Simulation of auditory analysis of pitch: An elaboration on the DWS pitch meter

Michael T. M. Scheffers
Institute for Perception Research, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

(Received 11 November 1982; accepted for publication 11 July 1983)

A model was developed for estimating the pitch of complex sounds that are partially masked by background sound. Our ultimate aim is to obtain a model that can separate two simultaneous sounds on the basis of the harmonic structure of at least one of the sounds. The MDWS model is an extension of the Duifhuis, Willems, and Sluyter pitch meter (DWS) [J. Acoust. Soc. Am. 71, 1568–1580 (1982)] which is a practical implementation of Goldstein's optimum processor theory of pitch perception [J. Acoust. Soc. Am. 54, 1496–1516 (1973)]. The main modifications incorporated in MDWS consist of a more faithful modeling of auditory frequency analysis and of an alteration to the criterion used to decide which fundamental best fits a set of resolved components. Effects of the latter modification were investigated in a comparison between model estimates of the pitch of inharmonic complex signals and results obtained for humans. Furthermore, the accuracy of model estimates of the pitch of periodic signals (among which were synthesized vowel sounds), partially masked by noise, was compared with the just noticeable difference of fundamental frequency of these sounds for human observers. The results of these two tests show that the model estimates come close to human perception.

PACS numbers: 43.66.Hg, 43.66.Ba [FLW]

INTRODUCTION

In many normal speech communication situations the voice of the speaker reaches the listeners' ears somewhat disturbed by background sound. In his classical study on this "cocktail party problem" Cherry (1953) found that human listeners have a surprising ability to separate the target sound from the background, even when the background sound is the voice of another speaker. As part of our research effort to investigate perceptual separation of simultaneous sounds by exploring the auditory and cognitive processes involved, an attempt will be made to model this human ability. In this research we are primarily interested in the role of pitch differences between the two sounds in the separation process (e.g., Broks and Nooteboom, 1982). For the model we therefore examined whether an existing model of pitch perception could be modified to serve our purpose. One of the most important characteristics of such a model should be that it is largely indifferent to signal components that are not harmonics of the target signal: simultaneous sounds will comprise many such spurious components viz. those of the second signal.

Most models are related to one of three recent theories of pitch perception, i.e., Wightman's pattern transformation theory (Wightman, 1973), Goldstein's optimum processor theory (Goldstein, 1973; Gerson and Goldstein, 1978), and Terhardt's virtual pitch theory (Terhardt, 1974, 1979, 1980). These three theories share the concept of a central formation of pitch on the basis of information derived from the frequencies of components resolved in the auditory frequency analysis. All three theories have at present reached a state in which they can be (or even have been) translated into a computer program. Goldstein's optimum processor theory is in our opinion the most promising for our purpose. The stochastic nature of this theory makes it most apt for quantitative predictions of the pitch perceived in periodic signals. A practical implementation of this theory is made in the DWS pitch meter of Duifhuis et al. (1979, 1982). The occurrence of spurious components is explicitly mentioned and dealt with in their description of the determination of pitch on the basis of a set of aurally resolved components. Preliminary tests showed that this meter can successfully extract the pitch of speech signals at low signal-to-noise ratios. This means that it can discriminate between signal components and noisy ones, which can be considered as a first step in separating a signal from the background noise.

The DWS meter, however, is optimized for the analysis of the pitch of normal speech sounds. It can make errors in determining the residue pitch of signals comprising only a few high (harmonic number above 5) harmonics. When a speech signal is disturbed by background sound, the lower harmonics might be masked. Some modifications are therefore proposed to reduce these errors. They consist of (1) a more faithful modeling of the auditory frequency analysis, which also led to a different component detection procedure and (2) a modification to the decision on the "best fitting fundamental" of the set of resolved signal components.

This paper introduces a model—which will be referred to as the MDWS model (Modified DWS)—which is an elaboration on the DWS meter. Results of a test of the performance of the model on complex signals comprising a few high harmonics will be described. Second, a comparison will be made between model estimates of the pitch of signals in a noise background and human pitch perception. The results of a listening experiment on discrimination of fundamental frequency of periodic sounds in a noise background (Scheffers, 1982) will to this end be compared with the accuracy of model estimates of the pitch of the experimental stimuli.
I. DESCRIPTION OF THE MDWS MODEL

A. Introduction

Because MDWS is essentially a modification of the DWS meter, a brief outline of the latter will be given first, followed by an argumentation for the proposed modifications.

Three stages can be discerned in the DWS procedure: (1) a frequency analysis of a short time frame of the acoustic signal, (2) detection of signal components, and (3) estimation of the fundamental frequency that optimally fits the set of resolved component frequencies. According to the Goldstein theory this fundamental frequency corresponds to the perceived pitch.

In the first stage a 128-point FFT amplitude spectrum is calculated for frequencies from 0 to 2.5 kHz of a 10- to 40-ms frame of the signal. The resolution bandwidth of this analysis is about 20 Hz.

In the second stage the spectrum is scanned for peaks from low to high frequencies. Each peak is checked (a) to be above an absolute threshold, (b) to be above a masked threshold that depends on the nearest accepted peak with a lower frequency, and (c) to have an acceptable shape. If a peak fulfills these conditions it is accepted as a signal component and a more accurate estimate of its frequency than provided by the FFT resolution is made by a quadratic interpolation on the three FFT points that define the peak. The estimated frequency values of the lowest six detected components form the input to the third stage.

