Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-format stimuli

Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli

Reinier W. L. Kortekaas and Armin Kohlrausch
Institute for Perception Research/IPO, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

(Received 8 July 1996; revised 14 October 1996; accepted 1 November 1996)

This article presents two experiments dealing with a psychoacoustical evaluation of the pitch-synchronous overlap-and-add (PSOLA) technique. This technique has been developed for modification of duration and fundamental frequency of speech and is based on simple waveform manipulations. Both experiments were aimed at deriving the sensitivity of the auditory system to the basic distortions introduced by PSOLA. In experiment I, manipulation of fundamental frequency was applied to synthetic single-formant stimuli under minimal stimulus uncertainty, level roving, and formant-frequency roving. In experiment II, the influence of the positioning of the so-called ''pitch markers'' was studied. Depending on the formant and fundamental frequency, experimental data could be described reasonably well by either a spectral intensity-discrimination model or a temporal model based on detecting changes in modulation of the output of a single auditory filter. Generally, the results were in line with psychoacoustical theory on the auditory processing of resolved and unresolved harmonics.
[S0001-4966(97)03903-9]

PACS numbers: 43.66.Fe, 43.66.Ba, 43.72.Ja [WJ]

INTRODUCTION

Over the past few decades, considerable research activities have concentrated on instrumental modification of the fundamental frequency and duration of natural speech. These types of modification enable the manipulation of speech prosody: modification of duration typically alters speech rhythm and tempo, whereas modification of fundamental frequency changes intonation. Characteristics not strictly pertaining to prosody, such as phonemic content and voice quality, ideally remain unaffected by these modifications. Numerous techniques have been proposed with the general aim of both maximizing intelligibility and perceived synthesis quality and minimizing computational complexity. A class of digital-signal-processing techniques with generally low complexity is the so-called overlap and add (OLA) framework (Rabiner and Schafer, 1978). For instance, time-domain OLA, where all operations are performed on the waveform itself, has been successfully applied not only in speech manipulation (e.g., Roucos and Wilgus, 1985) but also in other fields such as music synthesis (e.g., Roads, 1988).

This article will focus on pitch-synchronous overlap-and-add (PSOLA: Moulines and Charpentier, 1990; Moulines and Laroche, 1995), which is a variant of time-domain OLA.[1] The main feature of PSOLA is that the OLA operations are aligned to the (quasi)-periodicity of the input speech signal. PSOLA has found widespread application, e.g., in modules for text-to-speech synthesis and as a tool for fundamental speech-perception research (Moulines and Laroche, 1995). PSOLA-manipulated natural speech is generally characterized not only by high intelligibility but also by high synthesis quality. This finding is remarkable given the fact that, as will be described in the following section, the technique is based on rather rough signal operations.

Despite the generally satisfactory synthesis quality of PSOLA, annoying artefacts are sometimes introduced. Although a strict categorization is difficult, these artefacts can often be described as hoarseness and roughness of the synthesized signal. In addition, artefacts similar to comb filtering are observed in practice. As far as we know, the occurrence of these artefacts cannot be predicted beforehand. This unpredictability is, in our opinion, caused to a great extent by the lack of knowledge of the perceptual effects of the (PS)OLA operations, a view that was also expressed by Moulines and Laroche (1995).

Even if PSOLA manipulation of a speech signal does not lead to the perception of either of the artefacts mentioned above, the manipulation does affect its spectral content. To explain the success of PSOLA manipulation, one may hypothesize that these spectral changes are either perceptually subliminal or, within the context of speech perception, phonetically less relevant (cf. Klatt, 1982). This paper addresses the first hypothesis by determining the detectability of the spectral changes and by deriving the auditory cues involved in this detection process. Such a psychoacoustical basis is probably important for the long-term aim of increasing the predictability of audible (and annoying) artefacts. In addition, psychophysical evaluation may also increase knowledge about the auditory processing of speech.

In the experiments, synthetic single-formant signals (Klatt, 1980) were PSOLA manipulated. Apart from their application in speech synthesizers, single- or multiple-formant signals have been used to determine, e.g., jnd's in formant frequency (for a recent overview, see Lyzenga and Horst, 1995) and in fundamental frequency (Flanagan and Saslow, 1958; Klatt, 1973). Single-formant signals are used here to derive the sensitivity of the auditory system to the

''basic distortions'' introduced by the PSOLA operations. To establish the link to ''classical'' psychoacoustics, the Appendix presents experimental results concerning the auditory sensitivity to basic distortions when manipulating pure tones.

In experiment I, the perceptual effects of fundamental frequency ($F0$) manipulation for three levels of stimulus uncertainty, mimicking particular aspects of natural speech, are investigated. In experiment II, the perceptual effects of the ''pitch-marker'' location are studied (see Sec. I). Both experiments focus on $F0$ modification only because, in practice, this type of manipulation is more likely to result in annoying artefacts than the manipulation of duration. In addition, the experimental data are compared with predictions of two model simulations: a model based on detecting intensity differences between excitation patterns (Durlach *et al.*, 1986; Gagné and Zurek, 1988) and a model based on the discrimination of modulation depth within a single auditory filter.

I. GENERAL METHODS

A. The PSOLA technique

The PSOLA technique is a time-domain variant of the so-called overlap–add (OLA) technique for analysis–synthesis (Rabiner and Schafer, 1978; Allen and Rabiner, 1977). OLA generally consists of three steps: (1) decomposition of a signal into separate, but often overlapping, segments; (2) optional modification of these segments; and (3) recombination of the segments by means of overlap–adding. PSOLA consists only of steps (1) and (3). A short introduction to PSOLA will be presented here; for further details the reader is referred to Moulines and Charpentier (1990) and Moulines and Laroche (1995).

Figure 1(a) shows the waveform of a synthetic single-formant signal as used in both experiments. This signal is decomposed into separate segments in analysis step (1) by windowing it at particular time instances. These instances, represented by vertical lines in Fig. 1(a), are positioned *pitch synchronously* and are called ''pitch markers.'' Pitch markers are determined either manually by inspection of the speech waveform or automatically by means of some local $F0$ estimation (e.g., Ma *et al.*, 1994; Smits and Yegnanarayana, 1995). Figure 1(b) shows two segments extracted from the input signal. The maxima of the Hanning (raised-cosine) windows coincide with the pitch markers. The window duration depends on the temporal spacing between pitch markers; consecutive windows have 50% overlap. Because adjacent windows samplewise add up to one, the input signal can be restored perfectly. Note that in natural speech windows will typically be asymmetrical due to variation in $F0$.

Segment recombination in synthesis step (3) is performed after *defining* a new pitch-marker sequence. In Fig. 1(c), the new sequence is represented by vertical lines. An output signal is synthesized by first assigning a decomposed segment to each of the new pitch markers and then performing the samplewise overlap–add operation. Manipulation of fundamental frequency is achieved by changing the time intervals between pitch markers. In Fig. 1(c), for instance, these intervals are increased, leading to the percept of a lower pitch. Modification of duration, on the other hand, is achieved by either repeating or omitting segments. Note that, in principle, modification of fundamental frequency also implies a modification of duration.

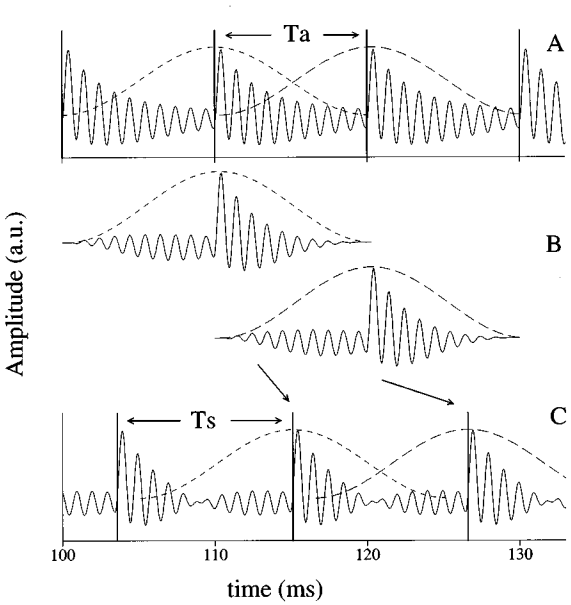

FIG. 1. Illustration of the PSOLA technique: Panel (a) shows the waveform of a synthetic 1000-Hz single-formant signal with a fundamental of 100 Hz. At the pitch-marker locations, indicated by thick vertical lines, the signal is decomposed by means of Hanning windowing. The interval between two pitch markers is indicated by T_a. Two segments are shown in panel (b). These segments are recombined by means of overlap-adding at the new pitch-marker positions indicated by thick vertical lines in panel (c). These pitch markers are regularly spaced at 11.5 ms, indicated by T_s, which gives a fundamental frequency of 87 Hz.

