Journal of Phonetics (2000) **28,** 111–135
doi:10.006/jpho.2000.0107
Available online at http://www.idealibrary.com on IDEAL®

Temporal distribution of information for human consonant recognition in VCV utterances

Roel Smits

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Received 4th February 1999, and in revised form 17th April 2000

The temporal distribution of perceptually relevant information for consonant recognition in British English VCVs is investigated. The information distribution in the vicinity of consonantal closure and release was measured by presenting initial and final portions, respectively, of naturally produced VCV utterances to listeners for categorization. A multidimensional scaling analysis of the results provided highly interpretable, four-dimensional geometrical representations of the confusion patterns in the categorization data. In addition, transmitted information as a function of truncation point was calculated for the features manner place and voicing. The effects of speaker, vowel context, stress, and distinctive feature on the resulting information distributions were tested statistically. It was found that, although all factors are significant, the location and spread of the distributions depends principally on the distinctive feature, i.e., the temporal distribution of perceptually relevant information is very different for the features manner, place, and voicing.

1. Introduction

During the past 50-odd years, a large number of studies have examined the relevance of various portions of the speech signal to the perception of consonants. Many of these studies investigated the effect of specific acoustic cues for consonant recognition by measuring listeners' responses to synthetic speech stimuli in which assorted acoustic parameters, such as formant frequencies or voice onset times (VOTs), were systematically varied (e.g., Cooper, Delattre, Liberman, Borst & Gerstman, 1952; Lisker, 1978). Other studies employed natural utterances that were systematically degraded in some fashion. Filtering was applied to examine the relevance of the information in various frequency bands (e.g., Fletcher, 1953; Miller & Nicely, 1955). The distribution of perceptually relevant information along the temporal dimension has been studied using truncated, or "gated" versions of natural utterances (e.g., 't Hart & Cohen, 1964; Öhman, 1966). In the gating technique only portions of the complete utterance are presented to listeners. In *forward gating* the initial part, in *backward gating* the final part of an utterance is

Address correspondence to Roel Smits. E-mail: roel.smits@mpi.nl.

0095–4470/00/020111 + 25 $35.00

presented. By varying the cutoff point before or after which the utterance is deleted, one can measure the perceptual relevance of various parts of the signal.

Many studies employing the gating technique have focused on a particular manner class, such as stops, nasals, or fricatives. By systematically presenting (or withholding) various temporally distinct acoustic structures, such as silence, release burst, and nasal murmur, the relevance of these structures for the perception of voicing or place within the respective manner class was measured (e.g., Schouten & Pols, 1983; Kurowski & Blumstein, 1987). Relevant results of these studies are discussed in the final section of this paper. Although these studies have provided useful knowledge on the temporal distribution of information distinguishing consonants within a manner class, they tell us little about how stops, nasals and fricatives are distinguished from each other. Besides, it is not clear to what extent listeners' use of acoustic information depends on whether or not the manner class is known beforehand.

A small number of investigations, namely Grimm (1966), Öhman (1966), and Furui (1986), have employed the gating technique for a large set of consonants, including several manner classes. Grimm (1966) presented final (backward-gated) portions of CV syllables spoken by American English speakers to listeners for identification. The consonants included stops, fricatives, and glides, and three vowel contexts were used. A peculiarity of Grimm's analysis was that he defined gate position relative to the instant of peak intensity of the vowel. The most important conclusions of Grimm's study were: (1) all perceptually relevant information for manner, place and voicing of the consonant resides in the signal portion up to 50 ms before the instant of peak intensity of the vowel; (2) place of articulation was most resistant to the removal of initial portions of the syllable, manner of articulation was least resistant, and voicing was intermediate; and (3) the identification curves for manner of articulation were shallower than those for place and voicing.

Öhman (1966) recorded Swedish /aCa/ utterances, with stress on the first syllable, where C could be one of 21 consonants, including stops, fricatives, nasals, liquids, and glides. Each disyllable was cut in two at various temporal locations and the resulting initial and final portions were presented to listeners for identification. The temporal reference points used to compare the responses for different utterances were the instants of closure and release. Öhman presented his results as percentages correctly perceived manner, place, and voice for individual phonemes, which makes it somewhat difficult to draw summary conclusions. Still, the following patterns could be discerned. For stops and nasals, the point of most rapid change in percent correct for all three distinctive features was located close to the instants of closure and release, respectively. For fricatives, this point was located some 20 ms after closure in VC and 20 ms before closure for CV syllables, indicating that some frication is required for correct recognition of fricative manner, place, and voicing. The point of most rapid change for liquids and glides was close to release for CV syllables, but 20 ms after closure for VC syllables. Overall, the curves for correct place perception were relatively shallow for the voiceless stops and the fricatives. Finally, the curves for correct place perception calculated across manner classes showed a clustering according to five places to articulation: bilabial, labiodental, dental, palatal, and velar.

Furui recorded all 100 phonotactically possible Japanese short syllables, including CVs CjVs and isolated vowels. Initially and finally truncated versions of these utterances, as well as both initially and finally truncated versions, were presented to listeners for identification of both consonant and vowel. For the correct identification curve of each

syllable, the "critical point", defined as the point where 80% correct was exceeded for the first time, was determined, and all curves were aligned at this point to calculate further summary functions. The results showed that information in very short portions of the speech signal can have a strong impact on correct consonant recognition. Performance on consonant recognition changed from 90 to 50% by shifting the initial truncation point from the critical point to 10 ms after the critical point. Shifting the final truncation point from 10 ms before the critical point to the actual critical point changed performance from 60 to 85%. Furthermore, Furui showed that the signal portions that were perceptually most important more or less coincided with regions of maximal spectral change. Thus, the study by Furui particularly emphasized the perceptual importance of acoustically dynamic regions.

Although, as the above discussion shows, these three studies have provided useful insights, much remains unknown. First of all, it should be noted that intervocalic consonants are much more common in conversational speech than initial or final consonants, as mentioned by Pickett, Bunnell & Revoile (1995). Therefore, in more natural settings listeners will generally have access to acoustic information on consonant identity at closure as well as release. Nevertheless, the study by Öhman is the only one employing intervocalic consonants, the other two use initial consonants only. Öhman's results are, however, difficult to generalize for several reasons. Most importantly, he used only a single speaker, vowel context, and stress condition. Second, all three studies concentrate on percentage correct classification, not on the confusion patterns themselves. Stated differently, they focus on the diagonal of the confusion matrices, paying less attention to the other cells. Studying the actual confusions may, however, reveal patterns in the responses that otherwise remain hidden.

The present study aims to increase our knowledge of the temporal distribution of perceptually relevant information for consonant recognition by avoiding some of the methodological drawbacks mentioned above. In order to be able to study the perceptual relevance of signal portions around closure as well as release of consonants, VCV utterances were gated both forward and backward in the present study. In the data analysis, the perceptually relevant information for various distinctive features, as well as patterns of confusions themselves were studied, rather than concentrating on percent correct scores only. Finally, in order to be able to draw conclusions that have a relatively high degree of generalizability, a large set of classification data was collected, using a relatively large number of utterances.

2. Method

2.1. *Stimuli*

2.1.1. *Original utterances*

Two speakers of British English, one male (speaker 1) and one female (speaker 2), each produced four tokens of 51 VCV nonsense words. The vowels /ɑ i u/ and the consonants /p t k b d g f θ s ʃ v ð z ʒ m n ŋ/ were used. The initial vowel was always identical to the final vowel. Two out of four tokens of each VCV combination were spoken with initial stress and two with final stress. Thus, 408 original utterances were obtained. The utterances were spoken in a soundproof anechoic room with a distance between mouth and microphone

of approximately 30 cm. The utterances were low-pass filtered at 10 kHz and quantized directly onto disk as 16-bit numbers at a sampling rate of 22.05 kHz. Next, the speech signals were digitally high-pass filtered at 30 Hz using a linear-phase filter in order to remove low-frequency components that may produce unwanted disturbances in combination with gating.

For each utterance, the closure and release landmarks were located visually using the waveform and wideband spectrogram. Generally, landmarks were put at instants where the amplitude of the first formant decreased or increased most rapidly. In difficult cases, like some of the voiced fricatives, a sudden increase or decrease in the amplitudes of the second and third formants was used as well. For all plosive consonants, the release landmark was put at the instant of burst release.

Appendix A presents statistics on the durations of consonantal closure, defined as the interval between the closure and release landmarks.

2.1.2. *Gating techniques*

Two types of gating were carried out as seen in Fig. 1. The perceptual relevance of the signal in the vicinity of the closure landmark was tested using *forward-gated* stimuli, that is, stimuli in which the part of the speech signal following an instant relative to the *closure* landmark was deleted. The perceptual relevance of the signal in the vicinity of the release landmark was tested using *backward-gated* stimuli in which the part of the speech signal preceding an instant relative to the *release* landmark was deleted.