In the final stage the harmonics sieve sifts out those components that bear a harmonic relation to one another and labels them with harmonic numbers. From the frequencies of the components that are accepted as harmonics and the harmonic numbers assigned to them, an optimum estimate is made of the best fitting fundamental frequency. This procedure will be more extensively described in Sec. I D.

As mentioned earlier, the DWS meter is optimized for a fast analysis of the pitch of speech signals. This necessitated some restrictions on, e.g., the frequency range, the maximum number of resolved components, and the lowest acceptable harmonic number. Because of these restrictions the meter may make errors in estimating the residue pitch of signals consisting of only a few high harmonics. The reason for stressing this point is that similar problems arise in estimating the phase relation between the individual components. How-

reasoning behind the choice of most parameter values for the model the reader is referred to the paper by Duifhuis et al. (1982).

B. Simulation of auditory frequency analysis

The signal is sampled at 10 kHz in the MDWS model, using a 12-bit resolution. A 256-point FFT power spectrum from 0-5 kHz is calculated of a 40-ms frame of the signal which is band-pass filtered from 50 Hz-4.5 kHz, shaped by a Hamming window and supplemented with 11.2 ms of silence. The levels of this spectrum are roughly corrected for sensation levels by shifting the whole spectrum upwards or downwards according to the sensation level of components around 1000 Hz. Auditory frequency analysis is then simulated by convolving this spectrum with stylized pure tone excitation patterns. This procedure is similar to the one applied in Wightman's pitch model (Wightman, 1973).

The excitation patterns are derived from the masking patterns used by Terhardt in his virtual pitch model (Terhardt, 1979; Terhardt et al., 1982). They have a triangular form on a dB versus log-frequency plot. For component frequencies above 400 Hz, the low-frequency slope equals 120 dB/oct and the high-frequency slope equals \(-20 \log_{10} L_{0} \) dB/oct with \( L_{0} \) the component level in dB SL. For frequencies below 400 Hz these slopes are adjusted to obtain a constant bandwidth in Hz of the patterns reflecting the frequency-independent critical bandwidth for lower frequencies (cf. Zwicker and Feldtkeller, 1967). The pattern of a pure tone is a convolution of the spectrum of the Hamming window and the stylized pattern. It is a little broader than the latter, has a rounded top, and a noise floor at about 43 dB (e.g., Harris, 1978) below the top. The bandwidth of a pattern will vary with the component level. For levels from 30 to 70 dB SL the equivalent rectangular bandwidth (ERB) for components with frequencies below 400 Hz is 70-120 Hz and the 3-dB bandwidth is 45-80 Hz. The ERB is 12%-24% of the frequency and the 3-dB bandwidth is 8%-16% for components with frequencies above 400 Hz. These properties are in close agreement with recent data on the width of the auditory filter (e.g., Patterson, 1976; Patterson and Nimmo-Smith, 1980; Pick, 1980). The simulation is illustrated in Fig. 1.

Because the shape of a pattern is not constant but dependent on the level and frequency of the component, it is more correct to describe the procedure as replacing the power level at each FFT point by the stylized excitation pattern of a pure tone with the same frequency and a corresponding sensation level. The excitation level at each point is then calculated by adding the power contributions of the excitation patterns for all other FFT points. Calculations are performed in floating point precision with a mantissa of 23 bits. The level resolution for the spectrum is therefore fully determined by the resolution of the A/D conversion (12 bits), yielding a value of less than \(-60 \) dB.

As usual in this kind of processing, phase is ignored because pitch perception seems to be largely independent of the phase relation between the individual components. However, this is true only for those components that can be resolved in the auditory frequency analysis. Because the com-
components of a periodic sound are correlated, phase effects might influence the detectability of a component. In a second version of the model, therefore, the excitation patterns are provided with corresponding minimum-phase characteristics in a first order approximation of the phase behavior within the pattern (cf. Goldstein et al., 1971). After calculating both FFT amplitude and phase spectrum each—now complex—point is replaced by its complex pattern. This leads to a set of amplitude and phase values at each individual FFT sampling frequency $f_i$ that reflect the influence of all components on the excitation waveform at this frequency. The time course of the waveform is reconstructed from these values on the assumption that all components are (quasi-) stationary sinusoids. Next, the rms value of the waveform is calculated over ten cycles of $f_i$. In this way the auditory system's increase in temporal resolution with increasing frequency is included. Figure 2 depicts the spectrum for frequencies from 0 to 5 kHz.

From a comparison between the power spectrum in Fig. 1(d) and the rms spectrum of Fig. 2 it can be seen that there is only a minor difference between the two versions. It was therefore judged not to be necessary to incorporate phase for the analysis of the pitch of signals with a single periodicity. Preliminary tests on signals consisting of two simultaneous periodic sounds showed major differences between the spectra produced by the two versions. This is probably because the harmonics of each of the sounds in signals obtained by adding two sounds with different fundamental frequencies are correlated, whereas in general no correlation will exist between the harmonics of the one sound and those of the other. This means that it might be useful to incorporate phase in an extension of the model to extract pitches of simultaneous sounds. It would also enable us to extract temporal information on the pitch from the excitation waveforms.

