Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Speech Communication 46 (2005) 53–72

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Phonological and statistical effects on timing of speech perception: Insights from a database of Dutch diphone perception

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

Max Planck Institute for Psycholinguistics, Postbus 310, 6500 AH Nijmegen, The Netherlands

Received 18 December 2003; received in revised form 26 January 2005; accepted 27 January 2005

Abstract

We report detailed analyses of a very large database on timing of speech perception collected by Smits et al. (Smits, R., Warner, N., McQueen, J.M., Cutler, A., 2003. Unfolding of phonetic information over time: A database of Dutch diphone perception. J. Acoust. Soc. Am. 113, 563–574). Eighteen listeners heard all possible diphones of Dutch, gated in portions of varying size and presented without background noise. The present report analyzes listeners' responses across gates in terms of phonological features (voicing, place, and manner for consonants; height, backness, and length for vowels). The resulting patterns for feature perception differ from patterns reported when speech is presented in noise. The data are also analyzed for effects of stress and of phonological context (neighboring vowel vs. consonant); effects of these factors are observed to be surprisingly limited. Finally, statistical effects, such as overall phoneme frequency and transitional probabilities, along with response biases, are examined; these too exercise only limited effects on response patterns. The results suggest highly accurate speech perception on the basis of acoustic information alone.

Keywords: Speech perception; Diphone; Timing; Dutch; Feature

1. Introduction

Listeners' recognition of speech requires decisions which are phonemic in nature: for example, that a speaker said *bit* and not *sit*, *but* or *bill*. The identification of phonemic information to motivate such decisions, however, is affected by a multiplicity of factors beyond the acoustic cues which—invariantly or otherwise—directly signal

* Corresponding author. Present address: Department of Linguistics, University of Arizona, P.O. Box 210028, Tucson, AZ 85721-0028 USA. Tel.: +1 520 626 5591; fax: +1 520 626 9014.

E-mail addresses: nwarner@u.arizona.edu (N. Warner), heersmits@hotmail.com (R. Smits), james.mcqueen@mpi.nl (J.M. McQueen), anne.cutler@mpi.nl (A. Cutler).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.01.003

phonemic identity. Thus identification responses are affected by the surrounding phonetic context in which a phoneme occurs, by the phoneme's position in a word or utterance and consequent differences in prosodic realization, as well as by listener expectations based on past experience, as when phoneme frequency effects or transitional probabilities play a role. Decades of speech perception research have been devoted to exploration of these factors (see Nygaard and Pisoni, 1995, for a review).

We here report analyses of these effects in a very large database of perceptual identifications. In speech research, very extensive databases have enabled important advances in our knowledge. Thus Miller and Nicely's (1955) database of perception of consonants in noise, Peterson and Barney's (1952) database on vowels and the subsequent work of Hillenbrand et al. (1995), as also the segment and syllable duration data of Crystal and House (1982, 1988a,b) have all proved treasurehouses for scholars working on a range of speech-related topics. Such extensive databases allow for comparison of many factors with experimental methods held constant, so that the information provided is directly comparable across segment types, stress positions, etc. The database which we describe here concerns perception of segments in Dutch in all possible immediately adjacent contexts. Collected via a gating task, the database gives a temporal view of how Dutch listeners perceive the sounds of every diphone (two-phoneme sequence) in the language, as acoustic information becomes available with each gate.[1]

The choice of diphones as the test set was motivated jointly by considerations of validity and feasibility. For validity, phonemic identification must be assessed in context. Clearly, the goal which listeners aim for in speech recognition is not apprehension of a sequential representation of phonemic units. Listeners want to know what the speaker wished to communicate, i.e. they are interested in meaning, and hence in recognizing the words which comprise an utterance. Phonemes are crucially relevant not because they are an end in themselves, but because they constitute minimal differences between words such as *bit* and *sit* or *but* or *bill*. We therefore wished to examine the uptake of phonemic information in all possible contexts. The larger the context, the better; but even triphone sequences would have presented us with a set of tens of thousands of stimuli, so on grounds of feasibility of data collection we chose diphone sequences. (Even then, there were over a thousand such possible sequences, and by varying stress and presenting the diphones in fragments of varying size, we ended up requiring our listeners to respond to over thirteen thousand stimuli, which took on average 27.9 test hours per listener.) Diphones thus offered the minimal contextual environment for a feasible study of natural perception of phonemic information in speech.

The database itself is publicly available: http://www.mpi.nl/world/dcspdiphones. Smits et al. (2003) describe in detail the methods used to collect the database. That methodological report contained however only the most summary statistics concerning the perceptual findings, namely percent correct judgments per gate for segments individually, and averaged across consonants, across vowels, and across all segments.

The data reported by Smits et al. (2003) nevertheless showed clearly how listeners progress in their perception of sounds, both for the first and the second sounds of a diphone. The most important patterns which Smits et al. observed for consonants were: (1) Stops were not recognized well until listeners could hear their bursts. (2) Voiced obstruents (both stops and fricatives) tended to be misperceived as the voiceless equivalent, but the confusion did not go in the opposite direction. (3) Fricatives could be recognized very well from the first third of the fricative, but not from the preceding vowel, so that improvement in perception of fricatives (both voiced and voiceless) was quite sudden at the first gate that included frication noise. (4) Useful information for perception of nasals was available both in the final portion of the preceding sound and, even more so, in the first

[1] Responses in the gating task, of course, represent listeners' conscious decisions about what sounds they have heard, rather than their online recognition of sounds as a part of spoken word recognition. See Norris et al. (2000) for extensive discussion of this distinction.

portion of the nasal itself. (5) Finally, the perceptual information for glides and liquids was more widely distributed in time than that for other consonants.

For vowels, Smits et al. (2003) found that listeners' confusions primarily reflected length distinctions and diphthongization, since these distinctions are cued by changes in vowels over time or by duration itself. Most phonemically short vowels were recognized well as soon as the first gate within the vowel was heard (one-third through the vowel). Only /ʏ/ and /ə/ were poorly recognized, because they were confused with each other. Among the long vowels of Dutch, some form a pair with a short vowel, while others have no short correlate. Those with no corresponding short vowel were recognized well early in the vowel, while those with a short vowel correlate were misperceived as the short vowel until the end of the vowel. Diphthongs, similarly, were misperceived as the nearest monophthong until the end of the diphthong.

A wealth of further information concerning the factors affecting phonemic identification can be gleaned from the data, and here we present analyses at a number of levels, ranging from factors associated with phoneme identity (phonological feature comparisons) through the effects of phonetic context and stress to higher-level influences of statistical factors such as phoneme frequency and transition probabilities.

2. Methods

Full detail of the data collection methods can be found in Smits et al. (2003). Here we summarize the general method and then focus on methodological points of relevance to the results in this paper.

All possible diphones of Dutch (1179 sequences, most of which occurred in two stress versions, giving in total 2294 diphones) were put into a phonotactically possible nonsense environment for recording, and produced by one phonetically trained female native speaker of Dutch. Each item was final-gated, in most cases with six gates for each item. (Diphones in which the first segment was a stop or affricate with no prevoicing had only four gates.) For segments that remain relatively steady throughout the segment (i.e. nasals, fricatives, monophthongs, etc.), the shortest gate allowed listeners to hear from the beginning of the item (including any preceding environment) up to a point one-third through the first segment of the target diphone. The next continued to two-thirds through the first segment, etc., and the final gate continued up to the end of the second segment of the diphone. For segments with substantial change during the segment, gate end points were located based on acoustic boundaries, as discussed by Smits et al. (2003). This produced a total of 13,570 stimuli. Ideally, one might wish to compare perception of similar stimuli produced by multiple speakers to rule out possible speaker effects. However, such an approach would have made the present study prohibitively large. The emphasis here is on variability of stimulus environments, rather than variability of speakers.

Eighteen listeners each heard all stimuli, with each listener hearing them in a different pseudo-random order. For each stimulus, listeners had to decide what the two segments of the diphone were, among a choice of all segments of Dutch. The resulting database encompasses 488,520 phonemic categorizations, and thus constitutes the largest database of information about timing of speech perception we know of for any language.

It is important to keep in mind that all possible diphones of Dutch were used, including those that can only occur across syllable boundaries or word boundaries. This includes CV, VC, CC, and VV diphones. For each diphone, judgments of both the first and the second sound were collected at all six gates. That is, listeners had to respond with what they thought the *two* sounds were when they heard up to one-third through the first sound, up to two-thirds through it, to the end of the first sound, up to one-third through the second sound, up to two-thirds through it, and up to the end of the diphone. Of course, at the first gate (one-third through the first sound), little or no acoustic information is available about the second sound, but listeners were forced to give responses for both the first and second sound for all stimuli, and were allowed to choose for each sound from the entire phonemic inventory of Dutch. Although all listeners heard all stimuli, stimuli were presented in a pseudo-random order (in which gates of the same diphone or of diphones beginning with the same sound never followed each other too closely). That is, listeners did not hear the various gates of a diphone in order of increasing length, and did not hear the various gates of a diphone together.

