Temporal envelope and fine structure cues for speech intelligibility[a)]

Rob Drullman
Department of Oto-rhino-laryngology, Free University Hospital, P. O. Box 7057, 1007 MB Amsterdam, The Netherlands

(Received 8 March 1994; accepted for publication 27 July 1994)

This paper describes a number of listening experiments to investigate the relative contribution of temporal envelope modulations and fine structure to speech intelligibility. The amplitude envelopes of 24 $\frac{1}{4}$-oct bands (covering 100–6400 Hz) were processed in several ways (e.g., fast compression) in order to assess the importance of the modulation peaks and troughs. Results for 60 normal-hearing subjects show that reduction of modulations by the addition of noise is more detrimental to sentence intelligibility than the same degree of reduction achieved by direct manipulation of the envelope; in some cases the benefit in speech-reception threshold (SRT) is almost 7 dB. Two crossover levels can be defined in dividing the temporal envelope into two equally important parts. The first crossover level divides the envelope into two perceptually equal parts: Removing modulations either x dB below or above that level yields the same intelligibility score. The second crossover level divides the envelope into two acoustically equal peak and trough parts. The perceptual level is 9–12 dB higher than the acoustic level, indicating that envelope peaks are perceptually more important than troughs. Further results showed that 24 intact temporal speech envelopes with noise fine structure retain perfect intelligibility. In general, for the present type of signal manipulations, no one-to-one relation between the modulation-transfer function and the intelligibility scores could be established.

PACS numbers: 43.71.Es, 43.72.Dv, 43.66.Mk

INTRODUCTION

The relevance of the temporal envelope in evaluating the quality of speech transmission has been elaborated in the concept of the modulation-transfer function (MTF; Houtgast and Steeneken, 1985). According to this concept, the detrimental effects of noise and reverberation on speech are adequately measured in terms of the reduction in modulation depth they produce in each of a series of frequency bands. The disappointing results in several studies on the benefit of amplitude compression in hearing aids were discussed by Plomp (1988) in terms of the importance of temporal modulations (intensity contrasts). Plomp used the MTF concept to argue that, just like noise, multichannel amplitude compression with small time constants reduces the intensity contrasts and will thus lead to reduced intelligibility.

In a comment on Plomp's paper, Villchur (1989) objected to the above way of reasoning. Villchur argued that the reduction of the MTF in itself is not the reason for reduced intelligibility. He said (p. 425): "It does not follow that if noise and compression each reduce the MTF, and noise reduces intelligibility, compression must reduce intelligibility equally." More specifically, adding noise to the speech signal causes the weaker elements (consonants) to be masked, which is not the case with compression, where this information is preserved.

The question underlying this discussion is whether the MTF concept holds for dynamic compression. In the present paper, we will not restrict ourselves to the matter of compression *per se*. Using various types of envelope manipulations, we will evaluate among other things whether the implication that a reduced MTF leads to reduced intelligibility is generally valid. As will be shown in the next sections, this is not always the case. It is possible to create conditions which have equal MTFs, but yield quite different intelligibility scores (or vice versa).

For each of a range of frequency channels, two features of the speech signal will be investigated: the temporal envelope and the fine structure (carrier signal). Of these two, the envelope appears to be most important for intelligibility. This is particularly illustrated by channel vocoders, where the amplitude modulations are preserved (up to about 25 Hz), whereas the fine structure at the receiver side is provided by a pulse or noise generator (Flanagan, 1972; O'Shaughnessy, 1987). Intelligibility is worse if the fluctuations in the temporal envelope are not transferred adequately. From previous experiments with filtered temporal envelopes (Drullman *et al.*, 1994a,b) we know that intelligibility at a critical signal-to-noise ratio (SNR) is virtually unaffected when amplitude modulations either above 16 Hz or below 4 Hz are reduced. In the extreme case of complete suppression of the modulations (flat envelope) in 24 $\frac{1}{4}$-oct bands, the intelligibility score for sentences in quiet drops to about 5%, demonstrating that the fine structure alone supplies insufficient information.

Concerning the importance of modulations, the next question is whether troughs are equally important as peaks. Commonly, it is assumed that most information is conveyed in the peaks of the speech signal. This can be inferred from

[a)]Part of this article was presented at the 127th Meeting of the Acoustical Society of America [J. Acoust. Soc. Am. **95**, 3009 (A) (1994)].

 0001-4966/95/97(1)/585/9/$6.00

TABLE I. Survey of the six processing strategies.

