9.9, 9.3

Received 17 July 1969

System for Automatic Formant Analysis of Voiced Speech

RONALD W. SCHAFER AND LAWRENCE R. RABINER

Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey 07974

A system for automatically estimating the lowest three formants and the pitch period of voiced speech is presented. The system is based on a digital computation of the cepstrum (defined as the inverse transform of the log magnitude of the z-transform). The pitch period estimate and smoothed log magnitude are obtained from the cepstrum. Formants are estimated from the smoothed spectral envelope using constraints on formant frequency ranges and relative levels of spectral peaks at the formant frequencies. These constraints allow the detection of cases where two formants are too close together in frequency to be resolved in the initial spectral envelope. In these cases, a new spectral analysis algorithm (the chirp z-transform algorithm) allows the efficient computation of a narrow-band spectrum in which the formant resolution is enhanced. Formant and pitch period data obtained by the analysis system are used to control a digital formant synthesizer. Results, in the form of spectrograms, are presented to illustrate the performance of the system.

INTRODUCTION

The acoustic theory of speech production[1] is the basis for most speech analysis–synthesis systems. The essence of this theory is that the speech waveform can be modeled as the output of a lumped parameter, linear, quasi-time-invariant system in response to an excitation which is either random noise (unvoiced sounds), a quasi-periodic pulse train (voiced sounds), or in some cases, a mixture of these sources (voiced fricatives). The implication of the model is that speech can be produced by proper excitation of a series or parallel connection of resonators whose complex natural frequencies vary slowly and continuously with time so as to approximate the time-varying eigenfrequencies of the vocal tract. These frequencies are called the formant frequencies, or simply formants.

This paper describes a system for automatically estimating pitch period and the lowest three formants of (nonnasal) voiced speech. By restricting the class of speech sounds to be analyzed, we have been able to develop a fully automatic system for estimating the desired parameters. It should be noted that the system presented here can be viewed as part of a larger system that would operate on a wider class of speech sounds. Such a system would incorporate a scheme for classifying speech sounds into several classes, each of which would then be analyzed by a different system.

The techniques presented in this paper may find application in a wide range of speech-processing problems. For example, if fully automatic analysis techniques can be developed for the other classes of speech sounds such as voiced and unvoiced fricatives, stops, and nasals; these can be combined into a formant–vocoder system. More immediate applications are foreseen in the areas of speaker recognition and speech recognition. Another area of application, which in a sense was the original motivation for studying the problem of automatic formant estimation, is as a tool in the evaluation and design of systems using synthetic speech as an output—such as a computer voice-response device. For example, in development of a system for speech synthesis-by-rule, automatic means for estimating formants and pitch could be used to analyze the utterances of a cross section of speakers each speaking a given utterance. These data could be compared to the formants and pitch generated by the synthesis-by-rule system. Such comparisons would provide a means for designing new synthesis strategies and for improving existing ones.

In the remainder of this section, some of the details of the model for production of voiced speech are reviewed, and the relationship of our work to previous research on the estimation of formant frequencies is discussed. The remaining sections discuss details of the approach and results that have been obtained.

[1] C. G. M. Fant, *Acoustic Theory of Speech Production* (Mouton and Co., 's-Gravenhage, The Netherlands, 1960).

aded 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/te

A. Model for Voiced-Speech Waveforms

The model upon which the analysis is based is shown in Fig. 1(a). The samples of the speech waveform (10-kHz sampling rate) are viewed as the output of the discrete-time system of Fig. 1(a). Since a digital computer is used to perform both analysis and synthesis, it is preferable to think in terms of a discrete-time model and discrete-time-analysis techniques rather than the more familiar analog model and corresponding analysis methods.

In Fig. 1(a), the box labeled "Impulse Train Generator" provides a train of unit samples whose spacing is τ, the pitch period, which is a function of time. The multiplier A is a gain control, which varies with time and controls the intensity of the output. The cascade of networks $G(z)R(z)$ is an approximation to the combination of the glottal-source spectrum and the radiation load spectrum. This cascade is a discrete linear time-invariant network of the form

$$G(z)R(z)=\left(\frac{1-e^{-aT}}{1-z^{-1}e^{-aT}}\right)\left(\frac{1+e^{-bT}}{1+z^{-1}e^{-bT}}\right), \qquad (1)$$

where a and b are constants that characterize the speaker (and possibly the utterance). Representative values of a and b are 400π and 5000π, respectively. More accurate values for a given speaker may be determined from a long-term average spectrum for that speaker. For the work presented in this paper, fixed values for a and b have been used throughout. Figure 2 shows a plot of the spectrum for the network $G(z)R(z)$ using these values for a and b.

The system labeled $V(z)$ in Fig. 1(a) is a linear quasi-time-invariant discrete system which consists of a cascade connection of digital resonators as shown in Fig. 1(b). That is,

$$V(z)=\prod_{k=1}^{4} H_k(z), \qquad (2)$$

where

$$H_k(z)=\frac{(1-2e^{-\alpha_kT}\cos\omega_kT+e^{-2\alpha_kT})}{[1-(2e^{-\alpha_kT}\cos\omega_kT)z^{-1}+e^{-2\alpha_kT}z^{-2}]} \qquad (3)$$

and $\omega_k=2\pi F_k$, and $T=0.0001$ sec.

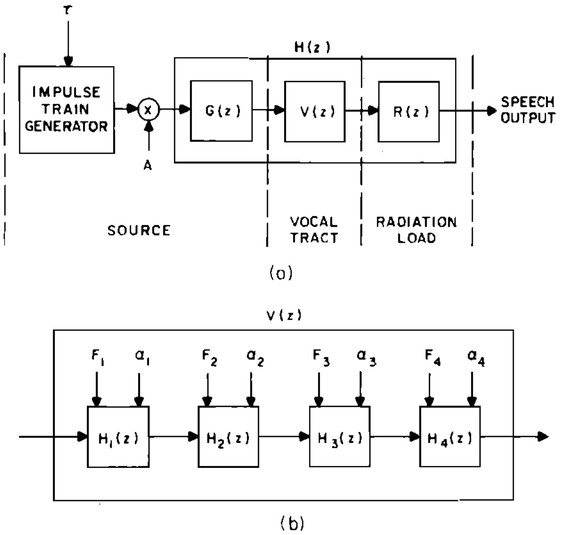

FIG. 1. (a) Digital model for voiced speech. (b) Detailed diagram of model for vocal tract transmission.

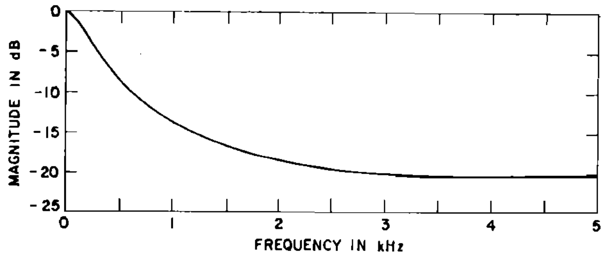

FIG. 2. Evaluation of the fixed system function $G(z)R(z)$ on the unit circle.

The purpose of the system $V(z)$ is to model the vocal tract transmission characteristics. In synthesizing speech or in modeling the speech waveform, the formant frequencies F_k and bandwidths $2\alpha_k$ vary with time, corresponding to the fact that the vocal tract changes shape as different sounds are produced. Since these parameters vary slowly with time, the spectrum of a short segment of speech is characterized, to a good approximation, by Eqs. 1–3 with fixed values of F_k and α_k. Over long time intervals, these equations should be interpreted as specifying the sequence of numerical operations required to compute the samples of an approximation to the speech waveform. Because the input speech is filtered sharply to contain no frequencies above 4 kHz, it is only necessary for $V(z)$ to consist of three time-varying digital resonators and a fourth fixed resonator to ensure proper spectral balance at high frequencies.

According to the acoustic theory,[1] another pole and a zero in series with the configuration of Fig. 1(a) are required to represent nasal consonants. In the system described in this paper, no provision is made for estimating this pole and zero. Preliminary synthesis results indicate that this extra pole and zero may not be necessary for acceptable synthetic speech.

The analysis of the speech waveform involves the determination of the time-varying parameters of the model: i.e., the formant frequencies F_k, the source-radiation parameters a and b, the formant bandwidths $2\alpha_k$, the pitch period τ, and the gain A. The analysis system discussed here determines, as a function of time, only the lowest three formants, the pitch period and the gain. In resynthesizing the speech for comparison with the original speech, F_4, a, b, and all the α_k's are held fixed.

d 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/in

It has been shown[2] that the model of Fig. 1(a) is at least as effective in approximating the spectral properties of speech signals as the more conventional continuous-time models that employ higher pole-correction networks. Figure 3 shows the results of evaluating the z-transform

$$H(z)=G(z)V(z)R(z) \tag{4}$$

on the unit circle, i.e., for $z=e^{j\omega T}$, where T is the sampling period. As noted in the figure, the peaks in spectral envelope $|H(e^{j\omega T})|$ occur at frequencies which are generally quite close to the formant frequencies. The basic approach of the analysis scheme is to estimate the formant frequencies from the peaks in a computed approximation to $|H(e^{j\omega T})|$. To facilitate the estimation of the formant frequencies, constraints on formant frequency ranges and relative levels of the formant peaks are imposed. The method of cepstral analysis and a new spectral-analysis algorithm are techniques employed in the analysis scheme.

