Institute of Phonetic Sciences
University of Amsterdam
Proceedings 18 (1994), 19-39

Classification and discrimination of single, complex, and interpolated speechlike stimuli*

Astrid van Wieringen and Louis C.W. Pols

Abstract

Perceptual processing of single and complex (multi-formant) CV-like and VC-like sounds, as well as interpolated natural speech-based syllables is examined in forced-choice classification and ABX discrimination tasks. The sounds have short and rapid plosive-like vocalic transitions, and are preceded or followed by an /a/-like or /u/-like stationary part. Twelve seven-item /b/-/d/ continua were created with interstimulus step sizes based on the difference limens in endpoint frequency (for the formant stimuli). It was expected that the processing strategy would depend on the stimulus complexity, i.e., that the number of discriminable or identifiable stimuli would decrease with increasing stimulus complexity. It is found that resolution varies with stimulus complexity and methodological paradigm, but that it is not controlled by speech specific properties. Perception of the single formant stimuli can approach the limits of the auditory system, because these stimuli are processed in an analytical listening mode. Complex stimuli are perceived less analytically, presumably because the additional formants partially mask the varying cue, and because the 'speechlike' character of the sound hinders within-category discrimination. Our results suggest that listeners process the stimuli in a temporary memory: internal representations for each of the stimuli of the continua under test are created, which enables them to distinguish the stimuli within the /b/ and /d/ categories. Although the interpolated speech-based sounds are perceived more categorically than the formant stimuli, the data give no clear evidence that these stimuli are processed by a long-term phoneme-labelling mechanism.

Introduction

The purpose of the present paper is to examine the perceptual resolution of single formant, complex formant and interpolated speech-based stimuli. It is generally ackowledged that performance is influenced by memory factors (e.g., Macmillan *et al.*, 1988; Schouten and Van Hessen, 1992) and it is, therefore, expected that the increased speechlikeness of the stimulus affects the discriminability and identifiability of sounds that are labelled similarly. In the present paper we first examine whether the stimuli of the seven-item single, complex, and interpolated (speech-based) continua

* Partly published in chapter 8 of the first author's Ph.D.-thesis "Perceiving dynamic speechlike sounds: psycho-acoustics and speech perception", that will be completed in April 1995.

can be *classified* consistently as either /b/ and /w/ or /d/ and /j/ in a one-interval 2-AFC task (in the absence of the release bursts and voice bars the vocalic transitions related to the plosives /b/ and /d/ also sound like the semivowels /w/ and /j/, because they have similar trajectories). Next, we examine *discriminability* of the different kinds of speechlike sounds. Comparison of the classification and discrimination data should give insight into the underlying processing strategies. It is expected that listeners cannot discriminate between those sounds that are retrieved from the same class of sounds from long-term memory; for these stimuli perception is based less on acoustical cues than on linguistic experience. However, the less complex the sound the less speechlike it sounds, and the less likely is the reference to long-term memory storage. It is expected that discrimination of these less speechlike stimuli is based more on acoustical cues.

Phoneme labels are evident response labels for the complex formant and interpolated speech-based stimuli, however not for the single formant stimuli; despite the speechlike features, subjects probably perceive the one-formant stimuli just as rising or falling transitions. We did use speech labels though for the single formant stimuli as well, as the subject would otherwise have to be trained to use various different response alternatives.

It is examined whether the classification scores depend on the position of the transition. We have found that listeners are significantly more sensitive to final transitions than to initial ones (Van Wieringen *et al.*, 1993, submitted), but this (sensory) difference may not appear in a more cognitive task. It is also examined how stimulus identity depends on the direction of the transition. For the /a/-like stimuli the transition direction changes halfway, creating a potential natural division between /b/-like and /d/-like stimuli (figures 1a and 1c). The change in transition direction cannot cue the /u/-like stimuli in a similar manner, as they already change after the first stimulus for the complex formant stimuli (figure 1d). The transitions of the single formant /u/-like stimuli all have the same direction per continuum (figure 1b). The subjects are required to use existing speech labels, such as /b/ and /d/ for both the single, the complex, and the interpolated speech-based CV-like and VC-like stimuli. The classification task as used here does not measure how well the listener can label a sound, but how consistent the subject is at categorizing the different sounds. The responses are not right or wrong.

The 30-ms single and complex formant syllables, and the interpolated /b/-/d/-like stimuli were classified in a one-interval 2-AFC task by 15 subjects (most of whom were inexperienced).

Stimulus generation

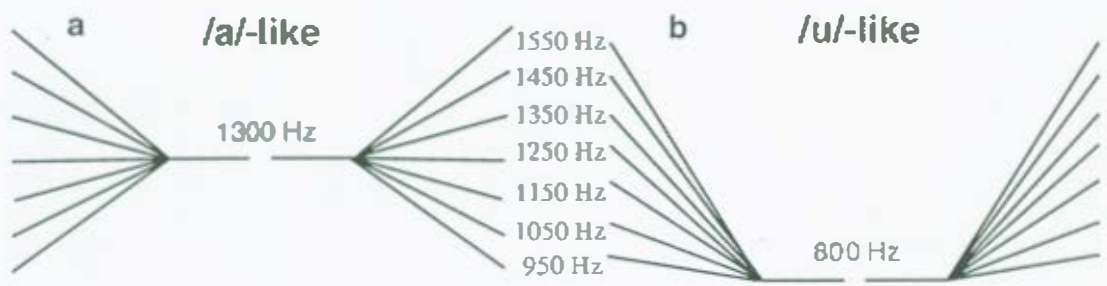
General parameters

The single and complex formant transitions are abstract representations of vocalic transitions of stop consonants in speech stimuli. They vary in offset or onset (hence: endpoint) frequency at a fixed 30-ms transition duration, and are either preceded or followed by an 80-ms /a/-like or /u/-like formant pattern. Together with the interpolated natural speech-based syllables six seven-syllable continua, varying along the bilabial-to-alveolar dimension, were generated for each of the two formant patterns. In the following section the (generation of) stimuli will be described in more detail for each condition separately.

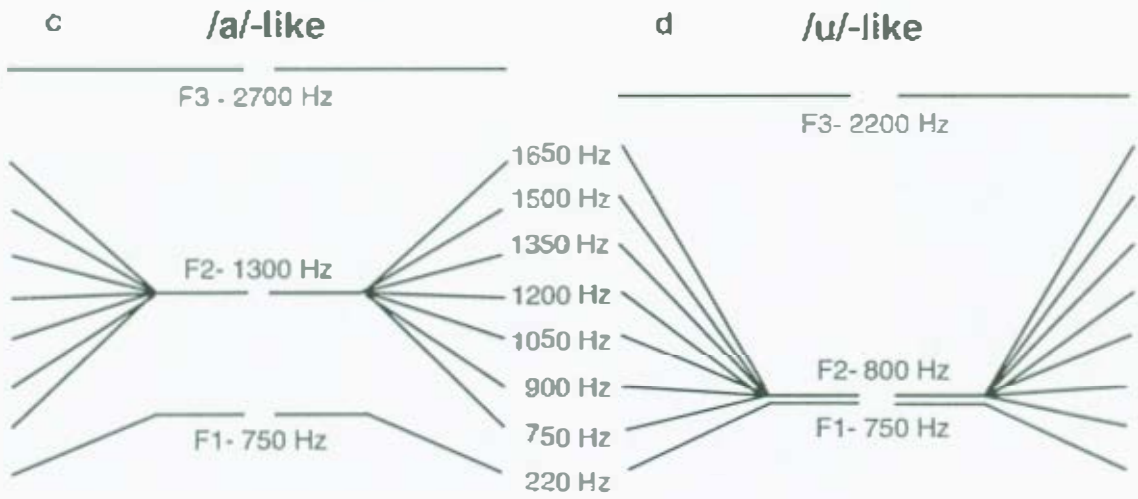
Formant synthesis

The single and complex CV-like and VC-like syllables were generated by a digital formant synthesiser (Weenink, 1988). A 110-Hz pulse was used as glottal source. To ensure a precise generation of these formant transitions the stimuli were sampled at 1.2 MHz. After low pass filtering, they were downsampled to 20 kHz (16 bit resolution). The formant frequency values were updated every 1 ms. Although the first period of the stimulus always started on a zero crossing, all stimuli were preceded and followed by a 2-ms cosine window to avoid clicks. The formant bandwidth was proportional to the changing formant frequency (10%). The stimuli were generated real-time by means of an OROS-AU22 DSP board with D/A converter.

The *single* formant syllables had 30-ms transitions, preceded or followed by 80-ms stationary portions (figures 1a and 1b). The first-formant and second-formant transitions of the *complex* stimuli were also 30 ms, preceded or followed by an 80-ms steady-state. A stationary third formant, and a 20-ms voice bar were added to make the stimuli sound more speechlike (figures 1c and 1d). The fixed F1-transitions of the complex syllables rose or fell from 220 Hz to 750 Hz and the F3 was fixed at 2700 Hz for the /a/-like stimuli, whereas the F1 changed from 220 Hz to 330 Hz (and v.v.) for the /u/-like stimuli. The F3 of the /u/-like stimuli was fixed at 2200 Hz. The single formant transitions as well as the second-formant transitions of the complex stimuli varied in endpoint frequency at a fixed transition duration. The step size in endpoint frequency, either 100 Hz (single) or 150 Hz (complex), was the average difference limen in frequency for the single and complex formant stimuli (Van Wieringen and Pols, submitted). Although the difference limens as found in that earlier study depend strongly on the position of the transition and on its frequency extent, a fixed step size was chosen per stimulus type to limit the number of variables in the design. The transitions of the single syllables varied in endpoint frequency from 950 Hz to 1550 Hz in steps of 100 Hz each (figures 1a and 1b) and those of the complex syllables varied from 750 Hz to 1650 Hz in 150 Hz steps each (figures 1c and 1d). The transitions preceded or followed a steady-state with either an /a/-like or /u/-like formant pattern. For the C/a/-like and /a/C-like stimuli the steady-state frequency was 1300 Hz, for the C/u/-like and /u/C-like stimuli it was 800 Hz. In total, eight seven-item continua were generated by varying the endpoint frequencies of the transitions in initial and final position.



