Open access • Journal Article • DOI:10.1121/1.387811

Measurement of pitch in speech : an implementation of Goldstein's theory of pitch perception — Source link

Hendrikus Duifhuis, Lei Lf Willems, Robert Johannes Sluyter

Published on: 01 Jun 1982 - Journal of the Acoustical Society of America (Acoustical Society of America)

Topics: Relative pitch, Pitch detection algorithm and Pitch (Music)

Measurement of pitch in speech : an implementation of Goldstein's theory of pitch perception

Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception[a)]

H. Duifhuis[b)] and L. F. Willems

Institute for Perception Research IPO, Den Dolech 2, Eindhoven, The Netherlands

R. J. Sluyter

Philips' Research Laboratories, Eindhoven, The Netherlands

(Received 31 August 1979; accepted for publication 10 March 1982)

Recent developments in hearing theory have resulted in the rather general acceptance of the idea that the perception of pitch of complex sounds is the result of the psychological pattern recognition process. The pitch is supposedly mediated by the fundamental of the harmonic spectrum which fits the spectrum of the complex sound optimally. The problem of finding the pitch is then equivalent to finding the best harmonic match. Goldstein [J. Acoust. Soc. Am. **54**, 1496–1516 (1973)] has described an objective procedure for finding the best fit for stimuli containing relatively few spectral components. He uses a maximum likelihood criterion. Application of this procedure to various data on the pitch of complex sounds yielded good results. This motivated our efforts to apply the pattern recognition theory of pitch to the problem of measuring pitch in speech. Although we were able to follow the main line of Goldstein's procedure, some essential changes had to be made. The most important is that in our implementation not all spectral components of the complex sound have to be classified as belonging to the harmonic pattern. We introduced a harmonics sieve to determine whether components are rejected or accepted at a candidate pitch. A simple criterion, based on the components accepted and rejected, led to the decision on which candidate pitch was to be finally selected. The performance and reliability of this psychoacoustically based pitch meter were tested in a LPC-vocoder system.

PACS numbers: 43.70.Gr, 43.70.Ny, 43.66.Hg, 43.66.Ba

INTRODUCTION

By and large the problem of how to determine the time course of pitch in continuous speech is treated as a purely technical issue. The problem can be formulated as follows: given an (acoustic) waveform which is almost periodic, determine the "pitch period." An ancillary task is to discriminate between aperiodic and (almost) periodic waveforms (unvoiced/voiced). Several pitch detection algorithms aiming at solving the problem have been discussed and evaluated by Rabiner *et al.* (1976).

The process of data reduction, which transforms an acoustic waveform into a single number that characterizes its pitch, obviously requires decision criteria to specify what information is to be retained/extracted and what to be discarded. On the whole those criteria have been chosen on the basis of optimal signal processing, treated as an engineering problem. These studies tend to pay little attention to perceptual aspects of pitch.

There is, however, an alternative approach to the problem, which, in our belief, can be highly successful. To begin with, pitch (e.g., of speech) is a subjective quantity. Therefore one might argue that the pitch meter which operates according to the principles of the human pitch extractor (the auditory system) will attain the optimum level of performance. This is undoubtedly the case if the optimization concerns the simulation of subjective pitch perception. However, many pitch meters find an implementation in vocoder systems. Here pitch information is used to trigger the "glottal pulses" in the synthesis part of the vocoder. Because pitch is not related in a simple way to glottal pulse period, the optimization for pitch perception performance is not necessarily equally effective in a vocoder context. The present study, which explores this effectiveness, has been set up with the hope that the distinction between pitch and glottal period measurement would be largely academic. We work from the point of view that a pitch meter, which performance relies on perceptual data, is a useful tool in vocoder techniques. The development of theories of pitch perception over the last decade provides support for optimism about the results of this approach. The vast amount of published data on pitch of complex tones (residue, repetition pitch, musical pitch, virtual pitch; see de Boer, 1976, for a review) formed a solid basis for this theoretical work. Although the theories are based on results of psychoacoustical experiments with "laboratory signals" which are usually much simpler than speech sounds, the extrapolation of these results to speech sounds would seem to be justifiable (see, e.g., Schouten, 1962). In one aspect speech sounds are simpler than the complex sounds used in psychoacoustic experiments: they contain more frequency components and in general evoke an unambiguous pitch percept. On the other hand, a difficulty of the speech sound is that pitch in speech is continuously varying, and psychoacoustic experiments have so far mainly been concerned with stationary stimuli. This difficulty can be dealt with in a pragmatic way. The related question is how coarsely the pitch contour can be sampled without affecting the perceived melodic line. This constraint

[a)] Some preliminary results have been presented at the EBBS workshop "Hearing Mechanism and Speech" April 1979, Göttingen, and to the 97th ASA meeting, June 1979, Cambridge, MA, paper Y7.

[b)] Present address: Department of Biophysics, Laboratory for General Physics, Westersingel 34, Groningen, The Netherlands.

 0001-4966/82/061568-13$00.80

touches upon the question of analysis window and processing time, and thus on the question of "real time" measurement of pitch (see Sec. IIA).

A successful quantitative theory of the subjective perception of the pitch of complex tones has been developed by Goldstein and his associates (e.g., Goldstein, 1973; Gerson and Goldstein, 1978; Goldstein *et al.*, 1978). We propose that (1) this theory is also applicable to the (subjective) perception of pitch in speech and (2) that the theory can be put into the form of an (objective) algorithm which will produce pitch values that have a psychophysical validity as well as practical applicability. This validity stems from the fact that the data reduction in the algorithm proposed here is based on constraints known from hearing theory, which in turn relies on psychoacoustical and physiological data.

In this paper we will not go into the details of the psychoacoustics of pitch. We restrict ourselves to a description of Goldstein's theory. We shall then discuss the additional steps that are involved in its application to speech material. Finally, the resulting algorithm is presented together with some data on its performance. The algorithm will briefly be compared with existing algorithms. As an example we present results of a direct comparison with the parallel processing pitch detector (PPROC) by Gold and Rabiner (1969).

I. GOLDSTEIN'S THEORY ON THE PITCH OF COMPLEX SOUNDS

A. Introductory remarks

The long-standing issue as to whether pitch is mediated through temporal aspects or frequency content of the acoustic waveform has reached an important milestone during the last decade. In particular the experiments by Houtsma and Goldstein (1972) revealed that residue pitch is perceived when the frequency components of the stimulus are separated and presented to different ears of the listener. This implies that residue pitch is the result of a synthesis which takes place at some level *after* the cochlea, where auditory frequency analysis occurs. The synthesis can be considered a spectral pattern recognition process. On different grounds essentially the same interpretation had been proposed by de Boer (1956) and Whitfield (1970). In the beginning of the last decade several theoretical studies appeared aiming at describing this pattern recognition process in detail. In addition to Goldstein's (1973) theory two other theories were published by Terhardt (1972, 1974) and Wightman (1973). However, their models of the spectral pattern recognizer are not specific enough to allow straightforward quantitative predictions to be made. In other words, they could not be translated into a working algorithm. de Boer (1977) has attempted to unify these views, but in our opinion the original theory of Goldstein (1973) is more transparent. It is acknowledged that Goldstein's theory, and thus our pitch extractor, does not account for phenomena such as the effects of level and partial masking on pitch, which are accounted for in Terhardt's theory. However, the most elaborated and quantitative theory proves to be best suited for practical implementation. Recently Terhardt (1979) has reformulated his theory in a more quantitative way. In this current form it contains some elements that are virtually identical to parts of our procedure. These will be indicated in Sec. IV.

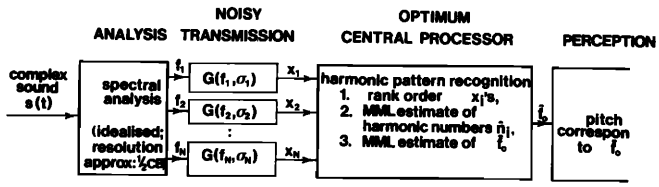

FIG. 1. Schematic block diagram of Goldstein's optimum processor theory for the "central formation" of pitch of complex sounds. The spectral analyzer resolves components that are less than approximately 1/2 CB (Fig. 2) apart and measures the frequencies. These are transmitted through independent noisy channels to a central processor. The central processor optimally fits a harmonic pattern to the received frequencies. The fundamental of the harmonic pattern corresponds to the wanted pitch (after Goldstein, 1973).