B. Previous Research on Formant Estimation

Since the pitch period and formant frequencies are the essential parameters in the acoustic model for the production of voiced speech, it is not surprising that a great deal of previous research has been directed toward estimation of these parameters from the acoustic waveform. Generally, the estimation of pitch period and the estimation of formant frequencies have been treated as two distinct problems. However, since the main emphasis of this paper is on estimation of the formant frequencies, and since estimation of pitch period and formant frequencies do not require independent analyses in the system proposed here, there is no need to review work on pitch period estimation. In contrast, some of the principles upon which the present system is based are basic to many of the previous schemes for estimation of formant frequencies. Therefore, it seems worthwhile to review some of the major efforts in this area that have influenced the work presented here. It should be emphasized that this section is not meant to be a comprehensive review of the literature on formant estimation. Its purpose is simply to place the present work in context.

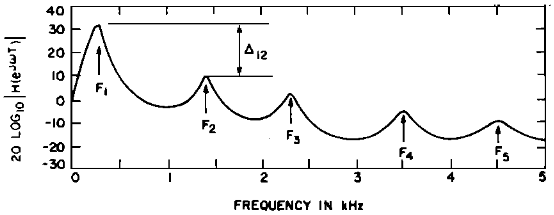

FIG. 3. An example of the evaluation of the system function $G(z)R(z)V(z)$ on the unit circle. Arrows mark the formant frequencies.

Although virtually all methods of formant estimation have been based on essentially a continuous-time version of the model discussed earlier, the techniques for estimating formants have differed widely. A majority of the previous systems have used frequency domain techniques. One such approach could be termed "peak picking," i.e., finding the location of spectral peaks in the short-time amplitude spectrum. For example, Flanagan[3–5] proposed two formant vocoder systems of this type in which a short-time spectrum was computed using an analog filter bank. The first three formants were estimated by locating spectral peaks in frequency ranges appropriate for the first three formants.

In another frequency-domain approach, termed "analysis by synthesis,"[6–9] a spectrum was computed using a filter bank[6–8] or pitch synchronously (i.e., using only a single period of the speech of the speech waveform) using a digital computer.[9] The computer synthesized a "best" spectral match (i.e., minimum mean-square error) by systematically varying the formant frequencies and bandwidths. The success of this technique depends on the accuracy of the speech model. Therefore, its performance is theoretically best for vowels.

A formant vocoder system employing analysis-by-synthesis techniques was discussed by Coker.[10] In this system, the differences between individual speakers were accounted for by recording an average curve of the difference between the measured and synthesized spectra. This difference curve was then used as a correction to improve spectral balance. Formant continuity constraints were also incorporated into the system as an aid in searching for the formants.

The use of low-order spectral moments[11] is another

[2] B. Gold and L. R. Rabiner, "Analysis of Digital and Analog Formant Synthesizers," IEEE Trans. Audio Electroacoust. **AU-16**, 81–94 (1968).

[3] J. L. Flanagan, "Automatic Extraction of Formant Frequencies from Continuous Speech," J. Acoust. Soc. Amer. **28**, 110–118 (1956).

[4] J. L. Flanagan, "Evaluation of Two Formant-Extracting Devices," J. Acoust. Soc. Amer. **28**, 118–125 (1956).

[5] J. L. Flanagan, *Speech Analysis Synthesis and Perception* (Academic Press, Inc. New York, 1965).

[6] C. G. Bell, F. Poza, and K. N. Stevens, "Automatic Resolution of Speech Spectra into Elemental Spectra," *Proceedings of the Seminar on Speech Compression and Processing*, W. Wathen-Dunn and L. E. Woods, Eds., AFCRC-TR-59-198, Vol. 1. Paper A-6, Dec. 1959.

[7] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Amer. **33**, 1725–1736 (1961).

[8] A.P. Paul, A. S. House, and K. N. Stevens, "Automatic Reduction of Vowel Spectra; an Analysis-by-Synthesis Method and its Evaluation," J. Acoust. Soc. Amer. **36**, 303–308 (1964).

[9] M. V. Mathews, J. E. Miller, and E. E. David, Jr., "Pitch Synchronous Analysis of Voiced Sounds," J. Acoust. Soc. Amer. **33**, 179–186 (1961).

[10] C. H. Coker, "Real Time Formant Vocoder, Using a Filter Bank, a General-Purpose Digital Computer, and an Analog Synthesizer," J. Acoust. Soc. Amer. **38**, 940(A) (1965).

[11] J. Suzuki, Y. Kadokawa, and K. Nakata, "Formant-Frequency Extraction by the Method of Moment Calculations," J. Acoust. Soc. Amer. **35**, 1345–1353 (1963).

led 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info

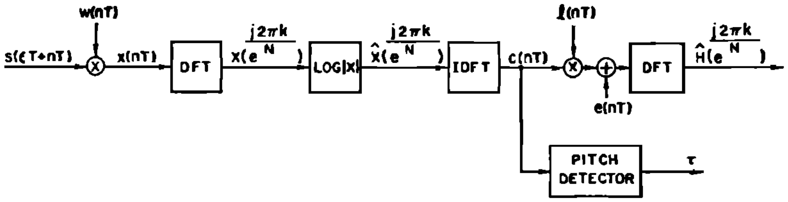
w(nT)
s(ξT+nT)
x(nT)
DFT
$X(e^{\frac{j2\pi k}{N}})$
LOG|X|
$\hat{X}(e^{\frac{j2\pi k}{N}})$
IDFT
c(nT)
ℓ(nT)
e(nT)
DFT
$\hat{H}(e^{\frac{j2\pi k}{N}})$
PITCH DETECTOR
τ

FIG. 4. Block diagram of the system for estimating formant frequencies and pitch period.

frequency-domain approach that has been tried. Although this approach did not produce very accurate formant data, it did yield approximations that were useful in more refined analyses.

One example of a time domain technique for estimating formants was the application of the analysis-by-synthesis approach directly to the acoustic waveform.[12] By systematic variation of the amplitudes, phases, damping and oscillation frequencies of a sum of complex exponential functions, a portion of the acoustic waveform was approximated in a minimum mean-square error sense. This technique was applied to the estimation of the formants of vowels.

The systems discussed in this section are not the only ones which have been proposed. However, they illustrate many of the basic principles involved in estimation of formant frequencies. It seems reasonable to state that none of the previous systems represents a completely successful solution to the problem of automatic estimation of formant frequencies of voiced speech. It may also be said, however, that no matter how sophisticated the analysis, the validity of the results reflects the validity of the model upon which the analysis is based. Although the model which has been discussed is not as good for voiced stops, voiced fricatives, and nasals as for vowels, glides, and semivowels, it was felt that the flexibility and efficiency of the formant representation of voiced speech justified a further effort toward a fully automatic formant-estimation system. The system proposed in this paper incorporates many of the principles just discussed. In addition, new spectral-analysis techniques and constraints based on a digital-speech model have been combined into a system which has produced accurate formant data for vowels, glides, and semivowels, and provided a good spectral match in the case of voiced stops, voiced fricatives, and nasals.

I. DESCRIPTION OF THE ANALYSIS SYSTEM

In this section, an asynchronous system for determining (as a function of time) the pitch period, the gain, and the lowest three formants of voiced-speech sounds is presented. Since the system is asynchronous, exact determination of a "pitch period" is not required. Instead, the analysis is applied to several periods of speech at a time. This has the advantage of eliminating the difficult problem of accurately determining pitch periods in the acoustic waveform. A disadvantage of this method, however, is that formant and pitch period estimates are in a sense "averaged" over the analysis window. As is demonstrated by the results, this "averaging" is not significant for the formant transitions and pitch period changes which are encountered in normal male speech. The analysis involves two basic parts:

(1) The estimation of pitch period and the computation of the spectral envelope,

(2) the estimation of formants from the spectral envelope.

Each of these procedures is discussed in detail in the following two sections.

A. Estimation of Pitch Period and Spectral Envelope

The pitch period and the spectral envelope are estimated from the cepstrum of a segment of the speech waveform. Since a variety of definitions of the cepstrum exists,[13] it is important to state our definition here. In this paper, the cepstrum of a segment of a sampled speech waveform is defined as the inverse transform of the logarithm of the z-transform of that segment.

Cepstral techniques for pitch period estimation have been discussed by Noll.[13] Oppenheim and Schafer[14] and Oppenheim[15] have discussed cepstral analysis techniques and their application to speech analysis within the context of a theory of generalized linear filtering. These previous investigations have shown that the logarithm of the Fourier transform (the z-transform evaluated on the unit circle for sampled data) of a segment of voiced speech consists of a slowly varying component attributable to the convolution of the glottal pulse with the vocal-tract impulse response, plus a rapidly varying periodic component due to the repetitive nature of the acoustic waveform. These two additive components can be separated by linear filtering of the logarithm of the transform. The assumption that the log magnitude is composed of two separate components is supported by the model for the speech waveform given previously. A

[12] E. N. Pinson, "Pitch Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths," J. Acoust. Soc. Amer. 35, 1264–1273 (1963).

[13] A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Amer. 41, 293–309 (1967). Research at Bell Telephone Laboratories I, Proc. Int. Congr. Acoust., 5th, Liège, 1965, Paper A21.

[14] A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," IEEE Trans. Audio Electroacoust. AU-16, 221–226 (1968).

[15] A. V. Oppenheim, "A Speech Analysis–Synthesis System Based on Homomorphic Filtering," J. Acoust. Soc. Amer. 45, 458–465 (1969).

ed 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info

detailed analysis of such a discrete model is given by Oppenheim and Schafer.[14]

The details of the computation of the cepstrum are summarized in Fig. 4. A segment of speech $s(\xi T+nT)$ is weighted by a symmetric window function $w(nT)$ such that

$$\begin{aligned}x(nT)&=s(\xi T+nT)w(nT)\\&=[p(\xi T+nT)*h(nT)]\cdot w(nT),\quad 0\leq n\ \ M,\end{aligned}\qquad(5)$$

where $*$ denotes discrete convolution, and ξT is the starting sample of a particular segment of the speech waveform.