B. Terminology

In the experiments manipulation was investigated for signals having a constant $F0$. This means that the pitch markers in the decomposition and synthesis phase are positioned at regular intervals. These intervals will be denoted by T_a and T_s, respectively. Analogous with the fundamental frequency, we introduce the ''window rates'' $F_{wa} = 1/T_a$ and $F_{ws} = 1/T_s$. In the experiments the analysis window rate F_{wa} was fixed and F_{ws} was the experimental parameter. In what follows, experimental results will be presented as a function of ΔF given by

$$\Delta F = \frac{F_{ws} - F_{wa}}{F_{wa}} \times 100\%.$$

For positive and negative values of ΔF, the symbols ΔF^+ and ΔF^- will be used.

Under some experimental conditions the perceptual effects of pitch-marker location were investigated. The pitch-marker location will be denoted by the parameter ΔP. As will be described in Sec. II A 1, the single-formant signals are generated by exciting a formant filter with a regular pulse train. The parameter ΔP indicates the shift of the pitch markers relative to the excitatory pulses. This shift will be given as a percentage of T_a. In Fig. 1(a), for instance, the pitch markers coincide with the formant-filter excitations so that

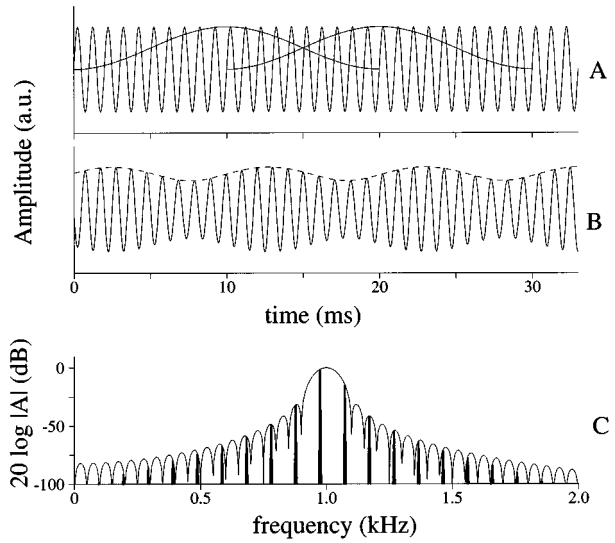

FIG. 2. A pure-tone signal of $f_c=1000$ Hz, shown in panel (a), is decomposed into segments by windowing the signal at a rate of $F_{wa}=100$ Hz. These segments are recombined at a rate of $F_{ws}=97.56$ Hz to synthesize the signal in panel (b). The dashed line represents the Hilbert envelope of the synthesized signal. Panel (c) shows the log-amplitude spectra of a single segment (thin solid line) and the synthesized signal (thick vertical lines).

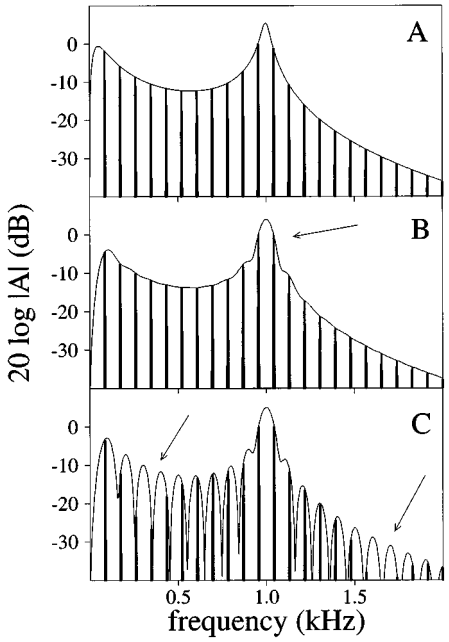

FIG. 3. Panel (a) shows the log-amplitude spectrum of a 1000-Hz single-formant signal with a fundamental of 87 Hz. The thin line represents the amplitude transfer function of the formant filter including pre-emphasis. Panel (b) shows the log-amplitude spectrum for a PSOLA-manipulated single-formant signal, shifted in fundamental frequency from 100 to 87 Hz. Parameter ΔP was set to 0%. The thin solid line now represents the log-amplitude spectrum of a single segment decomposed from the input signal. Arrows indicate frequency regions of maximal difference between the spectra in panels (a) and (b). Panel (c) shows the spectrum of a PSOLA-manipulated signal but now with ΔP set to 50%. Arrows indicate notches introduced in the log-amplitude spectrum.

$\Delta P=0\%$. Because the formant filter is minimum phase, the amplitude maxima of the signal in Fig. 1(a) are only slightly delayed relative to the maxima of the Hanning window. On the other hand, if $\Delta P=50\%$ the pitch markers are located *between* excitations of the formant filter. In that case maxima of the input signal and the Hanning windows are maximally misaligned.

C. Distortions in pure tones

First, we consider PSOLA manipulation of a single pure tone which is thought of as a component of a harmonic spectrum. Figure 2(a) depicts a pure tone of carrier frequency $f_c=1000$ Hz, assumed to be the tenth harmonic of a 100-Hz fundamental. After decomposition at intervals of $T_a=10$ ms and overlap–adding, the signal shown in Fig. 2(b) is synthesized where $\Delta F=-2.44\%$ ($F_{ws}=97.56$ Hz, $T_s=10.25$ ms). In contrast to the original pure tone, this signal shows amplitude modulation (AM) in its envelope and frequency modulation (FM) in its fine structure. For a sinusoidal input signal these two changes are the basic distortions introduced by PSOLA. Experimental results relating to the auditory sensitivity to these distortions will be presented in the Appendix. The AM of the envelope is partly caused by the fact that adjacent Hanning windows do not sum up to one if $T_a\neq T_s$. This can be compensated for by using a ''synthesis window.'' The perceptual relevance of using such a window will be discussed in Sec. III.

Alternatively, we can describe the distortions in the spectral domain. Time-domain multiplication (windowing) results in frequency-domain convolution of the spectra of the Hanning window and the pure tone (e.g., Rabiner and Schafer, 1978). The thin solid line in Fig. 2(c) depicts the log-amplitude spectrum of a single segment decomposed from the original pure tone of Fig. 2(a). The overlap–adding operation in synthesis, which extends the signal periodically in the time domain, is equivalent to *resampling* the (complex) spectrum of a single segment (Moulines and Laroche, 1995). The log-amplitude spectrum of the synthesized signal in Fig. 2(b) is shown by the line spectrum in Fig. 2(c). The spectral lines are harmonics of $F_{ws}=97.56$ Hz. For example, the strongest harmonic has a frequency of $10\times F_{ws}=975.6$ Hz. In other words, the introduction of AM and FM has a spectral counterpart in terms of the interaction of introduced components (Goldman, 1948).[2]

D. Distortions in single-formant signals

The experiments deal with the discrimination of PSOLA-manipulated and -unmanipulated single-formant signals. Such signals intrinsically have a harmonic structure so that the introduction of side components *per se* cannot be a cue for discrimination. In fact, cues may be changes in both spectral envelope and phase relations between harmonics. These changes will be illustrated below.

Figure 3(a) shows the log-amplitude spectrum of an unmanipulated single-formant signal with an $F0$ of 87 Hz, a formant frequency of 1000 Hz, and a formant bandwidth of 50 Hz. This spectrum also shows the effects of pre-emphasis applied to the formant signal (see Sec. II A 1). The corresponding phase spectrum (not shown here) is approximately linear except for a phase jump of π rad around the formant frequency.

Figure 3(b) shows the log-amplitude spectrum of a PSOLA-manipulated signal obtained by generating a 1000-Hz formant signal with an $F0$ of 100 Hz, decomposing it at $F_{wa}=100$ Hz, and resynthesizing it at $F_{ws}=87$ Hz. Here, ΔP is set to 0%. The Hanning-windowing operation has ''smeared out'' the spectral envelope: the bandwidth of the pronounced formant is increased to approximately F_{wa} Hz. The spectral slope, however, remains almost unaffected. Changes to the phase spectrum (not shown) are a phase shift of approximately $\pi/4$ rad for the two harmonics around the formant frequency. Figure 3(c) shows the spectrum of a signal synthesized with ΔP set to 50%. Its spectral envelope is clearly discontinuous which introduces pronounced notches in spectral envelope after resampling. The notch depth depends monotonically on ΔP. The corresponding phase spectrum (not shown) is discontinuous as well.

II. EXPERIMENTS

The main questions in these experiments are: (1) What are the thresholds for the discrimination of PSOLA-manipulated and -unmanipulated single-formant signals; (2) what is the influence of pitch-marker location on discrimination performance; and (3) how do the discrimination results relate to psychoacoustical models.

