

# **Unfolding of phonetic information over time:**

## **Perception of gated Dutch diphones**

Roel Smits\*, Natasha Warner\*†, James M. McQueen\* and Anne Cutler\*

\*Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

†Department of Linguistics, University of Arizona, Tucson, Arizona, U.S.A.

### **Address for correspondence:**

Roel Smits

Max Planck Institute for Psycholinguistics

Postbus 310

6500 AH Nijmegen

The Netherlands

Telephone: + 31 24 3521374

Fax: + 31 24 3521213

Email: roel.smits@mpi.nl

**Running head:** Perception of gated Dutch diphones

Received:

## **Abstract**

We present the results of a large-scale study on speech perception, assessing the number and type of perceptual hypotheses which listeners entertain about possible phoneme sequences in their language. Dutch listeners were asked to identify gated fragments of all 1179 diphones of Dutch, providing a total of 488,520 phoneme categorizations. The results manifest orderly uptake of acoustic information in the signal, and virtually no effect of higher-level influences such as phoneme occurrence frequencies or transition probabilities. Differences across phonemes in the rate at which fully correct recognition was achieved arose as a result of whether or not potential confusions could arise with other phonemes of the language (long with short vowels, affricates with their initial components, etc.). These data can be used to develop models of speech comprehension in which acoustic-phonetic information is mapped continuously onto the mental lexicon.

PACS number: 43.71.Es

## I. INTRODUCTION

The central component of human speech comprehension is word recognition. Only when listeners have decoded the words in an utterance can they derive a speaker's message. Listeners must therefore map the acoustic-phonetic information in the speech signal onto stored knowledge about the phonological forms of words in their language. A common view in psycholinguistic models of spoken-word recognition (e.g., TRACE, McClelland and Elman, 1986; Shortlist, Norris, 1994) is that this decoding process is essentially phonemic: the speech signal is mapped onto intermediate phonemic representations, which in turn are used to access phonemically coded lexical representations. Spoken-word recognition is however more fine-grained, both informationally and temporally, than can be captured by a discrete phonemic analysis of the signal.

Several types of evidence indicate that speech recognition does not involve a discrete phonemic stage of processing preceding lexical access. First, acoustic information relevant for phoneme recognition is strongly smeared over time by coarticulation, such that it overlaps with acoustic information associated with neighboring phonemes (see e.g., Fowler, 1984; Martin and Bunnell, 1981, 1982; Smits, 2000, 2001; Whalen, 1989). Thus Smits (2000) showed that perceptually relevant information for consonant identification in English VCV utterances extends for well over 80 ms into preceding and following vocalic segments. Phoneme information therefore does not become available to the listener as a string of discrete labels. Instead, the information unfolds gradually over time, and at any one moment information on several upcoming phonemes is available to varying degrees.

Second, acoustic realization of phonemes depends on phonetic environment (see, e.g., Stevens, 1998). For example, the acoustic information associated with /t/ is very different in the strings [str] and [utu]. As a result, the unfolding of perceptually relevant information for a given phoneme must also depend on the phonetic environment of that phoneme. Furthermore,

acoustic realizations of phonemes are different in stressed versus unstressed syllables – in particular, segment duration correlates strongly with stress, both for vowels and consonants (e.g., Lindblom, 1963; Van Son and Pols, 1999). Both vowels and consonants are also less intelligible in unstressed syllables than in stressed syllables (e.g., Van Bergem, 1995; Van Son and Pols, 1997). The variability in acoustic realization of phonemes due both to phonetic environment and to prosodic factors again argues against the discrete phonemic approach.

A third problem with viewing speech recognition in terms of discrete phonemic categorization is that if categorical phonemic decisions were made prior to lexical access, word recognition would be prone to error (Klatt, 1980). It would be more difficult to recover the intended word if the phonemic analysis misclassified a particular segment and passed only that incorrect decision forward to the lexicon, than if a number of different alternatives were passed forward to the lexicon. In other words, information that is useful for lexical selection would be thrown away if categorical decisions were made prelexically.

Finally, lexical-level processes are sensitive to subphonemic variation in the speech signal. Andruski *et al.* (1994), for example, found that variation of Voice Onset Time (VOT) in English stop consonants influences the activation of English words beginning with those consonants. Words beginning with unvoiced stops were more weakly activated when the VOTs in their stops were shorter than normal (i.e., were more like voiced stops) than when the VOTs were of normal duration. Studies of the perception of words containing mismatching acoustic-phonetic information (Dahan *et al.*, 2001; Marslen-Wilson and Warren, 1994; McQueen *et al.*, 1999; Streeter and Nigro, 1979; Whalen, 1984, 1991) also show that the lexical level is sensitive to subphonemic variation. Marslen-Wilson and Warren (1994), for example, showed that phonemic and lexical decisions were slower to the word *job* when the [dʒp] was taken from a token of *jog* (and thus contained formant transition information consistent with a velar [g], mismatching with the final bilabial stop release burst) than when the [dʒp] was spliced from another token of *job* (and thus had no mismatching information).

It therefore seems clear that discrete categorical decisions about phonemes in an utterance are not reached before contact is made with the mental lexicon. Acoustic-phonetic information appears to be passed to the lexicon continuously, as it becomes available in the speech signal. This is not to say, however, that phonemes have no role to play in lexical access. The view that speech decoding is essentially phonemic captures the important fact that differences between words are phonemic differences. By definition, a word's nearest lexical neighbors will differ from that word by only one phoneme. It is thus phonemic differences in the input which signal speakers' intentions. The processes and representations involved in lexical access must thus be sensitive to phonemic differences. It is perhaps for this reason that psycholinguistic models of spoken-word recognition have assigned a greater role in lexical access to phonemic distinctions than to the continuous unfolding of phonemically relevant information. If these models are to reflect human speech recognition more accurately, however, they must be able to capture the continuity of prelexical speech processing.

With this ultimate aim, the present experiment was devised to provide more complete and temporally detailed perceptual data than has hitherto been available. The unfolding of phonetic information over time is measured most naturally via the gating technique (Cutler and Otake, 1999; Ellis *et al.*, 1971; Grosjean, 1996; Ohala and Ohala, 1995; McQueen, 1995; Smits, 2000; Warner, 1998; Warren and Marslen-Wilson, 1987). In gating, portions of a natural utterance are presented to listeners for classification. In the present study we used only “final” (or “forward”) gating, in which the final portion of an utterance is deleted and the initial portion is presented. This method best captures how information becomes available to the listener over time, as it does in normal listening situations.

We strove to measure phoneme activations in all possible left and right phonetic contexts. Ideally this would be achieved by presenting gated versions of all possible triphones in a given language. This was impractical in Dutch given the size of the triphone set. We therefore presented gated versions of all possible Dutch diphones. Listeners were asked to

identify the first and second phoneme of each of these diphones. To enable us also to assess effects of stress on acoustic realizations of phonemes, most diphones were recorded in multiple stress conditions. There were a total of 2,294 sequences (1,179 diphones, most of which occurred in multiple stress conditions).

Each listener heard six gates of each sequence, based on six gating points, three in each sound of the diphone. The shortest gate therefore included only the first third of the first sound; each subsequent gate included another sixth of the entire diphone. These gates were not presented sequentially, by increasing size within each diphone. Instead, the entire stimulus set (all gates from all diphone sequences) was presented to each listener in a different pseudo-random order. Each of the 19 participants who completed the experiment provided 27,140 identification responses (on average 2,294 sequences times 5.915 gates – for some stimuli some gates were omitted, see below). This is the largest source of data on phonetic perception in Dutch or any other language that is currently available.

## II. METHODS

### A. Materials

#### 1. *Choice of diphones*

We first compiled a list of all possible diphones of the Dutch language. For this purpose, we considered the phonemic inventory of Dutch to be as in Tables I and II.

----- insert Tables I and II about here -----

Decisions as to what constitutes a single phoneme vs. a sequence of two phonemes were based on the CELEX database (Baayen *et al.*, 1993). We did not, however, include all phonemes and diphones used by CELEX (see Appendix A for explanation of exceptions). We

then formed a list of diphones consisting of all possible combinations of any two of these phonemes. Appendix B lists the rules we applied. In Appendix C is a list of the 2,294 diphones included in the experiment, and reasons for exclusion of missing diphones.

## 2. *Recording*

Each diphone in Appendix C was placed in a nonsense environment which, with the diphone, formed a phonotactically legal sequence in Dutch. In some cases this was necessary because the diphone by itself is not a syllable (CC diphones) or is not phonotactically legal (e.g., C-short vowel diphones, since short vowels cannot be syllable-final). For VV diphones where both vowels are either stressed or unstressed, inclusion of additional syllables made the sequences easier for the speaker to produce with correct stress. The nonsense environment always included at least one phoneme after the target diphone, so that the diphone would not be final to the item. This prevented excessive lengthening within the diphone, as would for example apply to the vowel in a CV diphone if it were recorded in isolation. The environments for CV and VC diphones were also varied; this prevented subjects from learning to predict the category of diphone from the preceding environment, as they could if, for example, the preceding environment /a-/ was only used for CC diphones. Table III lists the environments in which the various diphones were recorded.

----- insert Table III about here -----

All items (diphones in their environments) were transcribed phonemically, with stress and syllable boundaries marked. A phonetically trained female native speaker of standard Dutch was recorded on DAT tape reading all of the items from this transcription. The recording was made in a sound-treated recording booth using high quality equipment. Any items which were initially mispronounced were re-recorded.

### 3. *Stimuli for the perception experiment*

Each item was final-gated at six points during the target diphone (with exceptions for initial stops and affricates, see below), to create stimuli consisting of the entire item up to the gating point, including any preceding context. The total number of stimuli was 13,570. Items were gated to 300 ms of a 500 Hz square wave, with a 5 ms period during which the speech signal was ramped down and the square wave simultaneously ramped up. A square wave was used rather than noise or silence, since it is not misperceived as a speech sound, and thus does not introduce cues for a fricative or a voiceless stop at the gating point (Warner, 1998).

