UvA-DARE (Digital Academic Repository)

Perception of highly dynamic properties in speech

van Wieringen, A.; Pols, L.C.W.

Publication date
2006

Published in
Listening to speech. An auditory perspective

[Link to publication](#)

Citation for published version (APA):

van Wieringen, A., & Pols, L. C. W. (2006). Perception of highly dynamic properties in speech. In S. Greenberg, & W. A. Ainsworth (Eds.), *Listening to speech. An auditory perspective* (pp. 21-38). Lawrence Erlbaum Associates.

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

2

Perception of Highly Dynamic Properties in Speech

Astrid van Wieringen

Laboratorium Experimentele Otorhinolaryngologie, KU Leuven

Louis Pols

Institute of Phonetic Sciences, Amsterdam

Speech sounds vary continuously as a consequence of the changing configurations of the vocal tract, but the auditory system appears to be well equipped to integrate these spectro-temporally dynamic properties in our daily speech communication. On a global level, speech sounds are distinguished by the gross distribution of spectral energy (e.g., Blumstein, Isaacs, & Mertus, 1982), and by temporal envelope characteristics, such as duration and rise and fall times. On a more detailed level, they are characterized by their time-varying formant frequencies (e.g., Van Tassel, Soli, Kirby, & Widin, 1987). It is difficult to examine the perceptual importance of the various spectro-temporal properties in speech because the speech signal, in addition to being dynamic, is also highly complex by nature. As a result, it is difficult to isolate or manipulate single properties of the speech signal, especially when studying the characteristics on a detailed level. Under such conditions it is necessary to make use of such nonspeech analogs as tone and formant stimuli.

Over the past several years, several approaches have been taken to examine how complex dynamic events affect speech perception. For instance, in the field of profile analysis, detection of spectral changes in complex stimuli is usually examined by increasing the intensity of one frequency component relative to the other frequency components in a multitone complex signal (e.g., Green & Mason, 1984; Green, Mason, & Kidd, 1983; Versfeld & Houtsma, 1991). Such experiments have shown that detection improves as the number and density of the components increase, suggesting that comparisons are made across widely separated frequencies, not merely across adjacent channels.

Although both spectral and temporal cues are important for speech perception, dynamic temporal cues are more robust, and are least affected by background effects and speaker variation. Speech experiments by Drullman (1994) convincingly show that the envelope variations of the dynamic temporal envelope contain the essential information for phoneme perception. In his experiments, the original envelope is low-pass filtered, and the carrier signal is modulated according to the modified

envelope, leaving the fine structure intact. Speech experiments show that amplitude fluctuations can be limited to about 16 Hz without substantial reduction of speech intelligibility. In a much broader perspective, Greenberg (1996) explained why spectral analyses fail to account for the perceptual properties, and he presented a framework for speech perception, arguing, from both a perceptual and physiological point of view, in favor of dynamic temporal cues, in particular the low-frequency portion of the modulation spectrum.

In a series of experiments, we have examined how stimulus complexity and task load affect perception of spectro-temporal properties associated with short (20–50 ms) and rapid vocalic transitions in speech—those that carry a great deal of information about stop consonants (Fig. 2.1). The first formant (transition) specifies manner of articulation, whereas the onset or offset frequencies (also called the endpoint frequency in this study) and the direction of the second-formant transition specify place of articulation (e.g., Delattre, Liberman, & Cooper, 1955). Like any other acoustic signal, these rapid transitions undergo a psychoacoustic analysis during perception, which is limited by the sensitivity of the auditory system. However, much less is known about the dynamics of speech stimuli than about the perception of stationary vowel sounds. This is partly because of the difficulty of examining perceptual cues of sounds that consist of three covarying dimensions (frequency or extent, duration, and rate of frequency change), and partly because other physical parameters, such as amplitude and bandwidth, are difficult to control with dynamic sounds.

Following pioneering studies on the auditory basis of stop consonant perception (Schouten, 1985, 1986; Schouten & Pols, 1989), this study examines the perception of these highly dynamic spectro-temporal cues, from auditory analysis to plosive perception. The auditory basis of short and rapid vocalic transitions is examined by increasing the complexity of the stimulus (from tone sweeps to interpolated speech-based stimuli), and by varying the cognitive load of the task (from same–different classification to phoneme classification). Our approach is to examine how stimulus and task affect perceptual resolution of short and rapid vocalic transitions by creating a continuum of conditions measuring both psychoacoustic and speech-perception capabilities.

The next section of this chapter deals with the psychoacoustic capabilities of the normal hearing listeners by means of discrimination experiments that describe how sensitive normal hearing listeners are to changes in frequency, duration, and rate of frequency changes in short

FIG. 2.1. Spectrogram showing the rising first formant transition (F_1) and the falling second formant transition (F_2) in the second syllable of [ada].

and rapid transitions that vary in complexity from simple tone glides to complex formant transitions. In addition, the potential perceptual importance of some spectro-temporal properties is illustrated and discussed. The tone and formant transitions in initial and final position are short and rapid, corresponding acoustically to [b]-like and [d]-like vocalic transitions in speech. The psychoacoustic data were collected to gain (fundamental) knowledge on the limits of the auditory system to rapid spectral changes, as well as to provide (sensitivity) rules for speech synthesis-by-rule systems, and to serve as a basis for the speech perception experiments. We then discuss how speech tasks affect the perceptual resolution of different types of speech-like transitions, formant synthesis, and natural speech-based interpolated syllables. In addition, we examine whether higher order linguistic experience affects perception on a more central level under different experimental conditions. Finally, the stimulus continuum is completed by examining the spectro-temporal dynamic properties of natural speech transitions, of which the consonantal portion is successively deleted. In this section, the data pertaining to the psychoacoustic and speech experiments are integrated to determine how stimulus complexity and language background can affect perceptual processing of highly dynamic spectro-temporal properties in speech.