Pols & Schouten (1978), among others, have argued that a sudden offset or onset of gated stimuli can lead to deterioration in performance not only because part of the information is removed but additionally because of the "click sensation" introduced by the gating, which is distracting to listeners or introduces auditory masking effects. Others have suggested that the sudden offset may suggest the presence of a (labial) stop in the signal (e.g., Ohala & Ohala 1995; Smits, Ten Bosch & Collier, 1996). Pols & Schouten showed that the deterioration in performance was greatly reduced by properly ramping the stimuli at onset/offset and replacing the deleted part of the signal with noise, rather than silence. Following their advice, stimuli used in the present study were ramped at the cutoff point using a linear 10 ms ramp. The deleted portion of the speech signal was replaced by a 500 ms pink noise signal which was, like the speech signals, low-pass filtered at 10 kHz and high-pass filtered at 30 Hz. The noise was scaled to a fixed level such that its maximum instantaneous amplitude was 0.031 times (or 30 dB below) the maximum instantaneous amplitude of the speech signals across all utterances. For each stimulus a different 500 ms pink-noise portion was selected randomly from a 10 s signal. A linear 10 ms ramp was applied at the edge of the noise adjoining the gated speech signal, and the two signals were overlap-added with the midpoints of the two ramps coinciding. Both for the forward- and backward-gated stimuli the cutoff ramp was centered at -80, -60, -40, $-20, 0, 20, 40, 60$, and 80 ms relative to the landmark, thus creating nine forward-gated stimuli (henceforth *closure* stimuli) and nine backward-gated stimuli (henceforth *release* stimuli) from each original utterance. 7344 gated stimuli were created in total. The gate positions are illustrated in Fig. 1 for an utterance /ɑgɑ/ with stress on the second syllable, spoken by speaker 1 (male). The closure stimuli contained the first syllable from its onset to one of the boundaries marked in Fig. 1, while the release stimuli began at one of the boundaries and continued to the end of the second syllable.

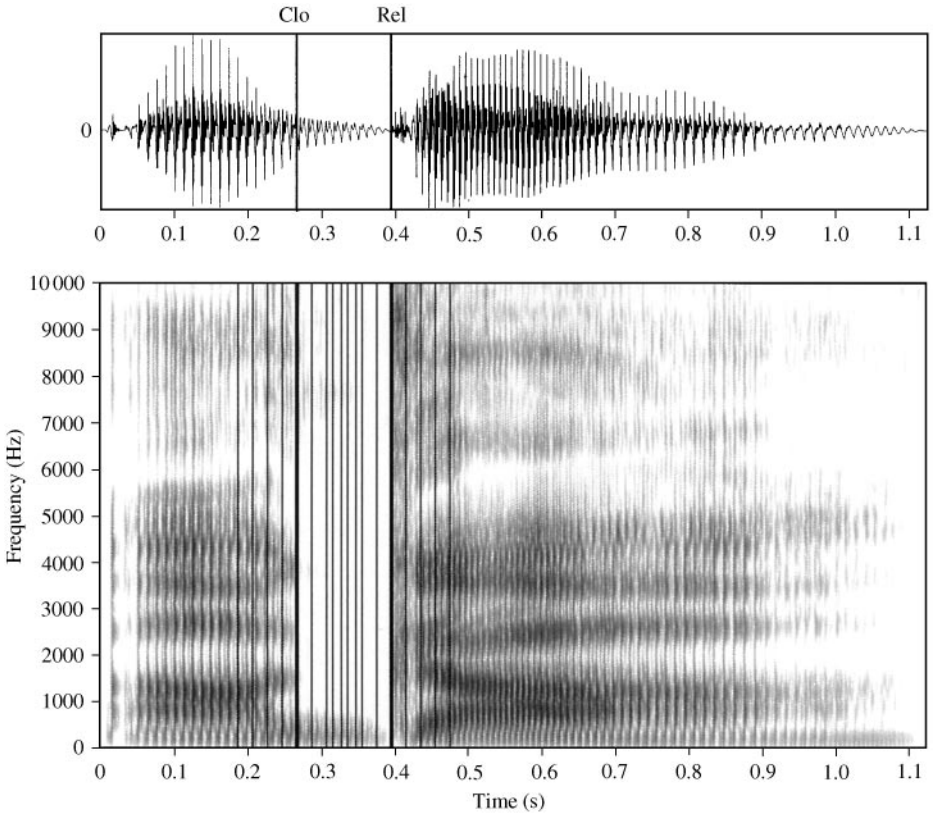

Figure 1. Illustration of the placement of gates for utterance /ɑgɑ/ spoken with final stress by speaker 1 (male). The top panel shows the waveform with closure and release landmarks. The bottom panel shows the spectrogram with the closure and release landmarks (thick lines) as well as the additional eight gate positions for each landmark (thin lines).

For 36 of the 408 original VCV utterances, the closure duration was less than 80 ms. Most of these utterances contained voiced stops and had stress on the initial vowel. For these utterances, the final gate (-80 ms for release stimuli, $+80$ ms for closure stimuli) contained both closure and release.

2.2. *Subjects*

Sixteen subjects took part in the experiments. All subjects were native speakers of British English and had normal hearing (within 20 dB HL for octave frequencies between 250 Hz and 8 kHz). None of the subjects were trained in phonetics or other speech-related sciences. The ages of the subjects varied between 18 and 38.

2.3. *Procedure*

The total set of gated stimuli was subdivided into four subsets: (1) speaker 1, initial stress; (2) speaker 1, final stress; (3) speaker 2, initial stress, and (4) speaker 2, final stress. Four

subjects were assigned to each of these subsets. Each subject took part in 11 experimental sessions. In the first session, the subjects were presented with the original utterances. They were asked to indicate the consonant by pressing one of 17 buttons. The labels on the buttons were ordered alpabetically: b, d, dh for /ð/, f, g, k, m, n, ng for /ŋ/, p, s, sh for /ʃ/, t, th for /θ/, v, z, and zh for /ʒ/. All subjects were trained briefly in the use of the labels dh, th, zh, and sh in a 4-alternative forced choice task. Three subjects could not master the use of the labels "dh" and "th" with an accuracy of at least 80% within 20 min of training. They were replaced by others. When a subject was sufficiently trained in the use of the labels, he or she was presented with five replications of each original utterance of the respective subset ordered randomly. All subjects passed the criterion of recognizing every consonant with an accuracy of 80% or higher. As in other studies with response sets of similar size (e.g., Miller & Nicely, 1955; Öhman, 1966; Furui, 1986), there were no indications that the subjects had difficulties with the size of the response set.

In the next 10 sessions, the subjects were presented with the gated stimuli. In every group, two subjects were presented with the closure stimuli in the first five sessions and with the release stimuli in the last five sessions. The order for the other two subjects in the group was reversed. In one session, all stimuli of the particular subset were presented once in random order. In the next session, all stimuli were presented again in a different random order, etc. The subjects were told that they would be presented with portions of the VCV utterances they had heard earlier, and they were instructed to indicate the consonant in the original VCV they thought the segment was taken from. Subjects were unaware of the purpose of the study. Before each batch of five sessions, the subjects were familiarized to the stimuli in a 10-min practice session, the results of which were discarded. On average, each subject spent about 20 hours on the 11 sessions of the experiment. Sessions were spread out over several days. Each session was divided into three runs by 10-min breaks. When all experiments were finished, each gated stimulus had been classified 20 times, resulting in 146 880 classifications in total.

Subjects were tested in a soundproof room one at a time. The stimuli were played directly from computer disk and were presented over headphones at a comfortable level. The experiments were self-paced. After the subject had pressed a response key, 2 s of silence followed, after which the next stimulus was played.

3. Results

In this section, descriptive data are presented which give a summary picture of the levels of correct categorization in the experiments. More detailed analyses of the confusion patterns and statistical tests of the influence of various experimental factors will be presented in later sections.

Fig. 2 presents levels of correct consonant recognition pooled across listeners, tokens, speakers, vowel contexts, and stress conditions. The left-hand panels give results for the closure stimuli, the right-hand panels for the release stimuli. From top to bottom the figures present results for all stimuli, fricatives only, nasals only, and stops only, respectively. Table I gives the 25%, 50%, and 75% correct points, rounded to 10 ms precision, for each of the panels in Fig. 2.

A first observation is that all data points in Fig. 2(a) and (b) are significantly above chance ($p < 0.0001$, chance level corresponds to 5.9% correct). Apparently, there is perceptually useful information present in the speech signal earlier than 80 ms before

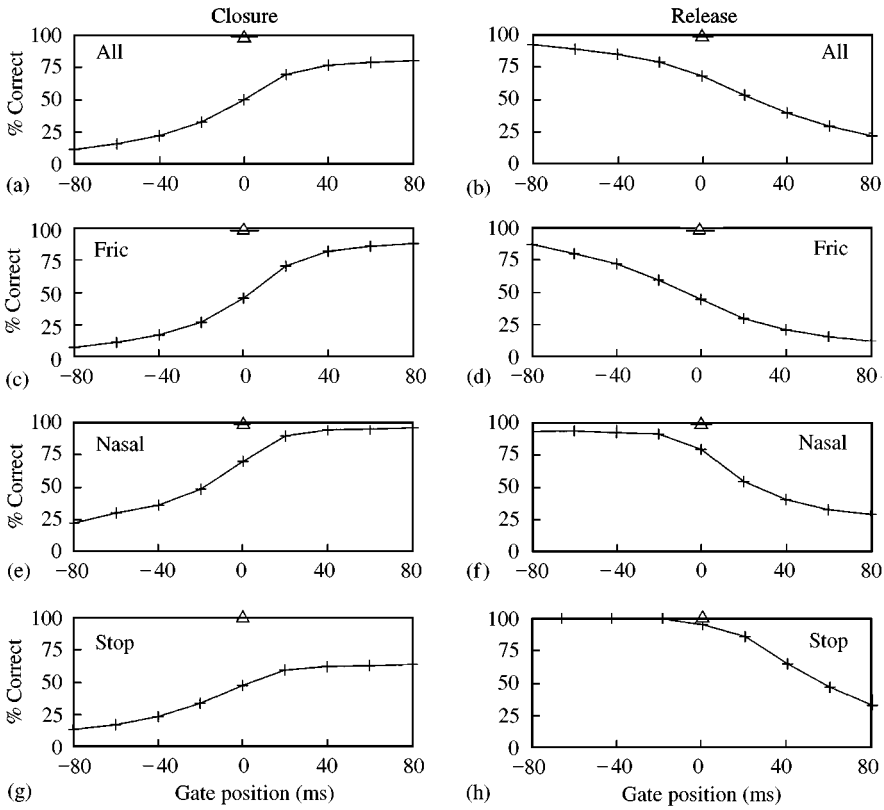

Figure 2. Percentage correct rates as a function of gate position for the closure (a, c, e, g) and release (b, d, f, h) stimuli. Triangular symbols at 0 ms indicate performance for the full VCV utterances. Rates are calculated across all consonants (a and b), for fricatives only (c and d), for nasals only (e and f), and for stops only (g and h). The temporal origins coincide with the relevant landmarks.

