A High Quality Speech Analysis/Synthesis System

by

John C. Hardwick

Submitted in Partial Fulfillment
of the Requirements for the
Degree of

Bachelor of Science
in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology
May 1986

© John C. Hardwick 1986

The author hereby grants to M.I.T. permission to reproduce and
to distribute copies of this thesis document in whole or in part.

Signature of Author _____
Department of Electrical Engineering and Computer Science
May 15, 1986

Certified by _____
Jae S. Lim
Thesis Supervisor

Accepted by _____
David Adler
Chairman, Department Committee

MASS. INST. TECH.
JUN 18 1986
LIBRARIES

Archives

A High Quality Speech Analysis/Synthesis System

by

John C. Hardwick

Submitted to the

Department of Electrical Engineering and Computer Science

on May 19, 1986 in partial fulfillment of the requirements

for the Degree of

Bachelor of Science

Abstract

In this thesis a high quality speech analysis/synthesis system is presented. The system functions by comparison of a speech signal to an underlying model. The model parameters are then extracted and used in the synthesis of the original signal. In order to achieve high quality speech a new model is used in this process. Each speech segment is characterized in the frequency domain by the product of an excitation spectrum and a spectral envelope. The spectral envelope is taken to be a smoothed version of the actual speech spectrum. The excitation spectrum is specified by a fundamental frequency, or pitch period, and a voiced/unvoiced decision for each harmonic of the fundamental. This model is much more flexible than previous models. As a result synthetic speech can be produced with much fewer artifacts. This is especially true of noisy speech, where the "buzziness" typically associated with it has been eliminated. One of the principal applications which is being explored for this system is in speech coding. An 8000 bits/sec system has already been developed and future work is being directed at further reductions in the bit rate. As a conclusion several means of accomplishing this are presented.

Thesis Supervisor: Jae S. Lim

Title: Associate Professor of Electrical Engineering

Dedication

To my parents,

Willis and Jean Hardwick,

whose kind and loving support

have made this possible.

Acknowledgments

I would like to thank all of the people who have helped me both directly and indirectly during the completion of this thesis.

In particular I would like to thank Dan Griffin for his time and patience during the past several months. It is to his credit that these concepts were ever originated, and his words of wisdom provided the answer to many a question. In addition I would like to thank my thesis advisor, Professor Jae Lim, for his guidance and comments during the course of this thesis. The entire membership of the Digital Signal Processing Group here at MIT deserves praise for their patience and helpfulness. The ability to learn in a friendly environment is a feature which cannot be overly appreciated.

The people who have helped me indirectly also deserve attention. This is particularly true of my family, which has been extremely supportive and attentive. The education which I have recieved is a result of their sacrifice, and my appreciation and love will always be with them. I am also thankful for the friendships which have made my education here a truly rewarding experience. The members of the fraternity of Phi Gamma Delta are prominant among this group. Friendship is truly the sweetest influence.

Contents

1 Introduction **7**

1.1 Motivation 7

1.2 Background 8

2 New Speech Model **11**

3 Speech Analysis **14**

3.1 Pitch Estimation 15

3.2 Voiced/Unvoiced Determination 19

4 Speech Synthesis **22**

4.1 Voiced Synthesis 22

4.2 Unvoiced Speech Synthesis 26

5 Conclusion **27**

5.1 Performance 27

5.2 Future Developments 29

List of Figures

1 Vocoder Block Diagram 10

2a Original Speech Spectrum 13

2b Spectral Envelope 13

2c Periodic Spectrum 13

2d Voiced/Unvoiced Information 13

2e Noise Spectrum 13

2f Synthetic Speech Spectrum 13

3 Error vs. Fundamental Frequency 18

4a Original Speech Spectrum 20

4b Synthetic Speech Spectrum 20

4c Comparison of Original and Synthetic Speech 20

Chapter 1

Introduction

1.1 Motivation

The analysis and synthesis of speech is an area which has recieved a considerable amount of interest. This is due primarily to the large number of applications for analysis/synthesis systems (vocoders), and the need for improvement in existing systems. A vocoder functions by comparing speech to an underlying model and extracting out the essential parameters of the model. These parameters are then stored, transmitted, etc. and finally recombined in a manner which will approximate the original speech. In most major vocoder applications it is desirable to minimize the number of parameters which must be extracted. An example of this lies in the coding of speech for either transmission or storage. Speech is generally regarded to have an effective bandwidth of 4 kHz. If speech is sampled at the Nyquist frequency of 8 kHz and quantized using 8 bit samples then 64 Kbits/sec are required. Significant reductions in this can be accomplished through the use of a vocoder. The number of bits which are needed to characterize the model parameters is considerably less than those for the actual samples. This results in less memory for storage purposes, or lower bandwidth requirements for a communication channel. The exact amount of improvement which is gained is dependent upon the required

speech quality and the particular vocoder which is being used. Reduction of the bit rate past a certain point will cause intolerable defects in any vocoder system. In addition performance is ultimately restricted by the model on which it is based. For this reason a model must be used which corresponds very closely to actual speech. In addition it must be flexible enough to operate in realistic environments in the presence of noise. Attainment of these features will distinguish between vocoders capable of high quality speech and those which are intelligible but produce noticable degradations in quality.

1.2 Background

The model on which most current vocoders are based views speech as the response of a time varying linear filter to some excitation sequence. Since speech is not a stationary signal it is necessary to analyze speech over time intervals which are small compared to the time rate of change of the linear filter. This is accomplished by multiplying the speech signal by a window which is zero outside of some designated interval. The length of this interval is chosen such that the linear filter remains relatively constant over its duration. By shifting the location of the non-zero portion of the window over the speech signal, it is possible to segment the entire signal. The vocoder must then analyze each segment and extract the parameters which characterize the linear filter and the excitation sequence for that segment. This is facilitated by dividing speech segments into two classes depending on the nature of the excitation sequence. In voiced speech the excitation is modeled as a periodic impulse train, while in unvoiced speech a white noise sequence is used as the excitation. Using this distinction the excitation parameters for each segment reduce to a voiced/unvoiced decision and the pitch period for voiced speech. The synthesis system uses these excitation parameters to generate either a white noise sequence for unvoiced speech or an impulse train with the desired period for voiced speech.

This sequence is then passed through a system which is specified by the linear filter parameters, and the output is taken as the synthesized speech. In Figure 1 an overview of this process is shown. The workings of a vocoder can essentially be broken into several functional blocks. The actual workings of the blocks will be discussed in the later sections of this thesis.