DISCRIMINATION OF RAPID TONE AND FORMANT TRANSITIONS

Difference limens (DLs) in endpoint frequency, duration, and rate of frequency change for different kinds of short (20–50 ms) and rapid tone and formant transitions were determined in same–different paired comparison tasks (van Wieringen, 1995; van Wieringen & Pols, 1994, 1995, 1998). Four types of transition signals were used:

1. *Tone glides*: These transitions were 40-ms, fixed-duration, sinusoidal tone glides, varying in endpoint frequency between 1850 Hz and 2050 Hz, and preceded or followed by a 200-ms stationary part.
2. *Isolated F_1 transitions*: Subsequently, DLs in duration and frequency were determined for 20-, 30-, and 50-ms isolated F_1 transitions (no stationary part) in the first-formant frequency range (550 Hz–1050 Hz).
3. *Single F_2 transitions*: 20-, 30-, and 50-ms F_2 transitions were preceded or followed by an 80-ms steady state.
4. *Complex F_2 transitions*: These same F_2 transitions could also be part of a complex signal consisting of a fixed, steady-state F_1 transition, a stationary third formant, and a 20-ms voice bar. The duration of the F_1 transition was the same as that of the F_2 transition.

The initial (CV-like, consonant–vowel) and final (VC-like, vowel–consonant) F_2 transitions were either [ba]-, [ab]-, [bu]-, and [ub]-like, or [da]-, [ad]-, [du]-, and [ud]-like (the stationary vowel-like component was 800 Hz for [u] and 1300 Hz for [a]). The complex formant transitions sounded like good prototypes of [ba], [da], and so on, whereas the single formant transitions and tone glides were perceived more like glissandos. Stylized representations of some of the stimuli are illustrated in Fig. 2.2.

As a consequence of covariation of frequency extent, rate, and duration in the dynamic stimuli, three conditions were tested:

1. The endpoint frequency was varied at a fixed duration to examine how DLs in frequency vary with stimulus complexity.
2. DLs in duration of isolated transitions (with no stationary part) were determined by changing the duration of the transition in 1-ms steps over a constant frequency range (van Wieringen & Pols, 1994).

FIG. 2.2. Schematic representation of 20-ms, 30-ms, and 50-ms complex formant transitions varying in F_2 endpoint frequency. The isolated transition has no steady state nor additional formants; the single-formant stimulus is just the second formant transition with steady state, but without additional formants.

3. The perceptual importance of duration and transition rate were tested with the complex transitions by keeping the rate of frequency change constant and varying the endpoint frequency and duration of the transition (as well as the steady-state portion) to maintain a constant stimulus duration (van Wieringen & Pols, 1995).

Four normal hearing individuals (between 23 and 36 years of age) participated in the experiments. They received approximately 2 hr of practice before actual data collection began. Two individuals were paid for participating.

RESULTS OF THE EXPERIMENTS

Dynamic Spectral Cues: Average DLs in Endpoint Frequency. Average DLs for endpoint frequency are plotted in Fig. 2.3 as a function of transition duration. The stimuli used in these experiments included fixed-duration tone glides, isolated formant transitions, and single as well as complex transitions with a steady-state component. The data are averaged over the parameters pertaining to participants, formant pattern ([a] and [u]), frequency range, and higher or lower rate of frequency change, because none of these factors proved statistically significant (see van Wieringen, 1995, for further details). Figure 2.3 illustrates three main (and statistically significant) findings for the condition in which endpoint frequency varies at a fixed duration for different stimulus types. First, discrimination improves with increasing duration for all types of transitions (see also Elliott, Hammer, & Carrell, 1991; Elliott, Hammer, Scholl, Carrell, & Wasowicz, 1989; Porter, Cullen, Collins, & Jackson, 1991). This is probably because the less the formant frequency changes spectrally, the easier it is to detect a change in (endpoint) frequency. Second, DLs increase as the stimuli become more speech-like. Perception of the varying transition is increasingly affected by surrounding acoustic information. In a tone glide only one harmonic varies in frequency, whereas in a formant transition the amplitudes of several harmonics change as the formant rises or falls in frequency. Third, final transitions are more discriminable than initial ones. This perceptual asymmetry, which has been encountered in both psychoacoustic and speech experiments, is discussed in more detail later.

Dynamic Temporal Properties: Average DLs in Duration

Short and rapid spectro-temporal changes not only signal place, but also manner of articulation. For example, a [w] is cued by transitions longer than 60 ms, and a [b] by transitions shorter than 40 ms (Miller & Liberman, 1979). We also examined auditory sensitivity to changes in transition duration, both with isolated formant transitions (van Wieringen & Pols, 1994), and with single and complex formant transitions in which the total duration remained fixed (van Wieringen & Pols, 1995, 1998). In the latter condition, the rate of frequency change remained constant and each change in transition duration was compensated for by a change in the duration of the stationary vowel part. This was also done to preclude loudness from being used as a cue.

Contrary to frequency-DLs, duration-DLs increase with increasing duration for a constant frequency range, because the time required for temporal discrimination increases as the duration of the standard increases (Creelman, 1962). Three frequency ranges in two frequency regions were examined. Transitions in the lower frequency region were similar to the first-formant transitions of voiced plosives (Liberman, Harris, Hoffman, & Griffith, 1957), and varied from 220 Hz to 550 Hz, 650 Hz, and 750 Hz. Transitions in the higher frequency region had no linguistic reference, and varied from 520 Hz to 850 Hz, 950 Hz, and 1050 Hz. Average DLs in duration were 2.7 ms, 4.5 ms, and 4.9 ms for the 20-, 30-, and 50-ms standard isolated transitions (Fig. 2.4; van Wieringen & Pols, 1994). These DLs are relatively larger than those of 50-ms stationary stimuli that were determined for the sake of comparison (see the cross symbol in Fig. 2.4). When transitions vary in transition duration, temporal as well as spectral properties are used. The shorter the transition the better the temporal resolution of the auditory system, as DLs increase with increasing transition duration. Transition duration is one of the major cues for distinguishing spectro-temporal dynamic properties in speech, and this persists throughout the different levels of processing.

