form at larynx during speech by means of an inverse filter," in *Proc. Stockholm Speech Commun. Seminar*, Royal Inst. Technol., Stockholm, Sweden, Sept. 1962.
[15] T. V. Ananthapadmanabha and B. Yegnanarayana, "A decomposition technique for composite signals," to be published.
[16] E. A. Robinson, *Statistical Communication and Detection.* London, England: Griffin, 1967, ch. 9, p. 253.
[17] T. Y. Young, "Epoch detection—A new method for resolving overlapping signals," *Bell Syst. Tech. J.*, vol. 44, pp. 401–425, Mar. 1965.
[18] A. Papoulis, *Systems and Transforms with Application in Optics.* New York: McGraw-Hill, 1968, ch. 7.
[19] B. F. Cron, "Phase distortion of a pulse caused by bottom reflection," *J. Acoust. Soc. Amer.*, vol. 37, pp. 486–492, 1965.
[20] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra.* New York: Dover, 1968.
[21] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, 1971.
[22] R. W. Schafer, "A survey of digital speech processing techniques," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 28–35, Mar. 1972.
[23] J. D. Markel, "Digital inverse filtering—A new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129–137, June 1972.
[24] J. I. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman, Inc., Cambridge, Mass., BBN Rep. 2304, Aug. 1972.

A Semiautomatic Pitch Detector (SAPD)

CAROL A. McGONEGAL, LAWRENCE R. RABINER, SENIOR MEMBER, IEEE, AND AARON E. ROSENBERG, MEMBER, IEEE

Abstract—**The purpose of this paper is to describe a technique for semiautomatically determining the pitch contour of an utterance. The method is significantly more sophisticated than the standard technique of hand tracking of pitch periods from a waveform display of the utterance and leads to a fairly robust measurement of the pitch period. This technique utilizes a simultaneous display (on a 10 ms section-by-section basis) of the low-pass filtered waveform, the autocorrelation of a 400-point segment of the low-pass filtered waveform, and the cepstrum of the same 400-point segment of the wideband recording. For each of the separate displays (i.e., waveform, autocorrelation, and cepstrum) an independent estimate of the pitch period is made on an interactive basis with the computer, and the final pitch period decision is made by the user based on results of each of the measurements. The technique has been tested on a large number of utterances spoken by a variety of speakers with very good results. Formal tests of the method were made in which four people were asked to use the method on three different utterances, and their results were then compared. During voiced regions, the standard deviation in the value of the pitch period was about 0.5 samples across the four people. The standard deviation of the location of the time at which voiced regions became unvoiced, and vice versa was on the order of half a section duration, or 5 ms. The major limitation of the proposed method is that it requires about 30 min to analyze 1 s of speech. However, the increased accuracy and robustness of the results indicate that the tradeoff of time for accuracy is a good one for many applications.**

I. Introduction

FOR SOME applications, an extremely accurate and reliable measurement of the pitch contour of an utterance is required. One such application is a comparison and evaluation study of a variety of pitch detection algorithms which has recently been performed at Bell Laboratories [1]. Another application was a study of the inter and intra speaker similarities in pitch contours for several utterances [2]. Other applications include studies for determination of linguistic rules for pitch generation for use in speech synthesis applications [3], [4]. Although a large number of pitch detection algorithms have been proposed in the literature, none of them is able to achieve the performance of a human who is knowledgeable in the area of speech communications with a fairly sophisticated interactive display of the speech waveform.

The usual method of manual pitch detection is for the user to mark pitch periods on a period-by-period basis, directly on a display of the speech waveform. Although such a technique is often quite good, there are segments of some speech sounds during which the waveform periodicity is not clearly visible in the waveform due to rapid spectral changes in the sound [5]. During such intervals a rough indication of the pitch period can be obtained from the waveform, but due to the changing spectrum, the pitch period estimate can be off by several samples. It is the purpose of the paper to describe a semiautomatic pitch detection (SAPD) technique which is significantly more sophisticated than the standard manual pitch tracking method described above, and which has been found to yield reliable, and repeatable estimates of the pitch period across a variety of speakers and utterances.