For CV, VC, and VV diphones, stress was manipulated where possible. This means that for all CV sound sequences (except those with /ə/ as the vowel, since it cannot be stressed), one diphone with the vowel of the CV stressed and another with it unstressed were used. For VC sequences, if the vowel of the diphone was stressed, and the vowel following the diphone not, the consonant is referred to as a "post-stress" consonant. If the vowel of the diphone was unstressed, and the vowel following the diphone stressed, the consonant is "pre-stress." For VV sequences, all four possible stress combinations (i.e. both vowels stressed, both vowels unstressed, stressed–unstressed, and unstressed–stressed) were used. The speaker inserted a glottal stop (sometimes realized as creaky voice with no silence) between the vowels in VV diphones. For CC diphones, stress location was not manipulated. Nevertheless, certain CC diphones were spoken in /'CCV/ context (pre-stress), whereas others were spoken in /'ɑC-Cə/ context (post-stress).

All diphones were recorded with a surrounding environment, to make them easier for the speaker to pronounce and to prevent them receiving excessive final lengthening. For some types of diphones, more than one environment was used so that listeners could not develop strategies by learning, for example, that VC diphones always followed a particular environment. The length of the environments ranged from just a following /ə/ for one half of the VC diphones to /'abVV'ke/ for the weak–weak VV diphones. (In that case, a stressed syllable on both sides of the VV diphone made it easier for the speaker to pronounce the sequence of unstressed vowels.) Listeners never heard any following environment, since the last gate ended at the end of the diphone. They did hear initial environments (where present), and in these cases the phonemes of the initial environment were printed on the response screen in such a way as to indicate that those sounds preceded the two sounds to which listeners should respond. Details of the environments are available in Smits et al. (2003). All gates were followed by a 300 ms, 500 Hz square wave, which is not perceived as speech (Warner, 1998) and which minimized the illusory perceptions which can arise when speech is truncated to silence.[2]

3. Results

3.1. Perception of phonological features

3.1.1. Consonants

For the purposes of analyzing this experiment, we consider the "features" of consonants to be place, manner, and voicing, rather than the more detailed feature systems (coronal, anterior, consonantal, sonorant, continuant, etc.) used in formal phonology (e.g. Kenstowicz, 1994). We consider the values of "place" within the Dutch consonant inventory to be labial/labiodental, alveolar, postalveolar/palatal, velar/uvular, and glottal. This classification strikes a balance between a very detailed phonetic inventory of places, which would have very few consonants at many places, and a gross classification into only labial, coronal, and dorsal. The values of "manner" we use are stop, fricative, affricate, nasal, glide, and liquid. "Voicing" has two values. Smits et al. (2003) provide a table of the Dutch phoneme inventory, as well as explanations of choices about which sounds to include in the inventory, and the sounds' featural values.

[2] Although the original speech signal was cut off after the transition to the square wave, so that no further information about the speech was available after the cutoff point, it is possible that listeners could interpret the following square wave as masking a speech signal. However, the relative amplitudes of the square wave and the preceding speech were such that the square wave could not have effectively masked most speech sounds. Furthermore, in a previous experiment using gating to square waves (Warner, 1998), listeners did not judge the gated phoneme based on the possibility of additional acoustic cues occurring, masked, during the square wave, but rather responded based on only those cues that they actually heard.

Fig. 1 presents recognition rates as a function of gate for the features manner, place and voice. The data is presented as the percentage of information transmitted, rather than raw percent correct. Information can be viewed as a measure of the level of 'structure' in the occurrence of a number of items such as presented or perceived feature values. If a confusion matrix is maximally structured, i.e., all cells contain either the maximal possible count or zero, the response is fully predictable from a stimulus presentation, therefore all information has been transmitted from stimulus to response (TI is 100%). If, on the other hand, the matrix is minimally structured, i.e., to each of the stimuli each response is equally likely, no information has been transmitted and TI equals zero. Expressing recognition levels in terms of % TI instead of percent correct has the advantage that chance level performance leads to 0% TI, irrespective of stimulus biases or the number of possible responses (see also Smits, 2000; Smits et al., 2003). All our calculations started from pooled confusion matrices. So to calculate, for example, consonant manner, we first summarized the data in a six-by-six confusion matrix (six rows for stimulus manner, six columns for response manner) from which TI was then calculated using well-known equations (e.g., Miller and Nicely, 1955).

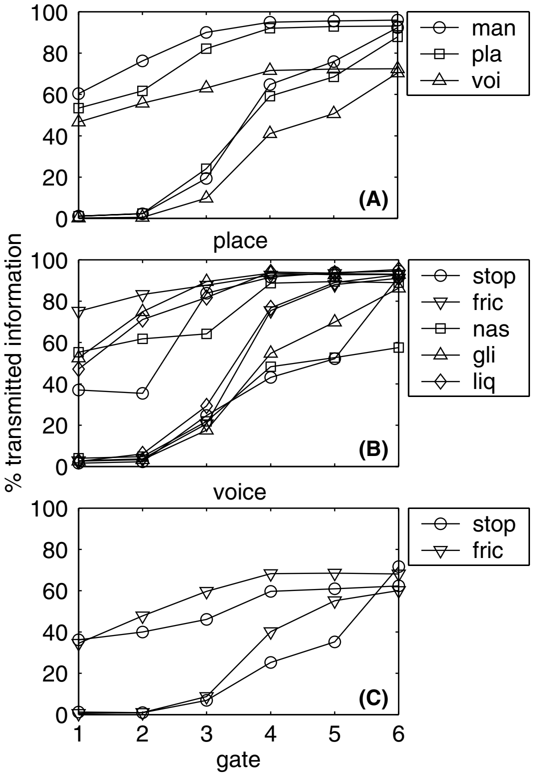

Fig. 1. Percent transmitted information for the consonantal features, by gate. In each panel, the higher set of curves represents phonemes in first position in the diphone, and the lower set of curves represents phonemes in second position in the diphone. (A) Percent transmitted information of the features manner ('man'), place ('pla'), and voice ('voi'), over all segments. (B) Percent transmitted information of the place feature plotted separately for stops ('stop'), fricatives ('fric'), nasals ('nas'), glides ('gli'), and liquids ('liq'). Dutch affricates occur in only one place and are therefore not represented. (C) Percent transmitted information of the voicing feature for stops and fricatives (the two manners that distinguish voice) separately.

TI was calculated for manner (over all manners), place (over all places), and voice (over the two voicing categories), as shown in Fig. 1A.[3] Calculations were made separately per listener and were then averaged for the purpose of the figures. The three lines grouped at the top of the graph are for perception of consonants which are the first segment of the diphone, and the three lines grouped at the bottom are for perception of consonants which are the second segment of the diphone. Fig. 1B shows the percent transmitted information for perception of place, calculated separately for consonants of each manner. Fig. 1C shows the percent transmitted information for perception of voicing for the stops and fricatives (the only manners with a voicing distinction) separately.

[3] There are, of course, many phonotactic constraints that affect the possible range of sounds that could occur in a particular diphone, including constraints involving syllable structure. But the listeners in this experiment did not know the syllable affiliation of most target sounds, since they heard only the preceding environment. It is therefore reasonable to assume that they considered a wide range of possible responses. In any case, since all and only the possible diphones of Dutch were used in the materials, the percent transmitted information measure allows for evaluation of how much information has been perceived, regardless of what the possible segments are.The potential effect of phonotactic probability is tested in Section 3.4.3 below.

Examination of Fig. 1A reveals that listeners have already perceived 40–60% of the information about each of manner, place, and voice by the time they have heard one-third of the sound. The greatest progress in perception of consonants in second position is made during the first third of the consonant (between gates three and four). For consonants in both first and second position, substantial additional information is conveyed during the final third of the consonant (between gates two and three for first position, and five and six for second position). This probably reflects the fact that this gate contains the bursts of stops and affricates, which in turn indicates that the burst provides considerable information. Some information is also conveyed during the middle third of the consonant.[4]

Fig. 1A also shows that manner information is generally transmitted slightly better than place information, with voicing information faring the worst. Each pairwise comparison of manner, place, and voice was then carried out separately for each gate for the first and second segment of the diphones. In these analyses, listener functioned as a random variable. These analyses show that manner, place, and voice each differ significantly from each other at nearly all gates. Only the comparisons of place and voice for the first segment at gate one, all three features for the second segment at gate one, and manner and place for the second segment at gate two fail to reach significance. All statistical tests are two-tailed pairwise *t*-tests. For all tests, the Bonferroni correction was applied to compensate for the large number of comparisons. For the evaluation of data in Fig. 1A, this means that each comparison was evaluated at a corrected α level of 0.00139 (0.05 divided by 36, the number of comparisons made).

