Effect of temporal envelope smearing on speech reception

Rob Drullman, Joost M. Festen, and Reinier Plomp
Department of Oto-rhino-laryngology, Free University Hospital, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands

(Received 17 September 1992; revised 30 June 1993; accepted 20 September 1993)

The effect of smearing the temporal envelope on the speech-reception threshold (SRT) for sentences in noise and on phoneme identification was investigated for normal-hearing listeners. For this purpose, the speech signal was split up into a series of frequency bands (width of $\frac{1}{4}$, $\frac{1}{2}$, or 1 oct) and the amplitude envelope for each band was low-pass filtered at cutoff frequencies of 0, $\frac{1}{2}$, 1, 2, 4, 8, 16, 32, or 64 Hz. Results for 36 subjects show (1) a severe reduction in sentence intelligibility for narrow processing bands at low cutoff frequencies (0–2 Hz); and (2) a marginal contribution of modulation frequencies above 16 Hz to the intelligibility of sentences (provided that lower modulation frequencies are completely present). For cutoff frequencies above 4 Hz, the SRT appears to be independent of the frequency bandwidth upon which envelope filtering takes place. Vowel and consonant identification with nonsense syllables were studied for cutoff frequencies of 0, 2, 4, 8, or 16 Hz in $\frac{1}{4}$-oct bands. Results for 24 subjects indicate that consonants are more affected than vowels. Errors in vowel identification mainly consist of reduced recognition of diphthongs and of confusions between long and short vowels. In case of consonant recognition, stops appear to suffer most, with confusion patterns depending on the position in the syllable (initial, medial, or final).

PACS numbers: 43.71.Es, 43.71.Gv

INTRODUCTION

The speech signal is characterized by a spectrum that varies in time. This is clearly illustrated by the spectrogram: The distribution of light and dark spots in vertical direction (frequency) changes continuously in horizontal direction (time). These variations contain the information that is essential for the identification of phonemes, syllables, words, and sentences. For this identification we need a detector which is able to perceive the spectrotemporal differences. Our ear is such a detector.

The ear's resolution in both frequency and time is sufficiently high to perceive the essential acoustical features of the various speech sounds. Depending on the speech material, we even have a reserve capacity. This reserve capacity is rather small for isolated phonemes, but large for sentences. For normal-hearing listeners, the speech-reception threshold (SRT) in noise, defined as the speech-to-noise ratio at which 50% of short everyday sentences are reproduced correctly, is about −5 dB (Plomp, 1986).

An interesting question is: How critical is the resolution in frequency and time for the intelligibility of speech? Recently, ter Keurs *et al.* (1992, 1993) investigated the effect of smearing in the frequency domain, as a way to reduce spectral contrast. They smeared the envelope of the spectrum over bandwidths varying from $\frac{1}{8}$ to 4 oct. The effect of this operation can be considered as a blurring of the formant structure. The results indicate that the SRT for sentences in noise increases as spectral energy is smeared over $\frac{1}{2}$ oct and more, thus exceeding the ear's critical bandwidth.

In the present study we focus on the *temporal* envelope. Temporal modulations of the speech signal have been described in terms of the modulation index (Houtgast and Steeneken, 1985). In all octave bands the most important modulation frequencies (i.e., where the modulation index reaches its peak value) are 3–4 Hz, reflecting the syllable rate in speech. Taking the frequency for which the modulation index is reduced to half its peak value (comparable to the −6-dB point of a filter), one can find relevant modulation frequencies up to about 15–20 Hz in undisturbed speech. The ear's sensitivity for temporal modulations shows a lowpass characteristic, with a 6-dB down point corresponding to a frequency roughly between 25 Hz (Festen and Plomp, 1981; Plomp, 1984) and 100 Hz (Rodenburg, 1977; Viemeister, 1979). From these data we may conclude that for normal hearing the ear's capacity to detect temporal modulations is not a limiting factor in speech perception.

What is the effect of reducing the degree to which temporal fluctuations are present in the speech signal? In case of reverberation, resulting in attenuation of fast temporal modulations (due to "filling" of the minima in the waveform by reflected speech), experiments have demonstrated a reduction in intelligibility for sentences (Duquesnoy and Plomp, 1980). With regard to multichannel amplitude compression, Plomp (1988) argues that with small time constants intensity fluctuations (particularly at low modulation frequencies) are attenuated in every channel, resulting in reduced intelligibility. Plomp states that this reduction increases as the compression ratio and the number of channels increase. In a comment on Plomp's paper, Villchur (1989) claims that infinite peak clipping (i.e., 100% compression) in a two-channel compression system would hardly affect intelligibility for normal-hearing listeners. One of the goals of the present paper is to quantify this effect as a function of the number of frequency bands.

 0001-4966/94/95(2)/1053/12/$6.00

The significance of the various modulation frequencies for speech communication can be compared with the significance of the various audiofrequencies. For example, in designing channel vocoders, we need to know not only the frequency range (e.g., up to 4 kHz) to be covered by the channels, but also the upper limit of the envelope frequencies required to preserve intelligible speech. Similarly, in applying alternative presentation of speech information to the deaf, we need to know up to which envelope frequency the (tactile, visual) channel must transfer the signal faithfully. The range of modulation frequencies most relevant for speech, as mentioned above, has been determined by means of physical/acoustical measurements, and not by any formal perceptual evaluation. In much the same way, a 25-Hz limit for temporal modulations in up to 100 filter bands was applied in early channel vocoders (cf. Flanagan, 1972). There have been measurements of consonant and vowel intelligibility scores, mostly for diagnostic purposes. But, as far as we know, the limit for temporal modulations has never been determined explicitly by means of intelligibility tests.

As to phoneme perception, several investigators have studied the information contained in the temporal envelope for consonant recognition in cases of limited spectral cues. With noise stimuli modulated by the amplitude envelope of /aCa/ syllables, Van Tasell *et al.* (1987) found poor identification scores, with highly variable performance across (untrained) subjects. Several features (voicing, amplitude, and burst) could be derived from the envelope, but for modulation frequencies up to 20 Hz, they accounted for only 19% of the transmitted information. When /aCa/ stimuli are masked by white noise with the same temporal envelope as the speech waveform, Freyman *et al.* (1991) have shown that nonlinear amplification of the envelope (a 10-dB increase of the consonant portion) has no effect on overall consonant recognition, but it can alter confusion patterns for specific consonant groups. In cases without limited spectral information, Behrens and Blumstein (1988) found that interchanging the amplitude of various voiceless fricatives in CV syllables resulted in few or inconsistent place of articulation errors. They concluded that, at least for voiceless fricative noise, compatibility of the spectral properties and of formant transitions dominated the effect of amplitude manipulations. It should be noted that the results of all these studies are based on *wideband* amplitude envelopes.

With the present perception experiments we investigated the extent to which speech intelligibility depends on the details (fast modulations) in the temporal envelope of the signal. In a first experiment the intelligibility for sentences in quiet and the SRT for sentences in noise were measured as a function of temporal smearing. In a second experiment the effects on vowel and consonant identification in nonsense syllables were studied.

I. METHOD

A. Temporal envelope smearing

Before describing the actual signal processing applied in this study, we have to consider the possible methods for envelope smearing. In general, there are two ways to reduce envelope variations: Convolution applied to the fine structure (carrier signal) on the one hand, or convolution (low-pass filtering in case of smearing) applied to the envelope on the other.

The essential feature of the former method is energy splatter. An example of this is reverberation, characterized by convolution of the carrier signal (within a frequency band) with a causal one-sided exponentially decaying impulse response. In signal processing, however, one could apply any impulse response, also symmetrical ones. A consequence of convolving the fine structure is that low amplitude and silent intervals are filled with carrier energy imported from adjacent phonemes. The same is true for the troughs of the fast (and weaker) amplitude modulations. As a result of this, the higher modulation frequencies in the temporal envelope are attenuated (typically by 6 dB/oct; Plomp, 1984), yielding a smeared envelope.