Vocoders based on the model described above vary primarily by the manner in which they extract the model parameters. Different techniques are exemplified by linear prediction vocoders, homomorphic decoders and channel vocoders. While some techniques out-perform others due to their superior ability to estimate the model parameters, they are all limited by the validity of the underlying model. Although this model has proved sufficient to allow the synthesis of intelligible speech, it has shown itself unable to produce high quality speech. In an attempt to circumvent this limitation, considerable work has been done to improve the correlation between the model and actual speech. In [1] a new model was presented in which speech is no longer viewed as either voiced or unvoiced. Instead each speech segment is modeled as partially voiced and partially unvoiced. This allows for a more flexible excitation sequence, and consequently, higher quality speech can be generated. In addition a new pitch detection algorithm has been developed in [2] which offers superior performance in certain cases. By combining these techniques a more robust and accurate vocoder can be generated. The improvements made have a significant effect on the quality of the synthesized speech, and they open up a number of applications requiring high quality speech.

This thesis focuses on the theory and implementation of a high quality speech analysis/synthesis system. Although one implementation currently exists, this work was done in order to familiarize myself with the system and to facilitate future research in the area. There is currently an interest in using the techniques presented here in a speech coding application. An 8000 bps. system has already been demonstrated, and lower bit rates are thought possible. This topic, however, is beyond

the scope of this thesis and will be the subject of future research.

In the next chapter the new speech model is described. Later chapters deal with the use of this model in the analysis/synthesis system. Specifically, chapter 3 discusses speech analysis, while chapter 4 is dedicated to the synthesis of speech from the model parameters. In chapter 5 the results are presented for both clean and noisy speech. Also future research areas are described in this chapter.

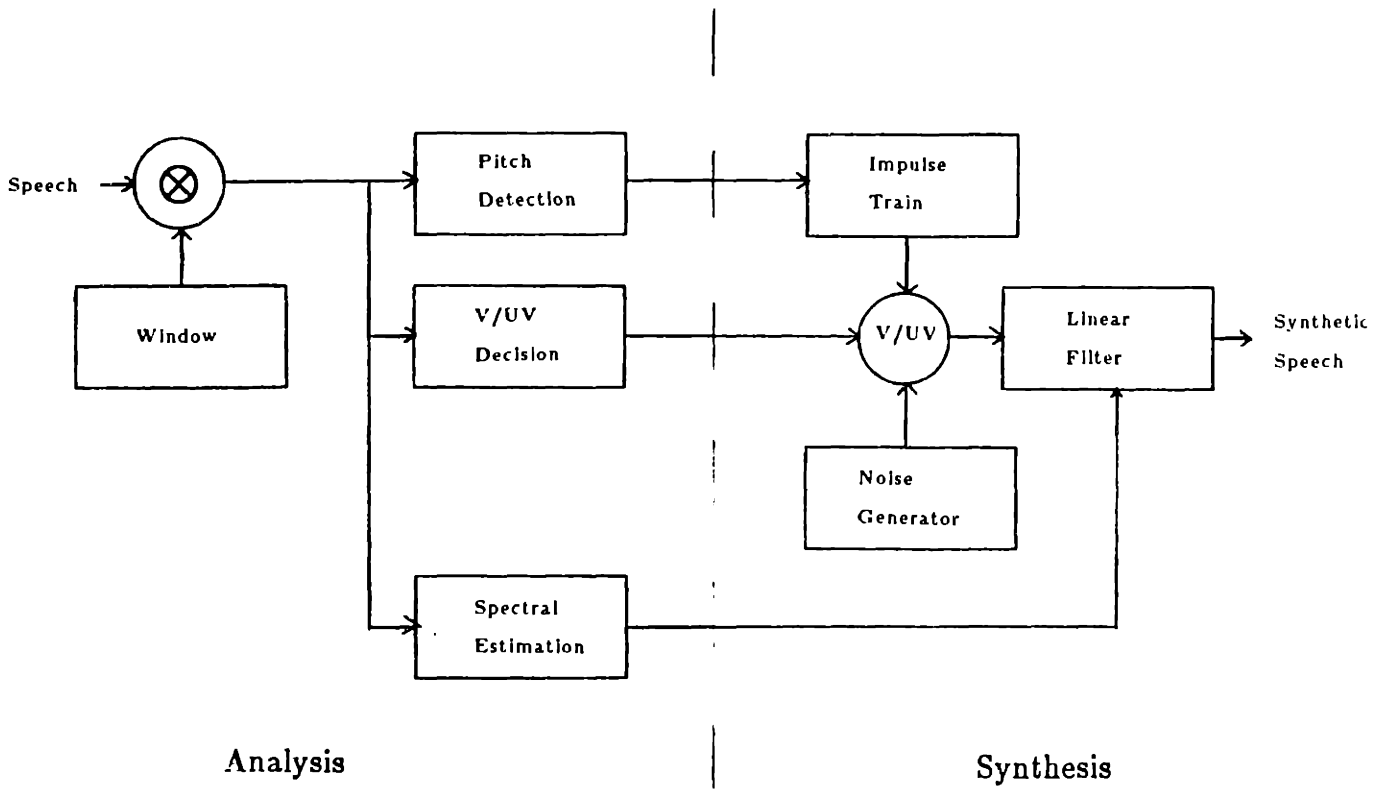

Figure 1 Vocoder Block Diagram

Chapter 2

New Speech Model

The model which is used is based around the spectral properties of speech. The Fourier Transform $S_w(\omega)$ of a windowed speech segment $s_w(n)$ can be modeled as the product of an excitation spectrum $E_w(\omega)$ and a spectral envelope $H_w(\omega)$. The primary difference between this model and previous ones lies in the form of the excitation spectrum. In previous models this was entirely specified by the pitch period n_o and a voiced/unvoiced decision for the entire speech segment. For voiced speech $E_w(\omega)$ was equal to $P_w(\omega)$, the Fourier Transform of a windowed impulse train with spacing equal to n_o. If the effects of aliasing are ignored, this signal corresponds to the Transform of the window centered at each harmonic of $w_o = 2\pi/n_o$. Speech segments which did not exhibit this periodic property were declared unvoiced. $E_w(\omega)$ for these regions of the speech signal was modeled as the spectrum of a white noise sequence. This approach allowed only two different excitation sequences which presented a serious limitation in the model.

