Perceptual linear predictive (PLP) analysis of speech

Hynek Hermansky[a)]
Speech Technology Laboratory, Division of Panasonic Technologies, Inc., 3888 State Street, Santa Barbara, California 93105

(Received 21 August 1989; accepted for publication 27 November 1989)

A new technique for the analysis of speech, the perceptual linear predictive (PLP) technique, is presented and examined. This technique uses three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum: (1) the critical-band spectral resolution, (2) the equal-loudness curve, and (3) the intensity-loudness power law. The auditory spectrum is then approximated by an autoregressive all-pole model. A 5th-order all-pole model is effective in suppressing speaker-dependent details of the auditory spectrum. In comparison with conventional linear predictive (LP) analysis, PLP analysis is more consistent with human hearing. The effective second formant $F2'$ and the 3.5-Bark spectral-peak integration theories of vowel perception are well accounted for. PLP analysis is computationally efficient and yields a low-dimensional representation of speech. These properties are found to be useful in speaker-independent automatic-speech recognition.

PACS numbers: 43.72.Ar, 43.70.Fq, 43.71.Cq, 43.72.Ne

INTRODUCTION

The autoregressive all-pole model $A(\omega)$ of the short-term power spectrum of speech $P(\omega)$, estimated by linear predictive (LP) analysis (see, e.g., Makhoul, 1975), is widely used. The all-pole model can be described in several different parametric spaces. Relatively simple and often computationally efficient transformations between parametric spaces are available (see, e.g., Vishwanathan and Makhoul, 1975). When the order of the model is well chosen, $A(\omega)$ approximates the areas of high-energy concentration in $P(\omega)$ while smoothing out the fine harmonic structure and other less-relevant spectral details. The approximated high-energy spectral areas often correspond to the resonance frequencies of the vocal tract (formants). The LP model assumes the all-pole transfer function of the vocal tract with a specified number of resonances within the analysis band. When this assumption is violated, $P(\omega)$ still approximates the spectral envelope of speech but might be more corrupted by analysis artifacts.

Once we view LP analysis as a means for obtaining the smoothed spectral envelope of $P(\omega)$, we can see that one of the main disadvantages of the LP all-pole model in speech analysis is that $A(\omega)$ approximates $P(\omega)$ equally well at all frequencies of the analysis band. This property is inconsistent with human hearing. Beyond about 800 Hz, the spectral resolution of hearing decreases with frequency. Furthermore, for the amplitude levels typically encountered in conversational speech, hearing is more sensitive in the middle frequency range of the audible spectrum. Consequently, the spectral details of $P(\omega)$ are not always preserved or discarded by LP analysis according to their auditory prominence.

Several techniques have been proposed to alleviate this inconsistency. Itahashi and Yokoyama (1976) warp the spectrum of the high-order LP model into the mel scale and preemphasize it through the equal-loudness curve prior to a second (6th order) LP modeling. Makhoul and Cosell (1976) try several spectral-warping functions on $P(\omega)$ prior to its approximation by $A(\omega)$. Strube (1980) proposes mel-like spectral warping through all-pass filtering in the time domain.

Hermansky (1982) studies a class of spectral transform LP techniques that modify the power spectrum of speech prior to its approximation by the autoregressive model. The current paper adopts this approach to study auditorylike spectral modifications. The all-pole modeling is applied to an auditory spectrum derived by: (a) convolving $P(\omega)$ with a simulated critical-band masking pattern, followed by, (b) resampling the critical-band spectrum at approximately 1-Bark intervals; (c) pre-emphasis by a simulated fixed equal-loudness curve; and (d) compression of the resampled and preemphasized spectrum through the cubic-root nonlinearity, simulating the intensity-loudness power law. The low-order all-pole model of such an auditory spectrum is consistent with several phenomena observed in speech perception. Further, such a model can be employed with advantage in automatic speaker-independent speech recognition.

The paper is organized as follows. Section I describes the implementation details of the method, which we call perceptual linear predictive (PLP) analysis. The second section describes experiments aimed at finding the optimal model order with respect to modeling the linguistic information in speech. Section III shows that a 5th-order PLP analysis is consistent with the sensitivity of human hearing to changes in several important speech parameters. Section IV shows that a 5th-order PLP analysis is consistent with two theories of vowel perception: (a) the effective second-formant theory (Fant and Risberg, 1962) and (b) the 3.5-Bark spectral peak integration theory (Chistovich *et al.*, 1978). Section V discusses some results that support auditory normalization in speech perception. Section VI compares the performance of PLP analysis with conventional LP analysis in speaker-independent digit recognition. Section VII contains the con-

[a)] Current affiliation: US WEST Advanced Technologies, Science and Technology, 6200 S. Quebec Street, Englewood, Colorado 80210.

 0001-4966/90/041738-15$00.80

clusions. FORTRAN 77 code for PLP analysis is given in the Appendix.

I. THE PLP TECHNIQUE

In the PLP technique, several well-known properties of hearing are simulated by practical engineering approximations, and the resulting auditorylike spectrum of speech is approximated by an autoregressive all-pole model. A block diagram of the PLP method is shown in Fig. 1.

A. Spectral analysis

The speech segment is weighted by the Hamming window

$$W(n) = 0.54 + 0.46\cos[2\pi n/(N-1)], \quad (1)$$

where N is the length of the window.

The typical length of the window is about 20 ms.[1] The discrete Fourier transform (DFT) transforms the windowed speech segment into the frequency domain. Typically, the fast Fourier transform (FFT) is used here. For a 10-kHz sampling frequency, a 256-point FFT is needed for transforming the 200 speech samples from the 20-ms window, padded by 56 zero-valued samples.

The real and imaginary components of the short-term speech spectrum are squared and added to get the short-term power spectrum

$$P(\omega) = \mathrm{Re}[S(\omega)]^2 + \mathrm{Im}[S(\omega)]^2. \quad (2)$$

B. Critical-band spectral resolution

The spectrum $P(\omega)$ is warped along its frequency axis ω into the Bark frequency Ω by

$$\Omega(\omega) = 6\ln\{\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5}\}, \quad (3)$$

where ω is the angular frequency in rad/s. This particular Bark–hertz transformation is due to Schroeder (1977).[2] The resulting warped power spectrum is then convolved with the power spectrum of the simulated critical-band (Fletcher, 1940) masking curve $\Psi(\Omega)$. This step is similar to spectral processing in mel cepstral analysis (Bridle and Brown, 1974; Mermelstein, 1976), except for the particular shape of the critical-band curve. In our technique, the critical-band curve is given by

$$\Psi(\Omega) = \begin{cases} 0 & \text{for } \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & \text{for } -1.3 \leqslant \Omega \leqslant -0.5, \\ 1 & \text{for } -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)} & \text{for } 0.5 \leqslant \Omega \leqslant 2.5, \\ 0 & \text{for } \Omega > 2.5. \end{cases} \quad (4)$$

This piece-wise shape for the simulated critical-band masking curve is our approximation to the asymmetric masking curve of Schroeder (1977). It is a rather crude approximation of what is known about the shape of auditory filters. It exploits Zwicker's (1970) proposal that the shape of auditory filters is approximately constant on the Bark scale. The filter skirts are truncated at -40 dB.

The discrete convolution of $\Psi(\Omega)$ with (the even symmetric and periodic function) $P(\omega)$ yields samples of the critical-band power spectrum

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega). \quad (5)$$

The convolution with the relatively broad critical-band masking curves $\Psi(\Omega)$ significantly reduces the spectral resolution of $\Theta(\Omega)$ in comparison with the original $P(\omega)$. This allows for the down-sampling of $\Theta(\Omega)$. In our method, $\Theta(\Omega)$ is sampled in approximately 1-Bark intervals. The exact value of the sampling interval is chosen so that an integral number of spectral samples covers the whole analysis band. Typically, 18 spectral samples of $\Theta[\Omega(\omega)]$ are used to cover the 0–16.9-Bark (0–5-kHz) analysis bandwidth[3] in 0.994-Bark steps.

C. Equal-loudness preemphasis

The sampled $\Theta[\Omega(\omega)]$ is preemphasized by the simulated equal-loudness curve

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)]. \quad (6)$$

The function $E(\omega)$ is an approximation to the nonequal sensitivity of human hearing at different frequencies (Robinson and Dadson, 1956) and simulates the sensitivity of hearing at about the 40-dB level. Our particular approximation is adopted from Makhoul and Cosell (1976) and is given by

$$E(\omega) = [(\omega^2 + 56.8\times10^6)\omega^4]/[(\omega^2 + 6.3\times10^6)^2 \times(\omega^2 + 0.38\times10^9)]. \quad (7)$$

Equation (7) represents a transfer function of a filter with asymptotes of 12 dB/oct between 0 and 400 Hz, 0 dB/oct between 400 and 1200 Hz, 6 dB/oct between 1200 and 3100 Hz, and 0 dB/oct between 3100 Hz and the Nyquist frequency. For moderate sound levels, this approximation is reasonably good up to 5000 Hz. For applications requiring a higher Nyquist frequency, an additional term representing a rather steep (about -18 dB/oct) decrease of the sensitivity of hearing for frequencies higher than 5000 Hz might be found useful. Equation (7) would then become

$$E(\omega) = [(\omega^2 + 56.8\times10^6)\omega^4]/[(\omega^2 + 6.3\times10^6)^2 \times(\omega^2 + 0.38\times10^9)(\omega^6 + 9.58\times10^{26})]. \quad (7')$$

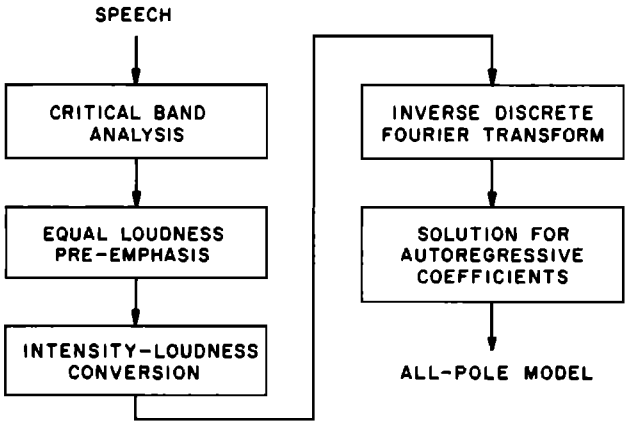

FIG. 1. Block diagram of perceptual linear predictive (PLP) speech analysis.

Finally, the values of the first (0 Bark) and the last (Nyquist frequency) samples (which are not well defined) are made equal to the values of their nearest neighbors. Thus $\Xi[\Omega(\omega)]$ begins and ends with two equal-valued samples.

D. Intensity-loudness power law

The last operation prior to the all-pole modeling is the cubic-root amplitude compression

$$\Phi(\Omega) = \Xi(\Omega)^{0.33}. \tag{8}$$

This operation is an approximation to the power law of hearing (Stevens, 1957) and simulates the nonlinear relation between the intensity of sound and its perceived loudness. Together with the psychophysical equal-loudness preemphasis, this operation also reduces the spectral-amplitude variation of the critical-band spectrum so that the following all-pole modeling can be done by a relatively low model order.