In the second method the temporal envelope is smeared directly, viz. by low-pass filtering the original envelope and modulating the carrier signal according to this modified envelope (i.e., multiplying the fine structure by the ratio between the filtered and the original envelope at each point in time). With this method there is no energy splatter. So, contrary to the previous method, there is no filling in case of silent intervals (zero carrier amplitude). However, as the digitized speech material we used was originally recorded on analog tape (with a signal-to-noise ratio of about 50 dB), tape noise provided just enough carrier signal in "silent" intervals.

We adopted the second method for smearing the temporal envelope. It has the advantage that the fine structure remains intact and that we can control the process of filtering the envelope by selecting the cutoff frequency and the slope of the filter. In this way it is known how the temporal modulation spectrum changes, and the intelligibility can be evaluated as a function of the temporal envelope cutoff frequency.

Smearing of the temporal envelope should not be done on the wideband speech signal. It is known that those modulations can be reduced considerably, without affecting the intelligibility in a major way. Since the temporal fluctuations of speech are only partly correlated over frequency (the more two frequency bands are separated, the lower their correlation, cf. Houtgast and Verhage, 1991), the wideband signal does not include all temporal amplitude variations in the different frequency bands. Therefore, the speech signal has to be split up into several frequency bands, so that the temporal envelope of each individual band can be modified.

B. Signal processing

For the signal processing, an analysis-resynthesis scheme for smearing the amplitude envelope of digitized speech was developed. A block diagram of the processing is shown in Fig. 1. First, the original wideband speech signal (digitized into 16 bits at a sampling rate of 15 625 Hz) is led through a filter bank with linear-phase FIR bandpass filters, covering the range 100–6400 Hz. The slopes of these

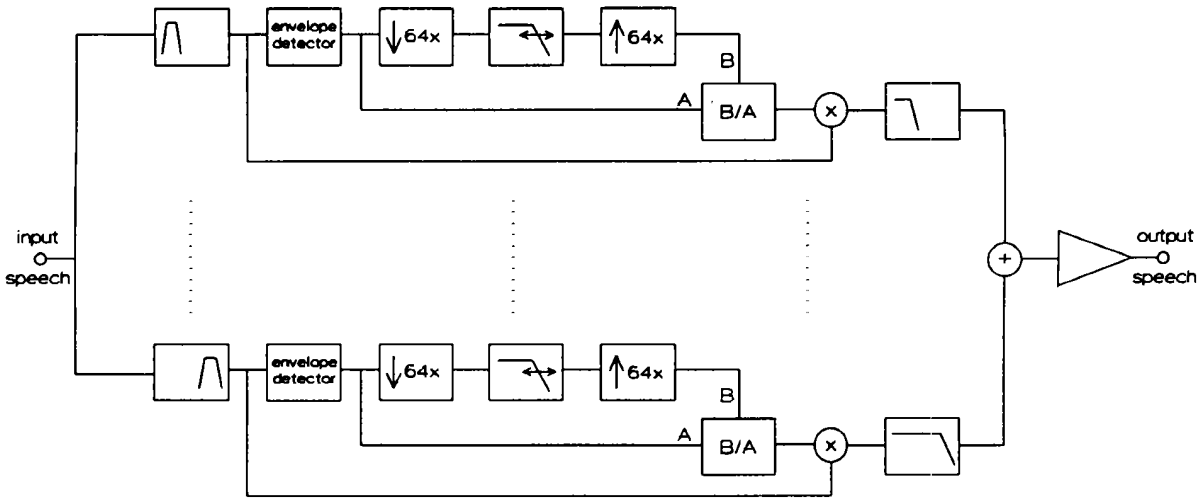

FIG. 1. Block diagram of the speech processing. The wideband input speech signal is split up into several frequency bands. For each band the amplitude envelope is determined, downsampled by a factor 64, low-pass filtered, and upsampled. The new band signal is obtained by modulating the original band signal according to the modified envelope. Each new band signal is low-pass filtered to eliminate undesired high-frequency noise. After adding all modified bands, the wideband signal is rescaled to match the rms level of the original speech.

filters are at least 80 dB/oct. The amplitude envelope from each band is computed by means of a Hilbert transform (Rabiner and Gold, 1975). The next step is a low-pass filtering of this original envelope to get the modified envelope. In the experiment the cutoff frequencies (−6-dB points) of these low-pass FIR filters ranged from 0.5 to 64 Hz. These filters have to be sufficiently steep at very low cutoff frequencies and, at the same time, manageable for implementation, i.e., their impulse response should have only a few hundred points instead of several thousand. In order to achieve that, the envelope is downsampled by a factor 64 (to a sampling rate of 244 Hz) before filtering.[1] The slopes of the low-pass filters were empirically set to approximately −40 dB/oct, so that the modified envelope will not become negative. After filtering, the envelope is unsampled again. The modified band signal is obtained by multiplying the original band (fine structure) by the ratio of the filtered envelope and the original envelope at each corresponding point in time.

As a result of the envelope filtering (especially at low cutoff frequencies), parts of the original band signal having low amplitude are amplified in the modified band signal, particularly just before and after periods with a high amplitude. These modified parts sometimes contain amplified quantization noise of high frequencies (not belonging in the frequency band), causing sharp, clicking sounds. To eliminate these, the modified band signal is low-pass filtered, using a FIR filter with a cutoff frequency 5% above the upper cutoff frequency of the corresponding bandpass filter.[2] Finally, all modified band signals are added and the level of the new wideband signal is adjusted to have the same rms as the original input signal.

All signal manipulations are performed (non-real-time) on an Olivetti PCS 286 computer, using an OROS-AU21 card with TMS320C25 signal processor. Figure 2 shows an example of the various stages during the processing of one $\frac{1}{4}$-oct band of a short sentence.

II. EXPERIMENT 1: SENTENCE INTELLIGIBILITY

A. Stimuli, design

The first experiment was set up in order to assess the contribution of various modulation frequencies to the in-

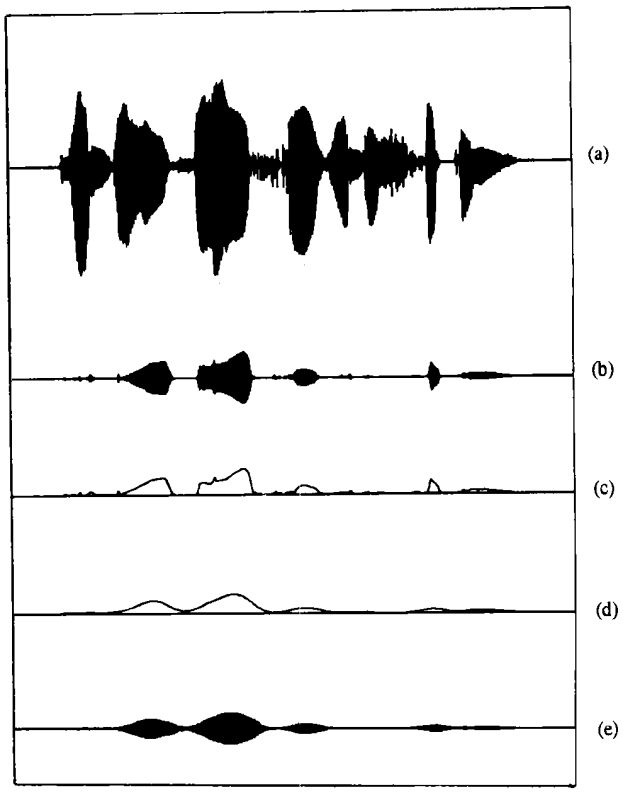

FIG. 2. Example of temporal smearing for one sentence in a single frequency band. (a) original wideband signal; (b) $\frac{1}{4}$-oct band signal (283–336 Hz); (c) amplitude envelope of (b); (d) envelope (c) low-pass filtered with 4-Hz cutoff frequency; (e) resulting modified band signal, amplitude modulated according to (d).

telligibility of sentences. Ten lists of 13 everyday Dutch sentences of eight to nine syllables read by a female speaker were used (Plomp and Mimpen, 1979). For the SRT measurements to be described below, a masking noise with the same spectrum as the long-term average of the 130 sentences was used.