The six gating points for a diphone included three in each of the target phonemes. For phonemes which lack abrupt acoustic changes during the segment, such as nasals, fricatives, and vowels in most environments, beginning and end points of the segments were identified by hand, and gate end points were placed automatically at one third and two thirds through the duration of the segment as well as at the end of the segment. For a vowel-nasal diphone, for example, gate 1 allowed subjects to hear from the beginning of the item to a point one third through the vowel, gate 2 to two thirds through the vowel, gate 3 to the end of the vowel, gate 4 to one third through the nasal, gate 5 to two thirds of the nasal, and gate 6 to the end of the diphone. Any preceding environment was always included in the stimuli, but following environment was never included. For segments with abrupt acoustic changes within the segment, such as stops and affricates, gate end points were determined relative to those abrupt changes. Appendix D lists criteria for identifying segment beginnings and ends.

### **B. Subjects and procedures**

Twenty-two listeners participated in the experiment, and 19 completed it. Data from the three subjects who did not finish the entire experiment were excluded. All subjects were native speakers of Dutch who had grown up in the Netherlands, and had no known hearing

impairment. Most were students at the University of Nijmegen. Subjects were paid for each hour of participation, with a bonus on finishing the experiment.

The task involved identifying the two phonemes of the target diphone. Subjects were tested individually in a sound treated booth. Stimuli were presented over closed headphones, with stimulus presentation and collection of responses controlled by NESU experimental control software. As each stimulus was played, a response screen appeared on a computer screen visible through the booth window. The response screen showed two panels, each containing buttons for each phoneme used in the experiment. Subjects used a computer mouse to click on one button of the left-hand panel for the first sound of the diphone, and one of the right-hand panel for the second sound. If the stimulus included preceding context (/a/, /ɑ/, /b/, or /ab/), the letters “aa,” “a,” “b,” or “aab,” respectively, appeared on the screen to the left of the left-hand response panel to inform subjects that those sounds were not the ones to which they should respond. The response buttons for these phonemes were also crossed out in the left response panel to remind subjects not to respond to the preceding environment.

Before beginning the experiment, subjects were trained on the set of symbols to use for responses. Since Dutch orthography is straightforward, most phonemes could be represented orthographically (with double vowels used for long vowels and single vowels used for short vowels); special symbols were necessary only for /ə/ (“@”) and /g/ (“G”). Examples of each phoneme were provided, and special attention was called to phonemes which appear only in loan words. Subjects were told that they would hear the beginning of a nonsense word followed by a beep, and that they should identify the two sounds of the nonsense word using the mouse. They were informed about possible additional initial sounds which they were not to respond to, and warned that they would sometimes hear very little of the nonsense word, making it difficult to identify the two sounds. A native Dutch speaker instructed each subject and checked subjects’ understanding of the mapping of response symbols to sounds.

Subjects then completed a practice session, comprising 185 stimuli drawn from the actual experiment. Diphones containing potentially problematic phonemes, such as /ə, ɪ/ and phonemes occurring only in loan words, were well represented in the practice session. The experimenter evaluated subjects' performance on stimuli which included these sounds or a vowel in their entirety to ensure that subjects could perform the task. No subjects were excluded at this stage, since none had difficulty with the task.

Subsequently, subjects completed a series of one-hour experimental sessions, with a break during each session. Subjects returned for as many sessions as needed to respond to all 13,570 stimuli, an average of 27.9 sessions. The total set of stimuli was divided into four blocks. For each subject a different pseudo-random order of stimuli within blocks was generated and different subjects received the blocks in a different order. Two gates of the same diphone were separated by at least six stimuli, stimuli from diphones beginning with the same phoneme were separated by at least four stimuli, and no stimuli which appeared in the practice session or other gates of those diphones occurred within the first 1200 experimental stimuli. In total 488,520 phoneme categorizations were collected.

### III. RESULTS

#### A. Summary results

----- insert Figure 1 about here -----

One subject performed much worse than the others in correctly recognizing the first phoneme at gates 1 to 3. For these gates this subject's recognition rates were more than four standard deviations below the mean recognition rates for all other subjects. The data of this subject were therefore excluded. Figure 1 shows average phoneme recognition rates as a function of

gate, pooled across the remaining 18 subjects, for consonants, vowels, and all phonemes. At gate 1, that is, one third into the first phoneme of the diphone, the first phoneme (top line) is recognized at about 60% correct. With increasing gates, this level rises smoothly to 92% at gate 4 and thereafter hardly increases any further. The recognition rate for the second phoneme (bottom line) starts close to chance level (which is 2.6%) at gate 1 and rises smoothly to 88% at gate 6. One-tailed *t*-tests (with subject as the random variable, as in all subsequent analyses) show that at all gates average recognition rates for both phonemes are significantly above chance level as well as below perfect performance (all *p*'s < .0005).

Recognition rates for gates 4, 5 and 6 of the second phoneme are quite similar to those for gates 1, 2 and 3 of the first phoneme. This shows that the fact that a longer preceding context was available to listeners for the second phoneme did not improve recognition much compared to the first phoneme.

The recognition curves for vowels and consonants are not very different. Globally, the main difference is that correct recognition of consonants starts somewhat earlier but rises more slowly compared to correct recognition of vowels. Overall, the information necessary for correct phoneme recognition is thus more spread out for consonants than for vowels.

Figure 2 shows correct recognition rates separately for the 22 consonants, grouped by manner and voicing, while Figure 3 presents responses for all 16 individual vowels as a function of gate, grouped partly according to vowel features and partly according to similarities of the individual curves. Tables IV and V present confusion matrices for consonants and vowels, respectively, summed across listeners, contexts, and stress conditions, in responses to gate 1 for the first phoneme and to gate 4 for the second phoneme.

## **B. Consonants**

----- insert Figure 2 and Table IV about here -----

(1) *voiceless stops /p t k/* (Figure 2A). As shown in Table III, some diphones were recorded with preceding context and some without. For those without preceding context, gates 1 and 2 were not presented because they contained only silence. Gates 1 and 2 in Figure 2A therefore represent only responses to gated diphones with preceding context – that is, the vowel /a/ with formant transitions plus respectively half or all of the following stop closure. It is clear that subjects could recognize the stops from these portions, with recognition rates between 50 and 80%. Recognition of /t/ was somewhat poorer than of /p/ and /k/. This is supported by a *t*-test ( $p < .001$  for all comparisons between /t/ and /p/ or /t/ and /k/ at gates 1 and 2). The difference was mainly caused by more place and voicing errors for /t/ than for /p/ and /k/ (see Table IV). Gate 3 included the release burst, which strongly improved recognition.

Recognition of the second phoneme (i.e., voiceless stops in second position in the diphone) at gates 4 and 5 is considerably worse than recognition of the first phoneme at gates 1 and 2 ( $p < .005$  for all six comparisons) - data for the second phoneme at gates 4 and 5 include all possible preceding contexts, not only /a/. The raw data show that, on average, /a/ as preceding context led to better recognition of the following stop than other preceding contexts. This agrees with reports of Dorman *et al.* (1977) and Smits *et al.* (1996) that formant transitions in /a/ are generally more informative on place of articulation of an adjacent consonant than transitions in other vowels. At gate 6, when the stop burst is audible, recognition levels exceed 90%.

(2) *voiced stops /b d g/ and the voiced affricate /dʒ/* (Figure 2B). In first position, recognition is poorer than for voiceless stops ( $p < .005$  for all 18 comparisons). Note that the data for gates 1 and 2 include those for stimuli which contained only half or all of the voice bar preceding the release of stops with no preceding context. The generally poor scores for these stimuli depress the overall mean for gates 1 and 2. /b/ fares better than /d/ and /g/ for gates 1 and 2 ( $p < .001$  for all four comparisons). Thus our data reconfirm the findings of, amongst others, Pols and Schouten (1978) and Smits (2000) that an isolated voice bar sounds

more like a /b/ than a /d/ or /g/. For later gates, place and voicing confusions are the main source of errors (see Table IV). Voiced stops are more often confused with their voiceless counterparts than vice versa. Especially /b/ was classified relatively frequently as /p/ up to gate 6. The voiced affricate /dʒ/ is not recognized reliably in either first or second position until burst and frication are audible. At earlier gates /dʒ/ is mainly confused with /j/ and /d/.

(3) *voiceless fricatives* /f s ʃ x/ (Figure 2C). At gate 1 of the first phoneme, which includes one third of the frication noise, recognition is already quite good, with levels between 60 and 90%. Recognition gradually improves with increasing amounts of frication and subsequent context. Remaining confusions of /f s ʃ/ are with their voiced counterparts. In addition, there is some confusion between /s/ and /ʃ/ (see Table IV). The voiceless velar fricative /x/ is recognized very well at all gates. Note that /x/ has no voiced counterpart in most regional variants of Dutch, including that of our speaker. Recognition levels for gates 4 to 6 of the second phoneme resemble those for gates 1 to 3 for the first. Note the marked jump in recognition between gates 3 and 4, that is, when some frication noise becomes audible.

(4) *voiced fricatives* /v z ʒ h/ (Figure 2D). These, like the stop consonants, are recognized less well than their voiceless counterparts. This is, however, only partially supported by statistical analysis. In initial position, /v z ʒ/ are recognized significantly worse than /f s ʃ/ in only 3 out of 18 cases at the  $p = .05$  level, while in second position, this holds for 8 out of 18 cases. Although the pattern is thus less clear than for the stop consonants, it has the same cause, namely asymmetric confusions on the voicing feature. Voiced fricatives are categorized as their voiceless counterparts more often than the reverse (see Table IV).

The glottal fricative /h/ is recognized better than the other fricatives (initial position:  $p < .05$  for 16 out of 18 comparisons; second position:  $p < .005$  for all 18 comparisons). In first position recognition already exceeds 90% at gate 2. Note that /h/ has no voiceless counterpart, so if manner and place of articulation are recognized, there is no room for voicing errors.

In second position /h/ is recognized surprisingly well at gate 1. This is however an artefact of the gating method: some subjects used a default /h/ response for the second phoneme when they had no information. As the second phoneme sometimes actually was /h/, this response bias increased recognition rates for the early gates of /h/ in second position.