FIG. 2.3. Difference limens in endpoint frequency (Hz) of tone glides, isolated formant transitions, as well as single and complex formant transitions with steady state. Difference limens in endpoint frequency of CV-like stimuli are marked by unfilled symbols, and those of VC-like stimuli are marked by filled symbols.

FIG. 2.4. Difference limens in duration of 20-ms, 30-ms, and 50-ms isolated formant transitions and of a 50-ms stationary stimulus.

Discrimination of Constant-Rate Transitions

In contrast to the release burst and the vocal murmur (which are often absent in natural speech) the vocalic transitions are always present as a consequence of the changing configuration of the articulators. It was therefore anticipated that transition rate would be an important cue for the perception of vocalic speech transitions. The potential perceptual importance of transition rate as a cue for discrimination of short formant transitions was examined by determining DLs in endpoint frequency for two different rates of frequency change: 5 Hz/ms and 10 Hz/ms, for equally long transitions. Because temporal cues strongly influence discrimination of isolated formant transitions (see the discussion pertaining to the previous experiment in the preceding section), temporal sensitivity was reduced by varying both the duration of the transition and the steady-state component. DLs for the endpoint frequency of complex transitions varying at 10 Hz/ms were 51 Hz, 52 Hz for a 30-ms standard transition, and 50 Hz and 47 Hz for a 50-ms standard transition in either initial and final position, respectively (Fig. 2.5). The DLs were lower for a 50-ms standard transition with a 5 Hz/ms rate of frequency change—42 Hz and 38 Hz, in initial and final position, respectively.

Altogether, the experiments show that DLs for endpoint frequency are smaller for transitions varying in transition duration at a constant rate of frequency change (the experiment described in this section) than for those varying in transition rate at a fixed transition duration (Experiment 1). The smaller DLs in the constant-rate condition probably result from excellent temporal resolution of the auditory system. Although the total duration remained constant, discriminability can also be cued by changes in loudness if the participant listened to changes in the duration (and hence the energy) of the transition.

Our psychoacoustic experiments have also shown that transition rate is perceptually important. Decreasing the rate of frequency change from 10 Hz/ms to 5 Hz/ms for the same transition duration im-

proves discrimination, possibly because frequency perception is more accurate per unit of time for the lower rate transitions (less spectral variation). Also, when the frequency range is relatively narrow, single-formant transitions increasing in the rate of frequency change (relative to the standard transition) yield smaller DLs than those decreasing in rate of frequency change (not shown in this chapter, but cf. van Wieringen & Pols, 1995). However, as this effect of relative rate of frequency change does not occur with initial or final complex transitions (not shown in this chapter; cf. van Wieringen & Pols, 1995), we must conclude that frequency and bandwidth cues (i.e., the difference in bandwidth between the first and last frequency of the transitions) are perceptually more important than transition rate.

The main conclusions that can be drawn from our experiments are that (a) DLs in frequency increase markedly as the stimuli become dynamic and more complex (speech-like), and (b) the magnitude of the perceptual asymmetry between initial and final transitions decreases as the stimuli become less “analytical.” Moreover, basic sensitivity is comparable across a large range of frequencies—the data do not reflect nonlinearities at certain phoneme-specific regions.

Subsequent experiments by Ainsworth and Carré (1997) and Ainsworth (1999) also show that formant transitions before, between, and after vowels are less discriminable than steady-state vowels (Kewley-Port & Watson, 1994). The DLs of two-formant 50-ms transitions in Ainsworth’s (1999) study are somewhat larger than those described in this chapter, presumably because his participants were not trained prior to testing. Despite this difference in magnitude, both studies show that DLs of complex dynamic stimuli cannot be predicted from stationary sounds. This might be of importance for rules involving formant transitions in speech synthesis-by-rule systems (Klatt, 1980).

FIG. 2.5. Difference limens in endpoint frequency (Hz) of initial and final 30-ms and 50-ms complex formant transitions at a constant rate of frequency change of 5 Hz/ms and 10 Hz/ms.

In addition, Ainsworth (1999) also reported that postvocalic (final) transitions are more discriminable than prevocalic (initial) transitions. Possible factors influencing such a perceptual asymmetry are discussed in greater detail later. In the following section we discuss how perceptual resolution varies with stimulus complexity in speech tasks.

PERCEPTUAL RESOLUTION WITH SPEECH TASKS

In daily communication, acoustically different spectro-temporal realizations are perceived as the same phoneme. Likewise, short and rapid vocalic transitions provide important, and even sufficient, cues for perception of plosives (Delattre, Liberman, & Cooper, 1955). In the following series of experiments, the perceptual importance of dynamic spectro-temporal speech properties are examined under the following conditions: (a) two-alternative, forced-choice classification, (b) ABX discrimination, and (c) absolute identification. The stimuli used were single and complex formant transitions, as well as interpolated speech-based syllables. We were interested in determining which spectro-temporal properties remain salient in the perceptual continuum. We hypothesized that if speech processing were mediated by brain mechanisms specific to speech, then ABX discrimination should be at chance level for stimuli labeled as the same phoneme in the classification task, and performance should be significantly greater than chance for stimuli classified as different phonemes. Of interest is whether categorization arises from increasing complexity of the stimulus, and whether listeners make use of acoustic cues that differentiate the seven items in each continuum. In the discrimination experiments the listeners should be able to discriminate the subsequent stimuli in each continuum, because the physical step size in the stimulus continuum is slightly larger than a DL in frequency.