Manuscript received May 19, 1975.
The authors are with Bell Laboratories, Murray Hill, N.J. 07974.

II. The Analysis System

Fig. 1 shows a block diagram of the SAPD processing. The speech signal $s(n)$ sampled at a 10 kHz rate is processed to give three simultaneous displays for each section of speech. For two of the displays the speech is low-pass filtered by an $N = 99$ linear phase, finite impulse response (FIR), low-pass digital filter with a passband cutoff frequency of 900 Hz, and a stopband cutoff frequency of 1100 Hz. Fig. 2 shows a plot of the log magnitude frequency response of the filter. The passband ripple of the filter is about 0.03 and the stopband ripple is down about 50 dB. The low-pass filtered speech waveform

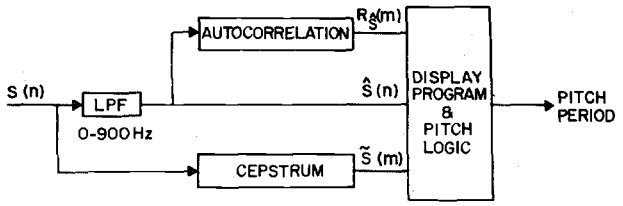

Fig. 1. Block diagram of the SAPD system.

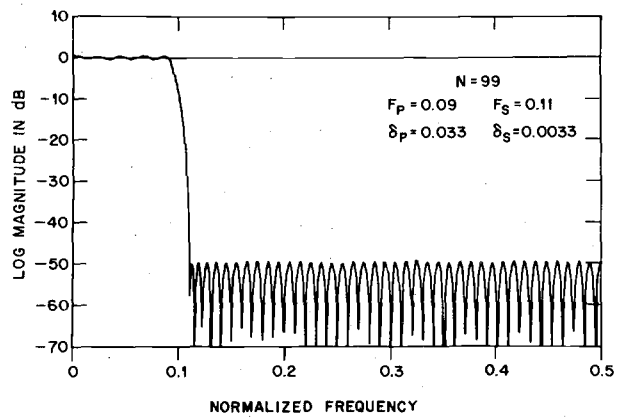

Fig. 2. Log magnitude frequency response of the low-pass filter.

$\hat{s}(n)$ is autocorrelated to give the first 250 points of the autocorrelation function $R_{\hat{s}}(m)$. Finally, the cepstrum of the wideband signal $\tilde{s}(m)$ is computed as the third display.

The choice of these three displays, i.e., the low-pass filtered waveform, its autocorrelation function, and the cepstrum of the wideband signal, was dictated by the desire to obtain several fairly independent estimates of the pitch period. It was expected that the autocorrelation display would provide a good estimate of the pitch period in cases where the waveform was unreliable due to the changing waveform phase since the autocorrelation function is essentially phase insensitive. The low-pass filtered waveform was used rather than the wideband signal because the high-frequency information present in the waveform tends to mask visual identification of the period rather than aid in the decision. Finally, the cepstrum was used to provide a frequency domain pitch measurement which again was fairly insensitive to changing phase in the waveform.

Fig. 3 shows a typical frame of the SAPD program, for a section of voiced speech. The top display shows 60 ms (600 samples) of the low-pass filtered waveform on a normalized amplitude scale. The solid vertical line at sample 1000 shows the center of the current section of speech being analyzed. The user manually (via cursor controls on the console) sets two pointers which delineate the pitch period closest to the solid marker, i.e., one marker precedes the solid marker, the other follows it. The computer draws a dashed line at each pitch period indication. Once the appropriate pitch period has been found, the computer indicates the sample number of the beginning of the period ($X1$ = 985) and the end of the period ($X2$ = 1063) with the difference being the first estimate of the current pitch period (PD = 78). It should be noted that all pitch markers from previous analysis frames are included in the low-pass waveform display to provide a continuity from period to period. It is also worth noting that experience has

