UvA-DARE (Digital Academic Repository)

Importance of temperol-envelope cues in consonant recognition

van der Horst, R.; Leeuw, A.R.; Dreschler, W.A.

DOI

[10.1121/1.426718](https://doi.org/10.1121/1.426718)

Publication date

1999

Published in

The Journal of the Acoustical Society of America

[Link to publication](#)

Citation for published version (APA):

van der Horst, R., Leeuw, A. R., & Dreschler, W. A. (1999). Importance of temperol-envelope cues in consonant recognition. *The Journal of the Acoustical Society of America*, 105, 1801-1809. <https://doi.org/10.1121/1.426718>

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.

Importance of temporal-envelope cues in consonant recognition

René van der Horst,^{a)} A. Rens Leeuw, and Wouter A. Dreschler

Clinical and Experimental Audiology, D2, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

(Received 4 August 1997; accepted for publication 16 October 1998)

The role of different modulation frequencies in the speech envelope were studied by means of the manipulation of vowel-consonant-vowel (VCV) syllables. The envelope of the signal was extracted from the speech and the fine-structure was replaced by speech-shaped noise. The temporal envelopes in every critical band of the speech signal were notch filtered in order to assess the relative importance of different modulation frequency regions between 0 and 20 Hz. For this purpose notch filters around three center frequencies (8, 12, and 16 Hz) with three different notch widths (4-, 8-, and 12-Hz wide) were used. These stimuli were used in a consonant-recognition task in which ten normal-hearing subjects participated, and their results were analyzed in terms of recognition scores. More qualitative information was obtained with a multidimensional scaling method (INDSCAL) and sequential information analysis (SINFA). Consonant recognition is very robust for the removal of certain modulation frequency areas. Only when a wide notch around 8 Hz is applied does the speech signal become heavily degraded. As expected, the voicing information is lost, while there are different effects on plosiveness and nasality. Even the smallest filtering has a substantial effect on the transfer of the plosiveness feature, while on the other hand, filtering out only the low-modulation frequencies has a substantial effect on the transfer of nasality cues.

© 1999 Acoustical Society of America. [S0001-4966(99)00302-1]

PACS numbers: 43.66.Mk, 43.71.Es [JWH]

INTRODUCTION

Speech intelligibility can be related to the temporal envelope of the acoustical speech signal (Houtgast and Steeneken, 1985). When analyzing the temporal domain of a speech signal, Rosen (1992) classified the temporal structure of speech into three categories of speech cues: envelope cues, periodicity cues, and temporal fine-structural cues. This classification is based on the time-scale of the dominant temporal structure in each category. Envelope cues consist of the frequencies 2 to 50 Hz, and are able to transmit the following phonetic and prosodic information: segment duration, rise-fall time (which can convey consonantal manner information), the presence or absence of voicing, syllabification, and stress.

Periodicity cues—containing the frequencies from 50 to 500 Hz—signal the strength and frequency of vocal fold vibrations, and therefore can convey phonetic information about consonant voicing and manner, as well as prosodic information about intonation and stress.

Temporal fine-structure. The periodicity of the temporal fine-structure ranges from 600 Hz to 10 kHz, and conveys information about the spectral distribution of energy of the speech signal. These cues are essentially the same as the spectral envelope structure, which cues consonant place and vowel quality.

A. Information contained in the speech envelope

Van Tasell *et al.* (1987) found that consonants can be recognized to some extent by the wideband speech envelope alone. In these experiments, they used noise signals modu-

lated with the temporal envelope of the wideband speech signal to be certain that the only phonetic information transferred is that contained in the speech temporal envelope. In this way, a score of 22-percent-transferred information was reached when the bandwidth of the temporal envelope was reduced to contain only the modulation frequencies up to 20 Hz. The transferred information in the 200-Hz and 2000-Hz conditions was 29% and 35%, respectively.