TABLE I. Gate positions at which levels of correct consonant recognition are 25%, 50% and 75%, for closure and release stimuli. The values in parentheses are based on extrapolations

	Gate position					
	Closure			Release		
Consonants (%)	25	50	75	75	50	25
All	−40	0	30	−10	30	80
Fricatives	−30	0	30	−40	−10	30
Nasals	−70	−20	0	0	30	(100)
Stops	−40	0	—	30	60	(100)

closure, and later than 80 ms after release in carefully uttered VCVs. Second, the location of the information, as captured by the 50% correct point, is somewhat different for fricatives, nasals and stops, especially around release. The 50% point for fricatives lies

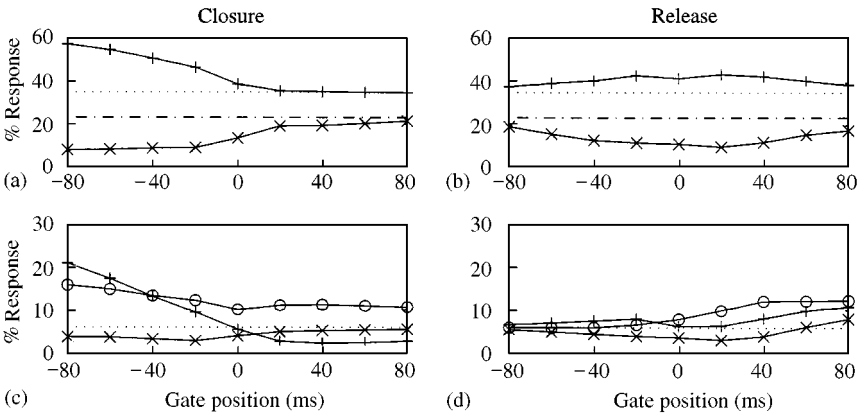

Figure 3. Percentages of stop ("+" symbols) and voiceless fricative ("×" symbols) responses as a function of gate position for closure (a) and release (b) stimuli. The dotted and dash–dotted lines indicate chance levels for stops (35.3%), and fricatives (23.5%), respectively. (c) and (d) give percentages of /p/ ("+" symbols), /b/ ("○" symbols), and /f/ ("×" symbols). The dotted line indicates chance level (5.9%).

10 ms before release, while that for stops lies 60 ms after release. This means that, in CV syllables, good fricative recognition requires at least some frication noise, while for good stop recognition some of the formant transitions plus the following vowel generally seems to be enough. A third observation is that the information for fricatives and nasals is more or less symmetrically distributed around closure and release. For stops, on the other hand, this is clearly not the case. Performance at closure levels off at some 65%, while at release performance is extremely reliable. Finally, Fig. 2(a) and (b) shows that the spread of information is somewhat wider around release than around closure. This is probably caused by the earlier discussed differences between the three manner classes around release. Pooling three categorization curves with different transition locations (50% points) will yield a curve with a relatively shallow slope.

In subsequent sections, more detailed analyses of these patterns are presented. First, however, the effects of the use of ramping and addition of pink noise preceding or following the gated stimuli will be briefly analyzed. As mentioned earlier, stimuli were ramped and pink noise was used in the present gating study with the purpose of minimizing deterioration in performance and perceptual biases introduced by the sudden onsets or offsets. The possibility remains, however, that the noise, which sounds somewhat like an /f/ frication, might have induced an /f/ percept, thus replacing one problem with another.

Fig. 3(a) and (b) present percentages of stop and voiceless fricative responses for all gates for the closure and release stimuli, calculated across all utterances. Fig. 3(c) and (d) give the percentages of responses /p/, /b/, and /f/. The following statements are evident in Fig. 3. First of all, there was never a bias towards /f/ or other voiceless fricatives. Secondly, there are no substantial bias effects for the release stimuli at all. Finally, there is evidence for a bias towards stops for early gates of the closure stimuli. Apparently, the use of noise following the gated stimuli has not completely prevented listeners from hearing something resembling a stop, although it may have reduced the effect. Note that in the study by Pols & Schouten (1978), which is the only one in which the effects

introduced by various gating techniques have been systematically investigated, only stop consonants were used. Obviously, no conclusions could be drawn about the effect of gating on manner perception. For the present study it is therefore concluded that the use of noise has probably reduced the additional place confusions, as in Pols & Schouten (1978), while the effect on manner recognition is unclear, although the use of noise has clearly not introduced a bias toward voiceless fricative responses.

4. Multidimensional scaling analysis

In order to interpret the principal confusion patterns in the data, a multidimensional scaling (MDS) analysis was carried out. MDS is a data analysis technique which can be used to construct a geometrical representation of stimuli on the basis of one or more confusion matrices. This representation can be interpreted as a psychological space, and the distances between the stimuli in the space corresponds to their perceived dissimilarity. An additional objective of the MDS analysis was to establish which confusions would be resolved at various stages of gating.

4.1. *Method*

4.1.1. *Input matrices and MDS settings*

For each gate, a single 17×17 confusion matrix was constructed by summing the categorization data across listeners, tokens, vowels, and stress. Separate confusion matrices were calculated for the two speakers and the two types of gating (forward around closure and backward around release). These confusion matrices were transformed into symmetrical distance matrices using the chi-squared distance measured D, defined in Appendix B.

A subset of 20 out of the total set of 36 distance matrices were selected on the basis of their degree of confusion. As MDS is based on the analysis of confusions, matrices which hardly display any confusion are unsuitable for entry in the analysis. Matrices which display very high levels of confusion, on the other hand, do not hold much information. The matrices for the following gates had intermediate levels of confusion and were selected:

Closure: -40, -20, 0, 20, 40 ms;
Release: -20, 0, 20, 40, 60 ms.

For each gating type, 10 matrices (5 gates $\times$ 2 speakers) were entered in a single nonmetric individual-differences MDS analysis using the ALSCAL program ("Alternating least-squares algorithm for individual differences scaling," Takane, Young & de Leeuw, 1977), applying the "matrix conditional" setting (see Appendix B). Originally, the ALSCAL analysis was designed for a set of matrices obtained for different subjects, leading to a universal underlying perceptual space for all subjects. Like Soli & Arabie (1979), however, matrices were entered for different conditions (in this case gates), rather than subjects. This approach is more suitable here because it causes patterns of confusions that differ between gates to be assigned to separate dimensions. For example, suppose that, for the closure stimuli, place of articulation confusions would become resolved for gates after -20 ms, while manner confusions would only become resolved for gates after

+20 ms. An individual differences MDS analysis carried out on separate matrices for different gates would assign place and manner confusions to different dimensions. That is, VCV stimuli containing consonants with different place of articulation would be close together (confusable) on, say, dimension 1, and well-separated (non-confusable) on dimension 2, while the reverse holds for consonants with different manner of articulation. Inspection of the dimension weights for various gates will reveal which confusions become resolved at which gates. Additionally, differences between speakers in the distribution of consonantal information across the time dimension will be revealed in the dimension weights.

4.2. *Results*

4.2.1. *Closure stimuli*

Stimulus configuration. The percentages of variance accounted for (VAF) by the two- to six-dimensional solutions for the closure stimuli were 62%, 66%, 76%, 81%, and 84%, respectively. The four-dimensional (4D) solution (VAF = 76%) was selected for two reasons. First, the largest break in VAF as a function of dimensionality occurs between dimensionality 3 and 4, suggesting that the 4D solution captures significant additional structure in the data that is not captured by the lower dimensions. Second, inspection of the 3D and 4D solution revealed that the 3D solution showed a circular pattern characteristic of solutions of too-low dimensionality (Takane *et al.*, 1977), and was less easily interpreted than the 4D one.