Figures 1 a/b. Schematic representation of the /a/-like (a) and /u/-like (b) *single* formant stimuli in initial and final position. The inter-stimulus step size is 100 Hz. The transitions are 30 ms, and the steady-states 80 ms (not to scale).



Figures 1c/d. Schematic representation of the /a/-like (c) and /u/-like (d) *complex* formant stimuli in initial and final position. The inter-stimulus step size is 150 Hz. The transitions are 30 ms, and the steady-states 80 ms (not to scale). The transitions of the complex formant stimuli were preceded or followed by a 20-ms voice bar in initial and final position, respectively (not shown).

Interpolated speech-based stimuli

To proceed to more speechlike conditions, we examined, in addition to the formant stimuli, identification of speech stimuli. The speech-based stimuli were created by interpolating the spectral envelope (Van Hessen, 1991) of two natural endpoints, e.g., /ba/ and /da/ (figure 2). The original /ba/, /da/, /ab/, /ad/, /bu/, /du/, /ub/, and /ud/ stimuli were segmented from CVC tokens pronounced by a native Dutch male speaker (F0 of about 110 Hz). Stimuli were digitised with a sample frequency of 20 kHz (cut-off frequency of the low-pass filter was 4.9 kHz; slope 96 dB/oct). All syllables were segmented to be 100 ms.

Four continua (/ba/-/da/, /ab/-/ad/, /bu/-/du/, and /ub/-/ud/) were created by interpolating the spectral envelope of two stimuli. This was done as follows: The two signals are divided into 25.6-ms frames, which are each multiplied by a hamming window (the frame is shifted forward by a quarter of the window). For each frame the peaks in the spectrum are estimated by means of a fourier transformation (fifty peaks for a spectrum ranging between 0-5000 Hz). Next, the amplitudes, frequencies, and the phases of the peaks of adjacent frames (of the two stimuli) are interpolated in a linear manner. The interpolated parts of the two stimuli are marked by placing the cursor at the beginning of the voice bar and at the end of the first period for the transition in initial position (or at the beginning of the last period and at the end of the voice bar for the transitions in final position). Only the portions between the lines, approximately 35 ms, were interpolated, although the entire signal is divided into frames, multiplied by a hamming window, and reconstructed with the modified spectral envelopes. In this way stimulus quality remains equal over the whole signal. In all the continua, the syllables were interpolated in the direction from /b/ to /d/, so that the interpolated stimuli contained the source characteristics of the /b/. In initial and final position the original syllables clearly sounded /b/-like or /d/-like. Contrary to the CV-syllables, VC-syllables contained no burst, only a short 20-ms voice bar (because the burst usually follows after a 60-ms voice bar).

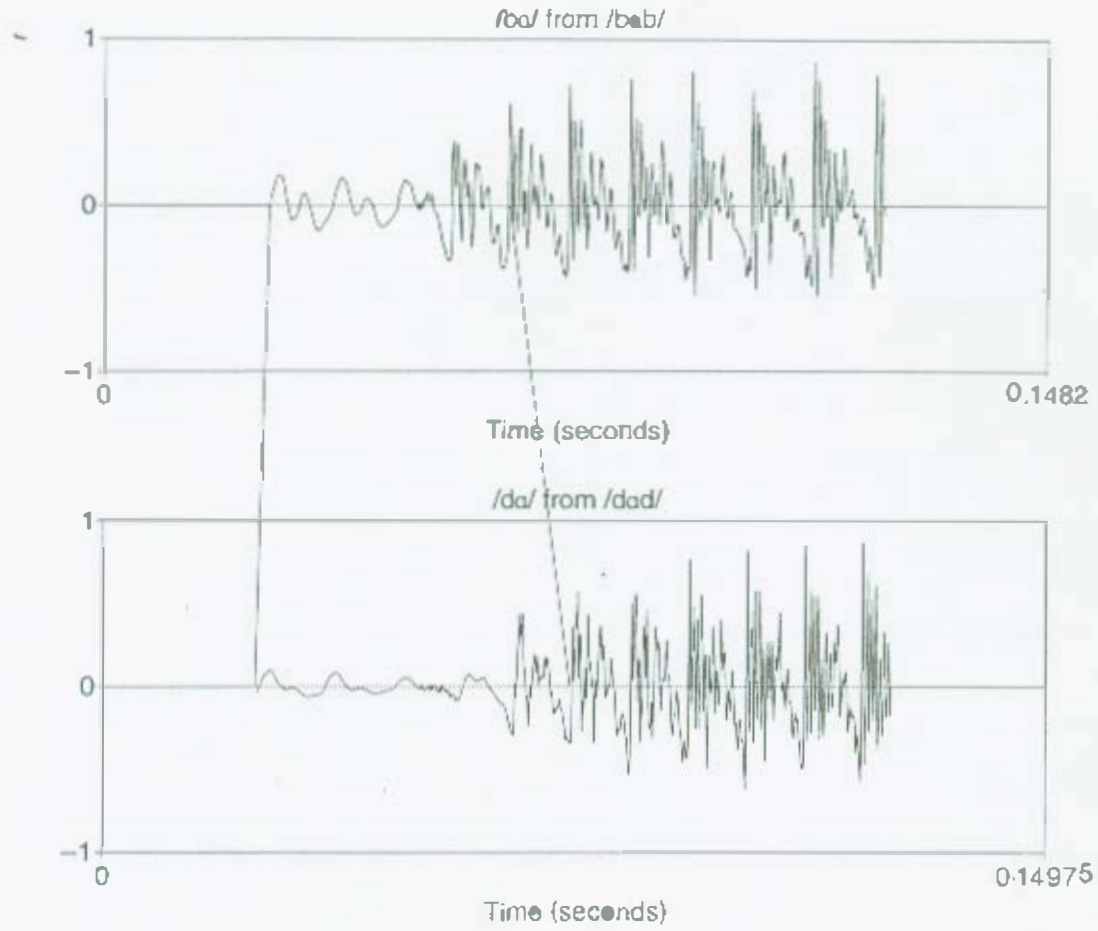


Figure 2. Example of two original stimuli, /ba/ and /da/, from which a stimulus continuum is created. The part between the lines, i.e., between the beginning of the voice bar and the end of the first period, is interpolated.

Procedure and subjects

Fifteen normal hearing subjects classified the single and complex transitions preceding or following the /a/-like and /u/-like formant patterns in four separate sessions (vowel timbres were not mixed). The subjects were untrained: training was not required since the physical cues measured in the classification task do not approach the borders of the auditory system. Table I lists the subsets under test. The first two sets contained either the /a/-like or the /u/-like single and complex formant syllables. Both sets contained six subsets of randomised continua, each with nine repetitions of each of the seven stimuli. The first two subsets consisted of single formant transitions in initial and final position, respectively, followed by the complex formant syllables in the third and fourth subsets. The single formant transitions were repeated in the last two subsets, as they hardly sounded as phonemes, but rather as chirps (subsets 5 & 6). Subjects were told that the single syllables would be difficult to perceive as phonemes, but that they had to listen whether the stimulus contained more /b/-like or /d/-like properties. The first two subsets (subsets 1 & 2) were not taken into account.

The test lasted approximately 25 minutes. All subjects remarked that the complex syllables were easier to classify than the single syllables and that the single and

complex formant transitions were easier to classify in final, rather than in initial position. Also, most subjects found the /u/-like transitions more difficult to classify than the /a/-like ones, presumably because transition direction could not be used as a perceptual cue. Moreover, single-formant /u/-like transitions were not perceived as /u/-like, but rather as /a/-like by many listeners (we learned this from the subjects after the test, since the actual vowel did not have to be classified). This is not so strange, as the second formant of an /u/-like stimulus approximates the first formant of an /a/-like stimulus.

Finally, the interpolated /a/-like and /u/-like speech stimuli were classified on two separate occasions (subsets 7 & 8). Subjects generally found the interpolated CV syllables easier to classify than the VC ones, presumably because the former contained a release burst. All of them also found the natural /a/-like syllables easier to classify than the /u/-like ones.

Table 1. Different subsets in the classification task. The vowel timbres were held separately, as well as the position of the transition for the single and complex formant stimuli. The single formant stimuli were classified twice; however, subsets 1 and 2 were disregarded for analysis. Each stimulus was repeated nine times for each of the 15 subjects.

Set		Subsets							
		1	2	3	4	5	6	7	8
1	initial/final /a/-like	single	single	complex	complex	single	single		
2	initial/final /u/-like	single	single	complex	complex	single	single		
3	initial/final /a/-like							interpolate	interpolated
4	initial/final /u/-like							interpolated	interpolated

All sets of randomised continua were classified in an interactive, self-paced, procedure. Subjects were seated in front of a terminal on which two alternatives were displayed, one square indicating 'b' or 'w' and the other indicating 'd' or 'j'. Although the stimuli were mostly perceived as plosives (as a result of their short durations) the semivowels were added to the response sets, because they are also likely responses for those stimuli that lack the characteristic voice bar or release burst. By pressing the rightmost mouse key, subjects heard the stimulus twice, they then could give their response by clicking the left mouse key on one of the two response squares. All stimuli, including the single formant transitions, were classified as 'b'-'w' or 'd'-'j'. At the beginning of each subset, subjects were informed about the position of the transition (either preceding or following the steady-state). Ten practice trials preceded every set of test trials to familiarise the subject with the stimuli.