Ten processing conditions were investigated. In the experimental conditions the envelope in each frequency band was low-pass filtered at nine different cutoff frequencies: 0 Hz (no modulations left), $\frac{1}{2}$, 1, 2, 4, 8, 16, 32, or 64 Hz. The 0-Hz condition was obtained by using the mean amplitude of each band for the entire duration of the utterance. An additional tenth condition, in which sentences were processed without modifying the envelope, acted as control condition. The number of bands on which envelope filtering took place was varied as well. The sentences were processed in 24 $\frac{1}{4}$-oct, 12 $\frac{1}{2}$-oct, or 6 1-oct bands (hereafter referred to as bandwidth), covering the range from 100 to 6400 Hz. The $\frac{1}{4}$-oct bandwidth was chosen because it is just smaller than the ear's critical bandwidth; $\frac{1}{2}$- and 1-oct processing bands were investigated in order to see whether this would be less detrimental for the intelligibility.

Measuring the SRT in noise requires sentences to be completely intelligible if presented at a comfortable level in quiet. This requirement was not met for cutoff frequencies up to 2 Hz (processed in $\frac{1}{4}$-oct bands), as was found in a pilot experiment in which sentences were presented at 70 dB(A) in quiet. Therefore, sentences in the 0-, $\frac{1}{2}$-, 1-, and 2-Hz conditions were presented in quiet and the number of correctly received sentences was scored. This will be referred to as the SIQ (sentence intelligibility in quiet) experiment.

B. Subjects

Subjects were 36 normal-hearing students and employees of the Free University, whose ages ranged from 18 to 30. All had pure-tone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz. They were divided into three groups of twelve, each group receiving the ten conditions for one of the three processing bandwidths.

C. Procedure

From the ten lists of 13 sentences, six were used in the SRT experiment and four in the SIQ experiment. Lists were presented in a fixed order. The sequence of the conditions was varied according to a digram-balanced Latin square—6×6 for the SRT and 4×4 for the SIQ experiment—to avoid order and list effects. Having twelve subjects in a group, each sequence was presented to two subjects in the SRT experiment and to three subjects in the SIQ experiment.

For the SIQ experiment, all four lists were presented in quiet, at an average level of 70 dB(A). Every sentence was presented once, after which a subject had to reproduce it as accurately as possible. Subjects were encouraged to respond freely, even if they would only understand fragments of the sentence. A response was scored only if the complete sentence was reproduced correctly.

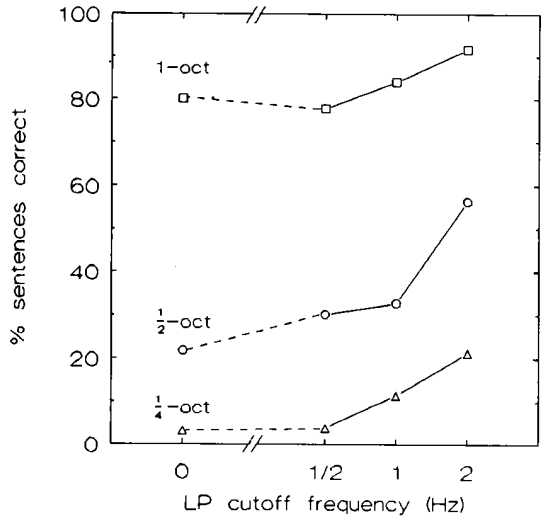

FIG. 3. Mean score of sentences in quiet as a function of cutoff frequency, with processing bandwidth as parameter.

For the SRT experiment, the level of the masking noise was fixed at 70 dB(A) for each condition; the level of the sentences was changed according to an up–down adaptive procedure (Plomp and Mimpen, 1979). The first sentence in a list was presented at a level below the reception threshold. This sentence was repeated, each time at a 4-dB higher level, until the listener could reproduce it without a single error. The remaining 12 sentences were then presented only once, in a simple up–down procedure with a step size of 2 dB. The average signal-to-noise ratio for sentences 4–13 was adopted as the SRT for that particular condition.

The sentences in both the SIQ and the SRT experiment were presented monaurally through earphones (Sony MDR-CD999) at the ear of preference in a soundproof room. Before the actual tests, a list of 13 sentences pronounced by a male speaker was presented, in order to familiarize the subjects with the procedure. For the SIQ experiment this list consisted of sentences in the 1-Hz condition; for the SRT experiment another list in the 16-Hz condition was used. All subjects started with the SIQ experiment, continuing with the SRT experiment after a short break. The entire session lasted about 1 h per subject.

D. Results and discussion

The mean results of the SIQ experiment for the four filtering conditions in the three bandwidths are plotted in Fig. 3. As expected, the overall performance improves with increasing cutoff frequency. With the $\frac{1}{4}$- and $\frac{1}{2}$-oct bands, however, the scores vary widely among subjects, resulting in large standard deviations. A two-way analysis of variance with repeated measures on the factor conditions, using arcsine transformed scores (Studebaker, 1985), revealed that both the effect of cutoff frequency and the effect

TABLE I. Mean SRT in dB with standard deviations in parentheses for the six filtering conditions in three bandwidths.

Bandwidth	Condition 4 Hz	8 Hz	16 Hz	32 Hz	64 Hz	Control
$\frac{1}{4}$-oct	−0.1(1.8)	−3.8 (0.9)	−5.0 (1.2)	−5.6 (1.1)	−5.5 (0.8)	−6.2 (0.8)
$\frac{1}{2}$-oct	0.5(2.5)	−3.3 (1.6)	−4.7 (1.4)	−4.7 (1.4)	−4.7 (1.4)	−5.8 (1.1)
1-oct	−0.7(1.4)	−2.6 (1.5)	−3.8 (1.2)	−4.4 (1.0)	−4.0 (1.6)	−5.1 (1.2)
Mean	−0.1(2.0)	−3.3 (1.4)	−4.5 (1.3)	−4.9 (1.3)	−4.7 (1.4)	−5.7 (1.1)

of bandwidth were highly significant ($p<0.001$), whereas the interaction was not. Pairwise comparisons (Tukey HSD test) of the mean scores for the three bandwidths showed the latter were all significantly different ($p<0.01$). So, for very low cutoff frequencies, intelligibility increases when the frequency bands become broader. As to the four cutoff frequencies, scores in the 2-Hz condition are significantly better than in the other three conditions ($p<0.01$); the 0-, $\frac{1}{2}$-, and 1-Hz conditions do not differ significantly.

The mean SRT for sentences in noise as a function of filtering condition and bandwidth is listed in Table I. For all bandwidths, the thresholds in the 4- and 8-Hz conditions are clearly higher than in the other conditions. A constant threshold is reached for the 16-, 32-, and 64-Hz conditions, although it is still about 1-dB higher than for the control conditions.

The raw data suggest a slightly different threshold for different processing bandwidths in all filtering conditions. Since this difference is also present in the control conditions, we inspected the stimuli more closely. It appeared that processing in $\frac{1}{4}$- and $\frac{1}{2}$-oct bands had resulted in slightly tilted long-term spectra. Because the same unprocessed masking noise was used in all conditions, actual signal-to-noise ratios were slightly more favorable for narrow processing bands. Therefore, we expressed the SRT in all experimental conditions relative to the mean SRT in the control condition for the corresponding bandwidth. This relative SRT is shown in Fig. 4. Although it would have been better if we had processed the noise in the same way as the sentences, the results are not essentially affected.

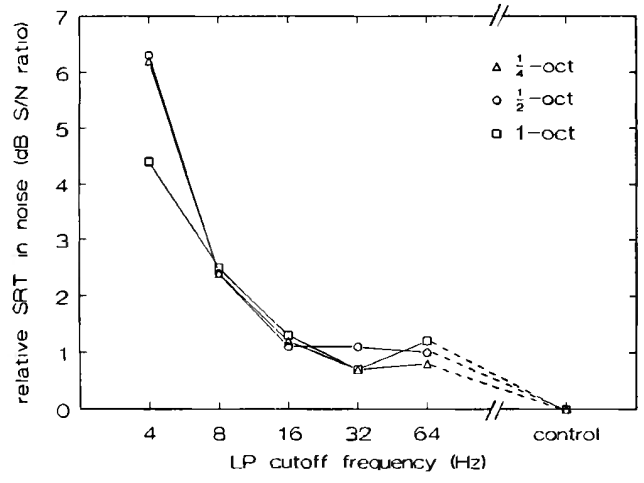

FIG. 4. Mean relative SRT in noise as a function of cutoff frequency, with processing bandwidth as parameter.