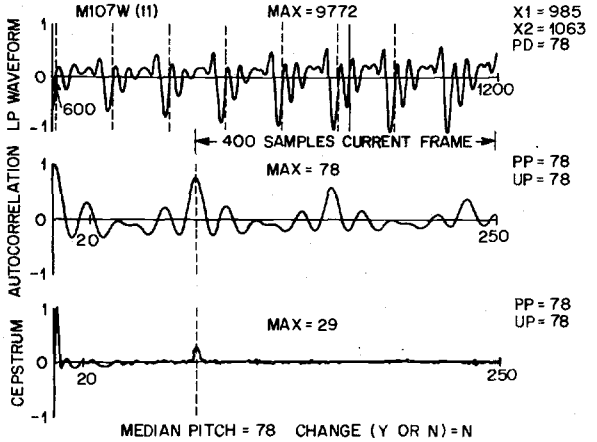

Fig. 3. Typical frame from the SAPD analysis during a voiced segment.

indicated that a reasonable place to put the pitch markers is at an upwards zero crossing where there is a sharp slope in the waveform. To this end the SAPD program aids the user in locating the exact place to put the marker, once the user has indicated the approximate location, by searching in the neighborhood of the indicated point to find the nearest place where an up crossing occurs. Thus extremely accurate placement of the markers is handled by the program, and is therefore not a stringent requirement on the user.

The second display in Fig. 3 is the first 250 points of the autocorrelation of the low-pass filtered waveform. The autocorrelation is computed from a 400-point section of the low-pass filtered waveform centered at the middle of the current frame, i.e., the 400 samples used in the computation are the last 400 samples displayed in the waveform display of Fig. 3. The SAPD program searches the autocorrelation function for the largest peak in the range m = 20 to m = 249, and draws a dotted line at the peak value, and indicates this pitch period estimate to the user (PP = 78). If the peak picker has made a mistake, or the user disagrees with the value chosen for the pitch period, he can manually move a curser to any point on the autocorrelation function and the program will select the maximum value in a 5-point range around the chosen value and indicate this new value to the user (UP = 78). This feature is especially useful in cases where the peak of the autocorrelation function at twice the period is somewhat greater than the peak of the autocorrelation function at the true period.

The final display in Fig. 3 is the cepstrum of the 400-sample section of the wideband speech centered at the middle of the current frame. The strongest peak in the cepstrum in the range m = 20 to m = 249 is automatically located, its position indicated by a dotted line, and this pitch period estimate is indicated to the user (PP = 78). If the user wants to choose a different cepstral point as the estimate of the pitch period, he can position a curser to a new value, and the program will find the maximum in a 5-point region around the indicated value (UP = 78).

For each display the program indicates something about the absolute signal level or the value at the estimated pitch peak. For the waveform display, the program indicates the maximum absolute signal level (MAX = 9772) relative to the peak possible

level of 32 767. This level gives a good indication of the amount of signal energy, and is helpful in making a decision as to whether a section of speech is voiced or unvoiced. For the autocorrelation function, and the cepstrum, the value of MAX on the displays is the value of the peak, relative to the value of the display at $m = 0$ which is arbitrarily set to a value of 100. Such indications serve as a measure of how good the pitch estimates are.

The final decision as to the pitch period for the current frame is made by the user based on the three estimates already made. The program indicates the median of the three pitch period estimates, and gives the user the option to change this value to any desired one. In the example of Fig. 3, all three estimates were identical (a period of 78 samples) and thus no changes were required. For unvoiced speech, the user changes the pitch period to a value of 0 and proceeds with the next frame.