Results

On average, each of the seven single, complex, and interpolated stimuli is classified consistently as /b/ or /d/. Even the one-formant stimuli contain sufficient information to allow subjects to group the stimuli into two phoneme classes. A fully factorial analysis of variance with formant patterns (/a/ & /u/), stimulus type (single formant, complex & speech-based), position (initial and final) and stimulus number as fixed factors was performed on the percentage /b/-responses (which are complementary to the /d/-scores). Statistically, there were no significant effects or interactions. The transition direction in the signal changes for the /a/-like stimuli between stimuli four and five, and for the complex /u/-like stimuli between stimuli one and two (see figure 1). The effect of transition direction, however, is not reflected in the response curves.

The classification functions are discussed below separately for the single, complex and speech-based stimuli. They are based on 135 responses per stimulus (15 subjects x 9 responses per stimulus per subject).

Results *single* transitions in initial and final position

The average classification functions of the single /a/-like and /u/-like syllables in initial and final position are shown in figure 3a. The percentage of /b/ or /d/ responses is plotted on the y-axis as a function of the seven different stimuli.

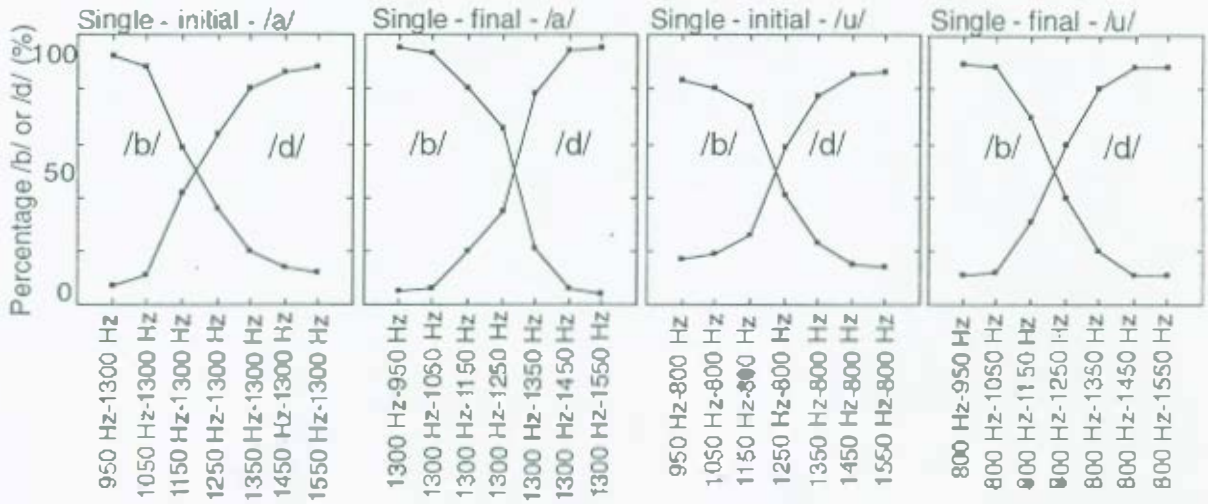


Figure 3a. Classification functions of the /a/-like and /u/-like single formant transitions in initial and final position. The data, which are based on 135 responses per point, are plotted in terms of the percentage of /b/ or /d/ responses as a function of the stimuli.

Although the single formant transitions are difficult to perceive as speech sounds, most of them are classified consistently as /b/ or /d/. Generally speaking, listeners reported that the CV-like syllables were more difficult to classify than the VC-like ones, and that the /a/-like ones were easier to classify than the /u/-like ones in the present condition (statistically not significant). Although the subjects reported that they were able to differentiate the /u/-like stimuli, but that they had to guess the choice itself (/b/ or /d/), none of the classification functions seem to be reverted. Despite individual differences, the average functions are comparable with respect to the cross-over point for the single /a/-like and /u/-like syllables. Transition direction is not a necessary cue to categorise single-formant stimuli into phonetic classes: the direction of the transition does not change at the cross-over points in our single formant continua (figure 1a/b)

Results *complex* syllables in initial and final position

The mean identification functions of the complex /a/-like and /u/-like syllables in initial and final position are shown in figure 3b. It is suggested that the endpoint of the transition is an important cue for plosive classification, as the endpoint stimuli of the complex /a/-like stimuli are identified 100% of the time as /b/ or /d/ (an endpoint between 700 Hz and 900 Hz is a good cue for /b/ and between 1500 Hz and 1650 Hz a good one for /d/). The frequency extent, i.e., the difference between the initial and

final frequency, may be perceptually important too, apart from the endpoint of the transition.

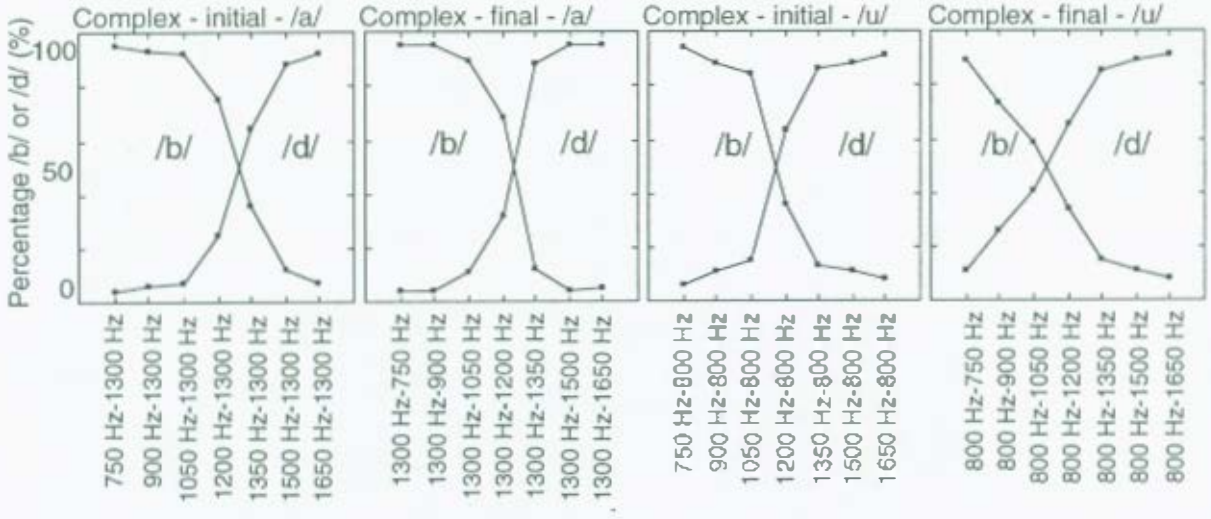


Figure 3b. Classification functions of the /a/-like and /u/-like complex formant transitions in initial and final position. The data, which are based on 135 responses per point, are plotted in terms of the percentage of /b/ or /d/ responses as a function of the stimuli.

Transition direction is probably not a very important perceptual cue because the subjects' responses for the /a/-like as well as for the /u/-like stimuli change abruptly between the 4th and 5th stimuli, whereas only the /a/-like stimuli, not the /u/-like ones, change direction at that point. The individual plots of the complex /u/-like syllables display slightly more variability than those of the /a/-like stimuli (not shown): two subjects identified all the complex formant /u/-like stimuli of one continuum as belonging to one category (one in initial position, the other in final position). It is not clear what causes the discrepancy (i.e., the higher uncertainty in classifying the /u/-like stimuli), as the variation does not seem to be related to a particular bias (such as /b/ before /u/). Moreover, most of the subjects are able to classify the syllables consistently.

Results *interpolated* speech-based syllables in initial and final position

The average classification functions of the interpolated /ba/-/da/-like, /bu/-/du/-like, /ab/-/ad/-like, and /ub/-/ud/-like stimuli in initial and final position are plotted in figure 3c. The interpolated speech-based stimuli are just numbered on the abscissa from 1-7 because the exact acoustical parameters are too numerous to specify. The /a/-like plots show that none of the listeners have any trouble in partitioning the responses into the two phonetic categories in initial or final position. Although the release burst was absent in final position, a bilabial or alveolar transition was clearly perceived. The same subjects performed poorer on the speech-based /u/-like stimuli, especially in initial position. It was expected that the subjects would be biased towards responding /b/ before a rounded vowel. However, the individual figures show that alveolar and bilabial cues were heard equally often (not shown here).

Both the CV and the VC transitions of the interpolated stimuli contain sufficient information to consistently classify the syllables as /b/ or /d/.

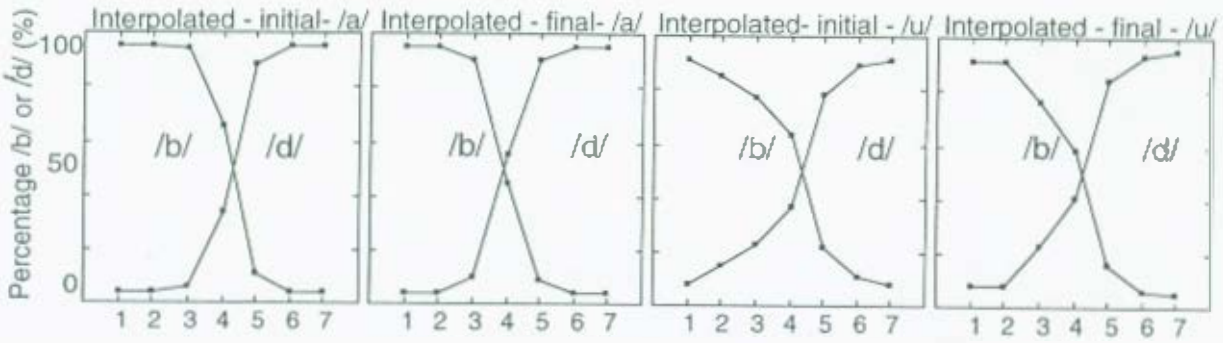


Figure 3c. Classification functions of the /a/-like and /u/-like interpolated speech-based formant transitions in initial and final position. The data, which are based on 135 responses per point, are plotted in terms of the percentage of /b/ or /d/ responses as a function of the stimuli.