Another approach was followed by Drullman *et al.* (1994). They first filtered the speech signal in frequency bands, and kept the original speech carrier. The frequency bands had a width of a quarter of an octave, half an octave, or one octave. They reduced the information in each temporal envelope by low-pass filtering, and modulated the carrier of each original speech band with this filtered envelope, in order to leave the temporal fine-structure intact. In this way, they were able to assess the relative importance of the fast and slow envelope variations. It was concluded that the envelope modulations above 16 Hz could be filtered out without reducing the speech intelligibility. But if all temporal variations were filtered out, the speech reception fell to a score of about 20 percent for consonant recognition. So, cues other than temporal envelope can account for about only 20% of a speech intelligibility score.

An important observation of this study was that the frequency width of the processing bands had an influence on intelligibility. For frequency bands wider than a quarter of an octave, higher scores were measured, and it was argued that subjects were able to resolve spectral information within each processing band and use that in the speech-recognition task. The envelope of a wideband signal is the sum of multiple narrow-band envelopes (i.e., the envelopes in every critical band). The auditory system can make use of multiple

^{a)}Electronic mail: R.vanderHorst@amc.uva.nl

envelopes within such a wideband signal. Therefore, filtering out the envelope modulations of a wideband signal does not guarantee that the narrow-band envelopes are filtered in the same way.

Shannon *et al.* (1995) combined aspects of the stimuli of Van Tasell *et al.* (1987) and Drullman *et al.* (1994). They used noise as the carrier, but extracted the envelope from the speech signal in one to four processing bands. The number of processing bands was an experimental variable. Consonant recognition increased from about 50% when using only one processing band, to 90 percent when using four bands. When envelope fluctuations were low-pass filtered at 16 Hz, which differed significantly from the other conditions (low-pass filtered at 50, 160, and 500 Hz), the consonant-recognition scores started at somewhat less than 40 percent for one processing band and increased to about 80% with four processing bands.

To summarize, when reviewing the conditions as used by Van Tasell *et al.* (1987), one can conclude that the envelope signal not only contained the envelope cues as defined by Rosen (1992), but also the periodicity cues (200-Hz condition) and even the cues of the spectral fine-structure (2000-Hz condition). Only the 20-Hz condition contained the envelope cues alone; this condition shows the relative importance of envelope information as extracted from a wideband speech signal. But this method uses a wideband envelope, and one cannot guarantee that any manipulation applied on this wideband envelope is equally applied on the envelopes in every critical band. With the method of Drullman *et al.* (1994) the envelopes were down-sampled to 244 Hz, so the bandwidths of the envelopes could be up to 122 Hz, this affects the envelopes extracted from filter bands wider than 244 Hz. It ensures that the envelope signals contain almost only the envelope cues. But Drullman *et al.* (1994) left the spectral fine-structure of the speech signal intact; thus, their stimuli contained temporal fine-structural cues as well as temporal envelope cues. Shannon *et al.* (1995) used noise as the carrier, so the spectral fine-structure of the speech signal was lost, but the processing bands were very broad in comparison with the critical bandwidth. They extracted the envelopes by half-wave rectification and low-pass filtering with 16, 50, 160, and 500 Hz as cut-off frequencies. So, in the 160- and 500-Hz conditions, the periodicity cues were still present.

Therefore, we designed a new experiment in which the influence of *only* the narrow-band temporal envelopes on the recognition of consonants was measured.

B. Temporal envelope manipulation

In order to degrade the temporal information in every critical band, the speech signal was transformed into a temporal-spectral envelope representation. Therefore, the stimuli were passed through a filter bank consisting of band-pass filters with bandwidths equal to the critical bandwidth of the human auditory system (Glasberg and Moore, 1990).

From each band of the speech signal, the temporal envelope was derived by means of a Hilbert transformation. Our hypothesis is that the temporal envelope in every critical

TABLE I. Center-frequencies and equivalent rectangular bandwidths of the band filters used.

