Investigation of perceptual constancy in the temporal-envelope domain

Marine Ardoint[a)]
Laboratoire de Psychologie de la Perception (CNRS—Université Paris 5 Descartes), Departement d'Etudes Cognitives, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France

Christian Lorenzi
Laboratoire de Psychologie de la Perception (CNRS—Université Paris 5 Descartes), Departement d'Etudes Cognitives, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France

Daniel Pressnitzer
Laboratoire de Psychologie de la Perception (CNRS—Université Paris 5 Descartes), Departement d'Etudes Cognitives, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France

Andreï Gorea
Laboratoire de Psychologie de la Perception (CNRS—Université Paris 5) Descartes, Université René Descartes, UFR Biomédical des Saints Pères. 45 rue des Saints Pères, 75006 Paris, France

(Received 2 April 2007; revised 21 December 2007; accepted 28 December 2007)

The ability to discriminate complex temporal envelope patterns submitted to temporal compression or expansion was assessed in normal-hearing listeners. An XAB, matching-to-sample-procedure was used. X, the reference stimulus, is obtained by applying the sum of two, inharmonically related, sinusoids to a broadband noise carrier. A and B are obtained by multiplying the frequency of each modulation component of X by the same time expansion/compression factor, α ($\alpha \in [0.35-2.83]$). For each trial, A or B is a time-reversed rendering of X, and the listeners' task is to choose which of the two is matched by X. Overall, the results indicate that discrimination performance degrades for increasing amounts of time expansion/compression (i.e., when α departs from 1), regardless of the frequency spacing of modulation components and the peak-to-trough ratio of the complex envelopes. An auditory model based on envelope extraction followed by a memory-limited, template-matching process accounted for results obtained without time scaling of stimuli, but generally underestimated discrimination ability with either time expansion or compression, especially with the longer stimulus durations. This result is consistent with partial or incomplete perceptual normalization of envelope patterns.
 [DOI: 10.1121/1.2836782]

PACS number(s): 43.66.Mk, 43.66.Ba [JHG] Pages: 1591–1601

I. INTRODUCTION

Normal-hearing listeners understand each other even when the rate of production of their spoken words is increased up to a factor of roughly 3 (e.g., Fairbanks and Kodman, 1957; Fu *et al.*, 2001; Versfeld and Dreschler, 2002). This form of *perceptual constancy* [which may be defined as the ability to listen to critical global aspects of speech and other complex nonspeech sounds, in contrast to the ability to listen to acoustic details (Li and Pastore, 1995)] seems to be based on the speech temporal envelope, as it is relatively independent of the audio carrier. Indeed, Fu *et al.* (2001) have shown that the deterioration of the spectral and temporal fine structure content of speech stimuli does not preclude their recognition after temporal compression or expansion. In addition, Ahissar *et al.* (2001) have shown that speech comprehension of time-compressed signals is correlated with the representation of the speech envelope in MEG (magnetoencephalography) signals. Hence, a possible explanation for the robustness of speech intelligibility to variation in presentation rate is that perceptual normalization is applied to the amplitude envelope of sounds, whether they are speech signals or not.

The general question asked in this paper was whether or not normal-hearing listeners show perceptual constancy for nonlinguistic amplitude envelopes presented at various time scales or, in other words, robust recognition of complex envelope patterns that are temporally compressed or expanded. Here, the nonlinguistic amplitude envelopes were obtained by summing two inharmonic, sinusoidal amplitude modulations. Temporal compression or expansion (i.e., temporal transposition) was achieved by mutiplying the frequency of each modulation component by a given index. Discrimination of the temporally transposed patterns was assessed as a function of their compression/expansion index. A similar approach was taken by Gockel and Colonius (1997) to study perceptual constancy following transposition of spectral patterns. In addition, discrimination of the temporally transposed patterns was assessed here for (i) two frequency ratios

[a)]Author to whom correspondence should be addressed. Electronic mail: ardoint@ens.fr

(i.e., two frequency spacing) of the two modulation components and (ii) three levels of amplitude compression/expansion applied to the complex envelopes, because both factors should be important determinants of envelope discrimination.

Manipulation of frequency ratio was intended to test the extent to which putative processing either within or across temporal modulation channels (Dau *et al.*, 1997a, b) affects a listener's resistance to temporal transposition. In other words, this manipulation attempted to assess the effect of the resolvability of the modulation components on the discrimination of the transposed envelopes.

The manipulation of the amplitude compression was intended to test the effect of the temporal envelope peak-to-trough ratio on performance. Previous experiments have revealed that this ratio is also an important determinant of speech identification (e.g., Fu and Shannon, 1999; Lorenzi *et al.*, 1999; Apoux *et al.*, 2001). In these experiments, the peak-to-trough ratio was modified by applying a power-law transform to the stimulus envelope. Overall, these experiments showed that increasing the peak-to-trough ratio yields significant improvements in phoneme identification performance in noise. It should be noted that all the speech perception studies investigating the effects of temporal compression/expansion (e.g., Fu *et al.*, 2001) have tested temporal compression/expansion constancy against a ~100% correct recognition performance for nontransformed (control) stimuli. The possibility remains that the observed constancy reflected in fact a ceiling effect. Accordingly discrimination of the temporally transposed patterns is assessed here for three peak-to-trough ratios (obtained by means of a compression/expansion of the envelope amplitude as in the studies cited previously) yielding different levels of discrimination performance with the highest still below perfect performance.

Current models of temporal-envelope processing in the auditory system do not include temporal normalization. One such model proposes that temporal-envelope detection or discrimination is achieved by cross correlating the outputs of amplitude-modulation channels with memory-stored templates according to an "optimal detector" scheme (Dau *et al.*, 1997a, b). This model accounts successfully for a variety of envelope detection data collected in masked and unmasked conditions. However, some form of a *normalization* process in the time domain might be required to account for the discrimination of complex temporal envelopes submitted to various levels of temporal compression or expansion. In the modeling part of the present study, we used a simplified front-end to the optimal detector approach to investigate whether listeners' performances can be predicted with a model that does not include a normalization stage.

II. EXPERIMENTS

A. Method

1. Listeners

Four listeners ranging in age between 20 and 33 years were tested. One of them was one of the authors (M.A.) and the other three were students. All listeners had absolute thresholds of less than 20 dB HL (Hearing Level) at audiometric frequencies between 0.125 and 8 kHz, and no history of hearing difficulty. Practice was given to each listener prior to data collection (see the following). All listeners were fully informed about the goal of the present study and provided written consent before their participation. The present experimental protocol is in accordance with the Helsinki declaration in 2004.