The configuration of the 17 consonants in the four-dimensional stimulus space is presented in Fig. 4. Extra lines have been added to the figure to facilitate interpretation. All four dimensions appear to be associated with distinctive features, although not in a straightforward way. Dimensions 2 and 4, depicted in Fig. 4(a), are associated with manner of articulation of voicing. Dimension 2 primarily separates nasals from other

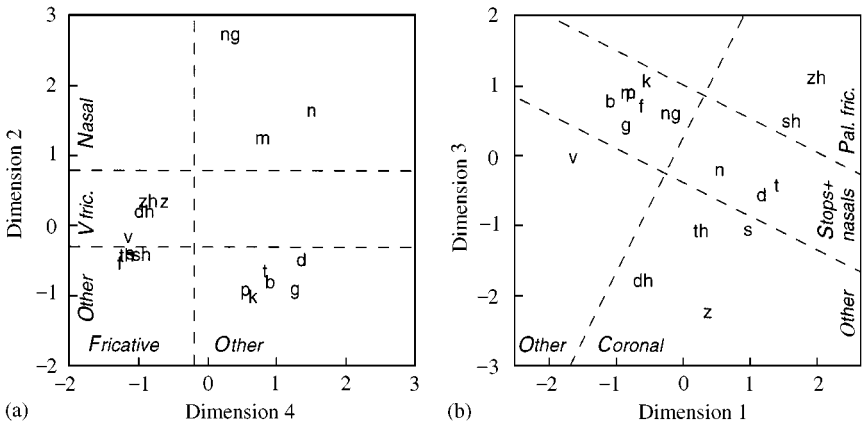

Figure 4. Stimulus configuration in the four-dimensional MDS solution for the closure stimuli. The dashed lines were added to facilitate interpretation. The labels "dh, th, zh, sh, ng" indicate consonants /ð θ ʒ ʃ ŋ/, respectively.

consonants. The latter group can be further subdivided into voiced fricatives *vs.* voiceless fricatives and stop consonants. In terms of voicing energy during consonantal constriction, nasals are the most voiced, followed by voiced fricatives, and finally other consonants. Thus, dimension 2 is associated with voicing. The interpretation of dimension 4 is very straightforward. It separates fricatives from stops and nasals. It is therefore associated with the feature continuant.

The interpretation of dimensions 1 and 3 is less clear-cut. It proved to be easiest to interpret Fig. 4(b) using a pair of dimensions that was slightly rotated with respect to the dimensions selected by the MDS analysis. Let us define dimension 1a as dimension 1 rotated clockwise over some 60°. This dimension separates coronal consonants from non-coronal ones. Dimension 3a, defined as dimension 3 rotated in the same way, can be interpreted as separating palatal fricatives /ʃ/ and /ʒ/ from other fricatives. Stops and nasals have neutral values along dimension 3a, and are accordingly located between the palatals and the other fricatives.

The question arises why MDS did not find dimensions 1a and 3a in the first place. The answer is that individual differences MDS (as used here) only finds dimensions that are special, or have preferred psychological status, if the relative weights of the dimensions involved changes sufficiently across the confusion matrices on which the analysis is based (Schiffman, Reynolds & Young, 1981). Translated to the present case, this means that the relative importance of dimensions 1 and 3 does not change significantly across gates, as will be validated in the next section. In such a case, the resulting orientation of the plane becomes highly sensitive to small, irrelevant aspects of the confusion matrices and is not meaningful.

Finally, it is useful to note that in the plane defined by dimensions 1 and 3, which is loosely associated with place of articulation, the consonants are less tightly clustered into groups than in the manner-and-voicing plane defined by dimensions 2 and 4.

Weighting of psychological dimensions over time. Fig. 5 shows the evolution of dimension weights as a function of gate position for the two speakers. The weight of a dimension can be interpreted as its importance in the classification process. A large weight effectively stretches a dimension, making objects more dissimilar, and therefore easier to recognize correctly. In the matrix-conditional ALSCAL analysis, information on the overall level of confusion is lost and only the proportions of the weights are preserved (MacCallum, 1977). For the interpretation of Fig. 5 it should be kept in mind that the overall level of confusion decreases rapidly across gates, although this is not visible in Fig. 5. Therefore, only the relative importance of the dimensions will be discussed.

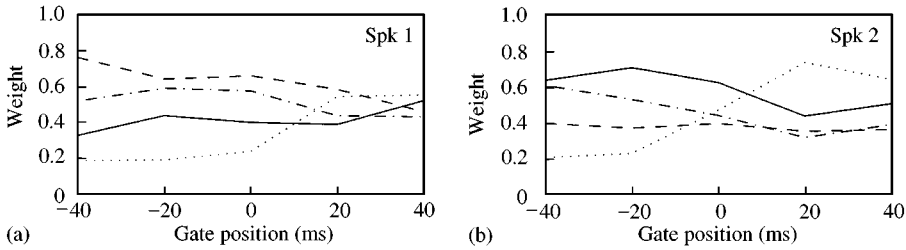

Figure 5. Dimension weights for the four-dimensional MDS solution for the closure stimuli for speaker 1 (a) and speaker 2 (b). Solid, dashed, dash-dotted and dotted lines are associated with dimensions 1, 2, 3, and 4, respectively. Refer to Fig. 4 for the interpretation of the dimensions.

First of all, Fig. 5 shows that the relative weight of dimension 4 rises abruptly at closure for both speakers. Dimension 4 is associated with the fricative–non-fricative distinction. This sudden increase in distinctiveness of fricatives *vs.* stops and nasals can be easily explained by the fact that when, during the production of a vowel-fricative sequence, the point of maximum constriction is reached (at 0 ms in Fig. 5), frication starts, which makes the fricative easily distinguishable from other consonants.

Secondly, the graphs of the weights for the place dimensions 1 and 3 are more or less parallel, both for speakers 1 and 2. As discussed above, it can therefore be concluded that the orientation of the plane defined by dimensions 1 and 3 is not well defined. This confirms that the definition of new dimensions 1a and 3a as linear combinations of dimensions 1 and 3 for the purpose interpretation was in principle allowed.

Fig. 5 also highlights a difference between the relative salience of consonant information for the two speakers. For pre-closure gates, the weights for the place-associated dimensions 1 and 3 dominate the weights for the manner/voice-associated dimensions 2 and 4 for speaker two. For speaker one, on the other hand, dimension 2, which mainly separates nasals from other consonants, has the largest weight up to 20 ms after closure. Apparently, upcoming nasals are easier to predict from pre-nasal vowels of speaker 1 than from speaker 2.

4.2.2. *Release stimuli*

Stimulus configuration. The percentages of variance accounted for (VAF) for the two- to six-dimensional solutions for the release stimuli were 71%, 82%, 86%, 88%, and 90%, respectively. As for the closure stimuli, the four-dimensional solution (VAF = 86%) was selected.

The configuration of the 17 consonants in the four-dimensional stimulus space is presented in Fig. 6. Again, all four dimensions can be interpreted in a phonetically meaningful way. Dimensions 1 and 3, depicted in Fig. 6(a), are associated with manner of articulation and stop voicing. Dimension 1 very crisply separates voiceless plosives from other consonants. Dimension 3 mainly separates nasals from other consonants and to a lesser extent separates fricatives and plosives, with the exception of /d/, which is located

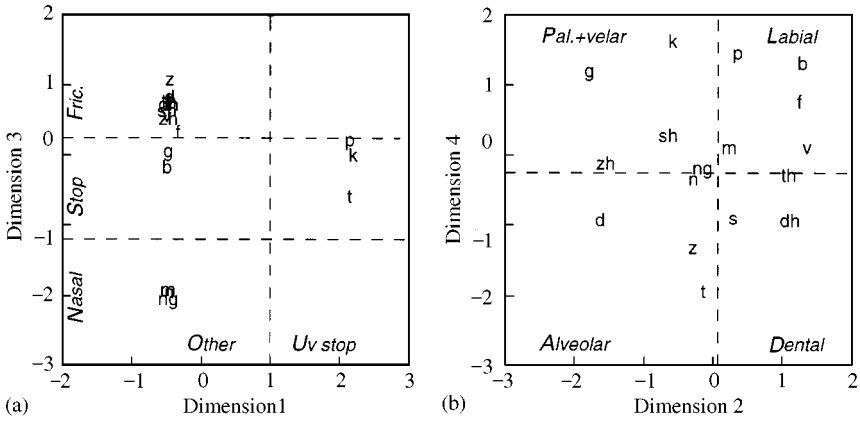

Figure 6. Stimulus configuration in the four-dimensional MDS solution for the release stimuli.

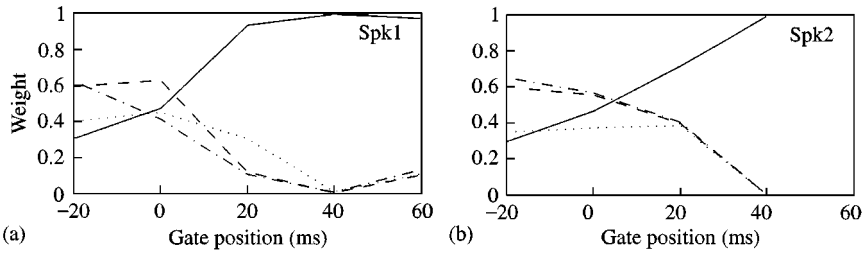

Figure 7. Dimension weights for the four-dimensional MDS solution for the release stimuli for speaker 1 (a) and speaker 2 (b). Solid, dashed, dash-dotted and dotted lines are associated with dimensions 1, 2, 3, and 4, respectively. Refer to Fig. 6 for the interpretation of the dimensions.

in the fricative area. Dimensions 2 and 4 in Fig. 6(b) are mainly associated with place of articulation. The consonants are more evenly distributed in Fig. 6(b) than in Fig. 6(a), and the distribution of consonants along dimensions 2 and 4 is best interpreted in the two-dimensional plane. The upper right-hand quadrant contains only labials, the lower right-hand quadrant contains the dental fricatives (plus /s/). The upper left-hand quadrant contains the palatal and velar consonants, while the lower left-hand quadrant contains the alveolars, except /s/.

