A measure for predicting audibility discrimination thresholds for spectral envelope distortions in vowel sounds

Preeti Rao
Advanced Center for Research in Electronics, Indian Institute of Technology, Bombay, Powai, Mumbai 400076, India

R. van Dinther, R. Veldhuis, and A. Kohlrausch[a)]
IPO, Center for User–System Interaction, 5600 MB Eindhoven, The Netherlands

(Received 13 April 2000; accepted for publication 18 January 2001)

Both in speech synthesis and in sound coding it is often beneficial to have a measure that predicts whether, and to what extent, two sounds are different. This paper addresses the problem of estimating the perceptual effects of small modifications to the spectral envelope of a harmonic sound. A recently proposed auditory model is investigated that transforms the physical spectrum into a pattern of specific loudness as a function of critical band rate. A distance measure based on the concept of partial loudness is presented, which treats detectability in terms of a partial loudness threshold. This approach is adapted to the problem of estimating discrimination thresholds related to modifications of the spectral envelope of synthetic vowels. Data obtained from subjective listening tests using a representative set of stimuli in a 3IFC adaptive procedure show that the model makes reasonably good predictions of the discrimination threshold. Systematic deviations from the predicted thresholds may be related to individual differences in auditory filter selectivity. The partial loudness measure is compared with previously proposed distance measures such as the Euclidean distance between excitation patterns and between specific loudness applied to the same experimental data. An objective test measure shows that the partial loudness measure and the Euclidean distance of the excitation patterns are equally appropriate as distance measures for predicting audibility thresholds. The Euclidean distance between specific loudness is worse in performance compared with the other two. [DOI: 10.1121/1.1354986]

PACS numbers: 43.66.Ba, 43.66.Cb, 43.71.Es [RVS]

I. INTRODUCTION

Two important problems in sound compression and speech synthesis are the prediction of whether two sounds are perceived as different and how to express supra-threshold quality differences. An objective distance measure for predicting audibility thresholds and supra-threshold quality differences is important in both areas of research. Although it is valuable to have an objective distance measure which can assess the subjective quality of an entire sentence or phrase, and which correlates well with subjective test scores, it is also useful to evolve a measure which can predict the quality of short steady segments. Such a measure can serve as the basis for an overall quality measure, and can be used in an analysis-by-synthesis framework where the difference between the reference sound and the synthesized sound needs to be estimated.

Commonly used basic objective distance measures, such as signal-to-noise ratios and spectral distances, are derived directly from differences in the waveforms or in the power spectra of the reference and test signals (Quackenbush *et al.*, 1988). However, because it is the perception of the distortion that needs to be quantified, it is expected that measures derived from models of the auditory system will provide the most accurate predictions.

This paper addresses the problem of finding a perceptual distance measure that predicts audibility discrimination thresholds of modifications to the spectral envelope of steady vowel-like sounds. Such sounds are completely specified by their power spectra which can be represented as a set of harmonic components at multiples of a specified fundamental frequency. Modifications to the power spectrum occur in the form of magnitude changes of the harmonic components. We do not consider the effect of phase changes in vowel spectra because it is known that phase distortion has a relatively minor effect on the sound quality of complex tones (Plomp, 1976). In the case of sounds with harmonic spectra, the spectral magnitude changes can be viewed as distortions of the spectral envelope of the harmonic components. Sources of this type of distortion are, for example, filtering by a nonuniform gain transfer function and the inaccurate modeling of the spectral envelope, for instance, in linear predictive synthesis. It is of interest to predict whether the modifications give rise to discriminable changes in perceived quality, and if so, to quantify the extent of perceptual degradation.

In the next section we review the past development of auditory distance measures for the distortion of the spectral envelope of vowel sounds. We motivate and propose a new distance measure based on partial loudness for the prediction of the discrimination threshold. In Sec. III, a brief overview is given of the loudness model recently proposed by Moore *et al.* (1997), which forms the basis of the present work. The adaptation of the partial loudness measure to the prediction

[a)]Also affiliated with Philips Research Laboratories, Eindhoven.

of audibility discrimination thresholds for arbitrary modifications of the spectral envelope is discussed. In Sec. IV, measured discrimination thresholds using a 3IFC adaptive procedure with a representative set of stimuli are used to validate the applicability of the model in this context. The experimental results are discussed and possible explanations for deviations from the predicted thresholds in some cases, are provided. Other previously proposed and commonly used auditory distance measures are also evaluated on the same set of experimental data and their performance is compared with that of the partial loudness based distance measure.

II. VOWEL QUALITY DISTANCE MEASURES

Previous work on the problem of prediction of perceptual differences for vowel sounds has been based on modelling, to various degrees, the differences in the internal representations of reference and modified sounds. The auditory system includes the auditory periphery as well as central processing. It is assumed that perceptual discriminability (under optimal listening conditions and using well-trained subjects) depends largely on the resolution properties of the auditory periphery, and should be predictable by any good model of peripheral auditory processing (Gagné and Zurek, 1988).

Such an assumption led to the work of Plomp (1976) in which the spectral levels of the input stimulus power are summed over 1/3-octave bands, approximating the critical bands of the auditory system, to obtain a spectral representation more closely matched to that assumed in auditory processing. The quadratic distance between the reference and test signal representations was used to predict subjective quality differences in a set of steady sounds. In a further refinement, perceptual distance measures based on auditory excitation patterns have been applied to explain a variety of subjective discrimination data by postulating a threshold difference in excitation levels for detectability. Excitation patterns, or the excitation level per critical band, were first proposed by Zwicker and Scharf (1965) as part of the ''power spectrum model'' for auditory processing. These are calculated from the power spectrum as the output of the auditory filters with centers distributed uniformly on a critical band scale. Excitation patterns were used by Gagné and Zurek (1988), who investigated resonance-frequency discrimination of single formant vowels. The difference in the excitation patterns of the reference and modified signals was used to derive a distance measure given by either the single, largest magnitude difference (single-band model) or by the appropriately combined differences across bands (multiband model). A similar approach is followed in Kewley-Port (1991) who reported on detection thresholds for isolated vowels and examined several detection hypotheses of vowel spectra, based on their excitation patterns. Sommers and Kewley-Port (1996) studied the modelling of formant frequency discrimination of female vowels and evaluated an excitation-pattern model for this purpose.

While the excitation pattern represents the distribution of excitation along the basilar membrane, the loudness per critical band (specific loudness) corresponds more closely to the distribution of neural activity. The specific loudness is closely related to the subjective perception of loudness. A perceptual measure based on specific loudness is justified by the fact that the specific loudness versus critical band rate represents the best psychoacoustical equivalent of the power spectrum (Zwicker and Fastl, 1990). Distance measures based on applying various Minkowski metrics to the difference between specific loudnesses have been used to predict subjective distances in vowel quality (Bladon and Lindblom, 1981). This approach has also been followed more recently to explain the variation in formant-frequency discrimination thresholds observed in steady-state vowels (Kewley-Port and Zheng, 1998). The Euclidean distance between the reference and modified signals' specific loudnesses is used as the distance measure. However, the approach of applying a distance measure to specific loudness suffers from two serious shortcomings as far as the prediction of discrimination thresholds is concerned.

(1) It lacks a sound basis for the mathematical form of the distance measure, e.g., Euclidean, area, etc. Such a development is possible only in a purely experimental manner by observing the correlation between the distance measure and subjective data in specific situations.
(2) There is no basis for selecting the numerical value of the threshold level of the distance metric for the prediction of audibility.

Because we are interested in predicting the thresholds of discrimination for wide-ranging modifications to the spectral envelope, it is of importance to have a relatively invariant threshold level for the distance measure, preferably one based on a large and diverse body of psychoacoustical data.