Fig. 1B represents recognition of place of articulation for each manner class. The affricate class, however, is not included because the inventory we used contained only one affricate, making it impossible to calculate percent transmitted information among affricates. Examination of this figure reveals several patterns. First, recognition of place for the second phoneme is fairly similar for all manners at all gates within the first phoneme (gates 1–3). Thus, regardless of manner, approximately the same amount of information about place of the upcoming consonant is transmitted during the preceding sound. (Place of liquids is significantly, but only very slightly, better perceived than that of fricatives, glides, or nasals at gate 3. Some very small differences at gate 2 are also significant.)

As for specific manners, stop place for the first phoneme is not recognized well at the first two gates. It is significantly worse than all other manners. This is partly due to the voiced stops, for which the listeners usually heard only (part of) the voice bar at these gates.[5] However, the poor performance on stops in second position at gates four and five (significantly lower than fricatives and liquids at gate four and also significantly lower than glides at gate five) demonstrates that perception of stop place is difficult even with preceding context. Stop place is only recognized well when the burst is included: at gate three for consonants in first position and gate six for those in second position, stops no longer differ from fricatives, glides, and liquids.

For nasal place such a jump for the first phoneme occurs one gate later (i.e., gate four), when the transitions from the oral release become audible (this is presumably why there is no such jump within the diphone for the second phoneme). Percent transmitted information for place of nasals is significantly lower than for other manners in the following comparisons: nasals are lower than glides and fricatives for the first phoneme at gate two, lower than all manners for the first phoneme at gate three, lower than fricatives and liquids for

[4] Throughout the study, we present graphs showing gradual increases in %TI across gates. We interpret this as showing gradual increases over time in how much acoustic information is available. However, it is possible that individual diphones show only categorical changes in perception, jumping from 0%TI to 100%TI within the space of one gate, and that the gradual curves appear when results from various diphones are averaged. We do not examine results for individual diphones in this study, but there is ample evidence in the previous literature to show that many diphones have a gradual improvement in perception over time (cf. Warner (1998) for results for individual diphones using a methodology similar to the current one).

[5] Approximately one-third of CV diphones were recorded with a preceding vowel environment, but two-thirds were utterance initial.

the second phoneme at gate four, lower than all manners except stops for the second phoneme at gate five, and lower than all manners for the second phoneme at gate six. (Here, the required α level with the Bonferroni correction is 0.00042.)

Place recognition for fricatives, liquids and glides grows smoothly with increasing gates. Place of fricatives is recognized quite well for the first phoneme even at gates one and two (significantly greater than all other manners). For the second phoneme, place of fricatives and glides is recognized significantly better than all other manners at gates five and six. Thus, the first third of a fricative contains considerable information about place, even if the fricative usually lacks any preceding sound. The early part of glides, however, contains a similar amount of place information only when there is a preceding context. A few smaller differences in Fig. 1B also reach significance.

Fig. 1C gives percentages of transmitted information of the feature voice for the two manner classes that distinguish voicing. It is clear that transmission of information about the voicing distinction is rather weak, even by the end of the sound following the stop or fricative. Percent transmitted information remains near 60% even at gate six when the stop or fricative in question is the first phoneme. Although both voiced and voiceless stops and fricatives are perceived with reasonable accuracy by the last gate (see Fig. 2 of Smits et al., 2003), nearly all of the remaining errors are errors of voicing, leading to the low percent transmitted information for voicing. Within Fig. 1C, the only significant differences between stops and fricatives are for the second phoneme at gates four and five, indicating that voicing of stops is perceived even less well during the closure than voicing of fricatives is (required α of 0.00417). This effect is not apparent for consonants in first position, but the lack of it there may be an artifact of the environments in which diphones were recorded. For stops initial to the diphone, if they had no preceding recording environment (as two-thirds did not), and they lacked prevoicing, the first two gates could not be presented because they would have contained only silence. Therefore, the only stops in first position included here are those with prevoicing or the one-third with a preceding vowel environment, i.e., the ones which carried relatively reliable voicing information. The stops in second position all had preceding context, but it often consisted of a consonant, which might provide less information about voicing of the stop than a preceding vowel does. This may have lifted the curve for stops in first compared to second position, and may thus have diminished the difference between the stop and fricative curves. We will present the broad implications of these featural results for consonants in Section 4.

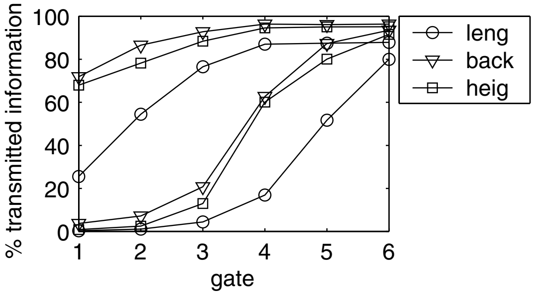

Fig. 2. Percent transmitted information for the vocalic features length ('leng'), backness ('back'), and height ('heig'). The upper set of curves are for phonemes in first position in the diphone, and the lower set for phonemes in second position in the diphone.

3.1.2. Vowels

Fig. 2 presents the percent transmitted information as a function of gate for the vowel features length, backness, and height. Length had three possible values: diphthong (/εi œy ɑu/), long (/a i u y e o œ/[6]), and short (/ɑ ε ɪ ɔ ʏ ə/). Place also had three possible values: front unrounded (/ε ɪ εi e i/), front rounded (/ʏ ə œ y œy/), and back (/ɑ ɑu ɔ a o u/). (/ə/ is not typically considered to be a front rounded vowel, but /ə/ and /ʏ/ were almost indistinguishable to listeners, so /ə/ is grouped with the front rounded vowels for further analysis.) Height also had three possible values: high (/i u y/), mid

[6] /i, u, y/ are not phonetically long, but they form a natural class with the other long vowels, and are traditionally grouped together with them (Booij, 1995; Gussenhoven, 1992). We have chosen to maintain this primarily phonological classification of vowels here, because there are many phonetic differences in duration among vowels. The phonological classification reflects partly duration, partly whether vowels can appear in open syllables, and partly degree of change in vowel quality.

(/e ɪ ɛ œ ʏ ə o ɔ/), and low (/ɑ a/). Because diphthongs change in height during their production, they were excluded from the height calculations.

The division into front unrounded, front rounded, and back is not entirely a matter of frontness/backness of vowels, as it is combined with rounding. However, previous research on Dutch vowels has indicated that the front rounded vowels are not very far forward in the vowel space, and are perhaps rather central instead of front (e.g. Pols, 1977, 1979; Warner, 2003). The classification of the Dutch vowels by height, backness, length, and rounding is somewhat complicated (Booij, 1995; Gussenhoven, 1992), and we have simplified the classification slightly here. For example, Booij (1995) uses four values of height, but we use three. Merging the backness and rounding distinctions is another such simplification. The classification presented here enables statistical analysis of featural differences in the perception results.[7]

Fig. 2 shows that the patterns of transmitted information for vowel backness and height as a function of gate are very similar. Perception of backness is slightly better than perception of height (significantly so at most gates), but this is a rather small effect. Perception of both these features improves quite quickly: 60–70% of the information about height and backness has already been transmitted by one-third of the way through the vowel. For vowels in second position, height and backness also show some improvement already at the third gate, which ends at the end of the preceding phoneme. This confirms that at least in some diphones, the preceding sound carries information about height and backness of an upcoming vowel.

Correct recognition of vowel length, on the other hand, is systematically worse than that of the other two features. For vowels as both first and second phoneme, percent transmitted information for length is significantly less than for each of the other two features at every gate (required α with Bonferroni correction is 0.00139). Length does reach high levels by the final gate, but remains worse than the other features, even for the vowel in first position at the sixth gate. That is, perception of vowel length is still worse than perception of other vocalic features by the end of the following phoneme. It is hardly surprising that perception of vowel length does not reach high levels until the full vowel becomes audible, but it is noteworthy that perception of length remains slightly worse than perception of other features long after the end of the vowel. Again, we will return to discuss the implications of these findings in Section 4.