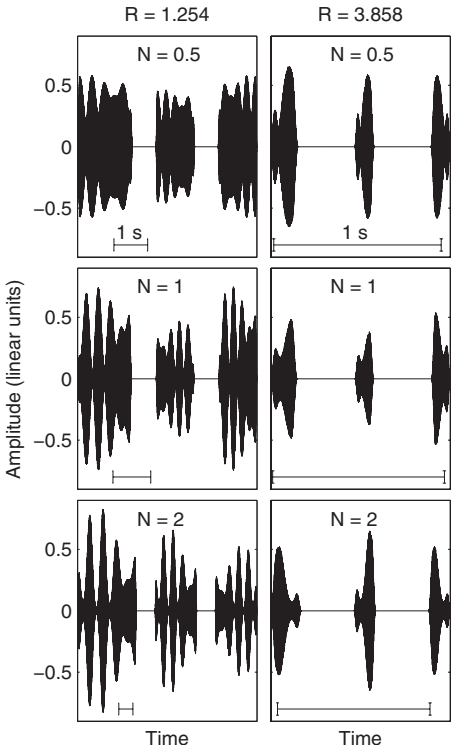

FIG. 1. Examples of wave form for stimuli in six typical trials, obtained with $R=1.254$ (left column) and $R=3.878$ (right column), $\alpha=1.414$, and $N=0.5$ (top panels), 1 (middle panels), and 2 (bottom panels). The center frequency and the modulation depths of the modulation components were varied across trials, whereas the global amplitude was varied independently for each stimulus within a trial. The 1-s time bars are different for different values of N because the center frequency of the modulation components is roved.

2. Stimuli

Examples of stimulus wave forms are illustrated in Fig. 1 for six typical trials. All stimuli were broadband noise audio carriers modulated by a complex temporal envelope equal to the sum of two temporal modulations:

$$S(t) = [1 + m_1 \sin(2\pi\alpha f_{m1} t + \varphi_1) + m_2 \sin(2\pi\alpha f_{m2} t + \varphi_2)]b(t), \quad (1)$$

with t being time, f_{m1} and f_{m2}, m_1 and m_2, and φ_1 and φ_2, respectively, are the frequencies (with $f_{m1}<f_{m2}$), depths and starting phases of the two components and with $b(t)$, the broadband noise carrier. Parameter α is a frequency-multiplication factor explained in the following. The stimuli were generated with a 16-bit digital/andlog converter (44.1 kHz sampling rate) under the control of a PC and delivered binaurally via a Sennheiser HD 600 headphone at a level of 65 dB SPL in a soundproof booth. The broadband

noises were non-Gaussian. They were generated in the time domain using a uniform distribution of amplitudes and were physically different within (i.e., across test and comparison stimuli) and across trials. The bandwidth of the noise was set to half the sampling rate.

Two f_{m2}/f_{m1} inharmonic ratios, R, were used so as to tap, presumably, the same ($R=1.254$, "unresolved components" condition), or two distinct temporal modulation channels [$R=3.879$, "resolved components" condition (e.g., Ewert and Dau, 2000; Lorenzi *et al.*, 2001)]. The two modulating frequencies, f_{m1} and f_{m2}, were symmetric (on a log scale) about a nominal central frequency f_c of 3 Hz [chosen because it corresponds to the most salient and critical frequency in the production and understanding of continuous speech (Houtgast and Steeneken, 1985)]. In order to prevent listeners from building over time a template of the stimuli and storing it in long-term memory, f_c was randomized across trials within a range of ± 0.5 octaves (i.e., 2.12–4.24 Hz). For the same reason, the phases φ_1 and φ_2 of the two modulation components were also independently randomized from trial to trial in a range of $0-2\pi$.

Both within and across trials, the modulation amplitudes m_1 and m_2 were each randomly varied between 0.25 and 0.5 so that their sum never exceeded 1.0 (i.e., overmodulation). The global amplitudes of the modulated noises of all stimuli (within and between trials) were independently randomized in a range of ± 3 dB SPL (with a 1-dB step) about the average 65 dB SPL. Two additional experimental conditions were obtained by elevating the envelope amplitudes to powers $N=0.5$ and 2 (amplitude compression and expansion, respectively). Although amplitude compression minimizes the peak-to-trough contrasts, amplitude expansion exaggerates them.

Manipulation of the factor α was central to the present study. In the *test* condition, it was used to produce the temporally compressed ($\alpha>1$) and expanded ($\alpha<1$) versions of the "reference" stimulus ($\alpha=1$). Factor α ($\alpha\neq 1$) was thus applied only to the two modulation frequencies, f_{m1} and f_{m2}, of the target and comparison stimuli. In the *control* condition, factor α ($\alpha\neq 1$) was also applied to the two modulation frequencies of the reference stimulus, so that in effect all stimuli had the same duration and the reference and target stimuli had identical envelopes. The following 7 α-values were used in all experimental conditions:.35,.5,.7, 1, 1.41, 2, and 2.82.

However, when N was equal to 1.0, 6 extra α-values (.42,.59,.84, 1.18, 1.68, and 2.37) were also used. To facilitate the analysis of the data and their comparison with previous studies, the α-values were converted to a compression/expansion index, CE, computed as $100|1-1/\alpha|$.

Stimuli duration, D, was equal to the period of the modulated envelope, i.e., $D=1/(\alpha f_{m2}-\alpha f_{m1})$, so as to prevent listeners from using more than one temporal envelope beat for their judgments. Obviously, D varies with both the compression-expansion factor, α, and with the frequency ratio, R. As can be seen in Table I, these durations (displayed for each α and R) range from as short a period as 81 ms ($\alpha=2.83$, $R=3.878$) to as long a period as 4199 ms ($\alpha=0.35$, $R=1.254$). Stimuli were ramped on and off with a cosine envelope whose temporal extent was equal to 50 ms for $\alpha=1$, and was proportional to α (ramp duration $=50/\alpha$ ms) when α departed from 1.

TABLE I. Stimulus duration for each value of α and R.

	Duration, D (ms)	
α	$R=1.254$	$R=3.878$
0.35	4199	652
0.42	3499	543
0.5	2939	456
0.59	2491	387
0.71	2070	321
0.84	1750	272
1	1470	228
1.19	1235	192
1.41	1042	162
1.68	875	136
2	735	114
2.38	617	96
2.83	519	81

3. Procedure

In the test condition, envelope discrimination performance (% correct) was measured by means of an XAB matching-to-sample procedure (see MacMillan and Creelman, 2005, Chap. 9) whereby X stands for the reference stimulus (with $\alpha=1$, i.e., a CE index of 0%), whereas A and B are its compressed or expanded temporal versions ($\alpha\lessgtr 1$, $\mathrm{CE}\neq 0\%$), one of which (randomized over trials) is a time-reversed (temporal mirror) rendering of X. The listener's task was to determine whether A or B matched the reference stimulus, X. The temporal interval between the three stimulus versions was 500 ms and the minimum interval between two successive trials was 2 s. Performance was measured in separated blocks for each combination of temporal compression/expansion (α), components frequency ratio (R), and amplitude compression/expansion (N). The control condition involved temporally noncompressed/expanded A and B versions of the reference X (see the following). Listeners completed the control and test conditions in random order. In both cases, one experimental block consisted of 50 trials and was repeated three times in a different random order for each listener. Hence, for each experimental condition, percent correct was computed out of 150 trials.

Before starting the main experiments, listeners passed 3–5 training sessions (i.e., 150–250 trials) with $\alpha=1$, $N=1$, and $R=1.254$ and 3.878. The training sessions were terminated once listeners reached a performance level of at least 75% correct (i.e., $d'=2$) which was achieved over a period of 2–6 h.