Processing	Fine structure	Temporal envelope
REF	speech+noise	speech+noise
SN	speech	speech+noise
FT	speech	speech in peaks, flat in troughs
FP	speech	speech in troughs, flat in peaks
BLK	speech	zero in troughs, flat in peaks
NFS	noise	speech

the idea that additional noise acts as a sort of "fence" where only the peaks of the speech can rise above. The higher the noise level, the less speech peaks can be perceived, until eventually the entire speech signal is masked. The peaks in a narrow frequency band are about 9–12 dB above the long-term average level (Pavlovic, 1987). So, on a dB scale, they are relatively unaffected for SNRs down to at least 0 dB (see Festen *et al.*, 1990, Fig. 2, for an example in an octave band), whereas the modulations in the troughs have disappeared. In an attempt to study the relative contribution of weaker speech components without using noise, Plomp and Van Beek (1990) eliminated energy below a certain level [3–6 dB above the long-term root-mean-square (rms) level, L_{eq}]. By presenting only the spectrotemporal peaks above that level, intelligible speech could be obtained. However, no formal listening tests were carried out.

In summary, two important factors that might affect speech reception have been described: envelope versus fine structure and peaks versus troughs. The aim of this study is to determine how intelligibility depends on the preservation of peaks and/or troughs in the temporal envelope with and without affecting the fine structure. Adding noise to the speech signal reduces the amount of temporal modulations by filling the troughs and, at the same time, changes the fine structure. However, to merely measure the effect of modulation reduction on intelligibility, the fine structure should remain intact. Therefore, a processing scheme was used that operates directly upon the temporal envelope. The present experiments were run to investigate the effects of (1) a noisy envelope alone, (2) removing modulations in the troughs, and (3) removing modulations in the peaks. Answers to (2) and (3) can give an estimate of the relative contribution of peaks and troughs. This gives a method of finding a critical level, subdividing the envelope into equally important peak and trough parts. In addition, while preserving the speech envelope, the effect of a noise fine structure was examined.

I. METHOD

A. Speech processing and experimental design

A total of 130 Dutch sentences of eight to nine syllables read by a female speaker were used as basic material (Plomp and Mimpen, 1979). All sentences were digitized at a sampling rate of 15 625 Hz and 16 bits resolution. An analysis–resynthesis algorithm was used for the signal processing. The basics of this algorithm served in previous experiments (Drullman *et al.*, 1994a,b). All signal manipulations were done (non-real-time) on an Olivetti M290S computer with an OROS-AU21 card with TMS320C25 signal processor.

The wide-band speech signal was split up into 24 $\frac{1}{4}$-oct bands, using a linear-phase FIR filter bank with slopes of at least 80 dB/oct, covering the range 100–6400 Hz. From the output of each channel the Hilbert envelope was determined (Rabiner and Gold, 1975). Each envelope was modified (see below) and the new narrow-band signal was obtained by multiplying each sample of the fine structure by the ratio of the modified and the original envelope. In this way all original amplitude modulations were eliminated. Finally, all modified channels were added and the level of the new wide-band signal was adjusted to have the same rms value as the input signal.

Four methods for envelope modification were investigated. They will be referred to by the acronyms FT, FP, BLK, and SN, respectively. Before explaining their meaning, one feature should be mentioned first: the *target level* within each $\frac{1}{4}$-oct band. For processing purposes, one may think of an imaginary horizontal line through the temporal envelope. It will be expressed in dB *re*: long-term rms value (which is based on the 130 sentences). Thus, when it coincides with the long-term-rms level (L_{eq}), we speak of a target level of 0 dB (one level per $\frac{1}{4}$-oct band accounts for all sentences); $-x$ dB means moving the target level downward; $+x$ dB moving it upwards. Table I gives an overview of all processing methods described below. Figure 1 shows schematic diagrams and examples of the FT, FP, and BLK envelope processing methods. The four methods for envelope modification are as follows:

(1) SN: envelope of speech+noise. The original speech envelope is replaced by the speech+noise envelope for a number of SNRs. Each $\frac{1}{4}$-oct speech+noise envelope is obtained by deriving the Hilbert envelope of the speech+noise signal for that band.

(2) FT: envelope with flat troughs. Envelope parts below the target level are set to this level, while parts above the target level remain unaffected. Thus the troughs are filled by amplifying the local fine structure (removing all original trough modulations). FT processing may be considered to create an artificial noise-floor envelope.

(3) FP: envelope with flat peaks. The complement of FT. Envelope parts above the target level are set to this level, while parts below the target level remain unaffected. In fact this method performs instantaneous 24-channel compression with an infinitely high compression ratio.