Prediction of the discrimination threshold is a part of the larger problem of quantifying the perceptual effect of a distortion of the signal. That is, treating the difference between the original and modified signals as the signal to be detected, we wish to quantify its audible significance or its perceived loudness. The type of distortion under consideration in this paper involves a spectral gain modification. Since no new frequency components are created, it constitutes a linear distortion. For the purpose of computing auditory distance measures, these can be treated as additive distortion with a power spectrum equal to the difference in the power spectra of the reference and modified signals (Schroeder *et al.*, 1979). We wish, then, to estimate the audibility of this additive distortion which can be viewed as the ''signal'' to be detected in the presence of the background ''noise'' representing the reference signal. The background sound generally reduces the perceived loudness of the signal, an effect known as partial masking. The loudness of the signal in the presence of the background noise, or the partial loudness of the signal, is then a valid basis for an objective distance measure between the original and modified power spectra. To assess the partial loudness it requires the availability of a computational procedure such as the one given by Zwicker's loudness model (Zwicker and Scharf, 1965; Zwicker and Fastl, 1990). Recently a modified version of Zwicker's loudness model incorporating a more analytical formulation, was introduced by Moore *et al.* (1997). This revised model has been shown to account more accurately for various subjective loudness data. An enhancement to the earlier model particularly rel-

evant to our problem is the quantification of subthreshold levels of partial loudness and the consequent outcome of a threshold of audibility in terms of a partial loudness threshold. With such a threshold definition, this model has been used to predict thresholds related to the detection of tones in noise backgrounds as measured in various masking experiments (Moore *et al.*, 1997). In the next section we discuss the implementation of the partial loudness model and its adaptation to the problem of the prediction of discrimination thresholds for arbitrary envelope modifications of steady harmonic complexes.

III. THE PARTIAL LOUDNESS MODEL

The loudness model of Moore *et al.* (1997) is based on the approximate stages of auditory processing representing the conversion of the input power spectrum to the excitation pattern on the basilar membrane and the subsequent transformation to a specific loudness density. In the case of a signal presented with a background sound or masker (henceforth referred to simply as the ''noise''), a partial specific loudness distribution is derived for the signal based on the computed excitation pattern of the signal as well as that of the noise. The overall partial loudness of the signal, in sone, is then given by the total area under the partial specific loudness distribution. While the loudness model is based on analytical formulations, which represent approximately the stages of physiological processing, the exact nature of the formulations and their various parameters have been optimized to fit a large body of psychoacoustical data on masked thresholds and partial loudness judgements for a variety of multitone and noise stimuli. We next review the structure of the stages of the model in some detail.

A. Computing the excitation pattern

The excitation pattern of a sound is calculated as the output of the auditory filters representing the frequency selectivity of hearing at specific center frequencies. Figure 1 shows the stages involved in obtaining the excitation pattern from the input signal power spectrum which is specified by the frequencies and power spectral levels in dB SPL of its components. The first two blocks describe transfer functions from the free field to the eardrum and through the middle ear, respectively. For sounds presented over headphones, the fixed filter modeling the transfer function from the free field to the eardrum is replaced by one with a flat frequency response. In the third stage the excitation pattern of a given sound is calculated from the effective spectrum reaching the cochlea. According to Moore and Glasberg (1987), excitation patterns can be thought of as the distribution of ''excitation'' evoked by a particular sound in the inner ear along a frequency axis. In terms of a filter analogy, the excitation pattern represents the output level of successive auditory filters as a function of their center frequencies. The excitation pattern is generally presented as a function of the ERB rate rather than as a function of frequency. ERB refers to the equivalent rectangular bandwidth of the auditory filter and is a function of the filter center frequency. The ERB rate is a value on the ERB scale, which is closely related to the critical-band scale of the auditory system. On this scale the auditory filters are uniformly spaced with the ERB rate related to the frequency in kHz through a approximately logarithmic relation (Moore *et al.*, 1997).

Auditory filter shapes, experimentally derived from notched-noise experiments, are characterized as rounded exponential (RoEx) filters with parameters that control the filter selectivity (Moore and Glasberg, 1987). The frequency selectivity depends both on the center frequency of the auditory filter and the input stimulus level. With increasing input level the lower slope of the filter becomes shallower. The contribution of each stimulus component to the excitation pattern is calculated with a filter shape particular to that component. The lower slope of a filter is determined by the total stimulus level within the one-ERB band surrounding the stimulus component under consideration (van der Heijden and Kohlrausch, 1994). Thus to calculate the excitation level corresponding to the output of a given auditory filter, the input power spectral components are each weighted depending on their level and distance from the filter center frequency and combined additively as depicted in Fig. 1. This is repeated for all filter center frequencies spaced at intervals of 0.1 ERB in the range of 50 Hz to 15 kHz. We thus obtain the complete excitation pattern as a density, i.e., in dB SPL per ERB.

B. Calculating the partial loudness

The next stage of the model is the transformation from excitation pattern to specific loudness, which is the loudness density in sone per ERB. The specific loudness is obtained from the excitation distribution versus ERB rate by a compressive nonlinearity. The partial specific loudness of a signal in a background noise refers to its reduced perceived loudness and hence depends on the excitation distributions of the signal as well as that of the noise background. The formulas in Moore *et al.* (1997) provide this mapping based on psychoacoustical studies of loudness (Stevens, 1957; Zwicker and Scharf, 1965) as well as several subsequent experimental data on loudness perception and discriminability thresholds. Figure 10 in Moore *et al.* (1997) shows plots of the model output in terms of partial specific loudness (sone per ERB) versus signal excitation level for a range of noise excitation levels. The center frequency influences the computations by way of the level of the threshold in quiet which is assumed to vary with frequency in the model. From an examination of this figure, several features become evident. (1) The partial specific loudness is related to the signal excitation by a compressive nonlinearity that increases in strength with increasing noise excitation levels. This arises from the increased levels of masking at higher noise levels. (2) At levels of signal excitation well above the noise excitation, the partial specific loudness curves for the various noise levels converge and approach the specific loudness for the signal in quiet. (3) For a given noise excitation level, as the signal excitation approaches its masked threshold, the partial specific loudness rapidly attains low values and continues to decrease in value with decreasing signal excitation level.

For a signal presented in a background noise, the calculation of partial specific loudness requires the computation of

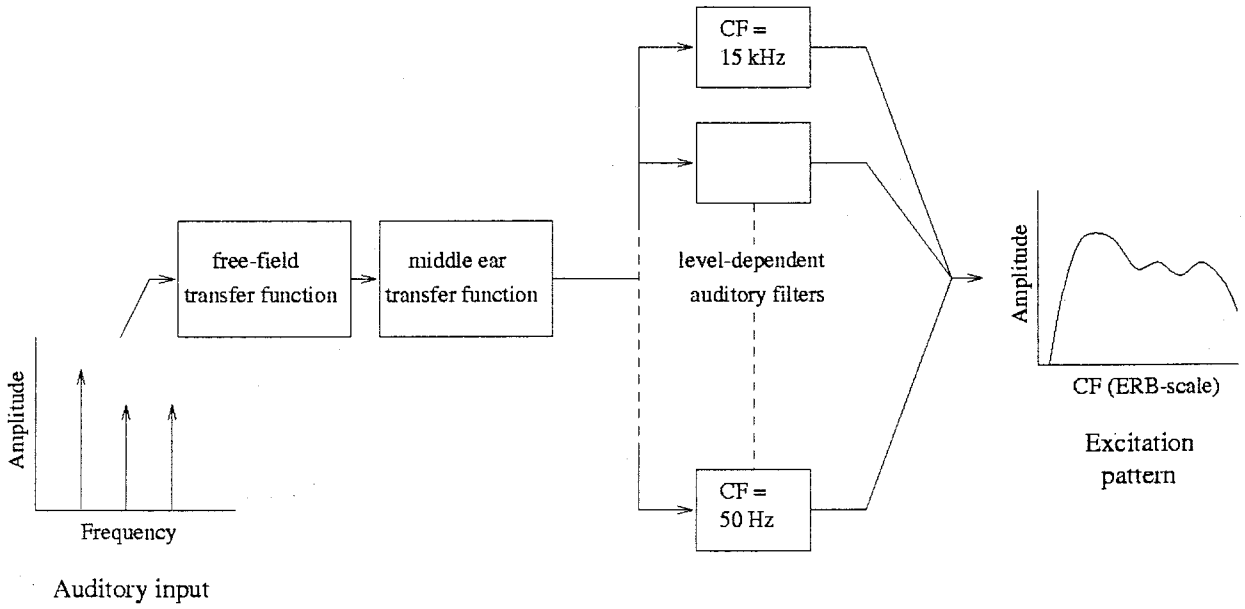

FIG. 1. Schematic diagram for calculating excitation patterns from the power spectrum of a sound represented by the frequencies and amplitudes of its harmonic components.

three excitation patterns. First, an excitation pattern is calculated for the total sound, that is, the signal plus the background noise. The auditory filter shape parameters obtained in the course of this computation are stored. Then, using these parameters, two further excitation patterns are calculated: one for the background noise and one for the signal. The partial specific loudness of the signal is next calculated using the formulas relating to the functions of Fig. 10 in Moore *et al.* (1997), at each ERB rate location as a function of the corresponding excitation levels of the signal and the noise as well as the threshold in quiet at that frequency location. The overall loudness of the given signal, in sone, is assumed to be the area under the specific loudness density. According to the model, the absolute or masked threshold of a sound corresponds to the level at which its partial loudness is 0.003 sone. Hence the model predicts, using the same transformation, both the subjective loudness and the discrimination threshold. The overall partial loudness as computed by the model is therefore a suitable candidate for quantifying the audible significance of the signal.