B. Outline of the theory

Given a complex sound (by definition a sound comprising more than one spectral component), the following steps can be distinguished (see Fig. 1).

(1) The peripheral ear performs a frequency analysis which reveals what frequency components are present. (The resolving power is limited, amplitude and phase information are removed.) The number of resolved components is N.

(2) Information on each resolved frequency component $f_i (i = 1, N)$ is conveyed stochastically to a "central processor." This provides the central processor with a set of independent stochastic representations (described with Gaussian probability density functions) of the component frequencies

$$f_i \rightarrow X_i, \quad p\,df(x_i) = G(f_i, \sigma_i), \tag{1}$$

where

$$G(f_i, \sigma_i) = (2\pi\sigma_i^2)^{-1/2} \exp[-(x_i - f_i)^2/2\sigma_i^2].$$

(3) The standard deviation σ_i depends only on the component frequency

$$\sigma_i = \sigma(f_i). \tag{2}$$

This is a result from matching the theory to psychoacoustical data rather than an *a priori* assumption.

(4) The central processor makes an optimum estimate (maximum likelihood estimation) of the unknown stimulus fundamental on the assumption that the stimulus frequencies are unknown harmonics. It turns out that this estimation can be split into two successive steps. The first optimally labels the frequencies with harmonic numbers $\hat{n}_i$, the second determines the maximum likelihood estimate of $f_0, \hat{f}_0$, based on the set of X_i's and corresponding $\hat{n}_i$'s.

(5) The residue pitch corresponds to the estimated fundamental $\hat{f}_0$.

By considering the central processor as a system that has to match a set of frequencies to a harmonic pattern, the relation to pattern recognition is emphasized. The pattern, however, is simple: given the harmonic structure it is fully determined by a single parameter, viz. f_0.

In the following subsections the steps in Goldstein's pitch extraction scheme are discussed in more detail.

C. Auditory frequency analysis

The inner ear performs an auditory frequency analysis which is roughly characterized by a bank of bandpass filters. The effective bandwidth of the filters is approximately equal to the so-called critical band. Although the audio frequency range is often divided into 24 successive critical bands, the peripheral ear actually works with 30 000 channels that innervate at least 3000 different inner hair cells. In other words, in so far as the critical bandwidth is a good characteristic of the selectivity of the channels, it is by no means an indication of the number of independent channels. So if we want to resolve the acoustic spectrum in a way similar to the auditory resolution we will have to work with bandwidths that are related to the critical bandwidth but with a spacing of tuning frequencies that is much narrower. Of course there will then be some correlation between information of neighboring channels, due to partially overlapping filter characteristics. The critical bandwidth is approximately 100 Hz for frequencies up to 500 Hz, and 20% of the tuning frequency above 500 Hz (Fig. 2, see Zwicker and Feldtkeller, 1967, p. 74 for precise data). According to Plomp (e.g., 1976, Chap. 1) the ear can identify components as long as their frequencies are separated by more than 15% to 20% with a minimum distance of about 60 Hz. This distance agrees reasonably well with the critical bandwidth. Goldstein uses a somewhat better resolution of 10% on the basis of an interpretation of available data in terms of his theory. The bandwidth determines two factors in the further analysis. First, of course, the frequency selectivity, but second, and not less important, the temporal resolution. The uncertainty relation in the frequency-time description states (Stewart, 1931; Gabor, 1947):

$$\Delta f \Delta t \geq 1 . \tag{3}$$

This means that a time window with an effective duration of 10 ms produces a spectral broadening of at least 100 Hz (effective bandwidth), and conversely, that a resolution of 100 Hz requires a time window with an effective duration of 10 ms. Assuming a worst case resolution (i.e., the narrowest bandwidth) of about 50 Hz (half the critical band) for component frequencies around and below 500 Hz one arrives at a time window (temporal integration time) of 20 ms. This being the effective duration, the total duration of a shaped time window will be about twice this size, i.e., 40 ms. Ideally, the time window should be shorter for frequencies above 500 Hz.

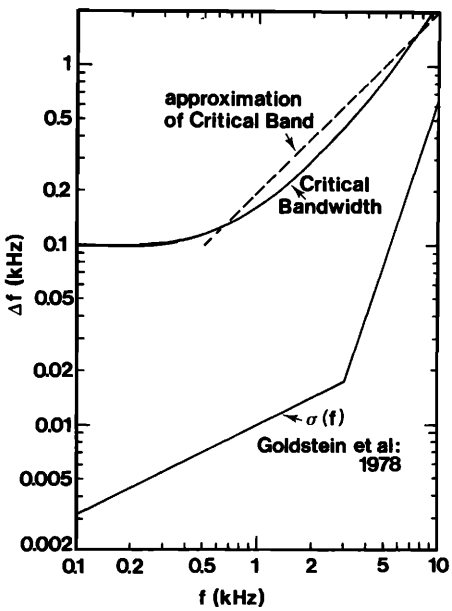

FIG. 2. A plot of the critical band (CB) against center frequency. The dashed line gives a simple approximation: $\Delta f = 100$ Hz if $f < 500$ Hz and $\Delta f/f \approx 20\%$ if $f > 500$ Hz. The lower function $\sigma(f)$ characterizes the noisiness of the channels in Fig. 1. The function is a stylized result of a fit to psychoacoustical data (Goldstein *et al.*, 1978).

D. Stochastic transduction

Whereas the peripheral frequency analysis determines the limits of resolving neighboring components, the accuracy with which frequencies become available to the central processor is determined by the noisiness in the stochastic channels. It turned out that the description in terms of Gaussian noise in the channels [Eq. (1)], characterized by a standard deviation that depends on frequency only [Eq. (2)], gives an acceptable account of the data. For σ Goldstein *et al.* (1978) propose the following schematic relation to f:

$$\sigma = 0.01 f^{1/2}, \quad f < 3 \text{ kHz} ,$$
$$\sigma = (0.01/9\sqrt{3}) f^3, \quad f \geq 3 \text{ kHz} \tag{4}$$

(σ and f in kHz).

For frequencies below 5 kHz, σ is one order of magnitude smaller than the critical bandwidth (Fig. 2). On the other hand, the value of σ is about one order of magnitude greater than the difference limen in frequency.

The assumption of independent stochastic channels is in line with the neurophysiological finding that responses in auditory nerve fibers from a single ear are stochastically independent (Johnson and Kiang, 1976). The only correlation found between responses in different fibers stems from the fact that the channels respond to the "same" stimulus in so far as their peripheral filters overlap.

E. The central processor

Given the representations X_i ($i = 1$ to N) of the frequencies f_i ($i = 1$ to N), which are harmonic, then the likelihood function to be optimized for the best estimate of f_0 is

$$L = \prod_{i=1}^{N} G(f_i, \sigma_i). \tag{5a}$$

Instead of maximizing L, it is standard practice to maximize $\Lambda = \log L$, which can be written as [using Eq. (1)]

$$\Lambda = -\frac{N}{2}\log 2\pi - \sum_{i=1}^{N}\log\sigma_i - \sum_{i=1}^{N}\frac{(x_i - n_i f_0)^2}{2\sigma_i^2}. \tag{5b}$$

The optimum estimates of n_i and f_0 ($\hat{n}_i$ and $\hat{f}_0$) are made when the terms in the right-hand part of Eq. (5b) are minimum. It is reasonable to assume that the second term is relatively insensitive to optimization of n_i and f_0 because σ varies slowly with f over the frequency range of most interest ($f < 3$ kHz). Maximizing Λ is then equivalent to minimizing the mean square error of "data" and matched harmonics:

$$\epsilon^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - n_i f_0)^2}{\sigma_i^2}. \tag{6}$$

Assume for a moment that the optimum values of n_i ($\hat{n}_i$) are known, then $\hat{f}_0$ follows from

$$\left.\frac{\partial\epsilon^2}{\partial f_0}\right|_{f_0=\hat{f}_0} = 0,$$

which, after some calculation, gives

$$\hat{f}_0 = \sum_{i=1}^{N}\frac{x_i\hat{n}_i}{\sigma_i^2} \Big/ \sum_{i=1}^{N}\frac{\hat{n}_i^2}{\sigma_i^2}. \tag{7}$$

Besides the value of the estimated fundamental, its accuracy is important. It turns out that errors in estimates of f_0 stem in practice almost entirely from errors in the estimated set of harmonic numbers. If we denote candidate sets for $\{\hat{n}_i\}$ as $\{m_i\}_l$, with $l = 1$ to L then the probability density function of f_0 will in general have L distinct modes, each of which is relatively narrow. For a typical value of $\sigma_i/f_i = 0.01$ and a number of components $N = 6$, the relative mode width is $\sigma_{0l}/f_{0l} \approx 0.004$, or 1 Hz for $f_{0l} = 250$ Hz. This meets the required accuracy range closely enough and is in good agreement with Ritsma's (1963) data on the accuracy of residue pitch. A systematic discussion on σ_0, including the basis for the above estimate, is given in Goldstein's (1973) paper.