(4) BLK: block pulse description of the envelope. Parts

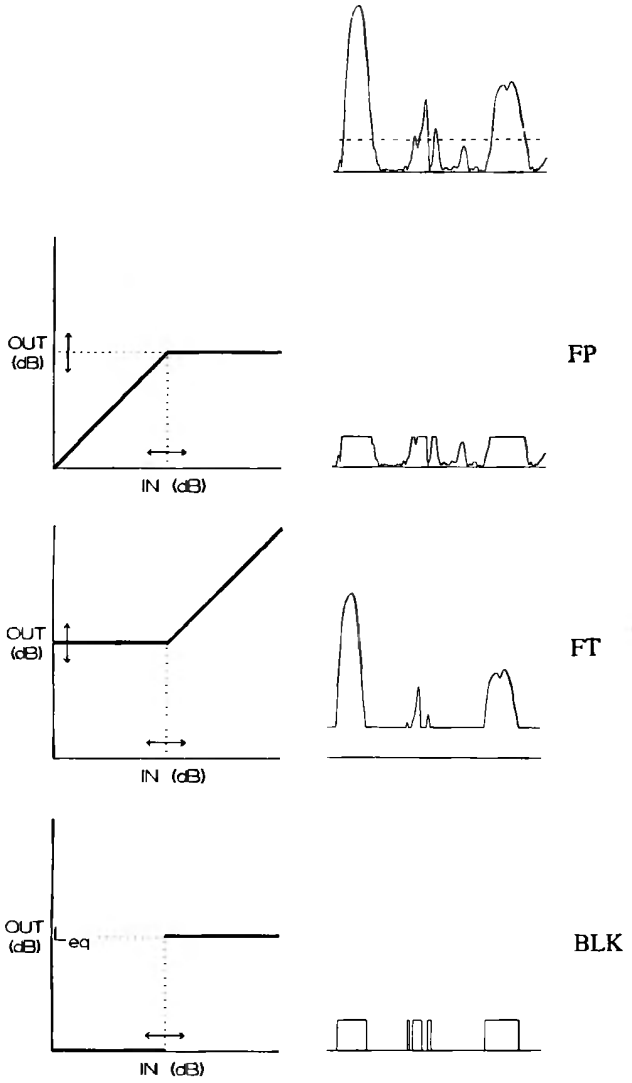

FIG. 1. From top to bottom: diagrams and examples of the modified temporal amplitude envelopes for FP, FT, and BLK, respectively. Top right is (a part of) an input envelope, with the target level drawn as a dashed line.

of the envelope at or above the target level are set to a fixed level, parts below the target level are set to zero. The fixed level always equals L_{eq} for that channel (although the target level may be higher or lower). This method actually performs a one-bid coding of the temporal envelope, as a means to find the optimum level for a cross section of the envelope. The edges of the block pulses are smoothed by means of a 0.5-ms half-cosine window. This prevents sudden signal on- and offset (which could result in clicking) without affecting the block character.

Besides these processed signals (with intact speech fine structure), ordinary speech+noise served as a reference (REF). The noise spectrum matched the long-term average spectrum of the 130 sentences. In the case of REF, the target level simply is the actual relative noise level (i.e., the negative SNR). These reference stimuli were unprocessed, except that both speech and noise had passed through the $\frac{1}{4}$-oct filter bank.

In the case of very high noise levels, SN reaches a limit with a noise envelope and a speech fine structure. Conversely, the effect of a loss of speech fine-structure cues can be assessed by a limit case with an unaffected speech envelope and a noise fine structure. Therefore, an extra type of processing called NFS (noise fine structure) was performed. This involved the combination of a random (noise) fine structure with a speech envelope. For this purpose, there was a parallel analysis (band filtering, envelope detection) of the separate speech and noise signals. For each $\frac{1}{4}$-oct band, the noise fine structure was multiplied by the ratio of the speech envelope and the noise envelope, sample by sample. In fact, NFS is a simple vocoder design (for a whispering voice).

As indicated in Table I, REF and SN differ only in fine structure, their envelopes being equal; FT, FP, SN, and BLK have the same fine structure (speech only) but different envelopes.

Six target levels (conditions) for REF were investigated, viz., 0 to 10 dB (i.e., SNRs of 0 to −10 dB), with a step size of 2 dB. Twelve conditions were used for the other four processing methods. For SN these conditions ranged from 0 to 22 dB, in 2-dB steps; for FT from −1 to +21 dB in 2-dB steps; for FP from −39 to −3 dB, in 4-dB steps, with extra measurements at −21 and −25 dB [for "fine tuning" in the critical (50%) intelligibility region]; for BLK from −35 to +9 dB, in 4-dB steps. The ranges for the conditions were based on the experiences from preceding (informal) tests. Clearly, only one condition existed for NFS.