(5) *liquids* /r l/ and *glides* /w j/ (Figure 2E). Recognition in first position is already good at gate 1, with recognition rates between 60 and 85%. At later gates recognition further increases to very high levels. Recognition of the labiodental glide /w/ is somewhat poorer than of the other glide and the liquids ( $p < .05$  for all 18 comparisons, except /w/ versus /j/ at gates 4 to 6); confusions occurred with the voiced labiodental fricative /v/ and the vowels /ʏ/ and /ə/ (N.B. the labiodental glide was hardly ever confused with the vowel /u/). Recognition of liquids and glides in second position gradually increases across all six gates. From gate 4 on, however, recognition of the glides is substantially lower than that of the liquids ( $p < .001$  for all 12 comparisons), and asymptotes at levels close to 80%. /w/ is again mainly confused with /v/ and /j/ is mainly confused with /i/, while the main confusions for /r/ are with /h/ and /ɑ/. The confusions for /l/ are rather scattered and include consonants /d j h r w/ and vowels /i ɪ ə/.

(6) *nasals* /m n ŋ/ (Figure 2F). For nasals in the first position it is striking that /ŋ/ is recognized much better than /m n/ at gates 1 to 3 ( $p < .002$  for all six comparisons). This is again an artefact: Because /ŋ/ cannot occur in syllable-initial position, recognition levels of /ŋ/ in initial position are based on tokens with preceding context /ɑ/, which therefore includes formant transitions into the nasal. In contrast, /m/ and /n/ occurred in initial position in two-thirds of the tokens, without informative preceding transitions. For nasals in second position a marked increase in correct recognition can be seen at gates 3 and 4, which include the speech signal up to oral closure and one third into the murmur, respectively. Table IV shows that at gate 1 in first position and at gate 4 in second, confusions are mainly across place, while at later gates the remaining confusions are across manner, while place is recognized reasonably

well. At gates 5 and 6, recognition of /m/ is some 15% lower than that of /n/ and /ŋ/. The raw data show that /m/ is often confused with /n/ at these gates.

### C. Vowels

----- insert Figure 3 and Table V about here -----

(1) *short vowels* /ɑ ɛ ɪ ɔ/ (Figure 3A). At gate 1, recognition of these vowels in first position is already very good, with levels close to 90% correct. In second position, recognition jumps to levels between 70 and 85% at gate 4 and rises further at subsequent gates. When listeners get one third or more of the target vowel, the remaining confusions are as follows. /ɑ/ is mainly confused with /a/, /ɛ/ is mainly confused with /ɛi/ and /ɪ/, /ɪ/ with /e/ and /i/, and /ɔ/ with /ɑ/ and /o/ (see Table V). That is, short vowels are confused with any nearby long counterpart.

(2) *long vowels* /i u y/ (Figure 3B). These, like the short vowels, are recognized well in first position at gate 1. Note that these vowels have no short counterparts (Booij, 1995). When a third or more of the vowels is audible, the remaining confusions are mainly with similar short vowels: /i/ is confused with /ɪ/, /u/ with /ʊ ə w/, and /y/ with /ə ʏ u/ (see Table V).

(3) *short vowels* /ɣ/ and /ə/ (Figure 3C). Recognition of /ə/ is poor, showing little improvement over the six gates and never exceeding 40% correct. /ɣ/ is recognized better, but still much worse than the short vowels in Figure 3A. As shown in Table V, /ɣ/ and /ə/ more or less form a single category: responses to both stimuli are very similar, and listeners seem to have selected at random between the two responses, with a bias in favor of /ɣ/. We therefore grouped stimuli and responses for these two vowels together and calculated recognition rates for the compound vowel class. The resulting recognition curve is displayed in Figure 3C with the label “Y/@”. Clearly, recognition for the new class is much better than that of either /ɣ/ or

/ə/ separately (initial position:  $p < .02$  for 11 out of 12 comparisons, second position:  $p < .02$  for all 12 comparisons), and is similar to that of the other short vowels. This shows that at gates where at least a third of the vowel is audible, the majority of confusions is indeed between /ʏ/ and /ə/. The remaining confusions are mainly with /œ/ (see Table V).

(4) *long vowels /e o œ/* (Figure 3D). In most regional variants of Dutch, including that of our speaker, these vowels are slightly diphthongized, ending in articulatory positions corresponding to /i u y/, respectively (Booij, 1995). In first position, the phonemes are initially not well recognized. At gate 1, recognition levels are between 15 and 25%, which is much lower than for other vowels discussed so far. At gate 1, /e/ and /o/ are mainly confused with /ɪ/ and /ɔ/, respectively, while /œ/ is mainly confused with /ʏ/ and /ə/ (see Table V). This is partly supported by Booij's (1995) position that the short counterparts of /e/ and /œ/ are indeed /ɪ/ and /ʏ/ (with /ʏ/ and /ə/ being highly confusable, as discussed earlier), while /o/ is higher than /ɔ/. Our data suggest, however, that, perceptually, the relation between /ɔ/ and /o/ is very similar to that between /ɪ/ and /e/. At gate 2, recognition levels are just above 70%, and the full three gates are necessary for recognition to exceed 90%. The recognition results for /e o œ/ in second position are very similar to those for the first position, shifted onwards by three gates.

(5) *vowel /a/* (Figure 3E). This vowel is depicted separately because it shows a pattern between that of /i u y/, which have no short counterpart, and that of /e o œ/, which do. This finding tallies with the description of /a ɑ/ as “almost” a long-short pair, with the qualification that both vowels are back, but /a/ is somewhat fronted compared to /ɑ/ (Booij, 1995). Another aspect which sets /a/ apart from the other long vowels is that its recognition asymptotes just below 90%, whereas the others are eventually recognized at levels close to 100%.

The raw data show that at all gates /a/ is recognized better when stressed than unstressed. When it is unstressed /a/ is mainly confused with /ɑ/ and to a lesser extent with /ə/ and /ɛi/. The pattern is, however, more subtle. When /a/ is part of a VV diphone (which always has a syllable boundary in the middle), and the stress pattern of this diphone is either weak-strong or strong-weak, the confusion with /ɑ/ is much less than when it is part of an unstressed CV or VC diphone, or a VV diphone with a weak-weak stress pattern. We hypothesize that when /a/ is stressed or it is possible to hear that /a/ is unstressed (by contrast to the syllable), listeners are more likely to choose the (correct) /a/ response. The data show that the same general pattern applies to /e/ and /œ/, but the effect is much weaker, possibly due to their slight diphthongization, which makes confusions with their short counterparts less likely.

(6) *diphthongs* /ɛi œy au/ (Figure 3F). The general picture is similar to that for the diphthongized long vowels (Figure 3E), but there is more variability. /au/ is recognized worse than the other two diphthongs ( $p < .0005$  in first position at gate 1,  $p < .05$  in second position at gates 2 to 5), with levels below 10% when a third of the vowel is audible. Not surprisingly, the majority of responses is /ɑ/ for these gates (see Table V). /œy/ is mainly confused with /ɑ/, /a/ and /ə/ at early gates, while /ɛi/ is mainly confused with /ɛ/ (see Table V). When these vowels are fully audible, recognition levels are close to 100%.

#### **D. Phoneme and diphone biases**

At early gates, listeners receive little information about the second phoneme. Categorization behavior at these gates can have several determinants. First, listeners may develop a strategy of choosing a fixed label for the second phoneme when very little acoustic information about it is available. Second, gating can introduce acoustic cues that were not part of the original

signal and thus can bias responses, usually towards plosive manner and labial place (Ohala and Ohala, 1995; Smits, 2000). Although we strove to minimize such biases by gating to a square wave and ramping the signal down over a window, we cannot exclude the possibility that some remained. Third, listeners may have used knowledge of phoneme frequencies and transitional probabilities in Dutch to bias their guesses toward the most frequently occurring phonemes, or the most probable second phoneme given a particular perceived first phoneme.

----- insert Figure 4 about here -----

Figure 4 displays the percentages of trials that subjects used each of the response categories /ə h m n p/ for the second phoneme as a function of gate. At gate 1, these five response categories were the most frequently used (the percentages for the remaining 33 phonemes are not shown). The figure shows that, in the absence of acoustic information, that is, at gates 1 to 3, subjects responded far from randomly. The fact that neither plosive nor labial responses were highly elevated for early gates shows that we succeeded in limiting biases due to gating-induced discontinuities. At gate 1, however, when there is very little information about the second phoneme, the response /h/ was given in over 25% of cases, while the response /ə/ was given in 8% of cases (see Tables IV and V). As we have already suggested, we interpret this not as evidence that in these cases subjects in fact heard an upcoming /h/, or /ə/, but rather as evidence that when subjects really did not know how to respond, they selected a default label, some using /h/, others both, yet others /h/ or /ə/ with another default. Others seemed to guess more randomly.

----- insert Figure 5 about here -----

Figure 5 shows observed response probabilities of all phonemes in second position at gate 1 (i.e., very little acoustic information about the phoneme is available) plotted against phoneme (token) probabilities from CELEX. If subjects guessed purely according to phoneme occurrence frequencies, the points would lie along the main diagonal. The actual correlation between log observed and log predicted frequencies is .66 ( $p < .0005$ ). This suggests that listeners did use phoneme frequencies in second-phoneme responses at gate 1. The correlations for subsequent gates gradually decrease, as do their significance levels, until, at gate 6, where the second phoneme is accurately recognized, the correlation is no longer significantly different from zero and the points are close to a horizontal line.

Consonants have a heavier component in the correlation than the vowels ( $r = .71$ ,  $p < .0005$  for consonants versus  $r = .55$ ,  $p < .05$  for vowels), with /dʒ ʒ ʃ/ carrying most of the correlation. These three consonants are originally foreign to Dutch, although /ʃ/ does play an important role in diminutives. If /dʒ ʒ ʃ/ are removed, the correlation is no longer significant. Note that the low response frequencies for /dʒ/ are probably at least in part caused by acoustic complexity of /dʒ/, which exhibits three distinct phases: closure, release, friction. In our discussion of Figure 2B we noted that listeners do not use the affricate response unless all necessary components are audible. Thus at early gates /dʒ/ responses will be rare, not because of a frequency bias, but because subjects choose phonemes that the input at that point is most similar to, namely /j/ or /d/ (see Ohala and Ohala, 1995, for a similar argument). On this account, a low score would also appear for listeners in whose native language /dʒ/ is frequent. In summary, the use of phoneme frequency in the responses to upcoming phonemes is, though significant, not strong, and is mostly limited to the originally non-native consonants /dʒ ʒ ʃ/.

