for articulation. To test what kind of phonological information is activated when sandhi words are processed but not overtly produced, prime and target conditions were reversed in Experiment 2.

In Experiment 2, primes were visually presented sandhi tonal variants or control distractor words, and targets were Tone 3 or Tone 2 pictures. As with Experiment 1, results showed facilit-

ation of both the tone category targets and the tonal contour targets with sandhi distractors compared to controls. This indicates that processing of the context-specific instantiation is automatic and is not simply a result of articulation preparation. Interestingly, the time course of activation differed from Experiment 1. When the tonal variants were processed visually, rather than overtly produced, it was the tone category that was facilitated both with simultaneous (0 ms) and delayed (83 ms) presentation of the distractor word. Facilitation from the tonal contour occurred only with simultaneous presentation. This suggests that during visual processing of tone sandhi variants, activation of the context-specific contour takes time relative to the speech category. When the distractor is presented simultaneously with the target, the contour is activated in time to benefit production. However, when presentation is delayed, it no longer benefits production of the contour target.

In an event related potential (ERP) study, Chapter 3 investigated processing of sub-phonemic segmental information in a typologically different language. Very little is known about sub-phonemic processing in reading aloud. This study used masked priming to investigate whether and when phonetic features are processed in Dutch reading aloud. In Dutch, the sound pairs t-d and p-b are produced at the same place of articulation (alveolar and bilabial, respectively), while the pairs t-p and d-b match in voicing (voiceless and voiced, respectively). EEG and reaction times were measured as participants read aloud real Dutch words (e.g. huid 'skin'). Each target word was preceded by a brief presentation of a masked non-word prime in which the final sound matched in voicing (*huib*), place of articulation (*huit*) or mismatched in both voicing and place (control condition, *huip*). Reaction times, analysed using linear mixed effects regression models, were significantly faster following voice-congruent primes, compared to control primes. Consistent with this, there was also reduced early negativity in the voice-match condition, compared to the control condition. The results indicate rapid processing of sub-phonemic voicing information in Dutch reading aloud. The facilitation cannot be due to orthographic or phoneme-level phonological information, since in match and mismatch conditions the

relation between prime and target was identical in terms of both letters and phonemes: all primes differed from the target by exactly one phoneme and one letter. Only at the sub-phonemic feature level was the overlap greater in the match conditions, compared to controls. These results also have implications for the way sub-phonemic features are represented. Due to 'final devoicing', Dutch voiced stops have two realisations. In word-initial position, voiced (e.g. /d/ and /b/) and voiceless stops (/t/ and /p/) are distinguished by voice onset time (VOT). But in word-final position, the VOT values of voiced and voiceless stops are very similar. Therefore, at the level of articulation, both prime types in this study are voiceless. However, facilitation in voice-match condition indicates that the voicing distinction is retained and processed during reading aloud. This suggests that, although the overt realisation is similar, voiceless and devoiced stops are processed as separate categories. This is consistent with data presented in Chapter 2 showing both category-level and context-specific processing of speech variants during speech production and visual word processing. Although this experiment did not specifically test for activation of a context-specific variant, it is consistent with the possibility that multi-level processing also occurs at the feature level.

Chapter 4 used ERP measures to address the question of how context constrains phonological processing during reading aloud. ERPs and reaction times were recorded as native Beijing Mandarin speakers read aloud two-character words, preceded by masked primes. The initial character of all critical primes was a Tone 3 character, and initial characters were identical between priming conditions. Only the second character differed between conditions. The phonetic context created by the tone of the second character determined whether the first character had a rising (sandhi) contour or the canonical, low contour. Critical targets were words beginning with Tone 2, which also has a rising contour. Therefore all primes mismatched in tone category, but the sandhi primes matched in tonal contour. The best-fit generalised additive mixed model (GAMM) included full random effects structure, and significant predictors of prime condition, prime frequency and target frequency over time, and their interactions.

In the mismatch condition, effects of prime and target frequency were relatively minor. However, in the contour match condition, the tonal contour did not discriminate between prime and target creating a conflict for the participant response. In this condition, the a priori probabilities of the prime and target came into play. When target frequency was low and prime frequency high, ERP amplitudes were greater, suggesting increased processing effort due to resonse conflict. The differential effects in the match compared to the match condition provide evidence that top-down processing from information provided by the phonetic context constrains activation of the alternative realisations of allophonic variants. Identical initial characters (between conditions) led to different amplitude in the EEG, depending on the tonal context of the following character. The finding that target word frequency interacts with other predictors in explaining neural activity suggests that it is informative to include item information in ERP studies of language processing.