In Eq. 5, $p(\xi T+nT)$ represents a quasiperiodic impulse train appropriate for the particular segment being analyzed, and $h(nT)$ represents the triple convolution of the vocal-tract impulse response with the glottal pulse and the radiation load impulse response [i.e., $h(nT)=v(nT)*r(nT)*g(nT)$, where $v(nT)$ is the inverse z-transform of $V(z)$, $r(nT)$ is the inverse z-transform of $R(z)$, and $g(nT)$ is the inverse z-transform of $G(z)$ in Fig. 1(a)].

The window function $w(nT)$ tapers to zero at each end in order to minimize the effects due to the inclusion of a nonintegral number of pitch periods within the window. Since $w(nT)$ varies slowly with respect to variations in $s(nT)$, $x(nT)$ is given approximately by[14]

$$x(nT)\approx h(nT)*p_w(nT),\qquad(6)$$

where

$$p_w(nT)=p(\xi T+nT)w(nT).\qquad(7)$$

That is, the purpose of multiplication of the speech by the window is to improve the approximation that a segment of voiced speech can be represented as a convolution of a periodic impulse train with a time-invariant vocal-tract impulse response sequence. Segments of speech are selected at 10-msec intervals along the waveform. The window we have used is a Hamming window, which is specified by the equation

$$w(nT)=\begin{cases}0.54-0.46\cos(2\pi nT/MT) & 0\leq nT\leq MT\\ 0 & \text{elsewhere.}\end{cases}\qquad(8)$$

The duration of the window, MT, is four times the maximum of the previous two estimates of pitch period and is the same for pairs of segments. Symmetry properties of the discrete Fourier transforms allow the computation of two real transforms in one computation. To take advantage of this, it is most convenient to change the window duration only for pairs of segments. The choice of the duration of the time window is governed by two conflicting considerations. In order to obtain a strong peak in the cepstrum at the pitch period, it is necessary to have several periods of the waveform within the window. In contrast, in order to obtain strong formant peaks in the smoothed spectrum, only about two periods should be within the window—i.e., the formants should not have changed appreciably within the time interval spanned by the window. The choice of four times the pitch period represents a suitable compromise.

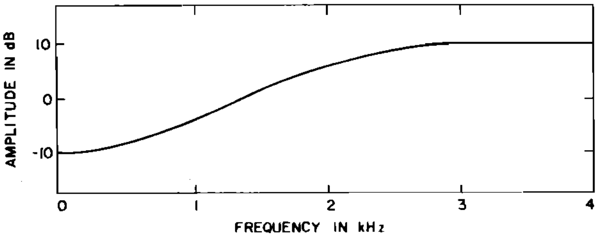

FIG. 5. Spectrum equalizing curve which is added to the smoothed spectral envelope.

The first three blocks of Fig. 4 depict the computation of the cepstrum $c(nT)$. Since a digital computer is used to compute the cepstrum, the discrete Fourier transform (DFT) and the inverse discrete Fourier transform (IDFT) are used in place of the z-transform and the inverse z-transform. [The DFT and IDFT are computed using the fast Fourier transform (FFT) algorithm.[16]]

Since the logarithm of the magnitude of the DFT is a sampled version of the logarithm of the magnitude of the z-transform, the input sequence $x(nT)$ must be augmented with a number of zeros sufficient to ensure that the logarithm of the magnitude is sampled sufficiently often to avoid aliasing in the cepstrum. A value of $N=1024$ for the DFT's and IDFT has been used so that (1024-M) zero samples must be appended to the windowed segment of speech samples.

The remainder of Fig. 4 depicts the estimation of pitch period and spectral envelope from the cepstrum. The cepstrum consists of two components. The component due primarily to the glottal wave and the vocal tract is concentrated in the region $|nT|<\tau$, whereas the component due to the pitch occurs in the region $|nT|\geq\tau$, where τ is the pitch period during the segment being analyzed. The pitch component consists mainly of sharp peaks at multiples of the pitch period. Thus, pitch period can be determined by searching the cepstrum for a strong peak in the region $nT>\tau_{\min}$, where $\tau_{\min}$ is the minimum expected pitch period. Noll[13] has given an algorithm that can be used for pitch-period estimation and voiced/unvoiced detection. The spectral envelope is obtained by low-pass filtering of the log magnitude of the discrete Fourier transform. This is accomplished by multiplying the cepstrum by a function $l(nT)$ of the form

$$l(nT)=\begin{cases}1 & |nT|<\tau_1\\ \frac{1}{2}\{1+\cos[\pi(nT-\tau_1)/\Delta\tau]\} & \tau_1\leq|nT|<\tau_1+\Delta\tau\\ 0 & |nT|\geq\tau_1+\Delta\tau,\end{cases}\qquad(9)$$

[16] J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," Math. Computation 19, 297–301 (1965).

led 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info

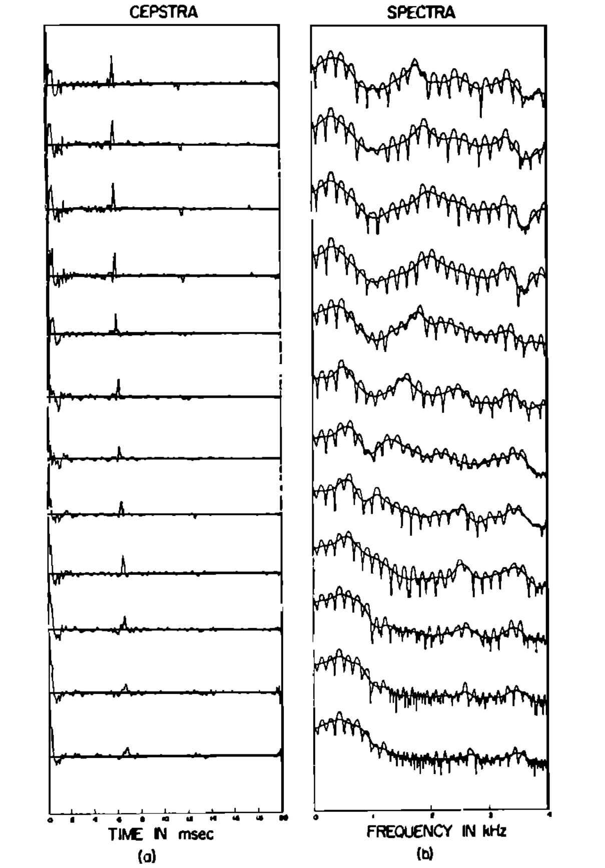

FIG. 6. (a) Cepstra for consecutive segments of speech separated by 20 msec. (b) Spectra and smoothed spectral envelopes corresponding to the cepstra on the left in (a).

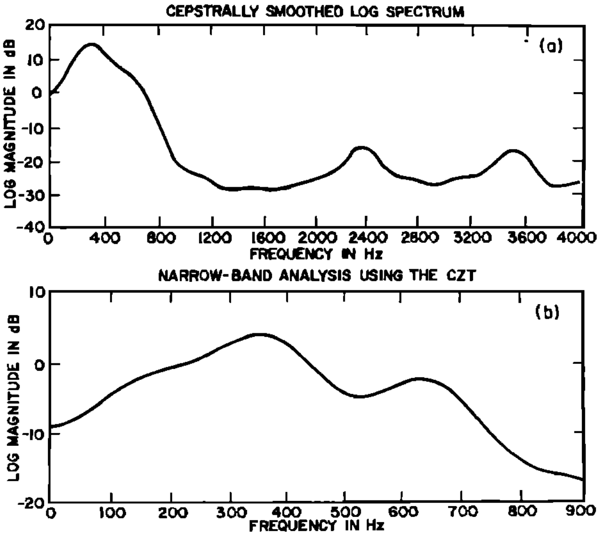

FIG. 7. (a) A smoothed spectral envelope in which F_1 and F_2 are too close to be resolved. (b) Analysis of the cepstrum over a narrow frequency band and on a contour inside the unit circle.

where $\tau_1+\Delta\tau$ is less than the minimum pitch period that will be encountered. The sequence $e(nT)$ is added to the sequence $c(nT)l(nT)$. The purpose of adding this component to the cepstrum is to add the curve of Fig. 5 to the smoothed spectral envelope. The curve of Fig. 5 is an approximation to the frequency response of a spectral equalizer which has been used by Flanagan *et al.*[17] to equalize formant amplitudes. Comparison of Fig. 5 with Fig. 2 shows that the effect of the equalizer is to remove approximately the contribution of the glottal waveform and the radiation load. The sequence $e(nT)$ consists of the Fourier coefficients of the curve of Fig. 5, viewed as a real, even, periodic function. Three nonzero Fourier coefficients are used. The resultant sequence $c(nT)l(nT)+e(nT)$ is transformed to produce the equalized spectral envelope.

If pitch period information is not required, or if pitch period is independently estimated, then the logarithm of the magnitude of the z-transform can be filtered using either a direct convolution or a recursive digital filter. If a simple recursive filter is used, considerable saving in computation may result.

Figure 6(a) shows a series of cepstral plots. The cepstra correspond to consecutive segments of speech; the midpoint of each segment is displaced 20 msec from the midpoint of the previous segment. (In actual operation, the window is moved in 10-msec steps.) From Fig. 6(a), it can be seen that there is a distinct peak in the cepstrum that moves from slightly less than 6 msec to slightly more than 6 msec as time increases. It can also be seen that the cepstrum is large for small values of nT, then drops off rapidly before the peak at the pitch period.