The extension to the model which is the subject of this thesis is to remove the well defined voiced/unvoiced distinction in favor of a more continuous division. Instead of making a single voiced/unvoiced decision for the speech segment, the new model breaks the excitation spectrum into voiced regions and unvoiced regions. For each harmonic of the fundamental frequency, w_o, a voiced/unvoiced decision is made.

The synthetic speech is created by using this information to selectively combine the two speech forms. Over harmonics which are declared voiced $P_w(\omega)$ is used in the excitation spectrum. The remaining frequency regions are then filled with noise energy.

The use of many voiced/unvoiced decisions allows the new model to have fine control over the makeup of the excitation spectrum. Speech segments which are completely noise dominated or completely periodic can still be accomodated, and in addition segments which are a mix of the two can now be handled. This approach was motivated by the observation that many speech segments have some frequency regions which are dominated by noise energy while other regions are periodic. This is especially the case in speech which has been degraded by the addition of noise. The flexibility provided by this model allows it to handle these cases in a manner which is superior to its predecessors.

The spectrum of each speech segment is only partially determined by the excitation spectrum. In addition the spectral envelope $H_w(\omega)$ determines the relative amplitude of each frequency component. Since the excitation spectrum is assumed to have a constant amplitude, the spectral envelope provides the scaling between $E_w(\omega)$ and the actual speech spectrum. In this manner $H_w(\omega)$ can be viewed as the frequency response which will map $E_w(\omega)$ into $S_w(\omega)$. Since $H_w(\omega)$ is generally slowly varying, it is often assumed to be a smoothed version of the actual speech spectrum $S_w(\omega)$.

A demonstration of the new model is shown in Figure 2. In Figure 2a the Spectrum of a typical speech segment is shown. This was obtained by windowing the speech signal with a 256 point Hamming window and then calculating the DFT of that sequence. Figure 2b shows the spectral envelope which has been found for the segment. One can see that this is a smooth contour containing the general shape of the spectrum shown in 2a. The pitch period which has been estimated for this segment is 78 samples at a 10 kHz sampling rate. $P_w(\omega)$ corresponding to this

pitch period is shown in Figure 2c. The voiced/unvoiced information is displayed in Figure 2d. A high value on this graph corresponds to a voiced harmonic and in this region of the spectrum $P_w(\omega)$ would be used in the excitation spectrum. Frequency regions having a low value in Figure 2d are unvoiced and noise energy as shown in Figure 2e is used in the excitation spectrum. This combination of voiced and unvoiced spectra is then multiplied by the spectral envelope to create the synthetic speech. This product is shown in Figure 2f.