Listeners classified (as either [b] or [d]) and discriminated (in an ABX task) single, complex, and interpolated speech-based, seven-item continua, ranging from [ba] to [da] and from [ab] to [ad] (and [bu] to [du] and from [ub] to [ud]). The step size in the single and complex formant continua were 100 Hz and 200 Hz, respectively (see Fig. 2.3) and were based on the average DLs in frequency for 30-ms transitions.

Two formant patterns were tested, those sounding [a]-like and those sounding [u]-like. Only the 30-ms transitions were taken into account. The interpolated continua were based on natural speech samples: the original [ba], [da], [ad] and [ab] stimuli (and [u]-like ones) were segmented from CVC tokens produced by a native Dutch male speaker (f_0 of about 110 Hz). Stimuli were digitized at a sample frequency of 20 kHz. All syllables were segmented to be 100 ms in duration. The stimuli were created by interpolating (van Hessen, 1992) the spectral envelope of the two natural endpoints in seven steps. The [a]-like and [u]-like continua were each discriminated in an ABX paradigm by three trained, normal hearing participants. No feedback was given during the test. Each of the four combinations per stimulus pair (ABA, ABB, BAA, BAB) was repeated 25 times, resulting in 100 observations per stimulus pair per participant. Fifteen normal hearing untrained listeners classified the single, complex, and interpolated speech-based stimuli as “b” or “d” on separate occasions. All tests were preceded by 10 test trials to familiarize the participants with the stimuli. Individuals were paid for participating.

Results Discrimination and Classification

Figure 2.6 illustrates the measured one-step ABX discrimination functions, as well as the classification functions, averaged over the three participants and two formant patterns (both of which were statistically nonsignificant factors), together with the predicted discrimination functions (Pollack & Pisoni, 1971).

The classification functions, which are based on 135 responses per stimulus, show that the seven items of the different single, complex, and interpolated speech-based continua are classified consistently as [b] or [d]. 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).

FIG. 2.6. Average classification and discrimination scores (both actual [solid] and predicted [dashed]), averaged over participants and formant patterns. The stimuli are indicated on the abscissa. The discrimination data apply to pairs of stimuli.

From a sensory point of view, listeners should be equally sensitive to acoustic differences of the single or the complex formant continua, because the step size is similar in a relative sense (being roughly one just noticeable difference, or JND). Figure 2.6 shows that the perception of initial and final transitions is, in large part, based on the spectro-temporal properties of the stimuli, and that it is not influenced by linguistic experience—discrimination is better than chance for sounds classified as the same phoneme. The less complex the stimulus, the better the different items of one stimulus continuum are differentiated (to the extent that discrimination of the single-formant transitions approaches the limits of the auditory system), whereas final transitions are more discriminable than initial ones. The more complex the stimuli, the more the responses are divided into two categories. However, there is no clear evidence of categorical perception (Liberman et al., 1957) in our study, not even with the interpolated, speech-based transitions.

Absolute Identification

To examine whether the speech data just described depend on the specific speech task, perceptual resolution was also examined in an absolute identification experiment. In this task, listeners are not given preassigned labels, but are first 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 provided about the nature of the stimuli under test. Feedback is given after each response to maintain a constant level of performance. Once again, we examined which dynamic spectro-temporal properties remain salient as the stimuli become more speech-like, and how the

physical properties relate to the perceptual dimensions. The less salient the dynamic spectro-temporal differences between the stimuli become, the more the perceptual continuum will be divided into two (phoneme) categories.

In total, 1,890 responses (270 repetitions of 7 stimuli) were collected for each stimulus complexity during the testing session. The data collected were analyzed in terms of d' , a signal-detection metric for measuring the distance between adjacent stimuli (Macmillan & Creelman, 1991), as well as in terms of the number of distinguishable categories per stimulus continuum that can be determined by means of a criterion. In the case of categorical perception, d' is 1.0 for those stimuli that are labeled similarly and higher than 1.0 for those labeled as different. We consider the value of 1.0 as our criterion for indistinguishable pairs of stimuli.

Results Absolute Identification

Figure 2.7 illustrates performance for the single, complex, and interpolated speech-based stimulus continua in initial (unfilled) and final (filled) position, averaged over three participants. Data are plotted in terms of d' as a function of the neighboring pairs of stimuli in the continua. The better the stimuli are identified, the smaller the number of confusions and the higher the d' . Higher d' values were found with single transitions in final position. Participants are very sensitive to the physical cues of the stimuli, as was also the case in the ABX discrimination paradigm (see Fig. 2.6). The figure shows that d' drops, on average, as the stimuli become more complex, and that the difference between initial and final transitions becomes smaller with increasing stimulus complexity.

The speech experiments presented in this section illustrate different levels of processing of spectro-temporal properties in speech. The underlying resolving power of the different kinds of stimuli and the acoustic information is extracted from all sounds in a similar manner. Recall that subsequent stimuli in the different continua were selected in such a way that they are discriminable in a same–different paired comparison task. However, not all stimuli are equally discriminable or identifiable in the actual speech experiments due to, for instance, additional attentional factors. Still, at all levels of processing, listeners make use of available acoustic information. In the identification experiments, listeners probably decide on the “goodness” of the syllables relative to a prototype, or they construct an internal representation of the stimuli in a temporary (context-coding) memory buffer in which they compare incoming signals (Macmillan, Braida, & Goldberg, 1987; Macmillan, Goldberg, & Braida, 1988). The data show that the differences between the seven stimuli of a continuum are remembered best with single-formant stimuli, and that these stimuli, which are relatively easy to perceive analytically, reflect the same perceptual asymmetry as found in the psychoacoustic experiments. At a higher level of processing, listeners are required to decide on the phonetic identity of the stimulus, albeit in a forced-choice task. Our data show that listeners are capable of ignoring the physical differences to classify the stimuli into predetermined categories.