Combination tones (CT's) are not incorporated in the standard version (without phase effects) of the model because they will usually be masked in speech sounds by the lower harmonics. They are, however, introduced after the component detection stage for some signals such as described in Sec. II A. If CT's would be included in the version of the model that incorporates phase effects, it should be done at the convolving stage. Because that version of the model is not used in the tests to be described, no effort is made for the moment to incorporate CT's there.

C. Detection of components and estimation of their frequencies

Because the excitation patterns are derived from masking patterns, an aurally resolvable component should produce a maximum in the convolved spectrum. In the detection stage that spectrum is therefore scanned for local maxima and minima. The level of each maximum is compared with the levels of adjacent minima and when both peak-to-valley ratios exceed the—rather arbitrarily chosen—value of 1 dB the maximum is considered as being caused by the presence of a signal component. Through parabolic interpolation on the levels at the three FFT points that form the maximum the peak frequency is estimated with an accuracy which is an order of magnitude better than the FFT resolution and more in accordance with that of the auditory system (Goldstein, 1973). The interpolation also
gives information about the peak shape so that very shallow peaks can be rejected because they are in general caused by more than one component. Maxima with a level more than 43 dB below the absolute maximum in the spectrum are also rejected because they cannot be distinguished from side lobes of the Hamming window used in the frequency analysis. Only the estimated frequency values of components resolved in this way are retained for the pitch estimation stage.

This procedure leads to estimated frequency values which differ somewhat from the "real" component frequencies, owing to masking and the level of the component. These effects are—qualitatively at least—the same as the pitch shifts incorporated in the virtual pitch model (Terhardt, 1979). In view of the order of magnitude of these effects and their variance between subjects the qualitative presence is judged sufficient for MDWS.

D. Pitch estimation

In determining the pitch of the set of resolved components two problems arise. First of all, it must be decided whether a component is a genuine harmonic or a spurious component due to an interfering sound. Secondly, when a component is accepted as a harmonic, the corresponding fundamental must be found. In other words, the component must be given the correct harmonic number. For this the harmonics sieve procedure is introduced in the DWS meter. Below, a general description of this DWS procedure will be given, followed by a proposed modification.

The harmonics sieve has meshes at the harmonic frequencies \( f_n = n f_0 \) with \( n = 1 \) to \( N \). Each mesh has a width of 8% of the frequency to which it corresponds in order to allow for frequency shifts up to 4%. Successive meshes are not allowed to overlap. This restricts the number \( N \) of meshes to 12. The set of harmonic numbers that best fits the set of resolved component frequencies, can be determined by using the sieve. To this end the sieve is successively set at a number of positions in respect to the components. Each position is fully characterized by the fundamental of the sieve which varies from 50-500 Hz. A step size between successive positions of 3% of the fundamental frequency is chosen so that there is a slight overlap to minimize the chance of a component being missed.

At each position \( i \) a criterion value is calculated for the match of the sieve in this position to the set of components. This criterion is basically a measure of the difference between the pattern of the sieve and that of the resolved components. The value is diminished for each component that passes a mesh and augmented (a) for each component that cannot pass a mesh (a spurious component) and (b) for each component through which no component passes (a missing harmonic). The last decision cannot be taken for components with frequencies above the frequency to which the 12th mesh corresponds. These components are therefore disregarded in the criterion. The same is done for empty meshes above the highest one through which a component passes.

The mathematics underlying the criterion are extensively described by Duifhuis et al. (1982). The criterion \( C_i \) is given in Eq. (1).

\[
C_i = \frac{[(H_i - P_i) + R_i - P_i]}{P_i}.
\]  

In this equation \( P_i \) equals the number of components that pass the sieve in this position \( i \), \( H_i \) equals the harmonic number of the highest mesh through which a component passes and \( R_i \) equals the number of resolved components minus the number of components with a frequency above that of the 12th harmonic at position \( i \). Note that \( (H_i - P_i) \) equals the number of missing harmonics for the fundamental of the sieve in this position and that \( (R_i - P_i) \) equals the number of components that are rejected as harmonics of that fundamental. That position of the sieve which obtains the lowest criterion value is regarded as the best fit. Each signal component that passes a mesh of the sieve in this position is labeled with the harmonic number of that particular mesh. A more accurate estimate of the best fitting fundamental than the fundamental of the sieve at the best fitting position, is then calculated by means of the maximum likelihood estimate of Eq. (2) (Goldstein, 1973).

\[
f_o = \frac{\sum_{j=1}^{p} X_j N_j}{\sum_{j=1}^{p} N_j^2}.
\]  

In this equation \( X_j \) equals the estimated frequency of the component that passes the mesh with harmonic number \( N_j \) and \( P \) equals the number of components that are accepted as harmonics.