Finally, we investigated whether listeners exploited transitional probabilities in second phoneme responses. A transitional probability  $p(\varphi_2 | \varphi_1)$  is the conditional probability of observing (or responding) phoneme  $\varphi_2$  given that the preceding phoneme was  $\varphi_1$  (e.g., Pitt and

McQueen, 1998). In our analysis we focused on gate 2, where listeners were generally able to make a reasonable guess at the first phoneme of the diphone, while the second phoneme was still difficult to recognize. For each phoneme in initial position, we calculated the correlation between the logarithm of the probabilities of all possible subsequent phonemes as predicted by CELEX, and the logarithm of the corresponding observed probabilities. Of 38 correlation coefficients, only the one for /au/ as initial phoneme proved significant at the  $p < .05$  level ( $r = -.68$ ,  $p < .01$ , which is negative). Inspection of the data revealed that when the first phoneme was recognized as /au/, listeners were likely to guess /w/ for the second phoneme. In CELEX, however, it is rare for /w/ to follow the diphthong /au/. It is possible that judgments in this particular case were contaminated by orthographic knowledge: In Dutch, although /au/ can be followed by coda consonants (as in *fout*, “fault”, or *kous*, “stocking”), in syllable-final position it is written as either “auw” or “ouw”. In any case, the data clearly show that transitional probabilities did not play a significant role in listeners’ response behavior.

### **E. Additional analyses**

We also investigated recognition rates as a function of flanking phonetic context; here we found only small and mostly insignificant effects. The influence of stress on recognition was even smaller than that of context. Subsequent reports will describe these analyses in full.

## **IV. DISCUSSION**

Our data set constitutes the largest speech perception database we know of (an order of magnitude larger even than the seminal and much-reanalyzed study of Miller and Nicely, 1955, who collected 68,000 phoneme categorizations). It will be of use in the testing of

hypotheses concerning the temporal unfolding of phonetic information, or dependencies between the categorizations of successive phonemes. In the near future we will issue a CDrom containing the raw data set to facilitate such studies.

Our results showed that the unfolding of perceptually relevant phonetic information differs substantially for different phonemes. Thus stop consonants are not recognized very well until the release burst is audible, and correct recognition of fricatives suddenly improves when a portion of the frication noise is audible. Voiced stops and fricatives were recognized consistently worse than their voiceless counterparts, mainly due to voicing errors. The recognition of glides and liquids improved very gradually with increasing gates.

In the recognition of vowels we found two major patterns. The long vowels /i y u/, which have no short counterparts, and the short vowels /a ε ɪ ə/ were recognized very well as soon as one third of their full extent was audible. The long vowels /e o œ/, which do have short counterparts, and the diphthongs /ɛi œy au/, on the other hand, were only recognized well if they were audible in full. The pattern for the vowel /a/ fell between that of the long vowels with and without short counterparts, except that at all gates there was some confusion with /a/, especially when it was unstressed. In addition, we found that the short vowels /ʏ ə/ were highly confusable with each other, and were effectively treated as a single category.

Finally we investigated the extent to which listeners made use of statistical knowledge of the Dutch phoneme inventory in the form of phoneme biases and transitional probabilities. We found that phoneme biases were only used only with regard to the low-frequency (and originally foreign) consonants /dʒ ʒ ʃ/. We found no evidence that listeners employed knowledge of transitional probabilities in their response behavior.

The ultimate purpose of the research presented in this paper is to make models of spoken-word recognition more realistic. As we described in the introduction, several lines of research suggest that human prelexical processing does not have a discrete phonemic stage

and that, instead, lexical access is continuous. The current data are consistent with this view. They also provide detailed measures of the perceptual hypotheses that listeners entertain as they listen to speech. These measures could be used to revise the estimates that models make about which words are activated, and when, by any given Dutch diphone in the speech input. Note, however, that the present data, and the other evidence on continuous lexical access, are not in fact inconsistent with models using phonemic prelexical and lexical representations. What that evidence suggests is that lexical access is continuous and probabilistic, rather than discontinuous and categorical. As McQueen *et al.* (1999) have argued, a word-recognition system with phonemic representations at both prelexical and lexical levels (like TRACE, McClelland and Elman, 1986, or Shortlist, Norris, 1994) could account for the data on continuous lexical access if information is allowed to cascade through the system. Models with other representational assumptions (e.g., features in the Cohort model, Marslen-Wilson, 1987, and in the Distributed Cohort Model, Gaskell and Marslen-Wilson, 1997; allophones in PARSYN, Luce, Goldinger, Auer and Vitevitch, 2000) can also account for those results, if again information is allowed to flow continuously up to the lexicon.

Spoken-word recognition theory, as instantiated in all these models, assumes that recognition involves the activation of multiple candidate words, followed by competition among those words until the word or words which best match the input can be selected. A central goal of psycholinguistics, therefore, has been to establish the characteristics of the lexical competitor set. Which words are activated, and to what extent, by a given input? The available models differ widely in the assumptions they make about goodness of fit in the lexical activation process, and in the dynamics of this process (when words become activated or deactivated). But because computation of goodness of fit is in general based on phonemes, it is too coarse-grained. In Shortlist, for example, words are activated depending on the number of matching and mismatching phonemes (Norris, 1994). But if information is indeed

mapped continuously to the lexicon, candidate-set membership will not depend on categorical phonemic differences. It will instead depend on subphonemic measures of goodness of fit.

Progress in modeling human spoken word recognition therefore depends on better estimates of what words will be activated, and when, by any given speech input. The present data can be used to provide such estimates. Given these estimates, categorical decision rules for lexical candidate-set membership could be abandoned. Data from perceptual confusions (Fletcher, 1953; Hillenbrand, Getty, Clark and Wheeler, 1995; Miller and Nicely, 1955; Peterson and Barney, 1952; Wang and Bilger, 1973) could in principle be used as a metric for goodness of fit at the lexical level, and indeed have been used in modeling spoken word recognition (Luce, 1986; Luce and Pisoni, 1998; Luce *et al.*, 2000). However, while the approach of Luce and colleagues has been extremely valuable, traditional confusion matrix data have two serious limitations. First, the perceptual identification data may be limited to a particular class of sounds (e.g., only consonants, Miller and Nicely, 1955, or only vowels, Peterson and Barney, 1952), or to materials with a particular structure (e.g., Consonant-Vowel-Consonant [CVC] syllables, Peterson and Barney, 1952). Such limitations are very problematic if the aim is to model spoken word recognition, where data is required on all types of sounds in all possible structural positions. Second, traditional confusion matrix data does not have sufficient temporal resolution. Identification responses based on entire syllables give no information on the subphonemic time-course of perception.

The current data set of responses to diphones thus provides a clear advance on previously available corpora, and enables a start to be made on modeling goodness of fit between competing word candidates in a more realistic manner.

## **Acknowledgments**

We are grateful to Dennis Norris for discussion of this material. We further thank Mattijn Morren, Keren Shatzman, Petra van Alphen, Niels Janssen, Tau van Dijck, Anne-Pier Salverda, and Aaju Chen for their great efforts in preparing and running the experiment.

## Appendix A

Reasons for selection of certain phonemes and diphones:

1. Besides the voiceless velar fricative /x/, both CELEX and Booij (1995) recognize the voiced velar fricative /ɣ/. We excluded this phoneme because many Dutch speakers - including the speaker for the experiment - neutralize the distinction, maintaining only /x/ (Gussenhoven 1992).
2. The vowels /i:, y:, u:, ɔ:, œ:, ε:/ occur only in unassimilated loanwords (such as “analyse, centrifuge, cruise, zone, oeuvre, serre” respectively) and differ from native vowel phonemes only in length. We excluded these nonnative vowels because Gussenhoven (1992) and Booij (1995) both state that these vowels are marginal and only occur in a small number of recent loanwords.
3. We did include a number of consonants which occur in Dutch only in unassimilated loanwords. We included the voiced velar stop /g/, the fricative /ʒ/ and the affricate /dʒ/ since they appear in a relatively large number of loanwords, many of them quite frequent (e.g., “goal”; “jam”, /ʒɛm/; and “jazz”).
4. There are certain inconsistencies in the CELEX inventory, such as the fact that [tʃ] is treated as a sequence of a stop and a fricative, /tʃ/, while [dʒ] is treated as a single affricate segment /dʒ/. In these cases we adhered to the CELEX standard.

## Appendix B

The following rules were applied in selection of the diphones:

1. For each sequence of two phonemes containing a vowel other than /ə/ (which is never stressed), one diphone was included with the vowel stressed, and another with it unstressed. For vowel-vowel diphones, all four stress combinations (stressed-stressed, unstressed-unstressed, stressed-unstressed, unstressed-stressed) were included.
2. We included diphones which can only occur across word or morpheme boundaries in Dutch (e.g., /ŋp/), but we excluded diphones which, because of phonotactic constraints, could never occur even across word boundaries.
3. In cases where phonotactic constraints were violated by large numbers of loanwords, we included the diphones. For example, Booij (1995) states that short vowels cannot be followed by a glide within the syllable, which is violated by loanwords like “timing,” “tranquilizer,” and “boiler”.
4. We excluded certain diphones which are possible (at least across morpheme boundaries) according to a phonemic transcription, but unlikely ever to be produced as a sequence of the two sounds, like /sʃ, ʃs, tɖʒ/.
5. We excluded all sequences of identical consonants ( $C_1C_1$ ), since Dutch phonology requires that these be degeminated within the prosodic word (Booij 1995), and they are likely to be reduced to a single consonant even across word boundaries unless produced with a pause.
6. A few diphones which probably never occur in Dutch, e.g., /ɑ, ε, ʏ/ followed by /ʒ/, were included simply because no known phonotactic constraint excludes them.

## Appendix C

Table C.I. Diphones included in the experiment, and reasons for exclusions. Each row of represents diphones  $X_1X_2$ , where  $X_1$  is each of the phonemes in the  $X_1$  column and  $X_2$  is each of the phonemes in the  $X_2$  column.