A two-way analysis of variance on the relative SRT with repeated measures on the factor conditions revealed a highly significant effect of cutoff frequency ($p<0.001$); there were no bandwidth or interaction effects. *Post hoc* tests (Tukey HSD) showed that increasing the cutoff frequency from 4 to 8 Hz and from 8 to 16 Hz significantly improved the subjects' performance ($p<0.01$), while the further increase to 16, 32, and 64 Hz did not. The 1-dB threshold shift between the 64 Hz and control condition is just significant ($p<0.05$). This difference can be explained from our envelope definition, viz. the Hilbert envelope. As a consequence, very high modulation frequencies can be found in the output of broad frequency bands (e.g., modulations by the pitch periods). Low-pass filtering causes these details to disappear. Although modulations above 64 Hz are very small, omitting them apparently results in a 1-dB increase of the SRT.

The results of the experiments indicate that the intelligibility increases progressively with low-pass cutoff frequency up to about 16 Hz. In other words, modulation frequencies in the amplitude envelope above 16 Hz (with lower modulation frequencies present) do not really contribute to understanding ordinary sentences. For cutoff frequencies above 4 Hz, the intelligibility appears to be independent of the processing bandwidth; the beneficial effect of a larger bandwidth is only demonstrated for cutoff frequencies below 4 Hz. An explanation for this, considering there is no 100% correlation between frequency bands, could be as follows. For the limit case of 0 Hz, the only information is in the variations of the energy distribution *within* each processing band. This can be referred to as spectral micro-information. This micro-information is quite useful within the 1-oct bands (and to a lesser extent in the $\frac{1}{2}$-oct bands), since it can be resolved by the ear, whereas it cannot be resolved within the $\frac{1}{4}$-oct bands. When increasing the envelope cutoff frequency, more and more of the spectral macro-information becomes available (variations in the overall spectral shape), dominating the role of the micro-information. The breaking-point appears to be around a cutoff frequency of 4 Hz; perhaps because then the information on the word/syllable structure is sufficiently present.

With respect to the infinite peak-clipping question raised in the introduction, 100% compression of the temporal modulations (0-Hz condition) in 1-oct bands still

yields a score of 80% correctly received sentences. The score drops dramatically to 22% and 3% if the number of frequency bands is doubled or quadrupled, respectively. This demonstrates clearly that zero-crossing information alone (no modulations in $\frac{1}{4}$-oct bands) is insufficient for speech intelligibility.

The above experiments do not answer the question of how individual phonemes are affected by temporal smearing. Therefore, vowel and consonant identification were studied in a second experiment.

III. EXPERIMENT 2: CONSONANT AND VOWEL IDENTIFICATION

A. Stimuli, design

The speech material consisted of two types of meaningless syllables. CVC syllables were used for the identification of initial consonants (C_i), vowels (V), and final consonants (C_f); VCV syllables were used for the identification of medial consonants (C_m).

The CVC words were obtained from 24 existing lists of 12 different syllables each (Bosman, 1989), read by the same speaker who had produced the sentences. Vowels and consonants were chosen from three sets of 12 phonemes, all of which appeared once in a list. C_i was choosen from /b, d, χ, h, j, k, l, n, t, v, w, z/, V from /ɑ, a, ɛ, e, I, i, ɔ, o, ø, u, ɑu, ɛi/, and C_f from /f, χ, j, k, l, m, n, ŋ, p, s, t, w/.

For the identification of C_m, we used VCV syllables spoken by the same speaker. Each syllable consisted of one of 16 consonants /b, d, f, χ, h, j, k, l, m, n, p, s, t, v, w, z/ surrounded by one of four vowels /a, i, u, œ/, yielding a total of 64 different syllables. The four vowels were selected to induce different envelope courses just before and after the consonant. Both CVC and VCV syllables were digitized with 16-bits resolution at a sampling rate of 15 625 Hz. They were normalized for rms level.

For all syllables the smearing was performed in 24 $\frac{1}{4}$-oct bands. There were six experimental conditions, cutoff frequencies 0, 2, 4, 8, or 16 Hz and a control condition. The choice of the filtering conditions was based on the results of the SRT experiments, viz. normal intelligibility (control and 16 Hz), reduced intelligibility (8 and 4 Hz), and low intelligibility (2 and 0 Hz). For the sake of convenience, we will write the filtering condition in parentheses following the set of phonemes to be identified. For example, $C_i(2)$ stands for initial consonants in the 2-Hz condition, V(con) stands for vowels in the control condition.

From the original 24 lists of 12 CVC syllables we made 36 randomized lists of 50 syllables. The first two syllables were copies of the last two and acted as dummy trials, so that there were 48 test stimuli in a list for the identification of C_i, V, and C_f. These lists were constructed in such a way that each initial consonant, vowel, and final consonant appeared four times in different contexts.

From the original four lists of 16 VCV syllables we constructed 36 randomized lists of 66 syllables for the identification of C_m. Each list contained all 64 VCV syllables and again the first two syllables were copies of the last two.

B. Subjects

Subjects were 24 normal-hearing students of the Free University, whose ages ranged from 19 to 28. All had puretone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz. They were divided into two groups of twelve, one group for the identification of C_i and C_f, the other for the identification of V and C_m.

C. Procedure

For both identification of C_i and C_f and of V and C_m, the 36 lists were assigned to the filtering and control conditions according to a 6×6 digram-balanced Latin square to avoid effects of measurement order. With twelve subjects per group, each sequence of conditions was presented to two subjects.

All stimuli were presented in quiet, monaurally through headphones (Sony MDR-CD999) at the ear of preference in a soundproof room. The level of presentation was 70 dB(A). The subject was seated in front of a video monitor and a response box. On this box a number of buttons (12 in case of vowels and 17 in case of consonants) were labeled with phonemes, in orthographic notation. After presentation of a stimulus (only once), the subject could take as much time needed to give a response by pressing one of the labeled buttons. The response was displayed on the monitor. After a response was given, there was a 1-s interval before the next stimulus was presented. If a mistake was made (which occurred only sporadically), the 1-s interval could be used to correct the response.

The order of the tests (i.e., whether the subject started with the C_i or C_f part, or with the V or C_m part) was counterbalanced. Before each test, two lists of 20 stimuli in the 4- and 2-Hz conditions were presented to familiarize the subjects with the experimental task. Subjects participating in the C_f test were told that the CVC syllables followed the Dutch phonological rules, e.g., they would not end in /b/ or /v/. An entire session of two tests lasted about 1 h per subject.

D. Results and discussion

In total, 48 identifications (12 subjects$\times$4 utterances) were obtained for each vowel and consonant in each condition. Due to an error in the automatic registration of the responses, however, we could not recover two subjects' V(con), one subject's V(16), and one subject's C_m(con). So for these conditions we only had 40, 44, and 44 identifications per phoneme, respectively. The mean score for consonants and vowels in each filtering condition is plotted in Fig. 5. As an example, confusion matrices for the four phoneme sets in the 2-Hz condition are given in the Appendix. A two-way analysis of variance on the arcsine transformed scores (Studebaker, 1985) with repeated measures on the factor condition showed significant effects of phoneme set, filtering condition, and interaction ($p < 0.001$). Because of the significant interaction, separate analyses were carried out for each of the six conditions and for each of the four phoneme sets. The results of these

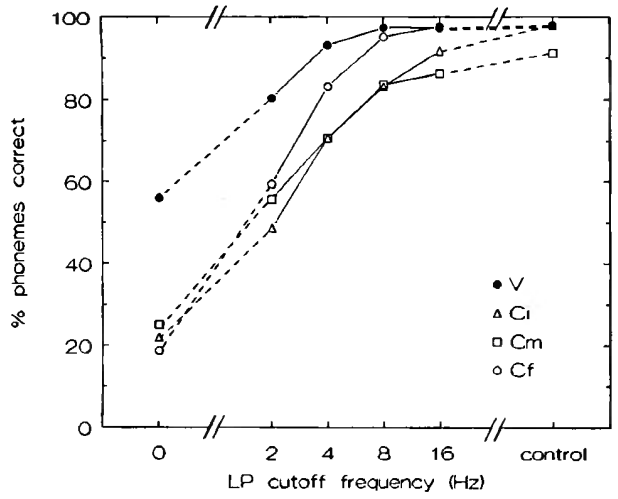

FIG. 5. Overall vowel- and consonant-identification score as a function of cutoff frequency, with phoneme type as parameter.