A. Method

1. Stimuli

For generation of the single-formant signals, a second-order digital resonator was implemented as proposed by Klatt (1980).[3] This filter was excited by a pulse train with an $F0$ of 100 or 250 Hz. The window rate F_{wa} was accordingly set to these values. Low-pass characteristics of natural voicing and high-pass radiation at the mouth opening, as described in Klatt (1980), were included as pre-emphasis. Formant frequencies f_r were 500, 1000, and 2000 Hz with -3-dB bandwidths of 50, 50, and 100 Hz, respectively.

As a baseline experiment, a minimal stimulus-uncertainty condition was investigated in which f_r was fixed and the overall level L was set to 70 dB SPL. To increase stimulus uncertainty, level roving between intervals was applied in the second condition. The level rove was uniformly distributed in the range ± 5 dB. As a third condition, the overall level L was fixed but the formant frequency f_r was roved uniformly over a range of $\pm 2\,\Delta f_r$. Here, Δf_r denotes one jnd in formant frequency for which Gagné and Zurek (1988) reported the following relation: $\Delta f_r = 0.079 f_r/\sqrt{Q}$, where Q is the Q factor of the formant filter. The range of f_r roving was within the range measured in a study by Pisoni (1980) in which subjects were instructed to ''reproduce'' steady-state synthetic vowels.

To investigate the perceptual effects of pitch-marker positioning, ΔP was set to 0% and 50% in experiment I. In experiment II, psychometric functions for ΔP were determined for two particular values of ΔF.

Stimuli were software generated on a Silicon Graphics Indigo workstation. The sampling frequency was 32 kHz. Apart from the built-in filters of the workstation, no additional anti-aliasing filtering was applied. Because the amplitude spectrum of the single-formant signals monotonically falls off to approximately -80 dB (relative to the formant peak) at the Nyquist frequency, no aliasing is to expected. After DA conversion signal levels were adjusted by means of analog attenuation. Stimuli were presented to the subject, seated in a soundproof booth, over Beyer DT 990 headphones. Subjects responded via a keyboard and received immediate feedback. Stimulus duration was 300 ms, the first and last 25 ms were ramped using a Hanning window. The interval separation was 200 ms.

2. Procedure

Psychometric functions were measured using a 3I3AFC odd-ball procedure with fixed levels of ΔF in each run. The odd-ball interval contained the PSOLA-manipulated single-formant signal. This signal was obtained by (1) generating a formant signal with an $F0$ of F_{wa} Hz; (2) decomposing this signal at a window rate of F_{wa} Hz; and (3) resynthesizing it at a rate of F_{ws} Hz. The reference intervals contained a single-formant signal generated directly with an $F0$ of F_{ws} Hz. For determination of the psychometric function, F_{ws} was varied according to

$$F_{ws}=\frac{1}{T_a+(n/4)}[\text{Hz}], \tag{1}$$

where $n = \pm 1$ ms, ± 2 ms, ± 3 ms, ...

Each run consisted of 15 trials. For each condition, i.e., a combination of ΔF, F_{wa}, and f_r, a total of five runs were performed of which the first run was omitted from the analysis. Each data point thus represents 60 trials. All conditions were measured once before the next set of runs was initiated. Mean values and standard deviations of the four runs are shown in the figures below. Instead of plotting percentage correct as a function of ΔF, the Pc values were converted to d' using a conversion table (MacMillan and Creelman, 1991).

3. Subjects

Three subjects (aged 25, 27, and 35) participated in the experiments. All subjects had normal pure-tone thresholds in quiet for the frequencies 500, 1000, and 2000 Hz. Unlike subject RK (the first author), subjects MB and KM had no or little experience in psychoacoustic listening experiments. All subjects performed experiments I and II for $f_r=1000$ Hz. Subjects KM and RK performed experiment I for $f_r=2000$ Hz. Results for $f_r=500$ Hz (experiment I) were obtained only for subject RK.

B. Experiment I: Influence of ΔF

1. Minimal stimulus uncertainty

Psychometric functions for minimal stimulus uncertainty are shown in the left-hand panels of Figs. 4–6. Figure 4 presents the data for $F_{wa}=100$ Hz and $f_r=1000$ Hz. Data points for $\Delta P=50\%$, indicated by filled squares, are generally far above the threshold $d'=1$ for all subjects. For $\Delta P=0\%$, however, the psychometric functions show a non-monotonic behavior. For all three subjects, subthreshold discrimination performance is found for $\Delta F=-16.66\%$,

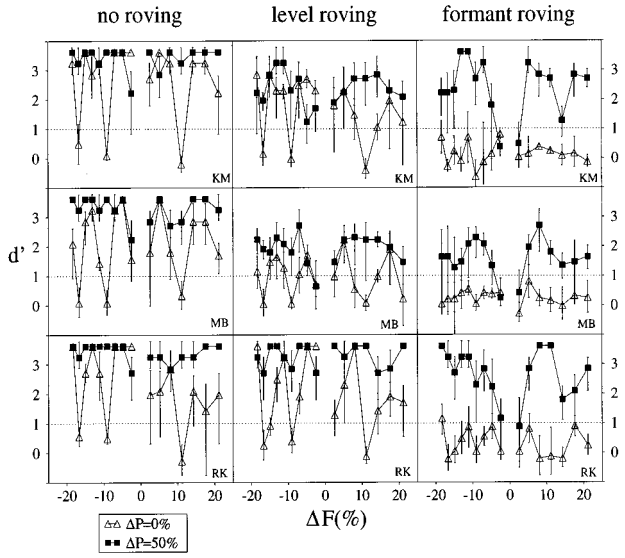

FIG. 4. Psychometric functions for discrimination of PSOLA-manipulated and unmanipulated single-formant signals with formant frequency $f_r=1000$ Hz and $F_{wa}=100$ Hz. Mean data for $\Delta P=0\%$ are shown by triangles, those for $\Delta P=50\%$ by filled squares. Standard deviations are indicated by vertical lines. Left-hand panels show results for the minimal-stimulus-uncertainty condition. Center panels show psychometric functions for the level-roving condition, and the right-hand panels data for roving of f_r.

-9.09%, and $+11.11\%$. These values correspond to values for T_s of 12, 11, and 9 ms, respectively, while T_a equals 10 ms. Because the 50-Hz formant bandwidth is rather small, it can be stated that each decomposed segment contains 20 periods of a 1000-Hz carrier (cf. Fig. 1). Setting T_s to an integer multiple of 1 ms, which is the period of the 1000-Hz carrier, thus results in an *in-phase* addition of the fine structure of adjacent windows. This results in minimal distortion of the temporal envelope of the signal. In spectral terms, setting T_s to an integer multiple of the carrier period results in a harmonic coinciding with the formant frequency, due to the resampling property of PSOLA.

Using the same line of reasoning for the case of

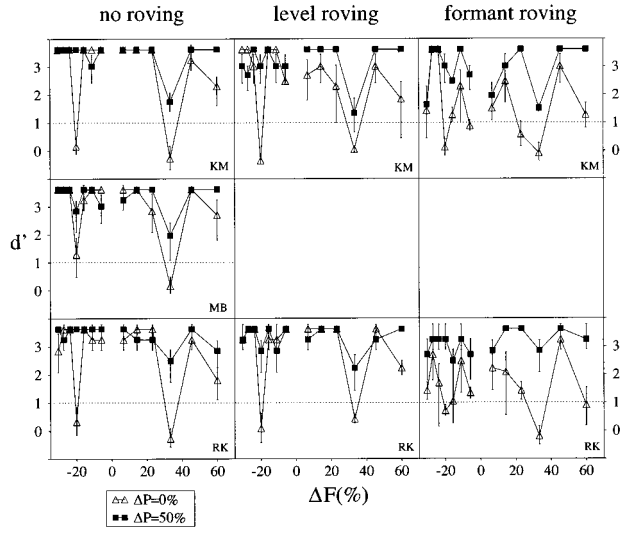

FIG. 5. Psychometric functions as in Fig. 4 but now for $f_r=1000$ Hz and $F_{wa}=250$ Hz.