Comparing physical and perceptual spacing

In the classification experiment listeners respond with either 'b' or 'd' by comparing the stimulus with an (internal) criterion. These stimuli, at least the formant-based ones, are physically equally spaced on the continua. It is of interest to know how the stimuli are spaced perceptually, i.e., how sensitive listeners are to adjacent stimuli and how sensitivity compares for the different kinds of stimuli. The perceptual spacing between adjacent stimuli is estimated by computing the cumulative *d'* of the classification responses of the seven different stimuli per continuum. The cumulative *d'* is the sensitivity distance between any stimulus and stimulus one (Macmillan and Creelman, 1991). The total sensitivity, i.e., the sum of these distances, indicates the listener's performance over the entire stimulus set.

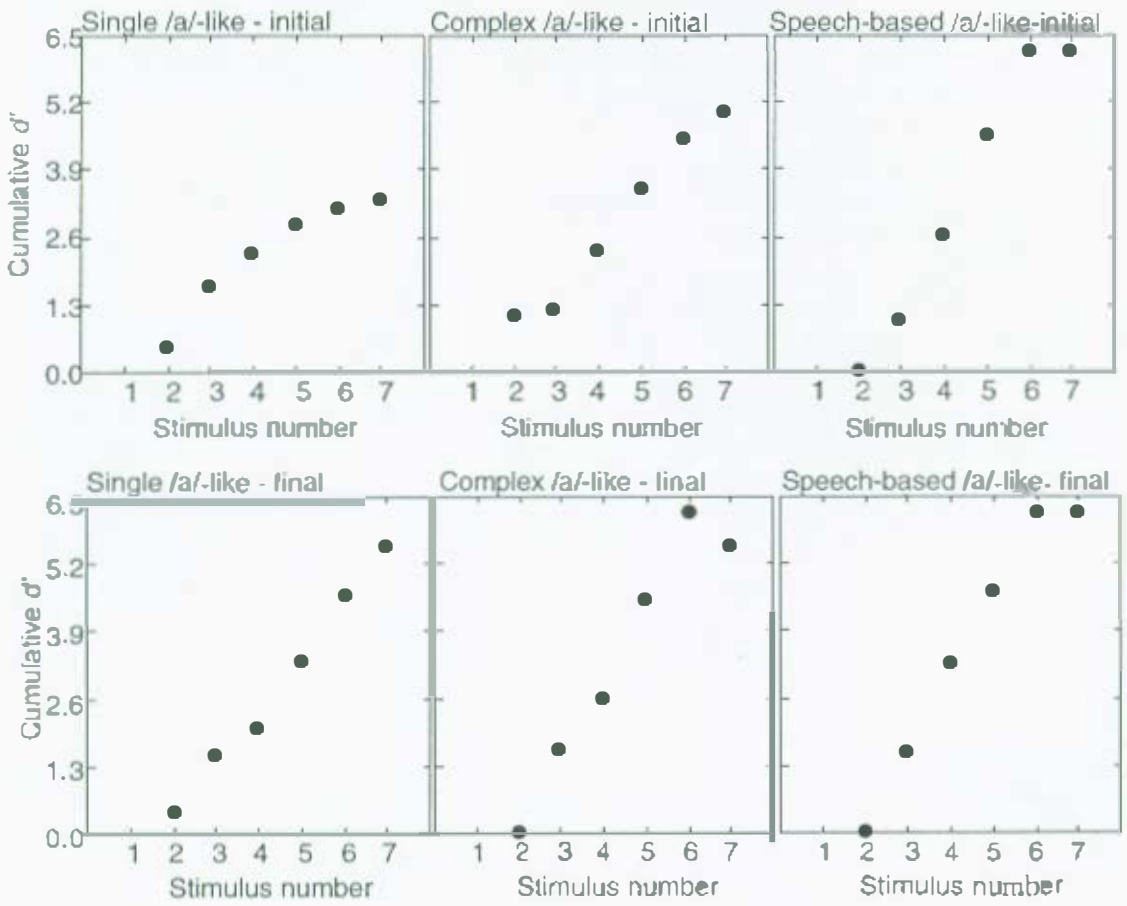
Cumulative *d'* is computed by transforming the percent correct responses ('b' or 'd') into z-scores (for computation see Appendix B in Van Wieringen, 1995). *d'* is the difference between z-scores, and cumulative *d'* is the perceptual distance between a certain stimulus and stimulus one. Cumulative *d'* of the last stimulus is equal to total *d'*. Figure 4 presents the perceptual spacing between the different kinds of /a/-like and /u/-like stimuli in initial and final position. Data are calculated from the average classification functions. The plots show that many of the stimuli are not equally spaced perceptually. Compare the perceptual spacing of the single, complex, and interpolated /a/-like stimuli in initial position. The perceptual spacing for the single transitions is largest between stimuli two and three, while the biggest step is between stimuli three and four, and four and five for the complex and interpolated stimuli. The boundary between the two stimuli, i.e., the largest spacing, also differs for the two formant patterns: for instance, the complex /u/-like stimuli in initial position show a relatively large perceptual spacing between stimuli three and four, while the complex /u/-like stimuli in final position show the larger space between stimuli four and five.

Total sensitivity also varies per condition. Total *d'* indicates how sensitive listeners are with respect to the entire stimulus continuum. Figure 4 shows that total sensitivity to the /a/-like stimuli in initial position increases as the stimuli become more complex. This is not the case in final position, where the stimuli of the continuum are spaced further apart perceptually. The total *d'* of the single, complex, and interpolated stimuli is high (the maximum of 6.5 is made to correspond to 100% /b/ or /d/ responses). Listeners appear to be less sensitive to stimulus changes of the different /u/-like

stimuli than to changes of the different /a/-like ones, because the perceptual spacing between the /a/-like stimuli is larger than between the /u/-like ones. The relatively low total *d'* of the single, complex, and interpolated /u/-like stimuli confirms the remarks made by the subjects that the /u/-like stimuli were more difficult to classify than the /a/-like ones.

Discrimination experiments clearly showed that listeners are more sensitive to changes in final transitions than in initial ones (Van Wieringen *et al.*, 1993, submitted). It was expected that the present classification experiments would not yield such a strong initial-final effect, as the acoustically different stimuli had to be categorised into two predetermined categories. In such a task the difference in sensitivity does not seem to be relevant anymore.

The sigmoids of figures 3a, 3b, and 3c show that listeners can classify the stimuli consistently, despite some individual differences. They do not show whether listeners are more capable of classifying the stimuli in final position, rather than in initial position. Figure 4, however, does show that sensitivity to the single and complex /a/-like stimuli is greater in final than in initial position, suggesting a perceptual difference between CV-like and VC-like syllables. This does not seem to be the case for the /u/-like syllables, as the perceptual spacings and the total d's of both the CV-like and VC-like stimuli appear to be equally large. As the CV and VC speech-based stimuli are perceptually similar, the initial-final effect probably results from the subject's ability to listen analytically to the acoustical changes of the /a/-like formant transitions in final position.