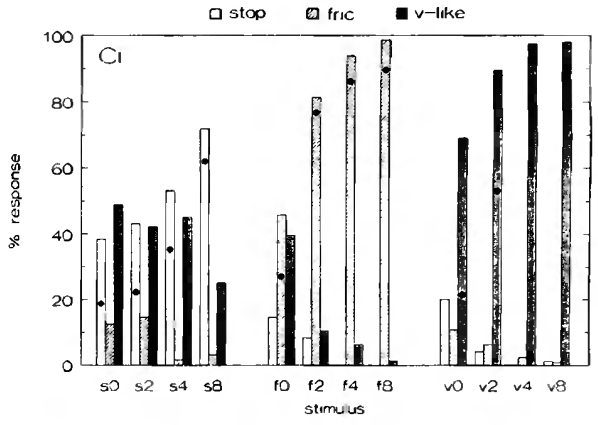

FIG. 6. Distribution of the responses for the initial consonants across the three categories stop, fricative, and vowel-like as a function of stimulus category and cutoff frequency ($s0$=stop/0-Hz condition, $s2$=stop/2-Hz condition, $f4$=fricative/4-Hz condition, $v8$=vowel-like/8-Hz condition, etc.). The black dots indicate the percentage of correctly identified consonants per category/condition.

analyses and subsequent *post hoc* tests (Tukey HSD) can be summarized as follows. The mean recognition improves for all phoneme sets as the cutoff frequency increases up to 8 Hz [$p<0.01$ for all pairwise comparisons, except for the V(4) and V(8) comparison, where $p<0.05$]. Only the initial consonants benefit from a further increase to 16 Hz ($p<0.01$). As to the different phoneme sets, in the 0-, 2-, and 4-Hz conditions the vowels are identified significantly better ($p<0.01$) than the consonants; in the 8- and 16-Hz conditions this still holds for the initial and medial consonants. The fact that the scores for C_m(con) are significantly lower than for C_i(con), V(con), and C_f(con) ($p<0.01$) may be due to the rather emphatic articulation of some medial consonants. This is particularly illustrated by the number of /w/-/v/ and to a lesser extent /z/-/s/ confusions, which also occur in the other conditions.

The results indicate that the vowels generally suffer less from temporal smearing than the consonants. At a cutoff frequency of 0 Hz, 56% of the vowels are still identified correctly, more than twice the average consonant score. An obvious explanation for the high vowel scores is that (the majority of) the momentary spectral information is well preserved because of the relatively long duration and high amplitude. After temporal envelope filtering, the distribution of the energy over most $\frac{1}{4}$-oct bands in the center of the vowel has not changed much (see Fig. 9).

Because of the very low error rates for V(4), V(8), and V(16), we will restrict ourselves to V(0) and V(2) for a discussion of the vowel confusions. The greater part of the errors can be attributed to two factors: Diphthong confusions (/ɛi/-/ɛ/ and /ɑu/-/ɑ/ or /ɑu/-/a/) and long–short/short–long confusions (/a/-/ɑ/, /e/-/I/, and /o/-/ɔ/). Of all errors in the 0-Hz condition (253), these factors account for 33% and 28%, respectively. In the 2-Hz condition (113 errors) these figures are 42% and 37%, respectively. The percentages in the 0-Hz condition are slightly less, since there is also a tendency to respond to /ø/ (15% of all errors), which is the most neutral vowel in the set. It is clear that the rapid spectral changes in a diphthong (the /u/ and /i/ states take only about a quarter of the entire duration of /ɑu/ and /ɛi/, respectively) are not modeled properly for these low cutoff frequencies. As to the confusions among long–short vowel pairs, the blurring of the temporal structure makes it difficult to perceive the beginning and/or end of a vowel and hence the vowel duration, so that listeners have to rely more on the spectral contents, which often leads to the observed confusions. It is also possible that smearing causes a reduction in the perceptually important vowel-inherent spectral change, i.e., slowly varying changes in formant frequencies. In the absence of these dynamic spectral changes long–short confusions may occur [cf. Nearey and Assman (1986), who found this for isolated monophthongs in English].

For an evaluation of the C_i, C_m, and C_f scores, the set of consonants was divided into three subsets (cf. Steeneken, 1992):[3] stops (/t, k, p, b, d/), fricatives (/f, s, χ, v, z/), and vowel-like consonants (/m, n, ŋ, l, w, j, h/). From the original confusion matrices, the data were collapsed into 3×3 matrices, which grouped the consonants according to the aforementioned categories. Figures 6–8 show the distribution of the responses across the three cat-

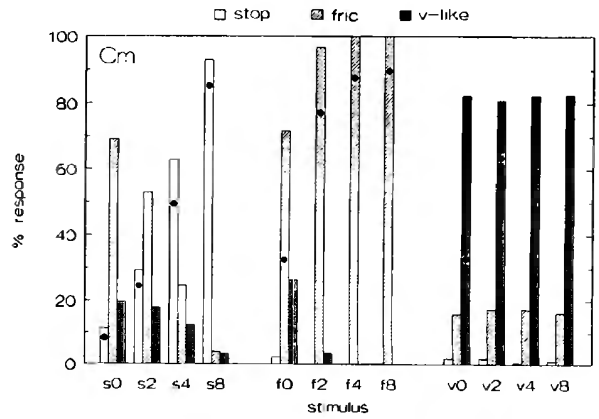

FIG. 7. As Fig. 6, for the medial consonants.

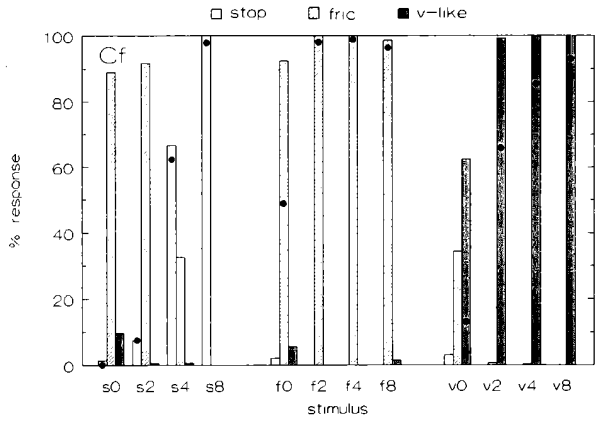

FIG. 8. As Fig. 6, for the final consonants.

egories in the 0-, 2-, 4-, and 8-Hz conditions for C_i, C_m, and C_f, respectively.[4] Within each category the percentage correct consonants is indicated by a black dot. As to the latter, the scores for stops are clearly lower than for fricatives and vowel-likes [except for $C_m(8)$ and $C_f(8)$], irrespective of consonant position. In general, stops are more affected by smearing because of their short duration.

As to the distribution of the responses, the patterns for the fricatives and vowel-likes are rather similar across consonant positions: In either case stops are rarely or never used as a response alternative; in the 2-, 4-, and 8-Hz conditions, the majority of the responses fall in the correct category (the C_m vowel-likes show a constant 17% responses for the fricatives, mainly caused by /w/-/v/ and /h/-/v/ confusions, probably as a result of emphatic articulation).

The confusions of the stop consonants show more variation, and depend on their position in the syllable. In initial position, stops are confused with other stops (place errors) and with vowel-likes. In addition to the bias toward /h/ responses in the 0- and 2-Hz conditions, we find /b/-/w/ and /d/-/w/ confusions. The perception of /w/ for /b/, at least, could be expected to result from temporal smearing. In medial and final position, stops are mainly confused with fricatives. In the 0-Hz condition the responses are rather scattered, but in the 2- and 4-Hz conditions we can distinguish some confusions that could be expected: /b/-/v/, /b/-/w/, /t/-/s/, /k/-/χ/, and /p/-/f/ (the last three especially in final position).