The criterion described above, works well on signals in which three or more low harmonics can be detected. However, errors occur on signals in which a number of low harmonics are missing, e.g., typical psychoacoustical stimuli, "telephone" speech, or vowel sounds that are partially masked by noise. It can be seen from Eq. (1) that for such signals a higher pitch than the correct one is favored by the criterion in order to reduce the number of empty meshes. The criterion value is regarded as the best fit. Each signal component that passes a mesh of the sieve in this position is labeled with the harmonic number of that particular mesh. A more accurate estimate of the best fitting fundamental than the fundamental of the sieve at the best fitting position, is then calculated by means of the maximum likelihood estimate of Eq. (2) (Goldstein, 1973).

\[
f_o = \frac{\sum_{j=1}^{p} X_j N_j}{\sum_{j=1}^{p} N_j^2}.
\]  

In this equation \( X_j \) equals the estimated frequency of the component that passes the mesh with harmonic number \( N_j \) and \( P \) equals the number of components that are accepted as harmonics.

The criterion described above, works well on signals in which three or more low harmonics can be detected. However, errors occur on signals in which a number of low harmonics are missing, e.g., typical psychoacoustical stimuli, "telephone" speech, or vowel sounds that are partially masked by noise. It can be seen from Eq. (1) that for such signals a higher pitch than the correct one is favored by the criterion in order to reduce the number of empty meshes. The criterion value is regarded as the best fit. Each signal component that passes a mesh of the sieve in this position is labeled with the harmonic number of that particular mesh. A more accurate estimate of the best fitting fundamental than the fundamental of the sieve at the best fitting position, is then calculated by means of the maximum likelihood estimate of Eq. (2) (Goldstein, 1973).

\[
f_o = \frac{\sum_{j=1}^{p} X_j N_j}{\sum_{j=1}^{p} N_j^2}.
\]  

In this equation \( X_j \) equals the estimated frequency of the component that passes the mesh with harmonic number \( N_j \) and \( P \) equals the number of components that are accepted as harmonics.

The criterion described above, works well on signals in which three or more low harmonics can be detected. However, errors occur on signals in which a number of low harmonics are missing, e.g., typical psychoacoustical stimuli, "telephone" speech, or vowel sounds that are partially masked by noise. It can be seen from Eq. (1) that for such signals a higher pitch than the correct one is favored by the criterion in order to reduce the number of empty meshes. The criterion value is regarded as the best fit. Each signal component that passes a mesh of the sieve in this position is labeled with the harmonic number of that particular mesh. A more accurate estimate of the best fitting fundamental than the fundamental of the sieve at the best fitting position, is then calculated by means of the maximum likelihood estimate of Eq. (2) (Goldstein, 1973).

\[
f_o = \frac{\sum_{j=1}^{p} X_j N_j}{\sum_{j=1}^{p} N_j^2}.
\]  

In this equation \( X_j \) equals the estimated frequency of the component that passes the mesh with harmonic number \( N_j \) and \( P \) equals the number of components that are accepted as harmonics.

The criterion described above, works well on signals in which three or more low harmonics can be detected. However, errors occur on signals in which a number of low harmonics are missing, e.g., typical psychoacoustical stimuli, "telephone" speech, or vowel sounds that are partially masked by noise. It can be seen from Eq. (1) that for such signals a higher pitch than the correct one is favored by the criterion in order to reduce the number of empty meshes. The criterion value is regarded as the best fit. Each signal component that passes a mesh of the sieve in this position is labeled with the harmonic number of that particular mesh. A more accurate estimate of the best fitting fundamental than the fundamental of the sieve at the best fitting position, is then calculated by means of the maximum likelihood estimate of Eq. (2) (Goldstein, 1973).

\[
f_o = \frac{\sum_{j=1}^{p} X_j N_j}{\sum_{j=1}^{p} N_j^2}.
\]  

In this equation \( X_j \) equals the estimated frequency of the component that passes the mesh with harmonic number \( N_j \) and \( P \) equals the number of components that are accepted as harmonics.

The criterion described above, works well on signals in which three or more low harmonics can be detected. However, errors occur on signals in which a number of low harmonics are missing, e.g., typical psychoacoustical stimuli, "telephone" speech, or vowel sounds that are partially masked by noise. It can be seen from Eq. (1) that for such signals a higher pitch than the correct one is favored by the criterion in order to reduce the number of empty meshes. The criterion value is regarded as the best fit. Each signal component that passes a mesh of the sieve in this position is labeled with the harmonic number of that particular mesh. A more accurate estimate of the best fitting fundamental than the fundamental of the sieve at the best fitting position, is then calculated by means of the maximum likelihood estimate of Eq. (2) (Goldstein, 1973).

\[
f_o = \frac{\sum_{j=1}^{p} X_j N_j}{\sum_{j=1}^{p} N_j^2}.
\]  

In this equation \( X_j \) equals the estimated frequency of the component that passes the mesh with harmonic number \( N_j \) and \( P \) equals the number of components that are accepted as harmonics.
obtains the highest $Q$ value is considered as the best fit.

$$Q_i = P_i / (M_i + R_i).$$  \hfill (3)

Figure 3 gives an illustration of the matching of the sieve to a set of component frequencies. The normalized $Q$ values ($Q_i / Q_{\text{max}}$) are shown in the left-hand graph.

In most cases only one optimal fit of the sieve is found. When the signal is heavily disturbed, however, two or more fits of the sieve sometimes obtain the highest $Q$ value. The lowest estimate is in these cases taken for the pitch. The other estimates, however, are also available in the output of this stage because pitch ambiguities can also occur in the perception in such conditions. The model does not make an estimate of the residue pitch when less than three components can be detected but produces the estimated frequencies of those components.

