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

Nixon, J.S.

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.