Apparently, then, it is important to select the right set of harmonic numbers. Goldstein (1973) and Goldstein *et al.* (1978) demonstrate that two factors determine the probability of selecting the right set. This is illustrated in Fig. 3, which, for successive harmonics, gives a plot (from Goldstein *et al.*, 1978) of $P(\{m_i\}_l = \{\hat{n}_i\})$ as a function of the lowest harmonic number n_1 and the number of components N. The trends are clear: the lower the value of n_1 and the larger the value of N, the greater will be the probability of estimating the proper set $\{m_i\}_l$ and hence the greater the probability that $f_{0l} = \hat{f}_0$. Although the result of Fig. 3 was determined for successive harmonics, it is fairly obvious that similar trends will apply to the situation where the harmonics are not successive. Figure 3 shows that, given a lowest harmonic number $m_1 \leq 7$ and the number of harmonics $N \geq 6$ the probability of finding the correct pitch is near 100%. It seems reasonable to assume that these conditions can usually be fulfilled in speech, so that virtually no mode errors are expected in the pitch of speech.

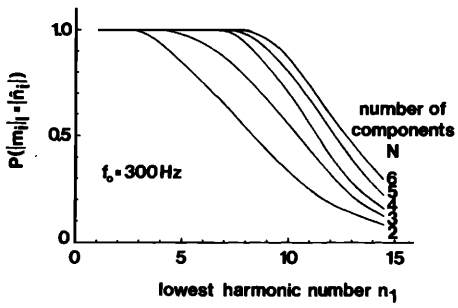

FIG. 3. The probability of correctly estimating the harmonic numbers of the components as a function of the lowest harmonic number presented. Parameter is the number of components. In this example, at $f_0 = 300$ Hz, it is assumed that all components are successive harmonics (after Goldstein *et al.*, 1978).

II. APPLICATION OF GOLDSTEIN'S PITCH THEORY TO CONTINUING SPEECH

A. General outline

The optimum pitch-measuring device can be thought to consist of two elements: a spectral analyzer that detects and measures the frequencies of the harmonic components, followed by an optimally functioning harmonic pattern recognizer (Fig. 4). The properties of analyzer and recognizer are matched to those of the model that describes human pitch perception (Sec. I). On the other hand they are adapted to current software and hardware techniques in digital signal processing. For the software algorithm we allow a nonreal-time solution provided that the prospect for a real-time hardware implementation would be left open and even considered feasible with present hardware technology. As we have seen that pitch is a subjective quantity that requires integration over a finite time interval, we have to allow for a delay of the order of this interval, i.e., of about 40 ms (Sec. IC). Updating of varying pitches may be required to be faster than this. For the moment we will assume an interval of 10 ms for this purpose.

Although it is common practice to smooth the measured pitches according to the expected pitch value, or, in other words, to determine the *a posteriori* pitch, we will not include such procedures in this study. Of course they are helpful in reducing error rates and in economizing the procedures. However, it was deemed

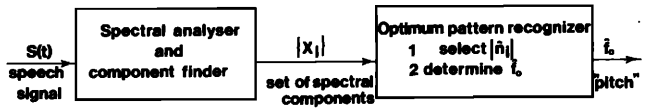

FIG. 4. Schematic block diagram of the pitch meter. First the spectrum of the speech signal is measured and component frequencies are determined. On the basis of the frequency values the pattern recognizer optimally estimates f_0.

more fruitful to try to optimize the *a priori* estimate of the pitch, so that the algorithm would give independent new estimates on successive samples. This aim had to be relaxed when we defined a voiced/unvoiced decision rule. A weak form of tracking was used which is based on the reliability of the computed pitches.

B. The spectral analyzer and component finder

1. Analyzer

Spectral analyzer and component finder have to produce the set X_i with an accuracy that is comparable to that characterized by the subjective $\sigma = \sigma(f)$ function. This implies a $\sigma \approx 3$ Hz at $f = 100$ Hz to $\sigma \approx 10$ Hz at $f = 1$ kHz. It is an obvious choice to use FFT for the frequency analysis. This, however, fixes σ for all frequencies. Therefore the resolving power in the FFT should be high enough to discriminate the harmonics of the lowest possible fundamental, which will be around 50 Hz. For Δf one thus has $\Delta f \leq 25$ Hz, which implies a time window of 40 ms. Since the frequency range which encompasses the relevant harmonics depends on f_0 and since the resolution required depends on f_0 very much like the ear's resolving power depends on frequency (Fig. 2), we introduced a feedback from f_0 to the time window duration T_w. The duration T_w was made inversely proportional to f_0 when f_0 was in the range from 100 to 400 Hz. For $f_0 \geq 400$ Hz we used $T_w = 10$ ms, for $f_0 \leq 100$ Hz $T_w = 40$ ms. This rule was applied only when a reliable pitch measurement had been made. In case of uncertainty T_w was set to 40 ms. This procedure is an *ad hoc* attempt to implement a resolving power which depends on frequency, in line with the size of the critical bandwidth (Sec. IC). In order to determine the frequencies of the maxima in the spectrum with sufficient accuracy, i.e., roughly a factor 10 better than the FFT, the peaks in the spectrum were located on the basis of parabolic interpolation of three neighboring spectral points.

In combination with Δf, the frequency range to be covered determines the number of points to be used in the FFT. The upper bound of the frequency range is determined by the product of the highest f_0 to be expected and the highest harmonic number that carries information, n_{max}. We expect f_0 not to exceed 500 Hz and n_{max} to be in the range of 10 to 15. However, we also expect that in the case of high fundamental frequencies the lowest harmonics will always be present. And even if $n_1 = 3$ a number of two successive harmonics would yield a 100% correct estimate of the set $\{n_i\}$ and hence of $\hat{f}_0$ (see Fig. 3). Therefore we decided to fix the maximum frequency to be analyzed at 2.5 kHz. It is noted that the existence region of the residue extends to 5 kHz (Ritsma, 1962). The value of 2.5 kHz, therefore, is somewhat small, but in practice we found it more than adequate. This sets the number of points at 256. With $f_{max} = 2.5$ kHz the sample frequency is 5 kHz, so that with 256 points the Δf becomes $\Delta f = 19.5$ Hz and the time window 51.2 ms. This window was filled with 10 to 40 ms of signal supplemented by 41.2 to 11.2 ms of silence (zeros).

The required word length in bits follows from signal-to-noise considerations. The Hamming window used produces a "noise" floor at 40 dB below the highest peak. This signal-to-noise ratio is roughly matched by a quantization into 8 bits, given a stationary amplitude. For our software simulation we have so far used an A/D conversion of 12 bits and a floating point FFT with a mantissa of 23 bits. This turned out to be sufficient to allow us to deal successfully with regular amplitude variations.

2. Component finder

So far Goldstein has not examined the effect of near-threshold components. He uses the simple rule that suprathreshold components count, independently of their amplitudes. In order to be applicable to natural sounds the theory requires the specification of a threshold. In fact even two thresholds will have to be specified. First, an absolute threshold, determined by the threshold of audibility, and second a relative threshold, which comes into operation in the context of other components or noise and which is determined by the psychophysical masked threshold. Apart from the requirement that the component amplitudes have to exceed both thresholds, the amplitudes play no role in the analysis.