PERCEPTUAL ASYMMETRY BETWEEN CV AND VC TRANSITIONS

One of the most salient phenomena encountered in both our psychoacoustic and speech experiments is the perceptual asymmetry between initial (CV) and final (VC) transitions (cf. Collins, Cullen, Jackson & Porter, 1990; Porter et al., 1991). Final transitions are more discriminable than initial ones, especially when listeners are capable of attending selectively to varying spectro-temporal changes, as in the case of a single varying transition. The more complex the stimulus and task, the less CV-like and VC-like transitions differ perceptually, probably as a result of masking effects by additional acoustical cues, and also because the listener’s attention is divided across a greater number of cues. Such findings suggest that the perceptual asymmetry should not occur with natural vocalic transitions, the most complex of all sounds used in our con-

FIG. 2.7. Absolute identification scores plotted in terms of d'_{ident} for neighboring pairs of stimuli. Data are averaged over three participants and two formant patterns ([a] & [u]). Initial (unfilled circles) and final transitions (filled circles) are presented separately for the single and the complex formant transitions, and for the interpolated speech-based stimulus continua.

tinuum. However, from the literature it is difficult to draw solid conclusions with respect to a possible perceptual asymmetry between CV and VC syllables in speech. Several speech experiments have shown VC transitions to be more consonant-specific than CV ones (e.g., Ohde & Sharf, 1977; Pols, 1979; Sharf & Hemeyer, 1972), although other studies show either just the opposite (e.g., Dorman, Studdert-Kennedy, & Raphael, 1977), or even a lack of asymmetry (Pols & Schouten, 1978, 1981). It is often argued that listeners may learn to depend more on VC transitions than on other consonantal cues, because plosives are often not released in final position. In addition, VC transitions are thought to add more to plosive classification CV transitions, as they converge into the consonant, whereas CV transitions move from consonant to vowel. Evidently, the results depend strongly on the specific segmentation procedure used (e.g., tape splicing, windowing, replacing deleted portions by noise), and on whether the different kinds of stimuli were presented in random order or in separate testing blocks.

It has also been shown that language background may cause initial and final transitions to be perceived differently (e.g., Fujimura, Macchi, & Streeter, 1978). In a joint study using Japanese speech with both Dutch and Japanese listeners we also found that Dutch listeners identified unvoiced stop consonants from the Japanese VC transition significantly better than Japanese participants, although both language groups identified the plosives even better from the CV transition (Kashino, van Wieringen, & Pols, 1992). All stimuli were extracted from natural Japanese VCV utterances. As Japanese is a CV-structured language, Japanese listeners find it even more difficult to identify the plosive from the VC transition alone.

Classification of Natural and Time-Reversed Speech Transitions

We also examined the perceptual asymmetry between initial and final transitions with natural transitions (van Wieringen, 1995). As CV and VC transitions are not entirely symmetric in natural speech, the original transitions were also time-reversed to be able to examine perception of acoustically symmetric transitions, namely the original CV transition and the time-reversed VC transition (which sounds CV-like), and vice versa. Voiced and voiceless transitions were excised from both CVC and VCV syllables, and identified by American English and Dutch listeners. Here we will mainly concentrate on a subset of our data of the cross-linguistic experiments, namely those of the original voiced and voiceless plosives excised from CVC syllables. Contrary to the comparison of Japanese and Dutch, the American English and Dutch languages are phonotactically similar, and only differ in the temporal structure on a segmental level. As American English voiced plosives generally have no voice bar, whereas carefully articulated Dutch voiced plosives do, American English listeners may be biased toward perceiving Dutch voiceless stimuli (that are unaspirated) as voiced. Contrary to American English, Dutch has no syllable-final voiced plosives. Dutch listeners may therefore be inclined toward perceiving the final voiced segments as voiceless. Or they may identify the [g] less well than the American listeners, as the voiced correlate of the [k] does generally not occur in the Dutch language.

Stimuli and Procedure

The three voiceless stops—[p], [t], [k]—as well as their voiced counterparts—[b], [d], [g]—were carefully pronounced in the context of three short Dutch vowels—[a], [o], [I]—as CVC syllables by one male native speaker of Dutch (average $f_0 = 110$ Hz). Stimuli were sampled at 20 kHz (low-pass filtered at 4.5 kHz, with a slope of 96 dB/oct, using a 12-bit resolution digital-to-analog converter), and five consonantal portions were successively deleted from equally long CV and VC syllables. The segmentation points of the voiced plosives were as follows:

1. The entire syllable.
2. Without the voice bar.
3. Without the release burst (the first transitional segment, referred to as A).

4. The second transitional segment (referred to as B).
5. The following period of the vowel.
6. The next two periods of the syllable.

The segmentation points of the unvoiced plosives were:

1. The entire syllable.
2. Without release burst (the first transitional segment, also referred to as A).
3. The second transitional segment (also referred to as B).
4. A period of the vowel.
5. Another period of the vowel.
6. The next two periods of the syllable.

Although segmentation occurred at the zero crossing, all segments, including the first (unmanipulated) one, were preceded or followed by 300 ms of low-level white noise to avoid abrupt onsets. All segments were also time-reversed. Some of the time-reversed stimuli sounded artificial because of an unnatural voice bar or release burst.

Twenty normal hearing Dutch and American English listeners classified the Dutch segments in a 6-AFC (alternative first choice) task ([b, d, g, p, t, k]) after hearing the stimuli once at a comfortable listening level in random order.

