PERCEIVING AND PROCESSING SPEECHLIKE SOUNDS

Astrid van Wieringen and Louis C.W. Pols
Institute of Phonetic Sciences, IFOTT, Amsterdam, The Netherlands

ABSTRACT

Perceptual processing of single and multi-formant CV-like and VC-like sounds, and of natural speech-based syllables is examined in three experiments to determine the extent to which perception of rapid transitions in speech can be explained by general auditory properties. These experiments show that resolution varies with the stimulus complexity and paradigm used, but that it is not controlled by speech-specific properties.

INTRODUCTION

Stop consonants in speech are cued by several properties including short and rapid transitions. Perceptual resolution of such rapid transitions is examined by asking listeners to classify (b or d), discriminate (ABX), and identify (by number 1-7) different kinds of speechlike transitions, which are preceded or followed by a stationary part. It was expected that perceptual processing would depend on the stimulus complexity, and that the number of discriminable or identifiable stimuli would decrease with increasing stimulus complexity: perception would be mediated more by long-term memory and less by acoustical properties with increasing speechlikeness of the stimuli.

STIMULUS GENERATION

Formant synthesis

The single and complex CV-like and VC-like syllables were generated by a digital formant synthesiser [6, 7]. 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, 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 actual 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 /a/-like (figure 1) or /u/-like (at 800 Hz) stationary portions. The transitions varied in endpoint frequency from 950 Hz to 1550 Hz in steps of 100 Hz, the average difference limen in frequency for these types of stimuli [5, 6].

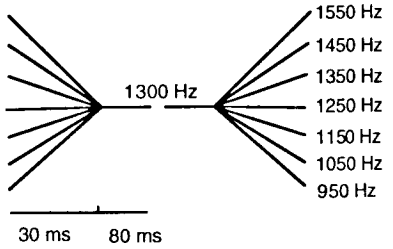

Figure 1. Schematic illustration of initial CV-like (left) and final VC-like (right) /a/-like single formant transitions (not to scale).

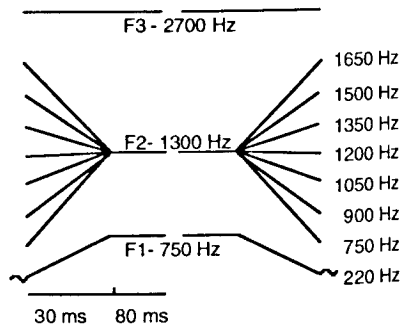

Figure 2. Schematic illustration of initial CV-like (left) and final VC-like (right) /a/-like complex formant transitions (not to scale).

The first-formant and second-formant transitions of the *complex* stimuli were also 30 ms, preceded or followed by an 80-ms /a/-like (1300 Hz) or /u/-like (800 Hz) steady-state. A stationary third formant, and a 20-ms voice bar were added to make the stimuli sound more speechlike (figure 2). The transitions varied in endpoint frequency from 750 Hz to 1650 Hz in 150 Hz steps, the average difference limen in frequency for these types of stimuli [5, 6]. 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 and at 2200 Hz for the /u/-like stimuli.

Interpolated speech-based stimuli

The speech-based stimuli were created by interpolating [4] the spectral envelope of two natural endpoints in seven steps, e.g., /ba/ and /da/. The original /ba/, /da/, /ab/, /ad/ 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.

In total twelve seven-syllable continua were created varying along the bilabial-to-alveolar dimension.

PROCEDURE

In the *ABX discrimination* task five subjects were tested individually in a quiet room. Three subjects listened to the /u/-like stimuli, three to the /a/-like ones (one of the subjects listened to both formant patterns). 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.

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. Each test, which was preceded by ten test triads, lasted approximately 10 minutes.

The same listeners also *classified* the single, complex, and interpolated speech-based stimuli as 'b' or 'd' on separate occasions.

In the *absolute identification paradigm* subjects are trained to assign a label (1-7) to each stimulus in a continuum. They have to learn the labels on the basis of their own criteria and therefore use numbers as response labels: no information is given about the nature of the stimuli under test. Feedback is given after each response, to maintain a constant level of performance.

Each subject made a total of 189 responses to the seven stimuli in each stimulus condition. Before each test series there were 63 test trials, which were not taken into account. Fourteen of these test series were collected for each stimulus complexity. Of these the first four were disregarded. Therefore, each of the six stimulus complexities per subject consisted of 1890 responses (270 x 7).

RESULTS

ABX discrimination

Figure 3 illustrates the 1-step ABX-discrimination functions and the classification sigmoids, averaged over the six subjects and two formant patterns (two statistically non-significant factors) together with the predicted discrimination function, based on the average classification sigmoids of these six subjects [2, 6]. 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. As basic sensitivity is comparable within a relatively large frequency range [6], perceptual discontinuities arise from attentional constraints due to the increasing number of cues with increasing complexity and speechlikeness of the stimuli.

In general, subjects discriminate better between subsequent pairs of stimuli than predicted from the classification sigmoids. Compared to the predicted functions, the experimental ones yield

higher percentage correct scores. As for classification performance, figure 3 shows that listeners were indeed able to classify the different kinds of stimuli consistently as 'b' or 'd'.