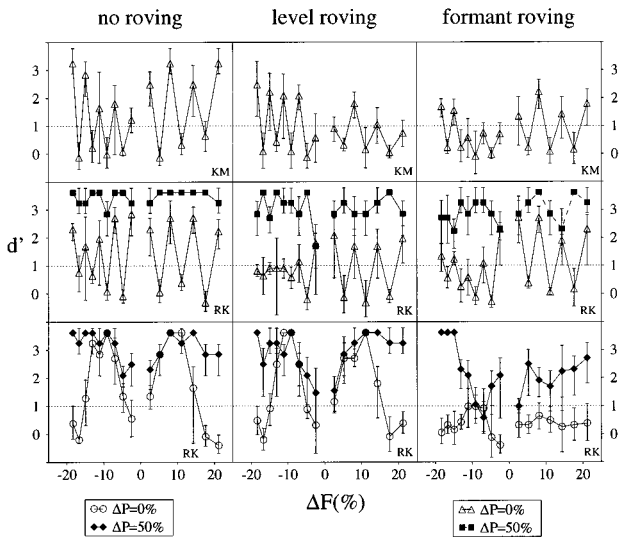

FIG. 6. Psychometric functions as in Figs. 4 and 5. Top and middle panels: psychometric functions for $f_r=2000$ Hz and $F_{wa}=100$ Hz. Bottom panels: psychometric function for $f_r=500$ Hz. Circles here are data for $\Delta P=0\%$, filled diamonds data for $\Delta P=50\%$.

$f_r=1000$ Hz and $F_{wa}=250$ Hz, subthreshold discrimination performance is predicted for $\Delta F=-20\%$ and $+33.33\%$. This is confirmed by the psychometric functions shown in the left-hand panels in Fig. 5. Note also that the data for other ΔF values show ceiling effects to even greater extent than the data in Fig. 4.

Data for the minimal stimulus-uncertainty conditions for $f_r=500$ and 2000 Hz are shown in the left-hand panels of Fig. 6 (subjects KM and RK only). The argumentation presented above also holds for these two formant frequencies. For $f_r=2000$ Hz, the general shapes of the psychometric functions are similar for both subjects, although the subjects are seen to have unequal difficulty in discrimination for ΔF^-.

2. Level roving

The center panels of Figs. 4–6 show psychometric functions for level roving. Starting with Fig. 4, i.e., $f_r=1000$ Hz and $F_{wa}=100$ Hz, it can be observed that discrimination performance is generally deteriorated relative to the minimal stimulus-uncertainty condition. Nevertheless, the general pattern of discrimination behavior is almost unaffected: The psychometric function for $\Delta P=50\%$ (squares) is almost always considerably above threshold. Moreover, the subthreshold data points for $\Delta P=0\%$ occur for the same ΔF values. Of all subjects, the performance of subject MB is seen to be affected most.

Level roving does not have a great influence on discrimination performance for $F_{wa}=250$ Hz and $f_r=1000$ Hz, as shown in Fig. 5. The middle panels of Fig. 6 show the psychometric functions for $f_r=2000$ and 500 Hz ($F_{wa}=100$ Hz). For $f_r=2000$ Hz, discrimination performance is reasonably affected, especially for ΔF^- for subject RK. On the other hand, performance for $f_r=500$ Hz is as good with level roving applied as without level roving.

3. Formant-frequency roving

The right-hand panels in Fig. 4 show that, for $f_r=1000$ Hz with $F_{wa}=100$ Hz and $\Delta P=0\%$, discrimination performance under f_r roving drops below $d'=1$ for *all* ΔF shifts. For $\Delta P=50\%$, performance is generally only moderately affected. A similar trend is observed for $f_r=500$ Hz (Fig. 6) but, here, performance is also deteriorated for $\Delta P=50\%$ for some values of ΔF. For $f_r=2000$ Hz, the alternating pattern is still observable for ΔF^+ if $\Delta P=0\%$. For ΔF^-, however, discrimination performance drops below threshold for almost all values. As is shown in the right panels of Fig. 5, roving of f_r has a moderate influence on performance for $F_{wa}=250$ Hz.

4. Discussion

The psychometric functions show a clear interaction between f_r, F_{wa}, and F_{ws}. The pattern of these functions is not greatly influenced by roving of overall level which also occurs in natural speech. Roving of formant frequency, on the other hand, can drastically affect performance in the sense that the distortions introduced by PSOLA apparently are no longer usable cues for discrimination. This suggests that the non-steady-state nature of natural speech may explain part of the success of PSOLA. Because setting ΔP to 50% provides strong and stable discrimination cues, the next experiment aims at determining discriminability as a function of ΔP.

C. Experiment II: Variation of the pitch-marker position

Psychometric functions as a function of ΔP were measured for $f_r=1000$ Hz and $F_{wa}=100$ Hz. Two values of the $F0$ shift were selected for which results for $\Delta P=0\%$ were below threshold: $\Delta F=-9.09\%$ and $+11.11\%$ (cf. Fig. 4). The parameter ΔP was varied in equal steps between -50% and 50% (note that these two values are identical for strictly periodic signals). Psychometric functions were obtained for the minimal stimulus-uncertainty conditions (all subjects) and for both roving conditions ($\Delta F=-9.09\%$, and subjects KM and RK only).

Figure 7 shows the data for $\Delta F=-9.09\%$ (squares) and $+11.11\%$ (circles) for the three stimulus-uncertainty conditions. The data in Fig. 7 do not show systematic differences for the two ΔF values. Also, the psychometric functions are seen to be symmetric around $\Delta P=0\%$. Thresholds are approximately reached at $|\Delta P|=25\%$, which means that pitch markers do not necessarily have to coincide exactly with either the filter excitation or the signal energy maximum. Moreover, these thresholds are reasonably stable under level and formant-frequency roving. The psychometric functions for subject KM become shallower with increasing stimulus uncertainty.

D. Model predictions

1. Intensity discrimination

Gagné and Zurek (1988) used an intensity-discrimination model (Florentine and Buus, 1981) to account for jnds in the resonance frequency of a single resonator. A

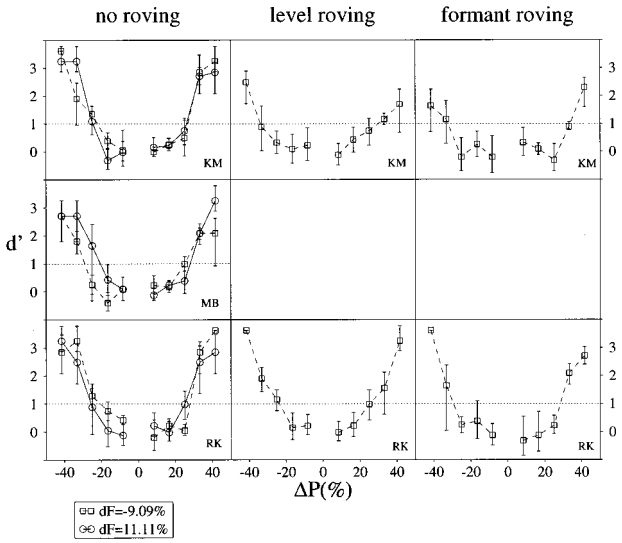

FIG. 7. Psychometric functions for $F_{wa}=100$ Hz and $f_r=1000$ Hz with ΔP as experimental parameter. Squares are data for $\Delta F=-9.09\%$, circles data for $\Delta F=+11.11\%$. Data for $\Delta P=0\%$ and $\pm 50\%$ have already been shown in Fig. 4.

model of this kind takes only spectral cues into account. The model is based on a channelwise determination of level differences between the excitation patterns of a reference and a signal. Channels here refer to critical bands. The model assumes that the partial sensitivity d_i' in channel i is proportional to the level difference $\Delta L_{E,i}$ between the two excitation patterns: $d_i'=k\cdot\Delta L_{E,i}$, where k is a constant which is the same for all channels. The overall sensitivity d' is derived from the partial sensitivities. In the single-band version of the model overall sensitivity d' is equal to the maximum of the partial sensitivities:

$$d'=\max_{i=1,N}(d_i')=k\cdot D_{\max}, \tag{2}$$

where N is the number of channels. In the multiband version partial sensitivities are optimally combined (Durlach *et al.*, 1986):

$$d'=\left(\sum_{i=1}^{N} d_i'^2\right)^{1/2}=k\cdot D_{\text{sum}}. \tag{3}$$

Gagné and Zurek found that resonance-frequency jnds could be best described by the single-band version of the model. In the present study both the single- and the multi-band model were implemented as a Gammatone filterbank (Patterson *et al.*, 1987). Simulation of absolute hearing threshold was included by adding an ''internal noise'' value to the power estimate at the output of each filter (cf. Moore and Glasberg, 1987).

According to formulas 2 and 3, d' is linearly related to $D_{\max}$ or D_{sum}. The predictive power of the models can thus be investigated by performing a linear regression on the experimental d' data in dependence on either $D_{\max}$ or D_{sum}. The regression equations were forced to intersect with the origin. As a measure of goodness of fit, the amount of explained variance, as expressed by the square of the correlation coefficient r, will be used.

TABLE I. For both intensity-discrimination models, the square of the correlation coefficient of the linear regression, r^2, is tabulated for each of the subjects. Here, the ΔP experimental data for the minimal stimulus-uncertainty condition are used.