Results: Classification of Natural and Time-Reversed Speech Transitions

Figure 2.8 presents the percentage-correct identification scores of the voiced and unvoiced original stimuli as a function of segment number (1–6), averaged over languages. The figure shows that identification does benefit from the presence of a release burst and a voice bar, but that the transitional segments, A and B, still contain sufficient cues to identify the plosives above chance level (i.e., 33%). In general, our data show no evidence that the VC transition is more consonant-specific than the CV one. The transitional segments, A and B, are slightly (although not significantly) higher for the CV syllables in both language groups (Fig. 2.8 shows only averaged data, however). In addition, Dutch and American English listeners identify [g] equally well, despite the absence of this plosive in the Dutch language. There is no evidence that American English listeners show a systematic bias toward perceiving Dutch voiceless segments as voiced (data not shown).

The transitional parts (A and B) of the time-reversed stimuli are also equally well perceived in initial and final position (not shown). As our original segments did not show a clear perceptual asymmetry between CV and VC transitions, our time-reversed segments were not expected to show a perceptual asymmetry either. Compared to the response functions of the original sounds, those of the time-reversed ones are less consistent. Plosive identification does not always decline as the time-reversed segments become shorter, probably because of a response bias. Some participants clearly had difficulty identifying plosives from time-reversed segments, probably because segments 1 and 2 (the voice bar and release burst) sometimes sound strange when they are reversed in time. However, the data show no evidence of a systematic perceptual asymmetry, probably because initial and final transitions appear to be equally prominent for identification of plosives in our study.

The Perceptual Asymmetry Between Initial (CV) and Final (VC) Transitions

In conclusion, contrary to four psychoacoustic results, the global pattern of responses shows no evidence that Dutch vocalic transitions are more consonant-specific in final than in initial position. The perceptual asymmetry encountered in the psychoacoustic experiments most probably occurs at least

FIG. 2.8. Percentage correct identification scores of the original voiced (left) and unvoiced (right) speech segments as a function of segment number (1–6) averaged over Dutch and American English listeners.

partly as a result of later information in the signal being remembered best, although there is also some evidence that the asymmetry already exists at an auditory level, due to instability effects at the onset of the CV-like stimulus (van Wieringen, 1995). The less complex the transition, the better the stimuli can be perceived analytically, and the more discrimination approaches the limits of auditory resolving power. This was found in both the psychoacoustic and speech experiments. From a different perspective it can also be argued that postvocalic (VC) transitions need to be more discriminable than prevocalic (CV) transitions, because the speech gesture is not completed yet in the postvocalic condition. Therefore, the direction of the transition that points toward the next target is perceptually important (Ainsworth, 1999; Carré & Ainsworth, 1996).

In contrast to tone and formant glides, natural speech transitions cannot be carefully controlled physically. Therefore, a possible perceptual asymmetry can also occur as a result of a production-based asymmetry. Acoustic analyses of different cues for consonant perception (release burst, aspiration, closure interval, etc.) showed that the perceptual advantage associated with initial consonants over final consonants is attributable to differences in production, resulting in, for instance, longer duration or higher amplitude of the initial consonant (Redford & Diehl, 1999). However, it is fairly unlikely that initial consonants would be significantly more identifiable than final ones as a result of the short and rapidly varying formant transition only. The data generated in speech experiments depend to a large degree on methodological factors, such as segmentation procedure and the presence or absence of noise (Pols & Schouten, 1978, 1981). The perceptual asymmetry between initial and final formant transitions observed in other studies does not necessarily stand in contradiction to the results of this study; rather it is not anticipated to occur in a carefully balanced laboratory experiment.

It may also be that the perceptual asymmetry is caused or enhanced by the vocalic portions preceding or following the transitions (Ohde & Sharf, 1977). Identification experiments with natural CV and VC segments, which were taken from continuous read speech, show that consonant identification benefits from the vocalic portions, especially when the consonants are in final position (van Son & Pols, 1995). Moreover, in a detailed study of acoustic cues, Smits, ten Bosch, and Collier (1996) showed that the relative perceptual importance of bursts and transitions also depends on the vowel context. In general, the release burst dominates place perception in front vowel contexts ([i]), whereas the transitions are perceptually more important in many nonfront contexts ([a], [u]).

DISCUSSION

In speech communication, listeners fortunately need not perceive all detailed spectro-temporal differences, such as the difference in the duration and extent of frequency change. If we consider our data on a perceptual continuum, with stimuli and tasks increasing in complexity to match speech conditions, the perceptual importance of psychoacoustic cues seems to diminish steadily with increasing stimulus complexity, due to partial masking effects and attentional constraints. Evidently, different experimental tasks trigger different modes of processing, during both spectral and temporal analysis, with the auditory system dealing with many different tasks, ranging from auditory acuity to perceptual organization.

At least two interesting findings emerge from the studies dealing with the psychoacoustics of speech-like stimuli. The first (and most important one) is that the perception of complex dynamic stimuli cannot be predicted from simple (or complex) stationary signals. DLs in frequency of stationary sinusoids (of 40 dB SPL) are about 1 Hz at 200 Hz, 2 Hz at 1000 Hz and 11 to 15 Hz at 4000 Hz (Wier, Jesteadt, & Green, 1977). DLs of more complex stationary vowel-like signals are on the order of 3 to 5% of the formant frequency (Flanagan, 1955) or range between 10 and 30 Hz (Kewley-Port & Watson, 1994). Compared to stationary stimuli, dynamic ones are far more complex to study, because frequency and duration covary with rate of frequency change (a decrease in transition rate also results in a narrower frequency range assuming the transition duration remains fixed). Both the decrease in transition rate and the narrower frequency range result in an increase in sensitivity (i.e., smaller DLs). For these stimuli, DLs cannot be expressed as a constant, but depend on the position (initial or final) and direction (rising or falling) of the transition. In this study, DLs in endpoint frequency of tone glides were, on average, 54 Hz in initial position and 23 Hz in final position. However, DLs in endpoint frequency of dynamic speech-like formant transitions ranged between 80 Hz and 220 Hz (depending on the complexity of the stimulus) in this study (see Fig. 2.3) or between 132 Hz and 647 Hz in the study of Ainsworth (1999). Similar results were obtained in gap detection experiments. Whereas minimal detectable gaps are on the order of 5 ms for spectrally similar markers, performance decreases by approximately tenfold as the stimuli become more complex, both for acoustic (e.g., Formby, Barker, Abbey, & Raney, 1993; Formby & Forrest, 1991; Formby, Sherlock, & Li, 1998) and electrical stimulation (van Wieringen & Wouters, 1999).