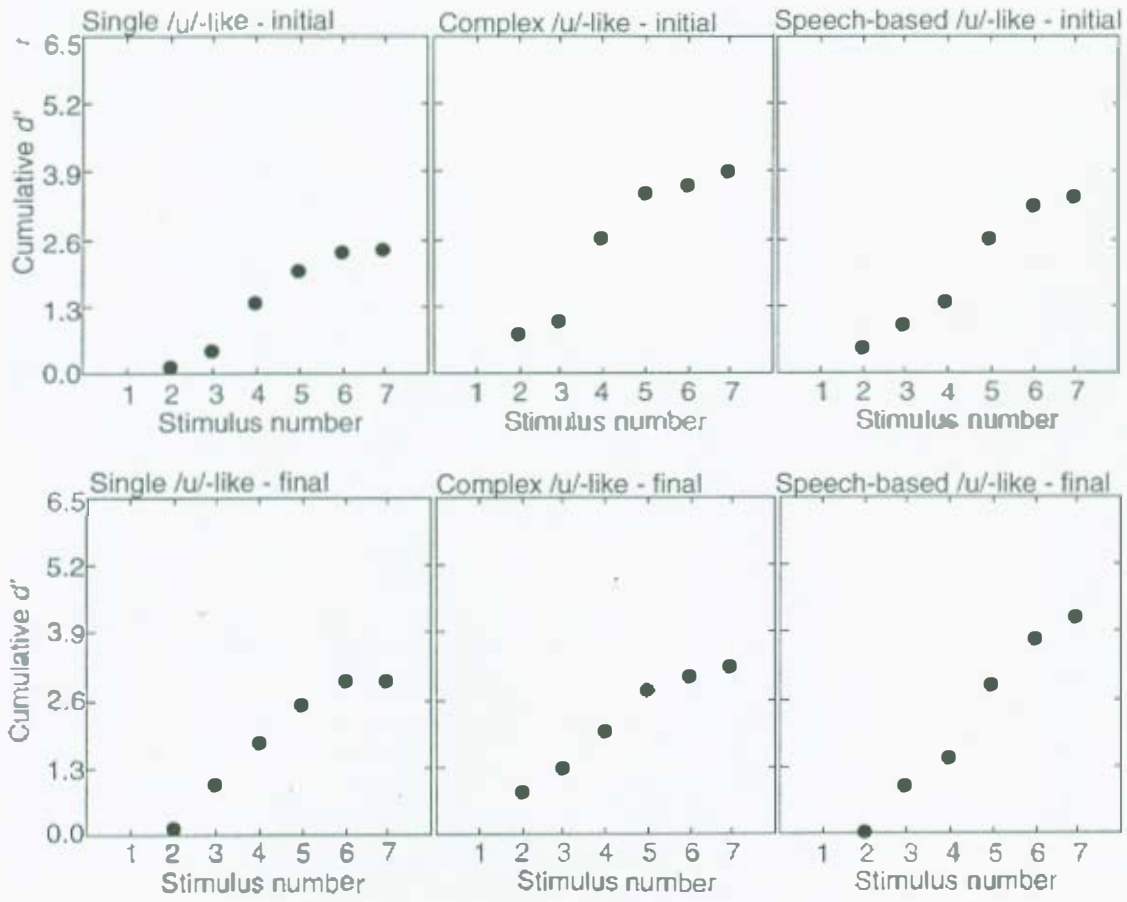


Figure 4. Cumulative *d'* as a function of the seven stimuli, for the single, complex, and speech-based stimuli separately. Although the (formant) stimuli are equally spaced physically, they are not always equally spaced perceptually. The slope and the total *d'* tells us how rapidly the perceptual effect grows with stimulus value, i.e., how sensitive the listener is to stimulus changes.

The classification experiments show that the single, complex, and the interpolated speech-based stimuli can be classified consistently as /b/ or /d/, even the single formant transitions, which would probably not have been labelled as 'b' or 'd' by choice. In our experiments the classification functions of the single stimuli are similar to those of the complex ones, in that the number of correct /b/ or /d/ responses is largest at the endpoints and approaches chance level in between. However, although speech labels are used in both stimulus conditions, it is not clear whether the single stimuli (and the complex stimuli) are perceived categorically in a discrimination task. Although just noticeable differences in endpoint frequency had been calculated to determine the physical spacing between the seven-item formant continua, *discrimination* data were also collected of the single, complex, and speech-based stimuli in an ABX-paradigm.

Predicted discrimination of single, complex, and interpolated speech-based stimuli

Categorical perception of speech stimuli implies that reference is made in memory to existing speech labels, and that, therefore, discrimination of physically equally spaced stimuli is at chance level for those stimuli that are labelled similarly and is higher than chance for those stimuli that are labelled into different categories (see figure 5). It is expected that the complex and speech-based stimuli, which sound much more like natural speech sounds, are perceived categorically while the single formant transitions are perceived analytically.

Discrimination data of the speech-based stimuli are also of interest from a sensory point of view: the discriminability of these stimuli is not known, as the physical properties of these stimuli cannot be controlled as systematically as in synthetic stimuli. As the physical spacing of the interpolated speech-based continua is not based on difference limens in frequency the perceptual spacing of these continua may differ from those of the formant stimuli.

Figure 5 illustrates with data from our classification experiments how classification and ABX discrimination are related to each other. In the classification experiments subjects listen to one stimulus at a time, after which they label them with one of the prescribed alternatives. In the ABX discrimination task two stimuli (A and B) are offered and then a third one (X), which is either A or B. The subject has to indicate whether X equals A or B. The measure of discriminability is the percentage of the time that X is correctly 'identified' as A or B. In figure 5 the dashed discrimination function is predicted from classification data of one of our subjects (for computation see Appendix E in Van Wieringen, 1995), while the measured discrimination function (solid line) illustrates the actual discriminability of that subject. Under the hypothesis of categorical perception discrimination is at chance level within the categories and higher than chance level at the cross-over boundaries. Actual discrimination is better than predicted and the cross-over boundary (peak) is shifted from stimuli 4-5 to stimuli 3-4.

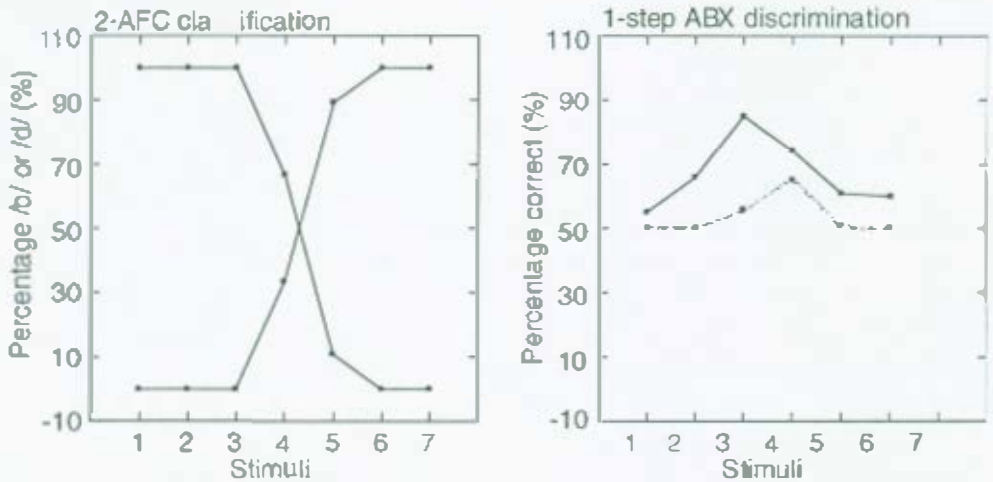


Figure 5. The plots illustrate the relation between the classification and discrimination data (of an individual subject in our study). The dashed discrimination function in the right-hand plot is predicted from the classification sigmoid in the left-hand plot (Pollack and Pisoni, 1971). Under the hypothesis of categorical perception stimuli are discriminated at chance level when they are classified similarly, and they are well discriminable when they are classified into different classes. The solid discrimination function illustrates the experimental data of the same subject: stimuli are discriminated higher than chance level within the categories and the measured boundary between the two classes differs from the one predicted from the classification data.

Measured discrimination of single, complex, and interpolated speech-based stimuli

It is generally acknowledged that *actual* discrimination is better than *predicted* discrimination: within-category performance is higher than chance level and across-category performance is higher than predicted (for an overview see Repp, 1984). Listeners use acoustical cues, both within and across phoneme boundaries, to differentiate between the stimuli.

To examine the relation between classification and discrimination in more detail, discrimination data of trained subjects were collected by means of the ABX paradigm. The discrimination test was first performed with the 15 naive subjects of the classification tests (to have a one-to-one relationship between classification and discrimination). However, results were unreliable, as there were too few data points per subject and we were not able to collect more data from these naive subjects. To test whether discrimination functions vary for the different kinds of stimuli, five subjects were trained for a short period of time, and the test was repeated.

The physical step size between the formant stimuli is appropriately chosen to be somewhat larger than the 'true' difference limens in frequency. From a psycho-acoustical point of view listeners should be equally sensitive to acoustical differences in the single or the complex formant continua if the step size is similar in a relative sense. This should also be the case in the ABX discrimination task, although the difference limens are somewhat poorer than in a same/different paradigm, which hardly requires memory strain (e.g., Saslow, 1967). However, a different pattern of results is expected if cognitive processes dominate sensory ones: it is then extremely difficult to apply an analytical listening strategy and to differentiate between the different stimuli of the continuum. Discrimination of the interpolated speech-based stimuli may be based on a long-term phoneme memory mechanism, as the natural quality of the sound is probably important for categorical perception (see also Schouten and Van Hessen, 1992).

The issue now is whether discrimination of single, complex and interpolated stimuli is based on sensory differences or on a phoneme labelling mechanism. These experiments are also of interest with respect to the discriminability of the interpolated speech-based stimuli. As the exact physical properties of these stimuli are unknown, difference limens in frequency cannot be determined.

Stimuli and procedure (ABX)

The stimuli were the same as the ones tested in the 1-interval 2-AFC experiments, i.e., the 30-ms /a/-like and /u/-like transitions with steady-state of the single, complex, and speech-based stimuli in initial and final position. Five subjects, who had also participated in the classification experiment, were tested individually in a quiet room. Three subjects listened to the /u/-like stimuli, three to the /a/-like (one of the six subjects listened to both formant patterns). Following the previous experiments they were seated in front of a terminal and heard three stimuli over Sennheiser headphones at a comfortable level. The inter-stimulus time between the three stimuli was 500 ms. By clicking the appropriate response square on the monitor, they could indicate whether they considered the third stimulus to be identical to the first or to the second, after which three new stimuli were generated. No feedback was given during the test.

There were six different conditions for each subject (single, complex, and speech-based stimuli in initial and final position). After a short training period, each of the four combinations per stimulus pair (ABA, ABB, BAA, BAB) was repeated 25 times,

resulting in 100 observations per stimulus pair per subject. All conditions were tested separately. The tests, which were preceded by ten test triads, lasted approximately 10 minutes. Subjects were paid for participating.

Results: ABX discrimination functions

Figure 6 illustrates the 1-step ABX-discrimination functions and the classification sigmoids, averaged over the six subjects and two formant patterns /a/ and /u/, together with the predicted discrimination function (based on the average classification sigmoids of these six subjects). The discrimination results are plotted in terms of percentage correct as a function of one pair of stimuli (one pair is averaged over ABA, ABB, BAA, and BAB). The two most striking results are 1) that the predicted and measured functions differ markedly and 2) that categorisation, if any, depends on stimulus complexity and on the position of the transition. In general, subjects are more able to discriminate between subsequent pairs of stimuli than predicted from the classification sigmoids. Compared to the predicted functions, the experimental ones yield higher percentage correct scores.