F_c	BW	F_c	BW
150	41	1306	166
193	46	1480	184
241	51	1674	205
294	56	1890	229
354	63	2131	255
420	70	2399	284
494	78	2697	315
576	87	3030	352
667	97	3400	392
769	108	3812	436
882	120	4271	486
1008	134	4782	541
1149	149	5352	602

band represents the way the auditory system analyzes the envelope information of speech sounds.

In this spectral-temporal representation, the temporal information was degraded by filtering each envelope. In order to remove all fine-spectral information from the stimulus, the resulting envelope was modulated on narrow-band noises, so all information about the speech carrier has been lost. These narrow-band noises were created by using a noise with a long-term spectrum equal to speech and feeding this signal through the aforementioned filter bank. The different processing bands were summed to compose the final stimulus.

In this way, we hope to assess the importance of the narrow-band temporal envelopes alone on the intelligibility of a speech signal, and assess the relative importance of different modulation frequencies of this narrow-band temporal envelope.

I. METHOD

A. Signal processing scheme

1. Analysis and synthesis

The following were the signal-processing steps used to construct the stimuli:

Filter bank. The speech signal was split up in the frequency domain in 26 different frequency bands, center-frequencies ranging from 150 to 5350 Hz. The filters were triangular shaped, their center-frequencies and bandwidths are described in Table I. The filters were implemented as time-aligned, linear-phase, finite impulse response filters.

Envelope detector. An envelope detector was implemented by means of the Hilbert transformation, using the analytic signal. Because the filtered speech signal is a real, causal, and band limited signal $[F_i(t)$ in Fig. 1], the Hilbert transform can be used to find the appropriate imaginary part $[X_i(t)$ in Fig. 1] of the analytic signal.

The signal $S_i(t)$ is thought of as an analytic signal

$$S_i(t) = R_i(t) + jX_i(t), \quad (1)$$

$$R_i(t) = F_i(t), \quad (2)$$

$$X_i(t) = \mathcal{H}\{R_i(t)\}, \quad (3)$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform.

FIG. 1. Signal scheme of the digital-signal processing per frequency band for the creation of the stimuli.

The magnitude of the complex signal, at any given moment, can be seen as a measure of the instantaneous envelope, and the angle as a measure of the instantaneous frequency of the speech signal. Thus, when using the Hilbert transform ($\mathcal{H}$), the envelope signal $[E_i(t)]$ in Fig. 1] would be

$$E_i(t) = |F_i(t) + j\mathcal{H}\{F_i(t)\}|, \quad (4)$$

where $F_i(t)$ denotes the filtered-speech band.

Modulation and summation. The manipulated envelopes were multiplied with narrow-band noise carriers $[N_i(t)]$ in Fig. 1], which were obtained by filtering a wideband noise with a long-term spectrum equal to speech $[N(t)]$, with the same filter bank as described above. The different frequency bands were summed to become the stimulus signal.

2. Manipulation of temporal information

The envelope signals were manipulated (filtered) to contain only certain modulation frequencies. When the envelope is low-pass filtered, one removes the transients in the envelope contour. When the envelope is high-pass filtered, one removes the global or coarse envelope variations.

In our study, the envelope is notch-band filtered in order to suppress the temporal-envelope information in a band of modulation frequencies.

Manipulations of the carrier also took place. Just as in the experiments from Van Tasell *et al.* (1987) and Shannon *et al.* (1995), the information contained in the spectral fine-structure of the signal was reduced by replacing the speech carrier with noise. In this manner, one can be sure that the subject is unable to use the information contained in the temporal fine-structure, which is the same as the instantaneous spectral envelope, and can only make use of the information as it occurs in the temporal envelope.