II. TESTING THE MODEL

A. Residue pitch of inharmonic signals

In order to test the performance of MDWS on signals comprising only a few harmonics, the pitch shift experiment by Schouten et al. (1962) was simulated on the computer. Stimuli in this experiment consisted of an AM signal with a modulation frequency $f_m$ of 200 Hz. Pitch matches were recorded as a function of the carrier frequency $f_c$ which varied from 1200 to 2400 Hz. Because it is generally assumed that aural combination tones affect the pitch perceived in this type of stimuli and cause the so-called second effect of pitch shift (Goldstein, 1973), the first two odd-order combination tones $2f_1 - f_2$ and $3f_1 - 2f_2$ were computed after the component detection stage of the model and added to the set of resolved components. Results of the simulation are shown in Fig. 4 together with the regression lines from Schouten et al. (1962, Fig. 2). When two or more positions of the sieve receive the maximum $Q$ value only the lowest $f_0$ estimate is normally produced by the model. Because we are in this case looking for pitch ambiguities, all $f_0$ estimates based on fits of the sieve that received the same (highest) $Q$ measure are plotted.

It should be mentioned that other "high quality" fits were generally found for pitch values near the edges of the sawtooth pattern in this figure. Schouten et al. (1962) also reported clusters of pitch matchings in those regions. The results show that the errors that DWS would make for carrier frequencies above 1400 Hz are avoided by the use of the new quality measure for a fit. The estimates by MDWS appear to be in close agreement with the experimental data. The deviations from the averaged data obtained for humans are of the same order of magnitude as the differences between the individual results.

B. Analysis of the pitch of periodic signals in a noise background

To test the performance of MDWS in noisy conditions, pitches of stimuli from an experiment on the discrimination by human listeners of fundamental frequency of periodic sounds in noise (Scheffers, 1982) were analyzed by the model. The stimulus set of this experiment contained synthesized vowel sounds and pulse trains with fundamental frequencies of 75, 150, and 300 Hz, and pure tones with frequencies of 150, 300, and 1000 Hz. The signals were masked by pink noise (3 dB/oct attenuation). For the verification of the model the means and standard deviations of pitch estimates by MDWS were determined. They were calculated from pitch estimates of 20 different presentations of the same stimulus at a fixed S/N ratio. This was done separately for the three vowel sounds and the pulse train at five S/N ratios, viz. 20, 15, 10, 5, and 0 dB S/N (the signal-to-noise ratio measured in a 10% band centered at the frequency where the highest 10%-band signal level was found, cf. Scheffers, 1982). The sd's will be compared with the just noticeable differences of fundamental frequency ($\text{jnd}_{f_0}$) for humans that were measured in the listening experiment. A similar comparison will be made between the frequency estimates of the pure tones by the component detection stage and the jnd's as a function of S/N ratio obtained for listeners.

After preliminary tests it was decided to modify the model further to bring it more into line with human perception. First, the accuracy with which human observers perceive characteristics such as the pitch and the loudness of a stationary sound increases with the duration of the signal (e.g., Cardozo, 1962; van den Brink, 1964; Henning, 1970). In the case of pitch this integration of information can be observed up to a duration of about 200 ms. The stationary part of the signals (which had a duration of 180 ms) was therefore divided in seven 40-ms frames, which overlapped by 20 ms. The "auditory" power spectra of these seven frames were computed and averaged. The component detection was performed on this averaged spectrum.

Second, it became clear that the accuracy of the model in estimating the component frequencies and thus the residue pitch was too good compared with that of human observers [see also Fig. 5(a)]. Goldstein (1973) has incorporated this accuracy in his theory by including the uncertainty in the conveyance of information on the component frequencies through stochastic channels. Similarly, an internal noise...
factor was therefore added to the frequency values of detected components in the present model. A random frequency value was drawn for each component from a Gaussian distribution with the estimated frequency of the component as mean and a standard deviation calculated from the sd \( f \) function of Eq. (4) (Goldstein, 1973, Fig. 7).

\[
\text{sd}(f) = \left( \frac{f}{10} \right)^{0.5}
\]

(sd and \( f \) in Hz).

Third, it is very probable that subjects used a priori knowledge on the pitch, because presentations started in the experiment at a high S/N ratio and the fundamentals were always in a small region during a run. Apart from the pitch estimation described in Sec. I D, the model therefore also made pitch estimates restricting the candidate fundamentals to regions from 10% below to 10% above the three stimulus fundamentals used viz. 75, 150, and 300 Hz. When analyzing signals of which the fundamental is not specifically known beforehand, the pitch estimate of the previous frame could, e.g., be used as the center frequency of such a “window.”