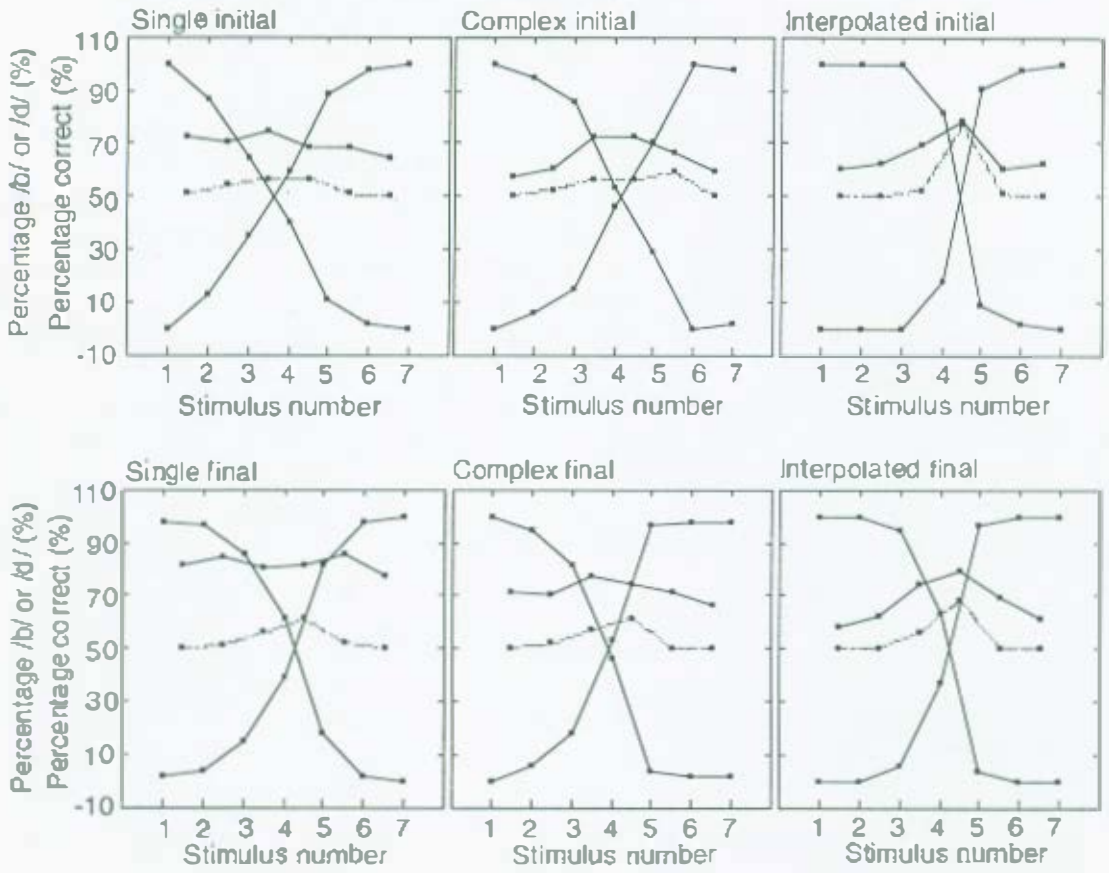


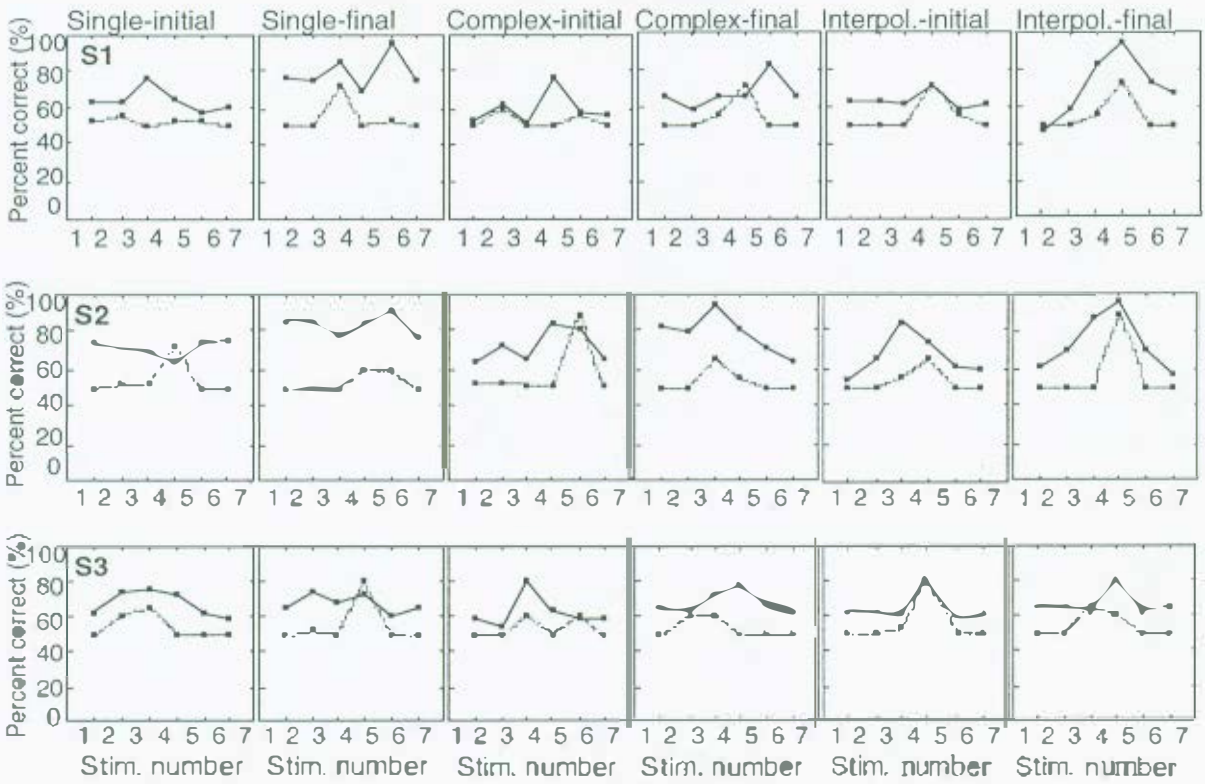
Figure 6. Classification sigmoids, predicted discrimination (dashed) and measured discrimination functions (solid) for the single, complex, and interpolated stimuli in initial and final position. Data are averaged over the six trained listeners and the two formant patterns (/a/&/u/). The stimuli are indicated on the abscissa (the discrimination data apply to pairs of stimuli).

Even when they show a clear perceptual boundary, performance is never at chance for those stimuli that are classified to the same category, and the across-category peak is higher than predicted. The more complex the stimulus, the more the perceptual

continuum seems to be divided into two categories. However, categorisation does not only depend on the stimulus complexity, but also on the position of the transition.

Compared to the single formant stimuli, more prominent peaks occur with the complex formant stimuli in initial position. In final position most subjects fail to hear two different percepts: the experimental response functions are flauer than in initial position. The present ABX ones show that listeners are capable of applying an analytical listening strategy to most of the stimuli. The more the stimuli sound like phonemes (or possibly any other kind of long-term prototypes), the more difficult it is to perceive them analytically. Our data show this to be the case for the complex formant transitions in initial position and for the interpolated speech-based stimuli in initial and final position.

The 1-step ABX-discrimination functions of the six subjects are plotted separately for each of the six stimulus conditions in figure 7. Each plot illustrates the individual discrimination performance, together with the subject's predicted discrimination function. S1 to S3 discriminated the /a/-like stimuli and S4 to S6 the /u/-like ones. These plots also show that the measured discrimination functions of the formant stimuli rarely show clear peaks indicating phoneme boundaries. Although categories seem to emerge with increasing complexity of the stimulus, it must be kept in mind that the physical spacing between the seven speech-based stimuli is not determined by difference limens, so that within-category discrimination may be auditorily impossible. Some of the discrimination functions of the single stimuli yield very high percentage correct responses (for instance, S5 and S6 in final position). These listeners were capable of applying (and maintaining) an analytical listening strategy: their performance approached the (difference) threshold level determined in the same/different AX paradigm (Van Wieringen and Pols, submitted). Recall that the physical spacing between the stimuli was slightly larger than the difference limens in frequency (100 Hz for the single stimuli and 150 Hz for the complex stimuli).