3.2. Context effects

3.2.1. Consonants in context

Fig. 3 shows consonant recognition accuracy, conditional on whether the preceding or following context is a consonant or a vowel. The data shown in the figure are based on responses to stimulus pairs with the same target consonant. For example, Fig. 3A shows recognition rates for consonants followed by consonants (Cc—the capital indicates the target phoneme) vs. consonants followed by vowels (Cv).[8] In both types of stimuli, the vowel following the consonant or consonant cluster is stressed. The recording environment for these stressed Cc diphones was /ˈCCa/ (see Table III in Smits et al., 2003). Dutch phonology allows only the consonants /b d f k p s ʃ t v z/ as the first consonant of a CC onset. In order to keep the two curves comparable, the data for the Cv stressed diphones, obtained using /ˈCV-kə/ utterances, was based on responses to the same restricted set of consonants as for the Cc diphones.[9]

[7] If a four-way height distinction were used, there would be too few vowels for many combinations of features to allow for statistical analysis.

[8] For all analyses of context effects, figures show percent correct rather than percent transmitted information. Percent transmitted information is used in the preceding analyses to remove effects of bias toward specific responses. However, in the analyses of context, responses to the same set of segments in varying environments are being compared, so bias toward specific responses will not create spurious differences.

[9] CC diphones were recorded in the item-initial /ˈCCa/ environment whenever the CC cluster formed a possible onset cluster of Dutch. CC diphones which cannot be onset clusters were recorded in the environment /ˈaCCə/. Thus, the Cc stressed diphones in Fig. 3A and B are exactly those that can be onset clusters, and the cC unstressed diphones in Fig. 3C are those that cannot appear as onset clusters. This distribution based on phonotactics reflects the fact that stress was not separately manipulated for any CC diphones.

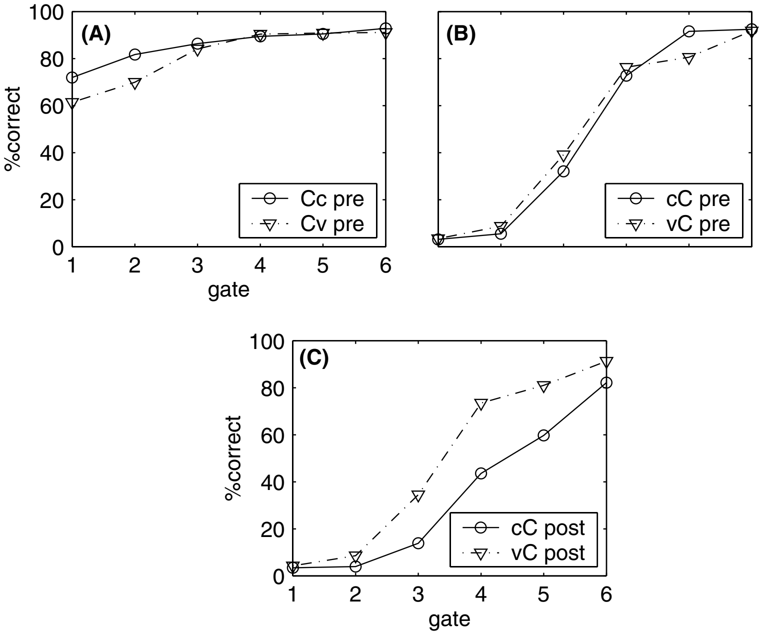

Fig. 3. Percent correct recognition for consonants (averaged over individual consonants) in various segmental environments. (A) Consonants as first phoneme of the diphone, followed by another consonant (Cc) or a vowel (Cv). The vowel in or following the diphone is stressed in both cases (pre-stress consonant). (B) Consonants as second phoneme of the diphone, preceded by a consonant (cC) or a vowel (vC). In both cases, the vowel following the target consonant is stressed (i.e. consonant is pre-stress). (C) Same as in B, but with the consonant preceded by a stressed vowel (i.e. consonant is post-stress).

Fig. 3A shows that recognition rates for initial consonants in stressed Cc and Cv diphones are remarkably similar at gates three to six. (Using the Bonferroni correction, only gates one and two meet the required α of 0.00833.) For the first two gates, initial consonants are significantly better recognized (by approximately 10%) in Cc diphones than in Cv diphones. This finding would seem surprising, since following environment should have more influence at later, rather than earlier, gates. Close inspection of the data reveals, however, that the difference for the initial two gates is almost entirely accounted for by the responses to stimuli beginning with /b/ and /d/. As mentioned above, the first two gates of initial voiced stops were only presented to listeners if a voice bar was actually present in the utterance; these gates were included whenever any voice bar was visible in the waveform, even if it was hardly audible. As discussed by van Alphen and Smits (2004), initial voiced stops in Dutch are more frequently realized with a voice bar in CV syllables than in CCV syllables. In the data of Fig. 3A, the proportion of diphones with initial /b/ or /d/ is thus much higher in Cv diphones than in Cc diphones; this depresses performance in the first two gates of the Cv stimuli. If responses to Cc and Cv stimuli beginning with /b/ or /d/ are removed, the two curves become very similar.

Fig. 3B shows recognition rates for pre-stress[10] consonants in second position, preceded by either a consonant or a vowel. The data for the cC diphones is based on responses to /'CCa/ utterances,

[10] The target consonant in these diphones is followed by a stressed vowel in its recording environment. Because of syllable structure restrictions and ambisyllabicity, it can be difficult to determine whether all the consonants in a particular type of diphone necessarily belong to the upcoming stressed syllable or not, particularly in VCV strings. We therefore refer to consonants only as "pre-stress" (preceding a stressed vowel) or "post-stress" (following a stressed vowel) where there is any ambiguity about syllable affiliation. Since we never had target consonants both preceded and followed by stressed (or unstressed) vowels, this coding system is consistent.

while that for the vC diphones derives from /V'Ce/ utterances (see Smits et al., 2003). Because Dutch phonology allows only the consonants /f j k l m n p r s t w x/ in second position in a consonant onset cluster, we use only responses to vC stimuli of the same restricted set. Fig. 3B shows that the effect of context on recognition rate is very small. The means are different with the Bonferroni correction (α level of 0.00833) only at gate five.

Finally, Fig. 3C shows recognition rates for post-stress consonants in second position preceded by either a consonant or a vowel. The data for the cC diphones is based on responses to /'aCCə/ utterances, while that for the vC diphones derives from /'VCə/ utterances.[11] Here we find a strong advantage of a preceding vowel compared to a preceding consonant, significantly so at gates two to six (α level of 0.00833). At gate four the advantage is almost 30%. The raw data shows that the lower performance for the cC diphones is distributed evenly across manner classes. The obvious explanation of this context effect would be that in vC diphones, the formant transitions at the end of the preceding vowel provide useful information for listeners on the upcoming consonant, whereas in cC sequences, such formant transitions are either absent, or (in liquids and glides), less informative than vowel formant transitions. It is noteworthy, however, that this pattern does not hold of pre-stress cC and vC diphones (Fig. 3B). It is likely that in the pre-stress vC diphones (Fig. 3B), the vowel of the diphone is reduced somewhat, and has shorter duration and less clear formant information than a stressed vowel. In the post-stress vC/cC comparison (Fig. 3C), it is likely that the stressed vowel of the vC carries substantial information about the consonant, while the first consonant of the cC cluster often does not. Thus the lack of a parallel effect in Fig. 3B is attributable to the short, unstressed vowel in the pre-stress vC diphones.

3.2.2. Vowels in context

We turn now to the influence of segmental context on vowel recognition. Fig. 4 presents the relevant comparisons. Fig. 4A gives recognition rates for stressed vowels in diphone-initial position conditional on whether they are followed by a consonant (Vc, stressed)[12] or an unstressed vowel (Vv, stressed). The Vc responses were obtained from / 'VCə/ utterances, while the Vv responses were obtained from /'bVVk/ utterances. Because Dutch phonology allows only long vowels and diphthongs in open syllables, like /'bV/ in /'bVVk/, we likewise only used responses to VC utterances with long vowels and diphthongs. Initially, the Vc diphones have an advantage over the Vv diphones (significant for the first three gates, at the required α of 0.00833). The most likely cause of this advantage is that, because the /'VCə/ utterances are pronounced with an initial glottal stop, the formants of the Vc diphones start at their target values, whereas those for the Vv diphones, originating from /'bVVk/ utterances, start with transitions from the /b/ closure. If this explanation holds, the difference in recognition rates at the initial gates is unrelated to the following context and thus irrelevant to the present discussion. At gates four to six, recognition rates are above 95% correct and the differences are very small, though significant at gate 5. The Vv diphones have a slight advantage over the Vc diphones (also significant at the α level of 0.00833). Both are, however, recognized at over 95% correct, so this difference is not worth further discussion.