For FT and SN, a noise floor at 35 dB below L_{eq} was added to the wideband speech signal before processing, to ensure there was just enough signal in the (silent) troughs to be amplified. To prevent sudden on- and offset of a sentence, the noise started 500 ms before and lasted until 500 ms after the speech. These initial and final noise parts were amplified by the processing. For the sake of similarity, the same procedure was undertaken for FP (where the processing did not yield amplified initial and final noise parts); for BLK and NFS, −35-dB noise was added after processing.

B. Subjects

Subjects were 60 normal-hearing students 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 five groups of twelve. Each group was assigned to an experimental condition (SN, FT, FP, BLK, or REF+NFS, respectively).

C. Procedure

The 130 sentences were divided into 12 lists of 11 sentences. Since there were two sentences too few to do this, the first sentences of lists 1 and 2 also served as the first sentences of lists 11 and 12. This did not matter, because only the last ten sentences in a list were used for the intelligibility scores. Six lists were used for REF; all 12 lists were used for FT, FP, SN, and BLK. Lists were presented in a fixed order. The sequence of the conditions was varied according to a digram-balanced 6×6 (REF) or 12×12 Latin square. So, each sequence was presented to one subject in the FT, FP,

SN, and BLK tests, and to two subjects in the REF test. For NFS, four lists of 11 sentences pronounced by a male speaker were used; each of these lists was presented to three subjects.

All stimuli were presented at a level of approximately 65 dB(A). Every sentence was presented once, after which a subject had to reproduce it as accurately as possible. Subjects were encouraged to respond at all times, even if they understood only parts of a sentence. A response was scored as correct only if the entire sentence was reproduced correctly. In the REF test, the level of the masking noise was fixed at 65 dB(A) and the level of the sentences varied according to the condition. The noise started 500 ms before and ended 500 ms after the sentence.

The sentences were presented monaurally through a headphone (Sony MDR-CD999) at the subject's ear of preference and in a soundproof room. Before the actual tests (except for NFS, which followed directly after REF), a list of 11 sentences pronounced by a male speaker in a representative condition was presented, in order to familiarize the subjects with the procedure. For each test and each condition the scores for sentences 2–11 were counted.

II. RESULTS and DISCUSSION

In presenting and discussing the results throughout the rest of the paper, the following abbreviation will be used for the different processing methods and experimental conditions: The condition (target level) will be written in parentheses after the acronym of the processing algorithm; e.g., REF(8), FT(10), FP(−15), SN(0), BLK(−7), etc.

The mean scores (percentages) as a function of condition for each of five processing algorithms are given in Table II. The figures in Table II are the arithmetical averages of the raw scores of 12 subjects. The mean score for NFS is 98.3%, which clearly shows that if the envelope is intact, the fine structure is of minor importance for intelligibility. For further statistical analysis of the data, arcsine transformed scores were used (Studebaker, 1985). The mean and standard error of the transformed scores was calculated for each condition; these means and standard errors were transformed back into percentages (Figs. 2 and 3).[1] All statistics consisted of a repeated-measures analysis of variance (ANOVA) with processing method as between-subjects factor and target level as within-subjects factor. In case of significant interactions, tests for simple effects (Kirk, 1968) were carried out.

A. Speech+noise envelope, artificial noise-floor envelope

Figure 2 displays the scores for REF, FT, and SN. The results for REF display the well-known intelligibility curve (see Plomp and Mimpen, 1979), with a speech-reception threshold (SRT, level for 50%) at a target level of about 5.5 dB. The scores for REF are never better than for SN and FT. In terms of the SRT, the latter yield about 6.5 and 12 dB, respectively. The effects of processing, condition, and the interaction between them are highly significant ($p<0.001$). *Post hoc* tests (Scheffé) on the simple effects showed significantly lower scores for REF than for SN for target levels above 2 to 3 dB ($p<0.01$; $p<0.05$ at 5 dB). From 4 to 10 dB the scores for REF and SN decrease rapidly. Their functions show almost the same slope; the slope for FT is flatter. The scores for FT and SN differ significantly over the range 5 to 15 dB ($p<0.01$; for 15 dB $p<0.05$).

TABLE II. Mean raw scores (percentages) as a function of condition (dB target level) for each of five processing algorithms.