We see that to predict the discrimination threshold, the model integrates the specific loudness contributions across the entire ERB-rate range and as such can be considered a ''multiband'' model. The model has been used successfully to predict threshold data from a number of previous experiments on multicomponent complex tones in noise by assuming the threshold in overall partial loudness to be at levels between 0.003 sone and 0.008 sone (Moore *et al.*, 1997).

C. Partial loudness of arbitrary spectral-envelope distortions

The partial loudness measure can be applied to the problem of discriminating modifications of the spectral envelope of a steady sound in the following way. The reference sound is intended to take the role of the background noise and the modified sound that of signal plus background noise. We assume that the linear spectral distortion can be treated as additive noise with a power spectrum given by the difference between the reference and modified power spectra. The distortion is then the signal to be detected and its partial loudness can be calculated as described earlier. The amount of distortion at which the partial loudness attains the value of 0.003 sone is taken as the discrimination threshold. There is a problem, however, in that such a procedure would be suitable only when the modification can be considered as a positive additive distortion of the power spectrum. Because we wish to study arbitrary changes of the spectral envelope, we need to incorporate the treatment of cases in which the spectral level may actually decrease, at least for some spectral components. Figure 2(a) shows an example of such a case. The spectral envelope of the vowel ''a'' is subjected to an decrease in spectral tilt by means of highpass filtering. It can be seen that the low-frequency components are attenuated while the higher-frequency components are amplified. Here we must compute the partial loudness of two distinct types of distortion, one being a positive change in spectral level and the other a negative change.

Our approach to the problem of computing the partial loudness of an arbitrary distortion of the spectral envelope is illustrated by Fig. 3. We first compute separately the excitation patterns of the reference and modified signals. Then based on the channel-wise comparison of these two excitation patterns, we redefine the signal and noise excitation patterns to be used in the partial loudness model as follows. Let E_1 be the excitation pattern of the reference sound and E_2 that of the modified sound. The excitation pattern of the background noise is then defined as $\min(E_1,E_2)$, that of the total sound as $\max(E_1,E_2)$ and that of the signal as $|E_1 - E_2|$. Negative changes are treated in the same way as positive ones, therefore only the absolute value of the difference

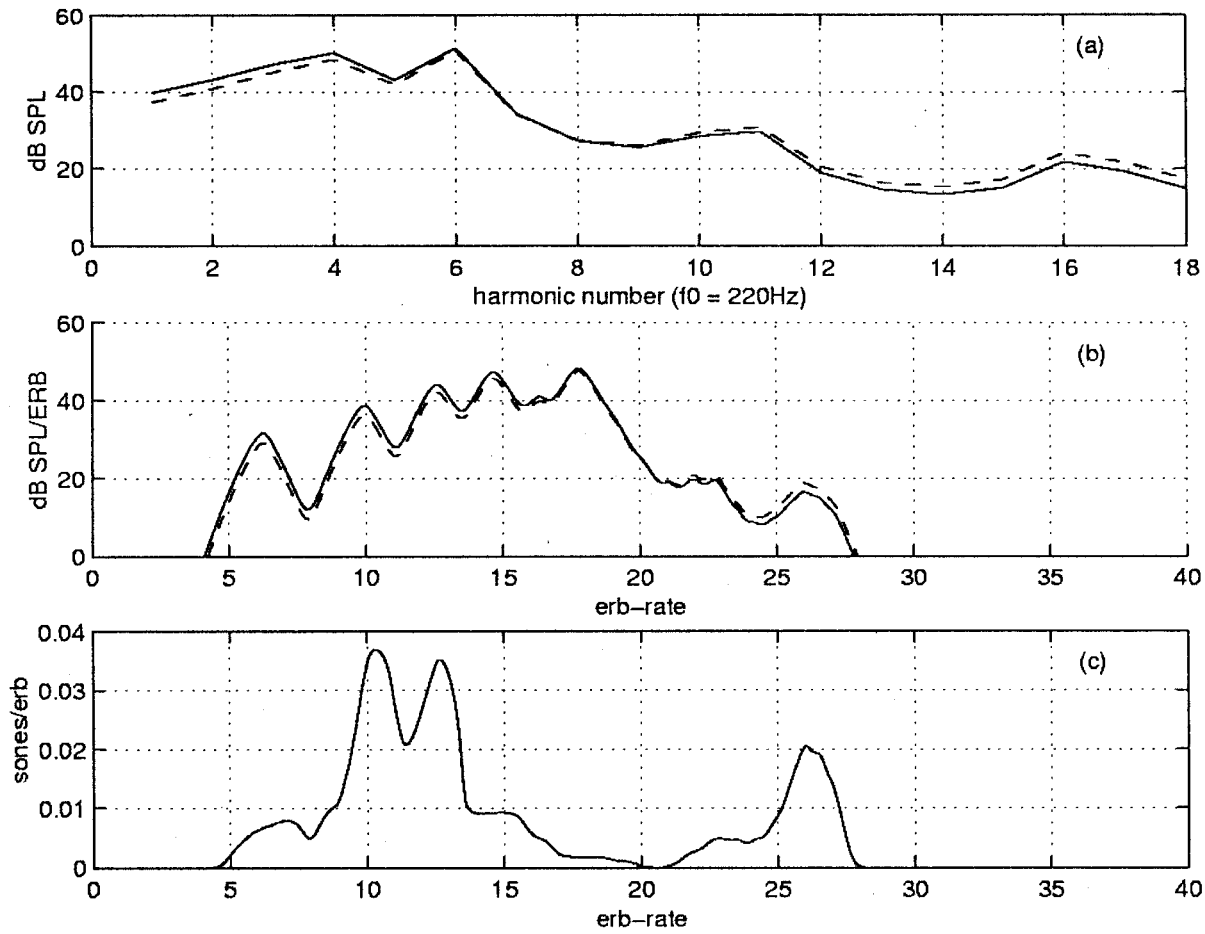

FIG. 2. An example of an arbitrary spectral-envelope modification. (a) The reference (solid line) and modified (dashed line) spectral envelopes of the simulated vowel /a/ with a fundamental frequency of 220 Hz. The modified sound is obtained by applying a single-pole, highpass filter to the reference sound. (b) The excitation patterns of the reference and modified signals. (c) The partial specific loudness distribution.

is of interest. The excitation patterns for the reference and modified signals of Fig. 2(a) are shown in Fig. 2(b). By applying these excitation patterns in the computation of the partial specific loudness of the distortion, we get the distribution shown in Fig. 2(c). The overall partial loudness is obtained by integrating the resulting (always greater-than-zero) values of partial specific loudness. It is the partial loudness measured in this way that we adopt as a measure for the perceptual distance between the sound with excitation pattern E_1 and the sound with excitation pattern E_2. Furthermore, we use a measure that is symmetric, i.e., when the reference sound and the modified sound are exchanged we obtain the same numerical value for the partial loudness of the difference. To what extent this distance measure is capable of predicting audibility thresholds in the context of spectral envelope distortions is investigated by means of the subjective experiment described in the next section.

IV. EXPERIMENT

A. Aim

The aim of the experiment is to validate whether partial loudness, computed according to the model presented earlier, can be used to predict audibility discrimination thresholds for arbitrary modifications of the spectral envelope of steady harmonic complexes. We also will compare our results with two alternative distance metrics, namely the Euclidean distance between excitation patterns, further denoted as the excitation pattern distance, and the Euclidean distance between specific loudnesses, further denoted as the specific loudness

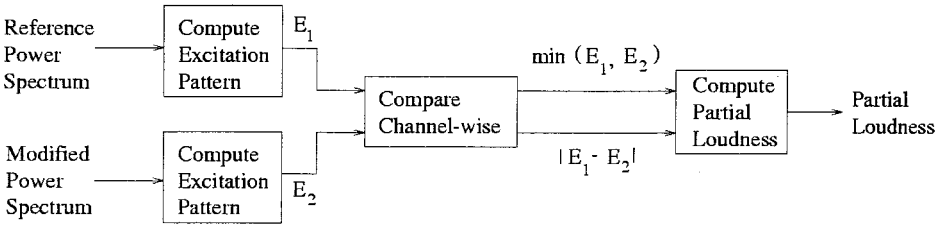

FIG. 3. Schematic diagram illustrating the computation of partial loudness from the excitation distributions of the reference and modified signals.