Generally, a waveform display program is used prior to the SAPD program to give a gross indication of the regions of unvoiced and voiced speech. Fig. 4 shows a typical section of 1024 samples at the beginning of an utterance. A gross indication of the point at which voicing begins is shown on this figure.

III. Examples of SAPD Analysis

Figs. 5-8 show typical examples of the types of behavior which are encountered using the SAPD program. In Fig. 5 the section of speech is at a transition from unvoiced-to-voiced speech. In this case about half the 400 samples of the current frame are unvoiced and half are voiced. Thus the autocorrelation and cepstrum indications of the pitch period are greatly in error (as further evidenced by the low values of the peaks), whereas the best estimate of the pitch period is directly from the low-pass filtered waveform in this case. Fig. 6 shows an example in which the waveform phase is changing within a voiced section of speech, in which case the waveform measurement is inaccurate, whereas the autocorrelation is more accurate, and even the cepstrum gives a fairly good pitch estimate.

Fig. 7 shows an example in which the speech section is undergoing a voiced-to-unvoiced transition. In this case each of the pitch period estimates are widely disparate with each other, and the final pitch period estimate is set to zero, indicating unvoiced speech. Finally, Fig. 8 shows a section of unvoiced speech in which all three estimates are indicative of unvoiced speech.

The SAPD analysis is performed every 10 ms, giving 100 estimates of pitch per second. Fig. 9 shows a typical analysis of an entire utterance. In this case the utterance was "every salt breeze comes from the sea" spoken by a male speaker. The upper plot in Fig. 9 shows the pitch estimates obtained directly from the analysis, and the lower plot shows a nonlinearly smoothed [6] version of the pitch contour.

IV. Error Analysis of SAPD Results

The SAPD analysis was intended to be used in a comparative study of several pitch detection algorithms [1]. As such, it was required to provide the most reliable estimates of pitch period which could be obtained in order to be used as the standard of

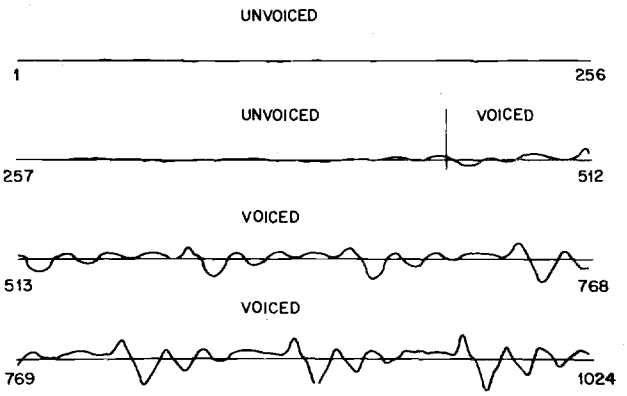

Fig. 4. Typical frame of 1024 samples of the low-pass filtered waveform.

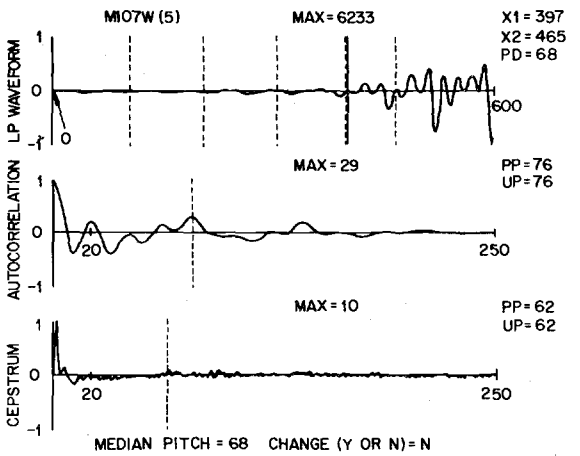

Fig. 5. Typical frame from the SAPD analysis during an unvoiced-to-voiced transition.