	Processing				Processing	
Cond.	FT	FP	BLK	Cond.	SN	REF
21	28.3			22	17.5	
19	30.8			20	16.7	
17	23.3			18	16.7	
15	32.5			16	17.5	
13	40.0			14	20.0	
11	55.0			12	18.3	
9	65.8		10.0	10	24.2	0.0
7	84.2			8	33.3	6.7
5	90.0		70.0	6	55.8	40.8
3	95.0			4	85.8	69.2
1	99.2		98.3	2	94.2	83.3
−1	96.7			0	98.3	92.5
−3		99.2	96.7			
−7		100.0	97.5			
−11		96.7	98.3			
−15		93.3	95.8			
−19		75.8	94.2			
−21		73.3				
−23		65.0	85.0			
−25		50.8				
−27		35.8	74.2			
−31		18.3	51.7			
−35		8.3	36.7			
−39		8.3				

In the light of the processing methods (see Table I), the difference in scores between REF and SN must be attributed to the unaffected fine structure in SN. Preservation of the fine structure if the original intensity contrasts are disturbed yields a 1-dB benefit of the SRT. In the case of extremely high target levels, when the envelope contains practically no speech information anymore, the fine structure seems to supply some minimal cues (0% for REF and 17% for SN). An explanation for the difference between the SN and FT scores must be based on the presence of nonrelevant fluctuations,

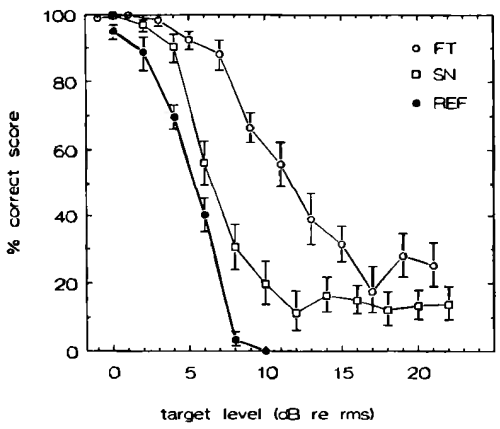

FIG. 2. Mean score and standard error (vertical bars) of FT, SN, and REF as a function of target level.

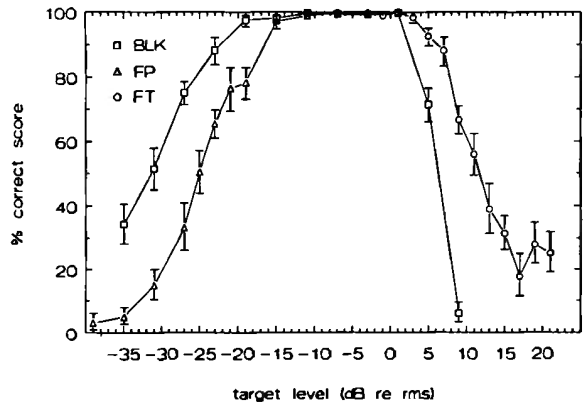

FIG. 3. Mean score and standard error (vertical bars) of BLK, FP, and FT as a function of target level.

originating from the noise, in the temporal envelope of SN. Apparently, these fluctuations interfere with those of the speech signal and leave the listener with a "sorting problem;" i.e., he/she is unable to separate the relevant (speech) modulations from the nonrelevant (noise) modulations.

In summary, flattening the troughs of the speech signal as a means to reduce the amount of temporal modulations (artificial noise-floor envelope) is less detrimental to intelligibility than the same modulation reduction brought about by the addition of noise. The benefit expressed in terms of the SRT is about 6.5 dB.

B. Removing envelope peaks and/or troughs

Figure 3 displays the scores for FP, BLK, and FT. The data for FP show that a considerable part of the temporal-envelope peaks (up to 15 dB below L_{eq}) can be "chopped off" before a detrimental effect on intelligibility is noticeable. The SRT for FP is around −25 dB. In contrast to the other processing methods, the curve for BLK is typically nonmonotonic. Apparently, complete intelligibility is reached for a wide range between −19 and +1 dB. At lower target levels, all frequency bands are filled more evenly, resulting in less temporal cues for the listener. At higher target levels, BLK only codes the upper peak parts; sentences in these conditions consist of only a few short-duration block pulses per $\frac{1}{4}$-oct band. The limit cases for low and high target levels are quite different, viz., severely compressed speech and complete silence, respectively. For high target levels both envelope and fine structure information is lost, so that intelligibility decreases rapidly (steep slope); for low target levels the loss of envelope information may be compensated for by relying more on the fine structure. The physical difference between FP and BLK may explain the discrepancy between their scores at low target levels. This difference consists of the presence (FP) or absence (BLK) of speech in the troughs. As for FP, at lower target levels the modulations in the troughs become relatively stronger. They may therefore disturb the perception of the peaks (or what is left from them) as separate entities. BLK speech does not have this problem; the peaks are well separated in time.

FP and FT have virtually the same slope (except for a different sign) from −31 to −15 dB and from 3 to 17 dB, respectively. The symmetry axis runs at a target level of about −6 dB. So, flattening the envelope x dB below (FT) or about (FP) that level results in virtually equal intelligibility (except for extreme deviations from this level). Simple one-bit coding of the envelope (BLK) retains perfectly intelligible speech when applied in a 20-dB range around L_{eq} −9 dB. FT vs FP on the one hand and BLK on the other are two approaches to estimate the relative contribution of peaks and troughs. Thus, the critical target level ("perceptual crossover level") for subdividing the envelope in two equally important parts is estimated at about 6–9 dB below L_{eq}.