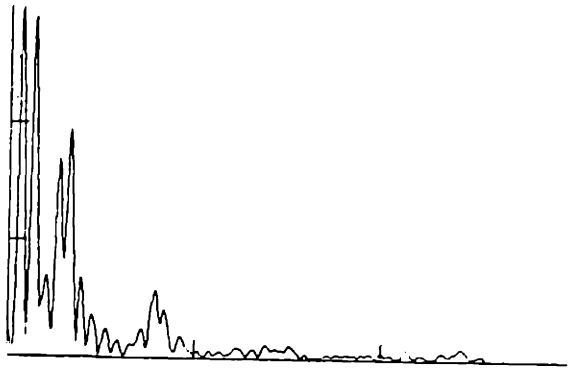

Figure 2a Original Speech Spectrum

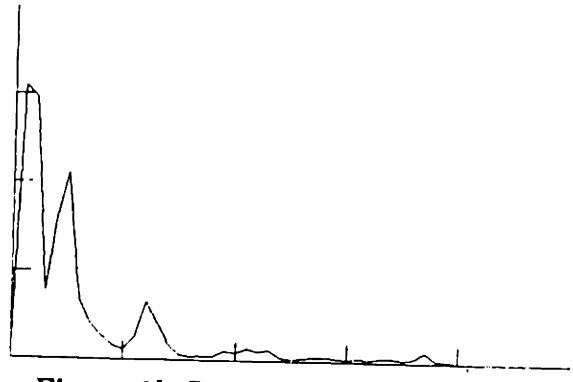

Figure 2b Spectral Envelope

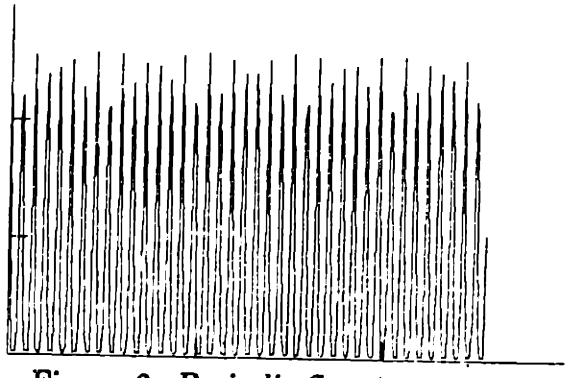

Figure 2c Periodic Spectrum

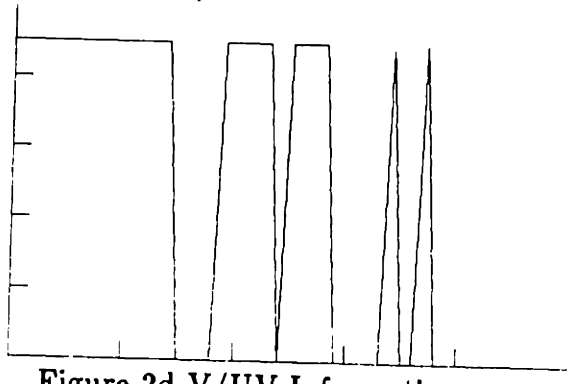

Figure 2d V/UV Information

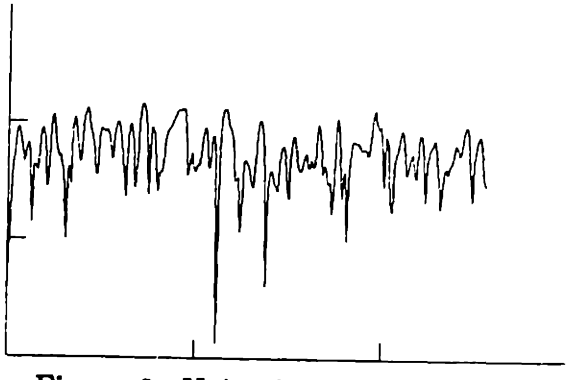

Figure 2e Noise Spectrum

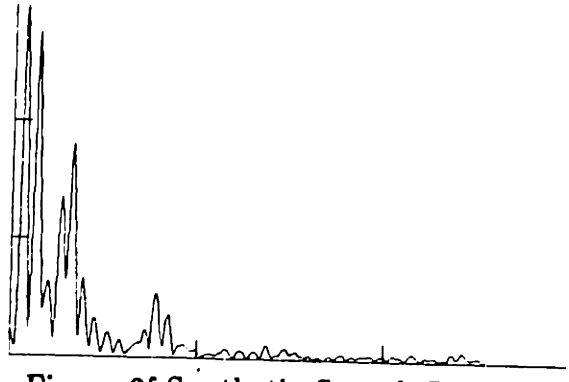

Figure 2f Synthetic Speech Spectrum

Chapter 3

Speech Analysis

The production of high quality speech is not only linked to an improved model, but is also related to the algorithm used to extract the model parameters. The technique which was used to analyze the speech differed from previous techniques in an important way. Instead of trying to extract the excitation sequence and the spectral envelope separately, these parameters are found simultaneously. This approach allows the error between the original and synthetic speech spectra to be minimized in a least squared sense. This ensures the closest possible fit which can be accommodated by the model.

In general each speech segment can be either voiced, unvoiced or a mixture of the two. For unvoiced speech, the pitch period or fundamental frequency is not relevant, since the excitation spectrum is the transform of a white noise sequence. However, since most segments will be at least partially voiced, a fundamental freqency will be needed to characterize the periodic portions of the excitation spectrum. In light of this, the approach which is taken is to first assume the segment to be entirely voiced. The fundamental frequency and spectral envelope are then found which minimize the difference between the original speech spectrum and the synthetic version. The result of this operation will produce small differences over voiced regions of the spectrum and large differences in the unvoiced regions. At this point

the spectrum is broken up into regions which correspond to the harmonics of the chosen fundamental frequency. The error is evaluated within each of these areas and is used as the basis for a voiced/unvoiced decision. The model parameters for voiced areas have already been determined through the error minimization process. For unvoiced regions the spectral envelope is taken to be the average value of the original speech segment evaluated over that region.

3.1 Pitch Estimation

In order to minimize the error difference between the synthetic and original speech, it was necessary to formulate the error as a function of the pitch period and the spectral envelope. In general the error between the two can be expressed as:

$$E = \frac{1}{2\pi}\int_{w=-\pi}^{\pi} G(\omega)|S_w(\omega) - \bar{S}_w(\omega)|^2 d\omega \tag{3.1}$$

where $\bar{S}_w(\omega)$ is the synthetic speech and $G(\omega)$ is a frequency dependent weighting function. To use this equation it must be rewritten in terms of the model parameters which are desired. Since $\bar{S}_w(\omega)$ is equal to the product of $H_w(\omega)$ and $E_w(\omega)$, (3.1) is identical to the following:

$$E = \frac{1}{2\pi}\int_{w=-\pi}^{\pi} G(\omega)|S_w(\omega) - E_w(\omega)H_w(\omega)|^2 d\omega \tag{3.2}$$

At this point in the analysis, the assumption that the speech segment is voiced is used to gain a simpler form of (3.2). For voiced speech $E_w(\omega) = P_w(\omega)$, which is the Fourier Transform of a windowed impulse train. In this case the spectral envelope can be reduced to a complex scale factor for each harmonic of $P_w(\omega)$. This set of scale factors is determined by error minimization around the region occupied by each harmonic. In this manner the error around the mth harmonic can be expressed as the following:

$$E_m = \frac{1}{2\pi}\int_{w=a_m}^{b_m} G(\omega)|S_w(\omega) - A_m P_w(\omega)|^2 d\omega \tag{3.3}$$

where A_m is the complex scale factor associated with the mth harmonic. The range of the integral is set such that the region goes from halfway to the previous harmonic to halfway to the next. In other words $a_m = (m - .5)\omega$ and $b_m = (m + .5)\omega$. By differentiating (3.3) with respect to A_m it is possible to the find the scale factor which minimizes the error over the region $[a_m, b_m]$. This leads to the following:

$$Re\{A_m\} = \frac{\frac{1}{2\pi}\int_{w=a_m}^{b_m} G(\omega)[S_w^*(\omega)P_w(\omega) + S_w(\omega)P_w^*(\omega)]d\omega}{\frac{1}{\pi}\int_{w=a_m}^{b_m} G(\omega)P_w(\omega)P_w^*(\omega)d\omega} \tag{3.4}$$

$$Im\{A_m\} = \frac{\frac{1}{2\pi}\int_{w=a_m}^{b_m} G(\omega)[S_w^*(\omega)P_w(\omega) - S_w(\omega)P_w^*(\omega)]d\omega}{\frac{1}{\pi}\int_{w=a_m}^{b_m} G(\omega)P_w(\omega)P_w^*(\omega)d\omega} i \tag{3.5}$$

Since a_m and b_m are dependent upon the choice of the fundamental frequency the values of A_m and E_m will change depending upon the value which is used. For a given w_o (3.4) and (3.5) can be used to generate the scale factors which minimize the error over a particular harmonic. This process can be extended to find the minimum total error as given by:

$$\bar{E}_T = \sum_m \bar{E}_m \tag{3.6}$$

where $\bar{E}_m$ is equal to (3.3) evaluated at the scale factor given by (3.4) and (3.5). In order to choose the best pitch period estimate it is necessary to calculate (3.6) for a wide range of fundamental frequencies. As this value is varied $P_w(\omega)$ is shifted in relation to the actual speech spectrum. The resulting error is calculated and used in determining the best choice of fundamental frequency. In this manner the original error minimization problem has been simplified. The need to minimize the error simultaneously against the fundamental frequency and the spectral envelope was eliminated. Instead the fundamental was assumed to be fixed and the error was minimized against just the spectral envelope. Repeating this step for a range of fundamental frequencies allows the minimum error associated with each to be evaluated. This measure is then used to determine the best fundamental.