	Single band	Multiband
KM	0.77	0.84
MB	0.78	0.82
RK	0.74	0.81

First, the data of ΔP experiment II were used for finding the slopes k of the linear regression equations, for each subject individually. The minimal stimulus-uncertainty data from Fig. 7 for both values of ΔF were used for this regression. Data points were discarded if $d' \geq 3.62$, i.e., $Pc \geq 99\%$. The linear regression results, in terms of r^2, are presented in Table I. For all subjects, the multi-band model yields the highest value of r^2. Differences between the two models are, however, small. All regression slopes are significantly different from zero at the $p<0.0001$ level. The slopes k for the multi-band model are: 0.18, 0.15, and 0.17 for subjects KM, MB, and RK, respectively. The threshold value for D_{sum}, yielding $d'=1$, is thus approximately equal to 6 dB. This value is a factor 2.5 higher than the value reported in Gagné and Zurek (1988). For the single-band model, the slopes k are 0.31, 0.24, and 0.28, respectively. At threshold $D_{\max} \simeq 3.5$ dB, which is also a factor 2.5 higher than the value reported by Gagné and Zurek.

Second, the data of ΔF experiment I were used for linear regression for D_{sum}. Here again, only the data for the minimal stimulus-uncertainty conditions (left-hand panels in Figs. 4–6) were used, with $\Delta P=0\%$. The results are listed in Table II. For $F_{wa}=100$ Hz, r^2 is in the range 0.4–0.6. The slopes k are similar across subjects and are approximately 1, 0.8, and 1.4 for $f_r=500$, 1000, and 2000 Hz, respectively. The threshold value for D_{sum} yielding $d'=1$ is thus approximately 1 dB, which is clearly at variance with the results for the ΔP experiment. For $F_{wa}=250$ Hz, however, k is approximately equal to 0.16, which is in good agreement with the slope found for the ΔP experiment. The high p levels, $p>0.1$, for subjects KM and MB are probably due to the small number of data points resulting from ceiling effects in the ΔF experiment. As nearly all data points were far above threshold for $\Delta P=50\%$, the corresponding r^2 values did not exceed 0.3 (not listed in Table II).

2. Modulation discrimination

This model is based on the discrimination of amplitude-modulation depth in the envelope of a single auditory-filter output. The auditory filter was simulated by a single Gammatone filter having a bandwidth of 1 ERB. The center frequency f_{cf} was varied over the range $[f_r - F_{WA}, f_r + F_{WA}]$, simulating off-frequency listening, in order to find the maximum difference between reference and signal. The maximum distances were mainly observed for filters centered at the boundaries of the f_{cf} range. The modulation indices M_{ref} and M_{sig} were calculated at the output of the filter.[4] The model is based on the assumption that sensitivity d' is governed by

$$d' = k \cdot |M_{\text{ref}}^2 - M_{\text{sig}}^2| = k \cdot D_{\text{mod}}, \tag{4}$$

where k is some constant. Moore and Sek (1992) found that, for low modulation rates (below 10 Hz), *detection* sensitivity was linearly related to the square of the modulation index. Wakefield and Viemeister (1990) used sinusoidally amplitude-modulated (SAM) noise and found almost linear relations between d' and $M_{\text{ref}}^2 - M_{\text{sig}}^2$. Because M_{ref}^2 may be smaller than M_{sig}^2 for the present signals, the absolute value of the difference was taken.

TABLE III. Values of r^2 for linear regression on the d' data of the ΔF experiment using the modulation-discrimination model. Only data for $F_{wa}=100$ Hz are shown. Significance levels p are indicated as in Table II.

f_r (Hz)	500	1000	2000
KM		0.52**	0.69***
MB		0.35*	
RK	0.33*	0.59**	0.92***

Table III shows the r^2 values for linear regression of D_{mod} on the experimental d' data of the ΔF experiment ($F_{wa}=100$ Hz and $\Delta P=0\%$ only). For $f_r=1000$ Hz and $\Delta P=0\%$, the slope k is found to be approximately 7 for all subjects. This corresponds to a modulation-discrimination threshold of 0.14 for D_{mod}. The explained variance for subject MB, however, is rather low. For $f_r=2000$ Hz, slope k is about 4 so that the threshold would be at $D_{\text{mod}}=0.25$. These D_{mod} values are in reasonable agreement with the data reported in Wakefield and Viemeister (1990) for SAM noise, provided M_{ref} is large, i.e., $10 \log (M_{\text{ref}}^2) \geq -5$ dB. For the present signals, M_{ref} is indeed in this range.

III. DISCUSSION

Although not explicitly verified experimentally, $|\Delta F|$ shifts as small as approximately 2% may lead to detectable distortions, as can be inferred from the region around $\Delta F=0\%$ in the psychometric functions in Figs. 4, 5, and 6. This finding agrees with the results presented in the Appendix for manipulation of pure tones. Remarkably, the psychometric functions for minimal stimulus uncertainty and level roving,

TABLE II. Values of r^2 for linear regression on the d' data of the ΔF experiment using the multiband intensity-discrimination model. Significance levels p, indicating the probability that the slope of the linear regression equation is equal to zero, are indicated as follows: *$p<0.05$, **$p<0.01$, ***$p<0.001$.

f_r (Hz)	500	1000	2000	1000 ($F_{wa}=250$ Hz)
KM		0.48*	0.45**	0.54
MB		0.56**		0.55
RK	0.61**	0.54**	0.40**	0.79**

provided $\Delta P = 0\%$, were nonmonotonic, revealing a clear interaction between f_r, F_{wa}, and F_{ws}. This finding is not in agreement with the intuitive expectation that distortions are more easily detectable for larger shifts in $F0$. We will try to explain the discrimination results in terms of spectral and temporal cues by first comparing the present results with data from the literature and then discussing our modeling results.

A. Comparison with the literature

As a result of PSOLA manipulation with $\Delta P = 0\%$, changes in the intensity of spectral components, in combination with phase shifts, occur in the spectral region of f_r, as was illustrated in Fig. 3. Changes in component intensities also occur due to changes in formant frequency. A number of studies (e.g., Gagné and Zurek, 1988; Kewley-Port and Watson, 1994; Lyzenga and Horst, 1995; Sommers and Kewley-Port, 1996) have explained formant-frequency jnd's in terms of profile analysis, i.e., in terms of discrimination of spectral shape (Richards *et al.*, 1989; Zera *et al.*, 1993). With minimal stimulus uncertainty and for $F_{wa} = 100$ Hz in the present study (absolute), component-level differences between signal and reference maximally amount to 2.5, 2.5, and 1.5 dB for $f_r = 500$, 1000, and 2000 Hz, respectively. These values are valid for the range of ΔF investigated in the experiments. The lower value for $f_r = 2000$ Hz is a consequence of the larger formant bandwidth of 100 Hz.

Thresholds for the detection of level increments of single components of a complex tone of equal-amplitude harmonics were reported by Zera *et al.* (1993). For complex tones consisting of 60 harmonics of 100 Hz, level-increment thresholds for harmonics at 500, 1000, and 2000 Hz were found to be approximately 2, 2.5, and 4 dB, respectively. This means that for $f_r = 500$ and 1000 Hz, the level changes of individual harmonics due to PSOLA manipulation would be near detection threshold. For $f_r = 2000$ Hz, level changes would be below threshold. The psychometric functions in Figs. 4 and 6, however, showed that discrimination sensitivity generally was above $d' = 2$. For the fifth harmonic of 200 Hz, on the other hand, Henn and Turner (1990) and Zera *et al.* (1993) reported level-increment thresholds of 2 dB. For $F_{wa} = 250$ Hz and $f_r = 1000$ Hz, spectral-envelope level differences are maximally 10 dB so that level differences of single components were potential cues for discrimination.

Instead of just a single harmonic, however, the intensities of a number of harmonics are changed both as a result of PSOLA manipulation and by changing the formant frequency. Sommers and Kewley-Port (1996) found that salient cues for formant-frequency jnds were mediated by the level changes of the three harmonics closest to the formant frequency. Using an excitation pattern model (Moore and Glasberg, 1987), they also found that formant-frequency jnds under different conditions resulted in more or less constant level differences between excitation patterns. Sommers and Kewley-Port (1996) only investigated formants at 500 and 1350 Hz with an $F0$ of 200 Hz, so that harmonics around the formant frequencies were likely to be resolved. Particularly for $f_r = 2000$ Hz and $F_{wa} = 100$ Hz in the present study, harmonics are unresolved so that temporal cues may have been used for discrimination. Lyzenga and Horst (1995), who used formantlike signals around 2000 Hz with an $F0$ of 200 Hz, also proposed a temporal mechanism, at least for part of their jnd data.