A second finding is that sufficient training can improve performance markedly, even when the stimuli are complex. In this study, participants were trained for at least 2 hr prior to data collection. Compared to the data of untrained listeners of Ainsworth (1999), DLs in endpoint frequency of (approximately) the same prevocalic (CV) and postvocalic (VC) stimulus conditions were much smaller in this study, that is, 142 Hz versus 352 Hz (Ainsworth, 1999) for CV-like stimuli and, on average, 131 Hz versus 224 Hz (Ainsworth, 1999) for VC-like stimuli. Similarly, the effect of training is also demonstrated in the gap detection study using electrical stimulation mentioned previously (van Wieringen & Wouters, 1999). The exact amount of training differs among cochlear implant subjects, but once the temporal cue is detected among other potentially confounding cues (some implant patients never succeed in performing the task, and others succeed only to a limited degree), gap thresholds of complex (across-channel) configurations approach those of simple (within-channel) ones. It is suggested that the increased sensitivity may not result from increased neural interaction, but may rather arise from the individual's ability to ignore the (confusing) perceptual differences between the pre- and postgap marker.

Although many psychoacoustic experiments have shown that DLs in endpoint frequency are generally smaller in final (VC) than in initial (CV) position (e.g., Ainsworth, 1999; Elliot, Hammer, Scholl, Carrell, & Wasowicz, 1989; Porter et al., 1991), this study also shows that the perceptual asymmetry between CV-like and VC-like transitions becomes less salient with increasing stimulus complexity. Other tests with vowel transitions have also shown that, within a certain range, delays between the onset of F_1 and F_2 transitions are not perceived (Ainsworth,

1995; Ainsworth & Carré, 1999). If the first and second transitions are equally long, the perceptual system requires at least a delay of 30 ms for the two sounds to be distinguishable. Greater delays are required for transitions of different durations. Analytical experiments, such as ours, illustrate that spectral cues are not very robust. Psychoacoustic cues are integrated into subsequent levels of processing and the perceptual importance for all cues may decrease for complex stimuli at higher levels of processing, not only because of the limited resolving power of the auditory system, but also because the listener has to divide his or her attention over a greater number of cues. In speech, this is compensated for by the combination of many cues and by additional stimulus properties that provide more cues for perception. Further experiments on vowel and consonant perception have shown that the perceptual cues are not only segmental, but also lie beyond the phoneme boundary (van Son & Pols, 1999).

This perceptual loss concerning details is not permanent. Listeners are able to use the (unmasked) cues by paying selective attention to them, for instance when phoneme labeling is ambiguous due to such factors as sloppy speech or non-native sounds. Considered in this way, speech perception, including the perception of highly dynamic vocalic transitions, is largely a matter of attention.

ACKNOWLEDGMENT

Part of this work was published in van Wieringen and Pols (1998), *Acta Acustica*, 84, 520–528.

REFERENCES

- Ainsworth, W. A. (1995). Perception of formant transitions in synthesized vowel pairs. In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Vol. 2, pp. 666–669). Stockholm, Sweden: 13–19 August 1995.
- Ainsworth, W. A. (1999). Discrimination of the shape of two-formant transitions before, between and after vowels. *Acta Acustica*, 85, 121–127.
- Ainsworth, W. A., & Carré, R. (1999). Perception of synthetic two-formant vowel transitions. *Speech Communication*, 21, 273–282.
- Blumstein, S. E., & Isaacs, E., & Mertus, J. (1982). The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 72, 43–50.
- Carré, R., & Ainsworth, W. A. (1996). Characteristics of vowel to vowel transitions in terms of speech gestures. *Acta Acustica*, 82, 128.
- Collins, M. J., Cullen J. K., Jackson, D. F., & Porter, R. J. (1990). Discrimination of formant-like frequency transitions preceded or followed by steady-states. In *Abstracts of the Thirteenth Midwinter Meeting of the Association for Research in Otolaryngology* (pp. 181–182).
- Creelman, C. D. (1962). Human discrimination of auditory duration. *Journal of the Acoustical Society of America*, 34, 582–593.
- Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27, 769–773.
- Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception and Psychophysics*, 22, 109–122.
- Drullman, R. (1994). *Intelligibility of temporally degraded speech: A study of the significance of narrowband temporal envelopes*. Unpublished doctoral dissertation, Vrije Universiteit van Amsterdam, Amsterdam, The Netherlands.
- Elliott, L. L., Hammer, M. A., & Carrell, T. D. (1991). Discrimination of second-formant-like frequency transitions. *Perception and Psychophysics*, 50, 1–6.
- Elliott, L. L., Hammer, M. A., Scholl, M. E., Carrell, T. D., & Wasowicz, J. M. (1989). Discrimination of rising and falling simulated single-formant frequency transitions: Practice and duration effects. *Journal of the Acoustical Society of America*, 86, 945–953.
- Flanagan, J. L. (1955). Difference limen for vowel formant frequency. *Journal of the Acoustical Society of America*, 27, 613–617.
- Formby, C., Barker, C., Abbey, H., & Raney, J. J. (1993). Detection of silent temporal gaps between narrow-band noise markers having second-formantlike properties of voiceless stop/vowel combinations. *Journal of the Acoustical Society of America*, 93, 1023–1027.