Experiments evaluating the relationship between the fundamental frequency and the total error have shown that the error varies smoothly with w_o. In addition the pitch periods associated with speech are generally in the range of 30 to 150 sample at 10kHz. These two facts allow the algorithm to reduce the number of frequencies at which the error must be evaluated. A first pass evaluates the error versus fundamental frequency on a coarse grid. This allows a rough pitch estimate to be made, which is then sharpened by a second, more detailed, evaluation. The primary difference between these two passes is that the second uses a much smaller frequency deviation over a smaller range of frequencies. In addition the frequency weighting function is changed between the two. In the initial pass it is desirable to look at the entire spectrum, and $G(\omega)$ is therefore equal to unity. On the second pass, however, $G(\omega)$ is a high pass filter. This is because the effect of an error in the fundamental frequency is much more apparent in the higher harmonics, and the energy in the signal is often dominatd by the lower harmonics. A high pass filter increases the sensitivity of the error to w_o and thus improves the algorithm.

Another feature of the pitch period detector is that it must distinguish between the desired period and its harmonics. In general the error associated with the multiples of the correct pitch period will be comparable. Since maximum performance is dependent upon rejecting the harmonics, the shortest pitch period with comparable error is chosen. In Figure 3 an example of this problem is shown. Since the fundamental frequency is proportional to the inverse of the pitch period the algorithm should choose the highest fundamental frequency with comparable error. However in Figure 3 the error is slightly lower at the harmonics of the desired pitch period. The correct fundamental is obtained by noticing that the error is comparable between all three minima.

Throughout the analysis process the model parameters are assumed to remain constant over each speech segment. In addition the parameters generally change in a slow and continuous fashion from segment to segment. This concept is exploited

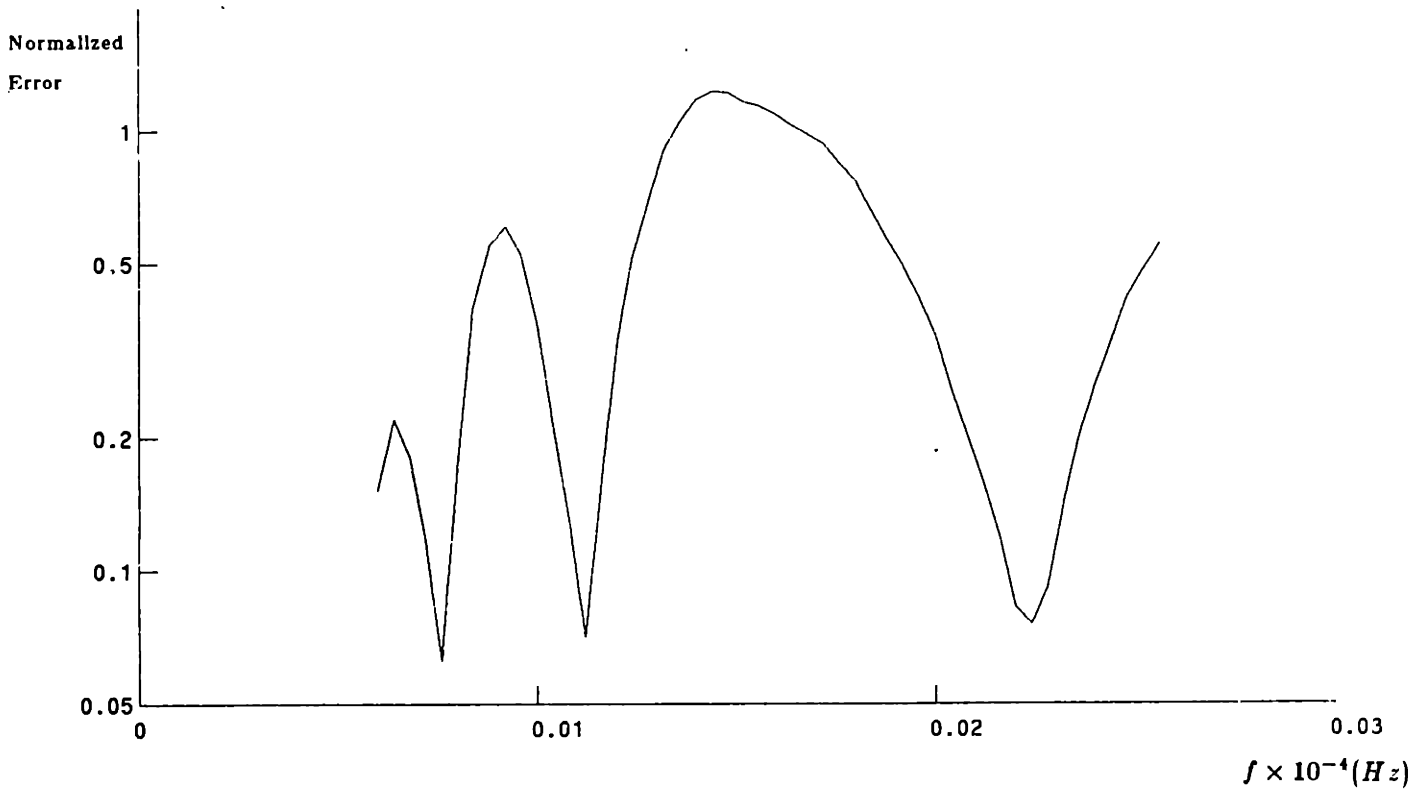

Figure 3 Error vs. Fundamental Frequency

in the pitch detection process by weighting each fundamental frequency by a distance measure. This measure favors frequencies which are in close proximity to the fundamental which was chosen by a neighboring segment. The algorithm first searches through the speech signal until a segment with a well defined pitch period is found. It then spreads to neighboring speech segments, using the previously found fundamental frequency in the distance measure. If a segment is reached which clearly cannot accomodate an estimate near its neighbor, the algorithm stops and looks for additional segments which vary more smoothly. These trouble segments are handled by declaring them unvoiced which is discussed in a later section.

A final procedure used in extracting the pitch period concerns the variation of the total error with pitch period over unvoiced segments. Experimentation has demonstrated a bias toward long pitch periods which is due to their spectral re-

semblence to white noise. Since long pitch periods correspond to closely spaced harmonics, the harmonic lobes will overlap. This causes $P_w(\omega)$ to flatten out and approach the spectrum of a white noise sequence. Since ideally, an unvoiced speech segment should not favor any single fundamental frequency, this effect must be corrected for. This is accomplished by experimentally determining the relationship between the error and pitch period for unvoiced speech. The total error for each speech segment is then correct for this to remove any unwanted bias.