Weighting of psychological dimensions over time. The evolution of dimension weights is given in Fig. 7. When examining Fig. 7, it is important to realize that going from negative to positive gate positions, the information for the listeners decreases and confusions increase. The most striking feature of Fig. 7 is that for late gates the picture is completely dominated by dimension 1 for both speakers. When only the final part of the utterance after closure is presented, listeners can easily hear the difference between voiceless plosives and other consonants, presumably on the basis of the aspirated formant transitions. For both speakers, the weights of all four dimensions are similar at closure (0 ms). This means that the psychological space displayed in Fig. 6 holds at closure, without much stretching or shrinking of axes. When some pre-release information is present (-20 ms), dimensions 2 and 3 gain in importance, indicating that manner recognition improves, as well as the front–back place distinction. Soon after release manner recognition (except for the voiceless stops) collapses for speaker 1, and somewhat later for speaker 2. For both speakers some place information is still present in the signal 20 ms after release.

Overall, the differences in the evolution of weights between the two speakers were deemed too small to warrant further investigation.

4.3. *Discussion*

The principle purpose of the MDS analyses was to describe major patterns of confusions in the classification data. In terms of manner of articulation and voicing, the closure stimuli were clustered as nasal consonants, voiced fricatives, voiceless fricatives, and plosives, while the release stimuli were clustered as nasals, fricatives, voiced plosives, and voiceless plosives. The distribution of consonants across the plane associated with place of articulation was less strongly clustered. However, the MDS solutions suggested that the following groupings of place articulation have perceptual relevance: (1) labials + velars, (2) dentals + alveolars, and (3) palatals for the closure stimuli; (1) labials, (2) dentals, (3) alveolars, and (4) palatals + velars for the release stimuli. These

patterns will be used in the definition of classes for the analyses of information transmission presented in the next section.

5. Information-theoretical analysis

5.1. *Method*

As described earlier, it is the goal of the study presented in this paper to describe the temporal distribution of perceptually relevant information for consonant recognition. This objective is achieved most directly by calculating *transmitted information* (TI) as a function of gate position. TI was chosen instead of percent correct (PC) for two reasons. First of all, TI = 0 when listeners perform at chance level, regardless of the number of possible responses. Second, in contrast with PC, TI is insensitive to strong response biases. For example, suppose that listeners would categorize all closure stimuli at early gates predominantly as stops. Results expressed in PC would show high PC for stops, suggesting that stop information is present in the speech signal long before closure. TI, on the other hand would reflect listeners' inability to reliably *discriminate* between stops and other consonants, and would be close to zero.

Usually, TI analyses of consonant recognition data are carried out in terms of distinctive features (e.g., Miller & Nicely, 1955). Many coding schemes of consonants in terms of distinctive features are possible, however, and it would seem that the suitability of competing feature-coding schemes for describing a data set is equally dependent on the particular acoustic manipulation adopted in the stimulus preparation as on properties of the human speech perception system. In particular, the grouping of consonants according to place of articulation differs greatly between authors and experiments. For example, the palatal fricatives /ʃ ʒ/ are sometimes assigned to the "back" place of articulation, along with velar consonants, sometimes to a "mid" category, along with alveolars, while at other times they are allocated a separate place category (for discussions, see Wang & Bilger, 1973; Singh, 1975). The multidimensional scaling study described in the previous section suggests that a relatively straightforward and intuitively appealing coding scheme is suitable for the analysis of the current data set. This coding scheme is given in Table II.

As Table II shows, TI was calculated in terms of manner, place, and voice, with three manner classes (fricative, nasal, stop), five place classes (labial, dental, alveolar, palatal, velar), and two voice classes (voiced, voiceless). However, on the basis of the MDS results, it was deemed useful to calculate TI for place and voice separately for different manner classes. Thus, six features were calculated: manner, place for fricatives, nasals and stops, and voice for fricatives and stops.

In order to have measures of location and spread of consonantal information along the time axis directly available for statistical testing, it was considered useful to fit a parametric curve to each set of TI points for nine gates, and enter the estimated parameters into a MANOVA. After inspecting the general shape of the TI-*vs.*-time curves it was decided to fit the following four-parameter sigmoid-based model to the data:

$$I(t) = C_b + \frac{C_t - C_b}{1 + \exp(-(t - P)/W)} \tag{1}$$

TABLE II. Feature coding scheme used in the analysis of information transmission

Feature	Values	Members
Manner	Fricative	/f θ s ʃ v ð z ʒ/
	Nasal	/m n ŋ/
	Plosive	/p t k b d g/
Place, fricatives	Labial	/f v/
	Dental	/θ ð/
	Alveolar	/s z/
	Palatal	/ʃ ʒ/
Place, nasals	Labial	/m/
	Alveolar	/n/
	Velar	/ŋ/
Place, stops	Labial	/p b/
	Alveolar	/t d/
	Velar	/k g/
Voice, fricatives	Voiced	/v ð z ʒ/
	Voiceless	/f θ s ʃ/
Voice, stops	Voiced	/b d g/
	Voiceless	/p t k/

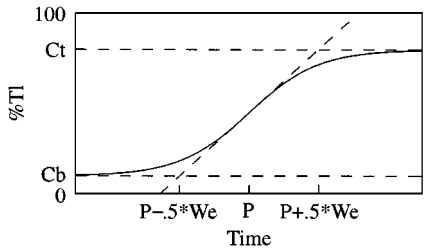

Figure 8. Graphic representation of Equation (1), which models transmitted information as a function of time. The dashed horizontal lines represent the top and bottom asymptotes C_t and C_b. The dashed slanted line is the tangent of the sigmoid function at its midpoint P. The intersection of the tangent with the sigmoid's asymptotes define the equivalent width W_e of the sigmoid.

where I represents transmitted information, t represents time (gate position), C_b and C_t represent the bottom and top asymptotes of the sigmoid function, and P and W represent the *position* and *width* of the sigmoid, respectively. For the data for the closure stimuli W is positive, for release data W is negative. For the purpose of interpretation the concept of *equivalent width* W_e of the sigmoid function $I(t)$ is introduced. W_e is defined as the interval between the two points where the tangent of $I(t)$ in the point $t = P$ intersects the bottom and top asymptotes of $I(t)$. This leads to $W_e = 4W$. Fig. 8 illustrates the sigmoid-based model for $I(t)$ and the interpretation of various model parameters.

In order to be able to carry out statistical tests for the influence of feature, vowel, stress, speaker, and listener on TI-*vs.*-time curves, 288 separate fits were made of the function defined in (1) to the closure data, and 288 to the release data. Function (1) generally fitted

the TI-*vs.*-time data very well. The RMS error computed across all sigmoid fits was 5.2%. In 23 out of the 288 fits for the closure data the position and width parameters assumed meaningless values because the TI values for all gates were very close to zero. These parameter values were not used in the statistical analysis.

Separate MANOVAs were carried out on the closure and release data, with the parameters C_b, C_t, P and W as dependent variables and feature, vowel, stress, speaker and listener as independent variables. Listener was defined as a random factor nested under speaker and stress. All two- and three-way interactions between feature, vowel, stress and speaker as well as the two-way interactions feature × listener and vowel × listener were used in the MANOVA model. Besides calculating separate ANOVAs for all dependent variables, the MANOVA also tests for the significance of main effects and interactions on the four dependent variables simultaneously.

5.2. *Results*

5.2.1. *Descriptive data*

Before the results of the statistical analyses are given, first some summary data are presented. Fig. 9 gives the percentage transmitted information as a function of gate position for the closure- and release stimuli. Here, TI is calculated on the basis of nine 17×17 consonant confusion matrices for closure and nine for release, pooled across vowel, stress, speaker and listener. Fig. 9 is similar to Fig. 2(a) and (b), in which percent correct was plotted rather than TI. Recall, however, that TI at chance level is zero.

Fig. 10 represents the percentage transmitted information as a function of gate position for the closure- and release stimuli calculated separately for the earlier defined distinctive features. Again, the classification data were pooled across vowel, stress, speaker and listener. As in Fig. 9, solid lines indicate the fitted sigmoid functions. Fig. 10(a) and (b) suggest that manner information becomes available very abruptly around closure and more gradually around release. The distribution of place information (Fig. 10(c) and (d)) is more similar around closure and release. In both cases, place information for fricatives is more spread-out than for plosives and nasals. Fig. 10(d) suggests furthermore that the transition regions of plosives are more informative on place than those of nasals and fricatives. Finally, Fig. 10(e) and (f) shows large differences in voicing

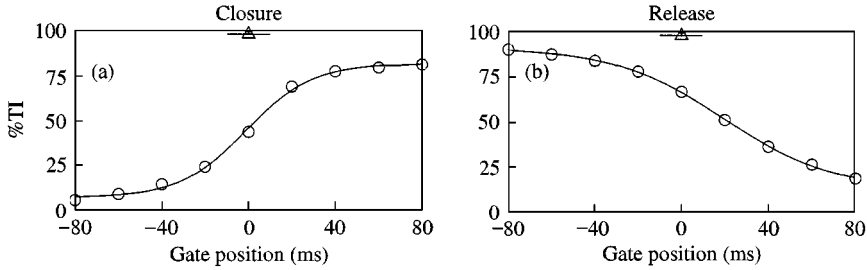

Figure 9. Percentage transmitted information as a function of gate position for the closure (a) and release (b) stimuli. The solid lines represent the sigmoid function fitted to the data. Triangular symbols at 0 ms indicate performance for the full VCV utterances. The origins on the time axes are set to coincide with the relevant landmarks.