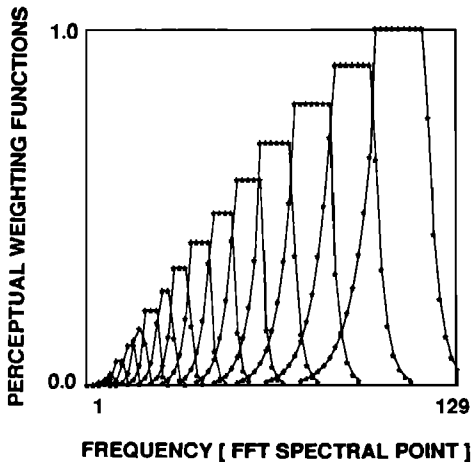

FIG. 2. The 16 weighting functions $w_i(\omega)$ used for computing 16 samples of the auditory spectrum $\Phi(\Omega)$ from the power spectrum $P(\omega)$ of 20-ms frames from speech sampled at 10 kHz.

E. Autoregressive modeling

In the final operation of PLP analysis, $\Phi(\Omega)$ is approximated by the spectrum of an all-pole model using the autocorrelation method of all-pole spectral modeling.[4] Details of the spectral all-pole modeling are sufficiently well described elsewhere (Makhoul, 1975), and we give here only a brief overview of its principle: The inverse DFT (IDFT) is applied to $\Phi(\Omega)$ to yield the autocorrelation function dual to $\Phi(\Omega)$. (Typically, a 34-point IDFT is used.) The IDFT is the better choice here than the inverse FFT, since only a few autocorrelation values are needed. The first $M + 1$ autocorrelation values are used to solve the Yule–Walker equations for the autoregressive coefficients of the M th-order all-pole model. The autoregressive coefficients could be further transformed into some other set of parameters of interest, such as cepstral coefficients of the all-pole model.

F. Practical considerations

In practice, the convolution and the preemphasis are carried out for each sample of $\Xi(\Omega_k)$ in the $P(\omega)$ domain by one weighted spectral summation per spectral sample $\Xi(\Omega_i)$. Thus the spectral sample $\Xi[\Omega(\omega_i)]$ is then given as

$$\Xi[\Omega(\omega_i)] = \sum_{\omega=\omega_{il}}^{\omega_{ih}} w_i(\omega)P(\omega). \tag{9}$$

The limits in the summation and the weighting functions w_i are computed from Eqs. (4), (6), and (10) using the inverse of (3), which is given by

$$\omega = 1200\pi \sinh(\Omega/6). \tag{10}$$

The weighting functions $w_i(\omega)$ are precomputed for the given sampling frequency and current size of the FFT. For illustration, the $w_i(\omega)$ for a 10-kHz sampling frequency are shown in Fig. 2. Some basic properties of the weighting can be seen in the figure. The width of $w_i(\omega)$, i.e., the spectral integration interval, increases with frequency as given by Eq. (3). The $w_i(\omega)$ are flat on the top with exponentially shaped skirts, with low-frequency slopes typically less steep than the high-frequency slopes, as given by Eq. (4) inverted in frequency by convolution and transformed from the Ω into the ω domain. The amplitude of the weighting increases with frequency as given by Eq. (7).

As shown later in Sec. VI, the computational requirements of PLP analysis are comparable to the requirements of conventional LP analysis. Computationally, the most expensive operation is the FFT spectral calculation, followed in cost by the critical-band spectral integration and the cubic-root compression. The cost of the autoregressive modeling is negligible due to the low number of spectral samples of the auditory spectrum to be approximated. A table lookup could be used to compute the root in the intensity-loudness conversion to save on the computational cost.

The FORTRAN 77 code of the subroutine that computes the PLP model of one frame of speech (written with emphasis on simplicity rather than on efficiency) is given in the Appendix.

G. Discussion

The underlying principle of PLP analysis is to approximate the auditory spectrum of speech by an all-pole model. In this section, we have described one computationally reasonably efficient way of obtaining the estimate of the auditory spectrum: convolving the FFT spectrum with the critical-band function, multiplying it by a fixed equal-loudness curve, and compressing its amplitude by a cubic-root function. The engineering approximations to psychophysical laws were our personal choices, often directed in the first place by computational efficiency. We consequently ignored a number of known phenomena, e.g., the dependency of the critical-band shape or the equal-loudness curve on sound intensity. However, our experience suggests that, with respect to our current applications of PLP in speech research, their inclusion would not make a significant difference. Our view is supported by Mason and Gu (1988) who have experimentally observed that the particular way of obtaining the auditory spectrum is not too critical and does not affect the fundamental properties of PLP analysis. Thus, depending on the available hardware and software tools or on the

personal preferences and beliefs of the user, a number of different ways of computing the estimate of the auditory spectrum can be used in PLP analysis.

One of weaker points of the current version of PLP analysis is the dependency of the result on the overall spectral balance of $P(\omega)$ (on formant amplitudes). The spectral balance is easily affected by factors such as the recording equipment, the communication channel or additive noise. The effect of the overall spectral balance can to some extent be suppressed *a posteriori* by a proper distortion measure, as discussed in Sec. II A. Its reduction *a priori* in the analysis is of current research interest.

II. CHOICE OF THE ORDER OF THE AUTOREGRESSIVE PLP MODEL

The choice of the model order specifies the amount of detail in the auditory spectrum that is to be preserved in the spectrum of the PLP model. With increasing model order, the spectrum of the all-pole model asymptotically approaches the auditory spectrum $\Phi(\Omega)$. Thus, for the autoregressive modeling to have any effect at all, the choice of the model order for the given application is critical.

In speech processing, we are often interested in representing the linguistic information in the speech signal. The following series of identification experiments[5] has been designed to determine the PLP model order that would be optimal for this task. The identification experiments resemble standard template-matching speaker-independent automatic speech recognition (ASR) experiments, except that, instead of using templates from a number of speakers sampled from the population of interest, the speech of one speaker is recognized using templates from only a *single* different speaker. Thus any extralinguistic information, e.g., speaker-dependent spectral factors, cannot be used to aid the identification. On the contrary, speaker-dependent factors in cross-speaker identification decrease the accuracy of the identification.

A. Spectral distortion measure for PLP

The group-delay distortion measure (Yegnanarayana and Reddy, 1979) is used in all our identification experiments using PLP analysis. This choice is based on our early ASR experiments that compared the conventional cepstral distortion measure with the group-delay distortion measure (Hermansky *et al.*, 1986). The group-delay measure (quefrency-weighted measure, index-weighted cepstral measure, root-power-sum measure) is implemented by weighting cepstral coefficients of the all-pole PLP model spectrum in the Euclidean distance by a triangular lifter. The cepstral coefficients are computed recursively from autoregressive coefficients of the all-pole model (see, e.g., Vishwanathan and Makhoul, 1975). The triangular liftering (the index weighting of cepstral coefficients) is equivalent to computing a frequency derivative of the cepstrally smoothed phase spectrum (Yegnanarayana and Reddy, 1979). Consequently, the spectral peaks of the model are enhanced and its spectral slope is suppressed (Yegnanarayana, 1977). For a minimum-phase model (such as the PLP all-pole model) computing the Euclidean distance between index-weighted cepstral coefficients of two models is equivalent to evaluating the Euclidean distance between the frequency derivatives of the cepstrally smoothed power spectra of the models. Thus the group-delay distortion measure is closely related to the spectral slope measure proposed by Klatt (1982) for evaluating critical-band spectra.

The group-delay distortion measure is given by

$$d_{\mathrm{GD}} = \sum_{i=1}^{p} i^2 (c_{i_R} - c_{i_T})^2, \tag{11}$$

where c_{i_R} and c_{i_T} are the cepstral coefficients of the reference and the test all-pole models, respectively, and p is the number of cepstral coefficients in the cepstral approximation of the all-pole model spectra. In all experiments with PLP reported in this paper, the number of cepstral coefficients was set to $p = 5$. The zeroth cepstral coefficient, i.e., the logarithmic gain of the model, was excluded.

One important difference between the cepstral and the group-delay measures is illustrated in Fig. 3. This figure shows how the investigated distortion measures reflect the difference in the frequency of one spectral peak in two compared all-pole models. As is evident, the group-delay measure is more sensitive to the actual value of the spectral peak width.

Hermansky and Junqua (1988) proposed that both the cepstral and the group-delay measures are special cases of the general exponential measure

$$d_{\mathrm{GEXP}} = \sum_{i=1}^{p} i^{2S} (c_{i_R} - c_{i_T})^2, \tag{12}$$

where $S \geqslant 0$ is a variable coefficient that allows for various degrees of peak enhancement. This exponential lifter is used for enhancing spectral peaks in the spectrograms shown in Sec. V.

B. Single-frame phoneme identification

Speech from two male and two female adult speakers, reading five repetitions of 104 words, corresponding to the names of typewriter keyboard characters, has been hand labeled at well-identifiable points of each phoneme (the most

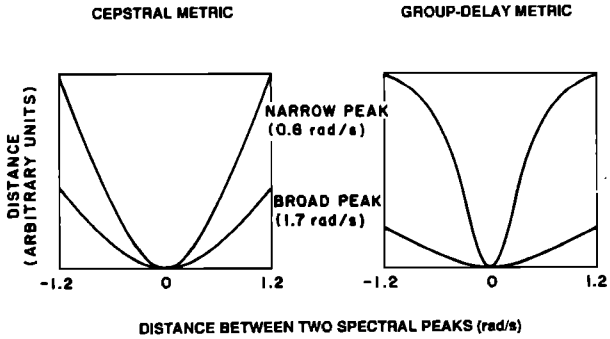

FIG. 3. Cepstral and group-delay distances between two all-pole models that differ in the position of one of their complex poles. The group-delay metric is more sensitive to distance between narrow peaks.

steady parts of sonorants, maxima of energy in stops, centers of transition in diphthongs). The speech was sampled at 10 kHz, and the labeled points were analyzed by PLP analysis. Clusters of analysis vectors with identical phonemic values from each of the speakers were formed. The sizes of the clusters varied from 100 vectors to 5 vectors, reflecting the phonetic balance of the database. The centroids of the clusters were defined and found as averages of each cluster in the autocorrelation domain. Thus each speaker was characterized by 41 phonemelike PLP vectors, representing the 41 phonemes occurring in the database.

The identification has been carried out with the phonemelike vectors of one speaker as reference templates and the phonemelike vectors of another speaker as the test. Note that, in this experiment, the identification relies completely on spectral cues and that no temporal cues are used. All possible combinations of speakers[6] were investigated, yielding 492 comparisons per test. The order of the PLP analysis model was varied from 1–14. Spectral distortions between the test phonemelike vector of one speaker and all the vectors representing the other speaker were computed. The identification was considered correct when the phonemelike vector with the identical phonetic value as the test vector was among the three closest vectors.[7] The percentage of correct choices averaged over all speaker combinations is shown in Fig. 4. For comparison, the result using conventional LP analysis (20-ms Hamming window, preemphasis by first-order 0.98 difference) with the cepstral distortion measure[8] is also shown by the dashed line.

