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

Nixon, J.S.

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.