TABLE I. The 12 experimental conditions with a description of the corresponding spectral modifications.

	Conditions and corresponding modifications			
	Stimulus descriptions			
Condition	$F0$ (Hz)	Phase	Vowel	Modifications
1	220	random	a	harmonic 6, positive
2	220	random	a	harmonics 10–11, positive
3	220	random	a	harmonics 12–15, positive
4	220	random	a	harmonics 12–15, positive and harmonic 6, negative
5	220	random	a	modification of spectral tilt, low pass filter
6	220	random	a	modification of spectral tilt, high pass filter
7	220	random	i	harmonics 1–2, negative
8	220	random	i	harmonics 4–8, positive
9	220	random	i	harmonic 12 positive
10	220	random	i	harmonics 4–8, positive and harmonics 1–2, negative
11	110	random	a	modification of spectral tilt, low pass filter
12	110	regular	a	modification of spectral tilt, low pass filter

distance, applied to the same experimental data. A requirement for a distance measure that is useful for a broad class of speech and musical sounds, is that it should be capable of predicting audibility thresholds for a large variety of spectral envelope modifications. Therefore we chose a representative set of modification conditions for the experiment, distributed over the spectra of two simulated steady vowels, /a/ and /i/. A good measure is expected to produce the same threshold values for distinct conditions, at least for each individual subject. A relative variation, quantifying the range of spread across conditions and defined as the standard deviation of the measured thresholds for the various conditions divided by their mean will, therefore, be used as an indication of the quality of the measures.

B. Stimuli

The reference sound spectra were derived from the amplitude spectra of the vowels synthesised by the cascade combination of an LF model glottal source and a formant filter based on the linear prediction coefficients (LPC) (Fant *et al.*, 1985). A constant overall level of about 55 dB SPL was maintained. The set of modifications was chosen in a way to encompass distinct types of gain changes of the spectral envelope. Table I gives an overview of all the modifications. Specifically, we considered localized spectral amplitude changes at the formant peaks and in the valleys, and also combinations of these changes, both in opposite and in equal directions. The amplitudes of the harmonics were modified by multiplication with a factor close to 1. When more than one harmonic was modified, each harmonic was multiplied by the same factor. For example, the condition 2 of Table I corresponds to a scaling of the harmonics 10 and 11 of the harmonic spectrum by a factor greater than 1. We also investigated modifications that are relatively broadband, or have more spectral spread, by varying the overall spectral tilt. This was achieved by either lowpass filtering to increase the spectral tilt, or by highpass filtering to reduce it. The filter parameters were adjusted so that the overall loudness of the sound was not changed significantly. Figure 4 depicts the set of stimuli and modifications by indicating which harmonic components are affected in each of the conditions.

It is generally accepted that amplitude changes in the spectrum of harmonic sounds are more detectable than phase changes. It was found that for complex tones with a fundamental frequency beyond 150 Hz the maximal effect of phase on timbre is smaller than the effect of changing the slope of the amplitude pattern by 2 dB/oct (Plomp and Steeneken, 1969). Therefore a fundamental frequency of 220 Hz for the vowel-like spectra was used. We applied random phase for the stimuli to maintain an equal distribution of energy within each pitch period. To investigate the influence of changing the fundamental frequency, one of the spectral envelope modifications was repeated at a fundamental frequency of 110 Hz. At this lower fundamental, however, there could exist phase effects, which could lead to temporal cues. Therefore adding a condition with a phase derived from the glottal-pulse model tested the influences of these effects.

In each of the conditions, the spectral amplitudes were modified in small steps corresponding to the calculated partial loudness of the distortion as given by the model. The reference and modified sounds were generated as the sum of harmonics with the specified amplitudes and random phases. Only in the final condition (number 12), were the actual phases provided by the vowel synthesizer applied. The duration of the stimuli was 300 ms with raised cosine ramps of 25 ms at the beginning and end of the signal. For each condition we measured the value of the partial loudness at which the subject was just able to discriminate between the reference sound and the modified sound.

C. Method

Four subjects (JB, JG, PR, and RD) participated in the experiments. The subjects' ages and sexes are presented in Table II. All were young adults with normal hearing and no reported history of hearing impairment. In addition, measurements in Sec. V C showed normal absolute thresholds for all subjects at a frequency of 1 kHz. The stimuli were presented binaurally over headphones at a level of approximately 55 dB SPL to subjects seated in a sound-proof booth. A 3-interval forced-choice adaptive procedure (Levitt, 1971) was used to obtain the thresholds. In this procedure each trial consisted of three stimuli, two stimuli representing the refer-

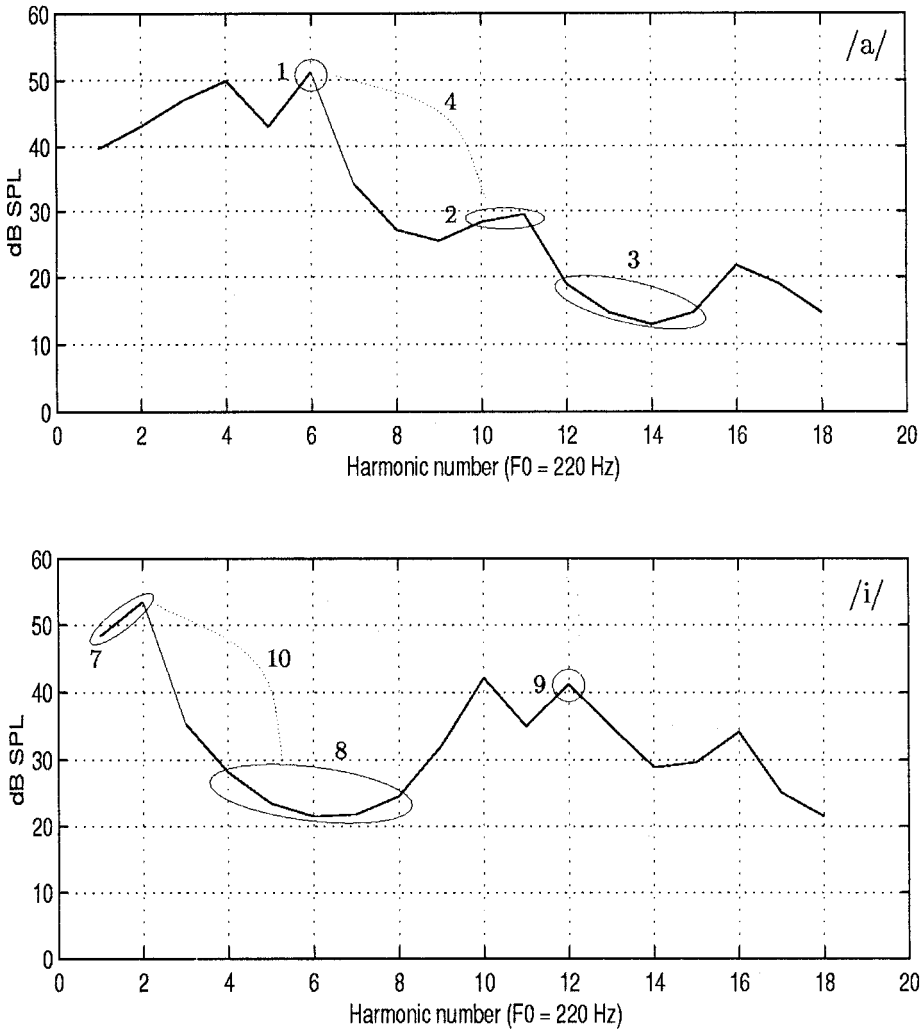

FIG. 4. The two panels show the reference spectra of the stimuli used in the subjective experiment. Top panel: Vowel /a/. Bottom panel: Vowel /i/, both with a fundamental frequency of 220 Hz. The encircled points indicate the modifications of harmonic amplitudes and the corresponding condition numbers. Not in the diagrams are the following four modifications of Table I: condition numbers 5, 6, 11, and 12.

ence sound and one the modified sound. The pause before one trial was 300 ms and the interstimulus interval was 400 ms. The assignment of the odd stimulus to one of the three intervals was randomized. The subject's task was to indicate the odd interval. Immediately after each response, feedback was given indicating whether the response was correct or incorrect. After two correct responses the amount of spectral modification was reduced by one step. After one incorrect response it was increased by one step. The spectral modifications of a stimulus were divided into 20 steps reaching from about 1 sone to 0.001 sone of partial loudness for the modification. A run began with a modification of the spectrum that produced an easily discriminable change. A test run was completed after 12 up–down reversals. A single-run estimate of the partial loudness at threshold was obtained by taking the median of the steps at the last eight reversals of the run. In this way the 70.7% correct detection threshold is measured. For each experimental condition, a final estimate of the partial loudness at threshold for each subject was based on the median of five single-run estimates taken over a period of several days.