Fig. 4B gives recognition accuracies for stressed vowels in second position preceded by either a consonant or a vowel. The cV and vV responses were obtained from /'CVkə/ and /'bV'Vkə/ utterances, respectively. During the first five gates, the vowels are recognized better in cV context than

[11] Context in this study is evaluated in terms of neighboring segment, but not in terms of syllabic affiliation. However, syllabic affiliation is controlled for here, since the post-stress cC diphones are exactly those where the cC is not possible as an onset cluster, so there must be a syllable boundary between the consonants (as explained in footnote 9). It is possible that syllable boundary information in running speech enables listeners to achieve even higher levels of accuracy than we have observed here. It would be ideal to compare all possible syllable configurations for all possible diphones, but the scope of the study precluded this manipulation.

[12] "Stressed" in the coding here refers to the target segment, that is, the initial vowel.

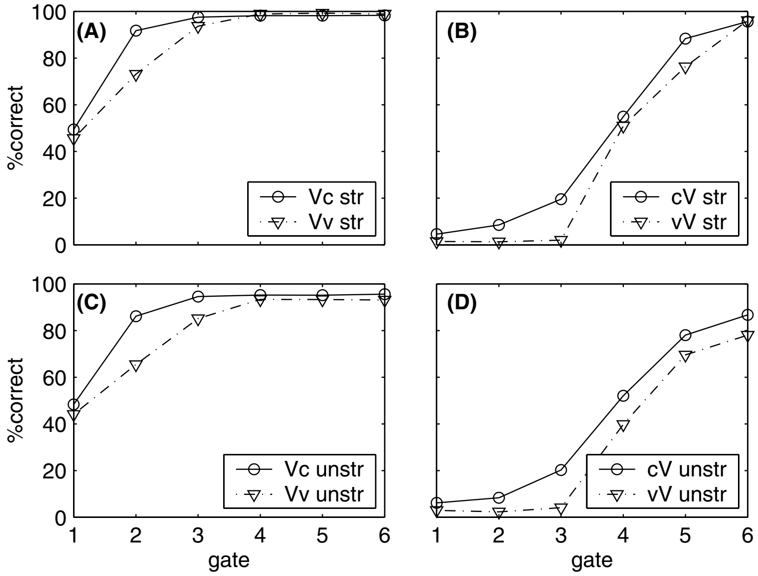

Fig. 4. Percent correct recognition for vowels (averaged over individual vowels) in various segmental environments. (A) Vowels as first phoneme of the diphone, followed by a consonant (Vc) or a vowel (Vv), in both cases with the target vowel stressed ('str'). (B) Vowels as second phoneme of the diphone preceded by a consonant (cV) or a vowel (vV), in both cases with the target vowel stressed. (C) Same as A, but with the target vowel unstressed ('unstr'). (D) Same as B, but with the target vowel unstressed.

in vV context (α level of 0.00833). The curves also follow qualitatively different patterns, however. While vowel recognition gradually improves for the cV stimuli during the first three gates, it stays close to chance level for the vV stimuli. This difference can be attributed to differences in coarticulation patterns in the CV and VV diphones in our experiment. The consonant and vowel in CV diphones are coarticulated, which allows the listener to make informed guesses about the upcoming vowel, for example from spectral transitions in fricative noise and formant transitions in liquids and glides. The two vowels in VV diphones, on the other hand, were consistently produced with a glottal stop or creaky voice between them. The glottal stop silence or creaky interval was located between the endpoints of gates three and four, i.e. the glottal stop or creaky voice was counted as part of the second vowel. The data shows that our speaker did not coarticulate across the gap between the two vowels provided by the glottal stop.

At gates four and five, the significant advantage of the consonantal context persists. This is somewhat surprising given our earlier point that vowels preceded by a glottal stop can be articulated in "target position" right from the onset of voicing, whereas, due to carryover coarticulation, vowels preceded by consonants start with a transitional phase. Since the interval from the third to the fourth gate in VV diphones consisted of the glottal stop silence or creaky voice, one might expect performance for vV diphones to remain low at the fourth gate, but this does not explain the continuing effect at gate five. The data suggests that the benefit from anticipatory coarticulation during the consonant outweighs any adverse effect of carryover coarticulation from the consonant into the vowel. This finding is in line with Bradlow's (2002) conclusion that consonantal coarticulation is an intentionally produced and useful feature of speech, rather than an unavoidable side-effect of moving articulators from one position to another.

Fig. 4C shows recognition rates of unstressed vowels in diphone-initial position when followed by a vowel vs. a consonant. The Vc and Vv responses derive from /V'Ce/ and /bV'Vk/ utterances, respectively. The results are very similar to those for stressed vowels in Fig. 4A, except that overall the levels are depressed (this effect of stress is discussed below). Again, the difference between Vc and Vv is significant for the first three gates, at the same α level (no significant difference at gate 5 this time). The same explanation for the difference that we offered for the first-position stressed vowels applies here, because the same types of utterances were used: the Vc diphones were pronounced with initial glottal stops, whereas the Vv diphones were preceded by /b/. So again we do not find any real effect of following context.

Fig. 4D shows recognition rates for unstressed vowels in second position in cV and vV diphones. The cV and vV data derives from /CV'ke/ and /'abVV'ke/ utterances, respectively. As we found for the comparison of Fig. 4A and B, the patterns shown in Fig. 4D are very similar to those found in Fig. 4C, except that overall the recognition levels are somewhat lower. At the first three gates, rates for cV diphones gradually rise, whereas those for vV diphones remain close to chance level. The difference in favor of cV diphones is significant for every gate here (at the same required α level).

3.3. Stress

3.3.1. Stress and consonant recognition

Fig. 5 presents recognition rates for consonants by stress of their adjacent vowel. Fig. 5A shows the data for consonants in initial position in Cv diphones, with the vowel of the diphone either stressed or unstressed. The data for the Cv stressed and Cv unstressed diphones derives from /'CVkə/ and /CV'ke/ utterances, respectively, with /ə/ excluded from both diphone types because it cannot be stressed. The similarity of the two curves is striking. Consonants before unstressed vowels are recognized nearly as well as those before stressed ones, with differences in recognition rates never exceeding 5% throughout. Nevertheless, the mean recognition rates are significantly different (using the Bonferroni correction and a required α level of 0.00833) at each of the first four gates. Thus, stress has very little effect on the perceptibility of initial consonants, but the small effect that is present is in the expected direction, and disappears by a point early in the following vowel.

Fig. 5B shows recognition rates for consonants in second position preceded by stressed vs. unstressed vowels. The vC post-stress diphones were obtained from /'VCə/ or /'bVCə/ utterances, while the vC pre-stress diphones were from /V'Ce/ or /bV'Ce/ utterances. Nevertheless, the two curves are again extremely similar, even more so than in Fig. 5A. The only significant advantages are for consonants followed by stressed vowels at gates three and four (same α level).

3.3.2. Stress and vowel recognition

Fig. 6 shows recognition rates for vowels differing in stress. Fig. 6A compares diphone-initial vowels in Vc context, with the vowel stressed vs. unstressed (from /'VCə/ and /V'Ce/ utterances, respectively). Although recognition is generally superior for the stressed vowels (significantly so at the required α level of 0.00833 for all gates ex-

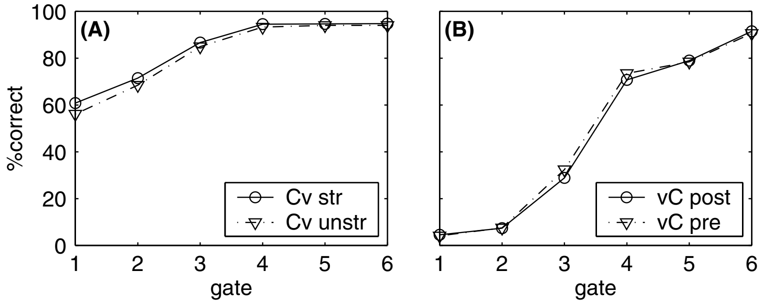

Fig. 5. Percent correct recognition for consonants (averaged over individual consonants) by stress. (A) Consonants in first position in the diphone, followed by vowels, where the vowel of the diphone is either stressed ('Cv str') or unstressed ('Cv unstr'). (B) Consonants in second position in the diphone, preceded by vowels, where the vowel of the diphone is stressed and the vowel following the diphone unstressed ('vC post'), or the vowel of the diphone is unstressed and the vowel following it is stressed ('vC pre').