The absolute accuracy of the identification in this difficult classification task is not very high. It is not, however, the absolute accuracy that is of interest here, but rather it is the dependency of this accuracy on the model order. As is evident, the PLP identification accuracy increases up to about the 5th order of the autoregressive model and then starts *decreasing* with further increases in the model order. The LP analysis identification accuracy steadily increases with the model order, but, even for the highest LP model order, it never reaches the accuracy of the 5th-order PLP optimum.

C. Isolated-word identification

A subset of the keyboard database, containing 36 alphanumeric words, was used in some isolated-word identification experiments. The 10-kHz sampled speech was analyzed by PLP analysis with a 10-ms frame advance. Word endpoints were found by hand. A conventional fixed-endpoint dynamic time-warping algorithm was used to compute distances between test and templates. As in the previous experiment, cross-speaker identification has been carried out with the templates of one speaker as reference templates and the templates of another speaker as the test. All possible combinations of speakers were investigated. Since there were $36 \times 5 = 180$ words available from each speaker (in comparison with the 41 phonemelike vectors in the previous experiment), the size of the experiment was significantly larger with 10 900 comparisons per test. In addition, speaker-dependent identification was also carried out, yielding 2880 comparisons per test.

The results of the experiment are shown in Fig. 5. Consistently with the result of the phonemelike identification, the PLP identification accuracy increases up to about the 5th order of the autoregressive model and then starts *decreasing* with further increases in the model order. The LP analysis identification accuracy steadily increases with the model order. However, even for the highest LP model order that we investigate, the accuracy again never reaches the accuracy of the 5th-order PLP.

As is shown in Fig. 6, the accuracy of speaker-dependent identification increases with the model order. For PLP,

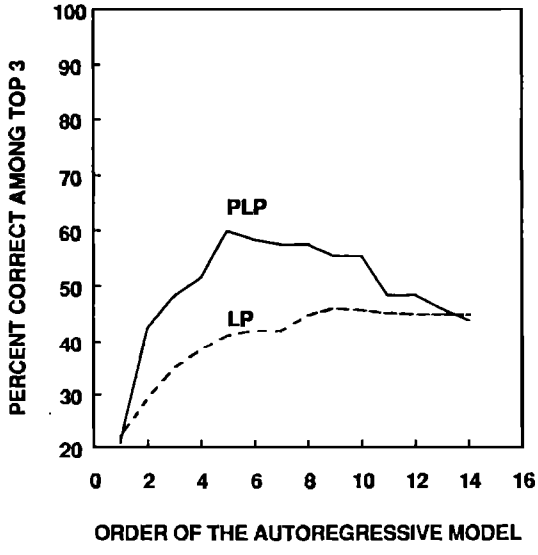

FIG. 4. Average accuracy of identification of an unknown phonemelike all-pole model of speech of one speaker using distances to all phonemelike all-pole models of another speaker. The average is over four speakers. The accuracy is the highest for relatively low-order models.

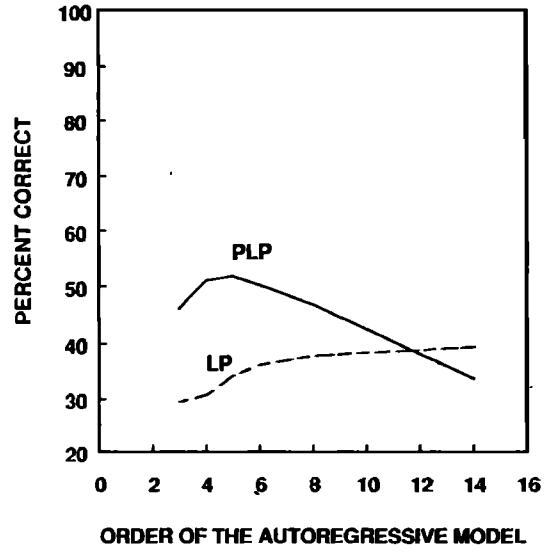

FIG. 5. Average accuracy of identification of an unknown alphanumeric word of one speaker from DTW distance to all alphanumeric words of another speaker (cross-speaker identification) as a function of the order of the autoregressive model in analysis. Average is over four speakers and five productions of the vocabulary by each speaker. As in the results of the phonemelike experiment, shown in Fig. 5, the accuracy is the highest for relatively low-order models.

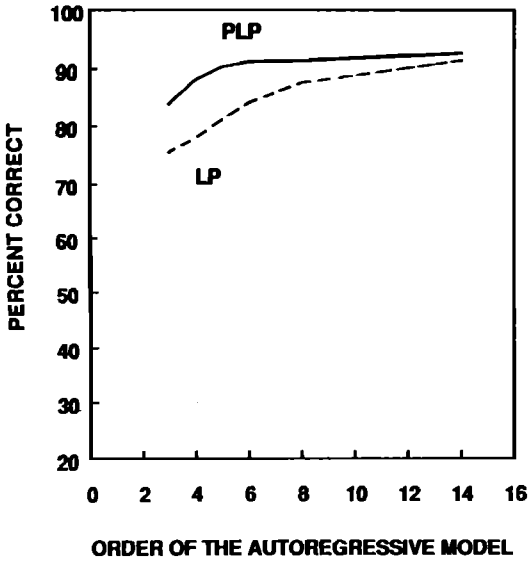

FIG. 6. Average accuracy of speaker-dependent ASR of alphanumeric words of four speakers as a function of the autoregressive model in analysis. For both analysis techniques, the accuracy increases with the model order, rather quickly up to the 6th order for the PLP analysis, relatively steadily for the LP analysis.

the increase is rather sharp up to about the 6th order. LP analysis yields a more gradual increase with the model order, but the accuracy is always lower than for the PLP analysis.

The same experiment was repeated for a *different* set of two male and two female speakers, with essentially the same results as found for the first four speakers (Hermansky, 1987a).

D. Discussion

Both the high-order PLP and the high-order LP are good techniques in the speaker-dependent experiment. In the cross-speaker experiments, the low-order PLP outperforms the conventional LP analysis.

The results of these experiments suggest the following: The formants of speech, approximated by the high-order LP analysis, carry both the linguistic message and speaker-dependent information. In the speaker-dependent experiment, both types of information contribute to the identification. On the other hand, in the cross-speaker experiment, the inclusion of speaker-dependent information decreases the identification accuracy. The advantage of the PLP technique over the conventional LP is that *it allows for the effective suppression of the speaker-dependent information by choosing the particular model order.*

We conjecture that the linguistically relevant speaker-independent cues lie in the gross shape of the auditory spectrum. This gross shape can be characterized by the one or two spectral peaks of the 5th-order PLP model. The finer details of the auditory spectrum, modeled by additional poles of the higher-order PLP models, carry more speaker-dependent information.

III. PLP AND HUMAN HEARING

It is easy to find some instances in which conventional LP analysis clearly contradicts our basic well-accepted knowledge of speech perception, and in which the low-order PLP analysis can bring some improvement. Thus it has been known since the early experiments of Flanagan (1955) that the just noticeable difference in the perception of the first three formant frequencies is approximately constant in relative frequency. This section shows that conventional LP analysis is in conflict with Flanagan's findings. Further, it shows that PLP analysis alleviates this deficiency. It also studies the sensitivity of the PLP and LP models to formant bandwidths, to spectral tilt, and to fundamental frequency.

We used synthetic speech for this study since it allows for exact specification of the acoustic parameters. The analysis and the metric are studied together as one component, the front-end module. The front-end module is presented as a three-port network (Fig. 7), one input of which is connected to a synthesizer whose parameters are held constant, the other to a synthesizer whose parameters are varied. The output of the network is the distance between two synthetic speech signals. If the parameters of both synthesizers are identical, the output distance is zero. In the experiments, the parameters of the variable synthesizer are varied one at a time, and the resulting distance is studied.

We studied two different configurations of the front-end module: (1) 14th-order standard LP analysis with a cepstral metric and (2) 5th-order PLP analysis with a group-delay metric, and we examined the response of the front-end module to changes in: (a) the first three formant frequencies, (b) the first three formant bandwidths, (c) the overall spectral tilt, and (d) the fundamental frequency F_0.

Nine synthetic vowel-like sounds, covering approximately the typical male formant frequency space, were used for the evaluation of the front-end module ($F1 = 300$, 500, 700 Hz; $B1 = 50$ Hz; $F2 = 1000$, 1500, 2000 Hz; $B2 = 90$ Hz; $F3 = 2500$ Hz; $B3 = 120$ Hz; $F4 = 3500$ Hz; $B4 = 150$ Hz; $F5 = 4500$ Hz; $B5 = 180$ Hz).

A. Formant frequency changes

Figure 8(a) shows how the output distance from the front-end module changes when one of the input speech signals changes one of its lower formant frequencies. Averaged results from all nine vowel-like sounds are shown. As pointed out by Kamm and Kahn (1985), in the conventional LP-based front-end module the relative change of the higher

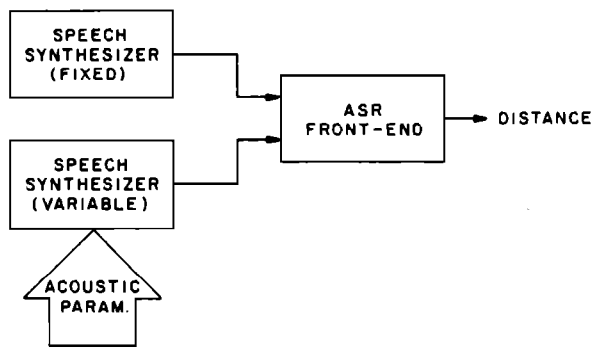

FIG. 7. Three-port network used in evaluating the ASR front ends by synthetic speech.