TABLE II. Characteristics of the subjects.

	Characteristics of subjects			
	JG	JB	PR	RD
Sex	male	male	female	male
Age	24	30	38	26

V. RESULTS AND DISCUSSION

A. Partial loudness

In Fig. 5 the medians of the partial loudness levels at threshold are given for each subject, where the data of JB, JG, PR, and RD are indicated with a circle, triangle, cross, and star, respectively. The interquartile ranges are indicated with bars.

An examination of the data shows that for most of the conditions, the subjects' thresholds are between 0.003 sone and 0.02 sone, which is close to the range of 0.003 to 0.008 sone used by Moore *et al.* (1997) to predict detection thresholds for simple psychoacoustic stimuli. We particularly note that the thresholds for the widely differing spectral modifications namely localized perturbations (conditions 1, 2, 7, 9) and spread perturbations (conditions 4, 5, 6, 10, 11, 12) fall within the same narrow range. Spectral modifications described by a combination of positive and negative changes appear to be adequately treated by the proposed procedure. Changing the fundamental frequency while maintaining the

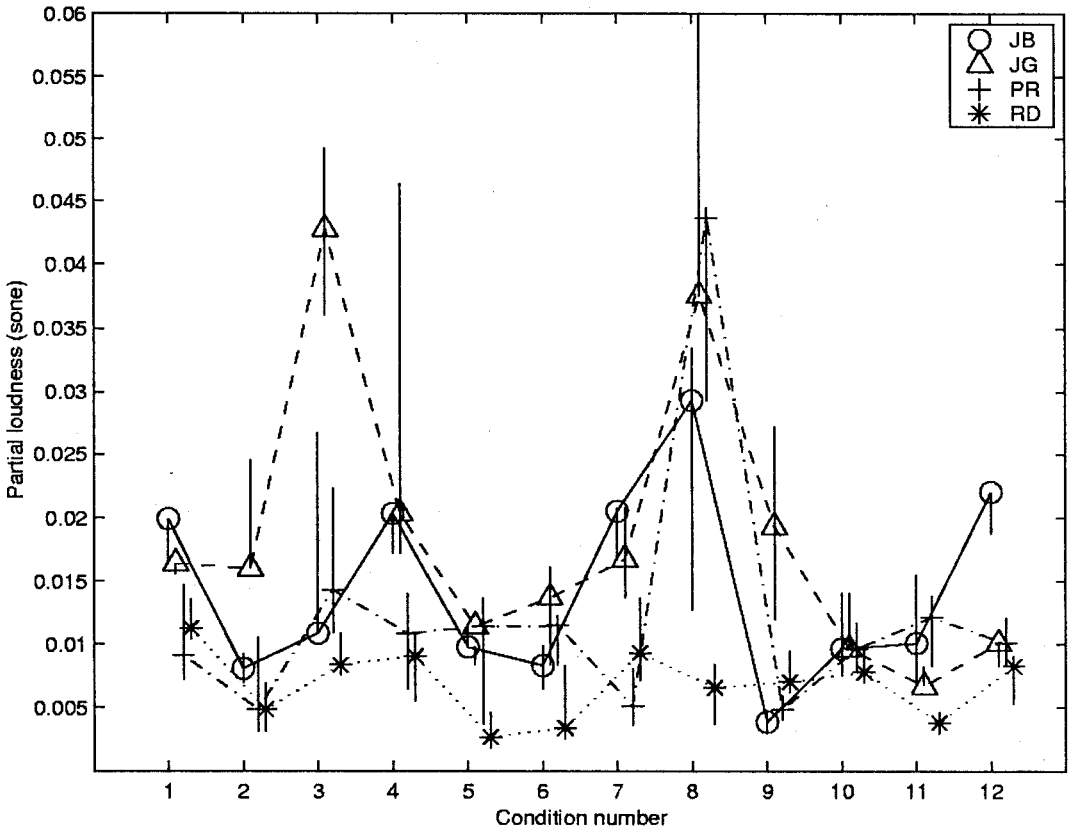

FIG. 5. Experimentally obtained thresholds plotted in terms of partial loudness. The bars indicate the interquartile ranges.

same spectral envelope (conditions 5 and 11) does not impact the accuracy of the predictions. The conditions 3 and 8 corresponding to modifications localized at the valleys of the spectral envelope are clear exceptions, however. For condition 8, three subjects show thresholds that are distinctly higher than thresholds measured for the other conditions while for condition 3, one subject shows high thresholds.

The partial loudness model of Moore *et al.* (1997) is based on the average of results of a large number of experiments involving listeners with normal hearing. The parameters of individual subjects, however, may vary from these average values. The greatest variability is expected in the selectivity of the auditory filters. Hence the predictions of the model cannot be expected to be accurate for all individual listeners. Later in this section we will attempt to correlate the large differences in the threshold levels with possible individual differences in auditory frequency selectivity. First we examine the performance of the Euclidean distance-based metrics on the same data.

B. Comparison with the excitation pattern and specific loudness distances

As discussed in Sec. II, the Euclidean distances between excitation patterns and between specific loudness have both been widely applied in the prediction of vowel quality differences. In contrast to the partial loudness measure, these measures are based on a direct comparison of the internal representations of the reference and test sounds.

The auditory model of Moore *et al.* (1997) considered in this paper was used in the computation of the excitation patterns and the specific loudness for the Euclidean metrics. For each subject and condition, the excitation patterns and the specific loudness of the reference sound and the modified sound corresponding to the just discriminable condition were computed. Figures 6 and 7 show the discrimination thresholds versus condition numbers for the Euclidean distances in the excitation patterns and the specific loudness, respectively.

Next, we compared the measures' performances. A requirement for a measure is that the distance values obtained at threshold for a large variety of spectral modifications are approximately constant for each individual subject. To compare the measures we therefore used the standard deviation of the measured thresholds for the various conditions divided by their mean, which is referred to as the relative variation. Figures 6 and 7 revealed that both the excitation pattern distance and the specific loudness distance display a range of overall variability of discrimination thresholds that is smaller than that of the partial loudness measure. The excitation pattern distance has a slightly smaller range than the specific loudness distance and shows a smaller variability across conditions. This would indicate that the partial loudness measure performs worse than the other two measures. Curves shown in Fig. 8 and Fig. 10 of Moore *et al.* (1997) indicate that both specific and partial loudness show an increased sensitivity with respect to excitation level when approaching the threshold of detectability. In a fair comparison of the measures'