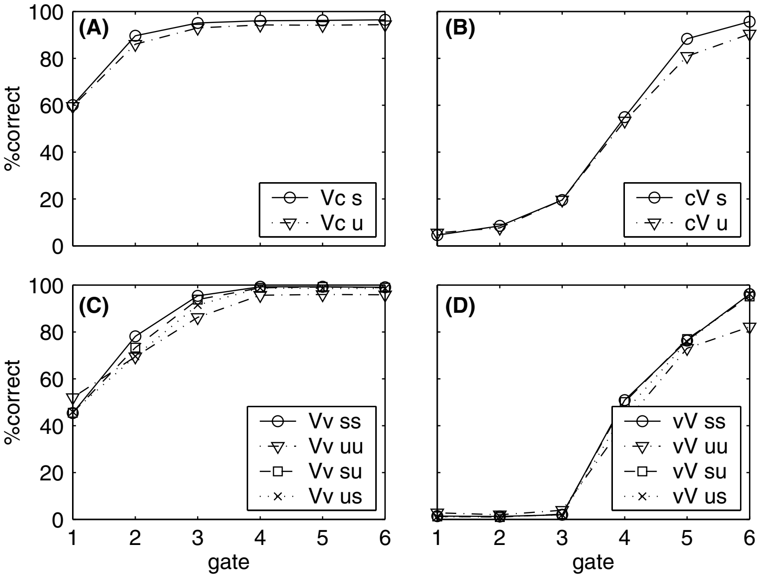

Fig. 6. Percent correct recognition for vowels (averaged over individual vowels) by stress. (A) Vowels in first position in the diphone, followed by consonants, where the vowel is either stressed (‘Vc s’) or unstressed (‘Vc u’). (B) Vowels in second position in the diphone, preceded by consonants, where the vowel is either stressed (‘cV s’) or unstressed (‘cV u’). (C) Vowels in initial position of vowel–vowel diphones, where the two vowels are both stressed (‘Vv ss’), both unstressed (‘Vv uu’), stressed–unstressed (‘Vv su’), or unstressed–stressed (‘Vv us’). D: Same as C for vowels in second position of vowel–vowel diphones (vV).

cept the first), the actual differences in recognition rates are extremely small. The existing small difference is mainly due to the vowel /a/, which is affected by stress somewhat more strongly than other vowels (see Smits et al., 2003, for discussion).

Fig. 6B presents recognition rates for stressed and unstressed vowels in second position preceded by consonants. The data for these stress conditions derives from /'CVkə/ and /CV'ke/ utterances, respectively. Vowel recognition is better in the stressed condition for gates five and six (α level of 0.0083), while the reverse holds for gate one (at the same required significance level). Only the effect in favor of stress at gates five and six is of any appreciable size, however. The raw data shows that the stress effect at these gates holds among most of the vowels.

Fig. 6C shows recognition rates for vowels in Vv context. Four stress conditions are contrasted: stressed–stressed, unstressed–unstressed, stressed–unstressed and unstressed–stressed, which were taken from /'bV'Vkə/, /'abVV'ke/, /'bVVk/, and /bV'Vk/ utterances, respectively. Although, as expected, the stressed–stressed diphones are generally recognized best, and the unstressed–unstressed ones worst, the differences are never very large (maximum of 10% at gates two and three). The unstressed–unstressed condition is significantly lower than each other condition at gates three, four, five, and six. Furthermore, the stressed–stressed condition is significantly greater than the unstressed–stressed condition at gates two and three only. These are the only significant effects in this comparison. (All pairs were evaluated using the Bonferroni test at a required α of 0.00139.) That is, perception of the first vowel in a vowel-vowel diphone is better if that vowel is stressed than if it is unstressed for gates ending during the first vowel. Once information about the second vowel is available, only sequences of two unstressed vowels show a deficit in identification of the first vowel. The raw data shows that the effect of stress is carried by the vowels /a e o œ εi œy ɑu/, which include all the diphthongs and relatively diphthongal vowels.

Finally, Fig. 6D shows recognition rates for vowels in vV context (second position) in the four possible stress conditions. Surprisingly, vowel recognition is better in the unstressed–unstressed condition than in the other three conditions at gates one and three (unstressed–unstressed significantly greater than unstressed–stressed and stressed–unstressed at gate one, and significantly greater than all other conditions at gate three). This effect is, however, very small and does not deserve much consideration because recognition rates are very close to zero. At gates four to six, on the other hand, recognition of the second position vowel is consistently worse in unstressed–unstressed diphones (unstressed–unstressed significantly different from all other conditions at gates four and six, and from stressed–unstressed only at gate five, required α of 0.00139).

This effect is larger, with a maximum difference of about 15% at the last gate. All vowels contribute to this difference. Thus, unstressed vowels after another unstressed vowel are more difficult to identify than either stressed or unstressed vowels following a stressed vowel. This influence of stress of the first vowel on perception of the second at late gates is interesting, because the very low recognition rates for all vV stress conditions during gates one to three show that little or no information about the second vowel is available during the first vowel. Thus, the effect of stress of the preceding vowel cannot reflect clarity of spectral information during the first vowel. This may be similar to the effect of stress on perception of length of /a/ discussed by Smits et al. (2003): either when a vowel is stressed, or when one can hear clearly that it is unstressed because the preceding vowel is stressed, perception of that vowel is better than when it follows an unstressed vowel.

3.4. Higher-level factors

3.4.1. Response strategies

At early gates, listeners received very little information about the second phoneme. Inspection of the raw data suggests that some of the listeners developed a strategy of choosing a fixed label for the second phoneme when very little acoustic information about it was available. Furthermore, it is generally accepted that gating introduces acoustic cues that are not present in the original signal and thus biases responses; specifically, a sudden offset of acoustic energy can induce listeners to hear plosive manner and/or labial place (Ohala and Ohala, 1995; Pols and Schouten, 1978; Smits, 2000; Warner, 1998). Although we strove to minimize such biases by gating to a square wave and ramping the signal down over a window (Smits et al., 2003), we cannot exclude the possibility that some biases remained.

Fig. 7 displays the percentages of responses for each of the response categories /ə h m n p/ for the second phoneme, as a function of gate. At gate one, these five response categories were the most frequently used. (The percentages for the remaining 33 phonemes are omitted). The figure shows that in the absence of acoustic information, i.e. at the early gates, subjects responded far from randomly. In fact, at the first gate, 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. Does this mean that in these cases subjects heard an upcoming /h/, or /ə/? We do not think so. The data suggests that when subjects really did not know how to respond, most of them selected a default label. Some subjects used /h/ as default label, others seemed to use two default labels, often including /h/ or /ə/ or both. Still others genuinely seemed to guess, and their confusion matrices for the early gates of the second phoneme were filled in a relatively homogenous fashion. However, the fact that neither plosive nor labial responses prove highly elevated for early gates shows that we succeeded in keeping artifact-induced biases within limits.

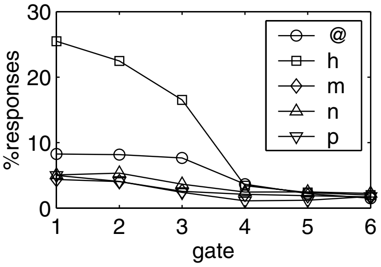

Fig. 7. Percent of all responses for the five overall most popular responses for the second phoneme as a function of gate ('@' indicates /ə/).

3.4.2. Phoneme frequency

Fig. 8 shows observed response probabilities of all second position phonemes at gate one (i.e., when little or no acoustic information about the phoneme is available) plotted against phoneme (token) probabilities derived from the CELEX corpus (Baayen et al., 1993). If subjects perfectly biased their guesses according to the frequencies of occurrence of the phonemes in the language, the points would be on the main diagonal. The actual correlation between the logarithms of the response frequency and the lexical frequency is .66 ($p < .0005$). This suggests that listeners did use phoneme frequencies in their second-phoneme responses at the first gate. The correlations for subsequent gates gradually decrease, as do their significance levels, until, at gate six, 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.

If we divide phonemes into vowels and consonants, it becomes clear that the consonants have a heavier component in the correlation than the vowels ($r = .71$, $p < .0005$ for consonants vs. $r = .55$, $p < .05$ for vowels). Examination of Fig. 8 suggests, furthermore, that it is in fact the consonants /dʒ ʒ ʃ/ which carry most of the correlation. These three consonants are originally foreign to the Dutch language, although /ʃ/ does play an important role in diminutives. If /dʒ ʒ ʃ/ are removed, the correlation is no longer significant even at the first gate.