B. Experimental design

An experimental variable was the amount of temporal-envelope information contained in the stimulus. There were ten different experimental conditions of band-rejection filtering:

- (i) no filtering as the control condition to assess the influence of the signal processing on the consonant recognition, and also be able to relate the data with the experiments of Drullman *et al.* (1994), Van Tasell *et al.* (1987), and Shannon *et al.* (1995).
- (ii) nine conditions of envelope notch filtering, around three different center frequencies and with three different notch widths. The center-frequencies used were

TABLE II. Center-frequencies and notch widths of the envelope filters used.

nr	F_c	BW	Condition
0	...	0	reference (R)
1	8	4	low-narrow (LN)
2	8	8	low-medium (LM)
3	8	12	low-wide (LW)
4	12	4	center-narrow (CN)
5	12	8	center-medium (CM)
6	12	12	center-wide (CW)
7	16	4	high-narrow (HN)
8	16	8	high-medium (HM)
9	16	12	high-wide (HW)

8, 12, and 16 Hz; the notch widths were 4-, 8-, and 12-Hz wide. (See Table II.)

In this way, one can assess the relative importance of the different sources of speech information under testing, i.e., temporal-envelope information and spectral fine-structural information.

As an example of the signals, the envelope representation of the VCV syllable /aba/ is plotted in Fig. 2(a). This set of envelopes represents the speech signals as they were extracted from the original speech recording. These envelopes were used to construct the stimulus in the reference condition. As an example of notch filtering, the narrow-band envelopes from the LW condition are plotted in Fig. 2(b). These envelopes were used to construct the stimuli by modulating a set of narrow-band noises. The resulting waveforms of the /aba/ stimulus in the reference- and the LW condition are plotted in Fig. 2(c) and (d), respectively.

C. Procedure

Vowel-consonant-vowel syllables (VCV) were used to measure consonant recognition. The stimuli consisted of 18 Dutch consonants in three different vowel contexts (/a/, /i/, /u/), resulting in a list of 54 different VCV stimuli. The consonants used were five plosives: (/b/, /d/, /p/, /t/, /k/), 6 fricatives: (/f/, /x/, /s/, /ʃ/, /v/, /z/), 2 nasals: (/m/, /n/), three semivowels: (/l/, /w/), and (/l/, /n/, /h/).

The speech material was a female voice recorded on digital audio tape (DAT). One token of each VCV word was sampled with a computer at 16-kHz sampling frequency in order to be manipulated. The stimuli were created off-line according to the signal-processing scheme described in the above section. The stimuli were equalized to have an equal rms level, converted to a 16-bit format, and stored on hard disk.

These stimuli were presented to the subjects monaurally via an audiometer Interacoustics AC5 and a TDH39P headphone at a level of 70 dB SP. On the computer display, 18 buttons were presented with all possible consonant responses labeled. The subject "pressed" the response buttons with a mouse. Consonant recognition was measured by counting the number of correctly identified consonants in an 18-alternative, forced-choice procedure.

Every subject was first made comfortable with the experimental task and the possible response categories by be-

FIG. 2. Envelope representations of the stimulus /aba/ in the reference condition (a) and the LW condition (b). These are plotted similar to spectrograms: larger amplitude is depicted as darker gray. These envelope sets were used to modulate the narrow-band noises; the waveforms of the resulting stimuli are plotted in (c) and (d), respectively.

ing presented two lists of stimuli of the reference condition in a random order, with feedback of the presented consonant when the subject responded incorrectly. To avoid order effects of the conditions, every subject was presented a different sequence of conditions according to a digram-balanced Latin square (Wagenaar, 1969). In every experimental condition, two lists (108 stimuli) were presented to the subject in a random order. In the measurements, only correct/incorrect feedback was returned by means of green/red light on the computer display.

D. Subjects

Ten subjects with hearing thresholds better than 20 dB HL for frequencies from 125 Hz to 8 kHz were tested in a soundproof booth. The subjects were between 18 and 40 years old, and were paid. Three of the subjects had some experience with psychophysical experiments.

II. RESULTS

A. Number of correct consonant identifications

The proportions of correct responses in test and retest for all subjects are plotted in Fig. 3 as a function of the notch

width of the envelope filtering and with the center-frequency of the notch as a parameter. The influence of the signal processing can be seen in the result of the reference condition; the average of all subjects in this condition is 71%. The results of all other conditions are lower than the reference condition. Furthermore, the following trends can be seen:

- (i) higher center frequencies of the notch give a higher score.
- (ii) wider notch widths result in either no effect (high- and middle-center frequencies) or a large detrimental effect (low center frequencies) on the score.