Listeners were provided with visual feedback in all sessions (training and testing) and experimental conditions (test and control conditions). Listeners' performance is presented as sensitivity (d') scores obtained from the assessed percentages correct (Macmillan and Creelman, 2005; Table A5.3: Differencing model).

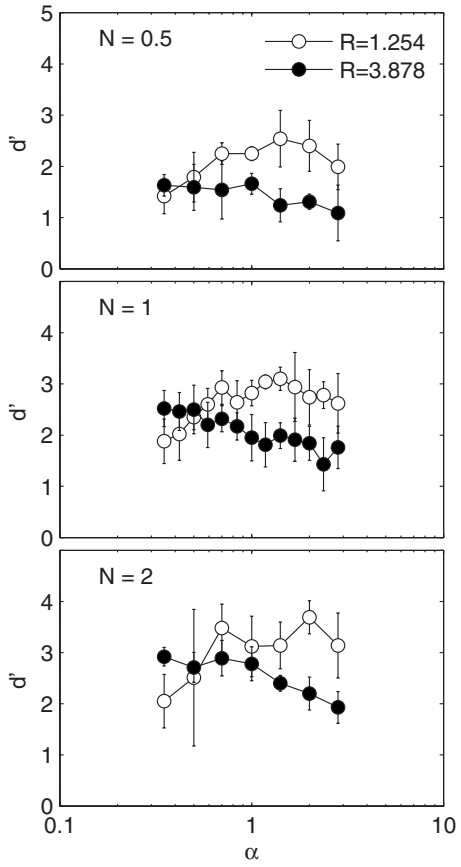

FIG. 2. Mean discrimination sensivity (d') for four listeners obtained in the control condition. Discrimination performance is plotted as a function of the time compression/expansion factor, α. Here, the time compression/expansion factor is applied to all envelopes (i.e., *X*, *A*, and *B*). Error bars represent ± 1 standard deviation across listeners. In each panel, open and filled circles correspond to cases where the frequency ratio, *R*, of the two modulation components of the complex envelopes is 1.254 and 3.878, respectively. The top, middle, and bottom panels show the data obtained with $N=0.5$ (all envelopes are compressed in amplitude), 1 (all envelopes are left intact), and 2 (all envelopes are expanded in amplitude), respectively.

B. Results: Control performance

In the control experiment, the expansion/compression factor α was applied to *all three* envelopes of the XAB sequence (with A or B being the time-reversed version of X) so that its manipulation was only meant to assess the dependence of envelope discrimination performance on the center frequency f_c of the envelopes, or, equivalently, on the duration of the stimuli. As a reminder, an $\alpha=1$ is equivalent to a nominal $f_c=3$ Hz with the two extreme α-values for all listeners ($\alpha=0.35$ and 2.83) yielding nominal f_c-values of 1.1 and 8.5 Hz.

All four listeners behaved similarly in this task. Therefore, for each experimental condition, discrimination sensitivity (d') was averaged across listeners. Figure 2 displays these average data as a function of α with envelope frequency-component ratio (*R*) the parameter [$R=1.254$ (open circles); $R=3.878$ (closed circles)]. The average data are shown for each of the three envelope amplitude compression/expansion indices, *N* (top panel: $N=0.5$; middle panel: $N=1$, bottom panel: $N=2$).

Overall, the discrimination of identical, time-reversed envelopes yields the following main characteristics: (1) it generally peaks for $\alpha=1.41-2$ (i.e., for $f_c=4-6$ Hz) when the modulation frequencies are close ($R=1.254$) and decreases monotonically as a function of α when the modulation frequencies are spaced apart ($R=3.878$); (2) it is globally better for the proximal rather than distal spacing of modulation components, particularly so within the medium-to-high f_c-range and independently of the amplitude expansion/compression index, *N*; (3) it increases with the amplitude expansion index, *N*; (4) it yields a maximum d' of about 3.69 (i.e., 93% correct). Overall, the present discrimination scores are within the range of those obtained for the discrimination of noise modulated envelopes (Takeuchi and Braida, 1995; 78–99% correct with a similar XAB method).

The above-mentioned qualitative account is confirmed by a three-way (α[7], R[2], N[3]) repeated-measures analysis of variance (ANOVA). Each of the main factors yields a significant effect [α: $F(6,18)=3.29$, $p<0.05$; *R*: $F(1,3)=17.84$, $p<0.05$; *N*: $F(2,6)=115.3$, $p<0.0001$]. Of the three second-order interactions, only $\alpha\times R$ is significant [$\alpha\times R$: $F(6,18)=14.34$, $p<0.00001$; $\alpha\times N$: $F(12,36)=1$, NS; $R\times N$: $F(2,6)<1$, NS]. Finally, the third-order interaction is not significant [$F(12,36)=1.94$, NS].

In other words, the present experiment and statistical analysis point to the fact that the discrimination of nontransposed temporal envelopes depends on their central modulation frequency (f_c), is better for a small frequency spacing of modulation components (*R*), and increases with amplitude expansion factor (*N*). Moreover, f_c and *R* interact in such a way that discrimination as a function of f_c has roughly an inverted U-shape for low values of *R* and a monotonically decreasing function for higher values.

C. Results: Test performance (control versus temporally expanded/compressed envelopes)

Again, as all four listeners behaved similarly for each experimental condition, discrimination sensitivity (d') was averaged across listeners. Figure 3 displays the average discrimination scores, d'_{Test}, in the same format as Fig. 2. In order to isolate the effect of the temporal transposition factor (α) from that of the envelope's central frequency (f_c; assessed in the first experiment), the d'_{Test} scores were normalized with respect to those obtained in the "control" experiment (d'_{Control}) and are expressed as $d'_{\text{Test}}/d'_{\text{Control}}$ ratios in Fig. 4.

With this format, a perfect perceptual invariance to temporal transposition will translate into flat functions relating $d'_{\text{Test}}/d'_{\text{Control}}$ to α. It should be also noted that if the effects of the two additional factors studied (*R* and *N*) were the same in the control and "test" experiments, computing $d'_{\text{Test}}/d'_{\text{Control}}$ ratios should cancel them out. Deviations from these predicted null effects of *R* and *N* would therefore indicate contributions of these factors different from those observed in the nontransposed case.

Based on the d' ratios shown in Fig. 4, the discrimination of temporally tranposed envelopes can be characterized as follows. It is an inverted U-shaped function of the transposition factor (with a peak at or just below $\alpha=1$) whatever

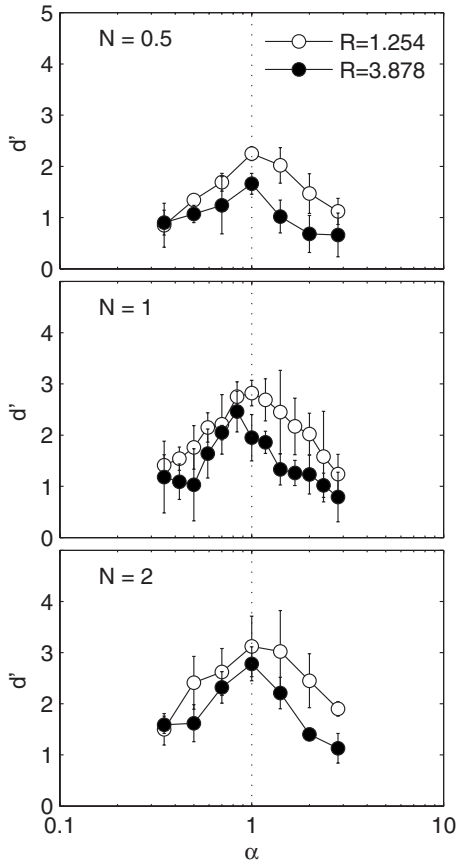

FIG. 3. Mean discrimination sensitivity (d') for four listeners obtained in the test condition. Discrimination performance is plotted as a function of the time compression/expansion factor, α. Here, the time compression/expansion factor is applied to the envelopes of A and B, only. For each value of R, vertical dotted lines indicate $\alpha=1$ Otherwise as in Fig. 2.