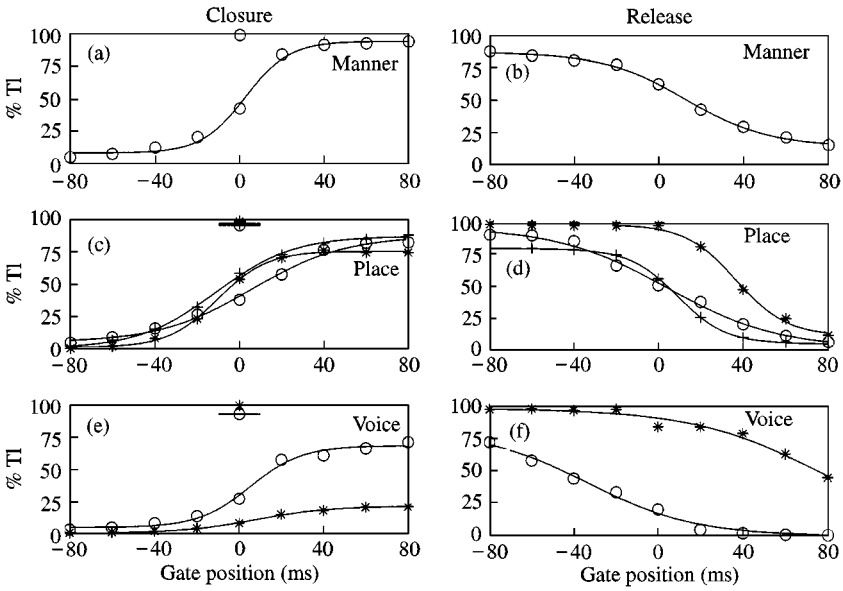

Figure 10. Percentage transmitted information as a function of gate position for the closure (a, c, e) and release (b, d, f) stimuli, calculated separately for the distinctive features manner place and voice. Again, the solid lines represent the sigmoid functions fitted on the data and the temporal origins coincide with the relevant landmarks. The isolated symbols plotted at 0 ms in panels a, c, and e indicate performance for the full VCV utterances. In panels (c) to (f), symbols ○, +, ∗, indicate fricatives, nasals, and stops, respectively.

perception for plosives and fricatives. Voicing information for plosives seems to be largely cued by the formant transitions after release. For fricatives, on the other hand, voicing perception seems to be based almost exclusively on the frication portion between the two landmarks.

5.2.2. *Statistical analyses*

The MANOVA on the parametrized sigmoid functions for the *closure* stimuli, using C_b, C_t, P and W as dependent variables, showed that vowel, feature, and listener were highly significant main effects ($p < 0.0001$). Speaker was not significant ($p > 0.2$), neither was stress ($p > 0.7$). The only significant interactions were speaker × feature ($p < 0.0001$), stress × feature ($p < 0.01$) and listener × feature ($p < 0.0001$). Fig. 11 shows the means of dependent variables position and width for all main effects, except listener. Non-significant differences, as revealed by Duncan *post-hoc* tests on individual dependent variables, are indicated by lines above the bars. Note that it may occur that a factor which is significant in the MANOVA does not have a significant influence on an individual dependent variable such as P or W.

The MANOVA on the data for the *release* stimuli showed that vowel ($p < 0.0001$), speaker ($p < 0.005$), stress ($p < 0.04$), feature ($p < 0.0001$), and listener ($p < 0.0001$), that is, all main effects, were significant. The only significant interactions were speaker × feature ($p < 0.0001$), speaker × stress × feature ($p < 0.04$) and listener × feature

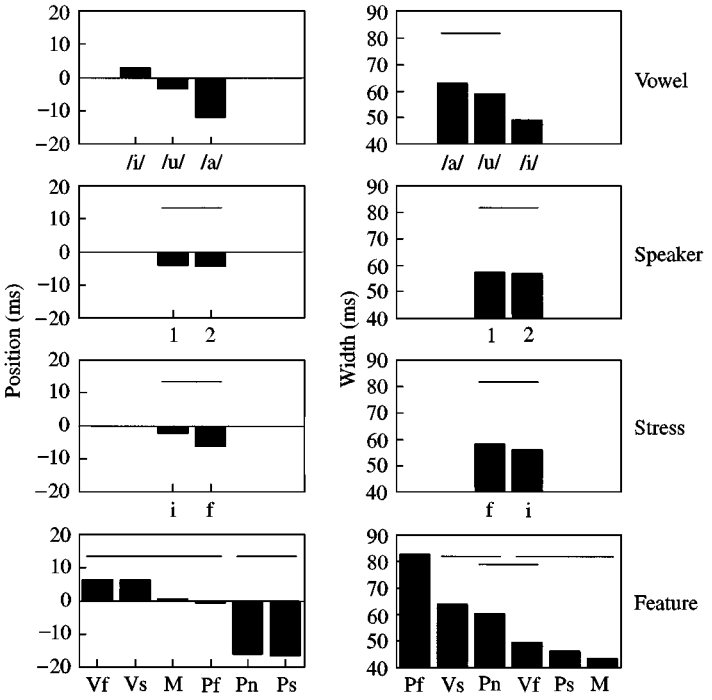

Figure 11. Closure stimuli: Means of position and equivalent width, both expressed in ms, of the sigmoid functions as a function of the factors vowel, speaker, stress, and feature. Non-significant differences, as revealed by Duncan *post-hoc* tests, are indicated by lines above the bars. Speakers are indicated by "1" and "2", initial and final stress by "i", "f". Manner, place for fricatives, nasals and plosives, and voicing for fricatives and plosives are indicated by M, Pf, Pn, Ps, Vf, Vs, respectively.

($p < 0.0001$). Fig. 12 shows the means of dependent variables position and width for all main effects.

Overall, the effects of vowel context, speaker identity and stress on the position and width of the information distributions around closure and release are small. Although the effect of vowel is significant for both position and width around closure, the effect is not large (in the order of 10 ms). The effects of speaker identity and stress, which are only significant for position around release, are even smaller. By far the most striking effect is that of feature. Around closure, the perceptually relevant information on place for nasals and stops is available earlier than information on all other features. Place information for stops and nasals is concentrated around roughly 15 ms before closure, while information for other features is concentrated around roughly 5 ms after closure. Place information for fricatives is more spread-out than that for other features. The equivalent width of the sigmoid representing the place information for fricatives is some 80-odd ms, while the equivalent width for the other features is roughly 50 ms. Around release the situation is different. As noted in the multidimensional scaling analysis, voicing information for plosives is available well into the second syllable. On average the information is concentrated around 70 ms after release. Information on place in stops is concentrated at 40 ms after release. Information on place in nasals and fricatives and on manner is located at

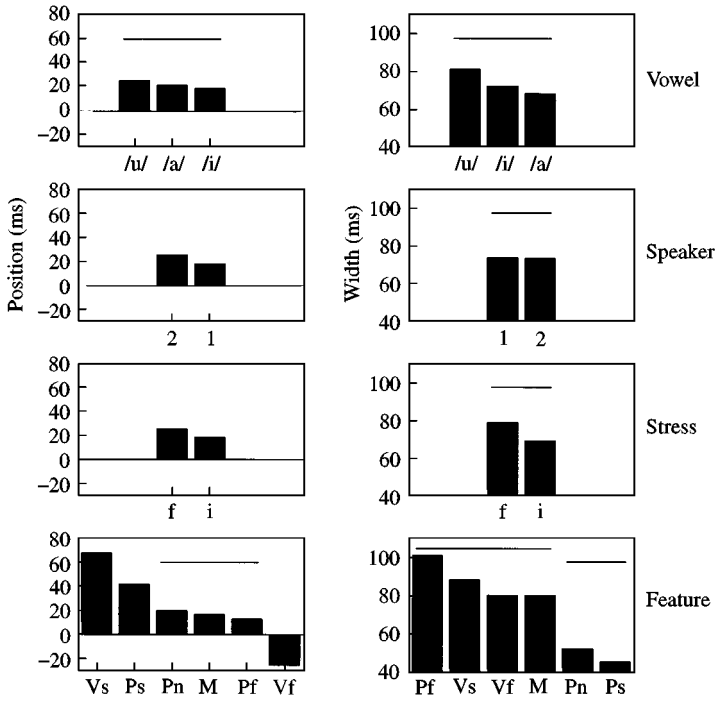

Figure 12. Release stimuli: Means of position and equivalent width, both expressed in ms, of the sigmoid functions as a function of the factors vowel, speaker, stress, and feature. Non-significant differences, as revealed by Duncan *post-hoc* tests, are indicated by lines above the bars. Speakers are indicated by "1" and "2", initial and final stress by "i", "f". Manner, place for fricatives, nasals and plosives, and voicing for fricatives and plosives are indicated by M, Pf, Pn, Ps, Vf, Vs, respectively.

roughly 15 ms after release, and voicing information for fricatives is located some 25 ms before release. The equivalent width of place information for stops and nasals is roughly 50 ms, while that for the other features is almost 90 ms.