Figure 6(b) shows the spectra corresponding to the same intervals of speech shown in Fig. 6a. Each of the rapidly varying curves is the unsmoothed spectrum corresponding to the cepstrum directly opposite to it in Fig. 6(a). The slowly varying curve is the corresponding smoothed spectrum. In the first curve, the first four formants are clearly in evidence. In the next three curves, F_2 has moved close to F_3, and in the third and fourth curves, F_2 and F_3 are not resolved. In the fifth through eighth curves, all the formants are clear, while in the last four curves, F_1 and F_2 are not resolved.

In cases where F_1, F_2, and F_3 are separated by more than about 300 Hz, there is no difficulty in resolving the corresponding peaks in the smoothed spectrum. However, when F_1 and F_2 or when F_2 and F_3 get closer than about 300 Hz the cepstral smoothing results in the peaks not being resolved. In these cases, a new spectral analysis algorithm called the chirp z-transform[18] (CZT) can be used to advantage. As discussed in Appendix A, the CZT permits the computation of samples of the z-transform at equally spaced intervals along a circular or spiral contour in the z-plane. In particular, if F_1 and F_2

[17] J. L. Flanagan, D. Meinhart, and P. Cummiskey, "Digital Equalizer and De-Equalizer for Speech," J. Acoust. Soc. Amer. 36, 1030(A) (1964).

[18] L. R. Rabiner, R. W. Schafer, and C. M. Rader, "The Chirp z-Transform Algorithm and its Application," Bell System. Tech. J. 48, 1249–1292 (1969).

ed 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info

are close, it is possible to compute the z-transform on a contour that passes closer to the pole locations than the unit-circle contour, thereby enhancing the peaks in the spectrum and improving the resolution. For example, Fig. 7(a) shows a smoothed spectral envelope in which F_1 and F_2 are unresolved. In this case, the parameters of the cepstral-window function $l(nT)$ were $\tau_1=2$ msec and $\Delta\tau=2$ msec. Figure 7(b) shows the result of a CZT analysis along a circular contour of radius $e^{-0.0314}$ over the frequency range 0–900 Hz with a resolution of about 10 Hz. This analysis was achieved using only two 128-point FFTs. The effect of analysis along a contour that passes closer to the poles is evident in contrast to Fig. 7(a). Figure 8 shows an identical analysis for the case when F_2 and F_3 are close together. A discussion of how this technique is incorporated into the algorithm for choosing formants from the smoothed spectra is given in the next section. Discussions of the CZT algorithm itself and its application directly to the cepstrum are given in Appendixes A and B.

Oppenheim[15] has recently discussed a new speech analysis–synthesis system that is similar to the present analysis system. That is, a cepstrum is computed as in the previous discussion, and the pitch period is estimated from the cepstrum. However, instead of transforming to the smoothed spectrum at the analyzer, a small number of cepstral samples (32 or less) are sent to the synthesizer. The synthesizer converts these cepstral values into an impulse response that is then convolved with either a quasiperiodic impulse train (voiced sounds) or a random polarity equally spaced impulse train (unvoiced sounds). In contrast, the approach in the present system is to estimate formant frequencies and pitch period for a formant synthesizer rather than retain the cepstral values. No physical comparison of the two systems has been performed. However, it seems reasonable to state that at least for voiced speech, the system discussed in this paper is related to that of Oppenheim in much the same way as a formant vocoder is related to a channel vocoder.

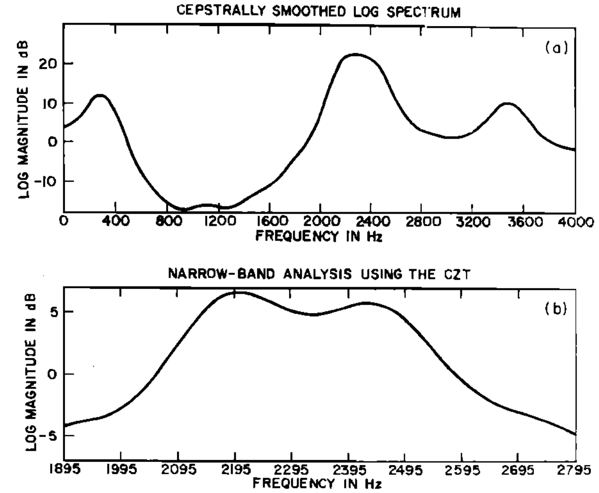

FIG. 8. (a) A smoothed spectral envelope in which F_2 and F_3 are too close to be resolved. (b) Analysis of the cepstrum over a narrow frequency band and on a contour inside the unit circle.

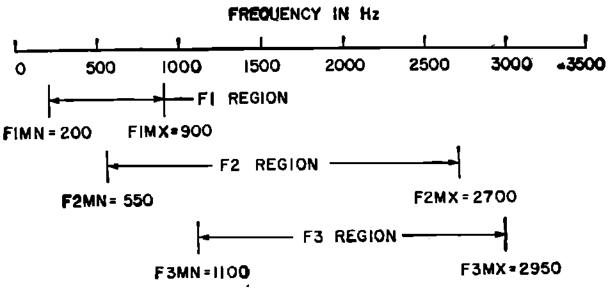

FIG. 9. Frequency ranges for the first three formants. (Empirically derived for male speech.)

B. Estimation of Formant Frequencies from Cepstrally Smoothed Spectra

Before proceeding to the details of the process of estimating the formant frequencies from smoothed spectra, it is necessary to present some data relating to the properties of the speech spectrum. Figure 9 shows the frequency ranges of the first three formants. These ranges were determined from experimental data on male speakers, and it should be noted that these ranges are somewhat more restricted than ranges that would apply to both male and female speakers. On the other hand, individual speakers may have formant ranges that are even more restricted than those of Fig. 9, and, if known, these ranges could be used for that speaker.

It is important to note the high degree of overlap between regions in which the formants may be located. The first formant range is from 200 to 900 Hz, however, for half of this range (550–900 Hz), the second formant region can overlap the first. Similarly, the second and third formant regions overlap from 1100 to 2700 Hz. Thus, the estimation of the formants is not simply a matter of locating the peaks of the spectrum in nonoverlapping frequency bands.

Another property of speech pertinent to formant estimation is the relationship between formant frequencies and relative amplitudes of the formant peaks in the smoothed spectrum. It has previously been noted[1,19] that the form of the model imposes constraints on certain features of the spectral envelope. In particular, it is clear that, for the model of Fig. 1, given the formant frequencies and bandwidths, the relative levels of the peaks in the spectrum are completely specified. The relationships between the formant levels can be very useful in the process of selecting formants from the spectral peaks. In particular, considerable importance is placed on a measurement of the level of the F_2 peak

[19] K. N. Stevens and A. S. House, "An Acoustical Theory of Vowel Production and Some of its Applications," J. Speech Hearing Res. 4, 303–320 (1961).

led 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info

relative to the level of the F_1 peak. The level measurement used Δ_{12} is defined as:

$$\Delta_{12}=\log|H(e^{j2\pi F_2 T})|-\log|H(e^{j2\pi F_1 T})|,$$

where F_1 and F_2 are the frequencies of the first and second formants and $|H(e^{j2\pi FT})|$ is the magnitude of the smoothed spectrum at frequency F hertz. Based upon the model of Fig. 1, a careful analysis shows that Δ_{12} depends primarily upon F_1 and F_2 and that it is fairly insensitive to the bandwidths of all the formants and to the higher formant frequencies. Figure 10 shows the dependence of Δ_{12} on F_2. The curve was derived from examination of a large number of spectra from several male speakers and from computations based on the model. This curve takes into account the equalization of the spectrum shown in Fig. 5 and served as a threshold against which the difference between the level of a possible F_2 peak and the level of the F_1 peak is compared. The dependence of Δ_{12} on F_1 is eliminated by assuming that F_1 is fixed at its lower limit, *F1MN*. If the F_1 dependence were to be accounted for, a family of curves, similar in shape but displaced vertically from the one in Fig. 10 would be required. For a value of F_1 greater than *F1MN*, the corresponding curve would be above the curve in Fig. 10. The shape of the curve is flat until 500 Hz, because F_2 is assumed to be above this minimum value. The curve then decreases until about 1500 Hz, reflecting the drop in F_2 level as it gets further away from F_1. However, above 1500 Hz, the curve rises again owing to the increasing proximity of F_2 and F_3. The curve continues to rise until F_2 gets to its maximum value $F2MX=2700$ Hz, at which point F_2 and F_3 are maximally close (according to the simple model of fixed F_3).

The process of estimating formants from the smoothed spectral envelope is depicted in Fig. 11. The first step is of course the computation of the smoothed spectrum as discussed in the previous section. Once the spectrum is obtained, all the peaks (maxima) are located, and the location and amplitude level of each peak is recorded. This table of peak locations and peak levels contains all the spectral information that is used in the estimation of the formants.

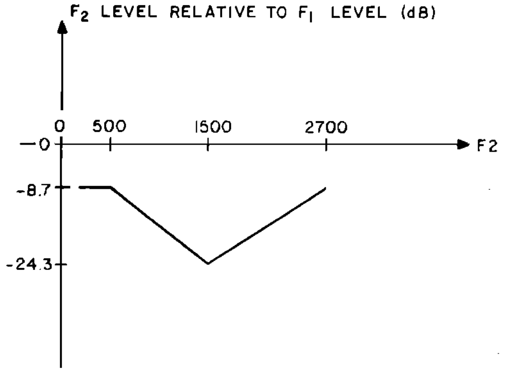

FIG. 10. Level of the F_2 peak relative to the F_1 peak for fixed F_1. (Derived from measurements and computations using the speech model.)