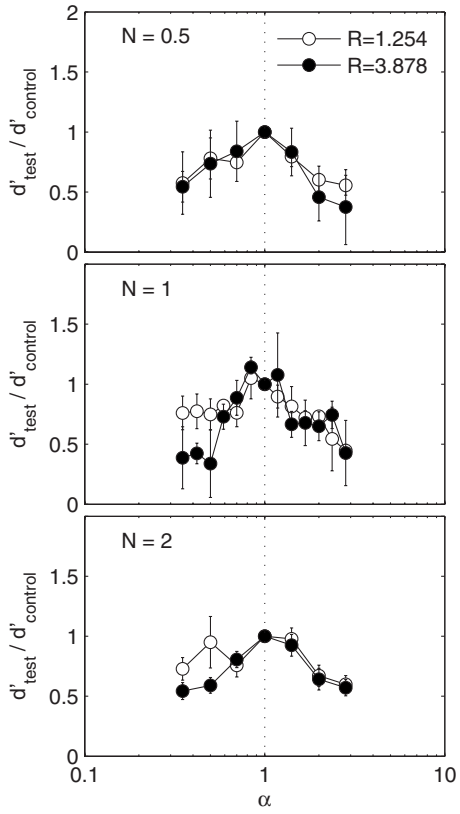

FIG. 4. Mean ratio of discrimination scores (d'_{Test} presented in Fig. 2) normalized with respect to those obtained in the "control" experiment (d'_{Control} presented in Fig. 1). The $d'_{\text{Test}}/d'_{\text{Control}}$ ratios are plotted as a function of the time compression/expansion factor, α. in each panel, the vertical dotted line indicates $\alpha=1$ Otherwise as in Fig. 2.

R or N. Given that perceptual constancy predicts that $d'_{\text{Test}}/d'_{\text{Control}}$ should be independent of α, ratios smaller than 1 indicate a sensitivity reduction due to the temporal transposition *per se*. For the extreme temporal expansion ($\alpha=0.35$) and compression ($\alpha=2.83$) values used, sensitivity drops by a factor of 1.32–2.7. The U-shaped functions of α are symmetrical for $R=3.878$ but temporal compression appears to be more detrimental than temporal expansion for $R=1.254$ (at least for $N=1$ and 2). With the exception of a limited temporal expansion range ($0.35<\alpha<0.5$), $d'_{\text{Test}}/d'_{\text{Control}}$ ratios are relatively independent of R, indicating that this factor contributes equally to the recognition of temporally transposed and non-transposed envelopes. $d'_{\text{Test}}/d'_{\text{Control}}$ ratios are also independent of N suggesting that this factor is also equally involved in the discrimination of temporally transposed envelopes and in the discrimination of nontransposed envelopes.

The previous observations are partially supported by a 3-way (α[7], R[2], N[3]) repeated measures ANOVA performed on the $d'_{\text{Test}}/d'_{\text{Control}}$ ratios. The effect of temporal compression/expansion factor α is significant [$F(6,18)=26.07$, $p<0.000001$], confirming the fact that temporally transposed envelopes are less well discriminated than nontransposed ones. Hence, contrary to previous studies that demonstrated a resistance of word or sentence identification to their temporally compressed/expanded versions (i.e., perceptual constancy; Fairbanks and Kodman, 1957; Fu *et al.*, 2001; Versfeld and Dreschler, 2002), the present data show a lack of temporal transposition constancy for nonlinguistic stimuli. For instance, Fu *et al.* (2001) showed that when a 32-channel vocoder was used to remove temporal fine-structure cues, time-expanded and time-compressed speech remained perfectly intelligible even at half (CE=100%) or two times (CE=50%) the normal speaking rate (equivalent to $\alpha=0.5$ and 2 in the present study, respectively). For such changes in α values in the present discrimination task, d' dropped by a factor of 1.3–2.4. The effect of the R-factor (presumably related to the resolvability of the envelopes' components) is also significant [$F(1,3)=20$, $p<0.05$]. This inference is qualified by the partial comparisons over the two R-levels showing a significant R-effect only for the largest temporal expansion ($\alpha=0.35$) and for the amplitude expanded ($N=2$) envelopes [$F(1,3)=12.22$, $p<0.05$]. These partial comparisons are in line with the significant $\alpha\times R$ interaction [$F(6,18)=3.15$, $p<0.05$]. The effect of the amplitude compression/expansion factor, N, is not significant [$F(2,6)=3.48$, $p=0.1$] and neither is the $\alpha\times N$ interaction [$F(12,36)=1.28$, $p=0.27$] or the $R\times N$ interaction [$F(2,6)=1.52$, $p=0.29$]. Finally, the triple interaction $\alpha\times R\times N$ is not significant either [$F(12,36)=1.02$ $p=0.45$]. Overall, the statistical analysis shows that perfect perceptual constancy is

not maintained for temporally transposed, nonlinguistic envelopes.

III. INTERIM DISCUSSION

The main results of the present study can be summarized as follows. The discrimination of two-component temporal envelopes *equally* compressed/expanded in time is maximized when their two modulation components are close in frequency and centered around 4–6 Hz, but is a monotonically decreasing function of the frequency of their modulation components when the latter are spaced apart (along the modulation frequency axis). Overall, discrimination scores are enhanced when the frequency spacing between the two modulation components is decreased and when the envelopes are expanded in amplitude. Discrimination of temporally transposed envelopes appears to preserve globally these characteristics while displaying a significant drop related to the amount of transposition (whether compression or expansion). Hence, at odds with previous studies that used linguistic stimuli, the present data suggest an absence of perfect perceptual constancy over temporal transpositions.

Effects of resolvability (R) *and duration* (D). The dependence of envelope discrimination on the frequency spacing of modulation components, R, is consistent with the existence of distinct temporal modulation filters. Indeed, the temporal-reversal discrimination task requires the encoding of the phase of the modulation components. On the assumption that temporally modulated signals are discriminated via a comparison (or cross correlation) of their temporal profiles, discrimination based on the phase of their components is possible as long as they feed into the same modulation filter but not otherwise. The R values used in the present experiment were chosen so that the two envelope components tap one ($R=1.254$) or two distinct ($R=3.878$) modulation filter(s) as they have been inferred from modulation masking experiments (e.g., Ewert and Dau, 2000). For these conditions, the modulation filterbank model hence predicts that discrimination of phase-reversed envelopes should be better for the smaller R, just as presently found. Sek and Moore (2003) who have measured the discrimination of two envelopes that differed only in the phase of one of their three components found a similar dependence on the frequency ratio of these components.