While the previous chapters investigated processing of speech contrasts, contrastive features and regular variation, Chapter 5 investigated effects of the degree of acoustic noise on processing of speech contrasts. Eye movements were recorded as native Cantonese listeners saw pictures corresponding aspirated and unaspirated word pairs and heard acoustic stimuli that contained either a relatively large amount of variation (the wide distribution condition) or relatively little variation (the narrow distribution condition). Analysis using generalised additive mixed modelling (GAMM) allowed complex, non-linear effects and interactions to be modelled, without the need to discretise continuous measures, such as time and frequency. The best-fit model revealed a clear shape of the distribution in the narrow (low-variability) condition. There was differential looking behaviour at category means, boundaries and peripheries. In contrast, in the wide (high-variability) condition, the distribution was flatter, particularly in the latter part of the trial. In other words, the effect of VOT was weak in the wide condition, especially after 600 ms, when the distribution appeared very flat across all VOTs. These results show that during online perception of speech contrasts, previous experience with the distribution an acoustic cue can affect

the degree to which it is used to predict a linguistic outcome. That is, subtle differences in the degree of variation of a particular acoustic cue affect the way that is perceived and utilised to discriminate speech contrasts. In sum, this thesis provides new insights into how human listeners process contrastive and non-contrastive acoustic information during a variety of language processing tasks. It seems that speakers are able to extract several types of regularity from speech, whether it is contrastive speech categories, regular allophonic variants, or sub-phonemic feature information. These regularities are processed during overt production, visual processing of written words that are not overtly produced and during reading aloud. Top-down processing of information provided by the phonetic context constrains activation. Finally, the degree to which acoustic information is relied on as a cue to discrimination of sound contrasts during speech perception depends on the informativity–that is, the shape of the statistical distribution–of the acoustic cues.

Nederlandse samenvatting

In dit proefschrift wordt bewijsmateriaal gepresenteerd dat door middel van verscheidene empirische en statistische methoden is verkregen om te onderzoeken hoe gezonde volwassen moedertaalsprekers de intrinsiek ruizige akoestische informatie verwerken waar de gesproken communicatie uit bestaat. Daar waar psycholinguïstische modellen (met name die van spraakproductie en van hardop lezen) fonologische verwerking doorgaans in termen van foneemsequenties hebben gemeten, toont dit proefschrift verschillende soorten sub-fonemische verwerking aan zoals die in verscheidene experimentele taken plaatsvindt. Eén van de hoofdonderwerpen van dit proefschrift is de manier waarop sprekers en luisteraars fonetische variatie verwerken. Belangrijk is dat een onderscheid wordt gemaakt tussen de reguliere informatieve variatie, zoals bij allofonische varianten, enerzijds, en anderzijds onvoorspelbare ruis, die de informativiteit van akoestische cues beperkt. Verder laten we zien dat zelfs in een taal waarbij geen informatie over spraakcontrasten in de spelling wordt weerspiegeld, bij zowel de overte productie als de visuele verwerking van geschreven woorden sprake is van verwerking van contrastieve spraakklankcategorieën.

In hoofdstuk 2 hebben we gebruik gemaakt van de plaatje-woord-interferentietaak om vast te stellen hoe toonvarianten worden verwerkt bij de spraakproductie van het Mandarijn. In het Peking-Mandarijn is het zo dat als Toon 3 wordt gevolgd door een karakter met nog eens Toon 3 (*sandhi* van toon 3), de eerste toon uiteindelijk als Toon 2 klinkt. Dit aspect van