III. GENERAL DISCUSSION

A. Number of channels and envelope definition

The present signal processing was performed on 24 channels of $\frac{1}{4}$-oct bandwidth, just smaller than the ear's critical bandwidth. Modification of the individual envelopes of narrow bands is perceptually more effective than modification in wider bands. This is caused by variations of the energy distribution within subbands that can be resolved by the auditory system. Thus part of the original modulations can still be perceived after manipulation, which is not the case with $\frac{1}{4}$-oct processing bands. Indeed, in the case of severe modulation reduction, previous studies (Drullman *et al.*, 1994a,b) have shown a significant increase in intelligibility if less channels with larger bandwidths ($\frac{1}{2}$- and 1-oct) are used. On the other hand, if noise bands are modulated with temporal speech envelopes, 24 channels provide a fairly detailed transmission of the spectrogram. The outcome of the NFS experiment may therefore not be very surprising. In a similar experiment Shannon *et al.* (1994) reported nearly 100% sentence intelligibility with only four channels, indicating that little spectral details is sufficient for speech recognition in quiet.

A second point is the definition of the temporal envelope. We adopted the Hilbert envelope, although many studies use a method of rectification and lowpass filtering. The Hilbert envelope has the clear advantage that it accurately follows all amplitude modulations of a frequency band, running smoothly over the actual waveform. As a consequence, any modification of the envelope can be performed without keeping unwanted (original) temporal modulations intact. In this way one can precisely control the envelope cues transmitted to the listeners.

B. Results in relation to the MTF

As mentioned in the Introduction, one aim of this study was to establish the relation between the intelligibility scores and (the prediction by) the MTF. For all processing methods, the MTF was measured for each of five octave bands with center frequencies at 0.25, 0.5, 1, 2, and 4 kHz. The measurements were based on the long-term envelope spectra (processed in $\frac{1}{4}$-oct bands versus unprocessed) of a 71-s speech fragment (30 concatenated sentences) and were performed according to the phase-locked MTF procedure described in Drullman *et al.* (1994b). For each octave band the

TABLE III. Average phase-locked MTFs of five processing algorithms for various target levels, based on a 71-s speech fragment.

Target level	Processing			
	REF/SN	FT	FP	BLK
20	0.01	0.07		
15	0.03	0.07		1.47
10	0.10	0.11		1.15
5	0.26	0.23		0.73
0	0.52	0.49	0.33	0.46
−5			0.24	0.31
−10			0.19	0.23
−15			0.16	0.18
−20			0.13	0.15
−25			0.11	0.13
−30			0.09	0.11

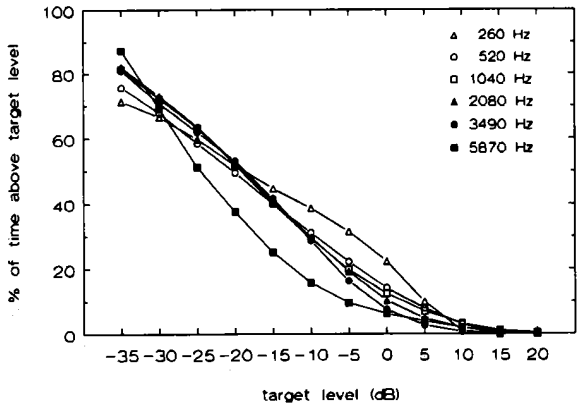

FIG. 4. Percentage of the time that the temporal envelopes in six $\frac{1}{4}$-oct bands of a 71-s speech fragment are above the target level.

mean modulation reduction (m) in the range 0.5–20 Hz was computed. To get the average MTF, a weighted average of the mean m per octave band was taken.[2] The results for all processing algorithms for a number of target levels are given in Table III.

First of all, the global trend shows that, of course, decreasing scores correspond to decreasing average MTFs, except for BLK. The MTFs for BLK(15) and BLK(10) need some clarification. For target levels above 2 dB, there is a sharp decrease in the scores (Fig. 3), whereas there is a major increase in the average MTF, even to values greater than 1. That this may happen with this type of processing can be explained by the fact that a block-pulse approximation with a few small pulses has stronger modulations than the relatively smooth original envelope.