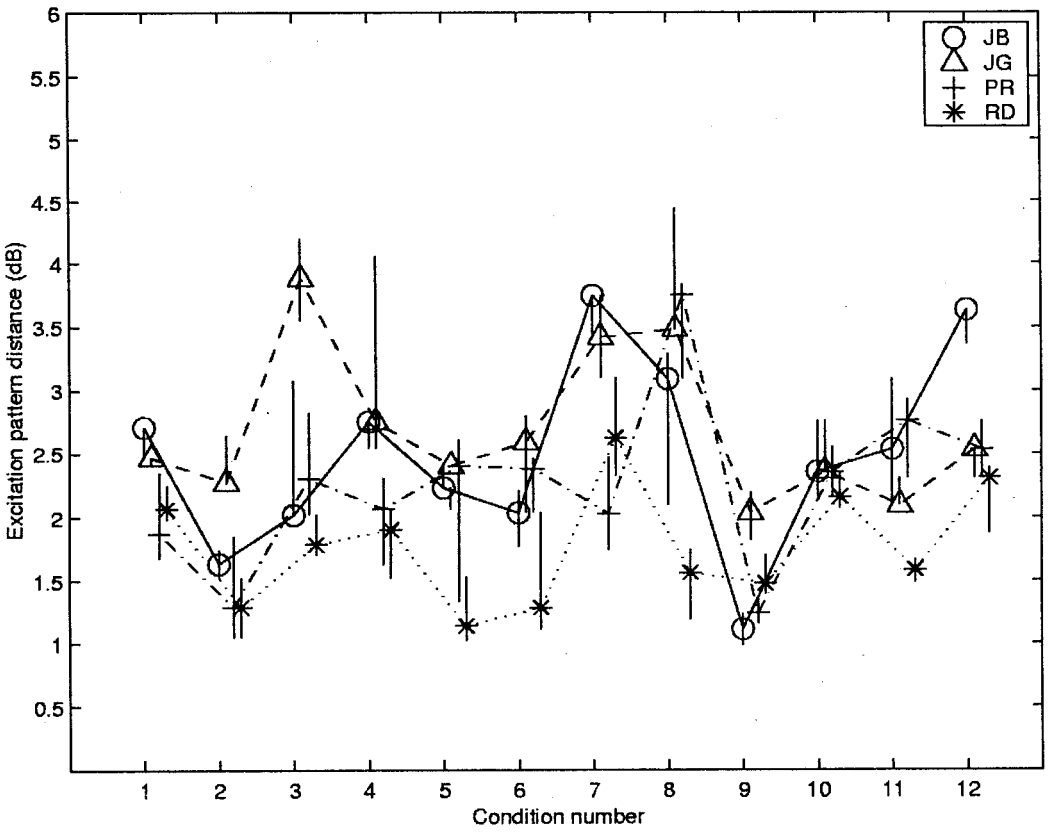

FIG. 6. Experimentally obtained thresholds plotted in terms of excitation pattern distance. The bars indicate the interquartile ranges.

performances, this difference in sensitivity must be compensated for. To objectively compare the relative variations of the three distinct distance measures, we carried out a normalization which takes into account the different sensitivities of the three measures near threshold. The sensitivities of the partial loudness and the specific loudness at threshold were normalized to match the sensitivity of the excitation patterns as follows. Plots of the log of partial loudness distance and the log of specific loudness distance as functions of the log of excitation pattern distance derived from the stimuli that were used, showed bundles of nearly parallel lines. This implies that there is a nearly constant proportional relation between relative variations in the excitation pattern distance and the other two measures. We could, therefore, use the means of slopes of theses curves at the various threshold points as estimates for two normalization factors ρ_{sl} and ρ_{pl}, by which the relative variations in the specific loudness distance and the partial loudness measure, respectively, were divided. The means and standard deviations of the normalization factors are plotted in Table III. Table IV presents the relative variations for the four subjects and the three distance measures before and after normalization. In the unnormalized case, the excitation pattern distance always has the lowest relative variation and it depends on the subject whether the partial loudness measure or the specific loudness distance performs second best. If we regard the mean over the subjects, presented in the last row of the table, the partial loudness measure comes out last and the excitation pattern distance first. In the normalized case, both the partial loudness measure and the excitation pattern distance come out best for two subjects. If we regard the mean over the subjects the partial loudness measure and the excitation pattern distance share the first position.

An error analysis or a presentation of confidence intervals for the results presented in Table IV would be in its place. However, such an analysis turns out to be analytically difficult. Therefore, we performed an error analysis by simulation. First of all, we assumed that the obtained data points had additive Gaussian errors with zero mean. The averages of the interquartile ranges across conditions were used to estimate the standard deviations of these errors for each subject and each distance measure. Then for each subject and each distance measure, the normalized relative variations were computed in 1000 simulation runs in which independent Gaussian errors were added to the computed threshold values. Figure 8 shows the resulting distributions of the relative variations, under the assumption that these distributions are also Gaussian. The distributions of the relative variations can be used to compute for each subject the probabilities that one distance measure performs better than another. These probabilities are presented in Table V. The notations P{PL>EPD}, P{SLD>EPD}, and P{PL>SLD} denote the probabilities that the partial loudness measure performs better than the excitation pattern distance, the specific loudness distance performs better than the excitation pattern distance and the partial loudness measure performs better than the specific loudness distance, respectively. The same probabilities have also been derived directly from the simulation data,

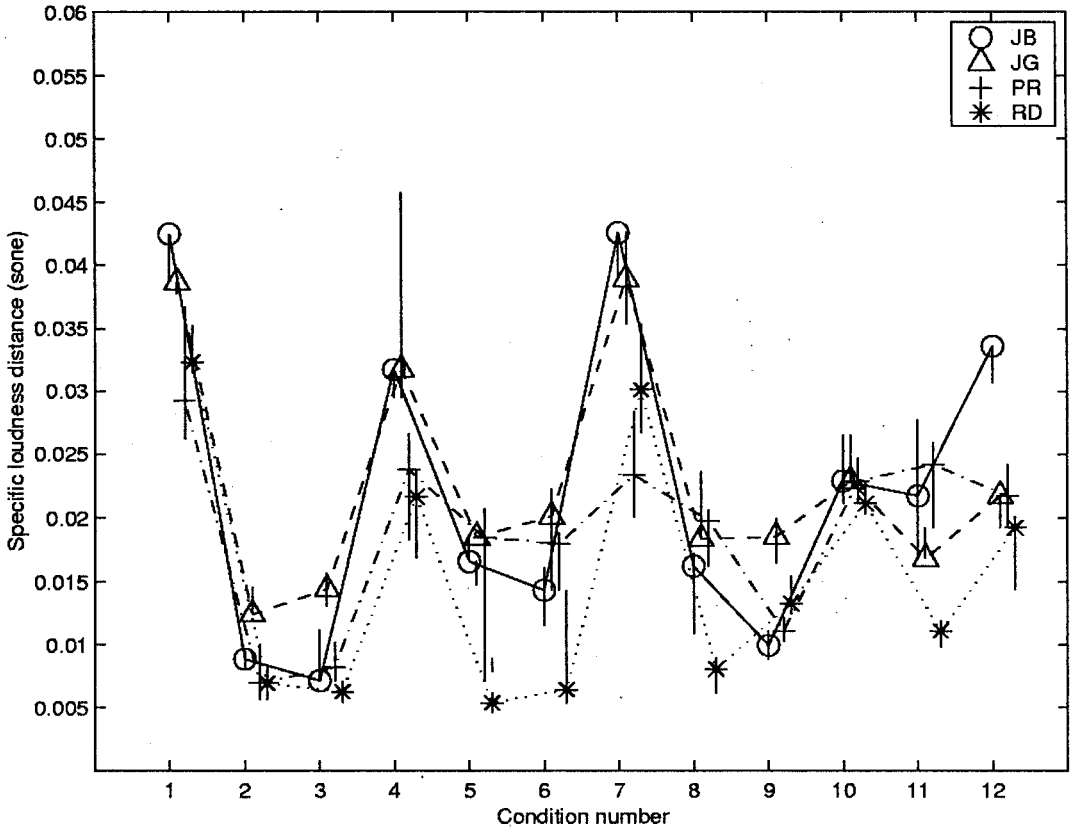

FIG. 7. Experimentally obtained thresholds plotted in terms of specific loudness distance. The bars indicate the interquartile ranges.

without the assumption of Gaussian distributions for the relative variations, but the results only differed in the first decimal of the percentages. These simulations confirm the results based on the data, namely that partial loudness measure and the excitation pattern distance perform equally well and clearly outperform the specific loudness distance.

C. Investigating the variations in partial loudness thresholds

We assumed that the explanation for the particularly high spread in partial loudness threshold values across subjects for the conditions 3 and 8 might be found in individual differences in auditory frequency selectivity. To support such an assumption, we investigated the effect of varying the auditory model filter parameters on the partial loudness levels at threshold, as well as looked for a basis on which any specific alteration of the model's auditory filter parameters may be justified. With this in mind, we picked condition 8 for further investigation.