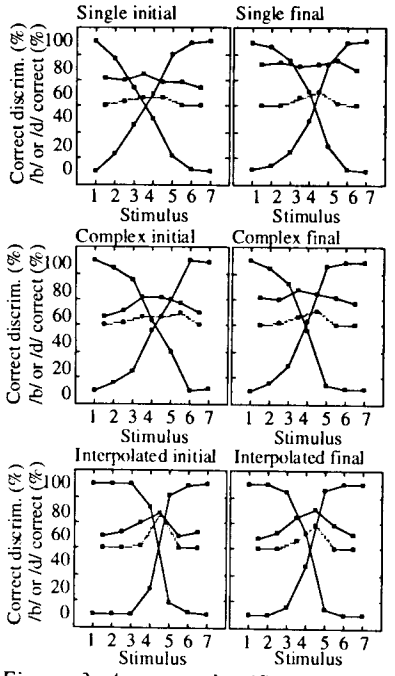

Figure 3. Average classification and discrimination scores (both actual (solid) and predicted (dashed)), averaged over subjects and formant patterns. The stimuli are indicated on the abscissa (the discrimination data apply to pairs of stimuli).

The issue is whether discrimination of single, complex and interpolated stimuli is based on sensory differences or on a phoneme labelling mechanism. From a sensory point of view listeners should be equally sensitive to acoustical differences of the single or the complex formant continua, because the step size is similar in a relative sense (being one JND). 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. Our study shows that listeners use acoustical cues to distinguish the seven stimuli of a continuum. Perception of the single formant stimuli can approach the limits of the auditory system, presumably because subjects can listen attentively to the varying acoustical cues. In these conditions the listener is more sensitive to acoustical differences in final than in initial transitions. The more complex the stimuli the more the responses are divided into two categories. However, there is no clear evidence of categorical perception, not even with the interpolated speech-based stimuli. In the case of categorical perception discrimination should be at chance level (50%) for those stimuli that are classified similarly and much higher than chance for those stimuli which are labelled differently.

Absolute identification

To determine whether the complexity of the stimulus induces 'speech categories', the data collected in the absolute identification experiment are also analysed in terms of *d'ident*, the perceptual distance between adjacent stimuli in an absolute identification experiment [1, 6]. Once *d'ident* is computed, the number of categories per stimulus continuum can be determined by means of a criterion. In the case of categorical perception *d'ident* is 1.0 (our arbitrary criterion for indistinguishable pairs of stimuli), for those stimuli which are labelled similarly, and higher than 1.0 for those which are labelled differently.

Figure 4 illustrates performance for the single, complex, and interpolated speech-based stimuli continua in initial (squares) and final (stars) position, averaged over three subjects. Data are plotted in terms of *d'ident* as a function of the neighbouring pairs of stimuli in the continua. The better the stimuli are identified, the smaller the number of confusions, and the higher the *d'ident*. High *d'ident's* were found with the single transitions in final position: subjects are very sensitive to the physical cues of these stimuli, as was also the case in the ABX-discrimination paradigm. The lower the *d'ident*, the less distinguishable the neighbouring pairs of stimuli are.

The figure shows that *d'ident* drops, on average, as the stimuli become more complex, and that the difference between initial and final transitions becomes

smaller with increasing stimulus complexity.

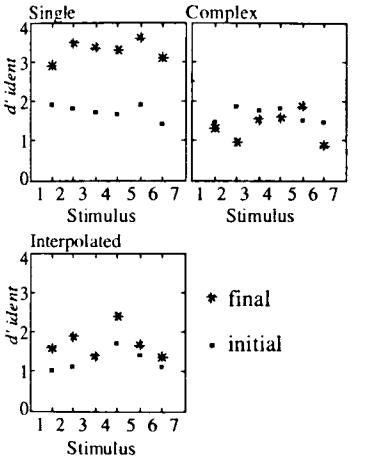

Figure 4. Absolute identification results, averaged over three subjects and two formant patterns. For more details see text.

It was expected that phonemic labelling would restrict the number of categories in a continuum, i.e., that the number of categories would decrease with increasing complexity of the stimulus. This does not appear from our data. As relatively few pairs of stimuli yield a *d'ident* lower than 1.0 we cannot conclude from our data that categorical perception occurs with increasing stimulus complexity. The interpolated speech-based stimuli show increased sensitivity between stimuli four and five, possibly as a result of a phoneme boundary in the stimulus continuum. Our results suggest that listeners perceive the single and complex stimuli in the so-called context-coding mode [3, 6]: they create internal representations of the continua under test and are capable of distinguishing 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.

In speech communication listeners (fortunately) do not need to perceive detailed acoustical cues. However, our study shows that the perception of vocalic transitions can, to some extent, be explained by general auditory properties, and that listeners can try to zoom in on certain levels of processing and discriminate ambiguous or new cues if they are not masked. In our study perception does not seem to be limited by a speech-specific mechanism based on long-term linguistic experience. In our study all the stimuli, including the interpolated speech-based ones, are discriminated better than predicted from the 2-AFC classification task, suggesting that listeners make use of additional acoustical cues. Experiments with natural speech transitions showed that the perceptual asymmetry between initial and final transitions decreases with increasing stimulus complexity, presumably because natural speech transitions contain redundant cues for plosive identification [6]. However, further study is necessary to understand how linguistic knowledge influences the perception of vocalic speech transitions.

REFERENCES

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

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

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

[4] Van Hessen, A.J. (1992): "Discrimination of familiar and unfamiliar speech sounds", *Ph.D.-thesis*, Univ. of Utrecht.

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

[6] Van Wieringen, A. (1995): "Perceiving dynamic speechlike sounds: psychoacoustics and speech perception", *Ph.D.-thesis*, University of Amsterdam.

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