C. Estimation of F_1

The formants are picked in sequence beginning with F_1. The process of estimating F_1 is depicted in Fig. 11(a). To start the process, the highest level of the spectrum in the frequency range 0 to *F1MX* is found, where *F1MX* is the upper limit of the F_1 region. This value of the spectrum is recorded as *FOAMP*. Generally, this value will occur at a peak in the F_1 region that will ultimately be chosen as the F$_1$ peak. However, sometimes there is an especially strong peak below *F1MN*, the lower limit of the F_1 region, which is due to the spectrum of the glottal-source waveform. In such cases, there may or may not be a clearly resolved F_1 peak above *F1MN*. In order to avoid choosing a low-level spurious peak or possibly the F_2 peak for the F_1 peak, when in fact the F_1 peak and peak due to the source are not resolved, a peak in the F_1 region is required to be less than 8.7 dB (1.0 on a natural log scale) below *FOAMP* to be considered as a possible F_1 peak. The frequency of the highest-level peak in the F_1 region which exceeds this threshold is selected as the first formant, F_1. The level of this peak is recorded as *F1AMP*. If no F_1 can be selected this way, the region 0–900 Hz is expanded and enhanced using the CZT algorithm, as discussed in the previous section and the appendixes. (This requires that about 40 values of the cepstrum be saved until after the formants are estimated.) This enhanced section of tne spectrum is then searched for the highest-level peak in the F_1 region. The location of this peak is accepted as F_1. If the enhancement has failed to bring about a resolution of the source peak and the F_1 peak, F_1 is arbitrarily set equal to *F1MN*, the lower limit of the F_1 region.

The quantity *F1AMP* is used in the estimation of F_2. If the F_1 peak is very low in frequency and is not clearly resolved from the lower-frequency peak due to the glottal waveform, *F1AMP* is set equal to (*FOAMP*−8.7 dB). This is done so as to effectively lower (because F_1 is very low) the threshold which is used in searching for F_2.

D. Estimation of F_2

The process of estimating F_2 is depicted in Fig. 11(b). The first step is to fix the frequency range to be searched for F_2. If F_1 has been estimated to be less than *F2MN*, the lower limit of the F_2 region, then only the region from *F2MN* to *F2MX* is searched. However, if F_1 has been estimated to be greater than *F2MN*, it is possible that the F_2 peak has in fact been chosen as the F_1 peak. Therefore, the combined F_1–F_2 region from *F1MN* to *F2MX* is searched so as to ensure that if this is the case, the F_1 peak will be found as the F_2 peak. After F_2 has been estimated, F_1 and F_2 are compared and their values are interchanged if F_2 is less than F_1.

d 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/inf

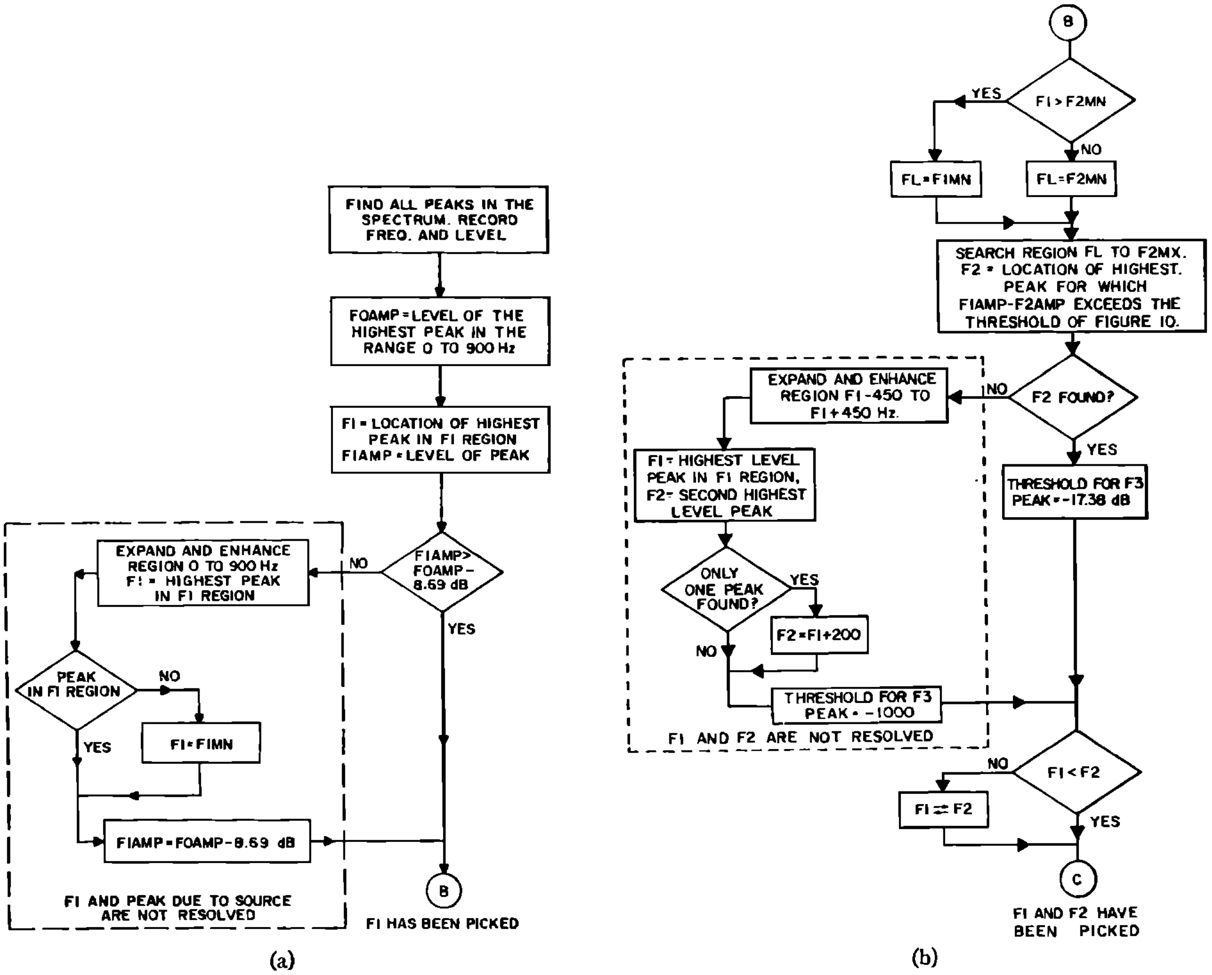
FIND ALL PEAKS IN THE SPECTRUM. RECORD FREQ. AND LEVEL
F0AMP=LEVEL OF THE HIGHEST PEAK IN THE RANGE 0 TO 900 Hz
F1=LOCATION OF HIGHEST PEAK IN F1 REGION F1AMP=LEVEL OF PEAK
F1AMP> F0AMP- 8.69 dB
NO
YES
EXPAND AND ENHANCE REGION 0 TO 900 Hz F1 = HIGHEST PEAK IN F1 REGION
PEAK IN F1 REGION
NO
YES
F1=F1MN
F1AMP=F0AMP-8.69 dB
F1 AND PEAK DUE TO SOURCE ARE NOT RESOLVED
B
F1 HAS BEEN PICKED
(a)
B
F1>F2MN
YES
NO
FL=F1MN
FL=F2MN
SEARCH REGION FL TO F2MX. F2 = LOCATION OF HIGHEST PEAK FOR WHICH F1AMP-F2AMP EXCEEDS THE THRESHOLD OF FIGURE 10.
F2 FOUND?
NO
YES
EXPAND AND ENHANCE REGION F1-450 TO F1+450 Hz.
F1= HIGHEST LEVEL PEAK IN F1 REGION, F2= SECOND HIGHEST LEVEL PEAK
ONLY ONE PEAK FOUND?
YES
NO
F2=F1+200
THRESHOLD FOR F3 PEAK= -1000
F1 AND F2 ARE NOT RESOLVED
THRESHOLD FOR F3 PEAK=-17.38 dB
F1<F2
NO
YES
F1⇄F2
C
F1 AND F2 HAVE BEEN PICKED
(b)

FIG. 11. (a) Flow chart depicting the process of estimating F_1 from the smoothed spectra. (b) Flow chart for estimation of F_2.

In deciding whether a particular spectral peak under investigation is a possible candidate for an F_2 peak, the threshold curve of Fig. 10 is used. The spectral peak is first checked to see if it is located in the proper frequency range. If so, the difference between the level of the peak under consideration and $F1AMP$ is computed. If this difference exceeds the threshold of Fig. 10, that peak is a possible F_2 peak; if not, that peak is not considered as a possible F_2 peak. The value of F_2 is chosen to be the frequency of the highest level peak to exceed the threshold. The level of this peak is recorded as $F2AMP$.

If not peaks are found that exceeded the threshold, further analysis is called for. The fact that no peaks are located has been found to be a reliable indication that F_1 and F_2 are close together as in Fig. 7(a). Therefore, the CZT algorithm is used to compute a high-resolution narrow-band spectrum over the frequency range (F_1-450) Hz to (F_1+450) Hz. (If $F_1<450$ Hz, the range is 0 to 900 Hz.) This spectrum is evaluated along a circular arc of radius $e^{-0.0314}$ in the z-plane. This analysis generally produces a spectrum such as shown in Fig. 7(b) in which the two formants F_1 and F_2 are readily seen.

The value of F_1 is reassigned as the frequency of the highest-level peak in the F_1 region and F_2 is the frequency of the next highest peak. If only one peak is found, F_1 is arbitrarily set equal to the frequency of that peak and $F_2=(F_1+200)$ Hz.