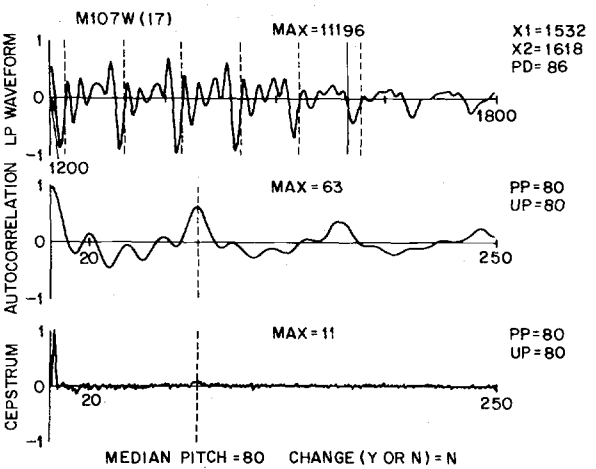

Fig. 6. Typical frame from the SAPD analysis during a sharp voiced transition.

comparison for the study mentioned above. To verify how reliable and robust the pitch estimates obtained from the SAPD analysis were, four users were given the same set of three utterances, and were required to do the entire analysis on each utterance.

The utterances were the sentence, "every salt breeze comes from the sea," spoken by a female, a male, and a low-pitched male. Table I shows part of the four analyses of the female utterance. It can be seen that each of the users obtained re-

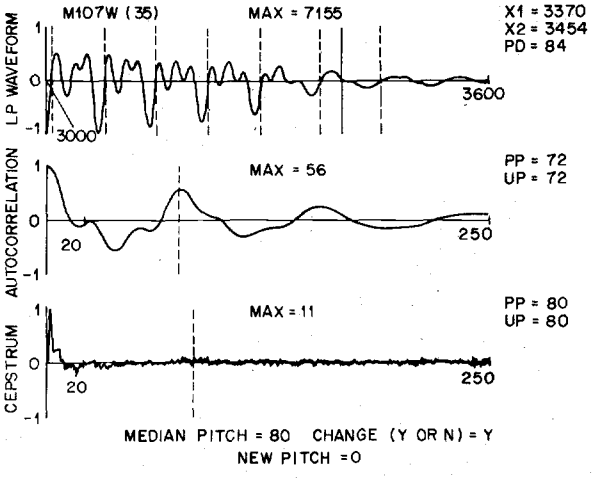

Fig. 7. Typical frame from the SAPD analysis during a voiced-to-unvoiced transition.

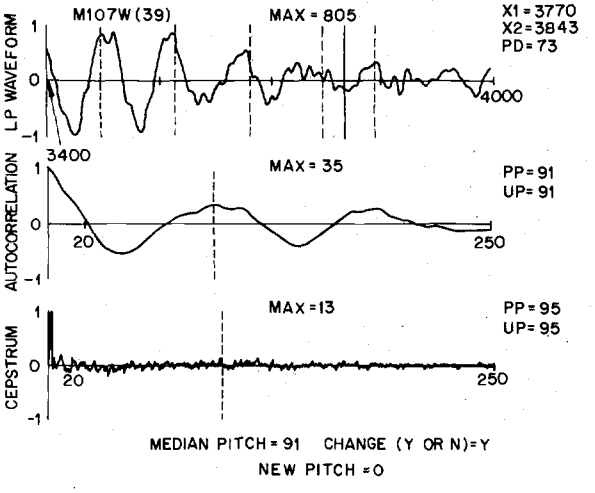

Fig. 8. Typical frame from the SAPD analysis during an unvoiced section.