An analysis of variance on the scores using the three factors of subjects, conditions, and test/retest gives the levels of statistical significance as depicted in Table III. It reveals that the subjects' scores were significantly different, the conditions gave significantly different results, and there was a significant learning effect between test and retest. Also, the interaction between subjects and conditions is significant. This means that different subjects had significantly different scoring patterns over the conditions. Furthermore, the interaction between subjects and repetitions is significant. So, different subjects had significantly different learning effects.

FIG. 3. Results of the consonant-intelligibility test as a function of the notch-width of the envelope filtering. The center frequency of the notch filtering is a parameter. Results for all subjects ($N=10$) and test and retest are averaged. Standard error of the mean is plotted as an error bar.

The interaction between conditions and test/retest is not significant.

The Wilcoxon signed-rank test was performed on all pairs of conditions in order to assess the significance of the differences between the conditions. All p values significant on a 0.01 level are tabulated in Table IV. Condition R, the reference condition, scores significantly higher than all other conditions except condition HW. Condition LM scores significantly lower than the high center-frequency conditions HN, HM, HW, and the condition CN, but not differently from conditions CM and CW. Condition LW scores significantly lower than all other conditions.

When looking per center frequency, one can see that there is no significant effect of a larger notch width around the high- and middle center frequencies (conditions HN, HM, HW, and conditions CN, CM, CW, respectively). For the low center frequency, there is only the aforementioned condition LW that differs from conditions LN and LM. Between conditions LN and LM there is no significant difference. When looking per notch width, one can see that the conditions with the widest notch widths (LW, CW, and HW) all differ significantly among themselves. But for the other notch widths, there is only a significant difference between the scores from the high- and the low center frequencies (conditions LN vs HN and conditions LM vs HM, respectively).

TABLE III. Analysis of variance on the scores. Dimensions in the design are subjects, conditions, and repetition. The highest interaction is taken as an error estimate.

Design	F_{prob}
subjects	$p < 0.001$
conditions	$p < 0.001$
subjects $\times$ conditions	$p < 0.001$
repetitions	$p < 0.001$
subjects $\times$ repetitions	$p < 0.005$
conditions $\times$ repetitions	$p = 0.29$

TABLE IV. One-tailed probabilities from the Wilcoxon signed-rank test performed on all pairs of conditions. Only significant values are displayed ($P < 0.01$). For clarity, larger P values are substituted with a dot ($\cdot$). The conditions are labeled with the abbreviations as used in Table II.

	LN	LM	LW	CN	CM	CW	HN	HM	HW
R	0.001	0.001	0.001	0.005	0.001	0.001	0.005	0.006	$\cdot$
LN		$\cdot$	0.001	$\cdot$	$\cdot$	$\cdot$	0.003	$\cdot$	$\cdot$
LM			0.001	0.001	$\cdot$	$\cdot$	0.001	0.001	0.001
LW				0.001	0.001	0.001	0.001	0.001	0.001
CN					$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
CM						$\cdot$	0.005	$\cdot$	$\cdot$
CW							0.002	$\cdot$	0.005
HN								$\cdot$	$\cdot$
HM									$\cdot$

B. Confusion between consonants

Confusion matrices were analyzed by means of multidimensional scaling and sequential information analysis.

1. Multidimensional scaling

The asymmetric confusion matrices had to be converted to symmetric similarity matrices in order to be analyzed by INDSCA (Carroll and Chang, 1970). This was accomplished by means of a procedure suggested by Houtgast and used by Klein *et al.* (1970). In this procedure, the cells of the similarity matrix, $s(i, j)$, are calculated not only from the four confusion elements $c(i, j)$, $c(j, i)$, $c(i, i)$, and $c(j, j)$, but from the total correspondence between the response distributions for the stimuli i and j . Mathematically

$$s(i, j) = s(j, i) = \frac{1}{2} * \left(\sum_{k=1}^n (c(i, k) + c(j, k)) - \sum_{k=1}^n |c(i, k) - c(j, k)| \right), \quad (5)$$

in which the dissimilarity of the response distributions, $\sum_{k=1}^n |c(i, k) - c(j, k)|$, is subtracted from the total number of valid responses on the stimuli i and j , $\sum_{k=1}^n (c(i, k) + c(j, k))$.