A demonstration of the pitch estimation algorithm is given in Figure 4. Figure 4a shows the spectrum of a single speech segment. The pitch period of this segment is 50 samples at a 10 kHz sampling rate. In Figure 4b the synthesized spectrum is shown for the same segment. The accuracy of the pitch detector is clearly seen in Figure 4c, where the spectra of 4a and 4b are overlaid. Even out at the higher frequencies the mismatch is on the same order as the background noise. The lack of any noticable pitch error is due to the high resolution of the estimation algorithm.

3.2 Voiced/Unvoiced Determination

Once the pitch period has been determined the error over each harmonic of the fundamental can be evaluated from (3.3). This can be used to find the normalized error through the following:

$$\xi_m = \frac{\bar{E}_m}{\frac{1}{2\pi}\int_{w=a_m}^{b_m} G(\omega)|S_w(\omega)|^2 d\omega} \tag{3.7}$$

Comparing the normalized error to a predetermined threshold, a voiced/unvoiced decision is made. If the normalized error is less than the threshold then the harmonic is declared voiced. Similarly if it is greater than the threshold the harmonic is declared unvoiced. The value of the threshold is determined experimentally by listening to the synthetic speech. A value is chosen which will accomodate a wide range of speakers under varying levels of background noise. If it is set to high the

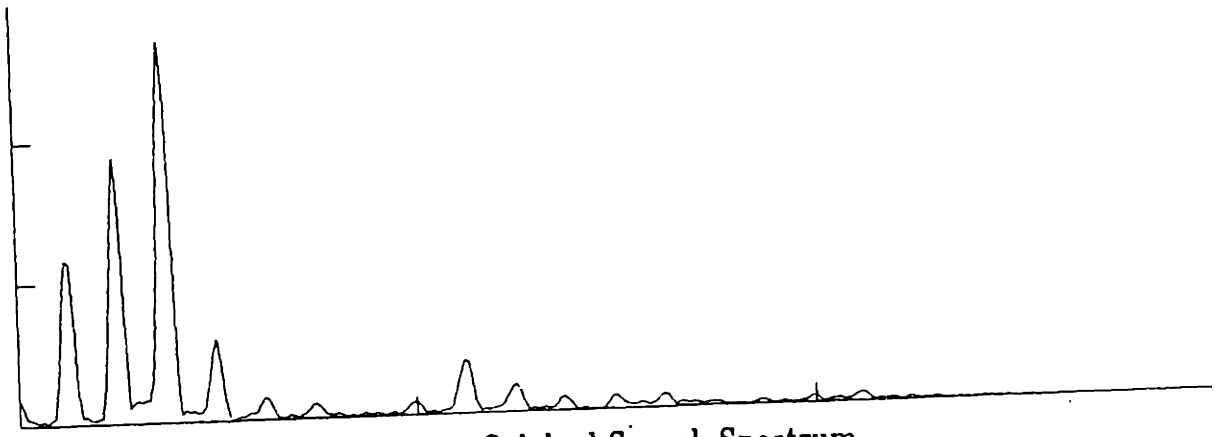

Figure 4a Original Speech Spectrum

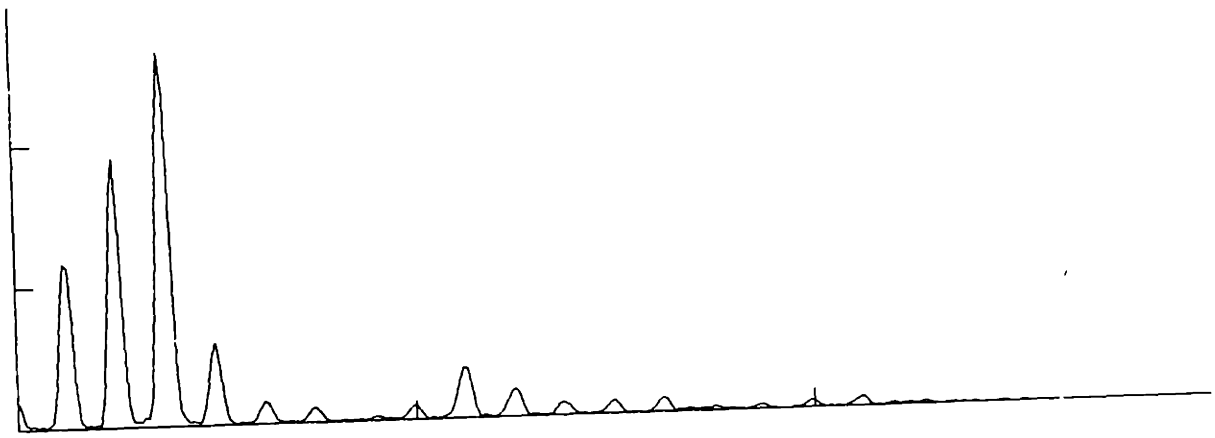

Figure 4b Synthetic Speech Spectrum

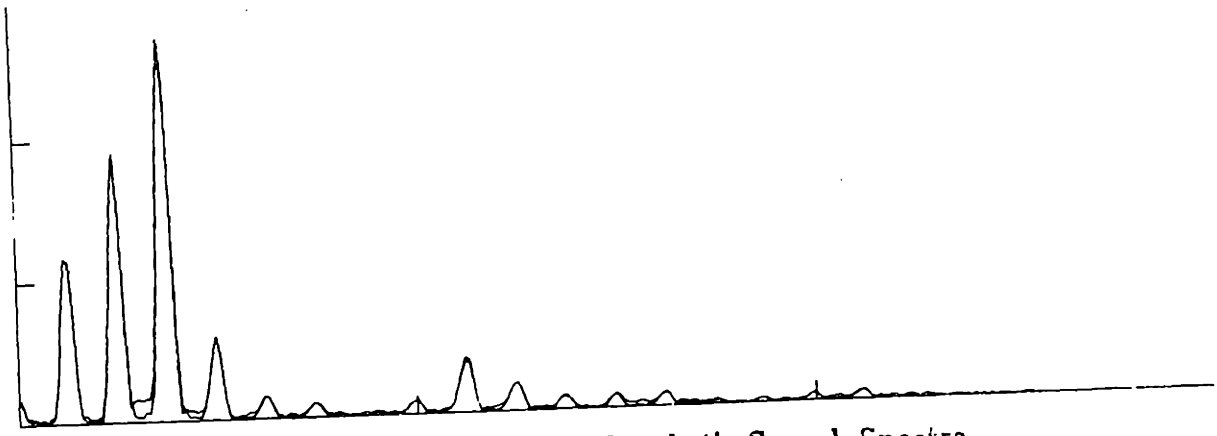

Figure 4c Original and Synthetic Speech Spectra

speech will have a "buzzy" sound, caused by the addition of undesired harmonics. On the other hand if the threshold is placed too low the speech will sound coarse, corresponding to too much noise energy in the signal.