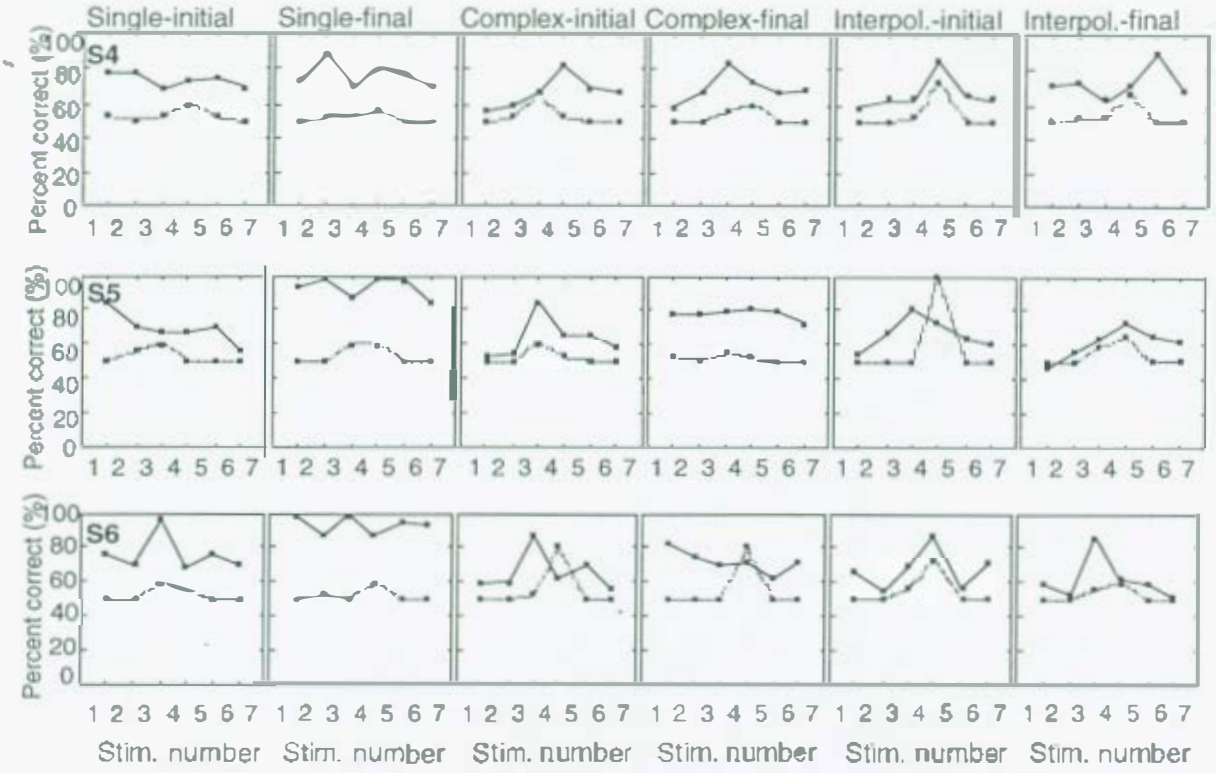


Figure 7. Discrimination functions of the single, complex, and speech-based stimuli in initial and final positions. Each plot illustrates the actual performance (solid) of one subject together with its predicted (dashed) function for one stimulus condition. S1 to S3 discriminated the /a/-like stimuli and S4 to S6 the /u/-like ones.

Comparison predicted and measured discrimination functions

A fully-factorial ANOVA was performed on the percentage correct discrimination scores of the experimental and predicted conditions. The fixed factors were predicted/measured, formant patterns (/a/ & /u/), position of the transition (initial & final), stimulus type (single, complex, interpolated speech-based), and stimulus number (6). The statistical analyses showed four significant main effects, namely predicted/measured [$F(1, 288) = 504.2$, $p < 0.001$], position of the transition [$F(1, 288) = 21.9$, $p < 0.001$], stimulus type [$F(2, 288) = 12.0$, $p < 0.001$], and stimulus number [$F(5, 288) = 26.4$, $p < 0.001$]. The following first-order interactions were significant: predicted/measured x transition position [$F(1, 288) = 26.9$, $p < 0.001$], predicted/measured x stimulus type [$F(2, 288) = 19.5$, $p < 0.001$], and stimulus type x stimulus number [$F(10, 288) = 5.5$, $p < 0.001$]. There were no significant higher-order interactions. The statistical analyses confirm our findings that performance depends on the complexity of the stimulus and the position of the transitions, and that discriminability is based on more cues than predicted from the assumption of 'absolute categorisation'. The data are comparable for the /a/-like and /u/-like formant patterns (statistically n.s.): performance does not depend on the different transition directions.

As a result of the large number of factors, statistical analyses were also conducted on the predicted and measured discrimination functions separately.

In the predicted condition, stimulus number (6) was the only significant main effect [$F(5, 48) = 327.4$, $p < 0.001$]. Neither transition position nor stimulus type were statistically significant.

In the measured condition, the three different stimulus types were analysed separately. The single formant transitions were statistically significant with regard to transition position (initial-final) [$F(1, 48) = 29.7$, $p < 0.001$], and formant pattern [$F(1, 48) = 11.4$, $p < 0.001$]. Statistical analyses of the complex formant stimuli showed transition position [$F(1, 48) = 11.5$, $p < 0.001$] and stimulus number [$F(5, 48) = 3.6$, $p < 0.05$] to be statistically significant, while those of the interpolated speech-based stimuli showed only stimulus number [$F(5, 48) = 9.6$, $p < 0.001$], not transition position to be statistically significant. There were no significant first-order interactions. In the next section we will discuss the cues underlying perception of single, complex and interpolated stimuli.

The measured discrimination data clearly yield better performance than predicted under the hypothesis of categorical perception: *within* categories, performance is higher than chance level, and *across* categories performance is higher than predicted from the classification functions. The significant main effect 'stimulus type' indicates that single, complex, and speech-based stimuli are discriminated differently. It is not expected that there are different processing strategies that operate independently from each other, but rather, that there are different levels of processing. The level of processing depends on both stimulus complexity and task. For instance, in a same/different discrimination task listeners are capable of applying an analytical listening strategy to the complex formant stimuli. In the ABX task the complex stimuli may also be perceived analytically, but long-term phoneme labels may also influence discrimination.

The physical spacing between the single and complex formant stimuli are based on the difference limens in endpoint frequency. If both the single and the complex stimuli were perceived similarly, their discrimination functions should be comparable with respect to within-category and across-category discrimination. However, the discrimination functions of the complex formant transitions are more categorical than those of the single formant stimuli, although there are not always peaks at the phoneme boundaries. It is hardly expected that the single formant transitions, although speechlike, elicit phoneme labels from long term memory. Rather, the single transitions are more likely to be perceived as rising or falling glides. As subjects can listen attentively to the stimulus changes, auditory sensitivity is better for the final than for the initial single formant transitions. The high performance may also result from the step size being somewhat too large for these stimuli (the step size was based on an average difference limens in endpoint frequency).

Once the transitions are incorporated into a multi-formant speechlike sound, the F2-transition is perceived more categorically (either because the additional parameters hinder analytical perception or because the stimulus is complex enough to elicit phoneme labels). Auditory perception stil influences discrimination of the complex stimuli when the varying parameters are perceived very clearly, as is the case when the transition is in final position. As the step size between the stimuli is similar for both the initial and the final seven-item continua (compare discrimination data of Van Wieringen *et al.*, 1993 with step size in figure 1) discrimination between the VC-like stimuli seems to be easier than of the CV-like sounds. This was also found for these stimuli in the same/different AX task. The more complex the stimuli the more comparable the predicted and measured discrimination functions (the factor predicted/measured remains statistically significant). Therefore, it is possible that phonemic labelling influences the perception of the interpolated speech-based stimuli, although acoustical cues also contribute to their discriminability.

Discrimination is considered to be comparable for both the /a/-like and the /u/-like formant patterns. The statistically significant interaction between subjects and formant

pattern is explained by the relatively high discrimination functions of one of the subjects (subject 5).

Perceptual spacing of ABX data

To examine the perceptual spacing of the different kinds of speechlike stimuli in the ABX discrimination experiment the response frequencies are computed in terms of *d'ABX* (Macmillan and Creelman, 1991, for calculation see Appendix F in Van Wieringen, 1995) The hit and false alarm rate probabilities of adjacent pairs of stimuli along the continuum are converted into *d's*, under the assumptions of normal distributions and equal variance. Figure 8 illustrates the *d'* of the three subjects per stimulus type.