A high value of calculated partial loudness at threshold implies that the modification is more difficult to detect than predicted by the model. A salient characteristic of condition 8 is that it involves the detection of a signal at a center frequency that is higher than that of the dominant masker. In such a situation it is natural to attribute the difference in detectability to a difference in the upward spread of masking. To follow this possible explanation an additional experiment was carried out to measure the upward spread of masking. We used a masker frequency of 440 Hz and a target frequency of 1000 Hz to create a similar situation as in condition 8. The masker levels were 70, 60 and 55 dB SPL. Table VI shows the medians of the masked thresholds of four sessions. The data in Table VI show a great variability in upward spread of masking and are in line with the assumption that the subjects' differences for condition 8 are due to differences in upward spread of masking. Subject PR shows

TABLE III. Means of normalization factors for the relative variations and their standard deviation.

	Mean	Standard deviation
ρ_{sl}	1.12	0.23
ρ_{pl}	2.19	0.44

TABLE IV. Relative variations of the unnormalized and normalized measures, excitation pattern distance (EPD), specific loudness distance (SLD), and partial loudness measure (PL).

	Relative variations					
	Unnormalized			Normalized		
	PL	SLD	EPD	PL	SLD	EPD
JB	0.53	0.56	0.31	0.24	0.50	0.31
JG	0.60	0.39	0.22	0.27	0.35	0.22
PR	0.84	0.36	0.29	0.38	0.32	0.29
RD	0.39	0.63	0.26	0.18	0.56	0.26
Mean	0.59	0.49	0.27	0.27	0.43	0.27

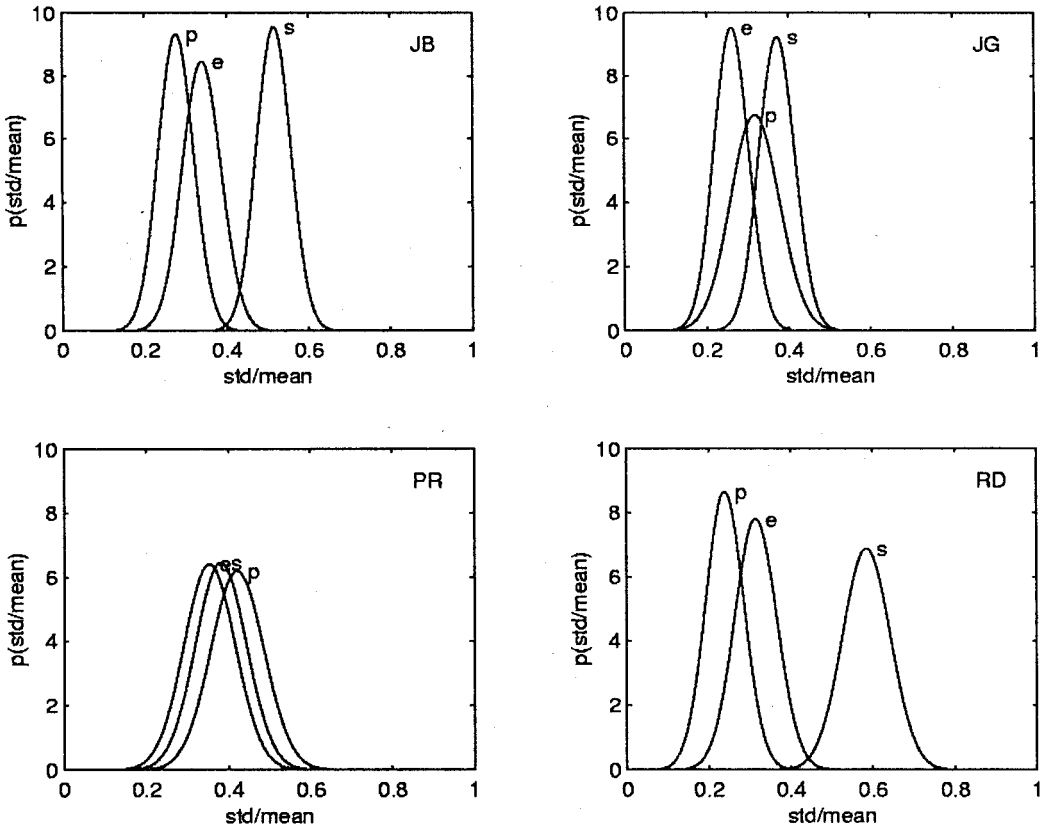

FIG. 8. Distributions of the relative variation for each subject, obtained by simulation.

high masked thresholds accompanied by a high score at condition 8, whereas subject RD shows low masked thresholds accompanied with a low score at condition 8. The subjects JB and JG have masked thresholds between those of PR and RD.

Individual differences in the upward spread of masking can be incorporated in the loudness model by modifying the auditory filter parameters. Decreasing the lower slope of the RoEx(p) filter by decreasing p is the most effective way to increase the predicted upward spread of masking. The filter slope influences the bandwidth (ERB) of the filter however. An alternative way to model the increased spread of masking is by introducing a small, non-zero value of ''r'' in the Roex(p,r) approximation of filter shape (Moore and Glasberg, 1987). The effect of this is to add a low-level skirt to the filter gain function while leaving its passband (upto 30 dB below the filter tip) essentially unchanged. The parameter ''r'' is thought to be related to absolute threshold effects which may vary among individuals (Moore, 1987). Both these approaches were considered separately by computing the parameter, ''*p*'' or ''*r*,'' required to fit the masked threshold data of Table VI for the subject PR, and then applying these modified parameters to calculate the partial loudness value at threshold for each condition. The results are shown in Fig. 9. We see that the value of partial loudness at threshold decreases for the conditions 3 and 8. Although the threshold levels for the other conditions too are affected to some extent, the threshold levels for modifications at the spectral valleys (including condition 2 in which the masker is primarily below the signal frequency) are clearly more sensitive to filter parameter changes. The conditions 1 and 7 can be characterized as being complementary to the spectral valley conditions and show the expected increase in the predicted threshold level with the increased upward spread of masking. So we see that while the modified parameters explain the high threshold of condition 8, they adversely impact the predictions for conditions 1, 2, and 3. However, it must be kept in mind that the modified parameter settings

TABLE V. Probabilities (in percentages) of relative performances of the measures per subject obtained by simulation.

	P{PL>EPD}	P{SLD>EPD}	P{PL>SLD}
JB	85	0	100
JG	21	3	78
PR	24	49	34
RD	86	0	100

TABLE VI. Median masked thresholds in dB SPL for three different masker levels in the upward-spread-of-masking experiment.

	Masked threshold			
	Subject			
Masker level	JG	JB	PR	RD
70	24.50	20.75	32.00	12.25
60	9.25	6.25	21.50	4.25
55	5.75	2.50	17.75	2.00

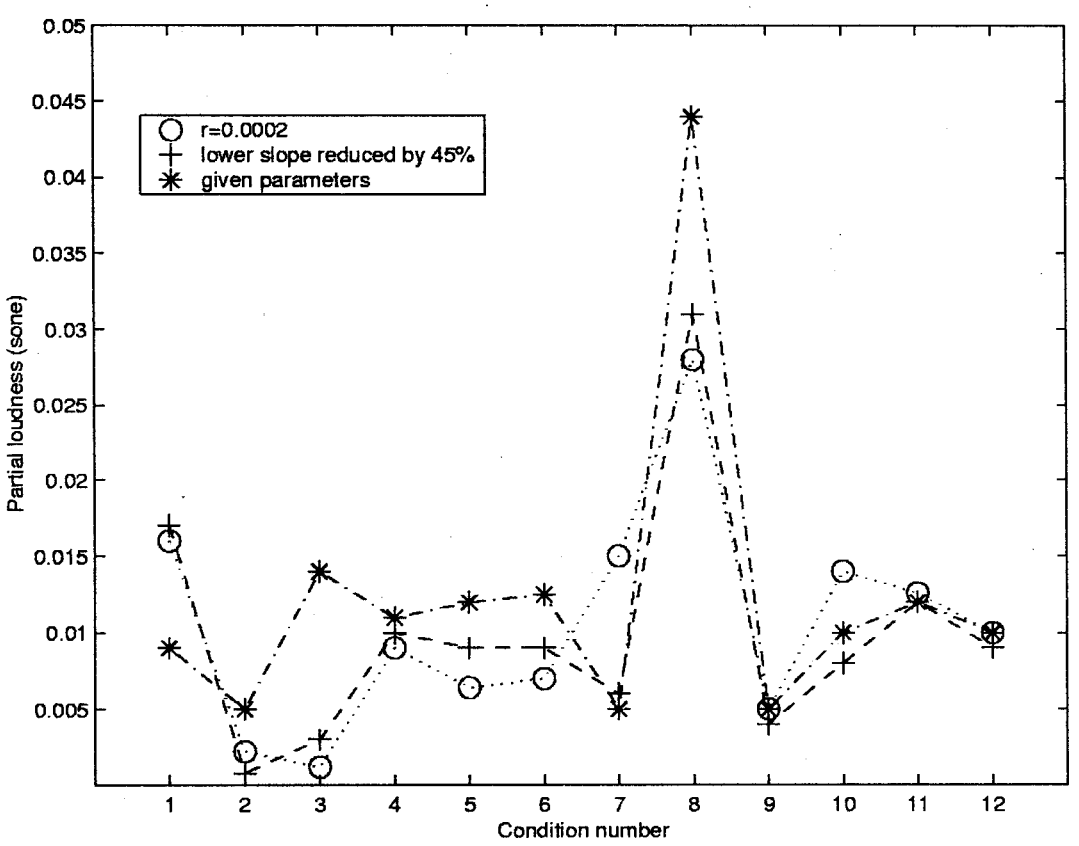

FIG. 9. Variation in partial loudness threshold levels with auditory filter selectivity computed for the experimental data of subject PR. The auditory frequency selectivity, modelled by the RoEx(p,r) function, was manipulated by changing the parameters p and r independently.