The results of INDSCAL analysis in two dimensions are plotted in Fig. 4.

In the stimulus space, the first perceptual dimension is always positive for the plosive phonemes, and always negative for the fricative phonemes. The nasals and semivowels are somewhat in-between. So, the first dimension seems to make a separation between “plosiveness” and “fricative-ness” of the phonemes. Looking at the second perceptual dimension, one can see the nasals and semivowels having a positive value, and the fricatives having a negative value with the /s/ phoneme at the extreme. The plosives have both positive (/b/, /p/) and negative values (/d/, /t/) for their second dimension. Therefore, it seems that the second dimension is related to the amount of low-frequency energy in the phoneme, because nasals and plosives like /b/ and /p/ are known to have a large amount of low-frequency energy, especially when compared with /s/ and /ʃ/, which consist of almost only high-frequency energy.

FIG. 4. The object and subject space of the INDSCAL analysis in two dimensions, related to the position of the stimuli and the weighting of perceptual information across conditions, respectively. “sj” is used as a symbol for the phoneme /ʃ/.

In the so-called subject space, each experimental condition is plotted according to the weight of the perceptual dimensions defined in the stimulus space. So, it can be regarded as a “condition space.” The quarter of a circle represents points for which 100% of the variance is explained. Obviously this is not the case, but for each condition, about 78% to 87% of the variance can be explained in two dimensions, even for the most difficult condition [LW in Fig. 4(b)]. For the reference condition R, both dimensions appear to be about equally important. The same is true for the conditions HN and HM [see Fig. 4(b)]. These conditions are characterized by the fact that all modulation frequencies up to at least 12 Hz are present. When the middle frequencies (8–12 Hz) are filtered out (conditions LN, CN, CM, CW, and HW), the weighting of the first dimension increases at the expense of the weighting of the second dimension. Thus, the ability to discriminate on the basis of the second perceptual dimension is reduced. So, filtering out modulation frequencies between 8–12 Hz blurs the distinction between high-frequency energy phonemes and low-frequency energy phonemes. If the lowest modulation frequencies (below 6 Hz) are filtered out (LM and LW), the opposite effect appears: The weighting of the second dimension increases at the expense of the weighting of the first dimension. Thus, the ability to discriminate on the basis of the first perceptual dimension is reduced. So, filtering out modulation frequencies below 6 Hz blurs the distinction between plosiveness cues and fricativeness cues.

2. Sequential information analysis

Sequential information analysis (Wang and Bilger, 1973) was applied for each experimental condition to assess the amount of information transfer from stimulus to response for a set of the most relevant phonetic features. In these analyses,¹ the sequence in which the information transmis-

sion per phonetic feature is analyzed influences the outcome. In eight out of ten conditions, the order of analysis as applied by the program was frication, plosiveness, place, nasality, and voicing. For reasons of comparison between experimental conditions, the order of analysis in the other two conditions was forced to be the same.

The total amount of transferred information is plotted in Fig. 5(a), and the information-transfer function for the phonetic features are plotted in the Fig. 5(b)–(f), respectively. The maximum amount of transferred information for 18 equally often presented stimuli is 4.17.

In general, the patterns in the total transferred information and the information transfer for the phonetic features are very similar to the pattern found for a simple tally of consonants correctly identified (Fig. 3), that is,

- (i) the reference condition has the highest value,
- (ii) the values for the high- and middle center frequencies do not differ much from this value and,
- (iii) are also not much dependent on the notch width
- (iv) only the values for the low center frequencies are strongly dependent on the notch width, and
- (v) only the LW (low center frequency, wide notch width) condition differs strongly from the other values.