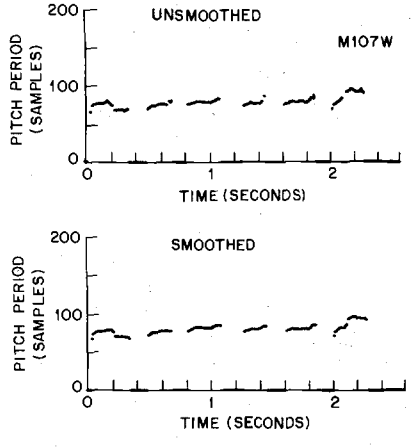

Fig. 9. Composite pitch period contours (both unsmoothed and smoothed) for an utterance analyzed using the SAPD method.

sults which were almost identical to those of the other users. Basically, there were two types of discrepancy in the detailed analysis of each of the users. These were the following.

1) During voiced regions there were some small differences in the estimated pitch periods.

TABLE I
TYPICAL EXAMPLE OF SAPD RESULTS FOR FOUR USERS

Female Speaker—F 107 M

	Frame Number									
User	1	2	3	4	5	6	7	8	9	10
MC	0	0	0	42	43	43	43	43	43	43
CM	0	0	0	43	43	43	43	43	43	43
LR	0	0	0	42	43	43	43	43	43	43
AR	0	0	0	42	43	43	43	43	43	43
Difference	0	0	0	1	0	0	0	0	0	0
	11	12	13	14	15	16	17	18	19	20
MC	43	44	44	45	44	44	43	42	40	40
CM	43	44	44	45	44	43	43	42	40	40
LR	43	43	44	45	44	43	43	42	40	40
AR	43	44	44	45	44	43	43	42	40	40
Difference	0	1	0	0	0	1	0	0	0	0
	21	22	23	24	25	26	27	28	29	30
MC	39	39	39	39	39	40	40	40	0	0
CM	39	39	39	39	39	40	40	0	0	0
LR	39	39	39	39	39	39	40	42	0	0
AR	39	39	39	38	39	40	40	40	48	0
Difference	0	0	0	1	0	1	0	42	48	0

TABLE II
VARIANCE OF VOICED-UNVOICED ONSET AND PITCH PERIOD FOR THREE UTTERANCES

	Female Speaker	Male Speaker	Low-Pitch Male Speaker
$\sigma_{v/u}$	0.4	0.5	0.4
σ_p	0.6	0.12	1.4 (0.23)[a]

[a]Unusually large discrepancy at the onset of one-voiced region disregarded.

2) There were some small differences in the estimated point at which voiced-unvoiced transitions occurred.

A simple statistical analysis was performed to obtain an idea as to the variance of the pitch estimates and the variance of the voiced-unvoiced switching time. (The average error was assumed to be zero across the four users.) Table II shows the results obtained for the three utterances. The variance of the voiced-unvoiced onset time was on the order of 0.5 intervals, or about 5 ms, across all three speakers. Thus the four users agreed on the location of the voiced or unvoiced onset to within one interval most of the time. The variance of the pitch period varied from 0.6 samples for the female speaker, to 0.12 samples for the male speaker, to 1.4 samples for the low-pitched male. The relatively large variance for the low-pitched male was due primarily to an unusual voicing onset for one voiced region in which a large discrepancy existed in the pitch period estimates. Without this large error, the variance was 0.23 samples for the remainder of the utterance. Thus Table II gives a strong indication that the SAPD technique yields reliable, and fairly robust estimates of the pitch period across a wide range of utterances and speakers.

One final note is worth making. This concerns the time required to perform the SAPD analysis. An experienced user required about 30 min to process 1 s (100 frames) of speech. Thus this technique is of value only if the amount of speech

to be analyzed is limited. In the comparative study of pitch detectors [1], a total of 56 utterances lasting about 80 s were analyzed, or a total of about 40 h on a computer were required.

In summary, the SAPD program is proposed as an alternative to simple waveform measurements for accurate and repeatable estimates of the pitch period for applications where extremely accurate analysis results are required.

REFERENCES

[1] M. Cheng, "A comparative study of several pitch detection algorithms," M.S. thesis, Dep. Elec. Eng., Mass. Inst. Technol., Cambridge, June 1975.