Class	$X_1$	$X_2$
CV diphones		
C=stop, affricate, nasal, liquid, or glide	p, t, k, b, d, g, dʒ, m, n, ŋ, r, l, j, w	all full vowels stressed, all vowels unstressed
C=fricative	f, v, s, z, ʃ, ʒ, x, h  f, v, s, z, ʃ, ʒ, x  h  Exclusion: */hə/ within the syllable, and /h/ cannot be syllable-final <sup>1</sup>	all full vowels stressed  all vowels unstressed  all full vowels unstressed
VC diphones		
C=stop, affricate, liquid or glide	all full vowels stressed, all vowels unstressed	p, t, k, b, d, g, dʒ, r, l, j, w
C=fricative	all full vowels stressed, all vowels unstressed  all long vowels and diphthongs stressed; all long vowels, diphthongs, and /ə/ unstressed  Exclusion: short vowels before /v, z/ not possible within the syllable, and short vowels cannot be syllable-final <sup>2</sup>	f, s, ʃ, ʒ, x, h  v, z

C=nasal	all full vowels stressed, all vowels unstressed all short vowels stressed; all short vowels and /ə/ unstressed  Exclusion: /ŋ/ cannot follow long vowels within the syllable <sup>3</sup> and cannot be syllable-initial	m, n  ŋ
VV diphones		
stressed-unstressed	all long vowels and diphthongs	all vowels
unstressed-stressed	all long vowels, diphthongs, and /ə/	all vowels except /ə/
unstressed-unstressed	all long vowels, diphthongs, and /ə/	all vowels
stressed-stressed	all long vowels and diphthongs	all vowels except /ə/
	Exclusion for all VV categories: short vowels cannot be V <sub>1</sub> because they cannot be syllable-final	
CC diphones		
C <sub>1</sub> =voiceless stop, nasal, liquid, or glide	p, t, k, m, n, ŋ, l, r, j, w  Exclusion: /ŋ/ cannot follow a stop or another sonorant within the syllable or be an onset	all consonants except C <sub>1</sub> =C <sub>2</sub> and /ŋ/
C <sub>1</sub> =voiced stop	b	d, g, dʒ, v, z, ʒ, n, l, r

	<p>d</p> <p>b, g, v, z, ʒ, m, n, r, j, w</p> <p>Exclusions for /b, d/: */bw, bj, bm, dl/ in syllable onset, and voiced stops must devoice if not in onset unless followed by a voiced obstruent<sup>4</sup>; cannot be followed by /ŋ/ because /ŋ/ cannot be an onset</p>
	<p>g</p> <p>b, d, v, z</p> <p>Exclusions: syllable-final /g/ without devoicing is only followed by these consonants, and /g/ is never word-final<sup>5</sup></p>
C <sub>1</sub> =fricative	<p>f</p> <p>all consonants except f, v, ŋ</p> <p>Exclusion: /fv/ too difficult for speaker to produce without assimilation</p> <p>s, ʃ</p> <p>all consonants except s, ʃ, ŋ</p> <p>Exclusions: /sʃ/ and /ʃs/ are unlikely, unless assimilated</p> <p>x</p> <p>all consonants except x, ŋ</p> <p>v</p> <p>b, d, g, z, ʒ, dʒ, n, l, r</p> <p>Exclusions: */vj, vw, vm/ as onsets and /v/ must devoice if not in onset</p> <p>z</p> <p>b, d, g, v, dʒ, m, n, j, w</p> <p>Exclusions: */zl, zr/ as onsets and /z/ must devoice if not in onset; /zʒ/ is likely to assimilate</p> <p>Exclusion for /v, z/: cannot be followed by a voiceless fricative within the syllable, and will devoice in coda position unless followed by a voiced obstruent</p> <p>ʒ</p> <p>w</p>

	Exclusions: /ʒ/ never occurs syllable-finally and in onset occurs only before vowels or /w/ (e.g., in “bourgeois”)  Exclusion for all fricatives: /ŋ/ cannot follow a fricative within the syllable and cannot be an onset	
C <sub>1</sub> =affricate	dʒ	m
	Exclusions: /dʒ/ never occurs word-finally, occurs syllable-finally only in the word “management,” and cannot be followed by any other consonant within an onset  Exclusion for all CC diphones: no geminates	

<sup>1</sup>CELEX does list three forms with /hə/, all based on the word “coherent.”

<sup>2</sup>Short vowel-/h/ diphones should be impossible, and thus should have been excluded, since short vowels cannot be syllable-final and /h/ cannot be in a coda. Also, although Booij (1995) states the prohibition of short vowels followed by /v, z/ within the syllable as a phonotactic constraint, another rule in the phonology voices underlying /f, s/ before a voiced stop (Booij 1995). Thus, a short vowel can be followed by [v, z] if a voiced stop follows, as in “zesde” [zɛzdə] ‘sixth,’ “afdeling” [ɑvdɛlɪŋ] ‘department,’ etc. These diphones should have been included.

<sup>3</sup>Although Booij (1995) states this phonotactic constraint, CELEX includes many words with long vowels followed by [ŋ]. However, the [ŋ] is always derived from underlying /n/ by assimilation to a following velar, e.g., “aangelegenheid” ‘affair,’ “woonkamer” ‘living room.’ Place assimilation in these cases tends to be optional.

<sup>4</sup>Booij (1995) states that coda voiced stops only remain voiced if followed by another voiced stop, not a voiced fricative or a sonorant. Since /bv, dz/ etc. are unlikely onsets, these diphones, as well as /bn, dm/ etc., may also be impossible. We included them since Booij mentions that some stop-fricative and stop-nasal onsets do occur in a few words. CELEX

lists words with the excluded diphones /bw/ (“clubwedstrijd” ‘club contest’), /bj/ (“objectief”), /bm/ (“schrabmes” ‘scraping knife’), and /dl/ (“woordloos” ‘wordless’), but in all these cases the voiced stop is in coda position and should be devoiced.

<sup>5</sup>/gr, gl/ do occur as onsets in some loanwords (e.g., “groupie, glamour”) and should have been included.

## Appendix D

The following criteria were applied in placing the gate endpoints:

1. The boundary between a *nasal* and a neighboring vowel or non-nasal consonant was identified as the point in the spectrogram with a sudden change in spectral distribution of the energy. For nasal-nasal diphones, the boundary between the nasals was located based on spectral changes during the nasalization.
2. The boundary between a *voiceless fricative* and a vowel was considered to be at the onset or cessation of voicing as identified from the waveform and the voice bar.
3. The boundary between *voiced fricatives*, including /h/, and vowels was positioned at the onset or cessation of the first formant of the vowel.
4. The *liquid* /r/ was most often produced as a trill, in which case the amplitude low point for the first tap of the trill, as determined from the waveform, was taken as the onset of trilled /r/. Trilled /r/ often had a slight burst at the end of the final tap, and the end of the /r/ was judged to be just after this burst, or at the amplitude low point if there was no such burst. When /r/ was produced as an approximant or a fricative, changes in formants or frication had to be used as the boundaries instead.
5. For the *liquid* /l/ in syllable onsets, a sudden change in the distribution of energy was usually visible in the spectrogram, as for nasals, and this was taken as the segment boundary. Coda /l/ was rather dark, and the moment of maximum decline of energy in the first and second formants of the preceding vowel was taken as the onset of /l/.
6. The boundary between a *glide* and a vowel was identified as the point halfway through the duration of the F2 transition. Boundaries between glides and most consonants could be determined based on the criteria for the other consonant, and for the boundary within a glide-glide diphone, the same criteria as for glide-vowel sequences were used. /l, j, w/

initial to the item sometimes had voiceless frication before onset of voicing, and this was included as part of the segment.

7. In *vowel-vowel* diphones, creaky voicing, the silence of a glottal stop, or both separated the two vowels. The end of the first vowel was identified as the onset of creaky voicing, or the beginning of silence for a glottal stop if no creaky voice was present. Gate end points were placed at one third and two thirds of the duration of the first vowel, and at the end of the first vowel. The first gate end point for the second vowel was placed at the end of creaky voicing or the end of the glottal stop if no creaky voice was present. The third gate end point was placed at the end of the second vowel, and the second gate end point was placed halfway between the other two gating points.
8. For *voiceless stops* which were recorded in an intervocalic environment, the boundaries of the stop were identified as the cessation of voicing for the preceding vowel and onset of voicing for the following one, as determined from the waveform and the voice bar. The second gating point within the stop was placed just before the beginning of the stop burst, and the first gating point was placed halfway through the silence. For voiceless stops preceding or following other consonants, the beginning of the stop was also identified as the beginning of the silence for the stop closure, and the end as the end of the burst and onset of the following segment. For voiceless stops initial to the item, only one gate end point was used for the stop itself, at the end of the stop (onset of voicing of following segment if voiced, or at onset of other criteria for the following segment, such as frication, if voiceless). This is because the usual earlier gate end points, during the stop closure, would produce stimuli containing no speech signal at all. Therefore, diphones with a voiceless stop as the first phoneme, if recorded without preceding environment, had only four gates instead of the usual six.
9. For phonemically *voiced stops*, whether intervocalic, initial to the item, or adjacent to a consonant, if prevoicing was produced, the beginning of prevoicing was identified as the

beginning of the stop. The end of the stop was defined as the end of the burst. If voicing ceased during the stop, the resumption of voicing for the following segment was counted as the end of the burst, otherwise the end of the burst had to be located from the spectrogram. For voiced stops with prevoicing, the first gate end point within the stop was placed halfway through the prevoicing, the second just before the beginning of the burst, and the last at the end of the burst. Thus, diphones with such a voiced stop had six gates. However, the stops /b, d, g/ in Dutch, although said to be fully voiced even initially rather than unaspirated as in English, are often produced without prevoicing (van Alphen, 2000). If no prevoicing was visible in the waveform at all in initial position, gate end points were placed as for a voiceless stop, producing four gates for the diphone, but if any prevoicing was visible, the stop was treated as a prevoiced voiced stop.

10. The only affricate, /dʒ/, was treated as a voiced stop for purposes of positioning gate end points (first gating point halfway through the prevoicing, second just before the burst, and third at the end of the affricate). No gate end point was placed at the end of the burst noise before the frication noise, since the burst without the affrication is rather brief.