The width of the window could be chosen on the basis of the \( Q \) value for that estimate (as a measure of its reliability) such as is done in the DWS meter in a similar way. Therefore, the application of these preference regions will henceforth be referred to as tracking. The value of 10% is based on two findings: (1) On the results of the listening experiments which indicated that \( f_0 \) differences greater than about 5% could be discriminated as soon as the signals were just detectable, and (2) on the fact that if two or more candidate fundamentals obtained the highest \( Q \) value, they always differed more than 12% from each other. The present value avoids the necessity of a decision what to do with such multiple estimates. It sets at the same time a theoretical upper limit of 6% (for a uniform distribution of estimates) to the relative sd’s to be found, which is in good agreement with the perceptual effects mentioned before. It is mentioned here that a value, larger than 12% would in general yield a few more estimates to be found for the lower S/N ratios. This would result in a somewhat greater sd, but would also give rise to multiple estimates. I found in a pilot study that multiple estimates occurred in about 5% of the signals if a value of 15% was chosen and in 15% for a value of 20%.

The standard deviations of the pitch estimates of the pulse train are plotted in Fig. 5 (filled symbols, solid lines) together with the corresponding jnd \( f_0 \)'s obtained for human listeners (open symbols, dashed lines). The data points are connected with lines only if the average number of pitch estimates on the basis of which the sd was calculated was at least 15. A number near a data point indicates the number of estimates on which the sd is based if this number does not equal 20. Figure 5(a) shows the results for estimates without the internal noise factor and without tracking. Figure 5(b) gives the results in the condition where internal noise was added to the component frequencies but no tracking was applied. In Fig. 5(c) the results are plotted in the condition where internal noise was added and tracking was used. For Fig. 5(a) and (b) the pitch estimates for each S/N ratio were sorted out on “the same pitch” on the basis of a 10% tolerance of the value. Thus, 75.3 and 77.9 for example were considered as estimates of the same pitch, while e.g., 73.8 and 112.5 (a typical MDWS error) were considered as estimates of different pitches. The mean and the standard deviation of each subdivision of estimates obtained in this way were calculated. From the first results it appeared that the sd for the same set of stimuli could vary by a factor of about 1.5, depending on the actual values of the internal noise. This was partly due to the fact that the pitch estimates were not always based on the same set of component frequencies because of the internal noise. Each measurement was therefore repeated 20 times, each time with different samples of the internal noise for each detected component. The sd’s plotted in Fig. 5(b) and (c) are the averages of these 20 measurements.

The results show that the sd’s are far smaller than the

![FIG. 5. Comparison between jnd \( f_0 \) of pulse trains in noise for listeners (dashed lines, open symbols) and sd’s of model predictions of pitch (solid lines, closed symbols). Panel (a) gives the model predictions without the addition of internal noise to the frequencies of resolved components and without the application of tracking. In panel (b) the average sd’s of pitch estimates are plotted in the condition where internal noise is added to the frequencies of resolved components but no tracking is used. Panel (c) gives the average sd’s of pitch estimates when internal noise is added and tracking is applied. A number near a data point indicates on how many estimates the sd is based if not on 20.](image-url)
jnd’s for humans when no internal noise factor is added. The sd’s come close to the jnd’s when the uncertainty in the internal representation of the component frequencies is taken into account. The S/N₁₀ ratio at which MDWS can still make a reasonable amount of “correct” estimates lies then about 9 dB above the S/N₁₀ ratio where human listeners can only discriminate f₀ differences greater than 5%. The use of tracking for the pitch estimates lowers this limit by about 6 dB.

Figure 6 shows the sd’s of pitch estimates of the three vowel sounds in the condition when internal noise was added and tracking was applied. Each sd is again the average of 20 measurements.

The results show that the model cannot make pitch estimates for the vowel sounds at low S/N₁₀ ratios. At higher S/N₁₀ ratios a conformity between the sd’s and the jnd’s can be seen for the vowel /a/ with fundamental frequencies of 150 and 300 Hz and for the vowel /u/ with fundamentals of 75 and 150 Hz. For the vowel /i/, however, the sd’s differ an order of magnitude from the jnd’s. Even at the higher S/N₁₀ ratios the fundamental of this vowel cannot always be estimated by the model.

In discussing the experimental results obtained with listeners (Scheffers, 1982) it was argued that for the vowel sounds in particular, there were two other cues besides residue pitch on the basis of which the subjects could discriminate the differences in fundamental frequency. First, the pitch of a strong low-frequency harmonic could be dominant. Second, due to the fact that the sampling frequency at which the vowel sounds were generated was varied to obtain differences in fundamental frequency, the spectral envelope of the signal was stretched when the fundamental frequency was increased. This could result in a perceptible change of the vowel quality.

For the vowel /u/ the pitch of the strong 300-Hz component appeared to be dominant. The change in this pitch was probably used by the subjects to discriminate the differences in fundamental frequency below an S/N₁₀ of about 20 dB. For the vowel /i/ the pitch of a harmonic in the first formant (225 Hz) or the change in the spectral envelope in the region of the second to fourth formant (2280–3200 Hz) could be used as a clue. For the vowel /a/ the pitch of a harmonic in the first formant (750 Hz) was probably used at low S/N₁₀ ratios. I therefore altered the model to produce the estimated frequencies of the detected components as well. These values were again sorted out on “estimate of the same component” with a 10% tolerance. The means and sd’s of these subdivisions were calculated. This of course raises a problem for the high harmonics that differ less than 10% in frequency. This problem could only partly be solved, viz. when a number of successive harmonics was consistently detected in those cases.