There are a few deviations from this general pattern:

- (i) the values for the transfer of the voicing feature are all very low (<0.2), which follows directly from the removal of the periodicity cues,
- (ii) for plosiveness, there is a large difference between the reference condition and the high- and middle center frequency conditions,
- (iii) for nasality, there is a large difference between all low center-frequency conditions and the high- and middle center frequency conditions,

FIG. 5. Results from the SINFA analysis plotted against the notch-width and with the center frequency as parameter. In (a), the total amount of transferred information is displayed. Information-transfer functions for the features frication, plosiveness, place, nasality, and voicing are plotted in (b)–(f), respectively.

- (iv) for nasality, the middle center frequency conditions seem as dependent on notch width as the low center frequency conditions.

In order to assess the influence of replacing the speech carrier by noise on the individual phonetic features, the unconditional measures of information transfer were calculated for each of the features. In this way, the influence of the signal processing on the information transfer of each pho-

netic feature can be assessed independently. When looking at the amount of information transfer for the reference conditions in Table V, it can be seen that this particular analysis–resynthesis method has a different impact on the transfer of the different phonetic features. While the frication and plosive cues are almost completely transferred (0.84 and 0.83, respectively), just two-thirds (0.64) of the place cues and only half (0.49) of the nasality cues get through, and more

TABLE V. Amount of information transfer for each phonetic feature in the reference condition, assessed independently.

Feature	Information transfer
frication	0.841
plosive	0.833
place	0.641
nasality	0.488
voicing	0.387

than 60% of the voicing information gets lost in the processing.

III. DISCUSSION

The goal of these experiments was to assess the relative importance of different regions of modulation frequencies for consonant recognition in speech signals containing only narrow-band envelope cues. Therefore, the speech signal was transformed in such a representation that it was possible to filter out different modulation frequencies in the region between 2 and 20 Hz. As can be seen in Fig. 3, the recognition of consonants is very robust for filtering of the narrow-band envelopes, and only large notch filtering around 8 Hz (condition LW) gives a substantial degradation of the score.

The results from Fig. 3 seem to suggest that a notch of a certain width is more detrimental to the score when it is around a lower center frequency; therefore, the results were also plotted with a logarithmic scale of notch width in Fig. 6.

When plotted in this way, it seems that the scores are directly dependent on the notch width expressed in octaves.

The lengths of the speech sounds were in the range of 0.5–0.7 s. Therefore, when filtering very low-modulation frequencies, one tends to smear the signal beyond its own length, resulting in a substantial lengthening of the signal and a gradual fading in and out of the signal [see also Fig. 2(b)]. So, the envelope cues gets “blurred” and the stimuli are hardly perceived as VCV syllables. This blurring is in-

FIG. 6. Results of the consonant-intelligibility test as a function of the notch-width of the envelope filtering expressed as octaves. Results for all subjects ($N=10$) and test and retest are averaged. Standard error of the mean is plotted as an error bar.

FIG. 7. Results of the consonant-intelligibility test as a function of the lower edge-frequency of the notch of the envelope filtering expressed as the $-\log 2(f_l)$. Results for all subjects ($N=10$) and test and retest are averaged. Standard error of the mean is plotted as an error bar.

herent to the filtering of very low frequencies but can also be considered an artifact of the signal processing.

The amount of smearing is directly related to the lower edge of the notch, and therefore the recognition scores were also plotted against this value (see Fig. 7). This plot is very similar to Fig. 6, and consequently does not give a better understanding of the results with respect to the underlying processes responsible for this degradation.

The relative amount of information transmitted in the reference condition is 0.69. This is three times as much as the relative amount of transmitted information as reported by Van Tasell *et al.* (1987) in their 20-Hz condition. This difference can only be explained by the fact that in this experiment, 26 narrow-band envelopes were used as the carriers of the information, thus conveying dynamic spectral-shape information.