In searching for F_3, a threshold on the difference in level between a possible F_3 peak and the F_2 peak is employed. In this case a fixed, frequency-independent threshold has been found satisfactory. If F_2 is located without the CZT analysis (i.e., F_2 is not extremely low), the threshold on the difference is set at -17.4 dB (-2.0 on a natural log scale). Otherwise, the threshold is effectively removed by setting it at -1000 dB.

E. Estimation of F_3

The estimation of F_3 from the smoothed spectrum is depicted in Fig. 11(c). Because of the equalization, there is a possibility of finding the F_3 peak as F_2. Thus, F_2 is

aded 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/t

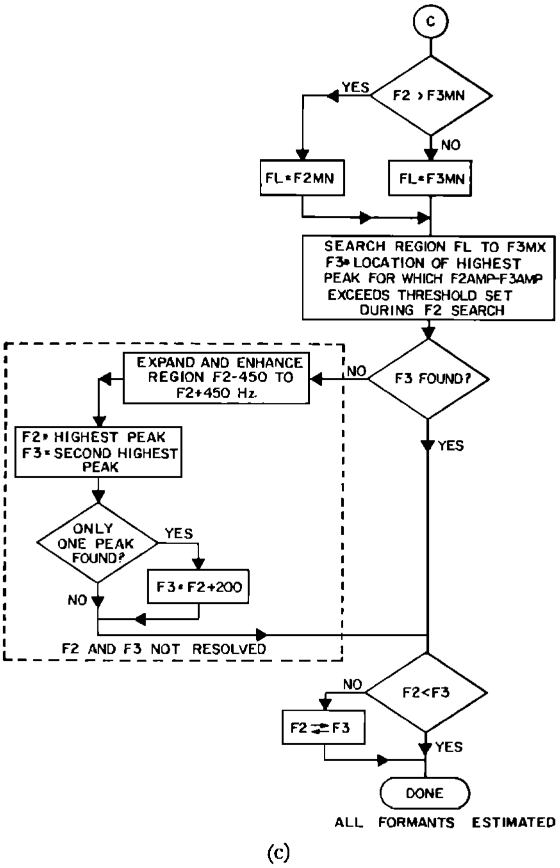

FIG. 11 (*continued*). (c) Flow chart for estimation of F_3.

checked to see if it is greater than $F3MN$, the lower limit of the F_3 region. If so, the search for F_3 is extended to cover the combined F_2–F_3 region from $F2MN$ to $F3MX$. Otherwise, the frequency region $F3MX$ to $F3MX$ is searched. As before, a spectral peak is first checked to see if it is in the correct frequency range. Then, the difference between the level of the peak being considered for an F_3 peak and $F2AMP$ is computed. The highest-level peak that exceeds the threshold is chosen as the F_3 peak. If no peak is found for F_3, further analysis is again called for. It has been found that this situation is generally due to F_2 and F_3 being very close together as in Fig. 8(a). As before, an enhanced spectrum is computed using the CZT, in this case over the frequency range (F_2-450)-(F_2+450) Hz. The result is normally a spectrum such as shown in Fig. 8(b) where F_2 and F_3 are clearly resolved. F_2 is chosen to be the frequency of the highest peak and F_3 to be the frequency of the next highest peak. If only one peak is found, that peak is arbitrarily called the F_2 peak and F_3 is set to (F_2+200) Hz. (This may sometimes result in estimates of both F_2 and F_3 that are slightly high.) The final step in the process is to compare F_2 and F_3 and interchange their values if F_2 is greater than F_3.

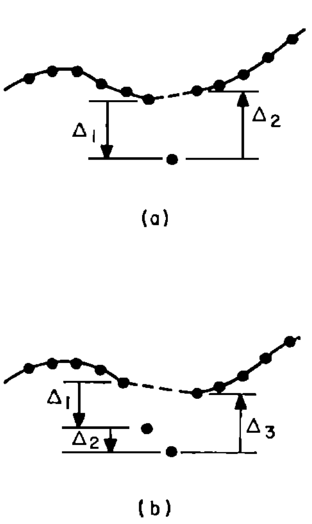

FIG. 12. Nonlinear smoothing applied to formant and pitch estimates. (a) One point "out of line." (b) Two points "out of line."

F. Final Smoothing

As can be seen from the preceding discussion, the three formants are estimated entirely from a single com-

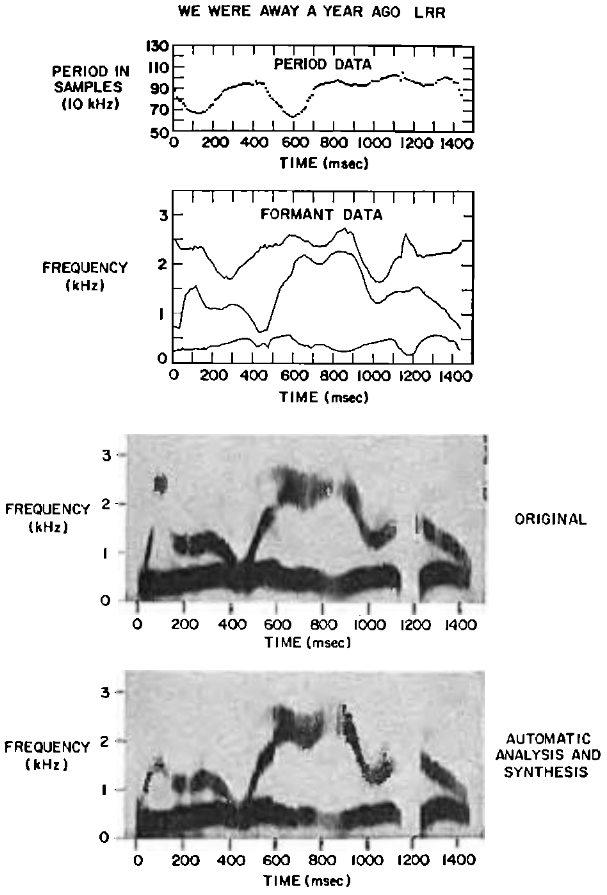

FIG. 13. Automatic analysis and synthesis of "We were away a year ago." Speaker LRR. (a) Pitch period and formant data as plotted by computer. (b) Wide-band spectrogram of original speech. (c) Wide-band spectrogram of synthetic speech generated from the data in (a).

ded 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/

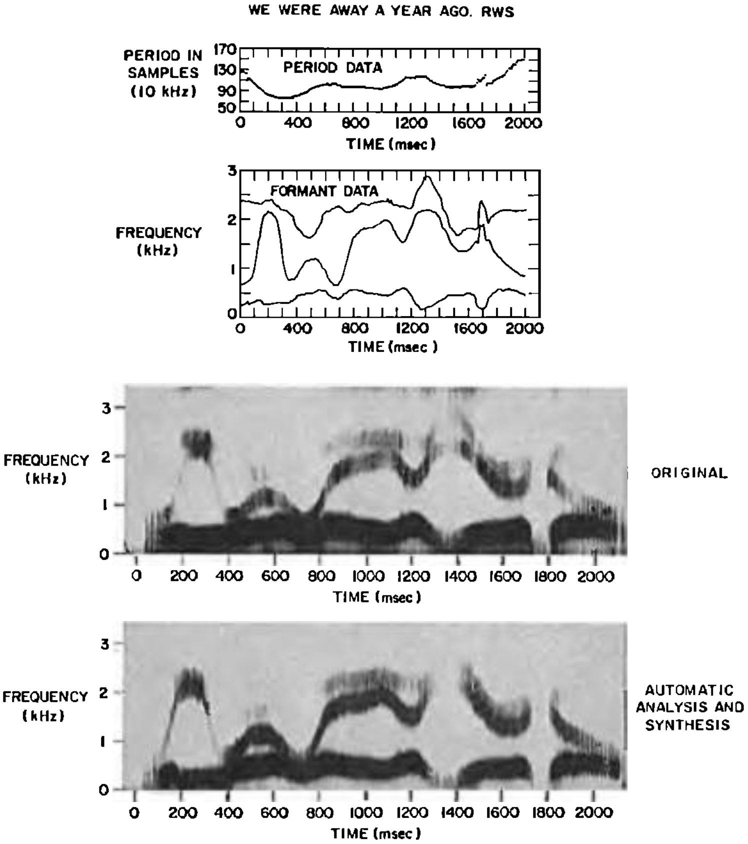

FIG. 14. Automatic analysis and synthesis of "We were away a year ago." Speaker RWS. (a) Pitch period and formant data as plotted by computer. (b) Wide-band spectrogram of original speech. (c) Wide-band spectrogram of synthetic speech generated from the data in (a).

putation of the short-time cepstrum. No attempt has been made to take advantage of the fact that the formants must vary with time in a continuous manner. Simple schemes for using continuity constraints as an aid in searching for the formant peaks were tried. It was found that such schemes generally led to situations in which it was difficult to recover from an erroneous estimate of one of the formants.

For this reason, the estimate of the formants at a given time depends only on the spectrum computed at that time. It is inevitable that some gross errors are made in the estimation of the formants in this manner. The continuity constraint provides a means of correcting such gross errors, i.e., points which are clearly out of line. Continuity constraints are incorporated into the system through the use of a simple, nonlinear smoothing operation applied to both the formant and pitch period data. Figure 12 illustrates the two types of smoothing corrections that are applied. Figure 12(a) shows the case of one point out of line. Here both Δ_1 and Δ_2 are of opposite sign, and both are greater in magnitude than a fixed threshold Δ_T. The point out of line is reassigned the value of the average of the values of the preceding and following points. No other points in the region are altered. Figure 12(b) shows the case of two points out of line. Here both Δ_1 and Δ_3 are of opposite sign and both are greater in magnitude than Δ_T. In this case, the points out of line are reassigned values on a straight line between the two end samples. No other type of smoothing is used. Values for Δ_T of 1 msec, 100 Hz, 150 Hz, and 200 Hz for the τ, F_1, F_2, and F_3 data, respectively, have been used.