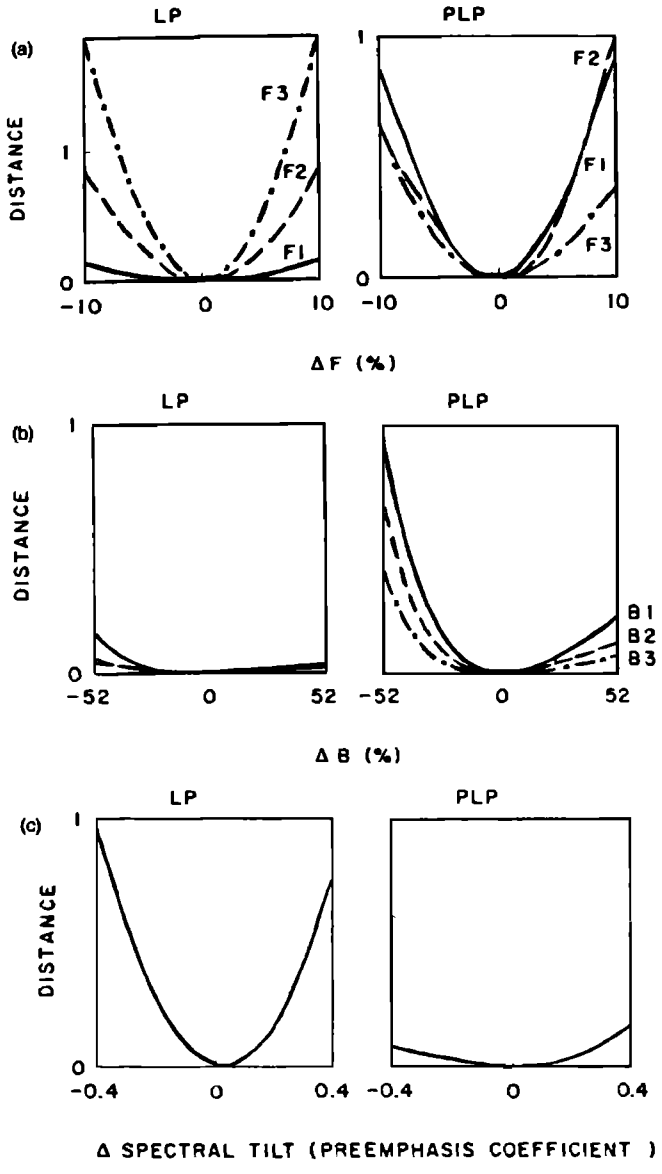

FIG. 8. Sensitivity of ASR front ends to changes in (a) the frequencies of the formants, (b) the bandwidths of the formants, and (c) the spectral tilt of the speech spectrum. The deemphasis coefficient 0.4 yields about −3-dB/oct spectral tilt; the preemphasis coefficient −0.4 results in about 3 dB/oct spectral tilt.

formant frequencies results in larger distances than does relative change of the lower formant frequencies. This is in conflict with Flanagan (1955). By contrast, PLP analysis results in distance curves that are much more similar to one another. This is in agreement with Flanagan (1955). The remedy seems to lie in the Bark frequency resolution, built into the PLP analysis.

Note that the PLP-based front end is sensitive to even very small changes in formant frequency–much smaller than the critical-band spectral resolution. The reason for this sensitivity is that, even when the frequency change occurs within the critical band, it influences the spectral balance of the whole critical-band spectrum and consequently also the shape of the all-pole model spectrum that approximates it.

B. Sensitivity to bandwidth changes

As shown in Fig. 8(b), the LP front end is less sensitive to the formant bandwidth changes than is the PLP front end. The formant bandwidth is usually not considered to be a primary phonetic cue. It is, however, an important cue in some phonetic distinctions (e.g., vowel nasality) and, when disregarded, some otherwise available phonetic information might be lost. The relative decrease of the formant bandwidth yields larger changes in the spectral distance than its relative increase.

C. Sensitivity to spectral tilt

The spectral slope metric of Klatt (1982), which is approximated in the group-delay distance metric, was originally proposed in order to alleviate the sensitivity of the standard spectral metric to the spectral tilt of the speech signal. The spectral tilt is easily influenced by recording conditions or by the glottal characteristics of a particular speaker and is often considered a phonetically nonessential or even disturbing factor.

The spectral tilt in our experiment was varied by preemphasis or deemphasis of the speech signal over approximately a 6-dB/oct range. Figure 8(c) shows that the PLP front-end module, due to its group-delay metric, appreciably suppresses the spectral tilt influence.

D. Sensitivity to F_0

Every automatic speech analysis technique is to a larger or lesser extent sensitive to the fundamental frequency F_0 of the speech signal. The PLP front-end module is more sensitive to F_0 than the LP front-end module. We have chosen for all our above described experiments a fundamental frequency $F_0 = 100$ Hz, i.e., all reference vowels had harmonic peaks coincident with their formant peaks. Had we chosen any other F_0 value, the sensitivity curves would have looked slightly different. As an example, Fig. 9 shows how the formant-frequency sensitivity curves of the PLP front-end

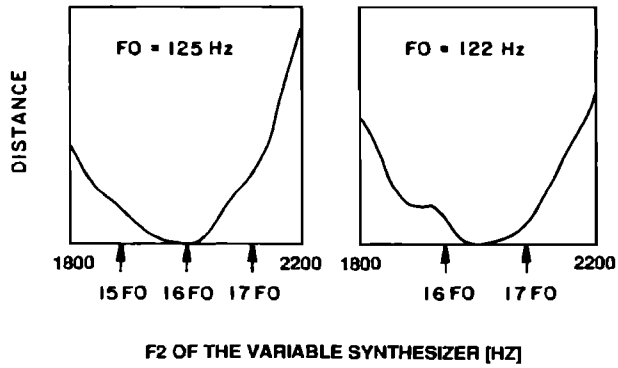

FIG. 9. Sensitivity of ASR front ends to changes in the fundamental frequency F_0 of synthetic voiced speech. Positions of F_0 harmonics are indicated by arrows. Changes in $F2$ have less effect in the vicinity of F_0 harmonics, indicating that the analysis estimates spectral peak shifted towards the nearest harmonic peak.

module change with different values of F_0. The distribution of harmonic peaks in the investigated frequency range is also indicated in the figure. The estimated position of the formant tends to be shifted towards the nearest harmonic peak, and the estimated bandwidth depends on the relative position of the formant and the harmonic peak (Fujisaki and Sato, 1973). This effect is reflected in the spectral distances shown in the figure. When the reference stimulus has its $F2$ coinciding with the 16th harmonic of the F_0 ($F_0 = 125$ Hz), the spectral distance increases monotonically as the $F2$ of the variable stimulus moves away from the $F2$ of the reference. On the other hand, when the reference stimulus has its $F2$ between the 16th and the 17th harmonics of F_0 ($F_0 = 122$ Hz), the movement of the $F2$ of the variable stimulus towards the nearer (16th) harmonic peak results in a rapid change of the spectral distance. In the vicinity of the harmonic peak, the spectral distance change slows down. Once the $F2$ of the variable stimulus escapes the influence of the harmonic peak, the spectral distance again starts increasing monotonically.

E. Discussion

The approximately equal sensitivity of PLP to changes in all the lower formant frequencies is consistent with Flanagan (1955) and represents an improvement over the conventional LP analysis.

Hearing seems to be about 3 (Flanagan, 1957) to 20 times (Carlson *et al.*, 1979) less sensitive to bandwidth changes than to changes in formant frequency. The sensitivity of PLP falls between those two extremes. Also, the experiments of Carlson *et al.* indicate that (consistently with the PLP front end) human hearing is more sensitive to the relative decrease in the bandwidth. However, if the bandwidth sensitivity depends on the formant structure of the vowel as indicated in Klatt's (1982) work, it is difficult to make a straight comparison because of the different vowel spaces used in their works and in the current study.

The low sensitivity of PLP to spectral tilt is consistent with Klatt's (1982) finding of the relative insensitivity of phonetic judgments to the spectral tilt of the stimulus. On the other hand, the spectral tilt is also a strong cue for voicing, often the only one available to current ASR systems. The role of spectral tilt compensation in ASR was studied by Hermansky and Junqua (1988), who showed that a slight increase in the sensitivity of the PLP front-end module to this speech parameter can improve recognition accuracy.

The sensitivity of PLP to F_0 seems to be in principle consistent with human hearing. Recent perceptual experiments (Hermansky, 1987b; Hirahara, 1988) indicate that the perceived formant peak is shifted towards the nearest harmonic peak. Figure 10 shows the results of one of the perceptual experiments (Hermansky, 1987b). Each curve shows how different two synthetic vowel-like sounds with fixed F_0 were perceived to be as the difference between their second formants $F2$ increased. This perceptual experiment is a replication of Flanagan's (1955) experiments aimed at finding just noticeable differences of formant frequencies. Here, the experiments are carried out for different values of

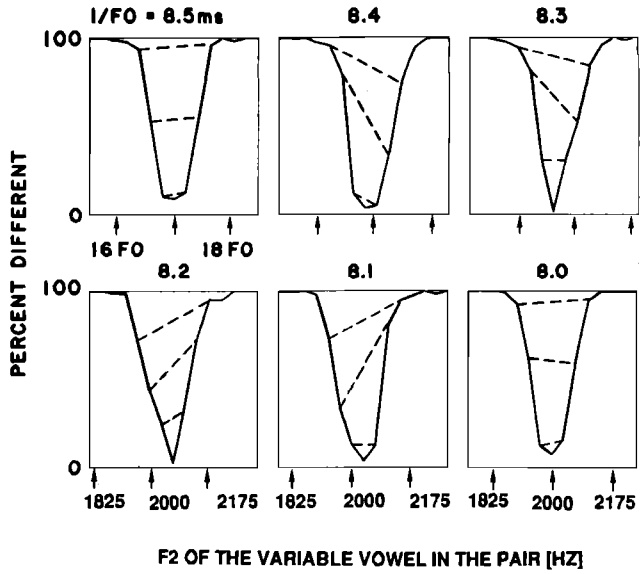

FIG. 10. Sensitivity of human hearing to changes in fundamental frequency F_0 of voiced speech as seen by the difference judgments of vowel-like synthetic pairs with different second-formant frequencies. $F2$ of the reference vowel was fixed at 2000 Hz, $F2$ of the other vowel in the pair varied as indicated on the abscissa. Positions of harmonics of the fundamental frequency F_0 are indicated by arrows. Dashed lines connect points with an equal deviation of $F2$ in the test stimulus from $F2$ in the reference stimulus. Therefore, their slopes reflect the asymmetry due to the F_0 effect. Compare to the sensitivity of the PLP front end shown in Fig. 9. The results are consistent with the notion of the perceived formant peak being shifted towards the nearest F_0 harmonic peak.

F_0 to investigate its effect on the perceived position of $F2$. Even though the difference judgments depicted in Fig. 10 are related to the perceived spectral distance through a (generally unknown) nonlinear psychometric function, the results of the perceptual experiment allow at least for a qualitative comparison with the spectral distance changes shown in Fig. 9. First, we can see that the curves are clearly dependent on F_0. For the symmetric distribution of the harmonic peaks around $F2$ of the reference stimulus, the curve is symmetric. As the relation between $F2$ and harmonics of F_0 becomes asymmetric, the resulting asymmetry of the difference judgment curve is quite obvious. As seen, changes in the curve shape for different values of F_0 are qualitatively consistent with the changes due to F_0 observed in Fig. 9. They are also consistent with similar asymmetries in Flanagan's (1955) curves. Further details of the perceptual experiment can be found in (Hermansky, 1987b).

IV. PLP AND VOWEL PERCEPTION

In the preceding section, we have shown some properties of PLP that are consistent with what we know about human auditory perception. Some of these properties, e.g., the equal importance of relative changes in the lower formant movements, come from modeling the *psychophysical properties of human hearing*, in this case mainly the Bark frequency scale of human hearing. One would expect any analysis technique that employs a similar scale to behave similarly.