In some circumstances it is desirable to declare an entire speech segment unvoiced. This is often the case during portions of speech which are highly noisy such as occurs during the space between words. The rationale behind this step is to completely remove the periodic nature of these speech segments. This ensures that the regions sound most natural, and do not exhibit any "buzziness".

For each harmonic region which is declared unvoiced, the value of the spectral envelope found by (3.4) and (3.5) is not appropriate. Since the portion of the spectrum is usually quite small it is fair to presume the spectral envelope remains nearly constant over this interval. By averaging the speech spectrum over this region the spectral envelope can be captured and used later in the synthesis procedure.

Chapter 4

Speech Synthesis

The problem of synthesizing the speech signal given the model parameters can be handled in several different ways. In particular two different approaches were tried during the course of this thesis, and the results from each were comparable. In both versions the synthesis of the voiced and unvoiced speech is handled separately. The final result is then obtained from a sum of the two. Producing the unvoiced speech is fairly straightforward, however, synthesizing the voiced speech raises a number of questions. The primary one involves how to account for the change in the model parameters from segment to segment. Since speech signals are continuous the synthesis procedure must smooth the transition between speech segments. If continuity is not maintained, unnatural sounds will be introduced which can be clearly discerned by the listener.

4.1 Voiced Synthesis

As discussed previously the analysis procedure extracts parameters for a windowed segment of the speech signal. These parameters consist of the fundamental frequency and information characterizing each harmonic of this frequency. This consists of the voiced/unvoiced decision and the value of the spectral envelope at the

harmonic frequency. Since the spectral envelope is a complex signal, it can be expressed in either rectangular or polar form. The synthesis procedure is oriented around the polar form, and, therefore, the magnitude and phase of each harmonic is used in the algorithm. Since voiced speech is modeled as being discrete in frequency the synthesis procedure can be implemented as a bank of tuned oscillators. For a particular speech segment, an oscillator is assigned to each harmonic which has been declared voiced. The output of the oscillator for the mth harmonic can be expressed as:

$$\bar{s}_{v_m}(t) = A_m(t)\cos\theta_m(t) \tag{4.1}$$

In order to ensure continuity between speech segments, the amplitude function $A_m(t)$ is a linear interpolation of model parameters found for the two segments. If the time between speech segments corresponds to $t = S$ then the amplitude function is given by:

$$A_m(t) = A_m(0) + [A_m(S) - A_m(0)]\frac{t}{S} \tag{4.2}$$

where $A_m(0)$ is the value of the spectral envelope for the speech segment occuring at $t = 0$, and $A_m(S)$ is the value for the corresponding harmonic at the $t = S$ segment. If the m_{th} harmonic for either of the two segments was declared unvoiced then its associated component in (4.2) is set to zero. The technique used to generate the unvoiced speech uses a similar procedure. This process ensures a smooth transition as a harmonic is changed from voiced to unvoiced between neighboring frames. For the case when the harmonic is unvoiced during both segments, $A_m(t)$ is set equal to zero. This region of the spectrum is then handled by the unvoiced synthesis routine.

The phase function for the m_{th} harmonic, $\theta_m(t)$, is determined by the model parameters in a manner very similar to the amplitude function. In general the phase function between two speech frames can be expanded to:

$$\theta_m(t) = \int_0^t \omega_m(\xi)d\xi + \phi_m \tag{4.3}$$

This equation involves determining two parameters which are an initial phase and a frequency track. At $t = 0$ or $t = S$, the frequency is defined as the m_{th} harmonic of

the fundamental frequency which was chosen for that segment. Since the frequency at these two points may not be identical, it is desirable to interpolate the frequency in between. This can be accomplished through the following:

$$\omega_m(t) = m\omega_o(0) + m[\omega_o(S) - \omega_o(0)]\frac{t}{S} + \Delta\omega_m \tag{4.4}$$

In order to match (4.3) with the known phase value at $t = 0$ and $t = S$, the variables ϕ_m and $\Delta\omega_m$ must be properly chosen. If the m_{th} harmonic has been declared voiced for both speech segments, then the initial phase, ϕ_m, is set equal to the measured phase at the segment corresponding to $t = 0$. The frequency deviation, $\Delta\omega_m$, then corresponds to the smallest value which will result in $\theta_m(t)$ having the phase which was measured at $t = S$. If the harmonic at either frame was declared unvoiced then ϕ_m is set equal to the measured phase of the voiced segment. Since the phase of the unvoiced segment is irrelevant the frequency deviation is set to zero.

Once the oscillator parameters, $A_m(t)$ and $\theta_m(t)$, have been calculated for each harmonic, the voiced portion of the speech signal between $t = 0$ and $t = S$ can be formed. This is accomplished by a simple summation over m, or symbolically,

$$\bar{s}_v(t) = \sum_m A_m(t)\cos\theta_m(t) \tag{4.5}$$

This result is then ready for the addition of the unvoiced speech which will complete the synthesis procedure.

The net result of this process is that for each voiced harmonic the amplitude, frequency and phase correspond to the model parameters at the point in the signal at which they were extracted. In addition, these parameters are constrained to move smoothly from the beginning of one segment to the next. If the segment offset, S, is small compared with the window length then the effect of parameter changes is negligible, and interpolation only complicates the synthesis procedure. In this case the technique described below is advantageous from a computational point of view. However for large window shifts, as is desired in coding applications, the method presented above is more desirable.