For each local maximum in the amplitude spectrum $\{AF(r)\}$, $r = 1$ to 128, where

$$AF(r) \geq AF(r-1) \cap AF(r) > AF(r+1), \tag{8}$$

it is checked whether $AF(r)$ is above threshold; then, by parabolic interpolation, amplitude and frequency of the peak are determined and finally the shape of the peak is checked. The expected peak shape for a stationary spectral component follows from the Fourier transform of the Hamming window (e.g., in Harris, 1978), it is straightforward to calculate the spectral sample values around a peak. Let a peak occur at $f_r = r\Delta f$, then the ratio $AF(r \pm 1)/AF(r) = 1 - \varphi(T_w)$, where $\varphi(T_w)$ runs from 0.03 to 0.4 as T_w changes from 10 to 40 ms. In general a peak occurs at $f = (r + \delta)\Delta f$, with $-0.5 \leq \delta < 0.5$. Parabolic approximation of the peak shape yields for the expected values around the peak

$$\hat{A}F(r+i) = [1 - \varphi(T_w)(i - \delta)^2]\hat{A}F(r + \delta), \tag{9}$$

where $i = -1, 0, 1$ for the points of interest, and $\hat{A}F(r + \delta)$ is the calculated peak level. We used as error measure for the goodness of peak shape

$$e^2 = \sum_i \frac{[\hat{A}F(r+i) - AF(r+i)]^2}{[\hat{A}F(r+\delta)]^2} = \sum_i \epsilon_i^2 [1 - \varphi(T_w)(i-\delta)^2]^2, \tag{10}$$

where the observed $AF(r+i) = \hat{A}F(r+i)(1+\epsilon_i)$. The error measure is a weighted sum of the squared relative differences between expected and observed spectral heights. A peak was accepted as component X_i whenever $e^2 < 1/4$. This rather lax threshold is required because spectral peaks in real speech signals tend to be broadened by nonstationarity.

As mentioned above, there are two thresholds for $AF(r)$ to exceed in order to qualify as a significant component. The first is the absolute threshold. Implementation of the auditory threshold would require a calibration of the system regarding sound pressure

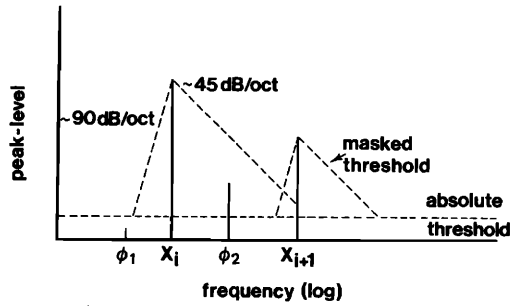

FIG. 5. After components are identified as local spectral maxima, it is checked whether they are above threshold. The components have to exceed an absolute threshold (determined by quantization noise, etc.) and a "masked" threshold, determined by masking slopes (stylized) connected with the spectral components. In the example, the peaks at X_i and X_{i+1} qualify. Those at ϕ are subthreshold and therefore rejected.

level. It is more practical to use a fluctuating threshold, related to the highest spectral peak or to the total energy of the sample. This takes care of window "splatter" and quantization noise (cf. Sec. IIB*1*). We set the first threshold level at 26 dB below the highest peak level, if this threshold exceeded a fixed minimum value. The automatic gain control involved in the updating of the threshold was of the fast-in–slow-out type; the decay time constant was 100 ms. The other threshold is the masked threshold. One of two components can be masked completely by the other. A simplified strategem that can be used is to assume that the presence of a component elevates the threshold to a ~45-dB/oct slope on the high-frequency side and to a 90-dB/oct slope at the other side (cf. Duifhuis, 1972). In the example in Fig. 5 the candidate ϕ_2 is masked by the component X_i, so that it does not count as a regular component. The values given for the slopes are to be considered as typical and as being roughly in accordance with auditory critical band filter characteristics. Actually the slopes of the masking pattern depend on component frequency as well as on component level. In practice the high-frequency side of the masking pattern (the 45-dB/oct slope) will present more consequences than the low-frequency side. In the results to be presented we used only this high-frequency slope.

Terhardt (1979) also uses absolute and masked thresholds as criteria for relevance of spectral components. His algorithm gives, at the cost of more complexity, a rather precise account of the dependence of the masking pattern on frequency and level.

The component finder starts looking for components at the low-frequency end of the spectrum, and it never looks for more than six components. The output of the component finder then consists of an array $\{X_i\}$, $i=1$ to N, with the parabolically interpolated peaks that fulfilled the several criteria. Formally, then, the number of components found, N, is restricted to the range $0 \leq N \leq 6$.

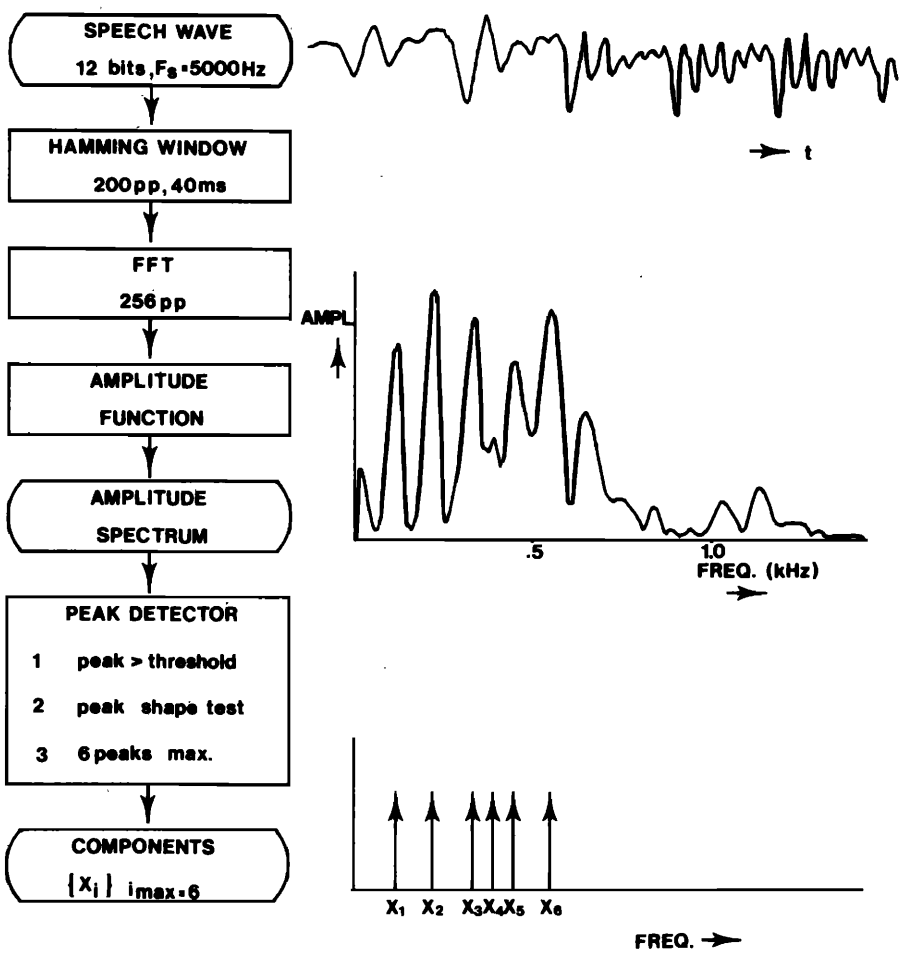

FIG. 6. Flow diagram of the spectral analyzer and component finder. The speech signal is low-pass filtered (at 2.5 kHz) and A/D converted as indicated. Every 10 ms, a 40-ms sample is spectrally analyzed (FFT). The amplitude spectrum is determined, $AF(r\Delta f)$, $r=1$ to 128, and local maxima are detected. For suprathreshold maxima, component frequency and amplitude are determined. Then it is verified whether the peak shape meets the wanted criterion (parabolic match), after which stage the amplitude information is discarded. If six components are found or if the entire spectrum is examined ($r \leq 127$), the process stops. The information on the components is carried on to the harmonic sieve.

A flow diagram of spectral analyzer and component finder is presented in Fig. 6.