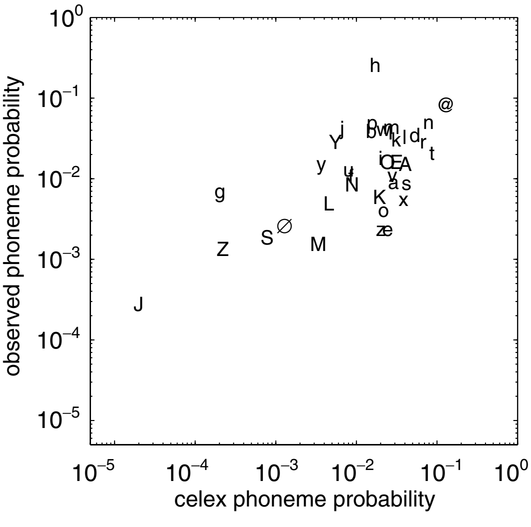

Fig. 8. Overall observed phoneme probabilities for gate 1 of the second phoneme plotted against phoneme probabilities estimated from the CELEX database. Phoneme symbols are in correspondence with IPA, except for A (indicating /ɑ/), E (/ɛ/), I (/ɪ/), K (/ɛi/), L (/œy/), M (/ɑu/), O (/ɔ/), @ (/ə/), ø(/œ/), S (/ʃ/), Z (/ʒ/) and J (/dʒ/).

It should be added that it is likely that the low response frequencies for /dʒ/ are, at least in part, caused by the fact that /dʒ/ is acoustically a complex consonant, in the sense that it goes through several distinct acoustic phases (closure, release and frication). Smits et al. (2003) found that listeners will 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 respond with phonemes that the stimulus is most similar to based on information available up to that point, in this case /j/ or /d/ rather than /dʒ/ (see Ohala and Ohala, 1995; for a similar argument). According to this reasoning, a low score would also be found 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 foreign consonants /dʒ ʒ ʃ/.

3.4.3. Transitional probabilities

Finally, we investigated whether listeners made use of transitional probabilities in their responses to the second phoneme. A transitional probability $p(\varphi_2|\varphi_1)$ is defined as the conditional probability of observing (or responding) phoneme φ_2 given that the preceding phoneme was φ_1 (e.g., Pitt and McQueen, 1998). In our analysis we concentrated on gate two, where listeners were generally able to make a reasonable guess at the first phoneme of the diphone, but there was still not much acoustic information for the second phoneme. For each phoneme in initial position, we calculated the correlation between the logarithm of the probabilities of all possible subsequent phonemes as predicted by the CELEX database, and the logarithm of the corresponding observed (response) probabilities. Out of 38 correlation coefficients, only one proved significant at the $p < .05$

level, namely the one for /ɑu/ as initial phoneme ($r = -.68$, $p < .01$), which is negative.[13] Inspection of the data revealed that when the first phoneme was recognized as /ɑu/, listeners were likely to guess /w/ for the second phoneme. In the CELEX database, however, it is rare for /w/ to follow the diphthong /ɑu/. It is possible that in this particular case, the judgments by the listeners were contaminated by orthographic knowledge. In Dutch, syllable-final /ɑu/ or /ɑuw/ is usually written as either "auw" or "ouw". However, the diphthong /ɑu/ can be followed by coda consonants (as in the frequent words "fout", *mistake*, or "kous", *stocking*), in which case the /w/ is lacking. Nevertheless, on hearing a stimulus ending in /ɑu/, the orthographic pattern may have led listeners to select /w/ as the second phoneme. In any case, the data clearly shows that transitional probabilities did not play a significant role in listeners' response behavior.

4. Discussion

This extensive database has enabled us to look in great detail at the effects of multiple factors on the identification of phonemes. The most notable results are perhaps the surprisingly limited extent of the effects of both context and stress. Statistical factors such as frequency and transitional probability also exercised only very limited influence on the responses.

4.1. Overview of timing of perception

The results show several overall patterns in the timing of perception of speech segments. First, most features of most speech sounds are already perceived fairly well by one-third through the segment, regardless of whether the segment is in first or second position in the diphone. Segments in second position (for which perception during the transition into the sound can be evaluated) show the most progress in perception between the third and fourth gates, when the first third of the segment itself becomes audible. Exceptions to this pattern are vowel length and stop place and voicing, which are perceived later, with the stops showing a large improvement in perception at the burst.

Examination of perception at the second and third gates for segments in second position (the gates surrounding the transition into the segment) show that more information spreads leftward from the second sound to the first in a CV sequence than in a VV sequence. This probably reflects the tendency to produce a glottal stop between two vowels in Dutch. The results also show that in VC and CC sequences, manner and place information spread into a preceding sound more than voicing information does. Furthermore, more information spreads into the preceding sound in VC sequences than CC sequences, but only if the stress precedes the second segment of the diphone (so that it is located on the vowel of the VC), rather than following it. This indicates that unstressed, reduced vowels cannot carry as much information about upcoming consonants as stressed ones can. Since reduction can be viewed as increased coarticulation with neighboring segments, one might expect reduced segments to carry more information about neighboring segments, while simultaneously carrying less information about the reduced segments themselves. This is not what our results indicate. Instead, they indicate that reduced segments also carry less information about neighboring segments, perhaps because the increased coarticulation is offset by lesser duration, amplitude, and acoustic clarity.

4.2. Perceptibility of phonological features

Among consonantal features, manner is perceived better than place, which is perceived better than voice. This result is particularly interesting in comparison with results from studies of perception of English CVC nonsense syllables in noise (e.g. Miller and Nicely, 1955; Benkí, 2003); under noise, perception of English consonant place of articulation is noticeably weaker than perception of voicing or manner, which receive similar identification scores. In our data, voicing is consistently perceived less well than either manner or place (except at gates where so little consonantal infor-

[13] If a correction for the large number of correlations were applied, even this correlation would likely not be significant.

mation is available that there is a floor effect). Furthermore, perception of manner is, in our data, only slightly better than perception of place. Two factors may underlie this difference between result patterns. The first and largest difference is that between perception of intact syllables in noise vs. perception of gated speech in silence. It is likely that noise has a particularly negative effect on the perceptual cues for place, particularly those cues located in the burst noise of stops or noise of fricatives, since added noise would obscure such relatively soft speech noise more than it would obscure the formants of a relatively loud periodic sound. Second, of course, these preceding studies were of English while our study is of Dutch, which could be of relevance given that some consonantal voicing distinctions are currently being lost in Dutch, particularly in the fricatives. Dutch also has final devoicing. We used only voicing distinctions which were maintained for dialects spoken in the area where the experiment was conducted, but it is possible that Dutch listeners pay little attention to consonantal voicing, because it is often neutralized in the speech they hear (in coda position for all obstruents, and in onset position for most fricatives in some dialects).

Our results also show some more detailed results regarding consonantal features: in order to perceive stop place well the burst is necessary, and in order to perceive nasal place well, the transition to the following sound is necessary. Particularly for stops, this finding may seem counterintuitive: it is well known that much information about stop place is conveyed by the formant transitions into the stop. As Fig. 1B shows, our results do show that information about stop place is transmitted during the preceding sound, particularly during the final third of it (between gates 2 and 3), even though not all second position stops have vocalic transitions into them since a large proportion of the stimuli are CC sequences. The degree to which place of stop is perceptible based on transition into the stop is similar to all other manners. What differs across manners is when further information is transmitted during the consonant itself. It is not surprising that not as much additional place information is transmitted during the stop closure as during other consonants.

Place of nasals is perceived rather badly during the nasal itself. This is expected, since the only articulatory difference between nasals is the length of the closed oral cavity, so the primary acoustic difference is frequency of the antiresonance (Stevens, 1998). This result is consistent with past findings about confusability of nasals (e.g. Ohala, 1975; Recasens, 1983; Repp, 1986). Voicing (of stops and fricatives, the only categories with a distinction) is also perceived rather poorly. For stops, this is consistent with the results of Smits (2000).

With respect to vocalic features, our principal result is that length is perceived less accurately than the other features (backness and height), even long after the vowel is finished. The dominance of length confusions was also reflected by the results for individual vowels reported by Smits et al. (2003). Not surprisingly, many long vowels tend to be misperceived as short vowels, but not vice versa. Benkí (2003) does not analyze vowel length (or the tense/lax distinction), but finds the opposite pattern from ours for height and backness. His results for perception in noise show substantially greater perceptual robustness of vowel height than backness. We find a slight advantage for backness over height. As with the consonantal features, this discrepancy too is likely to be a result of the difference between perception in noise vs. perception based on a gated signal in silence. Benkí (2003) in fact discusses evidence that height is better perceived than backness in noise, but backness is better perceived than height in silence. He attributes this to the effect of noise on F1 vs. other cues.

4.3. Segmental context and stress

Although the dataset in principle allows for fine-grained investigations of phonological context—indeed, part of the purpose of including every possible diphone of the language was to facilitate comparison across any desired phonological contexts—we here investigated the effects of phonological context in a broad sense, by comparing each type of segment with a preceding or following consonant vs. vowel.