Yet, it is perhaps even more interesting to mention some

instances in which the 5th-order PLP is consistent with some not yet fully accepted *concepts of speech perception.* These include the effective second formant $F2'$ concept (Fant and Risberg, 1962) and the 3.5-Bark spectral-peak integration theory (Chistovich *et al.*, 1978). This section shows that the 5th-order PLP is consistent with both concepts.

A. The effective second formant

According to Chiba and Kajiyama (1941), two or one "proper tones" characterize Japanese front and back vowels, respectively. Their "proper tones" do not necessarily coincide with lowest formants. A similar conclusion was reached by Delattre *et al.* (1954) who observed that two spectral peaks are all that are needed for simulating the phonetic qualities of front vowels, and that one spectral peak is sufficient for simulating back vowels. These observations are suggestive of the amount of spectral reduction that human hearing might be performing on the vowel spectrum. In similar two-peak simulations of Swedish vowels, Fant and Risberg (1962) propose that the first spectral peak be kept at the first formant frequency. For the second spectral peak, they use a so called *effective second formant* $F2'$. As already observed by Chiba and Kajiyama and by Delattre *et al.*, $F2'$ does not directly correspond to any particular formant, but is typically close to the second formant in simulating back vowels such as /a/ and is close to the third or even the fourth formant in simulating front vowels such as /i/. Later, Carlson *et al.* (1975), Bladon and Fant (1978), and Paliwal *et al.* (1983) used formulas for deriving $F2'$ as a weighted average of the first four formants. Carlson *et al.* (1975), Itahashi and Yokoyama (1976), Hermansky *et al.* (1985), and Shamma (1988) have observed that auditory models often exhibit a peak close to $F2'$.

We have replicated 18 synthetic cardinal vowels for which Bladon and Fant (1978) give values of $F2'$. A serial synthesizer excited by Rosenberg's (1970) type (c) differentiated glottal pulse with $F_0 = 80$ Hz, a 25% opening time and a 20% closing time was used in the simulation. The output speech was low-pass filtered using a digital filter with a double real pole at $z = 0.6$. Formant frequencies were as given by Bladon and Fant (1978), and formant bandwidths were computed from corrected formulas given by Fant (1972).

Spectra of the synthetic cardinal vowels together with the spectra of their 5th-order PLP models are shown in Fig. 11. As seen in this figure, the front vowels are typically approximated by two spectral peaks in the PLP model, and the back vowels by one spectral peak. Table I shows how the positions of the spectral peaks of the PLP models compare with the first formant $F1$ and the effective second formant $F2'$ of the respective vowels, as determined by Bladon and Fant's perceptual experiments. It gives numerical values in Bark [Eq. (4)] for all the true $F1$, the perceptually estimated $F2'$, and the PLP-estimated $\widehat{F}1'$ and $\widehat{F}2'$, together with the differences between the Bladon and Fant values and the PLP estimates.

As can be seen, the agreement between values from PLP and from Bladon and Fant is rather good. The largest error for the two-peak model is 1.2 Bark for the vowel /ʉ/. It is

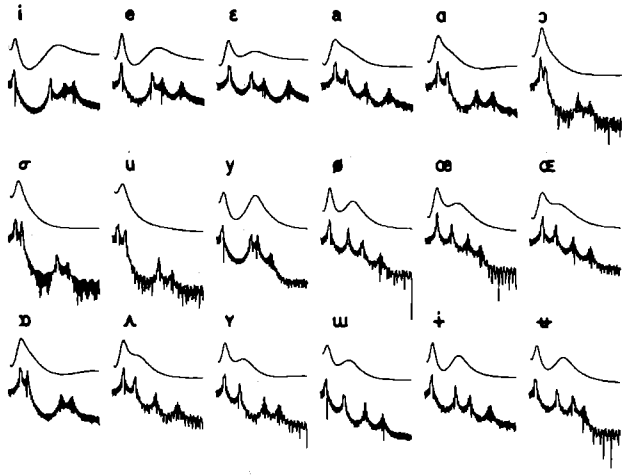

FIG. 11. Spectra of 18 synthetic cardinal vowels and their approximations by spectra of the 5th-order PLP models.

interesting that Bladon (1983) also encountered the most difficulties in predicting $F2'$ for this particular vowel. When the two peaks merge, the position of the single peak is either close to $F1$ (as, e.g., in /a/) or between $F1$ and $F2'$ (as, e.g., in /u/), depending on the spectral balance of the vowel. The results of this experiment verify the findings of Hermansky *et al.* (1985).

B. Spectral peak integration theory

Chistovich (1985) describes a series of experiments which all suggest that, in the perception of speech, two spectral peaks are integrated into one when they are closer than some critical distance $\delta_c \cong 3.5$ Bark.

As is evident from Fig. 11 and Table I, the two peaks of the 5th-order PLP model merge into one when approximating back vowels. This one peak is often positioned between $F1$ and $F2'$. The breakpoint for the existence of two peaks in the PLP spectrum is that the $F1$–$F2'$ distance should be larger than about 3–4 Bark. The merging of spectral peaks by the

TABLE I. Perceptually estimated (Bladon and Fant, 1978) and PLP estimated frequencies of perceptual formants of 18 cardinal vowels.

	Perceptual			PLP		Error	
Vowel	$F1$ (Bark)	$f2'$ (Bark)	$f2'$–$f1$ (Bark)	$F1'$ (Bark)	$F2'$ (Bark)	$F1'$–$F1$ (Bark)	$F2'$–$f2'$ (Bark)
i	2.9	14.1	11.2	3.4	13.3	0.5	−0.8
e	4.3	12.5	8.2	4.3	12.8	0.0	0.3
ɛ	5.8	11.7	5.9	5.3	11.7	−0.5	0.0
a	6.4	9.7	3.3	6.2	merged w/$F1$	−0.2	n/a
ɑ	5.7	8.2	2.5	5.6	merged w/$F1$	−0.1	n/a
ɔ	5.1	6.6	1.5	5.3	merged w/$F1$	0.2	n/a
ɤ	3.5	6.0	2.5	4.8	merged w/$F1$	1.3	n/a
u	2.8	5.8	3.0	4.7	merged w/$F1$	1.9	n/a
y	2.9	11.8	8.9	3.4	11.8	0.5	0.0
ø	4.2	10.1	5.9	4.3	10.7	0.1	0.6
œ	5.6	10.4	4.8	5.4	10.7	−0.2	0.3
oE	6.0	9.7	3.7	5.7	9.7	−0.3	0.0
ɶ	5.8	7.4	1.6	5.9	merged w/$F1$	0.1	n/a
ʌ	5.4	9.0	3.6	5.3	merged w/$F1$	−0.1	n/a
ɣ	4.2	9.2	5.0	4.3	9.7	0.1	0.5
ɯ	2.9	9.1	6.2	3.4	10.0	0.5	0.9
ɨ	3.6	10.8	7.2	3.8	11.0	0.2	0.2
ʉ	3.4	9.9	6.5	3.8	11.1	0.4	1.2

PLP model at about 3.5 Barks has been also reported by Hermansky *et al.* (1985).

C. The significance of the bandwidth $B2'$

Shortly after the $F2'$ concept was proposed, Fujimura (1967) argued against it by demonstrating that pairs of phonetically distinct vowels with identical $F1$ and $F2'$ exist. Such pairs of distinct vowels differ in the spread of the higher formant cluster. Later, Bladon (1983) attempted to alleviate this ambiguity by applying the 3.5-Bark spectral peak integration rule to the $F2$–$F3$–$F4$ cluster. In this section, we show that the 5th-order PLP model alleviates Fujimura's ambiguity.

We have synthesized six vowel pairs from the Bladon–Ladefoged experiment (Bladon and Ladefoged, 1982) using the synthesizer described in the Sec. IV A. Judgments of the perceptual differences between these vowel pairs are given by Bladon (1983). Two of his vowel pairs, /i/–/e/ and /i/–/o/, were used as the calibrating stimuli, the perceptually distinct pairs 1a–1b, 2a–2b, 3a–3b, and 4a–4b are his vowel pairs with identical $F1$ and $F2'$ but with different spreads of the higher formants.

The group-delay distances between the 5th-order PLP models of these vowel pairs are shown in Fig. 12 together with the median perceptual judgments from the Bladon–Ladefoged experiment. The group-delay distances, which represent differences in both the peak positions and the peak bandwidths of the compared models, agree well with the perceptual evaluations. In a similar experiment and with similar results, we have also evaluated Fujimura's (1967) ambiguous vowels.

D. Discussion

Both the $F2'$ concept and the 3.5-Bark spectral peak integration theory are difficult to explain on purely psychophysical grounds and are often affiliated with the so-called "speech mode" of hearing. Similarly, while operations to obtain the auditory spectrum prior the autoregressive modeling in PLP can be justified by psychophysical properties of hearing, the autoregressive modeling itself cannot. It is used to optimize the amount of detail that needs to be eliminated from the auditory spectrum in order to suppress speaker-dependent information.

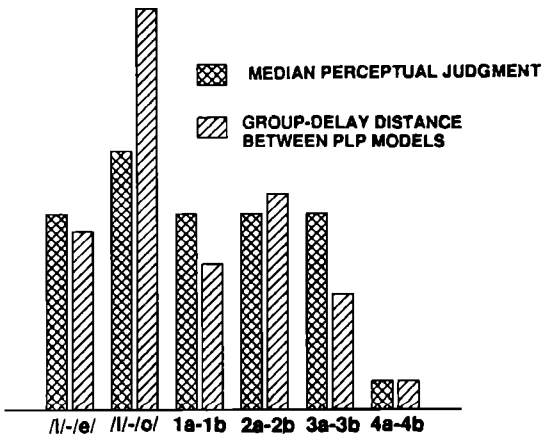

FIG. 12. Median judgments of perceptual differences between pairs of synthetic vowel-like sounds (Bladon, 1983) and the group-delay spectral distances between their 5th-order PLP models. The computed distances are fairly consistent with perceptual judgments.

The 5th-order PLP is consistent with the concepts of $F2'$ and 3.5-Bark integration because of its *combination of the psychophysical transformations and the low-order autoregressive modeling*. The fact that the 5th-order PLP model yields the best results in identifying linguistic information, as shown in Sec. III, seems to support the above-mentioned theories of speech perception.

The autoregressive model further reduces the spectral resolution of the auditory spectrum from its 1-Bark critical-band resolution to about the 3–4-Bark spectral resolution conjectured to take an effect in speech perception (Chistovich, 1985). As shown in Sec. IV A, this does not mean that smaller spectral changes are ignored; it merely means that two spectral peaks within this range are not resolved by hearing as two peaks and that it is rather the overall shape of the auditory spectrum in this range that is used in decoding the linguistic message.

V. AUDITORY NORMALIZATION IN THE PLP METHOD

In spite of spectral differences among speakers, humans can relatively easily decode linguistic messages from the speech of different speakers produced under different circumstances. Although the decoding is most likely happening at several different levels of cognition, some kind of normalization of spectral differences is hypothesized on the auditory level (see, e.g., Bladon and Lindblom, 1981). Our experience with low-order PLP supports this notion.