## References

- van Alphen, P. (2000). "Does subcategorical variation influence lexical access?," in *Proceedings of the Workshop on Spoken Word Access Processes*, edited by A. Cutler, J. M. McQueen and R. Zondervan (MPI for Psycholinguistics, Nijmegen), 55-58.
- Andruski, J. E., Blumstein, S. E., and Burton, M. (1994). "The effect of subphonetic differences on lexical access," *Cognition* **52**, 163-187.
- Baayen, H., Piepenbrock, R., and Rijn, H. van (1993). *The CELEX Lexical Database (CD-ROM)*. (Linguistic Data Consortium, University of Pennsylvania, Philadelphia).
- van Bergem, D. (1995). *Acoustic and lexical vowel reduction*, *Studies in Language and Language use 16*. Unpublished doctoral dissertation (U. Amsterdam, Amsterdam).
- Booij, G. (1995). *The phonology of Dutch* (Clarendon Press, Oxford).
- Cutler, A., and Otake, T. (1999). "Pitch accent in spoken-word recognition in Japanese," *J. Acoust. Soc. Am.* **105**, 1877-1888.
- Dahan, D., Magnuson, J. S., Tanenhaus, M. K., and Hogan, E. M. (2001). "Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition," *Language and Cognitive Processes* **16**, 507-534.
- Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (1977). "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," *Perception & Psychophysics* **22**, 109-122.
- Ellis, L., Derbyshire, A. J. and Joseph, M. E. (1971). "Perception of electronically gated speech," *Language and Speech* **14**, 229-240.
- Fletcher, H. (1953). *Speech and hearing in communication* (Kreiger, New York).
- Fowler, C. A. (1984). "Segmentation of coarticulated speech in perception," *Percept. Psychophys.* **36**, 359-368.

- Gaskell, M. G., and Marslen-Wilson, W. D. (1997). "Integrating form and meaning: A distributed model of speech perception," *Language and Cognitive Processes* **12**, 613-656.
- Grosjean, F. (1996). "Gating," *Language and Cognitive Processes* **11**, 597-604.
- Gussenhoven, C. (1992). "Illustrations of the IPA: Dutch," *J. Intern. Phonet. Assoc.* **22**, 45-47.
- Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.* **97**, 3099-3111.
- Klatt, D. H. (1980). "Speech perception: A model of acoustic-phonetic analysis and lexical access," in *Perception and Production of Fluent Speech*, edited by R. A. Cole (Erlbaum, Hillsdale), pp. 243-288.
- Lindblom, B. (1963). "Spectrographic study of vowel reduction," *J. Acoust. Soc. Am.* **35**, 1773-525.
- Luce, P. A. (1986). "Neighborhoods of Words in the Mental Lexicon," *Research on Speech Perception Technical Report No. 6*, (Speech Research Laboratory, Psychology Department, Indiana University, Bloomington).
- Luce, P. A., Goldinger, S. D., Auer, E. T., and Vitevitch, M. S. (2000). "Phonetic priming, neighborhood activation, and PARSYN," *Percept. Psychophys.* **62**, 615-625.
- Luce, P. A., and Pisoni, D. B. (1998). "Recognizing spoken words: The Neighborhood Activation Model," *Ear and Hearing* **19**, 1-36.
- Marslen-Wilson, W. D. (1987). "Functional parallelism in spoken word-recognition," *Cognition* **25**, 71-102.
- Marslen-Wilson, W., and Warren, P. (1994). "Levels of perceptual representation and process in lexical access: Words, phonemes, and features," *Psychol. Rev.* **101**, 653-675.
- Martin, J. G., and Bunnell, H. T. (1981). "Perception of anticipatory coarticulation effects," *J. Acoust. Soc. Am.* **69**, 559-567.

- Martin, J. G., and Bunnell, H. T. (1982). "Perception of anticipatory coarticulation effects in vowel-stop-vowel sequences," *J. Exp. Psychol. Hum. Percept. Perf.* **8**, 473-488.
- McClelland, J. L., and Elman J. L. (1986). "The TRACE model of speech perception," *Cogn. Psychol.* **18**, 1-86.
- McQueen, J. M. (1995). "Processing versus representation: Comments on Ohala and Ohala," in *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvaniti (Cambridge University Press, Cambridge), pp. 61-67.
- McQueen, J. M., Norris, D., and Cutler, A. (1999). "Lexical influence in phonetic decision making: Evidence from subcategorical mismatches," *J. Exp. Psychol. Hum. Percept. Perf.* **25**, 1363-1389.
- Miller, G. A., and Nicely, P. E. (1955). "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.* **27**, 338-352.
- Norris, D. (1994). "Shortlist: a connectionist model of continuous speech recognition," *Cognition* **52**, 189-234.
- Ohala, J. J., and Ohala, M. (1995). "Speech perception and lexical representation: the role of vowel nasalization in Hindi and English," in *Phonology and Phonetic Evidence, Papers in Laboratory Phonology IV*, edited by B. Connell and A. Arvaniti (Cambridge University Press, Cambridge), pp. 41-60.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**, 175-184.
- Pitt, M. A. & McQueen, J. M. (1998). "Is compensation for coarticulation mediated by the lexicon?" *Journal of Memory and Language* **39**, 347-370.
- Pols, L. C. W., and Schouten, M. E. H. (1978). "Identification of deleted consonants," *J. Acoust. Soc. Am.* **64**, 1333-1337.
- Smits, R. (2000). "Temporal distribution of information for human consonant recognition in VCV utterances," *J. Phonetics* **27**, 111-135.

- Smits, R. (2001). "Evidence for hierarchical categorization of coarticulated phonemes," *J. Exp. Psychol. Hum. Percept. Perf.* **27**, 1145-1162.
- Smits, R., Ten Bosch, L., and Collier, R. (1996). "Evaluation of various sets of acoustical cues for the perception of prevocalic stop consonants: I. Perception experiment," *J. Acoust. Soc. Am.* **100**, 3852-3864.
- van Son, R. J. J. H., and Pols, L. C. W. (1999). "An acoustic description of consonant reduction," *Speech Communication* **28**, 125-140.
- van Son, R. J. J. H., and Pols, L.C.W. (1997). "The correlation between consonant identification and the amount of acoustic consonant reduction," *Proc. Eurospeech 97*, 2135-2138.
- Stevens, K. N. (1998). *Acoustic phonetics* (MIT Press, Cambridge).
- Streeter, L. A., and Nigro, G. N. (1979). "The role of medial consonant transitions in word perception," *J. Acoust. Soc. Am.* **65**, 1533-1541.
- Wang, M. D., and Bilger, R. C. (1973). "Consonant confusions in noise: a study of perceptual features," *J. Acoust. Soc. Am.* **54**, 1248-1266.
- Warner, N. (1998). *The Role of Dynamic Cues in Speech Perception, Spoken Word Recognition, and Phonological Universals*. Unpublished doctoral dissertation (U. California, Berkeley).
- Warren, P., and Marslen-Wilson, W. D. (1987). "Continuous uptake of acoustic cues in spoken word recognition," *Percept. Psychophys.* **41**, 262-275.
- Whalen, D. H. (1984). "Subcategorical phonetic mismatches slow phonetic judgments," *Percept. Psychophys.* **35**, 49-64.
- Whalen, D. H. (1989). "Vowel and consonant judgments are not independent when cued by the same information," *Percept. Psychophys.* **46**, 284-292.
- Whalen, D. H. (1991). "Subcategorical phonetic mismatches and lexical access," *Percept. Psychophys.* **50**, 351-360.

Table I. The 16 Dutch vowels used in the experiment<sup>1</sup>.

	Front unrounded			Front rounded			Central	Back		
	Diphthong	Long	Short	Diphthong	Long	Short		Diphthong	Long	Short
High		i			y				u	
Mid		e	ɪ, ɛ		œ	ʏ	ə		o	ɔ
Low									a	ɑ
	ɛi			œy				au		

<sup>1</sup>Compared to Booij (1995), we have simplified the vowel system slightly by combining upper and lower mid vowels into a single height.

Table II. The 22 Dutch consonants used in the experiment.

	Labial/ Labiodental		Alveolar		Postalveolar/ Palatal		Velar/ Uvular		Glottal	
	Voiceless	Voiced	Voiceless	Voiced	Voiceless	Voiced	Voiceless	Voiced	Voiceless	Voiced
Stops	p	b	t	d			k	g		
Nasals		m		n				ŋ		
Fricatives	f	v	s	z	ʃ	ʒ	x <sup>1</sup>			h
Affricate						dʒ				
Liquids				l				r <sup>2</sup>		
Glides		w <sup>3</sup>				j				

<sup>1</sup>This fricative is /χ/, but for ease of transcription we will use /x/.

<sup>2</sup>This liquid is /ʀ/, but for ease of transcription we will use /r/.

<sup>3</sup>This glide is /ʋ/, but for ease of transcription, we will use /w/.

Table III. Environments in which diphones were recorded (in phonemic transcription).

Syllable boundaries are marked by hyphens.

Diphone class	Environment	Proportion with each environment
CV (stressed)	<sup>1</sup> CV-kə	2/3
	a- <sup>1</sup> CV-kə <sup>1</sup>	1/3
CV (unstressed)	CV- <sup>1</sup> ke	2/3
	<sup>1</sup> a-CV-ke	1/3
VC (vowel stressed)	<sup>1</sup> V-Cə	1/2
	<sup>1</sup> bV-Cə	1/2
VC (vowel unstressed)	V- <sup>1</sup> Ce	1/2
	bV- <sup>1</sup> Ce	1/2
CC	<sup>1</sup> CCa	if CC is a legal onset
	<sup>1</sup> aC-Cə	otherwise
VV (stressed-unstressed)	<sup>1</sup> bV-Vk	all
VV (unstressed-stressed)	bV- <sup>1</sup> Vk	all
VV (stressed-stressed)	<sup>1</sup> bV- <sup>1</sup> V-kə	all
VV(unstressed-unstressed)	<sup>1</sup> a-bV-V- <sup>1</sup> ke	all

<sup>1</sup> For all diphones beginning with /ŋ/, /a/ was used as the preceding vowel instead of /a/

because /ŋ/ cannot follow long vowels.