C. The harmonic pattern recognizer

At this point it is necessary to note a fundamental difference between the problems of finding pitch in speech and finding pitch for a psychoacoustical stimulus. In our case the set of components $\{X_i\}$ is less clean. In speech as well as in psychoacoustical stimuli certain harmonic components may be lacking. However, in the speech spectrum one may also, despite the criteria mentioned in the above subsection, encounter spurious components that bear no relation to the harmonic signal. They arise either from irregularities in the speech waveform or from interfering background sound. Thus our problem is to find a best fitting harmonic pattern to the set $\{X_i\}$, without necessarily having to classify all N components.

We now describe a harmonic pattern recognition procedure which we will refer to as the harmonic sieve procedure. The purpose of the sieve is to establish which components are genuine harmonics and which are not. The latter will not pass through the sieve, but the harmonics will. The harmonics sieve is a one-dimensional sieve in the frequency domain (see Fig. 7). The sieve has meshes of a bandwidth $W = W(f)$ around the harmonic frequencies $f_j = jf_0$, with $j = 1$ to J. The value of J reflects that only the lower 7 to 15 harmonics contribute significantly to residue pitch, or $7 \leq J \leq 15$. So far we have used $J = 11$, in accordance with Goldstein (1973). In approximate accordance with auditory frequency resolution, the widths of the meshes are chosen to be proportional to their center frequencies, i.e., $W(f) = 2\alpha jf_0$. In order for the sieve to be effective at all meshes, successive meshes are not allowed to overlap. Since W increases with f, this implies

$$(1-\alpha)Jf_0 > (1+\alpha)(J-1)f_0$$

or

$$\alpha < 1/(2J-1) = 1/21 \approx 0.05. \tag{11}$$

Of course, $W(f)$ must be wide enough to allow for the errors that can arise in the component finder. These errors are denoted by $\sigma = \sigma(f)$, and should not exceed the value of Eq. (4). This leaves us with a value of α of a few percent. We will next find a bound for the minimum value of α.

The harmonic sieve procedure now amounts to successively setting the sieve to all possible values of fundamental frequencies, covering the entire range encountered in human speech (50–500 Hz). Of course the frequency domain is scanned in discrete steps (index l, $l = 1$ to L), the size of each being taken proportional to f. Obviously the step size should be smaller than $W(f)$ in order not to miss parts of the frequency scale. Minimizing the total number of steps, L, is equivalent to maximizing $W(f)$ or α. In general we used $\alpha \approx 5\%$ and a step size of 1/24 octave or approximately 3%.

At each position of the sieve, characterized by the fundamental frequency value f_{0l}, it is checked which components pass through the sieve, thus qualifying as candidate harmonics. A component X_i passing through mesh j is labeled with the candidate harmonic number $m_{kl} = j$. Let the total number of components passing through the sieve be $K_l (k_l = 1$ to K_l; l refers to the sieve position). If more than one component pass through the same mesh, then only the one closest to the center is labeled, the other is rejected. Figure 7 together with Table I illustrate the procedure with an example.

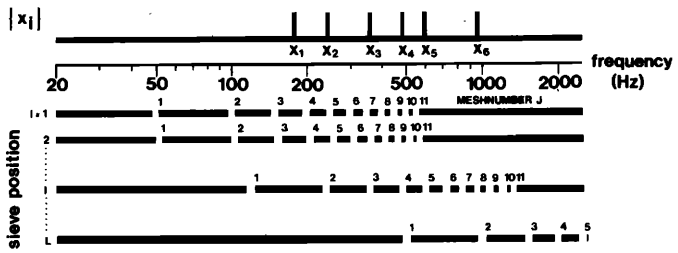

FIG. 7. Example of the harmonic sieve procedure: the component finder produced the set {177, 242, 360, 485, 600, 960 Hz}. The components are plotted on a log-frequency scale. The components are then sifted with a harmonics sieve, which has meshes 1 to 11 at harmonic intervals. The mesh width is approximately 8%. The position of the sieve is characterized by, for instance, that of mesh number 1, which starts 50 Hz. Then it moves to 500 Hz in steps of 3%. At each position it is checked which components pass through the sieve. Results for the present example are given in Table I.

On the basis of the results of the sifting we have to decide now which set of candidate harmonic numbers $\{m_k\}_l$ is the optimum set $\{\hat{n}_k\}$. This is equivalent to recognizing the harmonic pattern of which the set $\{X_i\}$ is a (noisy) sample. A common classifier in pattern recognition techniques is the so-called minimum distance classifier. Candidate set and reference set (ideal harmonic pattern) are both represented as vectors in a multidimensional space. The Euclidian distance between the endpoints of the vectors is a measure of the fit: the best fitting candidate is the one with minimum distance to the reference. The dimension of the space depends on $\{m_k\}$ and may differ from one sieve position to another (l). In order to compare adequately across l we consider the normalized distance, d, i.e., the distance divided by the "unit diagonal" (the square root of the dimension).

At position l the dimension of the space sufficient to encompass $\{m_k\}_l$ and the reference set is determined as follows: denote the highest candidate harmonic m_{Kl} as M_l. Then the dimension D is M_l plus the number of unclassified X_i's $(N - K_l)$, in order to allow orthogonal representation of all relevant components: $D = M_l + N - K_l$. The set $\{x_i\}$ is represented by the vector v, the elements of which are

$v_j = 1$ if $j \in \{m_k\}_l$, when X_i is accepted, or if $M_l < j \leq D$, when X_i is a rejected component

$v_j = 0$ otherwise.

The reference set is characterized by the vector u, given by $u_j = 1$ for $1 \leq j \leq M_l$ and $u_j = 0$ for $M_l < j \leq D$. The squared distance between u and v is

TABLE I. Example of classification by the harmonic sieve.

Sieve position	f_{0l}	Component frequencies (Hz) X_1	X_2	X_3	X_4	X_5	X_6	Effective input number	Total number classified	Highest harm. No. classified	Criterion
l	(Hz)	177	242	360	485	600	960	N_l	K_l	M_l	C_l
				Classified as							
1	50	...[a]	$m_{11}=5$	$m_{21}=7$	$m_{31}=10$	*[b]	*	4	3	10	14/3
2	53	...	...	$m_{12}=7$	...	*	*	4	1[c]	7	11/1
.											
.											
.											
.											
l	120	...	$m_{1l}=2$	$m_{2l}=3$	$m_{3l}=4$	$m_{4l}=5$	$m_{5l}=8$	6	5	8	14/5
.											
.											
.											
L	500	...	...	...	$m_{1L}=1$	...	$m_{2L}=2$	6	2[c]	2	8/2

[a]The three dots indicate that the component is rejected by the sieve.
[b]The star indicates rejection because the estimated harmonic number would be greater than 11. Components rejected with a star do not add to N_l.
[c]These fits are rejected immediately because K_l (the number of components classified) $<N_l/2$ (half the number to be recognized).

$$d_l^2=(M_l+N-2K_l)/(M_l+N-K_l). \tag{12}$$

It is straightforward to show that minimizing d_l or d_l^2 is equivalent to minimizing the quantity C_l defined as

$$C_l=(M_l+N)/K_l, \tag{13}$$

which form is somewhat simpler than Eq. (12).

The alternative approach of minimizing the angle between candidate vector and reference vector leads to a criterion that bears some relation to Eq. (13) and amounts to minimizing C_l^* defined as

$$C_l^*=M_lN/K_l^2. \tag{14}$$

However, in practice the criterion of Eq. (13) proved to perform slightly better.

The minimum of C_l over $l=1$ to L thus indicates the optimum set of harmonic numbers looked for. The best estimate of f_0 then follows from substitution of this set in Eq. (7). Actually in the algorithm used so far $\sigma(f)$ does not depend on frequency, so that Eq. (7) reduces to

$$\hat{f}_0=\sum_{i=1}^{K} x_i\hat{n}_i \Big/ \sum_{i=1}^{K} \hat{n}_i^2. \tag{15}$$

(This estimate is more accurate than simply taking $f_0=f_{0l}$ for the l that minimizes C_l; however, the additional accuracy may not always be needed.)

A minor complication arises if component frequencies are rejected because they lie above the highest mesh of the sieve. Such components may nevertheless be harmonic so they should not contribute to the distance in Eq. (12). This is remedied by defining an effective number of components at sieve position l as $N_l=N$ minus the number of X_i for which $X_i>(11+\alpha)f_{0l}$, and by replacing N by N_l in Eqs. (12) to (14). The overall, rather lax restriction that at least half of the components found should be classified as harmonics, or $K \geq N/2$ $(N>0)$, ascertains rejection of the trivial "zero solution" $N_l=0$.