IV. GENERAL DISCUSSION

In the experiments described above, we tried to assess the contribution of temporal modulations to intelligibility and identification. The manipulations of the speech signal are quite artificial and are not intended to model reduced temporal resolution by hearing-impaired listeners. Nor do they reflect any disturbances that may occur in communication channels, except maybe for early channel vocoders, as mentioned in the Introduction. The method provides information on the importance of temporal modulations without disturbing the fine structure, as is the case with noisy or reverberated speech.

Filtering of the envelope generally causes low amplitude regions to be amplified and high amplitude regions to be attenuated, yielding smaller modulation depths. Rapid changes in amplitude are flattened. As a result of this, short-duration phonemes (stops) will be more affected than long-duration phonemes (vowels, fricatives, vowel-like consonants), as was observed in the second experiment. By decreasing the cutoff frequency of the envelope filter, the amplitude variations in each processing band get smaller, eventually leading to a stationary sound. As noted earlier, for narrow processing bands this means that spectral content will ever more resemble the long-term average spectrum of the utterance. As an illustration, Fig. 9 (lower part) shows a narrow-band spectrogram of a sentence processed in $\frac{1}{4}$-oct bands in the 4-Hz condition. One can particularly see the fading of the stops and the filling of the "silent" parts before the voiceless stop-bursts with amplified noise.

It is worthwhile to view the results of the first experiment in the light of the MTF (Houtgast and Steeneken, 1985). The MTF gives the extent to which the frequency components of the *intensity* envelope are transferred. Since we filtered the *amplitude* envelope, the reduction of amplitude modulations should be translated into a reduction of intensity modulations.[5] It turns out that the filter slope in the intensity domain remains unchanged (approximately −40 dB/oct), but the point at which the modulation is reduced to half (the cutoff frequency) is about $\frac{1}{3}$ oct higher.

An interesting point is the relation between the intelligibility scores in the first experiment and the speech-transmission index (STI), based on the MTF. The STI can be calculated from the relative reduction of intensity modulations in 14 $\frac{1}{3}$-oct intervals of modulation (ranging from 0.63 to 12.5 Hz) within 1-oct frequency bands. The STI is a numerical index between 0 and 1, which bears a linear relation to signal-to-noise ratios from −15 to +15 dB (for a review of the STI concept, see Houtgast and Steeneken, 1985). We computed the STI for each of the nine experimental conditions.[6] The results are given in Table II. The STI values apply to the "clean" processed signal, i.e., without noise. In order to get an estimate of the SRT for each filtering condition, the signal-to-noise ratio which would give a STI of 0.33 was computed. A STI of 0.33 corresponds to a signal-to-noise ratio of −5 dB, which is a reasonable threshold for normal (unprocessed) speech. The estimated SRTs are shown in the second row of Table II. It is clear that no SRT could be computed for the 0-, $\frac{1}{2}$-, and 1-Hz conditions, since in these cases the STI is lower than 0.33. However, one can argue that these computed values are rather low, since an 80% correct score for sentences (0-Hz condition, 1-oct bandwidth) would already yield a STI of about 0.4. According to the STI in Table II, measuring the SRT in the 2-Hz condition would be possible; this seems plausible, given the 92% intelligibility score for sentences in quiet (see Fig. 3, 1-oct bandwidth). For the other conditions, the estimated SRT corresponds well to the actually measured SRT in the first experiment (see Table I, 1-oct bandwidth); although the estimated SRTs are systematically lower, the deviation is maximally 1.5 dB

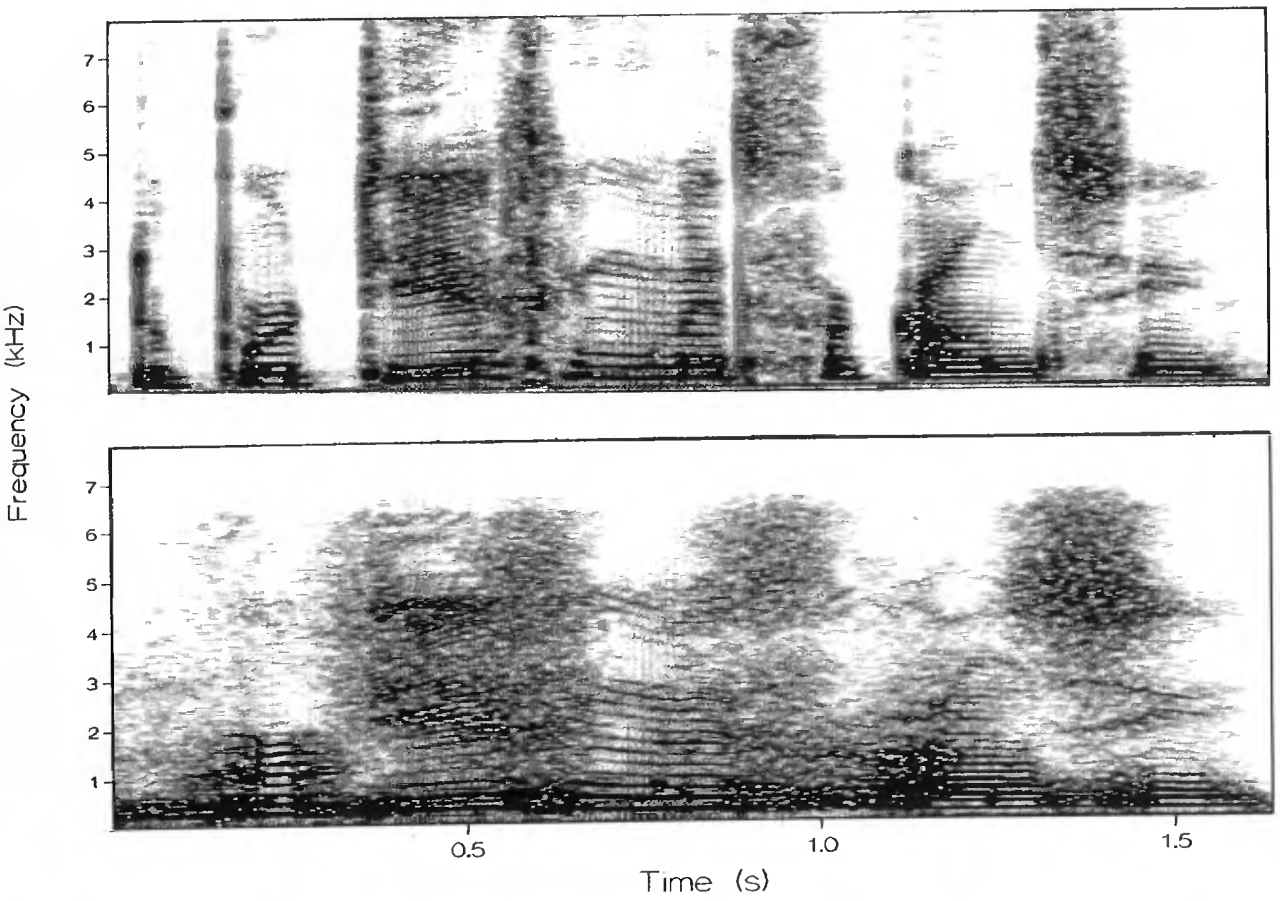

FIG. 9. Narrow-band spectrograms of the sentence "de portier ging met vakantie" (the porter went on holiday), original (above) and smeared in $\frac{1}{4}$-oct bands and 4-Hz cutoff frequency (below).

(in the 8-Hz condition). Since we found no effect of processing bandwidth for cutoff frequencies above 4 Hz, the estimated SRTs are also comparable with the measured SRTs in $\frac{1}{4}$- and $\frac{1}{2}$-oct bands (the deviation for the 4-Hz condition is about 2 dB). This is consistent with the octave-band specific STI concept, which does not account for effects within $\frac{1}{2}$- or $\frac{1}{4}$-oct bands (there is a bandwidth dependence for cutoff frequencies below 4 Hz, though). For cutoff frequencies of 16 Hz and higher, the good correspondence between the estimated and measured SRT is not surprising, since the modulations involved in the calculation of the STI do not exceed 12.5 Hz. In summary, considering that the STI was developed for time-domain distortions like reverberation, the SRT estimates down to 4 Hz are reasonable.