het tonale systeem behelst dat sandhi-woorden fonologisch zowel gerelateerd zijn aan Toon 3 als aan Toon 2. Dat stelde ons ertoe in staat om twee soorten fonologische informatie te manipuleren: de fonologische tooncategorie en de akoestische tooncontour. In hoofdstuk 2 zijn we ingegaan op de vraag welke van deze twee types van fonologische verbanden belangrijk is bij spraakproductie. Hoewel er een overvloed aan psycholinguïstisch onderzoek bestaat dat aantoont dat verwerking wordt ondersteund door de voor-activatie van congruente fonologische informatie, is de precieze aard van deze fonologische informatie nog niet opgehelderd. Gaat met spraakproductie het ophalen van de spraakklankcategorie gemoeid? Of gaat het om de activering van de eigenlijke akoestische realisatie? Hoofdstuk 2 laat zien dat spraakproductie een kwestie is van fonologische verwerking op meerdere niveaus (zie ook Nixon et al., 2014). De productie van afbeeldingsnamen in sandhi-conditie voltrok zich significant sneller wanneer de afleider en de doelafbeelding overeenkwamen in tooncategorie maar verschillende overte realisaties hadden *en ook* wanneer doelafbeelding en afleider in overte realisatie overeenkwamen maar onder andere tooncategorieën vielen (vergeleken met controleafleiders, die in toneem noch in contour met de doelafbeelding overeenkwamen). Interessant genoeg waren er verschillen in het tijdsverloop van de effecten. De tooncontour versnelde de productie zowel wanneer die gelijktijdig met de doelafbeelding werd gepresenteerd als wanneer die er 83 milliseconden achteraan kwam. De tooncategorie, echter, versnelde de productie alleen in het geval van gelijktijdige presentatie. Dit doet vermoeden dat de spraakcategorie in een vroeg stadium wordt geactiveerd (mogelijk op het moment van het ophalen van het woord), maar dat de daadwerkelijke akoestische realisatie in een later stadium van de verwerking van belang wordt (bijvoorbeeld bij de spraakvoorbereiding). Er zijn twee verklaringen denkbaar voor deze resultaten. Het kan zijn dat de gerealiseerde contour langer actief blijft dan de tooncategorie gedurende overte productie. Een andere mogelijkheid is dat de contour en de categorie beide actief blijven, terwijl de taak overgaat van lexicale retrieval naar de voorbereiding van de articulatie, maar dat alleen het samenkomen van articulatie en akoestische realisatie de productie ten goede komt. Er is meer onderzoek nodig om te be-

palen welk van deze twee mogelijkheden de voorkeur verdient. Een tweede vraag die zich opdringt naar aanleiding van de resultaten van Experiment 1 is of het verwerken van context-specifieke tooncontouren automatisch verloopt of dat het zich pas voordoet bij overte productie van spraakvarianten, in voorbereiding op de articulatie. Om te achterhalen wat voor soort fonologische informatie wordt geactiveerd wanneer sandhi-woorden worden verwerkt maar niet overt worden geproduceerd, zijn in Experiment 2 de prime- en doelcondities omgewisseld.

In Experiment 2 bestonden de primes uit visueel gepresenteerde aan sandhi onderhevige toonvarianten of afleiders als controlewoorden. Als doelwoorden dienden plaatjes met Toon 3 of Toon 2. Net zoals in Experiment 1 vond er versnelling plaats van zowel de doelwoorden van de tooncategorie als de doelwoorden van de tooncontour met sandhi-afleiders ten opzichte van de controlewoorden. Hieruit valt op te maken dat de verwerking van de context-specifieke verwezenlijking een automatisch proces is en niet simpelweg het gevolg van articulatievoorbereiding. Interessant genoeg verschilde het tijdsverloop van activatie met dat van Experiment 1. Wanneer de toonvarianten visueel werden verwerkt in plaats van overt geproduceerd was het de tooncategorie die werd versneld, zowel in het geval van de gelijktijdige (0 ms) als in het geval van de verlate (83 ms) aanbieding van het afleiderwoord. Versnelling van de tooncontour deed zich alleen voor bij de gelijktijdige aanbieding. Dit duidt erop dat de activatie van de context-specifieke contour gedurende de visuele verwerking van toon-sandhi varianten tijd kost in verhouding tot de spraakcategorie. Bij gelijktijdige presentatie van het afleiderwoord met het doelwoord wordt de contour op tijd geactiveerd om productie te kunnen versnellen. Wanneer de aanbieding echter later plaatsvindt kan het de productie van de doelcontour niet meer versnellen.

In hoofdstuk 3 hebben we met behulp van een onderzoek met event related potentials (ERP) de verwerking van sub-fonemische segmentele informatie in een typologisch andere taal bekeken. Er is erg weinig bekend over sub-fonemische verwerking bij het hardop lezen. In dit onderzoek hebben we ons bediend van gemaskeerde priming om na te gaan of en wanneer er fonetische