Inasmuch as the hypothetical modulation filters have a constant quality ratio, Q, the observed R-effect should be independent of the envelopes' central frequencies, f_c. The present data, however, show a significant $R \times f_c$ interaction, with the disappearance of the R-effect for the lower f_c values ($\alpha < 0.5$ that is $f_c < 1.5$ Hz: cf. first experiment and Fig. 2). To this we offer one possible interpretation relating to the duration of the stimuli. As noted in Sec. II A, in order to prevent listeners from using more than one envelope beat for their judgments, all envelopes were temporally windowed so that they included only one envelope beat period [$D = 1/\alpha(f_{m2}\text{-}f_{m1}) = 1/\alpha f_{m1}(R-1)$]. Hence, stimulus duration D was inversely proportional to both α and R (cf. Table I). Figures 5 and 6 replot the mean control and test data shown in Figs. 2 and 3, respectively, as a function of D (instead of α) in order to show the confounded effect of changes in duration on discrimination performance. In Fig. 5, the replotted control data (open and filled circles) reveal that discrimination performance is a nonmonotonic function of stimulus duration. More precisely, discrimination performance peaks at intermediate durations ranging from 735 to 1042 ms (corresponding to $\alpha = 1.41-2$, or $f_c = 4.2-6$ Hz). This seems consistent with the notion that, in the first (i.e., control) experiment, changes in stimulus duration are—at least partly—responsible for the effect of α or f_c (temporal compression/expansion). For instance, an increase in a listener's memory load or a decay of the sensory trace stored in auditory short-term memory could explain why envelope discrimination deteriorates for the longest duration. In Fig. 6, the replotted test data (open and filled circles) indicate that discrimination performance is a nonmonotonic function of stimulus duration. Discrimination performance peaks when all three stimuli of the XAB sequence have the same duration (as shown by vertical dotted lines), and degrades as a function of the departure in duration betwen the reference and comparison (target and standard) stimuli. Can changes in duration also account for the effect of R? For comparable duration intervals—that is for D between 326 and 521, 456 and 735, and 625 and 1042 ms—post-hoc comparisons (LSD–Least Significant Difference test) indicate that discrimination

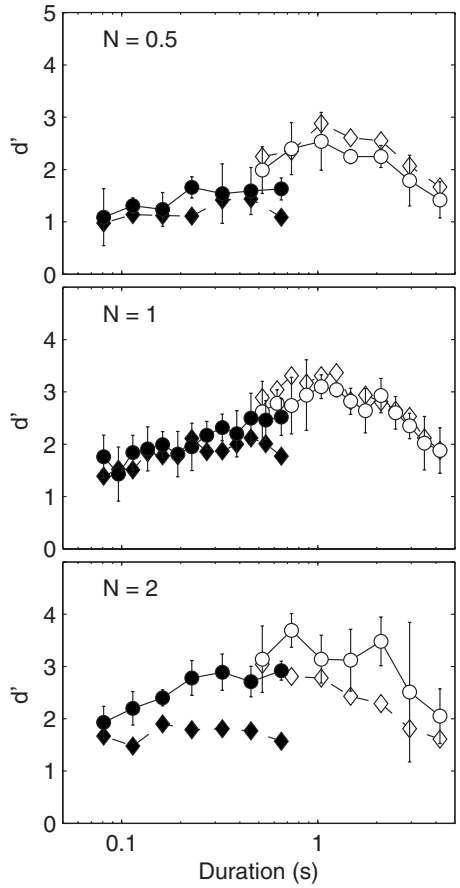

FIG. 5. Mean discrimination sensitivity (d'_{Control}) for the four listeners obtained in the control condition (open and filled circles). Discrimination performance is plotted as a function of stimulus duration, D (ms). Otherwise as in Fig. 2. Open and filled diamonds correspond to simulation data obtained with $R=1.254$ and 3.878, respectively.

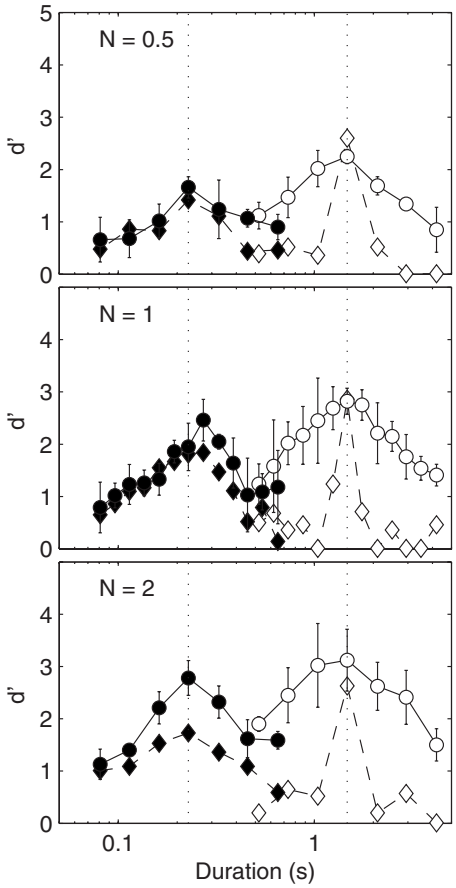

FIG. 6. Mean discrimination sensitivity (d'_{Test}) for the four listeners obtained in the test condition (open and filled circles). Discrimination performance is plotted as a function of stimulus duration, D (ms). Otherwise as in Fig. 2. Open and filled diamonds correspond to simulation data obtained with $R=1.254$ and 3.878, respectively. For each value of R, vertical dotted lines indicate $\alpha=1$.

scores obtained with $R=1.254$ are still significantly greater than those obtained with $R=3.878$ [$p<0.05$] for $N=0.5$, 1, and 2, except when $N=1$ and D is between 465 and 735 ms [$p=0.15$] and when $N=2$ and D is between 326 and 521 ms [$p=0.17$]. These tests hence sustain a genuine effect of components' resolvability. However, this effect is more apparent when the magnitude of envelope components is small (i.e., when $N=0.5$) and tends to disappear when envelope components are presented at levels (i.e., depths) well above detection threshold (i.e., when $N=1$ or 2). In addition, the effect of resolvability, when observed here, is relatively small in magnitude. Thus, in the present experiment, envelope discrimination performance seems to be more constrainted by stimulus duration (and presumably memory factors) than by envelope resolvability per se.

