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

Nixon, J.S.

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 subphonemic 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, soundproof 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

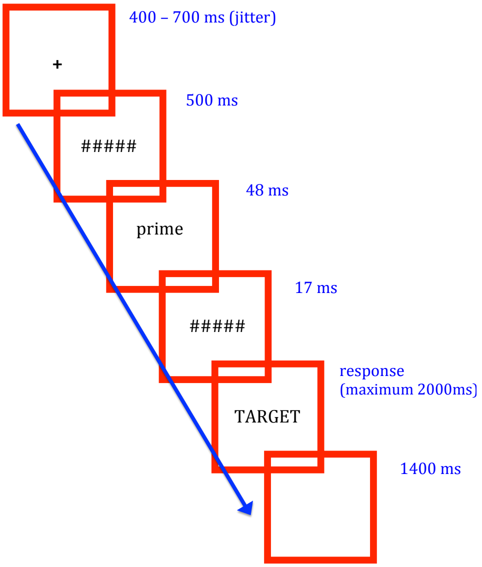

Figure 3.1: Trial procedure: a fixation cross is presented with a jittered duration, followed by a forward mask (500 ms), the non-word prime (48 ms), a backward mask (17 ms) and finally the target (until participant response or a maximum of 2000 ms). Errors are coded during a blank screen preceding the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

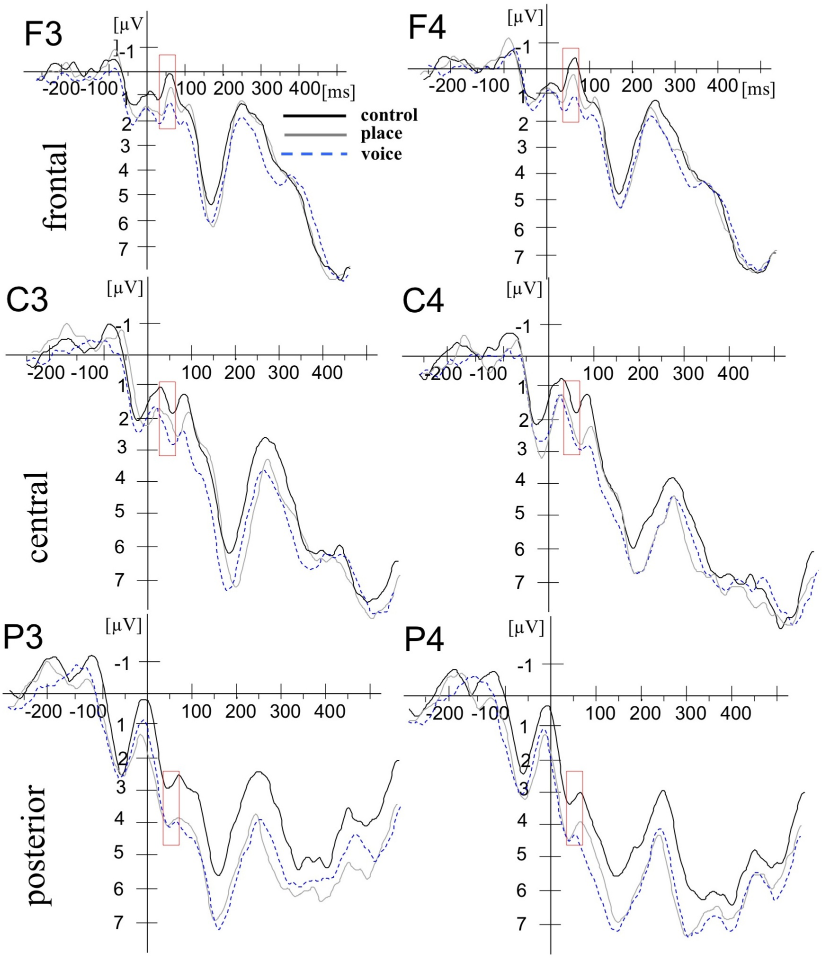

Figure 3.2: Average ERP responses for voice-match, place-match and control conditions.

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.