kenmerken worden verwerkt bij het hardop lezen in het Nederlands. De segmenten van de Nederlandse klankparen t-d en p-b hebben onderling dezelfde articulatieplaats (respectievelijk alveolair en bilabiaal), terwijl de segmenten van de paren t-p en d-b dezelfde stemhebbendheid hebben (respectievelijk stemloos en stemhebbend). We hebben EEG en reactietijden gemeten bij deelnemers die hardop bestaande Nederlandse woorden (bijv. *huid*) oplazen. Elk doelwoord werd voorafgegaan door een korte aanbieding van een gemaskeerde nonwoord-prime waarvan de laatste klank ofwel in stemhebbendheid overeenkwam (*huib*), dan wel in articulatieplaats (*huit*) of noch in stemhebbendheid noch in articulatieplaats (de controleconditie, *huip*). De reactietijden, geanalyseerd met linear mixed effects-regressiemodellen), waren significant korter bij primes die in stemhebbendheid overeenkwamen dan bij controleprimes. In lijn hiermee was de verminderde early negativity in de conditie met overeenkomstige stemhebbendheid vergeleken met de controleconditie. Deze resultaten wijzen op een snelle verwerking van sub-fonemische stemhebbendheid bij hardop lezen in het Nederlands. Deze versnelling kan niet het gevolg zijn van spellingsinformatie of informatie op foneemniveau aangezien de verschillen tussen het aantal letters en de fonemen van prime en doelwoord identiek waren in de condities met overeenkomstige en niet-overeenkomstige kenmerken: alle primes verschilden op precies één foneem en één letter van de doelwoorden. Alleen op het sub-fonemisch kenmerkniveau was de overlap groter in de overeenkomstconditie, in vergelijking met de controleconditie.

Deze resultaten zeggen ook iets over de manier waarop subfonemische kenmerken gerepresenteerd worden. Nederlandse stemhebbende plosieven kennen door 'final devoicing' (stemloos worden van een finale stemhebbende klank) twee realisaties. Stemhebbende (bijv. /d/ en /b/) en stemloze (/t/ en /p/) plosieven worden in woord-initile positie onderscheiden door voice onset time (VOT). Maar in woordfinale positie zitten de VOT-waardes van stemhebbende en stemloze plosieven heel dicht bij elkaar. Op het niveau van de articulatie zijn de beide primetypes daarom stemloos. De versnelling van verwerking in de conditie met overeenkomstige stemhebbendheid, echter, wijst erop dat het stemhebbendheidsonderscheid behouden is en wordt verwerkt

tijdens het hardop lezen. Dit duidt erop dat stemloze en stemloos geworden plosieven ondanks hun vergelijkbare overte realisatie als aparte categorieën worden verwerkt. Dit is in lijn met de data uit hoofdstuk 2 waaruit bleek dat gedurende spraakproductie en visuele woordverwerking spraakvarianten op het niveau van categorieën en in specifieke contexten een rol speelden. Ondanks het feit dat in dit experiment niet specifiek is getest of activatie van een context-specifieke variant plaatsvond, past het wel bij het scenario dat er zich ook op het kenmerkniveau meerlagige verwerking voordoet.

Hoofdstuk 4 beschrijft een ERP-experiment waarin de vraag centraal staat hoe de fonologisch verwerking van hardop lezen afhangt van de context. We hebben van moedertaalsprekers van het Peking-Mandarijn ERP's en reactietijden geregistreerd terwijl ze hardop woorden oplazen die een lengte hadden van twee karakters en die vooraf werden gegaan door gemaskeerde primes. Het eerste karakter van alle kritieke primes was een karakter met Toon 3 en de beginkarakters waren identiek voor alle primingcondities. Alleen het tweede karakter verschilde tussen condities. De fonetische context die gecre‘erd werd door de toon van het tweede karakter maakte uit of het eerste karakter een stijgende (sandhi) contour of de canonieke, lage contour had. De kritieke doelwoorden waren de woorden die begonnen met Toon 2, waarvan de contour ook stijgend is. Alle primes kwamen dus niet overeen in de tooncategorie, maar de sandhi-primes kwamen wel overeen in de tooncontour. Het best passende generalised additive mixed model (GAMM) dat we hebben toegepast had een volledige random effects-structuur en significante predictoren van conditie, trial en doelwoordfrequentie over de tijd, alsmede de interacties daartussen. Uit de predictor van condities over de tijd bleek dat er significante verschillen waren in ERP-amplitude binnen het tijdsvenster van 300-350 ms. Deze resultaten duiden erop dat de top-downverwerking van informatie vanuit de fonetische context de activatie van alternatieve realisaties van allofonische varianten beïnvloedt. Identieke beginkarakters (tussen condities) leidden tot verschillende amplitudes in de EEG afhankelijk van de tooncontext van het volgende karakter. Bovendien toonde de interactie tussen conditie en doelwoordfrequentie over de tijd aan dat de congruentie van de contour het ster-

kste was voor laagfrequente doelwoorden. Dit komt waarschijnlijk door de lagere resting state-activatie van laagfrequente woorden, die maakt dat ze ontvankelijker zijn voor primingeffecten. Gezien de bevinding dat doelwoordfrequentie in samenspel met andere predictoren een (deel van de) verklaring vormt voor de gevonden neurale activiteit heeft het meenemen van iteminformatie in ERP-onderzoek een toegevoegde waarde.