6. General discussion

The present study investigated the temporal distribution of perceptually relevant information for consonant recognition in naturally spoken British English VCVs. The information distribution in the vicinity of consonantal closure was measured by presenting initial portions of naturally produced VCV utterances to listeners for identification. Final portions of the VCVs were used to study information distribution around release.

The classification data from the experiment were analyzed in three ways. First, summary data were presented, giving overall percentages correct for stops, nasals and fricatives as a function of gate position relative to closure and release. Next, a multidimensional scaling analysis was carried out. This provided highly interpretable, four-dimensional geometrical representations of the confusion patterns in the data. It was found that, for release as well as closure stimuli, the dimensions of the stimulus space

were associated with distinctive features, although this relation was not straightforward. For both types of gating, two dimensions were related to manner of articulation and voicing, while the two other dimensions were related to place of articulation.

The patterns of confusions were different for the two types of gating. The closure stimuli were clustered as plosives, nasals, and fricatives. In the fricative group, voiced and voiceless fricatives were slightly separated. The release stimuli were clustered as unvoiced plosives, nasals and other consonants, while the other-consonants group was divided into a subgroup containing /b g/ and a subgroup containing all fricatives plus /d/. With respect to place of articulation, the closure stimuli could be separated into coronals *vs.* non-coronals. The non-coronals could be subdivided into palatal fricatives, stops + nasals, and dental + alveolar fricatives. The release stimuli could be subdivided into labial, dental, alveolar, and palatal + velar consonants. For both types of gating, the stimuli were distributed much more evenly along the place dimensions than along the manner/voicing dimensions.

Inspired by the confusion patterns in the multidimensional scaling representations, a set of six distinctive features was defined for further analyses. These features were manner, place, and voicing, where place and voicing were analyzed separately per manner class. The subsequent analysis consisted of three steps. First, the percentage TI as a function of gate position was calculated separately for different features, vowels, syllable stress, speaker, and listener. Next, a sigmoid function was fitted to each of 288 sets of TI data thus obtained for each type of gating. Finally, the sigmoid parameters were entered into MANOVAs which tested the influence of feature, vowel, stress, speaker and listener on the distribution of perceptually relevant information for consonant identity. It was found that, although vowel, stress, speaker, and listener had significant effects on the distribution of information, feature was the dominant factor both around closure and around release. This means that the location and spread of the information is very different for the different distinctive features. Below, the major findings reported in earlier studies will be briefly reviewed and compared to the findings of the current study.

6.1. *Closure*

Far fewer studies have been done on information distribution around consonantal closure than around release. The gating study by Öhman on Swedish VCCVs (Öhman, 1966) showed an overall rapid rise in correct *manner* identification when signal portions after closure become available. Plosive manner was well identified for all gates, probably because truncated voiced utterances generally sound like stop closures. As manner classifications for short VC portions of other consonants were close to 0% correct, it seems likely that no information on the original manner of articulation of the segment was actually transmitted to the listeners. The instant of most rapid increase in correct identification of fricative manner was about 30 ms after closure, while for nasal manner it was roughly at closure according to Öhman (1966).

Place information from closure transitions only seems to be comparable for fricatives, nasals, and plosives. When portions after closure become available, correct place identification for fricatives slowly rises, for nasals rapidly rises, and for plosives hardly rises at all (Malécot, 1958; Öhman, 1966; Sharf & Hemeyer, 1972; Pols, 1979; Ohde & Sharf, 1981; Schouten & Pols, 1983; Recasens, 1983; Repp & Svastikula, 1988).

The main source of *voicing* information around closure for plosives is the presence or absence of a voice bar (Lisker, 1978), but the use of this information by listeners seems to

be different in different languages. Swedish listeners make good use of this information, resulting in rapidly growing correct voicing recognition when portions after stop closure are presented (Öhman, 1966). American English listeners, on the other hand, do much worse with similar signal portions (Malécot, 1958). Not much is known about voicing perception in intervocalic and postvocalic fricatives. Öhman (1966) showed that correct voicing identification increases rapidly with increasing postvocalic frication portions, reaching near-perfect levels when about 50 ms of frication after closure is included. This is corroborated by Jongman's (1989) finding that voicing identification in initial portions of syllable-initial frication noise reaches its ceiling at a duration of about 60 ms.

The present study shows that, around closure, place information for nasals and plosives is available earlier (instant of maximum increase roughly 15 ms before closure), than information on any other features (instant of maximum increase roughly 5 ms after closure). This generally agrees with the findings discussed above, where correct manner as well as voicing identification was reported to increase at or after closure. The increase in information transfer for fricative place in the current data is somewhat later than in previous studies, where it was found that pre-closure formant transitions were equally informative for the different manners of articulation. Possibly, this discrepancy is caused by using 4 (rather than 3) distinct fricative places of articulation, and assigning the consonants /ʃ 3/ to a separate place class (palatal) rather than the velar class (Öhman, 1966, for example, assigned palatal and velar consonants to the same place class). The present study also showed that the information distribution for fricative place is wider (equivalent width of 80 ms) than that of any other feature (on average 50 ms). This also agrees with the findings of earlier studies. Finally, the present study shows that British English listeners have great difficulty in identifying plosive voicing when only signal portions preceding release are available, that is, they hardly make use of the information present in the voice bar during closure. Inspection of the individual data revealed that there were great differences between subjects in this respect. Some listeners simply classified all closure plosives as unvoiced, while others did a reasonable job in identifying plosive voicing. The American English listeners tested by Malécot (1958) seemed generally to be doing somewhat better than the listeners in the present study, while Öhman's (1966) listeners appeared to make very effective use of the voice-bar information. The difference with the Swedish listeners may be caused by Öhman's V-plosive-V utterances being unaspirated (Swedish plosives are aspirated only in stressed syllables), which necessarily shifts the perceptual weight to other cues, such as the presence or absence of a voice bar. In the present study, all voiceless plosives were aspirated.

6.2. *Release*

The understanding of information distribution around release is better developed than around closure, although this is mostly based on studies with CV rather than VCV utterances. The gating study by Öhman (1966) showed that correct *manner* identification changes abruptly around release, with nasal manner changing roughly at release, fricative manner changing before release and plosive manner after. Other studies indicate that fricative-vowel syllables with frication removed and nasal-vowel syllables with murmur removed are generally perceived as plosives (Manrique & Massone, 1981; Repp, 1986; Kurowski & Blumstein, 1987).

A large number of experiments have shown that removal of the release burst has a significant effect on correct *place* identification in plosives (Sharf & Hemeyer, 1972;

Pols, 1979; Ohde & Sharf, 1981; Schouten & Pols, 1983; Smits *et al*., 1996). Additional truncation of the subsequent formant transitions results in a rapid decline in correct place identification, reaching chance level at about 100 ms after release (Öhman, 1966; Pols, 1979; Furui, 1986). For place identification in fricatives information in both frication noise and formant transitions is used by listeners (e.g., Mann & Repp, 1980; Whalen, 1981). The frication portion is generally found to be of greater importance than the formant transitions, except for the distinction between /f/ and /θ/ (Harris, 1958). However, the level of correct place perception in fricative-vowel syllables from the vocalic portion only is comparable to place perception for the same portion in (voiced) plosive-vowel syllables (Sharf & Hemeyer, 1972; LaRiviere, Winitz & Herriman, 1975; Manrique & Massone, 1981; Jongman, 1989). For place identification in nasals it is generally found that the formant transitions are somewhat more informative than the nasal murmur, although the murmur does carry significant perceptual weight. However, a very short signal portion around release including both some murmur and some of the formant transitions is enough for good place recognition. Again, place recognition from transitions alone is comparable to plosive transitions (Repp, 1986; Kurowski & Blumstein, 1987).

Although many cues around release are perceptually relevant for *voicing* identification in plosives, VOT is generally dominant (Lisker, 1978). When increasing portions preceding the instant of voicing onset are included, correct identification gradually increases. The situation for fricatives is similar in that correct voicing identification gradually increases when increasing portions of frication (preceding release) are included.

In the present study it was found that, when increasing portions of the final part of the VCVs were presented, first plosive voicing information becomes available (70 ms after release), then plosive place (40 ms after release), then manner and nasal and fricative place (15 ms after release) and finally fricative voicing (25 ms before release). This partially agrees with findings reported in the literature. Plosive voicing being available well after release agrees with the reported finding that some information preceding voicing onset is generally enough for correct voicing identification. The finding that place information is available further after release for plosives than for fricatives and nasals seems somewhat contradictory to earlier studies reporting that the place information in post-release transitions is similar for plosives, nasals, and fricatives. However, while for fricatives and nasals the formant transitions start at release, plosives have a burst between the instants of release and (voiced or aspirated) transitions, which roughly accounts for the 25 ms difference. With respect to the spread of information for the various features, the present study found that the width of the information distribution for plosives and nasals (50 ms) is smaller than for other features (90 ms). The fact that the place distribution is narrower for nasals and plosives than for fricatives is supported by earlier studies. An explicit comparison with the manner and voicing features does not seem to be available in the literature, though.

It was the aim of this paper to increase our knowledge about the temporal distribution of information for human consonant recognition. One feature of the study was that a very large set of data was collected, using two speakers, three vowel contexts and two stress conditions, which increases the robustness of the reported findings. Besides the analyses presented here, many other data analyses, inspired by different research questions, may be applied to the experimental data, such as analyses of (in)dependence of feature evaluations, or the testing of quantitative models of consonant perception. For that purpose the experimental data are freely available from the author.