Roving of the *overall* stimulus level affected discrimination performance only to a small degree. This finding is in agreement with results presented by Farrar *et al.* (1987) on formant-frequency discrimination using noise sources as input to the Klatt synthesizer. Because the overall level of the intervals was normalized in the minimal stimulus condition, differences in loudness between signal and reference may have been a cue (cf. Lyzenga and Horst, 1995). Taking into account the rather small spectral-envelope differences mentioned above, however, it is more likely that discrimination performance under level roving is affected by the increase in distracting stimulus uncertainty.

The present results show that f_r roving can affect discrimination performance considerably. Roving of f_r results in spectral-envelope level differences near f_r, not only between signal and references, but also between references. The distribution of level differences has a standard deviation of approximately 4 dB for all f_r values and bandwidths under consideration. This value is of the same order of magnitude as the spectral-envelope differences introduced by PSOLA for $F_{wa} = 100$ Hz (see above). This means that if the harmonics around f_r are resolved, excitation-pattern differences between the two references are comparable to the differences between signal and references due to PSOLA. Discrimination performance can then be expected to drop below $d' = 1$, as observed for $f_r = 500$ and 1000 Hz. For $F_{wa} = 250$ Hz, level differences due to PSOLA can exceed those due to f_r roving (see above) so that discrimination is expected to be, at most, moderately influenced, which is in agreement with the experimental data. If, on the other hand, components are unresolved, as for $f_r = 2000$ Hz and $F_{wa} = 100$ Hz, the differences in the phase spectra between signal and references may become a cue. In other words, the effect of (in)coherent addition of subsequent segments is preserved in peripheral filtering. For ΔF^+, roving of f_r did indeed not deteriorate discrimination performance. It is not clear, however, why performance dropped below $d' = 1$ for ΔF^-.

In an additional, informal experiment, the phases of the components of the single-formant signal were randomized. For $f_r = 500$ Hz, phase randomization had only a small effect on discrimination performance. This was also observed for $f_r = 1000$ Hz with $F_{wa} = 250$ Hz. For $f_r = 1000$ and 2000 Hz ($F_{wa} = 100$ Hz), however, discrimination performance was below $d' = 1$. This provides additional evidence for the hypothesis that, for the latter two conditions, temporal cues played a dominant role.

In experiment II, the detection threshold was found to be $|\Delta P| \simeq 25\%$. The spectral ''notch depth'' at this value of $|\Delta P|$ is approximately 3 dB. Turner and Van Tasell (1984) found comparable thresholds for a notch with linear flanks on a dB scale, centered at 2120 Hz within the spectrum of a synthetic vowel with 120-Hz fundamental. If intensity discrimination determines detectability, however, then lower ΔP thresholds are to be expected if the components in the notches are resolved. The results of informal tests for $F_{wa} = 250$ Hz confirmed this expectation: with minimal

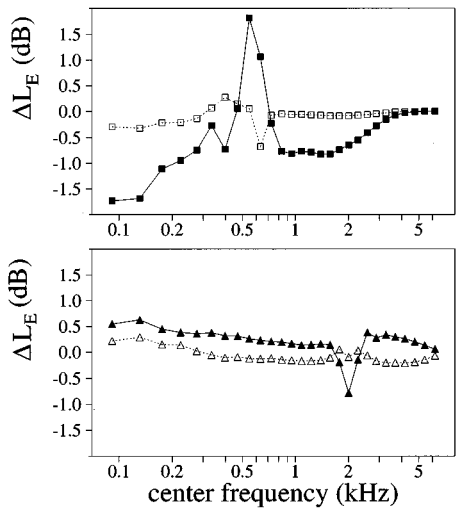

FIG. 8. Illustration of excitation-pattern level differences between unmanipulated and PSOLA-manipulated single-formant signals. The top panel shows the level differences $\Delta L_{E,i}$ for $f_r=500$ Hz. Filled squares indicate level differences for $\Delta F=-11.1\%$, open squares for $\Delta F=-2.4\%$. The corresponding values of D_{sum} are 1.9 and 0.5 dB, respectively. The bottom panel shows level differences for $f_r=2000$ Hz, where the filled and open triangles indicate differences for $\Delta F=8.1\%$ and 11.1%, respectively. For these ΔF values, D_{sum} is 0.8 and 0.5 dB, respectively.

stimulus uncertainty, ΔP thresholds were considerably smaller than 25%.

B. Models

In contrast to the findings of Gagné and Zurek (1988), the best results for the intensity-discrimination model were obtained here for the multiband version, although the differences between the single- and multiband model were small. The multiband model could describe the data for the ΔP experiment reasonably well, which suggests that discrimination was based on profile analysis. In the case of the ΔF experiment, the descriptive power of this model depended on whether harmonics around f_r were resolved by peripheral filtering. For both conditions in which harmonics were resolved, i.e., $f_r=500$ Hz ($F_{wa}=100$ Hz) and $f_r=1000$ Hz ($F_{wa}=250$ Hz), reasonable r^2 values were obtained. For the latter condition, the regression slopes for the ΔP and ΔF data were almost identical. For $f_r=1000$ Hz and $F_{wa}=100$ Hz, where harmonics 9–11 around f_r are only just resolved, the r^2 values were reasonable but the regression slope was much smaller than for the ΔP data. For $f_r=2000$ Hz and $F_{wa}=100$ Hz, where harmonics 19–21 are unresolved, r^2 values were lowest.

Figure 8 illustrates excitation-pattern differences between a PSOLA-manipulated signal and the unmanipulated reference for $f_r=500$ (top panel) and 2000 Hz (bottom panel). For both formant frequencies, ΔF conditions for which performance was above or below $d'=1$ are indicated by the filled and open symbols, respectively. As the data in Fig. 8 suggest, excitation-pattern differences for above-threshold stimuli are larger if the harmonics around f_r are resolved (top) than for unresolved harmonics (bottom).

For the modulation-discrimination model, r^2 values were highest for $f_r=2000$ Hz and lowest for $f_r=500$ Hz, as

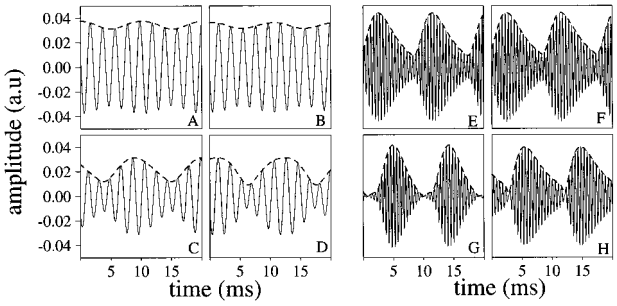

FIG. 9. Illustration of AM differences in the output of Gammatone filters. Panels A–D show portions of the output waveforms of a filter centered at 500 Hz ($f_r=500$ Hz): panel A shows the waveform for $\Delta F=-2.4\%$ for the PSOLA-manipulated signal, panel B for the unmanipulated reference. The modulation-depth measure D_{mod} is equal to 0.003. Panels C and D show the corresponding output waveforms for $\Delta F=-11.1\%$ ($D_{\text{mod}}=0.04$). Panels E–H show output waveforms of a filter centered at 2000 Hz ($f_r=2000$ Hz); $\Delta F=11.1\%$ in panels E and F ($D_{\text{mod}}=0.05$) and $\Delta F=8.1\%$ in panels G and H ($D_{\text{mod}}=0.7$).

would be expected on the basis of unresolved and resolved harmonics, respectively. Figure 9 illustrates this expectation by showing, for the same ΔF values as in Fig. 8, the output of Gammatone filters centered at the formant frequencies 500 and 2000 Hz. In the experiments the ΔF value of the top panels resulted in performance below $d'=1$. Accordingly, for both formants the difference in modulation depth is small. The ΔF value of the bottom panels resulted in above-threshold performance. Only for $f_r=2000$ Hz, however, a substantial difference in modulation depth can be observed.

As the r^2 values were moderate for both models for $f_r=1000$ Hz and $F_{wa}=100$ Hz, a multicue model might be reasonable. This was not verified, however, because we judged the amount of experimental data insufficient for performing multiple regressions reliably.

C. Synthesis window

Some of the effects of envelope modulation introduced by PSOLA (cf. Fig. 2) can be canceled by applying a so-called ''synthesis window.'' Such a window corrects for the fact that adjacent Hanning windows do not add up to one if $T_s \neq T_a$. A simple realization of such a window is to calculate the temporal envelope of the adjacent Hanning windows, spaced at intervals of T_s ms. By taking the reciprocal of this envelope and multiplying it with the PSOLA-manipulated (speech) signal, the degree of AM of the latter signal is reduced. Such an operation, however, does not correct for the AM introduced by out-of-phase addition of the fine structures of adjacent segments.