A second approach to the synthesis of voiced speech is to use the weighted overlap add procedure. In this method an oscillator is assigned to each harmonic in the speech frame beginning at $t = 0$ and for the one at $t = S$. The magnitude, frequency and phase of each of these oscillators is simply determined by the model parameter which was found for the corresponding harmonic. The oscillators are then summed to produce the unvoiced speech over the length of each segment. However, since the length of each segment is longer than S there will be regions in which several frames overlap. In general the output of the oscillators will not be identical in these regions, and some form of averaging must be done. In [[4]] an averaging scheme was presented which was shown to minimize the error between the actual and the desired signals. This procedure is referred to as the weighted overlap-add method, and it can be expressed as:

$$s(n) = \frac{\sum_{f=-\infty}^{\infty} w(fS - n) y_w(fS, n)}{\sum_{f=-\infty}^{\infty} w^2(fS - n)} \tag{4.6}$$

For the application under discussion $y_w(fS, n)$ corresponds to the windowed output of the harmonic oscillators for the f_{th} speech segment. The term $w(fS - n)$ is the segmentation window shifted to the location of the f_{th} speech frame. It essentially blends the output of the oscillators by windowing them twice and then summing over the overlapping portions of neighboring segments. The factor in the denominator is to remove the scaling effect of the window multiplications. This procedure removes the discontinuities which would normally occur at segment boudaries. It accomplishes the smoothing at the segment level, however, rather than with the parameters themselves. As mentioned earlier this can be significant as the window shift is increased.

4.2 Unvoiced Speech Synthesis

The creation of the unvoiced speech is using the overlap-add procedure which is described in the previous section. A Gaussian noise sequence is generated and used in the synthesis algorithm. For each speech segment the corresponding portion of the noise sequence is obtained and its Fourier Transform is taken. For each harmonic that was declared voiced the corresponding region of the noise spectrum is zeroed out. In the remaining regions the average magnitude is set equal to the value which was obtained during the speech analysis. The phase in these regions is not modified and, therefore, corresponds to the phase of the original noise sequence. The inverse transform of this spectrum is set equal to $y_w(fS, n)$ in (4.6). The overlap-add procedure is then used to calculate the unvoiced speech signal in the same manner as is discussed above. Since this technique tapers each speech segment toward its edges, it ensures that harmonics which transition from voiced to unvoiced are not over emphasized. Addition of this result to the voiced speech completes the synthesis procedure. The final output is high quality speech without many of the artifacts typically associated with vocoders.

Chapter 5

Conclusion

5.1 Performance

The actual implementation of the analysis/synthesis system was completed in an iterative manner in an effort to converge on the desired performance level. The design process was essentially oriented around the proper choice of a large number of parameters which needed to be set empirically. Examples of these included the voiced/unvoiced threshold, the pitch period dependent correction factor, the frequency weighting function and various others. Experimentation showed that the analysis algorithms produced markedly different results depending upon the value of these parameters. The approach which was taken was to choose their values, test the system and then examine the results. This process allowed the effect of parameter variations to be seen in terms of speech quality, and further iterations could then be made to improve it. Eventually a system capable of producing high quality speech was established.

The estimation of the algorithmic parameters was accomplished through two major techniques. The first involved listening to and comparing the synthesized and actual speech. This measure provided the final word on questions of speech quality, and it was used regularly to determine the audible effect of certain pa-

rameter changes. Visual comparison of the actual and synthetic speech was also used as a design aid. A number of difficulties emerged which were limited in extent to small portions of the speech signal. The graphical approach was well suited for these cases since it allowed a detailed comparison of the signals. In addition this technique permitted the spectra to be examined directly which facilitated the estimation process.

Experimentation with the vocoder showed that the speech quality was highly dependent upon the proper selection of the fundamental frequency. Major problems in the development were generally hinged to errors in this parameter, and several techniques have been described which were aimed at improving the capability of the selection algorithm. The final configuration allows the fundamental frequency to be estimated to within .5 Hz. Since for typical speech ω_o varies from 67 Hz to 333 Hz this provides resolution to greater than 1 percent. The price for accomplishing this accuracy is in terms of computation speed. The extraction of the fundamental frequency requires the majority of the computer power needed to operate the vocoder. For a 25.6 msec speech segment 35.6 sec of CPU time are currenty needed on a VAX 11/750. Although this can undoubtedly be improved through more efficient programming, real time processing would require an improvement by a factor of nearly 1400. In order to achieve this while still providing adequate frequency resolution more computer power will be needed. Some form of parallel approach would probably be desirable due to the segmented nature of the speech analysis.

The testing of the system included the processing of both clean and noisy speech. Four test sentences of clean speech, comprising both male and female speakers, were tested on the system. The results showed that the vocoder produced only minor differences between the actual and the synthetic speech. Proper selection of the voiced/unvoiced threshold ensured that the speech did not sound overly resonant or noisy. In addition it allowed the noisy speech to be processed without the characteristic "buzziness" typically associated with vocoder processing under

noisy conditions. Average signal to noise ratios of both 10 dB. and 20 dB. were tested with good results.

5.2 Future Developments

The system which has been described in this thesis is capable of analyzing a speech signal, estimating a number of model parameters and then using these to synthesize the original speech. As mentioned previously, one of the principle uses of an analysis/synthesis system is in speech coding. In [3] a 9600 bit/sec coding scheme is presented which is based on a previous implementation of the vocoder presented here. In addition a 8000 bit/sec has recently been demonstrated and there is interest in reducing the bit rate even further. Current concepts on the means by which this can be accomplished are based on the elimination of redundant information and more efficient coding. The primary targets for these improvements are the voiced/unvoiced decisions and the phase of the spectral envelope. Both of these parameters have statistical properties which could possibly yield a substantial bit reduction. A further technique which is more complex but could yield larger bit savings is to exploit the similarity of neighboring speech frames. Since speech is slowly varying the pitch period and spectral envelope are related from frame to frame. Differential coding schemes relating the difference between sequential frames may be able to use their similarity to advantage.

Bibliography

[1] Daniel W. Griffin and Jae S. Lim A New Model-Based Speech Analysis/Synthesis System *IEEE International Conference on Acoustic, Speech and Signal Processing*, Tampa, Florida, March 26-29, 1985, pp. 513-516.

[2] Daniel W. Griffin and Jae S. Lim A New Pitch Detection Algorithm *International Conference on Digital Signal Processing*, Florence, Italy, Sept. 5-8, 1984. 1985.

[3] Daniel W. Griffin and Jae S. Lim A High Quality 9.6 kbps Speech Coding System *IEEE International Conference on Acoustic, Speech and Signal Processing*, Tokyo, Japan, April 13-20, 1986.

[4] Daniel W. Griffin and Jae S. Lim Signal Estimation From Modified Short-Time Fourier Transform *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. ASSP-32, no. 2, pp. 236-243, April 1984.