Apart from this specific point, a unique relation between the values in Table III and the intelligibility scores for the different processing algorithms and target levels is missing. Clearly, since REF and SN have the same temporal envelope (with a $\frac{1}{4}$-oct resolution), they have the same average MTFs. These values are practically equal to the theoretical values.[3] However, the scores for REF and SN (Table II, Fig. 2) are different for most target levels. For example, an average MTF of 0.10 at 10 dB corresponds to a score of either 0% (REF) or 24% (SN). The MTFs of REF(10) and FT(10) are almost equal, unlike their scores of 0% and 60%, respectively.

Comparison of FT(0) and FP(0) is even more illustrative of why one has to proceed with caution in applying the MTF as a direct measure for intelligibility. Scores for these conditions are high (nearly 100%), whereas their MTFs differ substantially, viz., 0.49 and 0.33. The MTF can even drop as low as 0.16 for FP(−10) with a corresponding score of 95%.

These are just some examples; by further comparing the average MTF values with the measured intelligibility scores one can find more discrepancies. Let it be clear that the above is not meant to discredit the MTF concept in general. The MTF was mainly developed for practical purposes, and has proved to be very successful in studying the quality of speech transmission in the presence of disturbances like bandpass limiting, noise, reverberation, echoes, and (wideband) automatic gain control (Steeneken and Houtgast, 1980; Steeneken, 1992). But the experiments in this study have demonstrated that its use cannot simply be extended to any manipulation of the temporal speech envelope (see also Hohmann and Kollmeier, 1990; Verschuure *et al.*, 1993). Taking FP as an extreme example of fast multichannel compression and referring to the discussion mentioned in the Introduction, one must conclude that the relation between MTF, compression, speech+noise, and intelligibility is rather obscure. Indeed, as pointed out by Plomp (1988), fast multichannel compression (even in a more moderate form) reduces the MTF, as does the addition of noise. But the results of these experiments suggest that Villchur (1989) was right in stating that the effects of noise and compression are different and that reduction of the MTF does not automatically reduce intelligibility.

C. Temporal envelope statistics

For a better insight in the amount of information that is (physically) present after processing, some data about the temporal envelope statistics may be helpful. Using the same 71-s fragment as for the MTF measurements, the percentage of the time that the amplitude envelope exceeds the target level was measured for 6 $\frac{1}{4}$-oct bands, distributed regularly over the spectrum. The results for a range of target levels are shown in Fig. 4. Not surprisingly, all curves are monotonically decreasing, with slightly different paths for the lower and higher frequency bands. The steeper the slope, the narrower the peaks in that frequency band. By relating these data to the intelligibility scores, one can see how sensitive listeners are to minor differences in the temporal envelope. For example, at 10 dB, there are peaks for on average 2% of the time, the very amount that the envelopes contain after FT processing. Still, this results in a mean intelligibility score of 60%. At 15 dB there are on average 0.5% peaks, and the mean score for FT still is 33%. This sensitivity to a minimum peak remainder (even at 20 dB) may be the reason why the plateau for extreme FT conditions lies around a score of 25%–30%. In earlier studies, using really flat envelopes (Drullman *et al.*, 1994a,b), significantly lower intelligibility scores of 3%–7% were found.

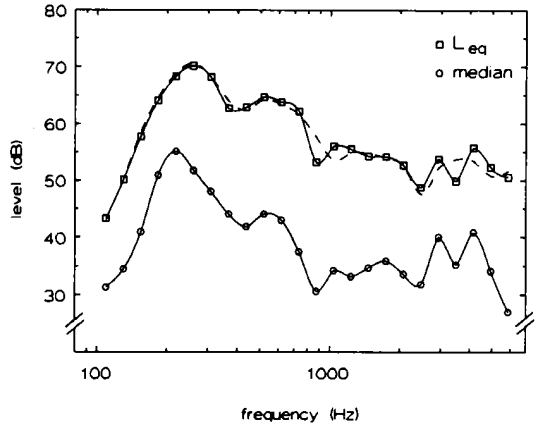

FIG. 5. L_{eq} and median for 24 $\frac{1}{4}$-oct bands, based on a 71-s speech fragment (30 concatenated female sentences). The dashed line reflects L_{eq} over all (130) sentences.

According to Fig. 4, the median envelope level is about 18 dB below L_{eq}. Figure 5 displays L_{eq} and the median for all 24 $\frac{1}{4}$-oct bands of the 71-s fragment. For completeness, the dashed line shows L_{eq} for all 130 female sentences. The difference between L_{eq} and the median is relatively independent of the frequency band. The mean difference over the 24 bands is 18.4 dB with a standard deviation of 3.5 dB. This level is much lower than the crossover level of 6–9 dB below L_{eq} that was found for the intelligibility scores of FT vs FP and BLK. So, for the present sentence material, one can say that there are two different crossover levels: (1) an acoustic crossover level of on average 18 dB below L_{eq} that divides the temporal envelope into two equal peak and trough parts, and (2) a perceptual crossover level of 6–9 dB below L_{eq} that yields equal intelligibility scores when removing all modulations either x dB below or above that level. The fact that the perceptual level is 9–12 dB higher than the acoustic level suggests that the envelope peaks are more important for intelligibility than the troughs.