Table IV. Confusion matrix for consonants. Responses were summed across subjects, contexts and stress conditions. For each stimulus, the first row gives responses to gate 1 for consonants in initial position in the diphone, whereas the second row gives responses to gate 4 for consonants in second position. The last column gives the number of vowel responses to each of the consonants.

		response																						
		p	t	k	b	d	g	dʒ	f	s	ʃ	x	v	z	ʒ	h	r	l	w	j	m	n	ŋ	vow
stimulus	p	325	6	3	62	0	0	0	0	0	0	0	0	0	22	0	0	4	0	9	0	0	19	
		331	8	34	187	13	16	0	11	0	0	0	11	0	0	58	2	10	43	15	15	13	5	20
	t	33	235	4	13	81	1	0	3	0	0	0	5	0	0	25	1	3	15	2	0	0	0	11
		28	258	7	12	340	2	9	2	1	1	1	10	0	0	35	3	5	11	13	6	26	3	19
	k	0	0	340	0	0	66	0	0	0	0	3	0	1	0	23	0	1	2	3	0	1	1	9
		26	18	399	11	7	120	1	8	1	0	2	6	0	0	77	6	9	28	13	3	5	5	47
	b	77	0	2	275	9	1	0	3	0	0	0	11	1	0	95	2	0	25	1	76	20	0	14
		18	2	0	566	32	18	0	2	0	0	0	10	2	0	11	2	3	98	4	92	10	4	8
	d	11	29	0	89	116	2	9	6	1	0	0	11	0	0	99	1	9	67	8	48	37	5	10
		6	5	3	45	571	2	1	1	0	0	3	3	0	0	25	0	40	61	8	6	71	3	28
	g	5	0	60	99	19	123	0	1	0	0	6	10	0	0	92	2	4	39	2	58	36	11	9
		6	1	75	82	33	394	4	4	0	1	22	9	1	1	30	4	8	88	22	8	16	11	62
	dʒ	8	46	1	95	35	2	49	5	0	0	0	16	0	0	113	2	2	49	21	70	51	6	5
		9	12	1	10	457	2	148	0	0	0	0	2	4	2	17	3	8	8	94	1	47	3	18
	f	4	0	0	0	1	0	0	646	1	0	9	172	0	0	43	3	0	6	0	2	0	0	13
		3	0	1	0	1	0	0	565	0	0	12	179	0	0	9	0	1	17	3	0	0	0	1
	s	6	2	0	0	0	1	0	1	670	46	0	0	141	3	27	2	0	0	0	1	0	0	0
		3	6	0	1	1	0	0	0	601	26	0	0	107	14	2	2	1	2	1	4	1	0	2
	ʃ	4	0	0	0	0	3	2	1	91	590	1	0	17	139	18	0	0	0	26	1	0	0	7
		2	5	1	1	0	1	19	2	112	522	2	2	15	66	5	0	0	2	10	2	1	0	4
	x	2	1	2	1	0	0	0	6	0	0	784	1	1	0	52	54	0	1	1	2	0	0	10
		0	0	1	0	0	1	0	3	0	1	709	2	0	0	47	19	0	1	1	0	1	0	6
	v	3	2	0	1	1	0	0	126	0	0	0	385	0	0	66	3	0	114	2	4	0	0	13
		1	0	0	0	1	0	0	116	1	0	3	445	2	0	8	4	1	84	9	0	2	0	7
	z	5	2	0	2	2	0	5	0	67	23	0	0	452	96	32	3	16	0	7	1	0	0	7
		2	2	0	1	5	0	7	1	106	20	0	1	394	90	12	0	7	5	16	5	1	0	27
ʒ	3	2	0	0	5	4	9	0	17	86	0	0	44	330	31	2	3	0	28	1	2	0	9	
	0	3	0	0	3	2	32	0	27	115	1	1	136	428	6	1	1	1	57	2	2	0	46	
h	10	1	1	9	1	0	0	25	1	1	16	20	0	0	386	2	6	29	16	5	6	0	5	
	2	0	1	2	1	0	0	10	0	1	6	2	0	0	683	3	7	21	12	0	2	0	57	
r	1	0	9	1	1	3	0	1	2	0	5	4	0	0	174	628	12	18	5	10	10	3	31	
	0	1	0	0	0	0	0	0	1	1	0	4	0	0	16	691	5	6	1	0	1	0	119	
l	5	1	4	2	0	0	0	1	0	0	1	1	0	0	67	0	758	10	21	5	3	0	39	
	0	0	0	0	1	0	0	0	0	0	1	3	0	0	5	6	759	11	5	1	5	2	29	
w	17	0	1	32	3	1	0	0	1	0	0	17	0	0	107	5	20	549	3	101	31	1	29	
	2	5	1	3	6	1	0	10	3	0	1	65	1	0	46	22	19	534	15	6	6	2	98	
j	0	1	1	3	1	0	0	1	4	0	1	3	5	0	84	0	12	4	683	5	4	3	103	
	4	2	2	4	0	0	0	0	3	0	0	0	0	0	20	3	8	15	591	1	3	0	172	
m	5	0	0	3	0	0	0	0	0	0	0	2	0	0	120	1	11	30	11	599	113	2	21	
	0	0	0	1	0	0	0	1	0	0	0	4	0	0	13	2	11	67	1	609	103	20	14	
n	4	0	2	3	2	0	0	0	0	0	0	5	0	0	108	2	9	17	15	140	579	11	21	
	1	0	0	2	3	0	0	2	1	0	1	3	4	0	10	1	25	41	4	88	648	7	23	
ŋ	0	1	2	0	0	1	0	0	0	0	1	1	0	0	18	6	3	2	5	0	28	810	58	
	0	0	1	0	0	0	0	0	0	0	0	0	0	0	8	1	1	2	1	1	13	166	4	

Table V. Confusion matrix for vowels. Responses were summed across subjects, contexts and stress conditions. For each stimulus, the first row gives responses to gate 1 for vowels in initial position in the diphone, whereas the second row gives responses to gate 4 for vowels in second position. The last column gives the number of consonant responses to each of the vowels.

	response																
	α	ε	ɪ	ɔ	ʏ	ə	i	u	y	e	o	æ	a	ɛi	œy	au	cons
α	640	0	0	4	1	9	0	0	0	0	0	0	26	0	1	11	28
	1275	5	1	27	7	13	0	1	0	0	1	0	131	3	29	45	10
ε	0	642	3	0	0	22	0	0	0	1	0	0	0	29	0	0	23
	42	1165	64	2	5	21	2	0	3	15	0	4	37	159	4	0	25
ɪ	2	1	611	0	1	6	32	0	0	30	0	0	0	0	0	0	37
	5	81	1125	2	25	18	90	1	17	127	0	6	4	4	0	0	43
ɔ	3	0	0	634	2	5	0	2	0	0	28	0	0	0	0	0	46
	92	1	2	1291	6	14	1	5	3	0	81	4	1	0	0	1	46
ʏ	0	1	6	1	450	144	0	0	20	1	0	59	0	0	0	0	38
	18	9	5	59	793	404	1	3	36	3	10	119	3	0	9	1	75
ə	10	5	20	8	439	259	4	5	51	4	0	46	12	1	4	0	86
	7	4	21	23	367	205	0	3	21	1	2	53	3	0	1	2	43
i	0	0	34	0	13	1671	0	5	2	0	0	2	0	0	0	0	145
	0	0	163	2	7	13	1260	1	12	3	0	5	1	1	0	0	80
u	0	0	1	18	4	29	0	1732	2	0	2	1	0	0	1	0	82
	0	1	3	47	21	21	1	1307	32	0	4	2	0	0	1	2	106
y	0	0	0	1	59	56	4	4	1588	0	0	4	0	0	0	0	156
	0	0	11	8	104	60	29	115	1048	0	1	8	0	0	4	2	158
e	0	30	1301	0	6	32	6	0	0	411	0	0	1	2	0	0	83
	1	179	989	2	7	11	30	2	0	289	1	2	0	3	1	0	31
o	4	2	0	1189	8	47	0	30	1	0	474	0	1	0	0	0	116
	23	1	0	1136	14	19	0	7	1	0	289	4	1	0	0	7	46
æ	0	9	9	2	1052	400	0	0	5	20	1	290	0	1	2	2	79
	13	4	10	18	814	373	8	5	28	7	4	191	0	0	0	1	72
a	426	23	2	0	0	66	1	1	1	0	0	0	1211	45	10	1	85
	431	90	2	0	8	7	3	0	1	0	1	1	841	76	51	1	35
ɛi	55	828	0	1	0	84	2	1	0	2	0	0	43	815	2	0	39
	149	602	4	0	5	19	3	0	1	3	0	2	248	457	18	3	34
œy	412	78	1	0	12	120	0	0	0	0	0	3	417	135	614	4	76
	602	48	3	2	24	33	1	0	1	1	0	3	452	34	306	12	26
au	1484	1	2	9	3	52	1	0	0	1	0	0	54	0	6	168	91
	1307	3	2	33	2	6	0	0	0	0	1	1	59	0	12	105	17

## Figure captions

Figure 1. Correct phoneme recognition rates as a function of gate, averaged across listeners. Results for vowels only, consonants only, and all phonemes, are given by separate lines. The upper and lower lines are associated with the first and second phoneme in the diphone, respectively. The dotted line indicates chance level (2.63%).

Figure 2. Correct consonant recognition rates plotted separately for each of the 22 consonants. Phoneme symbols are in accordance with IPA, except for J, S, Z, and N, indicating /dʒ ʃ ʒ ŋ/, respectively.

Figure 3. Correct vowel recognition rates plotted separately for each of the 16 vowels. Phoneme symbols are in accordance with IPA, except for A, E, I, O, @, and ø, indicating /ɑ ɛ ɪ ɔ ə œ/, respectively.

Figure 4. The five overall most popular responses for the second phoneme in the diphone as a function of gate.

Figure 5. Overall observed phoneme probabilities for gate 1 of the second phoneme plotted against phoneme probabilities estimated from the CELEX database of written Dutch. Phoneme symbols are in correspondence with IPA, except for A, E, I, K, L, M, O, @, and ø, indicating /ɑ ɛ ɪ ɛi œy au ɔ ə œ/, respectively.