A. Adult–child speech

It is instructive to compare spectrograms of the speech of an adult male with the speech of the 4-year-old child done by conventional LP and PLP analyses. The LP analysis was of 11th and 15th orders for the child and the adult speech, respectively; the PLP analysis was always of the 5th order. The spectrograms are shown in Fig. 13. Peaks in the spectrograms are enhanced by exponential cepstral lifter with $S = 0.6$ (Hermansky and Junqua, 1988). The LP model typically finds four formants for the adult speech and two or three formants for the child speech. The formants are in different positions for the adult and child speech. The child's $F2$ is tracking the adult male $F3$–$F4$ cluster in /j/; it is between the male $F2$ and $F3$ in /a/, /u/, and /o/. The child's $F1$ tracks the adult male $F1$ in /j/ and /a/ but is between the male's $F1$ and $F2$ in /u/ and the second /o/, and tracks the male $F2$ in the first /o/.

The PLP peak trajectories seem to be more similar across these two speakers. The largest differences are probably in the /u/ and the first /o/ that are represented by a one-peak PLP model in the adult speech and as the two-peak PLP model in the child speech, indicating a more frontlike quality in the child's production of these two vowels. However, these differences might indicate genuine differences in

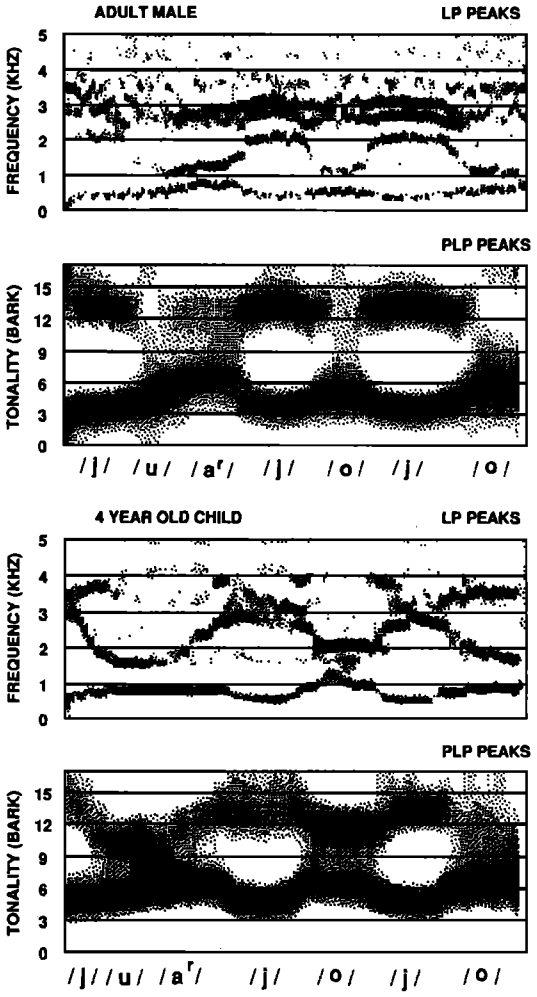

FIG. 13. Peak-enhanced spectrograms from conventional LP and PLP analyses of the utterance "You are yo-yo" uttered by adult male and a 4-year-old child. While for these two speakers LP analysis finds a different number of formants at different positions, PLP analysis finds the same number of peaks at similar positions.

phonetic quality. The greater similarity in the second /o/ of our two speakers supports this interpretation.

B. Voiced-fricative speech

The results of yet another experiment that argues for the existence of auditory normalization are shown in Fig. 14. Here, the voiced and the fricative (speech excited at the point of maximal constriction) productions of the sentence "Where were you a year ago?" were analyzed by 14th-order LP and 5th-order PLP analyses.

The LP analysis result confirms Kuhn (1975). The fricative sentence has only one resonance mode within the 5-kHz analysis band, seen on the LP peak trajectory of the fricative sentence. It is the resonance frequency of the cavity in front of the constriction. The resonance frequency of the front cavity seems to alternate its affiliation with the second and third formants of the voiced sentence.

The 5th-order PLP analysis does not approximate the higher formants directly but extracts their weighted combination, the effective second formant $F2'$. The PLP results

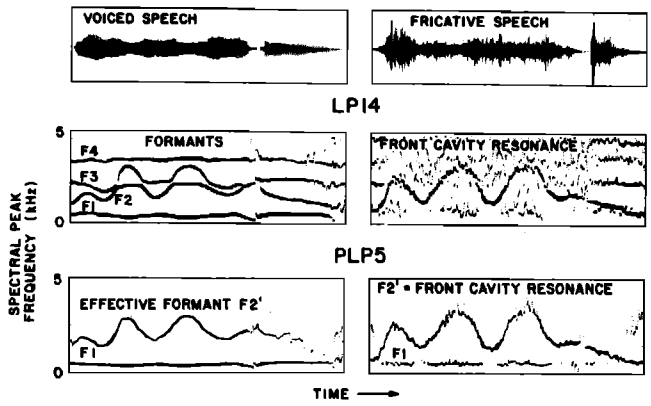

FIG. 14. Peak trajectories of the standard 14th-order LP model and the 5th-order PLP model of voiced and fricative productions of the sentence "Where were you a year ago?" The LP analysis indicates that the single resonance mode of the fricative sentence seems to alternate its affiliation with the formants of the voiced sentence (Kuhn, 1975). PLP analysis estimates the single resonance of the fricative sentence and the effective second formant $F2'$ of the voiced sentence and yields similar results for both sentences.

support Fant (1970) in his conjecture about $F2'$ coinciding with the resonance frequency of the front cavity.

C. Discussion

The main point of this section is that for two utterances with different spectra but with the same linguistic message, LP analysis gives two quite different results, while PLP analysis gives similar results. Thus, to the extent to which PLP analysis simulates the properties of human speech perception, our results support the notion of auditory-level normalization of spectral differences.

The results presented in this section would also be consistent with the hypothesis illustrated in Fig. 15. According to this hypothesis, the speaker-independent linguistic message is coded in the length and shape of the front cavity of the vocal tract. Message-induced changes in the back cavity would then be speaker dependent. Consequently, the formants of speech (which depend on the length and shape of the whole vocal tract) would carry both the linguistic message and the speaker-dependent information. Human speech perception would enhance the message-dependent front-

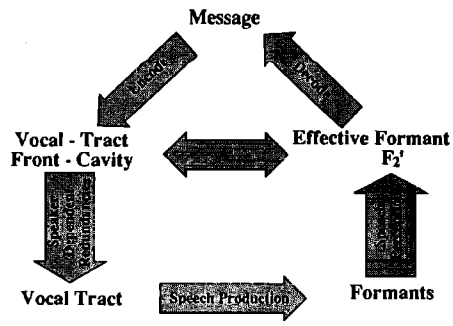

FIG. 15. Hypothesis about the role of $F2'$ and front cavity in speech communication process.

cavity contributions by extracting the front-cavity-affiliated effective second formant $F2'$. This possibility is supported by works of Kuhn (1975, 1978) and is the topic of some current research (Hermansky and Broad, 1989; Broad and Hermansky, 1989).

VI. PLP IN SPEAKER-INDEPENDENT ASR

The topic of the preceding section was how *similar* is the auditorylike representation of utterances with different acoustic qualities but with the *identical* linguistic message. The question remains whether the auditorylike PLP speech representation is *different enough* for the utterances with *different* linguistic messages. Though this question was to some extent addressed in Secs. II–IV, the speaker-independent ASR experiments described in this section further advance this issue.

Current speaker-independent ASR techniques typically handle speaker-to-speaker variations by extensive multispeaker training. A technique that would suppress speaker-dependent information, and would thus allow for reducing the amount of training, is needed. If, as indicated in the preceding sections, PLP could represent the linguistic information in speech better than does the conventional LP technique, it might be used with advantage in speaker-independent ASR.

A. Recognition experiment

To evaluate PLP in speaker-independent ASR, we used a multitemplate, dynamic-time-warping-based recognizer for isolated-digit ASR. The 11 digits (0–9 and the word "oh"), produced once by each of 48 male and 48 female speakers, formed the experimental database. Two front-end modules were used for comparison in the ASR experiments: (1) a 5th-order PLP analysis with the group-delay metric, and (2) a 14th-order autocorrelation LP analysis with the cepstral metric. Both front ends used a 20-ms Hamming window and a 10-ms frame advance. Preemphasis by a first-order 0.98 difference was used with the LP analysis. The word boundaries were hand corrected. Half of the database was used in training; the other half was used in test. No template clustering was used and the recognition was based on the nearest neighbor template. The number of templates per word varied between 2 and 23. In each experiment, 96 different template combinations were used. The examined template combinations were chosen at random. Thus the averaged ASR accuracy is based on more than 77 000 comparisons per test.

Figure 16 shows how the recognition accuracy varies with the number of templates per word. As expected, the recognition accuracy increases with the number of templates per word. The accuracy of the 5th-order PLP front-end module is consistently higher than the accuracy of the conventional LP front-end module.

Figure 17 summarizes the result of another experiment in which 48 different template combinations were compared for several orders of the autoregressive model in both PLP and LP analyses. The accuracies are plotted against the computational requirements. The 5th-order PLP yields the highest recognition accuracy while offering the most substantial computational savings.

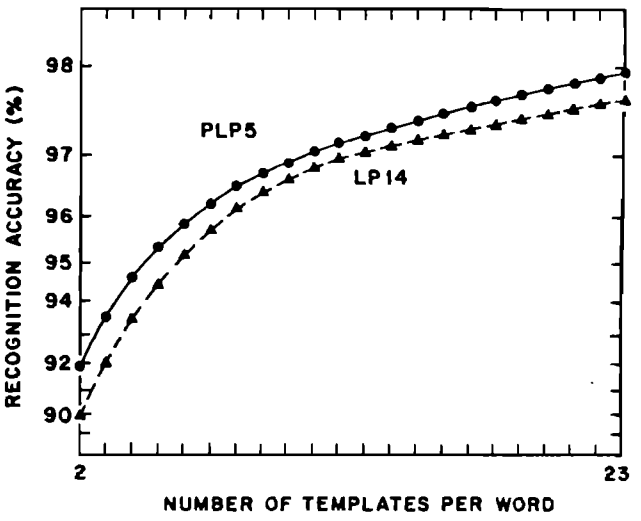

FIG. 16. Comparison of the accuracy of speaker-independent ASR of digits using conventional 14th-order LP analysis and 5th-order PLP analysis. Accuracy is plotted as a function of the number of templates per word. The PLP front end yields consistently higher accuracy.