A comparison between the sd’s calculated from Eq. (4) and of the jnd’s of pure tones (e.g., Moore, 1973) reveals a difference of an order of magnitude. This reflects the assumption that the component frequency analysis takes place at a lower level in the auditory system than the more central processing of pitch. The sd of the internal noise factor was therefore altered to that of Eq. (5) which is estimated on jnd’s for pure tones measured by Harris (1952), Walliser (1968), Moore (1973), and Wier et al. (1977) (for a review see Hoekstra, 1979).

\[ \text{sd}(f) = 0.03 \times f^{0.6} \]  

(5)

The sd’s of the frequency estimates of the pure tones and of the most dominant component of the vowel sounds are plotted in Fig. 7 together with the jnd’s for humans. A comparison between the sd’s calculated from Eq. (4) and of the jnd’s for pure tones measured by Harris (1952), Walliser (1968), Moore (1973), and Wier et al. (1977) (for a review see Hoekstra, 1979).

If we consider the results for the pure tones first, we see that the model can detect the signal consistently down to a few dB above masked threshold (indicated on the top horizontal axis). The sd’s follow a curve similar to the jnd’s, but are almost exactly a factor of 2 smaller.

Table I gives the number of harmonics, that are consistently detected by MDWS in each of the 20 stimuli used for determining the sd’s for the vowel sounds. The table also gives the number of harmonics that are detected in at least 15 of the 20 stimuli and the frequency of the most dominant harmonic f₀, viz. the harmonic that was detected most often down to the lowest S/N₁₀ ratios.

![Graph](image-url)
The 300-Hz harmonic was found to be the most dominant for the vowel /u/. The sd's of the frequency estimates of this harmonic follow a curve, more or less parallel to the experimental results. There is a difference of about a factor of 2 between the sd's and the jnd's, as is also found for the pure tones.

A similar observation can be made for the 225-Hz harmonic of the vowel /i/ with an f₀ of 75 Hz. The 300-Hz harmonic is the most dominant for the other two fundamentals. The sd's of the frequency estimates of this harmonic come close to the jnd's. The typical difference of a factor of 2 between the model predictions and the experimental results cannot, however, be observed for these two fundamentals.

The 750-Hz harmonic was detected most often by the model for the vowel /o/ with fundamental frequencies of 75 and 150 Hz. The 900-Hz harmonic appeared to be dominant for the 300-Hz fundamental. Again the sd's of the frequency estimates of these components differ by about a factor of 2 from the results obtained for listeners.

### III. DISCUSSION

The MDWS model was developed from the DWS pitch meter. Two main modifications were proposed in this paper, viz. a better approximation of auditory frequency analysis by convolving a high-resolution FFT spectrum with stylized excitation patterns and a modification to the criterion used to decide on the “best fitting fundamental” of a set of resolved components. This second modification was tested in a comparison between model estimates of the pitch of inharmonic complexes comprising three components and results obtained for humans. The results of this test show that errors induced by the absence of low harmonics are indeed removed. The model estimates are comparable to human perception. It is noteworthy that this practical model gives the same estimates as were predicted by Goldstein (1973) on a more theoretical basis. A limited test on normal speech utterances showed only minor differences between predictions by DWS and by MDWS.

### TABLE I. Number of harmonics of the vowel sounds that are detected by MDWS in each of the 20 stimuli (above the oblique), or in at least 15 of the 20 stimuli (below the oblique) and the frequency fₚ of the most dominant harmonic.

<table>
<thead>
<tr>
<th>vowel</th>
<th>/a/</th>
<th>/i/</th>
<th>/o/</th>
</tr>
</thead>
<tbody>
<tr>
<td>f₀ in Hz</td>
<td>75</td>
<td>150</td>
<td>300</td>
</tr>
<tr>
<td>20 dB S/N₁₀</td>
<td>2/3</td>
<td>2/4</td>
<td>2/2</td>
</tr>
<tr>
<td>15 dB S/N₂₀</td>
<td>1/3</td>
<td>2/3</td>
<td>1/2</td>
</tr>
<tr>
<td>10 dB S/N₃₀</td>
<td>1/1</td>
<td>1/2</td>
<td>1/1</td>
</tr>
<tr>
<td>5 dB S/N₄₀</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
</tr>
<tr>
<td>0 dB S/N₅₀</td>
<td>0/1</td>
<td>1/1</td>
<td>0/1</td>
</tr>
<tr>
<td>fₚ in Hz</td>
<td>300</td>
<td>300</td>
<td>300</td>
</tr>
</tbody>
</table>
The performance of the frequency detection stage of MDWS was investigated by a comparison between the jnd, of pure tones in noise for humans and the sd's of the frequency estimates of these signals by the model. The results of this test show that MDWS can identify pure tones in noise almost as far as subjective masked threshold. We consider the detection stage to be a good simulation of the auditory analysis at least for the signal durations that are used, particularly because in the listening experiment the noise onset preceded the signal onset by 300 ms, which facilitated the detection for the listener but was of no advantage for the model. The difference of a factor of 2 between the sd's of the model estimates of the frequency of the pure tones and the jnd, for humans are solely owing to the choice of the amount of internal noise. In view of the variance of jnd, measurements (see Hoekstra, 1979) one should not attribute too much importance to this difference. As for the effects of internal masking (the masking of a signal component by other components) the reader is referred to the papers by Terhardt (1979, 1982).