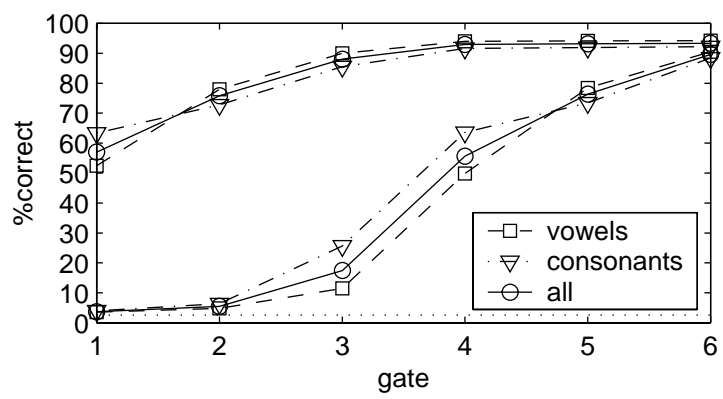


Figure 1

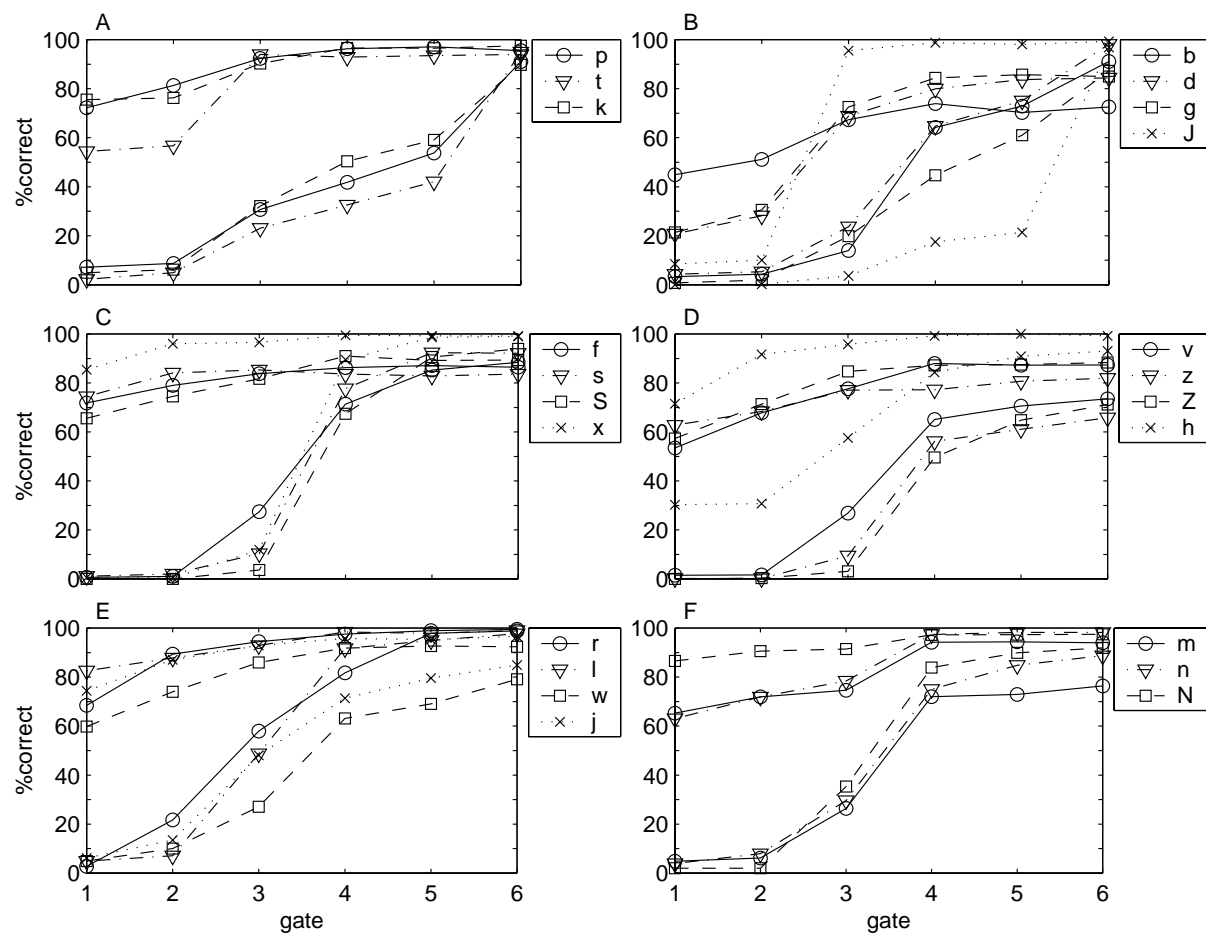


Figure 2

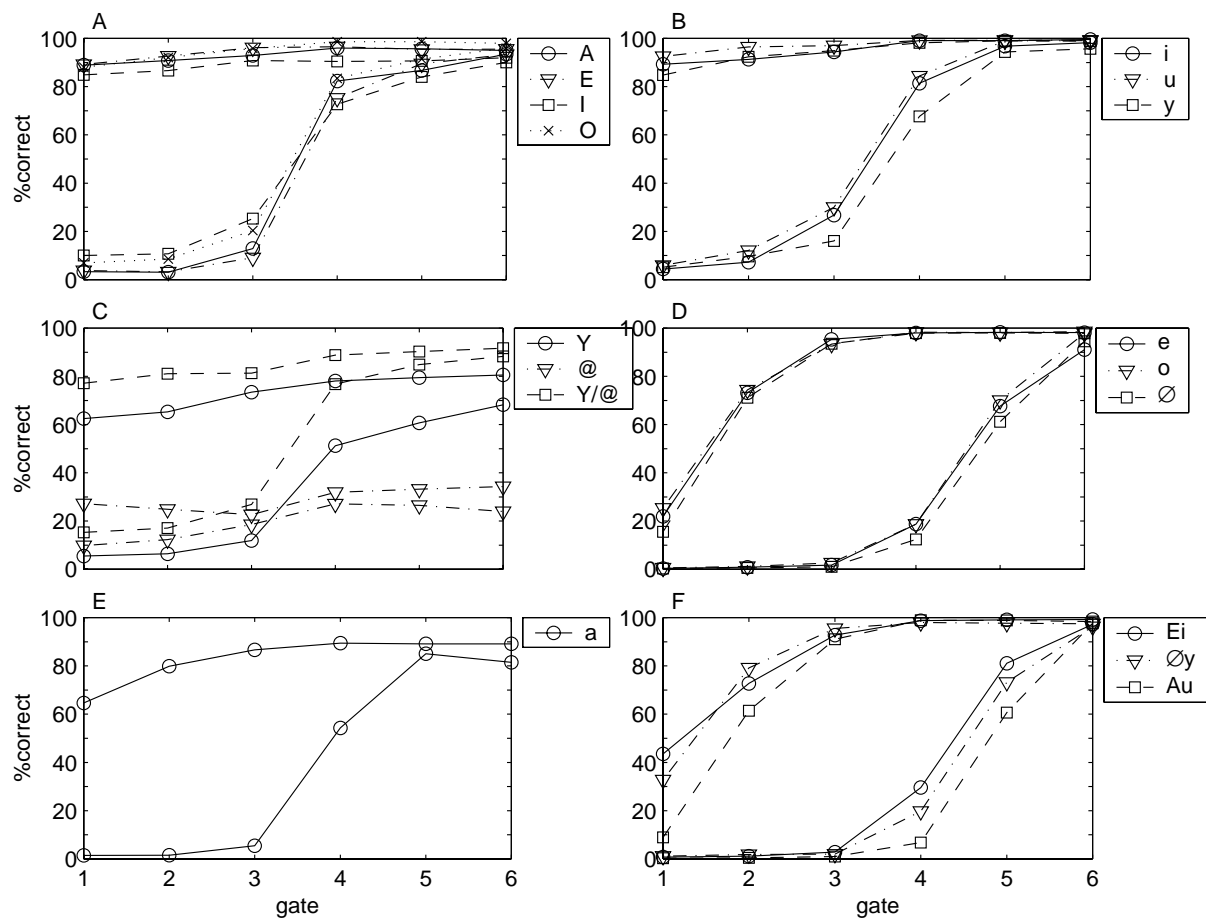


Figure 3

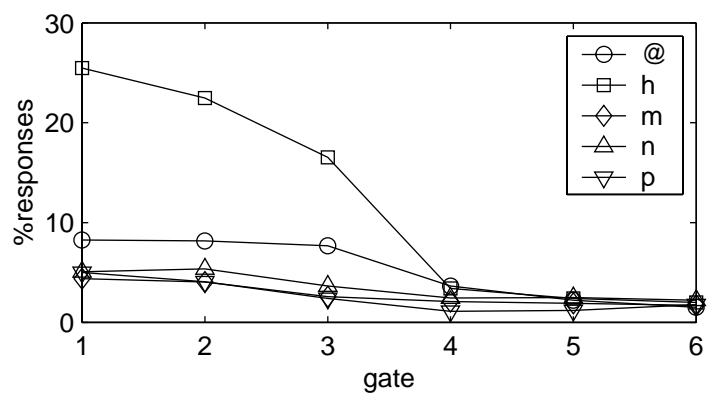


Figure 4

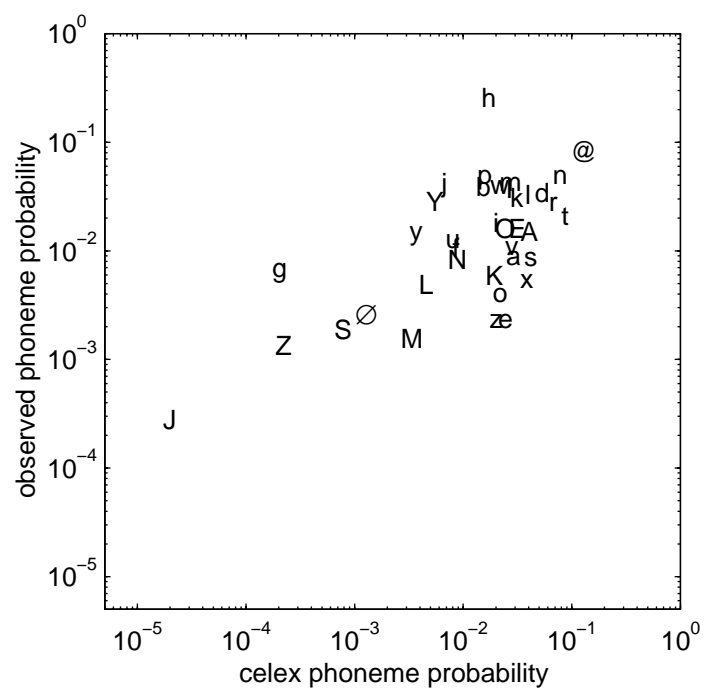


Figure 5