[2] H. Levitt and L. R. Rabiner, "Analysis of fundamental frequency contours," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 569–582, 1971.

[3] C. H. Coker, N. Umeda, and C. P. Browman, "Automatic synthesis from ordinary English," in *Proc. 1972 Conf. Speech Communication and Processing*, Newton, Mass., pp. 30–33.

[4] J. P. Olive, "Fundamental frequency rules for the synthesis of simple declarative English sentences," *J. Acoust. Soc. Amer.*, vol. 57, pp. 476–482, Feb. 1975.

[5] A. J. Goldberg, "Voiced speech in the absence of the Lanyngeal fundamental," M.S. thesis, Dep. Elec. Eng., Mass. Inst. Technol., Cambridge, 1967.

[6] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," this issue, pp. 552–557.

Numerical Determination of the Lip Impedance and Vocal Tract Area Functions

HISASHI WAKITA AND AUGUSTINE H. GRAY, JR., MEMBER, IEEE

Abstract—Two new methods based on a nonuniform acoustic tube model of the vocal tract are developed: a method for estimating the lip impedance zeros and poles directly from the speech wave and a method for computing the area function noniteratively from given lip impedance singularities.

I. INTRODUCTION

THIS PAPER describes two new methods which can be derived from an acoustic tube model of the vocal tract. The first is a method for estimating the lip impedance zeros and poles directly from the acoustic speech wave. The second is a simple noniterative method for determining a unique vocal tract area function from given lip impedance zero and pole frequencies.

There has been considerable interest in the problem of determining a unique vocal tract shape directly from acoustical measurements to avoid the drawbacks of X ray measurements, such as the limited safe dosage and the laborious task of processing data. In this context, Mermelstein and Schroeder [1], [2] first presented the problem as a first-order perturbation analysis of the Webster horn equation. The work was then extended by Mermelstein [3] and by Heinz [4]. Mermelstein presented the problem as an inverse eigenvalue solution for nearly uniform tracts, based upon the fact that two sets of eigenvalues for different boundary conditions give a unique area function. He chose the lip impedance zero and pole frequencies as two sets of eigenvalues and found a method for determining the area function by applying the perturbation theory iteratively so that the first few lip impedance singularities matched the measured values of these singularities. In the Schroeder-Mermelstein method, an impedance tube was employed to measure the lip impedance zero and pole frequencies.

In the meantime, another possibility for computing the vocal tract area function from acoustic speech waveforms has been investigated [5]. Following this line of research, Wakita finally showed [6] that a digital filter representation of a nonuniform acoustic tube model is identical to that of the linear prediction model of speech production if the boundary conditions of the acoustic tube model are properly chosen. Thus he succeeded in extracting reasonable vocal tract area functions directly from acoustic speech waveforms.

By use of the same acoustic tube model as in Wakita's method [6], it will be shown theoretically below that the discrete case of the Schroeder-Mermelstein method is equivalent to the completely lossless case of Wakita's method. Thus, by assuming a closed glottis condition, the zero and pole frequencies of the lip impedance needed in the Schroeder-Mermelstein method can be derived from speech waves by use of Wakita's model. Further, it will be shown that there exists a noniterative procedure to compute a unique discrete area function from those zero and pole frequencies of the lip impedance. Consequently, there are several alternatives in applying the Schroeder-Mermelstein method to estimate the vocal tract area functions. First, in determining the zero and pole

Manuscript received August 19, 1973; revised March 5, 1974 and March 20, 1975. This work was supported by the Office of Naval Research under Contract N00014-67-C-0118.

H. Wakita is with the Speech Communications Research Laboratory, Santa Barbara, Calif. 93109.

A. H. Gray, Jr. is with the Speech Communications Research Laboratory, Santa Barbara, Calif. 93109 and the Department of Electrical Engineering and Computer Science, University of California, Santa Barbara, Calif. 93106.