The harmonic sieve procedure is much more efficient than the straightforward optimum estimation procedure of calculating ϵ^2 for all possible permutations of harmonic numbers and selecting the solution that minimizes ϵ^2 (Gerson and Goldstein, 1978). Moreover, it is not overly sensitive to spurious components.

The implementation of tracking is described in the next subsection.

D. Voiced/unvoiced discrimination

Evaluation of the pitch meter in a vocoder setting requires an adequate voiced/unvoiced decision rule. For this purpose we developed a set of rules, which, however, has not been optimized to the same extent as the pitch analyzer. It is not clear whether hearing theory can provide insight into this point because a listener appears to be quite unaware of the voiced/unvoiced transitions during an utterance. Instead he perceives a continuous melodic line.

The starting point of our rules is that a speech sample which produces a good fit to the harmonics sieve, i.e., yielding a C_l [Eq. (13)] close to 2, is obviously voiced. The acceptable disparity from 2 was made to depend on the number of fitting components, K_l,

$$C_l \leq 2.1+0.1K_l, \quad \text{for } K_l>1. \tag{16}$$

A pitch for which the inequality is satisfied is judged reliable. The only acceptable sieve match for $K_l=1$ can occur for $N_l=1$, i.e., when the spectrum contains only one qualifying spectral component. It can be accepted either as fundamental, or, in case of tracking, as second or third harmonic.

Tracking is used in two ways. First, if the previous pitch was reliable according to Eq. (16), then a tracking range half an octave wide is centered around this pitch value. Within the tracking range potential matches are favored by using $C_l'=C_l/2$ instead of C_l

for optimizing f_{0i}. The best match within the range is accepted if $C_i \leq 3.5$, even though lower values of C_i might have been obtained outside the tracking range. Secondly, if the previous sample has been classified as voiced, then the current sample is called voiced as long as the best C_i is less than 3.5.

Any acceptable f_0 within the range from 50 to 500 Hz classifies the speech segment as voiced.

III. PERFORMANCE

We implemented the pitch-measuring algorithm described above in a FORTRAN IV computer program,[1] run on a P857 minicomputer. As mentioned in Sec. IIA, in this phase of the project we did not aim at real-time operation, and transparency of programs was favored to parsimony.

The speech material used in this study was borrowed from a set of Dutch test sentences developed for audiologic tests by Plomp and Mimpen (1979). Twenty-five sentences were copies of the original material (female speaker), 25 sentences were re-recorded with a male speaker. The speech waveform was low-pass filtered at 5 kHz and sampled at 10 kHz using a 12 bit A/D conversion, and then stored on disk. These signals were subjected to a tenth-order LPC analysis, yielding ten filter coefficients and the amplitude parameter. The LPC analysis operated on 25-ms segments, shaped with a Hamming window and pre-emphasized by a first-order filter $1-\mu z^{-1}$ with $\mu = 0.9$. The LPC analysis was executed every 10 ms.

The pitch analysis used the same stored signals, but they were low-pass filtered (digitally) at 2.5 kHz, and sampled down to 5 kHz. The signals are processed with the algorithm described in Sec. II, thereby creating pitch files and voiced/unvoiced parameters which line up with the LPC parameters.

For a comparative judgment of the performance of our pitch meter we also implemented the parallel processing pitch detector (PPROC) of Gold and Rabiner (1969), using the FORTRAN programs by Rabiner and McGonegal (unpublished report). It used the same material as our meter (which we will designate the DWS detector in this section). PPROC was used in this evaluation because it belongs to the set of pitch meters which has been evaluated objectively by Rabiner *et al.* (1976) as well as subjectively by McGonegal *et al.* (1977). PPROC ranked among the better algorithms (e.g., third in the subjective test) and it happened to be the test which was available in full detail so that a fair comparison was possible.

The pitch analysis results of DWS and PPROC were used in a software resynthesis of the test material. The comparative performance was evaluated in a preference test where each sentence was presented successively in each of the two versions, in random order. Twenty listeners took part in the test. Ten of them had experience in phonetics or in psychoacoustics, the others were naive listeners. Although some listeners interpreted the task as a two-alternative-forced choice task, with the response alternatives (prefer DWS; prefer PPROC), most listeners included a third response alternative, viz. (no preference).

The results of the preference test are presented in Table II. Four out of the 50 test sentences were used in an introductory session. The data in the table are based on the responses to the 46 remaining sentences, half of which are pronounced by a male speaker (m) and half by a female speaker (f). The overall result of the test indicates a marked 2.7 over 1 preference for DWS over PPROC. The "no preference" responses form a small category. In 92% of the presentations the listeners came up with a preference response. Dividing the responses in the "no preference" class equally over the two other classes results in the binary total response. The differences between results for male and female speakers and for experienced and inexperienced listeners are considered marginal. Interindividual differences are characterized by a standard deviation of approximately 10%. All subjects showed a greater than 50% preference for DWS (range 52%–85%).

In other words, the present test shows a clear preference for the DWS-pitch algorithm over PPROC. On the basis of this limited data it is, of course, not possible to make general claims on the performance of our meter as compared to other known algorithms, but the results

TABLE II. Results of the preference test, averaged across the test sentences and the subjects within the two categories.

	Prefer PPROC			No preference			Prefer DWS		
Speaker / Listener	m	f	av	m	f	av	m	f	av
	(in %)			(in %)			(in %)		
Experienced ($n=10$)	19	26	22	9	9	9	72	65	69
Unexperienced ($n=10$)	30	24	27	4	7	6	65	69	67
Total			25			7			68
Binary total			28						72

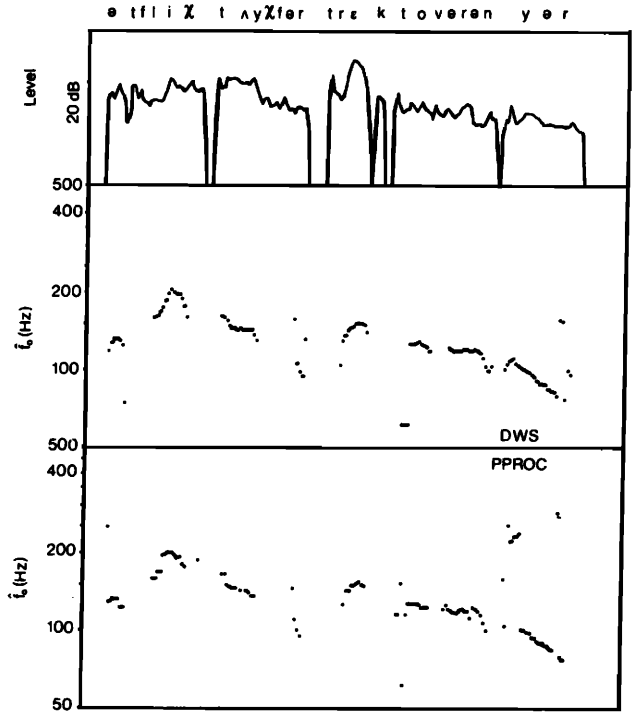

FIG. 8. Unsmoothed f_0 measurements from both DWS and PPROC pitch detectors of an utterance by a male speaker. The amplitude contour and a broad phonetic transcription are lined up with the f_0 contours.

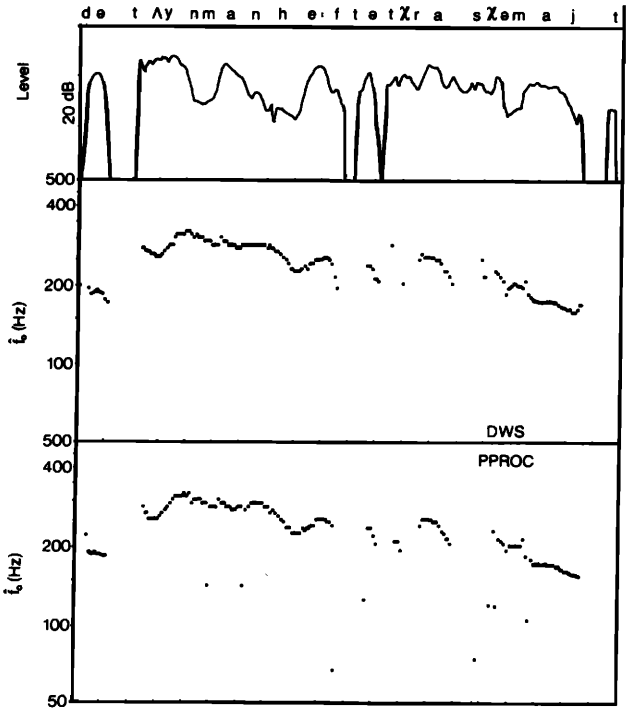

FIG. 10. As Fig. 8, female speaker.

obtained so far are promising. This statement is also based on informal results of a comparison with an advanced autocorrelation method used at our institute (Vogten and Willems, 1977).