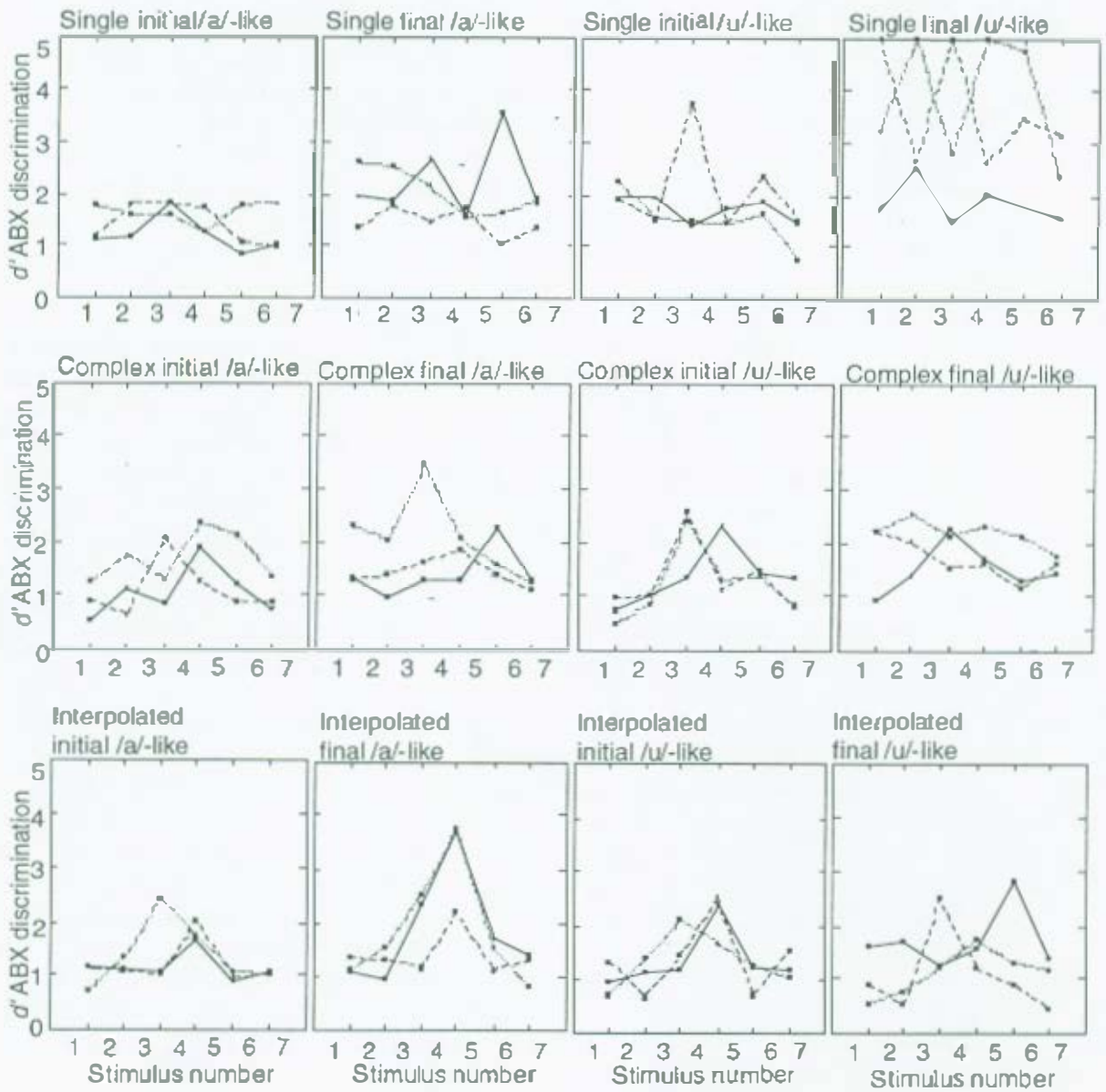


Figure 8. Response functions of the measured discrimination data in terms of *d'ABX*. Data of the single, complex, and interpolated /a/-like and /u/-like transtions in initial and final position are plotted separately. The three lines in each plot illustrate the performance of three different subjects.

In the case of categorical perception discriminability should be similar, for instance at a *d'* of 1.0, for those stimuli that are classified into the same category and it should then be higher than 1.0 for those stimuli that belong to different categories. Following figures 6 and 7 performance varies considerably: the data show no clear evidence of categorical perception. Only the data of the interpolated speech-based stimuli show peaks, suggesting phoneme boundaries (and possibly long-term phoneme cues). The single formant stimuli with /u/-like transitions in final position show increased sensitivity compared to the ones with transitions in initial position, with *d's* of 5.0! Yet, sensitivity for pairs of stimuli in final position also varies considerably: compare the performance of the three subjects of single /u/-like stimuli in final position.

Discussion

In daily communication acoustically different realisations can be perceived as the same phoneme. However, this does not necessarily mean that perception is based on a long-term phoneme labelling mechanism and that listeners cannot discriminate between stimuli that belong to the same category. The present experiments have shown that ABX discriminability of CV-like and VC-like formant stimuli and interpolated speech-based stimuli is based more on an analytical than on a labelling mechanism if the listener is capable of listening attentively to minor physical changes.

The less complex the stimulus the more capable the listener is of applying such an analytical strategy. Instead of /b/ or /d/ they probably make use of a larger range of alternatives, be it speech (/b/ /w/, /l/, /j/) or nonspeech (rising, level, etc.). This does seem to be the case with the interpolated speech based stimuli. Generally speaking their discrimination functions correspond to the ones predicted from the classification sigmoids. It is possible that the properties of these sounds are recovered from long-term memory and/or that the stimulus complexity does not allow analytical perception due to masking effects of other cues. In our experiments even the most complex stimuli are discriminated on the basis of acoustical cues: within-category performance is higher than chance level and across-category performance is better than predicted from the classification functions. It is expected that if feedback had been given during the experiment, subjects would have become even more aware of the acoustical differences. The ABX task requires a larger load on the memory than the AX one (e.g., Saslow, 1967), as the listener must make the double comparison of A to X and B to X.

General discussion

Two experiments were performed to examine how stimulus complexity, affected the perception of short and rapid transitions. The data suggest that perception of the speechlike stimuli is determined by a sensory mechanism rather than by a mechanism that extracts phoneme labels from long-term memory. In the AX discrimination task (Van Wieringen *et al.*, 1993, submitted), the listener compared and listened for differences between two stimuli presented in close temporal proximity. The listener was able to attend to the both the single and the complex stimuli in an analytical manner, because the task merely required a same/different judgement.

Analytical perception depends on the task and on the stimulus complexity. It was expected that long-term memory would interfere in the ABX discrimination paradigm, especially with the complex and interpolated speech-based stimuli, and that this would hinder analytical perception. This was not the case: in our study perception

was dominated by sensory factors, possibly because the formant stimuli were not as natural as speech sounds (Schouten and Van Hessen, 1992). However, the interpolated speech-based stimuli did not show clear evidence of categorical perception either. Contrary to the data of Schouten and Van Hessen (1992) which showed categorical perception for the consonants (not for the vowels), our speech-based data failed to show an equivalence between discrimination and phoneme labelling. Yet, although the peaks and valleys in the discrimination tasks can probably not be explained in terms of phoneme labelling behaviour, auditory sensitivity is not the same for the different kinds of stimuli of the continuum. Otherwise all the response functions such as in figure 7 would have been flat (as is the case with some of the single stimuli in final position).

Two factors can account for the differences in sensitivity: 1) categories result from sensory nonlinearities or 2) categories can be regarded as perceptual anchors in a temporary context-coding memory (Macmillan et al. 1987, 1988; Schouten and Van Hessen, 1992).

As for the first point: our psycho-acoustical experiments give no evidence of sensory nonlinearities (Van Wieringen *et al.*, 1993, 1994, submitted). Auditory sensitivity depends on transition duration, transition position and, to some extent, on rate-of-frequency change. However, the data show no enhanced discrimination at frequency loci of /b/ or /d/ stimuli. Therefore, sensory resolution of stimuli of one continuum that have the same transition duration and transition position should be the same (relatively small differences in frequency extent hardly affect discriminability, Van Wieringen and Pols, submitted).

As for the second point: if our formant stimuli are not natural enough to be retrieved from long-term memory, they could be retrieved from a temporary context coding memory that is created through training and feedback. Local maxima in sensitivity may occur near the extremes because the distance between the input and the anchor (a prototype) is estimated (the greater the distance the less accurately the perception of the stimulus). Perceptual differences occur as a result of the listener's inability to remember the stimuli precisely or a decay of the input as a function of time (Van Hessen and Schouten, 1992).

The more complex the stimulus, the more additional stimulus components seem to mask the varying transition under test (the less available the discriminative cues). Although it is not really possible to extrapolate a perceptual continuum between the formant stimuli and the interpolated speech-based ones, the lack of a perceptual equivalence between classification and discrimination leads us to conclude that perception of the speech-based stimuli too is influenced more by acoustical cues than by long-term memory.

Physical properties, such as the difference in frequency extent or the initial-final effect, remain perceptually salient in the speech paradigms with the simple stimuli, because perception of these stimuli is least affected by masking (there are, for instance, no other formant frequencies). Some of the single formant stimuli even yield similar results in the present ABX paradigm as in the same/different AX one (Van Wieringen and Pols, submitted). The ABX paradigm can also reflect basic sensitivity, although discrimination performance will be somewhat poorer due to a larger memory load.

Our study does not deny the role of linguistic experience in perception, it merely demonstrates that the perception of the stimuli in our experiments is strongly influenced by acoustical properties and that it is not necessarily based on speech-specific knowledge. More psycho-acoustical and speech perceptual research with the

most natural possible sounds are necessary to gain a thorough understanding of the processes underlying perception.

References

Hessen, A.J. van (1992): "Discrimination of familiar and unfamiliar speech sounds", Ph.D. *thesis*, University of Utrecht.

Hessen, A.J. van & Schouten, M.E.H. (1992): "Modeling phoneme perception. II: A model of stop consonant discrimination", *Journal of the Acoustical Society of America* 92, 1856-1868.

Macmillan, N.A., Goldberg, R.F.& Braida, L.D. (1988): "Resolution for speech sounds: basic sensitivity and context memory on vowel and consonant continua", *Journal of the Acoustical Societyof America* 84, 1262-1280.

Macmillan, N.A. & Creelman, C.D. (1991): *Detection theory: a user's guide* (Cambridge University Press).

Pollack I. & Pisoni, D.B. (1971): "On the comparison between identification and discrimination tests in speech perception", *Psychonomic Science* 24, 299-300.

Repp, B.H. (1984): "Categorical perception, issues, methods, findings", *Speech and language: Advance in basic research and practice* 10, 243-335.

Saslow M.G. (1967): "Frequency discrimination as measured by AB and ABX procedures", *Journal of the Acoustical Society of America* 41, 220-221.

Schouten, M.E.H. & Hessen, A.J. van (1992): "Modeling phoneme perception. I: Categorical perception", *Journal of the Acoustical Society of America* 92, 1841-1855.

Weenink, D.J.M. (1988): "Klinkers: een computerprogramma voor het genereren van klinkerachtige stimuli", *IFA-report nr. 100*.

Wieringen, A. van, Cullen, J.K. & Pols, L.C.W. (1993): "The perceptual relevance of CV- and VC transitions in identifying stop-consonants: cross-language results", *Proceedings EurosPeech'93* 2, 1499-1502.

Wieringen, A. van & Pols, L.C.W. (1994):"Frequency and duration discrimination of short first-formant speechlike transitions", *Journal of the Acoustical Society of America* 95, 502-511.

Wieringen, A. van & Pols, L.C.W. (submitted): "Discrimination of single and complex CV- and VC-like formant transitions", *Journal of the Acoustical Society of America*

Wieringen, A. van (1995): "Perceiving dynamic speechlike sounds: psycho-acouslics and speech perception", *Ph.D.-thesis*, University of Amsterdam.