B. Discussion

Some indications that 5th-order PLP might yield a better speaker-independent ASR front-end module than the conventional LP analysis were already presented in earlier sections of this paper. The cross-speaker identification experiments described in Sec. II indicated that 5th-order PLP is substantially better in suppressing the speaker-dependent cues in speech than the conventional higher-order LP. The improved consistency between low-order PLP analysis and

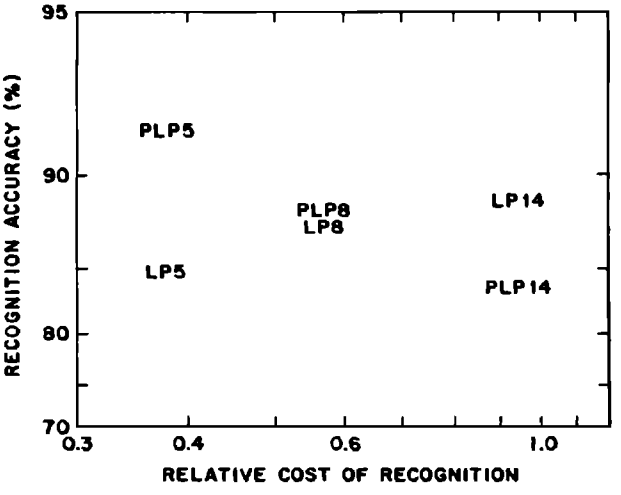

FIG. 17. Comparison of the accuracy and computational cost in speaker-independent ASR of digits using several orders of the autoregressive model in both the conventional LP and the PLP analyses. Two templates per word were used in this experiment. The 5th-order PLP front end gives the highest accuracy while offering substantial computational savings in template matching.

properties of human hearing was demonstrated in Sec. III. We have also described in Sec. IV several phenomena observed in human speech perception that are well represented in the 5th-order PLP analysis. Section V discussed the speaker-normalizing properties of PLP analysis. It was not clear, however, whether all this evidence would relate to real practical problems in speaker-independent ASR. The results of the experiments presented in this section show that the evidence from the previous sections does relate to ASR. PLP yields higher ASR accuracy in speaker-independent ASR than the conventional LP analysis. The improvement is consistent over different databases (Mason and Gu, 1988; Junqua, 1987), different pattern-matching techniques (Applebaum *et al.*, 1987), and different languages (Gu and Mason, 1987). The computational requirements for the PLP analysis itself are comparable to those for conventional LP, as shown in Table II. The use of PLP, however, implies considerably less computation and storage in an ASR system due to the smaller number of PLP parameters.

VII. CONCLUSIONS

A new technique for the analysis of speech, the perceptual linear predictive (PLP) analysis, has been described. The PLP technique uses engineering approximations for three basic concepts from the psychophysics of hearing: (1) the critical-band resolution curves, (2) the equal-loudness curve, and (3) the intensity-loudness power-law relation. In addition, (4) autoregressive modeling is applied to smooth out a certain amount of detail from the auditory spectrum.

In comparison to conventional speech analysis based on the power spectrum, such as LP analysis, even the crude approximations to the very basic and well-known psychophysical knowledge in PLP analysis allow for a different and often more comprehensible picture of the complex speech signal.

We have shown that the 5th order is the optimal order of the autoregressive model in suppressing speaker-dependent information from speech. The 5th-order PLP model is consistent with the human sensitivity to the frequency changes in the first three formant frequencies of speech and is not in conflict with the general tendencies in the human sensitivity to changes in formant bandwidths, in spectral slope, and in F_0. Further, it relates to the $F2'$ and the 3.5-Bark spectral peak integration concepts of speech perception. It expands the $F2'$ concept by modeling the spread of the higher formants, using the $B2'$ parameter in addition to their weighted average $F2'$. Further, we demonstrated some auditory-normalizing properties of the low-order PLP analysis. Finally, we demonstrated advantages of the PLP analysis in speaker-independent ASR.

TABLE II. Approximate cost of LP and PLP analysis of a 200-sample frame of speech (in number of multiplications per frame).

Standard LP14		PLP5	
Pre-emphasis	200	Window	200
Window	200	Fft	2100
Autocorrelation	2800	Critical band	450
Autoregressive model	200	Cubic root[a]	150
		Inverse DFT	30
		Autoregressive model	30
Total	3400	Total	3000

[a] Root through lookup table and interpolation.

ACKNOWLEDGMENTS

I wish to thank Dr. David Broad for numerous discussions and for generous use of his language skills on my behalf. Dr. Brian A. Hanson helped in early stages of implementation of the PLP technique.

APPENDIX

```
      subroutine plp (speech,nwind,m,a,rc,gain,sf)

c     computes 'm'-th order (max 15) PLP model, given by
c     m+1 autoregressive coefficients 'a' or by m reflection
c     coefficients 'rc' and model gain 'gain'
c     from 'nwind' samples (max 512) of speech signal 'speech'
c     with samplig frequency 'sf' (max 20 000 hz)

      real speech(512),a(16),rc(15),r(16),spectr(512),alp(17)
      real hwei(512),cb(900),wcos(23,16),audspe(23)
      integer ibegen(23,3)
      data icall/0/
      pal=4.0*atan(1.0)
      if(icall .eq. 0)then
      call hwind(hwei,nwind)
      nfft=ifix(alog(float(nwind))/0.693148)+1
      npoint=(2**nfft)/2+1
      call audw (npoint,nfilt,cb,ibegen,sf)
      call cost(m,nfilt,wcos)
      icall=1
      endif
      do 10 ii=1,nwind
10    speech(ii)=hwei(ii)*speech(ii)
      call fthr(speech,spectr,nwind,nfft)
      do 20 jfilt=2,nfilt-1
      audspe(jfilt)=0.0
      do 20 kk=ibegen(jfilt,1),ibegen(jfilt,2)
      icb=ibegen(jfilt,3)-ibegen(jfilt,1)+kk
20    audspe(jfilt)=audspe(jfilt)+spectr(kk)*cb(icb)
      do 25 ii=2,nfilt-1
25    audspe(ii)=audspe(ii)**0.33
      audspe(1)=audspe(2)
      audspe(nfilt)=audspe(nfilt-1)
      nspt=2*(nfilt-1)
      do 40 kk=1,m+1
      r(kk)=audspe(1)
      do 30 ii=2,nfilt
30    r(kk)=r(kk)+audspe(ii)*wcos(ii,kk)
40    r(kk)=r(kk)/nspt
      a(1)=1.0
      alp(1)=r(1)
      rc(1)=-r(2)/r(1)
      a(2)=rc(1)
      alp(2)=r(1)+r(2)*rc(1)
      do 50 mct=2,m
      s=0.
      mct2=mct+2
      alpmin=alp(mct)
      do 60 ip=1,mct
      idx=mct2-ip
60    s=s+r(idx)*a(ip)
      rcmct=-s/alpmin
      mh=mct/2+1
      do 70 ip=2,mh
      ib=mct2-ip
      aip=a(ip)
      aib=a(ib)
      a(ip)=aip+rcmct*aib
70    a(ib)=aib+rcmct*aip
      a(mct+1)=rcmct
      alp(mct+1)=alpmin-alpmin*rcmct*rcmct
50    rc(mct)=rcmct
      gain=alp(m+1)
      return
      end
```

```
      subroutine audw (npoint,nfilt,cb,ibegen,sf)

c     computes auditory weighting functions

      real cb(900)
      integer ipoint(23),ibegen(23,3)
      fnqbar=6.0*alog((sf/1200.0)+sqrt((sf/1200.0)**2+1.0))
      nfilt=ifix(fnqbar)+2
      f2samp=float(npoint-1)/(sf/2.0)
      zdel=fnqbar/float(nfilt-1)
      icount=1
      do 10 j=2,nfilt-1
      ibegen(j,3)=icount
      z0=zdel*float(j-1)
      f0=600.0*(exp(z0/6)-exp(-z0/6))/2.0
      fl=600.0*(exp((z0-2.5)/6)-exp(-(z0-2.5)/6))/2.0
      ibegen(j,1)=nint(fl*f2samp)+1
      if(ibegen(j,1).lt.1)ibegen(j,1)=1
      fh=600.0*(exp((z0+1.3)/6)-exp(-(z0+1.3)/6))/2.0
      ibegen(j,2)=nint(fh*f2samp)+1
      if(ibegen(j,2).gt.npoint)ibegen(j,2)=npoint
      do 10 i=ibegen(j,1),ibegen(j,2)
      freq=float(i-1)/f2samp
      x=freq/600.
      z=6.0*alog(x+sqrt(x**2+1.0))
      z=z-z0
      if (z.le.-0.5) then
      cb(icount)=10**(z+0.5)
      else if (z.ge.0.5) then
      cb(icount)=10**(-2.5*(z-0.5))
      else
      cb(icount)=1.0
      endif
      fsq=f0**2
      rsss=(fsq**2)*(fsq+1200.0**2)/(((fsq+400.0**2)**2)*(fsq+3100.0**2))
      cb(icount)=rsss*cb(icount)
      icount=icount+1
10    continue
      return
      end

      subroutine hwind (weight,npoint)

c     computes Hamming window weighting

      real weight(512)
      pai=4.0*atan(1.0)
      do 1 ii=1,npoint
1     weight(ii)=0.54-0.46*cos(2.0*pai*(ii-1)/(npoint-1))
      return
      end

      subroutine cosf(m,nfilt,wcos)

c     computes cosine weightings for IDFT

      dimension wcos(23,16)
      pai=4.0*atan(1.0)
      do 2 ii=1,m+1
      do 1 jj=2,(nfilt-1)
1     wcos(jj,ii)=2.0*cos(2.0*pai*(ii-1)*(jj-1)/(2*(nfilt-1)))
2     wcos(nfilt,ii)=cos(2.0*pai*(ii-1)*(jj-1)/(2*(nfilt-1)))
      return
      end
```

```
      subroutine ffthr(real,power,ll,m)

c     fft subroutine for computing the power spectrum

      dimension signal(512),sigima(512),power(512),real(512)
      k=0
      pai=4.0*atan(1.0)
      N=2**M
      do 30 ii=1,ll
      sigima(ii)=0.0
30    signal(ii)=real(ii)
      do 40 ii=ll+1, n
      sigima(ii)=0.0
40    signal(ii)=0.0
      n2=n/2
      n1=n-1
      j=1
      do 3 i=1,n1
      if(i.ge.j) go to 1
      t1=signal(j)
      t2=sigima(j)
      signal(j)=signal(i)
      sigima(j)=sigima(i)
      signal(i)=t1
      sigima(i)=t2
1     k1=n2
2     if(k1.ge.j)go to 3
      j=j-k1
      k1=k1/2
      go to 2
3     j=j+k1
      do 6 l=1,m
      le=2**l
      le1=le/2
      u1=1.0
      u2=0.0
      w1=cos(pai/(float(le1)))
      w2=(-sin(pai/(float(le1))))
      do 6 j=1,le1
      do 4 i=j,n,le
      id=i+le1
      t1=signal(id)*u1-sigima(id)*u2
      t2=sigima(id)*u1+signal(id)*u2
      signal(id)=signal(i)-t1
      sigima(id)=sigima(i)-t2
      signal(i)=signal(i)+t1
4     sigima(i)=sigima(i)+t2
      u3=u1
      if(k)10,9,10
9     u1=u1*w1-u2*w2
      u2=u2*w1+u3*w2
      go to 6
10    u4=w1*w1+w2*w2
      u1=(u1*w1+u2*w2)/u4
      u2=(u2*w1-u3*w2)/u4
6     continue
      if(k.eq.0)go to 12
      do 8 i=1,n
      sigima(i)=sigima(i)/n
8     signal(i)=signal(i)/n
12    continue
      do 50 ii = 1,n
50    power(ii)=signal(ii)*signal(ii)+sigima(ii)*sigima(ii)
      return
      end
```

[1] Parameter values given as typical in this section are used for the PLP analyses in all the experiments reported in this paper.