- Formby, C., & Forrest, T. G. (1991). Detection of silent temporal gaps in sinusoidal markers. *Journal of the Acoustical Society of America*, 89, 830–837.
- Formby, C., Sherlock, L. P., & Li, S. (1998). Temporal gap detection measured with multiple sinusoidal markers: Effects of marker number, frequency, and temporal position. *Journal of the Acoustical Society of America*, 104, 984–998.
- Fujimura, O., Macchi, M. J., & Streeter, L. A. (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech*, 21, 337–346.
- Green, D. M., Mason, C. R., & Kidd, G., Jr. (1983). Profile analysis: Critical bands and duration. *Journal of the Acoustical Society of America*, 75, 1163–1167.
- Green, D. M., & Mason, C. R. (1984). Auditory profile analysis: Frequency, phase and Weber's law. *Journal of the Acoustical Society of America*, 77, 1155–1161.
- Greenberg, S. (1996). Understanding speech understanding: Towards a unified theory of speech perception. In W. Ainsworth & S. Greenberg (Eds.), *Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception* (pp. 1–8). Keele, UK.
- Kashino, M., van Wieringen, A., & Pols, L. C. W. (1992). Cross-language differences in the identification of intervocalic stop consonants by Japanese and Dutch listeners. In J. J. Ohala, T. M. Nearey, B. L. Derwing, M. M. Hodge, & G. E. Wiebe (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (Vol. 2, pp. 1079–1082). Banff, Alberta, Canada.
- Kewley-Port, D., & Watson, C. S. (1994). Formant-frequency discrimination for isolated English vowels. *Journal of the Acoustical Society of America*, 95, 485–496.
- Klatt, D. H. (1980). Software for a cascade/parallel formant synthesiser. *Journal of the Acoustical Society of America*, 67, 971–995.
- Lieberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358–368.
- Macmillan, N. A., Braida, L. D., & Goldberg, R. F. (1987). Central and peripheral processes in the perception of speech and nonspeech sounds. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 28–45). The Hague, Netherlands: Martinus Nijhoff.
- Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.
- 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 Society of America*, 84, 1262–1280.
- Miller, J. L., & Liberman, A. M. (1979). Some effects of later occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics*, 25, 457–465.
- Ohde, R. N., & Sharf, D. J. (1977). Order effect of acoustic segments of VC and CV syllables on stop and vowel identification. *Journal of Speech and Hearing Research*, 20, 543–554.
- Pollack, I., & Pisoni, D. B. (1971). On the comparison between identification and discrimination tests in speech perception. *Psychonomic Science*, 24, 299–300.
- Pols, L. C. W. (1979). Coarticulation and the identification of initial and final plosives. In J. J. Wolff & D. H. Klatt (Eds.), *ASA*50 speech communication papers* (pp. 459–562). New York: Acoustical Society of America.
- 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., & Schouten, M. E. H. (1981). Identification of deleted plosive: The effect of adding noise or applying a time window (A reply to Ohde and Sharf). *Journal of the Acoustical Society of America*, 69, 301–303.
- Porter, R. J., Cullen, J. K., Collins, M. J., & Jackson, D. F. (1991). Discrimination of formant transition onset frequency: Psychoacoustic cues at short, moderate, and long durations. *Journal of the Acoustical Society of America*, 90, 1298–1308.
- Redford, M. A., & Diehl, R. L. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *Journal of the Acoustical Society of America*, 106, 1555–1565.
- Schouten, M. E. H. (1985). Identification and discrimination of sweep tones. *Perception and Psychophysics*, 37, 369–376.
- Schouten, M. E. H. (1986). Three-way identification of sweep tones. *Perception and Psychophysics*, 40, 359–361.
- Schouten, M. E. H., & Pols, L. C. W. (1989). Identification and discrimination of sweep formants. *Perception and Psychophysics*, 46, 235–244.

- Sharf, D. J., & Hemeyer, T. (1972). Identification of place of consonant articulation from vowel formant transitions. *Journal of the Acoustical Society of America*, 51, 652–658.
- Smits, R., ten Bosch, L., & Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants: I. Perception experiment. *Journal of the Acoustical Society of America*, 100, 3852–3864.
- van Hessen, A. J. (1992). *Discrimination of familiar and unfamiliar speech sounds*. Unpublished doctoral dissertation, University of Utrecht, Utrecht, Netherlands.
- van Son, R. J. J. H., & Pols, L. C. W. (1995). How transitions and local context affect segment identification. *Proceedings of the Institute of Phonetic Sciences*, 19, 51–69.
- van Son, R. J. J. H., & Pols, L. C. W. (1999). Perisegmental speech improves consonant and vowel identification. *Speech Communication*, 29, 1–22.
- Van Tassell, D. J., Soli, S. D., Kirby, V. M., & Widin, G. P. (1987). Speech waveform envelope cues for consonant recognition. *Journal of the Acoustical Society of America*, 82, 1152–1161.
- van Wieringen, A. (1995). *Perceiving dynamic speechlike sounds: Psycho-acoustics and speech perception*. Unpublished doctoral dissertation, University of Amsterdam, Netherlands.
- van Wieringen, A., & 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.
- van Wieringen, A., & Pols, L. C. W. (1995). Discrimination of single and complex CV- and VC-like formant transitions. *Journal of the Acoustical Society of America*, 98, 1304–1312.
- van Wieringen, A., & Pols, L. C. W. (1998). Discrimination of short and rapid speechlike transitions. *Acta Acustica*, 84, 520–528.
- van Wieringen, A., & Wouters, J. (1999). Gap detection in single- and multiple-channel stimuli by LAURA cochlear implantees. *Journal of the Acoustical Society of America*, 106, 1925–1939.
- Versfeld, N. J., & Houtsma, A. J. M. (1991). Spectral-shape discrimination of two-tone complexes. In Q. Summerfield (Ed.), *Hearing and speech* (pp. 459–479). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wier, C. C., Jesteadt, W., & Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America*, 61, 178–184.