Effects of amplitude compression/expansion (N). The beneficial effect of envelope amplitude expansion indicates that complex envelopes discrimination depends on their overall peak-to-trough ratio. It can also be related to the notion that envelope discrimination is at least partly based on listeners using local features of the envelope, particularly their local peaks, as suggested by a speech-perception study conducted by Drullman (1995). Indeed, the effect of raising the envelope amplitude by a power larger than 1 is equivalent to reinforcing its peaks (relative to troughs). Effects of amplitude expansion are also found for speech signals presented in noise (e.g., Fu and Shannon, 1999; Lorenzi *et al.*, 1999; Apoux *et al.*, 2001). Moreover, amplitude expansion is "naturally" observed in hearing-impaired listeners as a consequence of the loss of fast-acting cochlear compression. On a more audiological side, this suggests that peripheral amplitude compression (and its loss in the case of coch-le-ar lesions) affects not only detection (as shown previously for hearing-impaired listeners, e.g., Moore *et al.*, 1992) but also discrimination. The current results predict therefore that hearing-impaired listeners with loudness recruitment should show better-than-normal ability to discriminate between complex temporal envelopes of linguistic and nonlinguistic stimuli.

Perceptual constancy for envelope discrimination? The present study demonstrates a strong limitation in the discrimination of temporally compressed or expanded nonlinguistic envelopes regardless of their amplitude expansion. In fact, the data (Figs. 3 and 4) show a discrimination deterioration even for the smallest temporal compression/expansion used (CE: 16% and 18%; $\alpha=0.84$ or 1.18).

This lack of constancy for temporally transposed nonlinguistic temporal envelopes appears to be at odds with the constancy reported for both linguistic and musical signals. Indeed, identification of temporally transposed linguistic signals remains unaffected by transposition up to a compression/expansion index (CE) of 50% (e.g. Fairbanks and Kodman, 1957; Daniloff *et al.*, 1968; Vaughan and Letowski, 1997; Gordon-Salant and Fitzgibbons, 2001; Versfeld and Dreschler, 2002). Some studies on categorical perception of phonemes also seem to provide evidence for the existence of some form of temporal normalization (Summerfield, 1981; Miller and Volaitis, 1989). It may then be argued that the current discrepancy is related to the fact that linguistic signals are coded by a speech-specific system (Liberman and Mattingly, 1985) that may well be designed so as to resist temporal alterations. A resistance to temporal alterations has also been reported for musical sequences [as long as the duration of their component notes is within a 160–1280 ms interval (Warren *et al.*, 1991)] hence rejecting the singularity of the speech coding system.

The alternative, more plausible account of this discrepancy is that previous studies have compared categorization performance for reference and transposed signals under conditions where the former were always discriminable (e.g., Fairbanks and Kodman, 1957; Daniloff *et al.*, 1968; Fu *et al.*, 2001). It may then well be that, though degraded, performance for the transposed signal did not show a measurable drop due to a ceiling effect. This putative methodological concern was circumvented in the present study by utilizing a reference task in which performance was below 100% correct (i.e., a d' not larger than 4; see Fig. 2).

IV. MODEL PREDICTIONS

The present data show that the discrimination of nonlinguistic temporal envelopes is degraded by temporal transpositions. Hence, perfect perceptual constancy is not achieved

for time-stretched or time-compressed random envelopes. It is unclear, however, whether the observed degradation is consistent with the total absence of perceptual constancy in the envelope domain, or whether it still requires some sort of normalization mechanism. To investigate this issue, we now present a qualitative modeling study in which we compare listeners' performances with the predictions of an envelope cross correlator after auditory filtering. The cross correlator did not include any normalization stage. We could obtain a good fit to the control data, which indicates that envelope cross correlation was sufficient to explain behavioral performance when comparing stimuli with the same duration. The model failed however in the test conditions, strongly suggesting the need for an additional normalization stage when stimuli have different durations.

Model structure. The model was an envelope extractor with a limited memory store followed by a cross-correlation decision stage. The first stage was a single linear gammatone filter that simulated the band pass filtering at one locus on the basilar membrane (Patterson *et al.*, 1987). In a second stage, the temporal envelope of the band pass-filtered signal was extracted using half-wave rectification followed by lowpass filtering [cutoff=64 Hz, rolloff=6 dB/oct (see Viemeister 1979)]. The envelope obtained was then temporally windowed with an exponential function in order to simulate a decay of the memory trace. A similar approach to account for memory constraint was taken by Sheft and Yost (2005).

The decision stage was realized as a cross correlation between the windowed envelopes. On each trial, the output of the model for the three stimuli (X, A, and B) was computed. The windowed envelope of the reference stimulus, X, served as a "template" that was cross correlated with the windowed envelopes of A and B. The response was determined by the largest cross-correlation coefficient (X better correlated to A or X better correlated to B). Note that this differed from a simple Pearson product-moment correlation in two important ways. First, the correlation was applied on the envelopes including the direct current component. The measure was thus sensitive to modulation depth to some extent. Second, the whole cross-correlation function was computed so envelopes were effectively time shifted to find the highest correlation. This approach was very similar to that used by van de Par and Kohlrausch (1998) to model monaural and binaural envelope correlation detection, and it could be viewed as simplified version of the optimal detector described in Dau *et al.* (1997a, b).

Stimuli were generated as in the behavioral experiment, except that no level rove was applied. Center frequency f_c (or, equivalently, duration) was roved across trials just as in the behavioral experiment. Six hundred trials were simulated for each condition. To restrict the numbers of degrees of freedom in the model, no internal noise was added. The noise carrier, refreshed for each interval, was thus the sole source of variability in the predictions for a given set of stimulus parameters. The randomization of modulation depth, phase, and f_c are other sources of variability across trials. Percent correct was transformed into d'. The half-life of the exponential window and the gammatone center frequency were varied to fit the data in the control condition, where stimuli had identical durations within each trial.

Model results, control condition. Fits were obtained by minimizing the root mean square (rms) error between experimental data and model predictions for the two R values and for $N=1$. The best fit was obtained for a half-life of 1.2 s and a filter center frequency of 5 kHz (Fig. 5). Model predictions for these parameters (open and filled diamonds) and empirical data (open and filled circles) for the control conditions are shown in Fig. 5. The results have been replotted as a function of the duration of the stimuli. As indicated earlier, this duration covaried with R, except for a few values where the same duration could be obtained with two different R's. Most predicted values fell within the variability range of the empirical data for $N=1$. There was also a relatively good fit for $N=0.5$, even though the parameters were not optimized for this condition. The fit was poorer for $N=2$, where the model consistently underestimated performance.

The discrimination performance peaked at intermediate stimulus durations. In the model, this was because performance first increases with stimulus duration and then decreases because of the exponential weighting window, which limits the maximum stimulus duration that can be accurately stored. Performance would increase indefinitely with stimulus duration without such a window, because d' increases by the square root of duration for a correlation receiver. For any given duration, the model also predicted poorer discrimination for high R compared to low R. The poorer discrimination for high R was also observed in the listeners' discrimination scores, although the model overestimates the effect. It is noteworthy that the model predicted an effect of R without any modulation filterbank and thus without any notion of modulation frequency resolvability. We hypothesize that the model's behavior for these points is related to the complexity of the envelope pattern. For low R, there are more distinct features in the envelope, as illustrated in Fig. 1 (left versus right column). The decision stage of the model has to extract a signal template from the noisy stimulus, so having many peaks in the envelope will make this stage more resistant to noise. This is less the case for high R where the envelopes are broadly similar, without sharp features, and hence more susceptible to noise. Such an interpretation of the model's behavior remains speculative and should be verified by further testing.