This research was carried out while the author was with the Department of Phonetics and Linguistics, University College London, U.K. and was supported by a grant from NATO and the Netherlands Organization for Scientific Research (NWO) and by a Marie Curie fellowship granted by the European Commission. I am very grateful to Valerie Hazan and Stuart Rosen for advice, encouragement, and discussions, and to Anne Cutler, Doug Whalen and three anonymous reviewers for useful comments on an earlier version of this paper. A limited number of CD-ROMs with all stimuli and responses of the experiments are available from the author.

References

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. & Gerstman, L. J. (1952) Some experiments on the perception of synthetic speech sounds, *Journal of the Acoustical Society of America*, **24,** 597–606.

Fletcher, H. (1953) *Speech and hearing in communication.* New York: Krieger.

Furui, S. (1986) On the role of spectral transition for speech perception, *Journal of the Acoustical Society of America*, **80,** 1016–1025.

Grimm, W. A. (1966) Perception of segments of English-spoken consonant-vowel syllables, *Journal of the Acoustical Society of America*, **40,** 1454–1461.

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables, *Language and Speech*, **1,** 1–7.

't Hart, J. & Cohen, A. (1964) Gating techniques as an aid in speech analysis, *Language and Speech*, **7,** 22–39.

Jongman, A. (1989) Duration of frication noise required for identification of English fricatives, *Journal of the Acoustical Society of America*, **85,** 1718–1725.

Kurowski K. & Blumstein, S. E. (1987) Acoustic properties for place of articulation in nasal consonants, *Journal of the Acoustical Society of America*, **81,** 1917–1927.

LaRiviere, C., Winitz, H. & Herriman, E. (1975) The distribution of perceptual cues in English prevocalic fricatives, *Journal of Speech and Hearing Research*, **18,** 613–622.

Lisker, L. (1978) *Rapid vs. rabid*: *a catalogue of acoustic features that may cue the distinction.* Haskins Laboratories Status Report on Speech Research SR-54, pp. 127–132.

MacCallum, R. C. (1977) Effects of conditionality on INDSCAL and ALSCAL weights, *Psychometrika*, **42,** 297–305.

Malécot, A. (1958) The role of releases in the identification of released final stops, *Language*, **34,** 370–380.

Mann, V. A. & Repp, B. H. (1980) Influence of vocalic context on perception of the [ʃ]-[s] distinction, *Perception and Psychophysics*, **28,** 213–228.

Manrique, A. M. B. de & Massone, M. I. (1981) Acoustic analysis and perception of Spanish fricative consonants, *Journal of the Acoustical Society of America*, **69,** 1145–1153.

Miller, G. A. & Nicely, P. E. (1955) An analysis of perceptual confusions among some English consonants, *Journal of the Acoustical Society of America*, **27,** 338–352.

Ohala, J. J. & 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* (B. Connell & A. Arvaniti, editors), pp. 41–60. Cambridge: Cambridge University Press.

Ohde, R. N. & Sharf, D. J. (1981) Stop identification from vocalic transition plus vowel segments of CV and VC syllables: a follow-up study, *Journal of the Acoustical Society of America*, **69,** 297–300.

Öhman, S. E. G. (1966) Perception of segments of VCCV utterances, *Journal of the Acoustical Society of America*, **40,** 979–988.

Pickett, J. M., Bunnell, H. T. & Revoile, S. G. (1995) Phonetics of intervocalic consonant perception: retrospect and prospect, *Phonetica*, **52,** 1–40.

Pols, L. C. W. & Schouten, M. E. H. (1978) Identification of deleted consonants. *Journal of the Acoustical Society of America*, **64,** 1333–1337.

Pols, L. C. W. (1979) Coarticulation and the identification of initial and final plosives. In *ASA 50 speech communication papers* (J. Wolff & D. Klatt, editors), pp. 459–562. New York: Acoustical Society of America.

Recasens, D. (1983) Place cues for nasal consonants with special reference to Catalan, *Journal of the Acoustical Society of America*, **73,** 1346–1353.

Repp, B. H. (1986) Perception of the [m]-[n] distinction in CV syllables, *Journal of the Acoustical Society of America*, **79,** 1987–1999.

Repp, B. H. & Svastikula, K. (1988) Perception of the [m]-[n] distinction in VC syllables, *Journal of the Acoustical Society of America*, **83,** 237–247.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981) *Introduction to multidimensional scaling.* Orlando: Academic.

Schouten, M. E. H. & Pols, L. C. W. (1983) Perception of plosive consonants — the relative contributions of bursts and vocalic transitions. In *Sound structures*: *studies for Antonie Cohen* (M. P. R. van den Broecke, V. J. J. P. van Heuven & W. Zonneveld, editors), pp. 227–243. Dordrecht: Foris.

Sharf, D. J. & Hemeyer, T. (1972) Identification of consonant articulation from vowel formant transitions, *Journal of the Acoustical Society of America*, **51,** 652–658.

Singh, S. (1975) Distinctive features: a measure of consonant perception. In *Measurement procedures in speech, hearing and language* (S. Singh, editor). pp. 93–155. Baltimore: University Park Press.
Smits, R., Ten Bosch, L. & Collier, R. (1996) Evaluation of various sets of acoustical cues for the perception of prevocalic stop consonants: I, perception experiment, *Journal of the Acoustical Society of America*, **100,** 3852–3864.
Soli, S. D. & Arabie, P. (1979) Auditory versus phonetic accounts of observed confusions between consonant phonemes, *Journal of the Acoustical Society of America*, **66,** 46–59.
Takane, Y., Young, F. W. & de Leeuw, J. (1977) Non-metric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features, *Psychometrika*, **42,** 7–67.
Wang, M. D. & Bilger, R. C. (1973) Consonant confusions in noise: a study of perceptual features, *Journal of the Acoustical Society of America*, **54,** 1248–1266.
Whalen, D. H. (1981) Effects of vocalic formant transitions and vowel quality on the English [s]-[ʃ] boundary, *Journal of the Acoustical Society of America*, **69,** 275–282.

Appendix A: durations of consonantal closure

The duration of consonantal closure is defined as the interval between the closure and release landmarks. The average duration of consonantal closure across all original utterances was 143 ms. An ANOVA was carried out on the closure durations for all original utterances. The factors consonant, vowel, speaker, and stress were used in the analysis, as well as all possible interactions. All main effects were highly significant ($p < 0.001$). So were all interactions, except speaker × stress ($p = 0.2$), stress × vowel ($p = 0.4$), and speaker × stress × vowel ($p = 0.2$). Bonferroni *post-hoc* comparisons of means revealed the following:

1. The durations of the 17 consonants can be ordered into four non-overlapping groups: /d g b k t ð v p z n m ʒ/ (90, 94, 98, 103, 104, 108, 115, 116, 125, 131, 133, 138 ms, respectively), < /ŋ/ (183 ms) < /f θ/ (210, 212 ms, respectively), < /ʃ s/ (230, 243 ms, respectively). Note that the first group, with the shortest durations, contains all plosives, voiced fricatives and the nasals /m n/, while the last two groups contain the voiceless fricatives. Inspection of means for the two-way interactions showed that the nasal /ŋ/ being significantly longer than the other nasals is mainly caused by one of the speakers.
2. The ordering of consonantal duration according to vowel is: /ɑ/ (138 ms) < /i u/ (144, 147 ms, respectively).
3. The female speaker produced significantly longer closures (147 ms) than the male speaker (139 ms).
4. VCVs produced with final stress had significantly longer closure durations (157 ms) than those with initial stress (129 ms).

Appendix B: details of the MDS methodology

*B.*1. *Assumptions of non-metric individual differences MDS*

In non-metric individual differences MDS, a single multidimensional perceptual space is derived from a set of distance matrices using the assumptions that

1. The monotonic transformation of measured distances to scaled perceptual distances may differ per matrix.
2. Perceptual dimensions may be weighed differently for different matrices.

B.2. Conversion of frequencies to distances

The confusion matrices were converted into symmetrical distance matrices using the chi-squared distance measure D defined by

$$D = \sqrt{\sum_i \frac{(X_i - E(X_i))^2}{E(X_i)} + \sum_i \frac{(Y_i - E(Y_i))^2}{E(Y_i)}}$$

where X_i and Y_i represent the number of times stimuli X and Y have been assigned to response i, and $E(X_i)$ and $E(Y_i)$ represent the expected frequencies under the assumption of independence. In this case, $E(X_i) = E(Y_i) = \frac{1}{2}(X_i + Y_i)$, which leads to

$$D = \sqrt{\sum_i \frac{(X_i - Y_i)^2}{X_i + Y_i}}$$

B.3. Conditionality

In the ALSCAL program one can choose between several measures to be optimized, which is reflected in the *conditionality* option (Takane *et al.*, 1977). If comparisons between distances across matrices are meaningful, one should optimize the overall percentage of explained variance across all distance matrices ("unconditional"). If such comparisons are not meaningful, one should optimize the average percentage of explained variance per matrix ("matrix conditional"). In principle, comparisons between distances of different matrices are meaningful in this case. However, the absolute variances of distance matrices with low and high levels of confusion are generally quite different. As a result, the *unconditional* option caused the various matrices within one analysis to be weighed very differently, which produced results that were difficult to interpret. Therefore, the *matrix conditional* option was chosen for the analyses.