IV. CONCLUSIONS

The main conclusions of this study are as follows.

(1) Reduction of the temporal modulations in speech by direct manipulation of the envelope is less detrimental to intelligibility than the same degree of reduction caused by adding noise. Preservation of the speech fine structure alone (speech+noise envelope) results in a 1-dB decrease of the SRT. If in addition an artificial noise-floor envelope is used, the SRT decreases by an extra 5 to 6 dB. This indicates that fine structure cues play a less important role than envelope cues and that noise introduces spurious modulations, disturbing the perception of the relevant speech modulation.

(2) The most important fraction of the dynamic range, providing essentially 100% sentence intelligibility, is between 19 dB below and 1 dB above L_{eq}.

(3) A perceptual crossover level of 6–9 dB below L_{eq} is the critical level for which removing modulations either x dB below or above yields the same intelligibility score. This is substantially higher than the median envelope level of 18 dB below L_{eq} (acoustic crossover level), demonstrating the relative importance of the speech peaks for intelligibility.

(4) Intact temporal speech envelopes of 24 $\frac{1}{4}$-oct bands with random fine structure retains perfect intelligibility. Conversely, an intact fine structure and a random temporal envelope yields an average score of only 17%.

(5) For the present type of envelope processing methods no one-to-one relation between the MTF and the intelligibility scores could be established. Equal MTFs do not lead to equal intelligibility.

ACKNOWLEDGMENTS

This research was supported by the Linguistic Research Foundation, which is funded by the Netherlands organization for scientific research, NWO. The author wishes to thank Tammo Houtgast and Joost Festen for their useful comments on the manuscript.

[1]The standard errors (σ_M) in Figs. 2 and 3 were obtained by transforming the interval (mean-σ_M) to (mean+σ_M) in the arcsine domain back into percentages. Because of this (nonlinear) inverse transform, the mean is generally not in the middle of the interval.

[2]The weighting factors for the octave bands were derived from Steeneken and Houtgast (1980). Their original weighting factors account for a total of seven octave bands, viz., those used here and two with center frequencies of 125 Hz and 8 kHz. Because we omitted the latter two, the weighting factors for the remaining five bands were recalculated so that their sum equals 1: $W_{0.25}=0.196$; $W_{0.5}=W_1=0.157$; $W_2=0.255$; $W_4=0.235$.

[3]The theoretical values for the MTF in case of steady-state interfering noise are given by the formula $m=(1+10^{-\mathrm{SNR}/10})^{-1}$, where m is independent of the modulation frequency (Houtgast and Steeneken, 1985).

Drullman, R., Festen, J. M., and Plomp, R. (**1994a**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.

Drullman, R., Festen, J. M., and Plomp, R. (**1994b**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.

Festen, J. M., van Dijkhuizen, J. N., and Plomp, R. (**1990**). "Considerations on adaptive gain and frequency response in hearing aids," Acta Otolaryngol. Suppl. **469**, 196–201.

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

Hohmann, V., and Kollmeier, B. (**1990**). "Sprachverständlichkeit bei Dynamik-kompression," in *Fortschritte der Akustik—DAGA'90* (DPG-Kongress-GmbH, Bad Honeff), pp. 1115–1118.

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.

Kirk, R. E. (**1968**). *Experimental Design: Procedures for the Behavioral Sciences* (Brooks/Cole, Belmont, CA), 1st ed., Chap. 8, pp. 263–270.

O'Shaughnessy, D. (**1987**). *Speech Communication* (Addison–Wesley, Reading, MA), Chap. 7, pp. 305–309.

Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust Soc. Am. **82**, 413–422.

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.

Plomp, R., and Van Beek, J. H. M. (**1990**). "The spectrogram as an aid in studying speech intelligibility at low S/N ratios," J. Acoust. Soc. Am. Suppl. 1 **87**, S118.

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

Shannon, R. V., Zeng, F.-G., Wygonski, J., Kamath, V., and Ekelid, M. (**1994**). "Speech recognition with minimal spectral cues," J. Acoust. Soc. Am. **95**, 2876(A).

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

Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

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

Verschuure, J., Dreschler, W. A., de Haan, E. H., van Cappellen, M., Hammerschlag, R., Maré, M. J., Maas, A. J. J., and Hijmans, A. C. (**1993**). "Syllabic compression and speech intelligibility in hearing impaired listeners," Scand. Audiol. Suppl. **38**, 92–100.

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.