The results for the pitch estimates of the pulse trains and the vowels show that when the signal spectrum contains a number of harmonics which (a) are above the calculated masked threshold, (b) have successive harmonic numbers, and (c) lie within the existence region (Ritsma, 1962, 1963), the model can successfully predict the residue pitch and the sd's of the model estimates come close to the jnd, for listeners. These conditions are met down to S/N 10 ratios near subjective masked threshold for pulse trains. Higher S/N 10 ratios are required for vowel sounds to bring enough components above masked threshold so that the residue pitch can be estimated, because of the peaked spectrum of these signals. Residue pitch can therefore be estimated down to lower S/N 10 ratios for a vowel with a rather flat spectrum like the vowel /a/ than for the vowels /u/ and /i/ which have a more peaked spectrum or a flat part at frequencies outside the existence region.

One should note that even when the subjects in the experiment could perceive the residue pitch of a signal at a certain S/N 10 ratio, this does not necessarily mean that they used this pitch to discriminate the presented difference in fundamental frequency. In fact, subjects often reported performing the discrimination task on the pitch of a prominent harmonic rather than on the residue pitch in particular for the vowel sounds at low S/N 10 ratios. The sd's of the estimates by MDWS of the frequencies of dominant harmonics show that the results for the listeners can indeed often be predicted on that basis.

The results indicate two points which may need further investigation. First, the rather strong dependence of the sd values on the actual value of the internal noise. It appeared that the internal noise could be so strong that a harmonic was wrongly rejected by the harmonics sieve. This might be caused by the "all-or-nothing" decision of the sieve—viz. the shape of the meshes—or by the width of the meshes (especially for low frequencies). In view of the sd values of the internal representation of component frequencies as calculated from Eq. (5), one could argue that the meshes are too narrow for harmonics with frequencies below 250 Hz.

Second, the results for the vowels with a fundamental frequency of 75 Hz are poorer than were expected from the perceptual impressions. It is possible that the internal masking effects in the model are too strong. This may be caused by the use of a convolution followed by the component detection stage, as opposed to a detection stage followed by the incorporation of masking effects as is done in Terhardt's virtual pitch model (Terhardt, 1979; Terhardt et al., 1982). The convolution adds 10 to 20 Hz to the bandwidth of the excitation patterns which may result in too poor a resolving power for low frequencies. It is also possible that the level dependence of the high-frequency slope of the excitation patterns is too great.

IV. CONCLUSIONS

The modifications of the DWS pitch meter (Duifhuis et al., 1979, 1982) incorporated in MDWS have led to a practical model that can predict the pitch perceived by human listeners in normal speech signals equally as well as in typical psychoacoustical stimuli comprising only a few high harmonics.

The performance of MDWS on pulse trains that are partially masked by noise comes close to human perception. When tracking is applied, the model can successfully estimate the pitch of these signals down to about 3 dB above the S/N ratio where listeners apparently can just perceive residue pitch. Standard deviations of these pitch estimates are comparable with the jnd, for humans. Pure tones in noise can be identified down to about subjective masked threshold.

The comparison between the sd's of estimates by MDWS of the residue pitch of vowel sounds in noise and the jnd, for these signals for humans appears to be quite complex. Depending on the vowel sound, the model can predict the pitch of these signals down to 10 to 30 dB above masked threshold. It appeared that the jnd, at lower S/N ratios could in many cases be predicted on the basis of frequency estimates of a prominent low-frequency harmonic. This agrees well with perceptual impressions reported by the listeners in the experiment where jnd, were measured.

The success MDWS has in discriminating between signal components and noisy components leads us to believe that it will be a powerful tool in our study into the perceptual separation of simultaneous speech sounds.

ACKNOWLEDGMENTS

I would like to thank H. Duifhuis, S. G. Nooteboom, and R. D. Patterson for their stimulating support and for their assistance in the preparation of this paper. This research was supported by the Netherlands Organization for the Advancement of Pure Research (Z.W.O.) through the Netherlands Psychonomics Foundation, grant number 15-31-011.

1Actually, $C_i = (H_i + R_i)/P_i$ is used in the DWS meter. This reduced form is mathematically simpler, but functionally equal to that of Eq. (1). The latter version is considered to be more transparent because it gives a direct reference to the fit parameters.

2During the preparation of this paper, a similar solution was published by Sluyter, Kottmans, and Claasen [Proc. ICASSP 82, 188-191 (1982)].

3The sd's should in fact be compared with $\sqrt{2}$ times the jnd, in view of the
fact that the amount of internal noise that had to be added to the frequencies of resolved components to bring the model predictions into line with human perception is only a first approximation, this factor is neglected in the comparison.

I realize that this is also a form of tracking. It was, however, necessary in order to separate the two or more "optimal fits" that can occur for the same stimulus (see I D). It also suppresses the enormous increase of the sd as a result of a single octave error. The value of 10% was chosen for reasons given at the end of the previous paragraph. The frequency with which the estimates were compared was in this case not fixed but was gradually built up during each series of measurements by adjusting it to the average of the estimates that were previously accepted.