Overall, the simulations show that the effects of α, f_c (or duration) and R on complex envelope discrimination observed in the control experiment, where all stimuli within a trial have the same duration, can be accounted for reasonably well by a simple model of envelope cross-correlation with a memory limit.

Model results, test condition. Figure 6 shows empirical data (open and filled circles) and model predictions (with the half-life time parameter used to simulate control data) for the test condition, again plotted as a function of duration. Model parameters were kept identical to the ones used for the control condition. The model predictions (open and filled diamonds) always peaked at a value corresponding to $\alpha=1$ (indicated by vertical dotted lines for each value of R). Such a

value represents the case where reference and comparison stimuli have the same duration. It is not surprising that the model should perform well in these conditions as these are similar to control conditions. For other values of α, predicted performance decreased, for each value of R. The same trend was observed in listeners' performances. For $N=0.5$ and $N=1$ and for small durations (high R), the model accurately predicted the rate of decrease in performance due to the mismatch in durations. Crucially, however, for long durations (low R) the rate of decrease was much faster in the model's predictions than in the listeners' data. This suggests that, for long durations, envelope cross-correlation underestimates listeners' performance. A different mechanism or an additional, as yet unspecified normalization stage is thus needed to account for listeners' performance.

Model discussion. The aim of the present model was to illustrate the prediction of an envelope-correlation approach when comparing two random envelopes. Such an approach has been used before in the context of envelope perception (Dau *et al.*, 1997a, b, van de Par and Kohlrausch, 1998; Sheft and Yost, 2005). The main finding of the present study is that envelope correlation predicts well behavioral performance when stimuli durations are equal, but fails when durations are unequal. In order to keep the focus on the predictions of envelope correlation in the context of perceptual constancy, we tried to keep the model as simple as possible. For instance, no adaptation or compression front-end was used, even though such processing would affect model behavior (Derleth *et al.*, 2001). We now examine briefly this and other choices made in the modeling and show that they do not bear on our general conclusion.

No attempt was made to model the influence of N or of the level rove imposed on the stimulus. Adaptation is important to account for these parameters in at least two ways. First, static compression would change the effective peak to trough ratio, as well as the effect of the level rove. Second, dynamic changes in the adaptation characteristics would result in different behavior for forward and reversed envelopes. Accurately capturing these effects in a model would require adding a realistic front-end with respect to absolute and dynamical changes in level. Although this would be of interest for future modeling studies to better account for performance in the control conditions, it is unlikely that such a front-end would change anything regarding the failure to predict performance in the test conditions where stimuli durations are unequal.

The choice of the auditory filter considered was based on the optimization of the fit between model and experimental data, and it was found that a center frequency of 5 kHz provided the best fit. The relatively wide bandwidth (ERB $=564$ Hz) of the 5 kHz filter minimizes two disruptive effects on envelope perception resulting from band pass filtering, that is, envelope filtering, and masking produced by the intrinsic envelope fluctuations of the noise carrier. Figure 7 illustrates the quality of the fit between model and data when the half-life of the exponential memory window is varied, with filter center frequency as a parameter. Low-frequency filters produced a worse fit to the data (rms error, upper panel) and a lower performance overall (mean error, lower panel). We hypothesize that listeners would ignore the low-frequency filters and listen to frequency regions providing more reliable cues. The fit also got worse for filters with frequencies above 5 kHz, but for a different reason. In this case, the model predicted higher performance than observed behaviorally. Note however that the model did not include any source of internal noise. Results similar to what we are presenting could be obtained by choosing the most accurate envelope representation, that is the highest available auditory filter, and then adding a source of internal noise. Again, this would introduce additional parameters to the model without affecting the general conclusion.

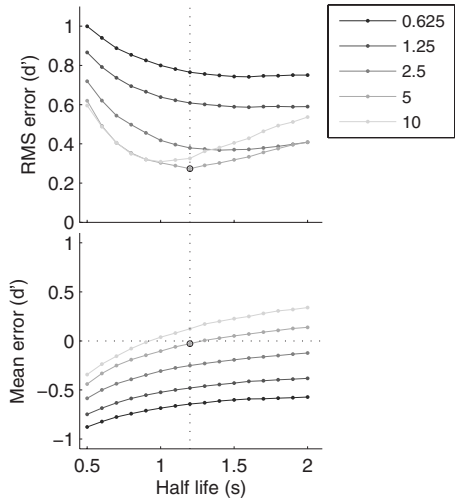

FIG. 7. Influence of model parameters on the quality of fit to the behavioral data, for $R=1.254$ and 3.878 and $N=1$. rms error (top) and mean error (bottom) are plotted as a function of the half-life of the exponential window applied to the envelopes. Each shade of gray indicates a different auditory filter frequency (0.625, 1.25, 2.5, 5, or 10 kHz). The best fit, half-life $=1.2$ s and center frequency$=5$ kHz, is indicated by circles.

A model applicable to the comparison of temporal patterns of different length has been proposed by Sorkin and colleagues (Sorkin and Montgomery, 1991; Sorkin *et al.*, 1994). Sorkin and colleagues used tone sequences rather than envelopes and, in the case of equal stimulus durations, listeners' performances could be predicted by cross correlating the sequences of onset times after introduction of an internal noise. In order to account for performance with stretched or compressed sequences, Sorkin and Montgomery (1991) assumed a normalization of all sequences to the same duration, but the internal noise was proportional to the amount of normalization required. Such a model is based on correlations of time-of-occurrence between salient features in the sequence, the tone onsets in the case of Sorkin and Montgomery (1991). Applying it to the comparison of temporal envelopes would require extracting 'features' from the continuous envelope function, as was in fact proposed by Sheft and Yost (2005). An interesting future direction for modeling the test data of the present experiment would thus be to apply a noisy normalization mechanism, similar to Sorkin and colleagues, to salient features of the internal envelope.

V. GENERAL DISCUSSION

The discrimination of nontransposed temporal envelopes (that is envelopes of identical duration) can be accurately

accounted for by an auditory model using an envelope extraction stage similar to that proposed by Viemeister (1979) followed by a template-matching process similar to that proposed by Dau *et al.* (1997a, b). Although beyond the goal of the present study, the empirical and simulated results obtained in these control discrimination experiments suggest that envelope filtering via selective modulation filters such as those described initially by Dau *et al.* (1997a, b) is not a necessary prerequisite for the discrimination of equal-duration time-reversed envelopes.

The poor discrimination of temporally transposed, nonlinguistic envelopes suggested—at a first sight—a total absence of perceptual constancy to be contrasted with previous work on speech recognition. The comparison of the current psychoacoustical and modeling data argues nevertheless in favor of the existence of some form of (incomplete) normalization, whose effects are mainly visible for envelopes of long duration and highly contrasted envelopes (as produced by amplitude expansion). Detailed inspection of the longest envelopes indicates that they display more "local" features (i.e., primary and secondary peaks and troughs) than the shorter ones. This suggests that within each trial, perceptual constancy may be achieved, although imperfectly, by comparing the temporal sequences of envelope peaks and troughs across stimuli when these local features are numerous and salient enough. These conjectures warrant further experimental investigation.