[2] While almost any function that compresses upper end of $P(\omega)$ is an improvement over the linear scale of the conventional LP analysis, yet more accurate approximation of the nonlinear frequency scale of hearing might be desirable here as our understanding of the frequency selectivity of human hearing advances.

[3] The upper limit of our analysis band was given by our speech database, sampled at 10 kHz. There is nothing that would prevent the user from choosing a different upper limit [with a possible correction of Eq. (7), as described in the next section]. One of the advantages of an analysis based on the critical bands is that an analysis bandwidth increase can often be done with only a moderate increase in the computational requirements.

[4] The sampling of $B(\Omega)$ results in aliasing of the related autocorrelation function (El Jaroudi and Makhoul, 1987). If $A(\Omega)$ contained sudden spectral changes, the higher autocorrelation values might have been significant. In this case, a higher sampling rate on $B(\Omega)$ or a computationally more expensive all-pole modeling method (such as spectral envelope interpolation LP (Hermansky *et al.*, 1984) or discrete all-pole modeling (El Jaroudi and Makhoul; 1987) might have been necessary. Fortunately, in speech analysis, $A(\Omega)$ is reasonably smooth and has relatively low spectral spread. We have experimentally observed that any further upsampling of $A(\Omega)$ above the 1-Bark rate does not yield any significant change in the resulting all-pole model.

[5] We would like to avoid the possible confusion of such experiments, which are aimed at evaluating the speaker-normalizing properties of the analysis and not necessarily at the highest ASR accuracy, with practical multitemplate, speaker-independent ASR.

[6] The multiple use of the database is not equivalent to the use of a larger database. The result is valid only within the limited speaker population on which it was obtained. However, the mean result is much less sensitive to the particular choice of templates than the result of only one template choice would be.

[7] The absolute number of correct choices is lower when only one nearest choice is considered and it increases when more than three choices are evaluated. However, the shape of the curve indicating the dependence of the percentage of correct choices on the order of the model and its maximum remains basically identical.

[8] In the current paper, we report only on the cepstral distortion measure with the conventional LP analysis. In speaker-independent ASR, the group-delay distortion measure with the conventional LP analysis yields consistently inferior results (Applebaum *et al.*, 1987; Hermansky, 1987).

Applebaum, T. H., Hanson, B. A., and Wakita, H. (1987). "Weighted cep-

stral distance measures in vector quantization based speech recognizers," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 87*, Paper 27.9, pp. 1155–1158.

Bladon, A. (1983). "Two-formant models of vowel perception: shortcomings and enhancements," Speech Commun. **2**, 305–313.

Bladon, A., and Fant, G., (1978). "A two-formant model and the cardinal vowels," STL-QPRS **1**, 1–8, Royal Institute of Technology, Stockholm.

Bladon, A., and Ladefoged, P. (1982). "A further test of a two-formant model," J. Acoust. Soc. Am. Suppl. 1 **71**, S104.

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

Bridle, J. S., and Brown, M. D. (1974). "An experimental automatic word recognition system," JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England.

Broad, D. J., and Hermansky, H. (1989). The front-cavity/*F2'* hypothesis tested by data on tongue movements," J. Acoust. Soc. Am. Suppl. 1 **86**, S13–S14.

Carlson, R., Granstrom, B., and Fant, G. (1970). "Some studies concerning perception of isolated vowels," STL-QPRS **2-3**, 19–35, Royal Institute of Technology, Stockholm.

Carlson, R., Fant, G., and Granstrom, B. (1975). "Two-formant models, pitch and vowel perception," in *Auditory Analysis and Perception of Speech*, edited by G. S. Fant and M. A. A. Tatham (Academic, New York), pp. 55–82.

Carlson, R., Granstrom, B., and Klatt, D. (1979). "Vowel perception: the relative perceptual salience of selected acoustic manipulations," STL-QPSR **3-4**, 73–83, Royal Institute of Technology, Stockholm.

Chiba, T., and Kajiyama, M. (1941). *The Vowel: Its Nature and Structure* (Tokyo Kaiseikan, Tokyo).

Chistovich, L. A., Sheikin, R. L., and Lublinskaja, V. V. (1978). " 'Centers of gravity' and spectral peaks as the determinants of vowel quality," in *Frontiers of Speech Communication Research*, edited by B. Linblom and S. Ohman (Academic, New York), pp. 143–157.

Chistovich, L. A. (1985). "Central auditory processing of peripheral vowel spectra," J. Acoust. Soc. Am. **77**, 789–805.

Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (1952). "An experimental study of the acoustic determinants of vowel color," Word **8**, 195–210.

El Jaroudi, A., and Makhoul, J. (1987). "Discrete all-pole modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 87*, pp. 320–323.

Fant, G., and Risberg, A. (1962). "Auditory matching of vowels with two formant synthetic sounds," STL-QPRS **4**, 7–11, Royal Institute of Technology, Stockholm.

Fant, G. (1970). *Acoustic Theory of Speech Production* (Mouton, The Hague), 2nd printing, p. 123.

Fant, G. (1972). "Vocal tract wall effects, losses and resonance bandwidths," STL-QPRS **2-3**, 28–52, Royal Institute of Technology, Stockholm.

Flanagan, J. (1955). "Difference limen for vowel formant frequency," J. Acoust. Soc. Am. **27**, 613–617.

Flanagan, J. (1957). "Estimates of maximum precision necessary in quantizing certain dimensions of vowel sounds," J. Acoust. Soc. Am. **29**, 533–534.

Fletcher, H. (1940). "Auditory patterns," Rev. Mod. Phys. **12**, 47–65.

Fujimura, O. (1967). "On the second spectral peak of front vowels: a perceptual study of the role of the second and third formants," Lang. Speech, **10**, 181–193.

Fujisaki, H., and Sato, Y. (1973). "Comparison of errors in formant frequencies obtained by various methods of formant extraction," Trans. Comm. Speech Res., Acoust. Soc. Japan, December 1973 (in Japanese).

Gu, Y., and Mason, J. S. D. (1987). "Vocal tract and auditory feature analysis using Chinese utterance in ASR system," in *Proceedings of International Conference on Chinese Information Processing*, Beijung, China.

Hermansky, H. (1982). "Improved linear predictive analysis of speech based on spectral processing," Ph.D. dissertation, University of Tokyo.

Hermansky, H., Fujisaki, H., and Sato, Y. (1984). "Spectral envelope sampling and interpolation in linear predictive analysis of speech" in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 84*, pp. 221–224.

Hermansky, H., Hanson, B. A., and Wakita, H. (1985). "Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain," Speech Commun. **4**, (1–3), 181–187.

Hermansky, H. (1987a). "An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory perception," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 87*, pp. 1159–1162.

Hermansky, H. (1987b). "Why is the formant frequency DL curve asymmetric?, " J. Acoust. Soc. Am. Suppl. 1 **81**, S18; full text in STL Research Reports No. 1, Santa Barbara, 1987.

Hermansky, H., and Junqua, J. C. (1988). "Optimization of perceptually based ASR front-end," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 88*, Paper S5.10, pp. 219–222.

Hermansky, H., and Broad, D. J. (1989). "The effective second formant *F2'* and the vocal tract front cavity," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 89*, Paper S10a.4, pp. 480–483.

Hirahara, T. (1988). "On the role of fundamental frequency in vowel perception," J. Acoust. Soc. Am. Suppl. 1 **84**, S156.

Itahashi, S., and Yokoyama, S. (1976). "Automatic formant extraction utilizing mel scale and equal loudness contour," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 76*, Paper 9.2, pp. 310–313.

Junqua, J. C. (1987). "Evaluation of ASR front-ends in speaker dependent and speaker independent recognition," J. Acoust. Soc. Am. Suppl. 1 **81**, S93; Full text in STL Research Reports No. 1, Santa Barbara, 1987.

Kamm, C., and Kahn, D. (1985). "Relationship between LP residual spectral distances and phonetic judgement," J. Acoust. Soc. Am. Suppl. 1 **78**, S82.

Kuhn, G. M. (1975). "On the front cavity resonance and its possible role in speech perception," J. Acoust. Soc. Am. **58**, 428–433.

Kuhn, G. M. (1978). "Stop consonant place perception with single-formant stimuli: Evidence for the role of the front-cavity resonance.", J. Acoust. Soc. Am. **65**, 774–788.

Klatt, D. (1982). "Prediction of perceived phonetic distance from critical-band spectra: a first step," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 82*, pp. 1278–1281.

Makhoul, J. (1975). "Spectral linear prediction: properties and applications," IEEE Trans. **ASSP-23**, 283–296.

Makhoul, J., and Cosell, L. (1976). "LPCW: An LPC vocoder with linear predictive spectral warping," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 76*, pp. 466–469, Philadelphia.

Mason, J. S., and Gu, Y. (1988). "Perceptually-based features in ASR," in *Proceedings of IEEE Colloquium on Speech Processing*, London.

Mermelstein, P. (1976). "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, edited by C. H. Chen (Academic, New York), pp. 374–388.

Paliwal, K. K., Lindsay, D., and Ainsworth, W. A. (1983). "A study of two-formant models for vowel identification," Speech Commun. **2 (4)**, 295–303.

Robinson, D. W., and Dadson, R. S. (1956). "A redetermination of the equal-loudness relations for pure tones," Br. J. Appl. Phys. **7**, 166–181.

Rosenberg, A. (1970). "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. **49**, 583–590.

Schroeder, M. R. (1977). *Recognition of Complex Acoustic Signals, Life Sciences Research Report 5*, edited by T. H. Bullock (Abakon Verlag, Berlin), p. 324.

Shamma, S. A. (1988). "The acoustic features of speech phonemes in a model of auditory processing: Vowels and unvoiced fricatives," J. Phon. **16**, 79–91.

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

Strube, H. W. (1980). "Linear prediction on a warped frequency scale," J. Acoust. Soc. Am. **68**, 1071–1076.

Vishwanathan, R., and Makhoul, J. (1975). "Quantization properties of transmission parameters in linear predictive systems," IEEE Trans. **ASSP-26**, 587–596.

Yegnanarayana, B. (1977). "Formant extraction from linear prediction phase spectra," J. Acoust. Soc. Am. **63**, 1638–1640.

Yegnanarayana, B., and Reddy, R. (1979). "A distance measure derived from the first derivative of the linear prediction phase spectra," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 79*, pp. 744–747.

Zwicker, E. (1970). "Masking and psychological excitation as consequences of ear's frequency analysis," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leyden, The Netherlands).