II. RESULTS

An algorithm for estimating the three lowest formant frequencies of voiced speech has been described in considerable detail. The algorithm has been designed to perform well on vowels, glides, and semivowels. No attempt has been made to incorporate into the algorithm any special provisions for dealing with voiced stop consonants or nasal consonants.

One way of evaluating the performance of a formant estimation system is through synthesis of speech from the estimated formants and pitch period. The original speech and the synthetic speech can be compared aurally and through visual inspection of wideband spectrograms. Only informal listening tests have been performed, so that it is not possible to present detailed perceptual results. However, it is possible to present some results in spectrographic form. These results are shown in Figs. 13–16. Each figure consists of: (a) formant and pitch-period plots as obtained from the computer, (b) a wide-band spectrogram of the original speech, and (c) a wide-band spectrogram of speech

aded 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/te

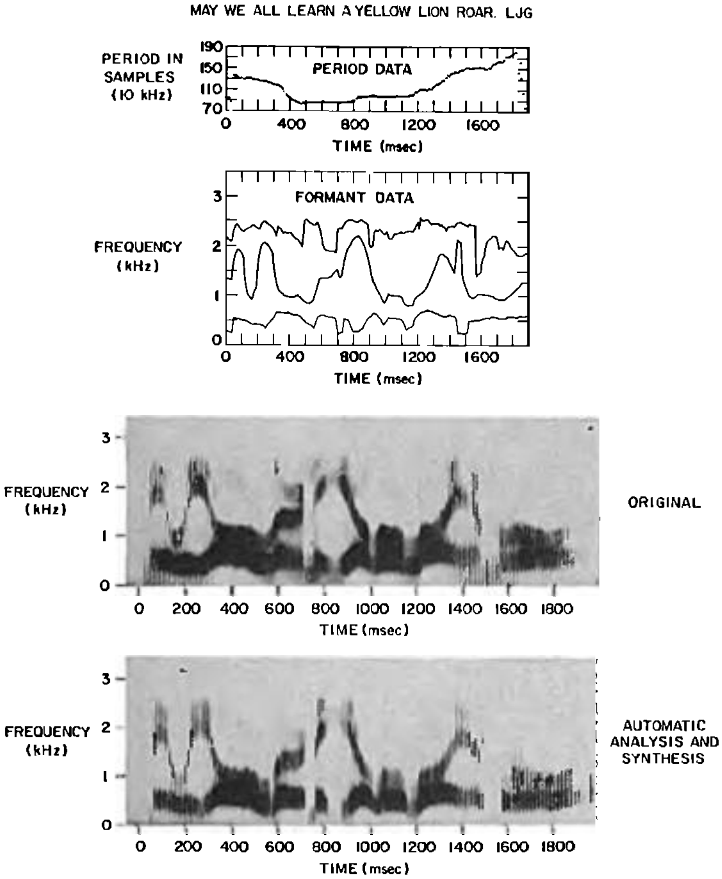

FIG. 15. Automatic analysis and synthesis of "May we all learn a yellow lion roar." Speaker LJG. (a) Pitch period and formant data as plotted by computer. (b) Wide-band spectrogram of original speech. (c) Wide-band spectrogram of synthetic speech generated from the data in (a).

synthesized from the data shown in (a) of that figure. Before discussing these results, it is appropriate to discuss some of the details of the generation of the synthetic speech.

A. Synthesis

The synthesizer configuration is shown in Fig. 1. For the specific results presented here, the bandwidth parameters of all the formants were held fixed at the following values: $\alpha_1=60\pi$, $\alpha_2=100\pi$, $\alpha_3=120\pi$, and $\alpha_4=175\pi$. The frequency of the highest formant was fixed at the value $F_4=4000$ Hz. The fixed spectral shaping is as shown in Fig. 2, and the parameters a and b are, respectively, 400π and 5000π. The gain parameter A was computed from the original speech waveform. A new value of A was computed for each estimate of the formant frequencies and pitch period. The value of A is directly proportional to the rms value of the original speech waveform within the analysis interval. This causes the rms value of each period of synthetic speech to be approximately the same as a corresponding period of the real speech.

In using the system of Fig. 1 for synthesis, the parameters were supplied at 10 msec intervals. However, the system parameters are only allowed to change at a time when an impulse is present at the output of the impulse generator, i.e., pitch synchronously.

B. Examples

Figure 13 shows the automatic analysis and synthesis of the utterance "We were away a year ago," spoken by Speaker LRR. Figure 14 shows the analysis and synthesis of the same utterance by Speaker RWS. Figure 15 shows the analysis and synthesis of the utterance "May we all learn a yellow lion roar," by Speaker LJG. Figure 16 shows the analysis and synthesis for "May we all . . ." by Speaker PDB. There are several comments on these examples and on our results in general that seem worthwhile.

- In all cases, it is clear that the spectrograms of original and synthetic speech compare favorably. As judged by informal listening, the intelligibility of the synthetic speech is equal to that of the original.

- The synthetic speech retains many of the qualities of the original speaker. However, in the case of Speakers LRR and LJG, the synthetic speech was judged by experienced listeners to sound more like the original than for the other two speakers. It is speculated that this is due to a better spectral balance for Speakers LRR

ded 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/t

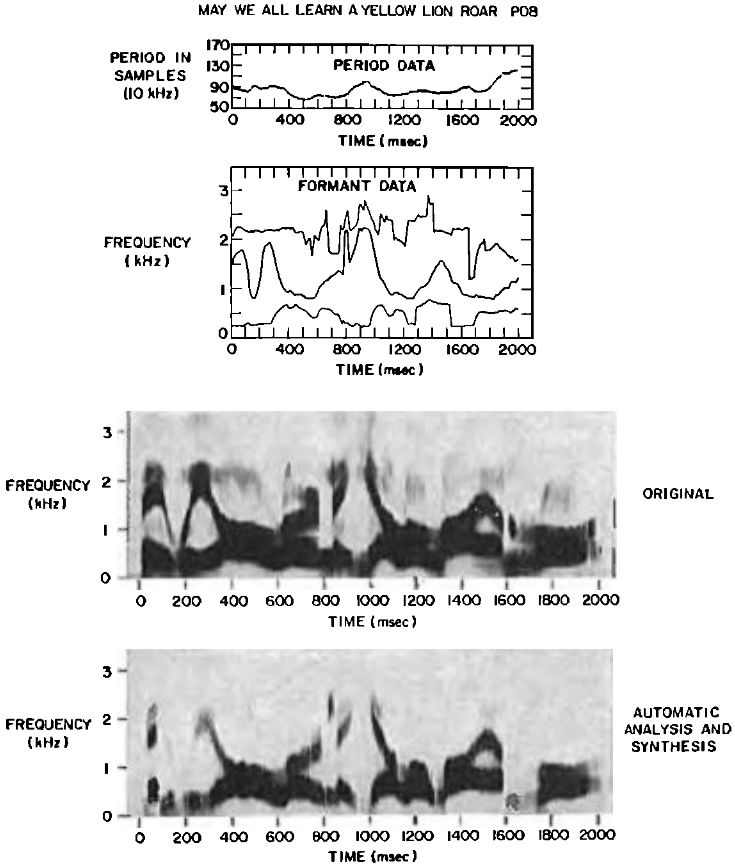

FIG. 16. Automatic analysis and synthesis of "May we all learn a yellow lion roar." Speaker PDB. (a) Pitch period and formant data as plotted by computer. (b) Wide-band spectrogram of original speech. (c) Wide-band spectrogram of synthetic speech generated from the data in (a).

and LJG, which can in turn be attributed to the fixed spectral balance network discussed earlier. It is conceivable that adjustment of the fixed parameters of this network would improve the spectral balance for other speakers.

- It is difficult to express quantitatively accuracy of the system in estimating formants. The synthetic speech was reanalyzed, and results were obtained that agreed with the synthesizer control data to within 40 Hz for all formants. It can be seen in Fig. 15 and especially Fig. 16 that the F_3 data vary erratically in places. This did not cause much difficulty in the synthesis. It should be noted that the places where the automatic scheme is reliable are the same places where it is almost impossible to locate the formant by eye on the spectrogram of the original speech. At these places, the formant has a low spectra level and is perceptually unimportant.

- It is stressed that this system was not specially designed to handle nasal consonants. Figures 15 and 16 show the spectrograms of the utterance "*M*ay we all lear*n* a yellow lio*n* roar." By comparing the spectrograms, it can be seen that the system tends to estimate formants that produce a spectrum which matches the original reasonably well. In all cases except the /m/ in Fig. 16, the synthetic nasals are properly perceived even though the extra pole–zero pair called for by the acoustic theory has not been included. Since the results on nasals are at present limited, it is not appropriate to place much importance on this finding.

- Within the class of speech sounds for which the system was designed, its major limitation seems to be that results of the quality of Figs. 13 16 have only been obtained for male speakers. Using appropriate formant frequency ranges, we have tested the system on speech from a female speaker. The results were not as good as for male speech in the sense that the formant data was not very smooth. However, speech synthesized from these data was quite intelligible but contained some distracting sections due to the roughness of the data.