Experimental results for $f_r=1000$ and 2000 Hz, obtained by including the synthesis window as described above, are shown in Fig. 10 by the open symbols ($F_{wa}=100$ Hz with level roving, subject RK only). The filled symbols indicate corresponding data from experiment I. The synthesis window seems to cancel the ceiling effects for $f_r=1000$ Hz, although the fact that performance is still above $d'=2$ suggests that this effect is perceptually less relevant. Even in the case of $f_r=2000$ Hz, a condition for which temporal cues

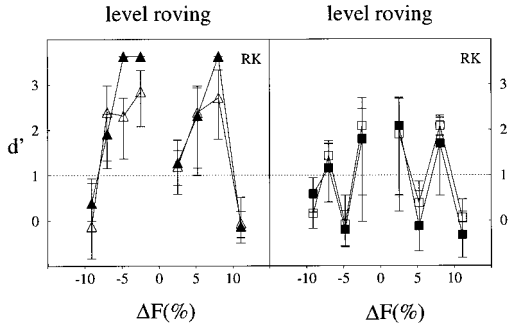

FIG. 10. Psychometric functions obtained by including a synthesis window (open symbols) and by the standard PSOLA operations (filled symbols, data already shown in Figs. 4 and 6). Data for f_r=1000 Hz are shown in the left-hand panel, data for f_r=2000 Hz in the right-hand panel. In both cases, level roving was applied.

presumably dominate detection performance, the two psychometric functions are basically identical.

D. Natural speech

In order to understand the perceptual effects of PSOLA manipulation of natural speech, the following aspects should, in our view, be additionally investigated. First of all, natural speech is generally characterized by the presence of at least 2 to 3 formants, at least in vowels. The use of multiple-formant signals may inform about the way in which cues occurring in several frequency regions are combined. Second, it should be investigated to what extent the detectability of distortions is influenced by fluctuations in both spectral content and $F0$. These fluctuations can be either of a random nature (e.g., jitter) or more deterministic (e.g., formant and $F0$ trajectories). In addition, the perceptual consequences of errors in $F0$ estimation in natural speech, leading to incorrect pitch-marker positioning, should be investigated. Third, the present experiments were performed under well-controlled acoustical conditions. The amplitude and phase transfer characteristics of (normal) playback rooms, however, will in their turn affect stimulus characteristics. It is conceivable that, under such listening conditions, the perceptual tolerance for the distortions introduced by PSOLA is actually increased.

IV. CONCLUSIONS

(1) Discrimination thresholds as a function of ΔF (the shift in fundamental frequency) for PSOLA-manipulated and -unmanipulated single-formant signals are found to be low: $|\Delta F| \leqslant 2\%$. Moreover, the psychometric functions reported here typically show an interaction between the formant and fundamental frequency.

(2) Roving of overall level does not seem to greatly affect discrimination performance as a function of ΔF. Roving of the formant frequency does impair performance: for formants at 500 and 1000 Hz (100-Hz fundamental), performance drops below $d'=1$ for all ΔF values tested. For a fundamental of 250 Hz, performance seems to be only moderately influenced by formant roving.

(3) Discrimination thresholds as a function of ΔP (the pitch-marker location) reported here are approximately a quarter of the fundamental period. The discrimination cues that occur due to incorrect positioning of the pitch markers seem to be robust under level and formant frequency roving. These findings apply to signals with a fundamental of 100 Hz; ΔP thresholds for higher $F0$ values are expected to be lower.

(4) The discrimination data as a function of ΔP can be described well with an intensity-discrimination model. This model can also account moderately well for the experimental data as a function of ΔF, in the case of resolved harmonics around the formant frequency (e.g., a formant at 1000 Hz with a 250-Hz fundamental). The modeled discrimination sensitivities, however, generally differ across the ΔP and ΔF conditions. For unresolved harmonics, such as for a 2000-Hz formant and a 100-Hz fundamental, the modulation-discrimination model matches the experimental data reasonably well. These findings are in agreement with the psychoacoustical notion of different modes of processing for resolved and unresolved harmonics.

(5) As for natural speech, distortions introduced to signals with higher fundamental frequencies are expected to be more easily detectable [see conclusions (2) and (3)]. In the case of low fundamental frequencies, the occurring phase cues are often subtle and may not be stable under different playback conditions.

ACKNOWLEDGMENTS

The authors thank Dik Hermes, Adrian Houtsma, Andrew Oxenham, Steven van de Par, Raymond Veldhuis, Rob Maher, and an anonymous reviewer for critically reading earlier versions of this article and for providing useful suggestions for improvement.

APPENDIX: PURE TONES

The main question of this baseline experiment was to what extent the basic distortions described in Sec. I C are detectable by the human auditory system. Pure tones with carrier frequencies f_c=500, 1000, and 2000 Hz were PSOLA manipulated. The carriers were thought of as harmonics of a fundamental of 100 or 250 Hz. In order to determine the detectability of the introduced side components, the references were unmanipulated pure tones having the same frequency as the strongest component in the manipulated-tone spectrum [cf. Fig. 2(c)]. Instead of taking all side component into account, the results presented below were obtained for signals consisting of the three strongest spectral components only. These results were compared with results for ''real'' PSOLA-manipulated tones and did not differ considerably. Three overall levels L were used: 45, 60, and 75 dB SPL in combination with a level rove uniformly distributed between -5 and $+5$ dB. Stimulus characteristics such as duration and stimulus generation were the same as described in Sec. II A 1.

A two-down, one-up 3IFC adaptive procedure was used in which ΔF was varied adaptively for the determination of discrimination thresholds. After a learning phase, three measurements for each condition were collected whose mean and

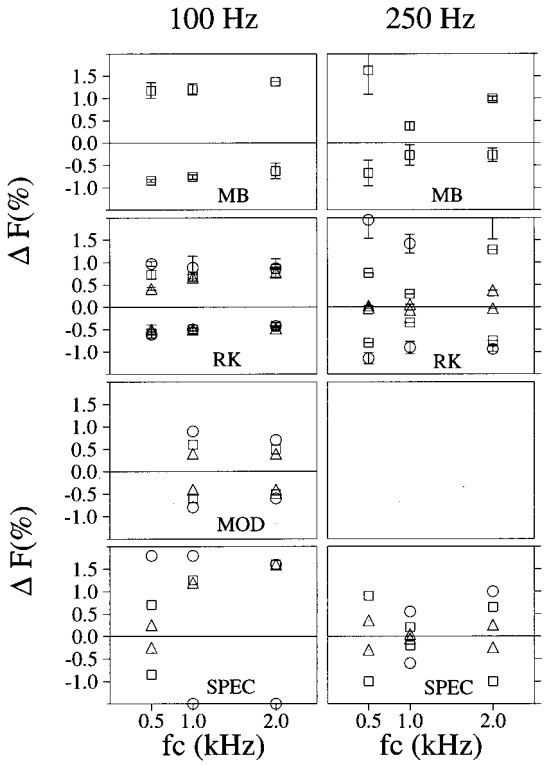

FIG. A1. Detection thresholds for subjects MB and RK for ΔF^+ and ΔF^- for the manipulation of pure tones. Left-hand panels give the data for $F_{wa}=100$ Hz, right-hand panels those for 250 Hz. Data for $L=45$ dB SPL are shown by circles, those for 60 dB SPL by squares, and those for 75 dB SPL by triangles. Standard deviations are indicated by vertical lines. Threshold predictions on the basis of a modulation-detection model are denoted by MOD. Predicted thresholds using a spectral masking model are marked SPEC.

standard deviation will be presented below. Two subjects (MB and RK) participated in this experiment.

Figure A1 presents both the ΔF thresholds for subjects MB and RK and predictions from two models (the models are different from the models of experiments I and II, see below). Squares indicate data for $L=60$ dB SPL (both subjects) and circles and triangles for $L=45$ and 75 dB SPL, respectively (subject RK only). Generally, thresholds for both subjects are low: $|\Delta F| \leqslant 2\%$. Thresholds for ΔF^- tend to be lower than for ΔF^+. Carrier frequency does not greatly influence thresholds for $F_{wa}=100$ Hz but does have an effect for $F_{wa}=250$ Hz. Likewise, level does affect thresholds for $F_{wa}=100$ Hz, except for $f_c=500$ Hz, but influences thresholds for $F_{wa}=250$ Hz.