The amount of medial consonant recognition as reported by Drullman *et al.* (1994) was around 90% in their reference condition and about 85% in their 16-Hz low-pass condition. In our experiment, only 71% of the consonants were correctly recognized by the subjects. This difference can only be accounted for by the lack of periodicity- and temporal fine-structure cues that were removed from the stimuli in this experiment.

The consonant-recognition scores in the experiment of Shannon *et al.* (1995) were almost 90 percent in the condition with four processing bands. In our experiment there were more processing bands, and thus more information. But the extra envelopes did not add more usable information for the subjects; it even degraded the recognition scores in comparison with the results from Shannon *et al.* (1995). The largest difference between the outcome of Shannon *et al.* (1995) and our results is in the information received on voicing. They found that temporal information in only two processing bands can transfer 90% of the voicing cues. In our experiment, only 39% of the voicing cues were transferred in our reference condition.

One aim of this experiment was to assess the importance of different modulation frequencies, but due to the choice of

filtering out modulations and, therefore, the inherent smearing in the low-wide condition, one cannot conclude decisively whether the low scores in that condition are caused by the large importance of the low-modulation frequencies or by this smearing effect.

Figure 6 suggests that the width of the notch, and therefore the amount of envelope information that is filtered out, can serve as an explanation for the results. On the other hand, Fig. 7 suggests that the lower edge frequency of the notch, and therefore the amount of smearing, is responsible for the degradation of the recognition scores. Because the trends in these figures are very similar, one cannot favor one explanation over the other based on this experimental data.

IV. CONCLUSIONS

- (1) Consonant recognition seems very robust for notch filtering the narrow-band envelope representation.
- (2) Only large notch filtering (condition low-wide) gives a substantial degradation of consonant recognition.
- (3) The perception of the modulations in speech sounds is better described with a logarithmic frequency scale of modulation rate than with a linear scale.
- (4) There are two explanations that describe the experimental results equally well. On the one hand, the amount of the modulation spectrum still present in the stimuli is directly proportional with the scores. This can be concluded from the fact that the amount of remaining modulation spectrum is inversely proportional to the width of the notch as expressed in octaves. On the other hand, the amount of temporal smearing is inversely proportional with the score. This follows from the fact that the temporal smearing is inversely proportional to the lowest edge-frequency of the notch filter.
- (5) More information in the form of more narrow-band envelopes does not result in higher scores.

- (6) The use of a high number of processing bands, as used in this study, has a detrimental effect on the transfer of voicing cues.

ACKNOWLEDGMENTS

We thank Stuart Rosen for his helpful comments on an earlier version of this paper. This study was supported by the Heinsius-Houbolt Fund, The Netherlands.

¹The sequential information analysis was carried out with the computer program FEATURE INFORMATION XFER (FIX) from M. Johnson, Dept. of Phonetics and Linguistics, University College, London.

- Carroll, J. D., and Chang, J. J. (1970). "Analysis of individual differences in multidimensional scaling via an n -way generalization of the 'eckart-young' decomposition," *Psychometrika* **35**, 283–319.
- Drullman, R., Festen, J. M., and Plomp, R. (1994). "Effects of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.* **95**, 1053–1064.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.* **47**, 103–138.
- Houtgast, T., and Steeneken, H. J. M. (1985). "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.* **77**, 1069–1077.
- Klein, W., Plomp, R., and Pols, L. C. W. (1970). "Vowel spectra, vowel spaces and vowel identification," *J. Acoust. Soc. Am.* **48**, 999–1009.
- Rosen, S. (1992). "Temporal information in speech: acoustic, auditory and linguistic aspects," *Philos. Trans. R. Soc. London, Ser. B* **336**, 367–373.
- Shannon, R., Zeng, F.-G., Kamath, V., Wyganski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- 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.
- Wagenaar, W. A. (1969). "Note on the construction of digram-balanced latin squares," *Psychol. Bull.* **72**, 384–386.
- Wang, M. D., and Bilger, R. C. (1973). "Consonant confusions in noise: a study of perceptual features," *J. Acoust. Soc. Am.* **54**, 1248–1266.