III. CONCLUSION

A new system for automatically estimating formant frequencies of voiced speech has been discussed in detail. This system, like many previous ones, attempts to obtain the short-time spectral envelope of the speech and estimates the formants by searching for peaks in this spectrum. Cepstral-analysis techniques and a new spectral-analysis algorithm as well as the theoretical properties of the spectral envelope have been used in the realization of this system.

The results obtained so far have been judged by experienced listeners to produce highly intelligible and, in

aded 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/te

some instances, very natural-sounding synthetic speech. It is anticipated that this approach may be extended to a wider class of speech sounds than it is presently used for.

A present limitation of the system is that high-quality results are produced only for speech from male speakers. The entire scheme has been programmed on a GE-635 computer and runs in about 120 times real time. That is, to estimate three formants and pitch period for 2 sec of speech and plot the result on microfilm requires about 4 min of computing time on a GE-635 computer. This cost is certainly not prohibitive if one is using the system as a tool in research on speech synthesis. Hardware realization of such a system is not out of the question given the existence of a special-purpose computer for computing the transforms.

Plans for future work primarily focus on extention to a wider class of speech sounds. This will require a reliable voiced/unvoiced detection and detection of different phoneme classes. It is felt that most of the analysis can be performed within essentially the present framework using a combination of the cepstrum, smoothed spectrum, and simple measurements on the acoustic waveform.

Appendix A. Chirp z-Transform Algorithm

The chirp z-transform algorithm[18] plays an important role in the method of estimating formant frequencies discussed in Sec. I. This Appendix discusses the algorithm and summarizes its important properties. Consider the z-transform of a finite sequence of samples $\{x_n,\ n=0, 1,\cdots, N-1\}$;

$$X(z)=\sum_{n=0}^{N-1} x_n z^{-n}. \tag{A1}$$

The CZT algorithm is an efficient means for evaluating Eq. A1 at the points

$$z_k=AW^{-k},\quad k=0, 1,\cdots, M-1, \tag{A2}$$

where M is an arbitrary integer and A and W are arbitrary complex numbers of the form

$$A=A_0e^{j2\pi\theta_0},$$
$$W=W_0e^{j2\pi\phi_0}.$$

The case $A=1$, $M=N$, and $W=e^{j(2\pi/N)}$ corresponds to the discrete Fourier transform which, when N is a highly composite number, can be very efficiently evaluated using one of the so-called fast Fourier transform (FFT) algorithms.[16] The more general z-plane contour specified by Eq. A2 is shown in Fig. A-1, The contour begins at the point $z=A$, and depending on W_0, spirals in or out with respect to the origin. If $W_0=1$, the contour is an arc of a circle of radius A_0. The angular spacing of the samples along this contour is $2\pi\phi_0$. This z-plane contour can be related to an s-plane contour through the relation $z=e^{sT}$. Thus, the equivalent s-plane contour begins at the point

$$s_0=\sigma_0+j\omega_0=(1/T)\ln A,$$

and the samples of the transform are evaluated at the points

$$s_k=s_0+k(\Delta\sigma+j\Delta\omega)=(1/T)(\ln A-k\ln W)$$

for $k=0, 1, \cdots, M-1$. The CZT algorithm is an efficient algorithm for evaluating the transform on such contours and for our purposes affords the following advantages:

- The number of time samples N does not have to equal the number of frequency samples M.

- Neither M nor N need be a composite number.

- The starting point $z=A$ is arbitrary. ($A_0<1$ causes enhancement of the spectral resonances, and θ_0 is chosen so as to center the analysis on the frequency region of interest.)

- The frequency spacing of the spectral samples depends on ϕ_0, which is also arbitrary. Thus, a fine-grain

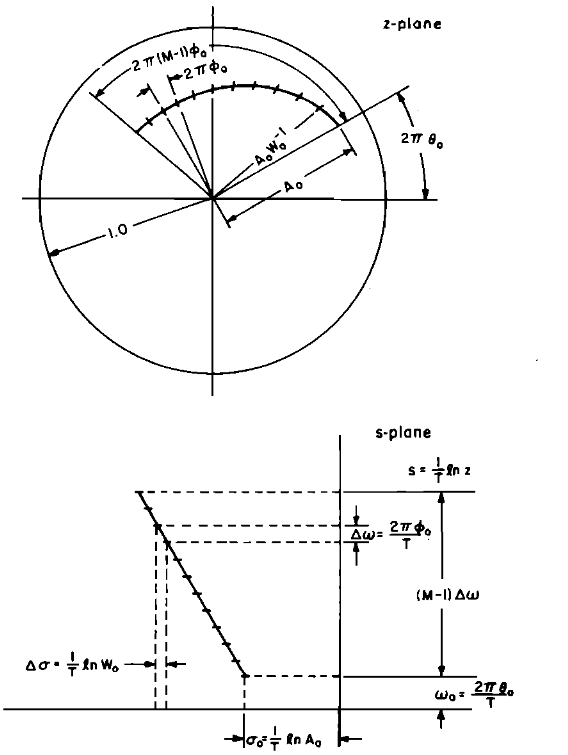
z-plane
$2\pi(M-1)\phi_0$
$2\pi\phi_0$
$A_0W_0^{-1}$
A_0
$2\pi\,\theta_0$
1.0
s-plane
$s=\frac{1}{T}\ln z$
$\Delta\omega=\frac{2\pi\phi_0}{T}$
$(M-1)\Delta\omega$
$\Delta\sigma=\frac{1}{T}\ln W_0$
$\omega_0=\frac{2\pi\theta_0}{T}$
$\sigma_0=\frac{1}{T}\ln A_0$

FIG. A-1. *Top*: z-plane contour for analysis using the chirp z-transform. *Bottom*: Corresponding s-plane contour.

d 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/inf

frequency analysis can be performed over a narrow band beginning at some arbitrary frequency specified by θ_0.

Along the contour of Eq. A2, Eq. A1 becomes

$$X_k = \sum_{n=0}^{N-1} x_n A^{-n} W^{nk} \quad k=0, 1, \cdots, M-1, \quad \text{(A3)}$$

which seems to require $N \cdot M$ complex multiplications and additions. However, through the substitution[A1]

$$nk = [n^2 + k^2 - (k-n)^2]/2,$$

Eq. A3 can be written

$$x_k = W^{k^2/2} \sum_{n=0}^{N-1} (x_n A^{-n} W^{n^2/2}) W^{-(k-n)^2/2},$$
$$k=0, 1, \cdots, M-1. \quad \text{(A4)}$$

This form of the equation can be evaluated with computation time roughly proportional to $(N+M)$ log $(N+M)$, since the convolution sum in Eq. A4 can be evaluated using the FFT.[A2] The steps in the programming of the algorithm are given in detail in Ref. 18.

[A1] L. I. Bluestein, "A Linear Filtering Approach to the Computation of the Discrete Fourier Transform," 1968 NEREM Rec. **10**, 218–219 (Nov. 1968).

[A2] T. G. Stockham, Jr., "High Speed Convolution and Correlation," 1966 Spring Joint Computer Conf. AFIPS Proc. **28**, 229–233 (1966).

Appendix B. Application of the Chirp z-Transform to the Cepstrum

The purpose of this Appendix is to show that applying the CZT algorithm directly to the cepstrum $c(nT)$ leads to a meaningful enhancement of the resonances in the smoothed spectrum. The complex cepstrum, $\hat{x}(nT)$, is defined as the inverse transform of the *complex* logarithm of $X(z)$, where $X(z)$ is the z-transform of the input sequence $x(nT)$[14,B1]. It can be shown that the cepstrum $c(nT)$ (as defined in this paper) is just the even part of $\hat{x}(nT)$; i.e.,

$$c(nT) = [\hat{x}(nT) + \hat{x}(-nT)]/2.$$

Thus, $c(nT)$ and $\log|X(z)|$ are transforms of each other. A sequence $y(nT)$ whose z-transform $Y(z)$ is minimum phase (has no poles or zeros outside the unit circle) has a complex cepstrum $\hat{y}(nT)$ that is zero for $n<0$.[14,B1] Furthermore, if

$$\log|Y(z)| = \log|X(z)|, \quad \text{(B1)}$$

then

$$\begin{aligned} \hat{y}(nT) &= 0 & n&<0 \\ &= c(nT) & n&=0 \\ &= 2c(nT) & n&>0. \end{aligned} \quad \text{(B2)}$$

Therefore, arguing backwards from Eq. B2, we can always compute a minimum phase sequence that has the same log magnitude as $X(z)$.

It can also be shown that if a sequence $y_1(nT)$ is defined by

$$y_1(nT) = A^n y(nT) \quad \text{(B3)}$$

with z-transform

$$Y_1(z) = \sum_{n=0}^{N-1} y(nT) A^n z^{-n} = Y(z/A), \quad \text{(B4)}$$

then

$$\hat{y}_1(nT) = A^n \hat{y}(nT). \quad \text{(B5)}$$

This computation of $\log|Y_1(e^{j\omega T})|$ is equivalent to evaluation of $\log|Y(A^{-1}e^{j\omega T})|$, which is in turn identical to $\log|X(A^{-1}e^{j\omega T})|$. Therefore, using the CZT algorithm to evaluate the transform of $\hat{y}(nT)$ as given in Eq. B2 on a circular arc of radius A^{-1}, the same result in the log magnitude is achieved as if the original evaluation of the z-transform had been made along that contour.

[B1] A. V. Oppenheim, R. W. Schafer and T. G. Stockham, Jr., "Nonlinear Filtering of Multiplied and Convolved Signals," Proc. IEEE **56**, 1264–1291 (1968).

ded 26 Oct 2010 to 202.120.37.167. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/t