In the spectral masking model (SPEC), the detection of the side components is determined by their masked threshold in the presence of the much stronger center component [cf. the ''No Summation Model'' of Hartmann and Hnath (1982)]. The masked thresholds were estimated using the data in Schöne (1979).[5] The modulation-detection model (MOD) is based on determining[6] the AM and FM indices of the PSOLA-manipulated pure tone as a function of ΔF. As in Zwicker (1952), threshold predictions were only derived for $F_{wa}=100$ Hz, with $f_c=1000$ or 2000 Hz (where the ''Phasengrenzfrequenz'' is approximately 80 and 140 Hz, respectively). Although the signals used actually had mixed modulation (Hartmann and Hnath, 1982; Moore and Sek, 1992), the lowest prediction based on detection of either AM or FM was taken.

In the case of resolved harmonics, i.e., for $f_c=1000$ and 2000 Hz with $F_{wa}=100$ Hz (harmonics 9–11 and 19–21, respectively), the predictions based on modulation detection are in reasonable agreement with the experimental data. The predictions based on the spectral masking model are less accurate, especially for ΔF^-, with the possible exception of $f_c=500$ Hz. For conditions with resolved harmonics (i.e., $f_c=500$, 1000, and 2000 Hz with $F_{wa}=250$ Hz, and $f_c=500$ Hz with $F_{wa}=100$ Hz), the spectral-masking model predictions are qualitatively similar to the experimental data, also showing level dependency. The actual predicted thresholds, however, do not exactly match the experimental data, especially for $L=45$ dB SPL.

[1]Although a frequency-domain (FD-PSOLA) variant has also been proposed, the time-domain version (TD-PSOLA) has been commonly preferred due to its computational efficiency.

[2]Strictly speaking, this only applies if the pure tone is a harmonic of F_{wa}.

[3]The implementation described in Klatt (1980) is based on a sample rate of 10 kHz. Cox *et al.* (1989) address the issue of using other sample rates, such as 32 kHz used here, which leads to differences in spectral envelope. They propose to introduce additional poles in the resonance-filter transfer function to compensate for the high-frequency attenuation. In the present study no compensation is taken into account because just a single formant is simulated (the high-frequency mismatch is more severe for multiple formants). Nevertheless, the spectral difference at 5 kHz between a signal generated at 10 kHz and at 32 kHz amounts to 10 dB. Around the formant frequency, differences are within 0.5 dB.

[4]Because the AM in the auditory filter output generally was not sinusoidal, the modulation index M was calculated:

$$M=\frac{\sqrt{2}\sigma_e}{m_e},$$

where σ_e and m_e are the standard deviation and average of the envelope of the output, respectively. The envelope is obtained by calculating a discrete Hilbert Transform. For an unmanipulated sinusoid of amplitude 1, $m_e=1$ but $\sigma_e=0$, so that $M=0$. For a 100% amplitude-modulated sinusoid of amplitude 1, $\sigma_e=1/\sqrt{2}$ and $m_e=1$, so that $M=1$, as expected.

[5]Predictions for spectral masking are based on Fig. 3 in Schöne (1979).

[6]Visual inspection showed that, for the range of ΔF used here, both the envelope and the instantaneous frequency were modulated in a sinusoidal fashion. The AM index m and FM index $\Delta f/f_{\text{mod}}$ were therefore calculated using their definition in Zwicker (1952).

Allen, J. B., and Rabiner, L. R. (**1977**). ''A unified approach to short-time Fourier analysis and synthesis,'' Proc. IEEE **65,** 1558–1564.

Cox, N. B. , Ito, M. R., and Morrison, M. D. (**1989**). ''Technical considerations in computation of spectral harmonics-to-noise ratios for sustained vowels,'' J. Speech Hear. Res. **32**, 203–218.

Durlach, N. I., Braida, L. D., and Ito, Y. (**1986**). ''Towards a model for discrimination of broadband sounds,'' J. Acoust. Soc. Am. **80,** 63–72.

Farrar, C. L., Reed, C. M., Ito, Y., Durlach, N. I., Delhorne, L. A., Zurek, P. M., and Braida, L. D. (**1987**). ''Spectral-shape discrimination. I. Results from normal hearing listeners for stationary broadband noises,'' J. Acoust. Soc. Am. **81,** 1085–1092.

Flanagan, J. L. and Saslow, M. G. (**1958**). ''Pitch-discrimination for synthetic vowels,'' J. Acoust. Soc. Am. **30,** 435–442.

Florentine, M., and Buus, S. (**1981**). ''An excitation-pattern model for intensity discrimination,'' J. Acoust. Soc. Am. **70,** 1646–1654.

Gagné, J., and Zurek, P. M. (**1988**). ''Resonance-frequency discrimination,'' J. Acoust. Soc. Am. **83,** 2293–2299.

Goldman, S. (**1948**). *Frequency Analysis, Modulation and Noise* (McGraw-Hill, New York).

Hartmann, W. M., and Hnath, G. M. (**1982**). ''Detection of mixed modulation,'' Acustica **50,** 297–312.

Henn, C. C., and Turner, C. W. (**1990**). ''Pure-tone increment detection in harmonic and inharmonic backgrounds,'' J. Acoust. Soc. Am. **88,** 126–131.
Kewley-Port, D., and Watson, C. S. (**1994**). ''Formant-frequency discrimination for isolated English vowels,'' J. Acoust. Soc. Am. **95,** 485–496.
Klatt, D. H. (**1973**). ''Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception,'' J. Acoust. Soc. Am. **53,** 8–16.
Klatt, D. H. (**1980**). ''Software for a cascade/parallel formant synthesizer,'' J. Acoust. Soc. Am. **67,** 971–995.
Klatt, D. H. (**1982**). ''Predictions of perceived phonetic distance from critical-band spectra: A first step,'' in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris (IEEE, New York), Vol. 2, pp. 1278–1281.
Lyzenga, J., and Horst, J. W. (**1995**). ''Frequency discrimination of bandlimited harmonic complexes related to vowel formants,'' J. Acoust. Soc. Am. **98,** 1943–1955.
Ma, C., Kamp, Y., and Willems, L. F. (**1994**). ''A Frobenius norm approach to glottal closure detection from the speech signal,''IEEE Trans. Speech Audio Process. **2**, 258–265.
MacMillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A Users Guide* (Cambridge U.P., New York).
Moore, B., and Glasberg, B. (**1987**). ''Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns,'' Hearing Res. **28**, 209–225.
Moore, B. C. J., and Sek, A. (**1992**). ''Detection of combined frequency and amplitude modulation,'' J. Acoust. Soc. Am. **92,** 3119–3131.
Moulines, E., and Charpentier, F. (**1990**). ''Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,'' Speech Commun. **9**, 453–467.
Moulines, E., and Laroche, J. (**1995**). ''Non-parametric techniques for pitch-scale and time-scale modification of speech,'' Speech Commun. **16**, 175–205.
Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1987**). ''An efficient auditory filterbank based on the gammatone function,'' in *Appendix B of SVOS Final Report: The Auditory Filterbank*, APU report 2341.
Pisoni, D. B. (**1980**). ''Variability of vowel formant frequencies and the quantal theory of speech: A first report,'' Phonetica **37**, 285–305.
Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).
Richards, V. M., Onsan, Z. A., and Green, D. M. (**1989**). ''Auditory profile analysis: Potential pitch cues,'' Hearing Res. **39**, 27–36.
Roads, C. (**1988**). ''Introduction to granular synthesis,'' Comput. Music. J. **12**, 11–13.
Roucos, S., and Wilgus, A. (**1985**). ''High quality time-scale modification for speech,'' in *IEEE ICASSP-85*, Vol. 2, pp. 493–496.
Schöne, P. (**1979**). ''Mithörschwellen–Tonheitsmuster maskierender Sinustöne,'' Acustica **43,** 197–204.
Smits, R., and Yegnanarayana, B. (**1995**). ''Determination of instants of significant excitation in speech using group delay function,'' IEEE Trans. Speech Audio Process. **3**, 325–333.
Sommers, M. S., and Kewley-Port, D. (**1996**). ''Modeling formant frequency discrimination of female vowels,'' J. Acoust. Soc. Am. **99,** 3770–3781.
Turner, C. W., and Van Tasell, D. J. (**1984**). ''Sensorineural hearing loss and the discrimination of vowel-like stimuli,'' J. Acoust. Soc. Am. **75,** 562–565.
Wakefield, G. H., and Viemeister, N. F. (**1990**). ''Discrimination of modulation depth of sinusoidal amplitude modulation (SAM) noise,'' J. Acoust. Soc. Am. **88,** 1367–1373.
Zera, J., Onsan, Z. A., Nguyen, Q. T., and Green, D. M. (**1993**). ''Auditory profile analysis of harmonic signals,'' J. Acoust. Soc. Am. **93,** 3431–3441.
Zwicker, E. (**1952**). ''Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones,'' Acustica **2**, Akustische Beihefte AB125–AB133.