The results of the second experiment indicate that consonants are affected most by temporal smearing. Gelfand and Silman (1979) investigated the effect of reverberation ($T=0.8$ s) upon consonant recognition in CVC syllables. As far as the reduction of fast modulations is concerned, this corresponds roughly to a situation between our 4- and 8-Hz condition. Gelfand and Silman found that initial consonants are on average less affected than final consonants, whereas our study does not show this difference. In reverberant speech, final consonants are masked by delayed energy of preceding segments, which does not hold for initial consonants. In our processing however, smearing of the envelope causes segments to integrate with preceding and following segments. Initial consonants will thus be corrupted by the following vowel, and will therefore also deteriorate.

As to the confusions, place of articulation errors did not occur regularly in fricatives, which seems to be in agreement with the findings of Behrens and Blumstein

TABLE II. Computed speech transmission index (STI) for the nine filtering conditions without noise. The second row (estimated SRT) gives the S/N ratio in dB that yields a STI of 0.33, corresponding to −5 dB S/N, the SRT for unprocessed speech.

	Condition								
	0 Hz	$\frac{1}{2}$ Hz	1 Hz	2 Hz	4 Hz	8 Hz	16 Hz	32 Hz	64 Hz
STI	0.00	0.09	0.25	0.47	0.68	0.88	0.98	1.00	1.00
est. SRT	···	···	···	+4.7	−1.1	−4.1	−4.9	−5.0	−5.0

(1988). Place errors were primarily found in initial stops, which may be related to a reduction of changes in the distribution of spectral energy from burst onset to vowel onset. In both natural and synthetic stop-V stimuli the relative amplitude of the burst and its (rapidly) changing spectrum have been shown to affect the perception of place of articulation (Ohde and Stevens, 1983; Lahiri *et al.*, 1984). We found manner feature recognition for fricatives and vowel-likes to be (far) less reduced and less position dependent than for stops (Figs. 6–8). The position dependence of the stop confusions (vowel-like responses for C_i and fricative responses for C_m and C_f) may be explained as follows. In initial position, smearing increases the integration of the stop with the immediately following vowel. This evokes the perception of a vowel-like consonant (strictly speaking, this accounts for the voiced stops /b/ and /d/). Final stops are more isolated from the preceding vowel, and tend to be longer in duration (word-final lengthening). Smearing will therefore not lead to integration with the preceding vowel. Because there is no speech sound following, the stop can be "spread out" entirely, resulting in a fricative perception. According to the above reasoning, smearing of the medial stops would evoke more vowel-like than fricative confusions. The opposite is true, however. The medial stops were pronounced rather emphatically, making them longer than usual, so that there was less influence of the following vowel after smearing, yielding fricative perception.

Finally, it is important to consider the effect of narrow-band smearing upon perception, compared to a wideband approach. For natural speech, Nittrouer and Studdert-Kennedy (1986) investigated the effect of interchanging the wideband amplitude envelope of /b/-V and /w/-V stimuli. They found that this had little effect on the consonant recognition (97% correct). Their results suggest that increasing the amplitude of occlusion and burst of /b/ does not automatically evoke the perception of /w/. The fact that we do find /b/-/w/ confusions with the initial consonants must be ascribed to the narrowband approach, smearing the envelopes of 24 $\frac{1}{4}$-oct bands.

V. CONCLUSIONS

The most important conclusions of this study are:

(1) Amplitude fluctuations in successive $\frac{1}{4}$-, $\frac{1}{2}$-, or 1-oct frequency bands can be limited to about 16 Hz without substantial reduction of speech intelligibility for normal-hearing listeners.

(2) Listeners can only partially understand speech in quiet when the amplitude fluctuations are limited to 2 Hz; performance improves as broader frequency bands are used. In case of 100% compression within $\frac{1}{4}$-oct bands (beyond the critical bandwidth), intelligibility drops dramatically. For envelope cutoff frequencies above 4-Hz intelligibility is independent of the processing bandwidth.

(3) SRT values obtained for envelope cutoff frequencies above 4 Hz appear to correspond well (maximal deviation of 1.5 dB) to those predicted by the speech-transmission index. However, for low cutoff frequencies (0–2 Hz) the computed STI is rather low and does not account for the observed dependence of the processing bandwidth.

(4) Phoneme identification with nonsense syllables shows that consonants are more affected by temporal smearing than vowels. Stops appear to suffer most, due to their short duration, with confusion patterns depending on the position in the syllable.

ACKNOWLEDGMENTS

This research was supported by the Linguistic Research Foundation, which is funded by the Netherlands organization for scientific research, NWO. We would like to thank the two reviewers for their helpful comments and suggestions.

APPENDIX: SUMMED CONFUSION MATRICES FROM THE VOWEL AND CONSONANT IDENTIFICATION EXPERIMENTS IN THE 2-Hz CONDITION

Summed confusion matrices are given in Tables AI through AIV.

TABLE AI. Summed confusion matrices for 12 subjects: Initial consonants.

	Stimulus/response																	
	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	3	20	5	3	·	2	·	6	2	·	·	·	·	·	·	·	7	48
k	1	20	2	2	·	·	2	3	·	·	·	1	·	1	·	2	14	48
b	·	·	·	12	·	·	·	1	4	·	·	·	·	·	23	1	7	48
d	·	1	·	7	7	1	·	1	4	2	1	·	·	2	14	1	7	48
v	·	·	1	1	·	·	·	·	33	·	·	·	·	·	9	·	4	48
z	·	·	·	·	·	1	3	·	·	42	·	2	·	·	·	·	·	48
χ	6	4	·	·	·	2	·	35	·	1	·	·	·	·	·	·	·	48
n	·	·	·	1	·	·	1	·	1	2	1	15	·	4	14	1	8	48
l	·	·	·	·	·	·	·	·	1	2	1	·	·	28	11	·	5	48
w	·	·	·	5	·	·	·	1	3	·	·	·	·	·	32	·	7	48
j	·	·	·	3	·	·	·	2	·	·	·	·	·	·	20	11	12	48
h	·	·	·	·	1	·	·	2	·	·	·	·	·	·	2	1	42	48
sum	10	45	8	34	8	6	6	51	48	49	3	18	0	35	125	17	113	576

TABLE AII. Summed confusion matrices for 12 subjects: Final consonants.

	Stimulus/response																	
	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	6	·	·	·	·	8	33	1	·	·	·	·	·	·	·	·	·	48
k	·	4	·	·	·	9	·	35	·	·	·	·	·	·	·	·	·	48
p	·	·	1	·	·	38	·	8	·	·	·	·	·	·	1	·	·	48
f	·	·	·	·	·	47	1	·	·	·	·	·	·	·	·	·	·	48
s	·	·	·	·	·	·	48	·	·	·	·	·	·	·	·	·	·	48
χ	·	·	·	·	·	1	1	46	·	·	·	·	·	·	·	·	·	48
m	·	·	·	·	·	·	·	·	·	·	25	15	3	·	5	·	·	48
n	·	·	·	·	·	1	·	·	·	·	6	25	9	3	3	·	1	48
ŋ	·	·	·	·	·	1	·	·	·	·	4	13	27	1	·	1	1	48
l	·	·	·	·	·	·	·	·	·	·	1	·	·	39	8	·	·	48
w	·	·	·	·	·	·	·	·	·	·	·	·	·	10	37	·	1	48
j	·	·	·	·	·	·	·	·	·	·	·	2	2	4	3	37	·	48
sum	6	4	1	0	0	105	83	90	0	0	36	55	41	57	57	38	3	576

TABLE AIII. Summed confusion matrices for 12 subjects: Medial consonants.