In light of the present results, the resistance to temporal alterations of speech and music signals reported in previous studies may result from the operation of these normalization and template-matching processes. Compared to the stimuli of the current study, the higher level of redundancy of speech and music may account for the improvement in perceptual constancy with these stimuli.. In addition or alternatively, the possibility still remains that the reported constancy reflected partly—as suggested in Secs. I and V—a ceiling effect artifact.

VI. CONCLUSIONS

The current research investigated perceptual constancy in the temporal-envelope domain using nonlinguistic stimuli. Taken together, the psychophysical results indicate that the discrimination of temporally transposed envelopes degrades continuously as a function of the degree of temporal transposition. At least for moderate temporal expansion/compression rates, this deterioration is only slightly modulated by manipulations of stimulus parameters (frequency spacing between envelope components, peak-to-trough ratio of the envelopes) shown to influence the discrimination of (nontransposed) complex temporal envelope patterns.

A quantitative model of temporal envelope processing using a memory-limited envelope extraction stage followed by a template-matching process accounts for the discrimination of equal-duration envelopes, but generally underestimates listeners' discrimination of temporally transposed envelopes for the longest stimuli. This suggests that the auditory system applies some form of incomplete normalization to the temporal envelopes of incoming sounds, whether linguistic or nonlinguistic in nature.

ACKNOWLEDGMENTS

This research was supported by a MENRT grant to M. Ardoint, a grant from the Institut Universitaire de France to C. Lorenzi, an ANR grant (ANR-06-NEURO-022-01) to D. Pressnitzer, and an ANR grant (ANR-06-NEURO-042-01) to A. Gorea. The authors thank two anonymous reviewers for helpful comments on an earlier version of this manuscript.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M. (**2001**). "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," Proc. Nat. Acad. Soc. **98**, 13367–13372.

Apoux, F., Crouzet, O., and Lorenzi, C. (**2001**). "Temporal envelope expansion of speech in noise for normal-hearing and hearing-impaired listeners: Effects on identification performance and response times," Hear. Res. **153**, 123–131.

Daniloff, R., Shriner, T. H., and Zemlin, W. R. (**1968**). "Intelligibility of vowels altered in duration and frequency," J. Acoust. Soc. Am. **44**, 700–707.

Dau, T., Kollmieier, B., and Kohlrausch, A. (**1997a**). "Modeling auditory processing of amplitude modulation. I. Dectection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2893–2905.

Dau, T., Kollmieier, B., and Kohlrausch, A. (**1997b**). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," J. Acoust. Soc. Am. **102**, 2906–2919.

Derleth, R. P., Dau, T., and Kollmeier, B. (**2001**). "Modeling temporal and compressive properties of the normal and impaired auditory system," Hear. Res. **159**, 132–149.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.

Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**, 1181–1196.

Fairbanks, G., and Kodman, F. (**1957**). "Word intelligibility as a function of time compression," J. Acoust. Soc. Am. **29**, 636–641.

Fu, Q.-J., Galvin, J. J., and Wang, X. (**2001**). "Recognition of time-distorted sentences by normal-hearing and cochlear-implant listeners," J. Acoust. Soc. Am. **109**, 379–384.

Fu, Q. J., and Shannon, R. V. (**1999**). "Recognition of spectrally-degraded speech in noise with nonlinear amplitude-mapping," Proceedings of the 1999 IEEE (Institute of Electrical and Electronics Engineers) ICASSP (International Conference on Acoustics, Speech, and Signal Processing) , Vol. **1**, pp. 369–372.

Gockel, H., and Colonius, H. (**1997**). "Auditory profile analysis: Is there perceptual constancy for spectral shape for stimuli roved in frequency?" J. Acoust. Soc. Am. **102**, 2311–2315.

Gordon-Salant, S., and Fitzgibbons, P. F. (**2001**). "Sources of age-related recognition difficulty for time-compressed speech," J. Speech Lang. Hear. Res. **44**, 709–719.

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.

Li, X., and Pastore, R. E. (**1995**). "Perceptual constancy of a global spectral property: Spectral slope discrimination," J. Acoust. Soc. Am. **98**, 1956–1968.

Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**, 1–36.

Lorenzi, C., Berthommier, F., Apoux, F., and Bacri, N. (**1999**). "Effects of envelope expansion on speech recognition," Hear. Res. **136**, 131–138.

Lorenzi, C., Soares, C., and Vonner, T. (**2001**). "Second order temporal modulation transfer functions," J. Acoust. Soc. Am. **110**, 1030–1038.

MacMillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide* (Cambridge University Press, Cambridge).

Miller, J. L., and Volaitis, L. E. (**1989**). "Effect of speaking rate on the perceptual structure of a phonetic category," Percept. Psychophys. **46**, 505–512.

Moore, B. C. J., Shailer, M. J., and Schooneveldt, G. P. (**1992**). "Temporal modulation transfer functions for band-limited noise in subjects with co-

chlear hearing loss," Br. J. Audiol. **26**, 229–237.
Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1987**). "An efficient auditory filterbank based on the gammatone function," presented at the Meeting of the IOC Speech Group on Auditory Modeling at RSRE (Royal Signals and Radar Establishment), 14–15 December.
Sek, A., and Moore, B. C. (**2003**). "Testing the concept of a modulation filter bank: The audibility of component modulation and detection of phase change in three-component modulators," J. Acoust. Soc. Am. **113**, 2801–2811.
Sheft, S., and Yost, W. A. (**2005**). "Minimum integration times for processing of amplitude modulation," *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, edited by D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet (Springer, New York), pp. 244–250.
Sorkin, R. D., and Montgomery, D. A. (**1991**). "Effect of time compression and expansion on the discrimination of tonal patterns," J. Acoust. Soc. Am. **90**, 846–857.
Sorkin, R. D., Montgomery, D. A., and Sadralodabai, T. (**1994**). "Effect of sequence delay on the discrimination of temporal patterns," J. Acoust. Soc. Am. **96**, 2148–2155.
Summerfield, Q. (**1981**). "Articulatory Rate and Perceptual Constancy in Phonetic Perception," J. Exp. Psychol. Hum. Percept. Perform. **5**, 1074–1095.
Takeuchi, A. H., and Braida, L. D. (**1995**). "Effect of frequency transposition on the discrimination of amplitude envelope patterns," J. Acoust. Soc. Am. **97**, 453–460.
van de Par, S., and Kohlrausch, A. (**1998**). "Analytical expressions for the envelope correlation of narrow-band stimuli used in CMR and BMLD research," J. Acoust. Soc. Am. **103**, 3605–3620.
Vaughan, N. E., and Letowski, T. (**1997**). "Effects of age, speech rate, and type of test on temporal auditory processing," J. Speech Lang. Hear. Res. **40**, 1192–1200.
Versfeld, N. J., and Dreschler, W. A. (**2002**). "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners," J. Acoust. Soc. Am. **111**, 401–408.
Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.
Warren, R. M., Gardner, D. A., Brubaker, B. S., and Bashford, J. A. (**1991**). "Melodic and nonmelodic sequences of tones: effects of duration on perception," Music Percept. **8**, 277–290.