Figures 8 to 11 present examples of the performance of the two pitch algorithms, which are selected from the set of 46 sentences used in the above test. In the upper part one finds the phonetic transcription of the utterance and a sound level measure based on the rms amplitude in each segment. The lower two panels give the f_0 measurements for the two algorithms. The utterances are judged unvoiced at the points where no pitch values are displayed.

It is clear that both PPROC and DWS have little difficulty in catching the overall melodic line in an utter-

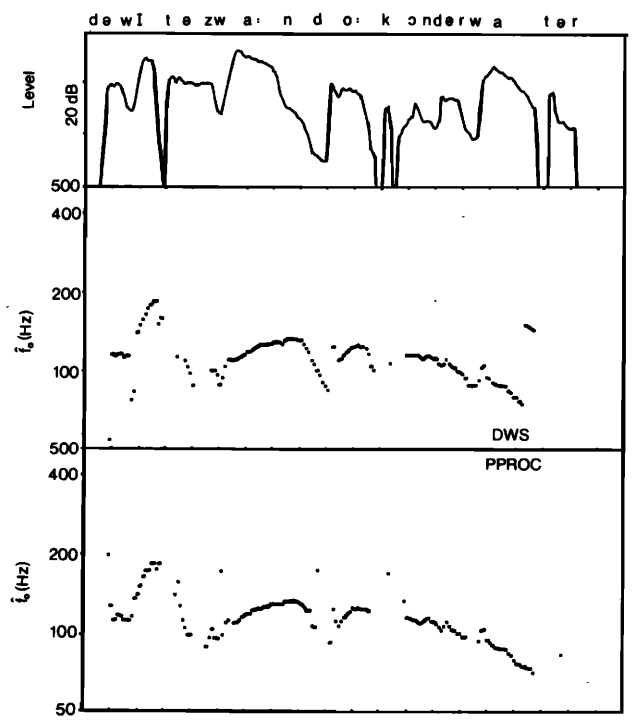

FIG. 9. As Fig. 8, male speaker.

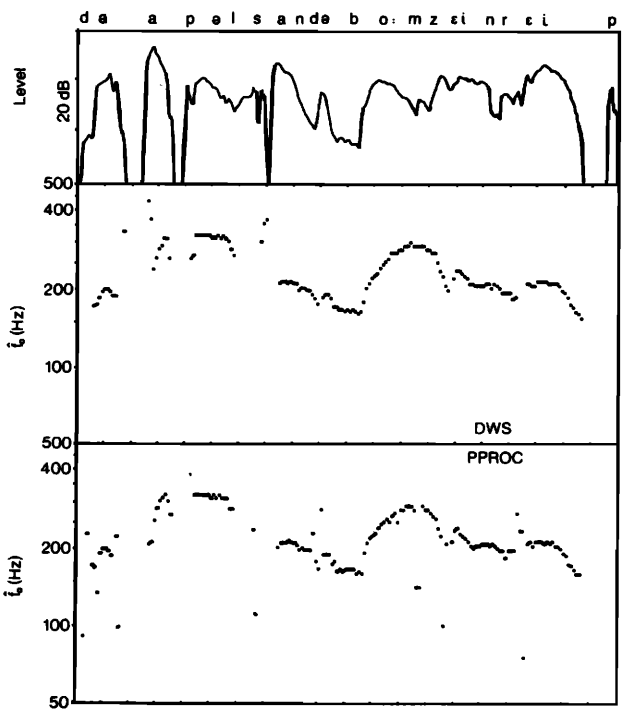

FIG. 11. As Fig. 8, female speaker.

ance. Problems arise almost exclusively in the voiced/unvoiced decisions. We argued that hearing theory does not have a ready solution to the voiced/unvoiced problem because in listening to a normal utterance one does not have a clear voiced/unvoiced percept. However, it is also apparent that the ear is quite sensitive to errors in the voiced/unvoiced decision in a vocoder system. This indicates that the ear has a clear picture of the expected value of the signal, given its past. It is not clear, as yet, whether this picture exists on a peripheral auditory level or on a more central level. We suspect, however, that the more fruitful application of perception to this issue goes beyond the level of classical hearing theory and would require a link with human speech recognition.

PPROC and DWS both show few octave errors. The total number of octave errors as judged by visual inspection from the pitch records as shown in Figs. 8 to 11 amounts to 19 for PPROC and 16 for DWS, i.e., 2.4% and 2.0%, respectively.

So far we have tested our pitch detector only in a limited set of conditions. Relevant tests include performance for different speakers and for poor signal to noise conditions. Unfortunately, these tests fall beyond the scope of the present study, which wants to emphasize the new approach rather than its complete evaluation.

IV. DISCUSSION

Performance of our pitch meter has been compared directly with performance of the parallel processing meter (PPROC) of Gold and Rabiner (1969). The latter method ranks amongst the best discussed by Rabiner *et al.* (1976) and by McGonegal *et al.* (1977). The results of the new pitch meter proved to compare favorably with PPROC. The primary differences are that our psychoacoustically based pitch meter (DWS):

(1) makes fewer octave errors;

(2) cannot be trapped at an incorrect pitch because it employs a weak form of tracking, and that only on the basis of "reliable" pitches;

(3) seems to perform better in poor signal-to-noise conditions such as low signal levels and voiced/unvoiced transitions (as judged from performance at amplitude ramps).

It is noted in passing that a final smoothing element added to our algorithm would produce a better pitch record in voiced speech. But, of course, this also enhances performance of most other pitch meters.

The first results of this evaluation lead to the conclusion that the pitch meter, based on Goldstein's (1973) theory on the perception of pitch of complex sounds, performs better than PPROC. The parallel processing meter PPROC operates exclusively on information in the time domain. Its mechanism is not directly comparable with the principle of our meter. A comparison with frequency domain pitch meters appears more appropriate, although some of the characteristics of frequency domain meters have, of course, their logical counterparts in the time domain. First we compare our procedure with the somewhat related "cepstrum" method. Three differences present themselves. The first difference is that in our method (as in Goldstein's) a severe data reduction takes place after the first spectral transformation. The actual spectrum is in fact reduced (or sharpened) to a line spectrum, where the lines occur at the suprathreshold peaks in the amplitude spectrum. The sharpening caused by this data reduction apparently leads to a "sharper" ultimate result. The second difference is that we base our estimate of f_0 on the lowest harmonics in the spectrum (highest harmonic number ≤ 11, highest harmonic frequency $\leq$ minimum of 11 f_0 and 2.5 kHz). Higher (measurable) harmonics are not represented with sufficient accuracy to allow a reliable estimate of their harmonic numbers and hence of their fundamentals. Therefore taking higher harmonics into account leads to a decrease in performance rather than an increase. They do not carry useful, retrievable information in the frequency domain. The third, and probably most important, difference is that our procedure makes explicit use of the fact that the frequency components stem from a harmonic sequence. Each estimate for f_0 uses specific harmonic number information of each component. The estimated harmonic numbers result from a simultaneous optimum fit of all components. One obvious restriction which follows from this is that harmonic numbers of different components have to be different. This, in combination with the criterion for best fit, reduce the occurrence of octave errors.