	Stimulus/response																	
	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	10	2	2	·	·	6	5	4	8	4	·	·	·	·	·	·	7	48
k	·	12	·	·	·	4	·	21	4	·	·	·	·	·	1	·	6	48
p	·	2	5	·	·	5	·	17	8	·	·	·	·	·	·	1	10	48
b	·	1	·	12	4	2	·	6	14	·	·	·	·	·	8	·	1	48
d	·	·	·	·	20	·	2	5	11	1	·	·	·	2	·	1	6	48
f	·	·	·	·	·	30	·	6	12	·	·	·	·	·	·	·	·	48
s	·	·	·	·	·	·	43	1	·	4	·	·	·	·	·	·	·	48
χ	·	·	·	·	·	3	·	33	8	·	·	·	·	·	·	·	4	48
v	·	·	·	·	·	6	·	·	38	·	·	·	·	·	·	·	4	48
z	·	·	·	·	·	·	7	·	·	41	·	·	·	·	·	·	·	48
m	·	·	·	·	·	·	·	·	1	·	25	12	·	2	8	·	·	48
n	·	·	·	·	·	·	·	·	·	·	2	35	3	7	·	·	1	48
l	·	·	·	·	·	·	·	·	·	·	·	1	·	46	1	·	·	48
w	·	·	·	2	·	1	·	1	24	·	·	·	·	·	17	·	3	48
j	·	·	·	·	·	·	·	·	1	·	·	4	·	1	1	41	·	48
h	·	·	·	3	·	1	·	7	14	·	·	·	·	·	3	·	20	48
sum	10	17	7	17	24	58	57	101	143	50	27	52	3	58	39	43	62	768

TABLE AIV. Summed confusion matrices for 12 subjects: Vowels.

	Stimulus/response												
	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	35	13	·	·	·	·	·	·	·	·	·	·	48
a	1	46	1	·	·	·	·	·	·	·	·	·	48
ɑu	19	3	24	·	·	·	2	·	·	·	·	·	48
ɛ	·	·	·	43	1	2	·	1	·	·	1	·	48
e	·	·	·	·	42	·	·	4	1	·	·	1	48
ɛi	·	·	·	23	·	24	·	·	·	·	1	·	48
ø	·	·	·	·	·	·	48	·	·	·	·	·	48
I	·	·	·	·	2	·	1	44	1	·	·	·	48
i	·	·	·	1	·	·	·	2	45	·	·	·	48
ɔ	1	·	·	·	·	·	2	1	·	24	15	5	48
o	·	·	·	·	·	·	·	·	·	7	41	·	48
u	·	·	·	·	·	·	1	·	·	·	·	47	48
sum	56	62	25	67	45	26	54	52	47	31	58	53	576

[1]A factor 64 was chosen for practical reasons. The OROS-AU21 signal processing card enables the use of fast decimator and interpolator routines by a factor 4, with automatic filtering. By calling these routines three times in a row, a factor of 64 is obtained. The sampling rate is then brought down to 244 Hz, which enables the implementation of sufficiently steep low-pass FIR filters with cutoff frequencies as low as 0.5 Hz.

[2]Formally, subsequent low-pass filtering of the fine structure may restore the original envelope to some extent, depending on the envelope cutoff frequency and the processing bandwidth. However, it was checked that this had no noticeable influence. In the worst case of a completely flat envelope within a 1-oct band (in which case there is a maximum difference from the original envelope), dominant modulations were still sufficiently attenuated by about 33 dB.

[3]The subdivision of the stimuli is based on Steeneken's study (Chapter 3) on the perceptual similarity of phonemes at various transmission conditions (among which distortion in the time domain).

[4]The percentages are based on the total number of stimuli within each category. The total number of stops, fricatives, and vowel-likes are for C_i: 192, 144, and 240; for C_m: 240, 240, and 288; for C_f: 144, 144, and 288, respectively.

[5]In order to get the modulation transfer in the intensity domain, the following procedure was undertaken: The modulation spectrum of the squared original amplitude envelope was compared with the modulation spectrum of the squared filtered amplitude envelope (for all cutoff frequencies used in the experiment). The original amplitude envelope was taken of a 30-s 1-oct filtered speech fragment (viz. 13 concatenated sentences used in the first experiment). The MTF was obtained by computing the ratio between the original and the modified spectrum.

[6]The calculations were verified by actual measurement of the STI with the so-called STITEL procedure (cf. Steeneken, 1992, pp. 133–139). Differences between the computed and measured STI values for the same condition were within 0.04.

Behrens, S., and Blumstein, S. E. (**1988**). "On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants," J. Acoust. Soc. Am. **84**, 861–867.

Bosman, A. J. (**1989**). "Speech perception by the hearing impaired," Ph.D. dissertation, University of Utrecht.

Duquesnoy, A. J., and Plomp, R. (**1980**). "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," J. Acoust. Soc. Am. **68**, 537–544.

Festen, J. M., and Plomp, R. (**1981**). "Relations between auditory functions in normal hearing," J. Acoust. Soc. Am. **70**, 356–369.

Flanagan, J. L. (**1972**). *Speech Analysis, Synthesis, and Perception* (Springer-Verlag, Berlin), 2nd ed., Chap. 8, pp. 323–330.

Freyman, R. L., Nerbonne, G. P., and Cote, H. A. (**1991**). "Effect of consonant-vowel ratio modification on amplitude envelope cues for consonant recognition," J. Speech Hear. Res. **34**, 415–426.

Gelfand, S. A., and Silman, S. (**1979**). "Effects of small room reverberation upon the recognition of some consonant features," J. Acoust. Soc. Am. **66**, 22–29.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MFT concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.

Houtgast, T., and Verhave, J. A. (**1991**). "A physical approach to speech quality assessment: correlation patterns in the speech spectrogram," in *Proceedings of the 3rd European Conference on Speech Communication and Technology*, edited by G. Pirani, Genova, September 1991 (IIC, Genova, Italy), Vol. 1, pp. 285–288.

Lahiri, A., Gewirth, L., and Blumstein, S. E. (**1984**). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," J. Acoust. Soc. Am. **76**, 391–404.

Neary, T. M., and Assman, P. F. (**1986**). "Modeling the role of inherent spectral change in vowel indentification," J. Acoust. Soc. Am. **80**, 1297–1308.

Nittrouer, S., and Studdert-Kennedy, M. (**1986**). "The stop-glide distinction: Acoustic analysis and perceptual effect of variation in syllable amplitude envelope for initial /b/ and /w/," J. Acoust. Soc. Am. **80**, 1026–1029.

Ohde, R. N., and Stevens, K. N. (**1983**). "Effect of burst amplitude on the perception of stop consonant place of articulation," J. Acoust. Soc. Am. **74**, 706–714.

Plomp, R. (**1984**). "Perception of speech as a modulated signal," in *Proceedings of the 10th International Congress of Phonetic Sciences*, edited by A. Cohen and M. P. R. van de Broecke (Foris, Dordrecht), pp. 29–40.

Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J. Speech Hear. Res. **29**, 146–154.

Plomp, R. (**1988**). "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," J. Acoust. Soc. Am. **83**, 2322–2327.

Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.

Rabiner, L. R., and Gold, B. (**1975**). *Theory and Application of Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ), Chap. 2, pp. 70–72.

Rodenburg, M. (**1977**). "Investigation of temporal effects with amplitude modulated signals," in *Psychophysics and Psychology of Hearing*, edited by E. F. Evans and J. P. Wilson (Academic, London, UK), pp. 429–437.

Steeneken, H. J. M. (**1992**). "On measuring and predicting speech intelligibility," Ph.D. dissertation, University of Amsterdam.

Studebaker, G. A. (**1985**). "A "rationalized" arcsine transform," J. Speech Hear. Res. **28**, 455–462.

ter Keurs, M., Festen, J. M., and Plomp, R. (**1992**). "Effects of spectral smearing on speech reception. I," J. Acoust. Soc. Am. **91**, 2872–2880.

ter Keurs, M., Festen, J. M., and Plomp, R. (**1993**). "Effects of spectral smearing on speech reception. II," J. Acoust. Soc. Am. **93**, 1547–1552.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). "Speech waveform envelope cues for consonant recognition," J. Acoust. Soc. Am. **82**, 1152–1161.

Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.

Villchur, E. (**1989**). "Comments on 'The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function' [J. Acoust. Soc. Am. **83**, 2322–2327 (1988)]," J. Acoust. Soc. Am. **86**, 425–427.