were derived from a masking situation applicable to the conditions 7, 8, and 10 and therefore may not be completely relevant for the other conditions.

The sensitivities of the Euclidean distance-based metrics to auditory filter parameter changes were also examined. It was found that the excitation pattern distance is also sensitive to the parameter changes but to a significantly lesser extent than the partial loudness measure. The specific loudness distance on the other hand is relatively insensitive to changes in filter parameter settings. These facts can also be seen in Figs. 6 and 7 where we observe less variation among subjects at any given condition in the specific loudness metric as compared to the excitation pattern metric. On the other hand, the specific loudness distance at threshold appears to be more dependent on the actual nature of the spectral modification.

The relative variations of the measures with the changed parameters were computed for subject PR. In Table VII the relative variations and performances of the measures are presented for the case of the lower slope reduced by 45%, the case of $r=0.0002$ and the original case. For the case of the lower slope reduced by 45% the normalization factors were recomputed as $\rho_{sl}=1.10$ and $\rho_{pl}=2.19$. For the case $r=0.0002$ the factors were recomputed as $\rho_{sl}=1.01$ and $\rho_{pl}=2.35$. The results in Table VII show that the performance of the partial loudness measure has improved slightly after adjusting the filter parameters. It can also be noted that the performance of the partial loudness measure has become better than the other two measures and that the results of subject PR have become more in line with the results of the other subjects.

VI. CONCLUSIONS

The partial loudness measure computed from the auditory model of Moore *et al.* (1997) was proposed and adapted for the problem of predicting perceptual differences caused by spectral envelope modifications of steady sounds. The partial loudness measure is based on a spectral model and does not take into consideration phase effects. The effectiveness of this measure for the prediction of discrimination thresholds of spectral envelope modifications in simulated vowel sounds was studied by means of subjective experiments. Our results indicate that the assumptions of the model

TABLE VII. Relative variations of the normalized measures and probabilities of the relative performances of the measures obtained by simulation for subject PR with adapted parameters of the auditory filter. Case 1 represents the reduced lower slope by 45%. Case 2 represents the parameter change $r=0.0002$.

	Relative variations			Relative performances of measures		
Case	PL	SLD	EPD	P{PL>EPD}	P{SLD>EPD}	P{PL>SLD}
1	0.35	0.44	0.36	69	47	87
2	0.31	0.44	0.39	82	50	85
Original	0.38	0.32	0.29	24	49	34

are justified and that the experimentally determined thresholds are reasonably close to the predicted values. Our results provide a range for the discrimination thresholds applicable to realistic data such as steady vowels. A score of 0.01 sone is a good estimate, although there appear easily variations of a factor of 0.5 to 2. Previously proposed vowel quality distance measures were also evaluated on the same experimental data. A relative variation, quantifying the range of spread across conditions, was defined in order to compare the measures. At a first glance, the Euclidean distance between excitation patterns gives a narrower range of spread in discrimination thresholds compared to the partial loudness measure. However, it was argued that the greater variability of the partial loudness measure and of the specific loudness distance is due to the sensitivity of these measures at threshold. Once normalized to match the sensitivity of the excitation pattern distance at threshold, the variability of both the partial loudness measure and the specific loudness distance is reduced and the performances of partial loudness measure and excitation pattern distance are similar and clearly better than that of the specific loudness distance.

The occasionally very large deviations from the predicted value were found to be related to individual differences in the upward spread of masking. It was found that the computed partial loudness measure is sensitive to changes in the auditory model filter parameters. This sensitivity is most pronounced for modifications localized at spectral envelope valleys. An attempt was made to model individual differences as measured in a masked threshold experiment and to link the results with the experimentally measured partial loudness thresholds for one subject. In this way we were able to bring the results of this subject more in line with the results of the other subjects. This effort illustrates an approach to explaining individual variation in behavioral results, which is potentially useful in the development of robust tools for use in clinical settings. It is worthwhile to investigate in how far careful tuning of filter parameters for subjects can improve the results in terms of the relative variation.

Summarizing our results on the prediction of the discrimination thresholds for spectral envelope modifications of vowel sounds, we see that the partial loudness measure as well as the excitation pattern distance are equally appropriate measures for predicting audibility discrimination thresholds.

Moore investigated extensively supra-threshold differences for sounds in noise and evaluated the predictability of partial loudness on this in particular. It will be of interest to extend the present work to evaluate the performance of partial loudness measure in the prediction of supra-threshold differences in vowel quality.

ACKNOWLEDGMENTS

The work described in this paper was carried out while the first author was visiting IPO. The authors would like to thank Steven van de Par and Jeroen Breebaart for their help with setting up the experiments.

Bladon, R. A. W., and Lindblom, B. (**1981**). ''Modeling the judgement of vowel quality differences,'' J. Acoust. Soc. Am. **69**, 1414–1422.

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). ''A four-parameter model of glottal flow,'' Speech Transmission Laboratory Quarterly Progress Report 4/85, KTH, 1–3.

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

Kewley-Port, D. (**1991**). ''Detection thresholds for isolated vowels,'' J. Acoust. Soc. Am. **89**, 820–829.

Kewley-Port, D., and Zheng, Y. (**1998**). ''Auditory models of formant frequency discrimination for isolated vowels,'' J. Acoust. Soc. Am. **103**, 1654–1666.

Levitt, H. (**1971**). ''Transformed up-down methods in psychoacoustics,'' J. Acoust. Soc. Am. **49**, 971–995.

Moore, B. C. J. (**1987**). ''Distribution of auditory filter bandwidths at 2 kHz in young normal listeners,'' J. Acoust. Soc. Am. **81**, 1633–1635.

Moore, B. C. J., and Glasberg, B. R. (**1987**). ''Formulae describing frequency selectivity as a function of frequency and level and their use in calculating excitation patterns,'' Hear. Res. **28**, 209–225.

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). ''A model for the prediction of thresholds, loudness and partial loudness,'' J. Audio Eng. Soc. **45**, 224–240.

Plomp, R., and Steeneken, H. J. M. (**1969**). ''Effect of phase on the timbre of complex tones,'' J. Acoust. Soc. Am. **46**, 409–421.

Plomp, R. (**1976**). *Aspects of Tone Sensation* (Academic, London).

Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. (**1988**). *Objective Measures of Speech Quality* (Prentice Hall, Englewood Cliffs, New Jersey).

Schroeder, M. R., Atal, B. S., and Hall, J. L. (**1979**). *Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception. Frontiers of Speech Communication Research* (Academic, New York).

Sommers, M. S., and Kewley-Port, D. (**1996**). ''Modeling formant frequency discrimination of female vowels,'' J. Acoust. Soc. Am. **99**, 3770–3781.

Stevens, S. S. (**1957**). ''On the psychophysical law,'' Psychol. Rev. **64**, 153–181.

van der Heijden, M., and Kohlrausch, A. (**1994**). ''Using an excitation-pattern model to predict auditory masking,'' Hear. Res. **80**, 38–52.

Zwicker, E., and Scharf, B. (**1965**). ''A model of loudness summation,'' Psychol. Rev. **72**, 3–26.

Zwicker, E., and Fastl, H. (**1990**). *Psychoacoustics—Facts and Models* (Springer, Berlin).