Waar de hoofdstukken tot nu toe ingingen op de verwerking van spraakcontrasten, contrastieve kenmerken en reguliere variatie, bogen we ons in Hoofdstuk 5 over de effecten van de hoeveelheid akoestische ruis op het verwerken van spraakcontrasten. We hebben de oogbewegingen gemeten van moedertaalsprekers van het Kantonees die een taak uitvoerden waarvoor ze naar plaatjes keken en naar akoestische stimuli luisterden. De plaatjes correspondeerden met geaspireerde en niet-geaspireerde woordparen en de akoestische stimuli bevatten ofwel een relatief grote hoeveelheid variatie (de conditie met de *wijde* distributie) ofwel een relatief kleine hoeveelheid variatie (de conditie met de *nauwe* distributie). Door gebruik te maken van generalised additive mixed modelling (GAMM) konden we complexe, niet-lineaire effecten en interacties modelleren zonder continue variabelen, zoals tijd en frequentie, discreet te hoeven maken. Uit het best passende model kwam een duidelijke vorm van de distributie naar voren in de conditie met de nauwe distributie (weinig variatie). Het kijkgedrag verschilde tussen gemiddeldes, grenzen en periferieën van categorieën. In de wijde conditie (veel variatie) was de verdeling platter, met name in het laatste stuk van de trial. Het effect van VOT was, met andere woorden, zwak in de wijde conditie, vooral na 600 ms, waar de verdeling erg plat was voor alle VOT-waarden. Wij maken uit deze resultaten op dat tijdens het online waarnemen van spraakcontrasten eerdere ervaring met de verdeling van een akoestische cue de mate waarin die cue wordt gebruikt om een talige uitkomst te voorspellen beïnvloedt. Anders gezegd zijn de subtiele verschillen in de mate van variatie van een bepaalde akoestische cue van belang voor de manier waarop die wordt waargenomen en gebruikt om spraakcontrasten van elkaar te onderscheiden.

Samengevat biedt dit proefschrift nieuwe inzichten in de manier waarop mensen contrastieve en niet-contrastieve akoestische in-

formatie verwerken bij verscheidene taalverwerkingstaken. Het lijkt erop dat sprekers verschillende soorten regelmatigheden uit het spraaksignaal opvangen, zoals contrastieve spraakcategorieën, reguliere allofoonvarianten en sub-fonemische kenmerkinformatie. Deze regelmatigheden spelen in het taalsysteem een rol bij overte spraakproductie, visuele verwerking van geschreven woorden die niet overt worden geproduceerd en hardop lezen. Verwerking gaat echter ook gepaard met de activatie van een context-specifieke realisatie. Activatie wordt beïnvloed door top-downverwerking van informatie over de fonetische context. Tot slot hangt de mate waarin akoestische informatie als cue houvast biedt aan het onderscheiden van klankcontrasten bij spraakperceptie af van de informativiteit–dat wil zeggen, de vorm van de statistische verdeling–van de akoestische cues.

About the author

Jessie Nixon was born in Auckland, New Zealand, but soon headed south to Takamatua, Banks Peninsula, at age 2 months and then to Christchurch two years later. After graduating from the University of Canterbury with a Bachelor of Arts in Chinese and Russian (1^{st}) and a Bachelor of Science in Linguistics (1^{st}), she took up a scholarship at Peking University to study Chinese and Chinese Linguistics and lived in Beijing for 4 years. She completed a Master of Science in Psycholinguistics at the University of Edinburgh (with Distinction) in 2009, before beginning her PhD research at the University of Leiden in the same year. In 2014, Jessie was awarded a grant from Asian Modernities and Traditions at the University of Leiden to conduct research on tone processing in speech. She spent the summer of 2014 as a post-doctoral researcher with Harald Baayen at the University of Tübingen and is a recipient of the 2014 Endeavour Research Fellowship for the position of post-doctoral researcher at the MARCS Institute, University of Western Sydney.

ISBN: 978-90-8891-975-6