The results of this analysis show two noteworthy patterns. First, consonants in the second position of the diphone are perceived more accurately if preceded by a vowel than by a consonant, as long as the vowel of the vC diphone is stressed. If the pre-consonantal vowel is stressed, and hence not reduced, it is able to carry more useful information about the upcoming consonant than a preceding consonant in a cC diphone can. Second, vowels in second position are perceived more accurately if preceded by a consonant than a vowel, regardless of stress. At early gates, this indicates that information about an upcoming vowel is present in many consonants, but not in vowels in VV sequences. At later gates, this pattern indicates that consonant-vowel coarticulation provides a long-lasting perceptual advantage, since vowels after consonants continue to be perceived more accurately than those after vowels even by the end of the (second position) vowel. Beyond these two results, there are few notable effects of phonological context in the sense of having a neighboring vowel vs. consonant.

The effect of stress is, in most cases, in the predicted direction but surprisingly small. For consonants, stress of the surrounding vowels has only a very small effect. There are somewhat larger effects of stress on perception of vowels: vowels in vowel-vowel diphones are less accurately perceived if both are unstressed than if one or both of the vowels are stressed, and vowels in CV diphones are more accurately perceived if they are stressed, but only when two-thirds or more of the vowel is heard.

4.4. Frequencies and response biases

Turning to influences on responses other than availability of acoustic information, we found that many subjects adopt a default response (often /h/ or /ə/) when little or no acoustic information is available about a segment. Perhaps the choice of these particular two sounds as default responses indicates that Dutch speakers consider them to be relatively neutral sounds. At early gates, listeners may only be able to tell that the upcoming segment might be a vowel, or that it might be a consonant. They may choose /h/ as the most general consonant, and /ə/ as the most general vowel.

Subjects are not greatly influenced by phoneme frequency: they may choose a default response when they truly do not know what a sound was, but they do not generally weight their responses to favor the more common phonemes of the language. Subjects do disfavor certain consonants that are not part of the native Dutch phoneme inventory, and these also happen to be relatively uncommon consonants, so there is some appearance of a phoneme frequency effect, but it does not appear that subjects generally weight their guesses based on frequency. We also found effectively no use of transitional probabilities. That is, once subjects have recognized the first phoneme of the diphone relatively well, they do not use this information to help them predict an upcoming sound they cannot hear yet.

The lack of an effect of overall phoneme frequency or transitional probabilities has interesting implications for speech perception models and spoken word recognition models more generally. These results suggest that listeners can do quite well at speech perception, and at recognizing individual sounds, from bottom-up information alone. Listeners certainly do not have to rely on higher-level information such as overall frequency or transitional probabilities in order to decide what sounds they are hearing. It may be that our experimental task discouraged the use of such information because the experiment was so long and repetitive and because it clearly did not involve recognizing real words. It could also be that the careful speech used for the stimuli contained clearer acoustic cues than connected speech does, making use of higher-level information less necessary. That is, listeners may make more use of phoneme frequency and transitional probabilities in perceiving normal connected speech than subjects did in our experiment.[14] Furthermore, transitional probabilities differ depending on whether the diphone spans a syllable boundary or not, and we

[14] van Son and Pols (1999) studied perception of segments in connected speech. Comparison of that work with ours suggests that if listeners use higher-level information more in listening to connected speech, it is not likely to be because of the availability of clearer acoustic cues in the careful speech of our stimuli. The most comparable data is for perception of vowels after a consonant, for which our error rate (Fig. 4B and D, cV stressed and unstressed, gate 5) is very similar to that found by van Son and Pols (their Fig. 2, VC condition, 1999, p. 8): both are approximately 15%.

have not examined use of transitional probabilities separately for stimuli within and across syllables.

However, the result does show that use of frequency information is neither necessary for accurate speech perception, nor an ineluctable component of the perceptual process. It is possible to perceive all combinations of sounds in a language at least reasonably well through bottom-up information alone. This finding is compatible with the arguments presented by Norris et al. (2000) that spoken-word recognition is a feedforward process. The lack of frequency effects in our current data suggests that the data is a relatively pure reflection of perception from acoustic information, supplemented with a default strategy or random guessing when no acoustic information is available. It is our hope that the present publicly available dataset will be useful for the analysis of many other questions about speech perception in the future.

Acknowledgement

We are grateful to Dennis Norris for discussion of this material, as well as to Rob van Son and three anonymous reviewers for their helpful comments. We further thank Mattijn Morren, Keren Shatzman, Petra van Alphen, Niels Janssen, Tau van Dijck, Anne Pier Salverda, and Aoju Chen for their great efforts in preparing and running the experiment, and Saskia Bayerl for assistance with preparing the database for the Internet.

References

Baayen, H., Piepenbrock, R., van Rijn, H., 1993. The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Benkí, J., 2003. Analysis of English nonsense syllable recognition in noise. Phonetica 60, 129–157.

Booij, G., 1995. The Phonology of Dutch. Clarendon, Oxford.

Bradlow, A.R., 2002. Confluent talker- and listener-oriented forces in clear speech production. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology, Vol. 7. Mouton de Gruyter, Berlin, pp. 241–273.

Crystal, T.H., House, A.S., 1982. Segmental durations in connected speech signals: Preliminary results. J. Acoust. Soc. Amer. 72, 705–716.

Crystal, T.H., House, A.S., 1988a. Segmental durations in connected-speech signals: Current results. J. Acoust. Soc. Amer. 83, 1553–1573.

Crystal, T.H., House, A.S., 1988b. Segmental durations in connected-speech signals: Syllabic stress. J. Acoust. Soc. Amer. 83, 1574–1585.

Gussenhoven, C., 1992. Illustrations of the IPA: Dutch. J. Internat. Phonet. Assoc. 22, 45–47.

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Amer. 97, 3099–3111.

Kenstowicz, M., 1994. Phonology in Generative Grammar. Blackwell, Cambridge, MA.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27, 338–352.

Norris, D., McQueen, J.M., Cutler, A., 2000. Merging information in speech recognition: Feedback is never necessary. Behav. Brain Sci. 23, 299–325.

Nygaard, L.C., Pisoni, D.B., 1995. Speech perception: New directions in research and theory. In: Miller, J.L., Eimas, P.D. (Eds.), Speech, Language and Communication. Academic Press, New York, pp. 63–96.

Ohala, J.J., 1975. Phonetic explanations for nasal sound patterns. In: Ferguson, C.A., Hyman, L.M., Ohala, J.J. (Eds.), Papers from a Symposium on Nasals and Nasalization. Language Universals Project, Stanford, pp. 289–316.

Ohala, J.J., Ohala, M., 1995. Speech perception and lexical representation: The role of vowel nasalization in Hindi and English. In: Connell, B., Arvaniti, A. (Eds.), Phonology and Phonetic Evidence: Papers in Laboratory Phonology, Vol. IV. Cambridge University Press, Cambridge, pp. 41–60.

Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. J. Acoust. Soc. Amer. 24, 175–184.

Pitt, M.A., McQueen, J.M., 1998. Is compensation for coarticulation mediated by the lexicon? J. Memory Lang. 39, 347–370.

Pols, L.C.W., 1977. Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words. Institute for Perception TNO, Soesterberg.

Pols, L.C.W., 1979. Some spectral and perceptual measurements on Dutch diphthongs. In: Anniversaries in Phonetics: Studia Gratulatoria Dedicated to Hendrik Mol. Institute of Phonetic Sciences, Amsterdam, pp. 270–285.

Pols, L.C.W., Schouten, M.E.H., 1978. Identification of deleted consonants. J. Acoust. Soc. Amer. 64, 1333–1337.

Recasens, D., 1983. Place cues for nasal consonants with special reference to Catalan. J. Acoust. Soc. Amer. 73, 1346–1353.

Repp, B., 1986. Perception of the [m]–[n] distinction in CV syllables. J. Acoust. Soc. Amer. 79, 1987–1999.

Smits, R., 2000. Temporal distribution of information for human consonant recognition in VCV utterances. J. Phonet. 27, 111–135.

Smits, R., Warner, N., McQueen, J.M., Cutler, A., 2003. Unfolding of phonetic information over time: A database of Dutch diphone perception. J. Acoust. Soc. Amer. 113, 563–574.

Stevens, K.N., 1998. Acoustic Phonetics. MIT Press, Cambridge.

van Alphen, P.M., Smits, R., 2004. Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing. J. Phonet. 32, 455–491.

van Son, R.J.J.H., Pols, L.C.W., 1999. Perisegmental speech improves consonant and vowel identification. Speech Comm. 29, 1–22.

Warner, N., 1998. The role of dynamic cues in speech perception, spoken word recognition, and phonological universals. Ph.D. thesis, University of California, Berkeley.

Warner, N., 2003. Rapid perceptibility as a factor underlying universals of vowel inventories. In: Carnie, A., Harley, H., Willie, M. (Eds.), Formal Approaches to Function in Grammar. John Benjamins, Amsterdam, pp. 245–261.