It is noted that the first two points mentioned above are, to a certain extent, also applied in the pitch detector proposed by Seneff (1978). Like ours, her detector consists of a peak-picker followed by some sort of pattern recognizer. The peak-picker covers a frequency range up to 1.1 kHz, which is still more than an octave narrower than ours. Her recognizer works on the basis of peak distances, taking into account the fact that in a harmonic spectrum the distances between peaks are all equal to f_0. This procedure, we would argue, does not make optimum use of the information carried by the frequencies. It only uses the restriction that successive peaks probably stem from successive harmonics, but it does not use the estimates of the harmonic numbers. Moreover, in a slightly inharmonic signal the pitch is not equal to that of the difference tone (Schouten, 1940) but to that of the fundamental of the best harmonic fit (e.g., de Boer, 1956; Goldstein, 1973). Pilot experiments by one of the present authors with a detector similar to Seneff's lead us to believe that our harmonic pattern recognizer gives a better performance.

The harmonic sieve procedure described in Sec. IIC is formally almost identical to Terhardt's (1979) procedure of finding near coincidences of subharmonics of the component frequencies. The basic difference lies in the specification of the criterion for the best fit. Terhardt maximizes the number of fitting components K_I, without taking account of noisy or missing components, as happens in DWS by maximizing $K_I/(M_I+N)$. Spurious noisy components are dealt with in an interactive (nonautomatic) way. As noted by Terhardt, his

pitch extractor is related to the period histogram method by Schroeder (1968) and the HIPEX system by Miller (1970). In these systems the near coincidence of subharmonics is determined in the time domain, although Schroeder also mentions the possibility to do the computation in the frequency domain. Schroeder is not specific on the criterion for coincidence, but the values used by Miller and by Terhardt are similar to the ones proposed in this paper. Besides the difference in optimization criterion, additional differences with DWS are that the separation of spectral components occurs much more coarsely (using a filter bank) and secondly that the level of spectral components plays a role. Miller (1970) claims that the performance of HIPEX is similar to that of cepstrum, which ranked among the lowest both in the evaluation by Rabiner *et al.* (1976) and in that by McGonegal *et al.* (1977).

Obviously not all constraints that were derived in Sec. II, and based on psychoacoustic data, are entirely new for technical pitch extractors. For those that are not (e.g., a 40-ms window) the discussion in Sec. II may provide an additional support.

There are reasons to believe that the robustness of our pitch meter in poor signal-to-noise conditions should result in a performance similar to human performance. This point needs further psychoacoustical study. Furthermore, the present procedure will probably prove successful in voice separation algorithms (cf. Parsons, 1976), albeit that the criterion for best fit will have to be adapted. It will have to reflect the hypothesis that, for example, two harmonic signals were presented.

V. CONCLUSION

We have implemented Goldstein's (1973) theory of pitch perception in a practical algorithm which measures pitch in speech. The application of the insights of hearing theory leads to a very successful pitch meter. Theoretically, performance would be equal to human performance; practically, this goal appears to be within reach if not reached yet.

The performance of the proposed procedure compares favorably to that of the parallel processing pitch detector (PPROC, Gold and Rabiner, 1969). Prospects for real-time hardware implementation [Sluyter *et al.*, (1980)] as well as for application to voice separation systems are promising.

[1]The texts of the programs which implement the DWS pitch detector are available on request as IPO report 394 by L. F. Willems.

de Boer, E. (1956). "On the 'residue' in hearing," Doctoral dissertation, Univ. Amsterdam.

de Boer, E. (1976). "On the 'residue' and auditory pitch perception," in *Handbook of Sensory Physiology*, edited by W. D. Keidel and W. D. Neff (Springer, Berlin), pp. 479–583.

de Boer, E. (1977). "Pitch theories unified," in *Psychophysics and Physiology of Hearing*, edited by E. F. Evans and J. P. Wilson (Academic, London), pp. 323–334.

Duifhuis, H. (1972). "Perceptual analysis of sound," Doctoral dissertation, Eindhoven University of Technology.

Gabor, D. (1947). "Acoustical quanta and the theory of hearing," Nature 159, 591–594.

Gerson, A., and Goldstein, J. L. (1978). "Evidence for a general template in central optimal processing for pitch of complex tones," J. Acoust. Soc. Am. 63, 498–510.

Gold, B., and Rabiner, L. (1969). "Parallel processing techniques for estimating pitch period of speech in the time domain," J. Acoust. Soc. Am. 46, 442–448.

Goldstein, J. L. (1973). "An optimum processor for the central formation of pitch of complex tones," J. Acoust. Soc. Am. 54, 1496–1516.

Goldstein, J. L., Gerson, A., Srulovicz, P., and Furst, M. (1978). "Verification of the optimal probabilistic basis of aural processing of pitch of complex tones," J. Acoust. Soc. Am. 63, 486–497.

Harris, F. J. (1978). "On the use of windows for harmonic analysis with the Discrete Fourier Transform," Proc. IEEE 66, 51–83.

Houtsma, A. J. M., and Goldstein, J. L. (1972). "The central origin of the pitch of complex tones: Evidence from musical interval recognition," J. Acoust. Soc. Am. 51, 520–529.

Johnson, D. H., and Kiang, N. Y. S. (1976). "Analysis of discharges recorded simultaneously from pairs of auditory nerve fibers," Biophys. J. 16, 719–734.

Miller, R. L. (1970). "Performance characteristics of an experimental harmonic identification pitch extraction (HIPEX) system," J. Acoust. Soc. Am. 47, 1593–1601.

McGonegal, C. M., Rabiner, L. R., and Rosenberg, A. E. (1977). "A subjective evaluation of pitch detection methods using LPC synthesized speech," IEEE Trans. Acoust. Speech, Signal Process. 25, 221–229.

Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. 60, 911–918.

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

Plomp, R., and Mimpen, A. M. (1979). "Improving the reliability of testing the speech reception threshold for sentences," Audiology 18, 43–52.

Rabiner, L. R., Cheng, M. J., Rosenberg, A.E., and McGonegal, C. A. (1976). "A comparative performance study of several pitch detection algorithms," IEEE Trans. ASSP 24, 399–418.

Ritsma, R. J. (1962). "Existence region of the tonal residue. I," J. Acoust. Soc. Am. 34, 1224–1229.

Ritsma, R. J. (1963). "On pitch discrimination of residue tones," Int. Audiol. 2, 34–37.

Schouten, J. F. (1940). "The perception of pitch," Philips Tech. Rev. 5, 286–294 [reprinted in *Five Articles on the Perception of Pitch* (IPO, Eindhoven, 1960)].

Schouten, J. F. (1962). "On the perception of sound and speech," 4th Int. Congr. Acoust., Copenhagen, Congr. Rep. II, 195–207.

Schroeder, M. R. (1968). "Period histogram and product spectrum: New methods for fundamental-frequency measurement," J. Acoust. Soc. Am. 43, 829–834.

Seneff, S. (1978). "Real-time harmonic pitch detector," IEEE Trans. Acoust., Speech. Signal Process. 26, 358–365.

Sluyter, R. J., Kotmans, H. J., and Leeuwarden, A. V. (1980). "A novel method for pitch extraction from speech and a hardware model applicable to vocoder systems," Proc. ICASSP 80, 45–48.

Stewart, G. W. (1931). "Problems suggested by an uncertainty principle in acoustics," J. Acoust. Soc. Am. 2, 325–329.

Terhardt, E. (1972). "Zur Tonhöhenwahrnehmung von Klangen I. Psychoakustische Grundlagen," Acustica 26, 173–186; "II Ein Funktionsschema," Acustica 26, 187–199.

Terhardt, E. (1974). "Pitch, consonance, and harmony," J. Acoust. Soc. Am. 55, 1061–1069.

Terhardt, E. (1979). "Calculating virtual pitch," Hear. Res. 1, 155–182.

Vogten, L. L. M., and Willems, L. F. (1977). "The Formator: A speech analysis–synthesis system based on formant extraction from linear prediction coefficients," IPO Annu. Prog. Rep. 12, 47–54.
Whitfield, I. C. (1970). "Neural integration and pitch perception," in *Proceedings of the 5th International Meeting of Neurobiologists*, edited by P. Anderson and S. K. S. Jansen (Universitetsforlaget, Oslo), pp. 277–285.
Wightman, F. L. (1973). "The pattern transformation model of pitch," J. Acoust. Soc. Am. 54, 407–416.
Zwicker, J. J., and Feldtkeller, R. (1967). *Das Ohr als Nachrichtempfanger* (Hirzel, Stuttgart), 2nd ed.