VOICE SIMULATION: FACTORS AFFECTING QUALITY AND NATURALNESS

B. Yegnanarayana
Department of Computer Science and Engineering
Indian Institute of Technology, Madras-600 036, India

J.M. Naik and D.G. Childers
Department of Electrical Engineering
University of Florida, Gainesville, FL 32611, U.S.A.

ABSTRACT

In this paper we describe a flexible analysis-synthesis system which can be used for a number of studies in speech research. The main objective is to have a synthesis system whose characteristics can be controlled through a set of parameters to realize any desired voice characteristics. The basic synthesis scheme consists of two steps: Generation of an excitation signal from pitch and gain contours and excitation of the linear system model described by linear prediction coefficients. We show that a number of basic studies such as time expansion/compression, pitch modifications and spectral expansion/compression can be made to study the effect of these parameters on the quality of synthetic speech. A systematic study is made to determine factors responsible for unnaturalness in synthetic speech. It is found that the shape of the glottal pulse determines the quality to a large extent. We have also made some studies to determine factors responsible for loss of intelligibility in some segments of speech. A signal dependent analysis-synthesis scheme is proposed to improve the intelligibility of dynamic sounds such as stops. A simple implementation of the signal dependent analysis is proposed.

I. INTRODUCTION

The main objective of this paper is to develop an analysis-synthesis system whose parameters can be varied at will to realize any desired voice characteristics. This will enable us to determine factors responsible for the unnatural quality of synthetic speech. It is also possible to determine parameters of speech that contribute to intelligibility. The key ideas in our basic system are similar to the usual linear predictive (LP) coding vocoder [1], [2]. Our main contributions to the design of the basic system are: (1) the flexibility incorporated in the system for changing the parameters of excitation and system independently and (2) a means for combining the excitation and system through convolution without further interpolation of the system parameters during synthesis.

Atal and Hanauer [1] demonstrated the feasibility of modifying voice characteristics through an LPC vocoder. There have been some attempts to modify some characteristics (like pitch, speaking rate) of speech without explicitly extracting the source parameters. One such attempt is with the phase vocoder [3]. A recent attempt to independently modify the excitation and vocal tract system characteristics is due to Senef [4]. Unlike the LPC method, Senef's method performs the desired transformations in the frequency domain without explicitly extracting pitch. However, it is difficult to adjust the intonation patterns while modifying the voice characteristics.

In order to transform voice from one type (e.g., masculine) to another (e.g., feminine), it is necessary to change not only the pitch and vocal tract system but also the pitch contour as well as the glottal waveshape independently. It is known that glottal pulse shapes differ from person to person and also for the same person for utterances in different contexts [5]. Since one of our objectives is to determine factors responsible for producing natural sounding synthetic speech, we have decided to implement a scheme which controls independently the vocal tract system characteristics and the excitation characteristics such as pitch, pitch contour and glottal waveshape. For this reason we have decided to use the standard LPC-type vocoder.

In Sec. II we describe the basic analysis-synthesis system developed for our studies. We discuss two important innovations in our system which provide smooth control of the parameters for generating speech. In Sec. III we present results of our studies on voice modifications and transformations using the basic system. In particular, we demonstrate the ease with which one can vary independently the speaking rate, pitch, glottal pulse shape and the vocal tract response. We report in Sec. IV results from our studies to determine the factors responsible for unnatural quality of synthetic speech from our system. After accounting for the major source of unnaturalness in synthetic speech, we investigate the factors responsible for low intelligibility of some segments of speech. We propose a signal dependent analysis-synthesis scheme in Sec. V to improve intelligibility of dynamic sounds such as stops.

II. DESCRIPTION OF THE ANALYSIS-SYNTHESIS SYSTEM

A. Basic System

As mentioned earlier, our system is basically same as that LPC vocoders described in the literature [2]. The production model assumes that speech is the output of a time varying vocal tract system excited by a time varying excitation. The excitation is a quasiperiodic glottal volume velocity signal or a random noise signal or a combination of both. Speech analysis is based on the assumption of quasistationarity during short intervals (10-20 msec). At the synthesizer the excitation parameters and gain for each analysis frame are used to generate the excitation signal. Then the system represented by the vocal tract parameters is excited by this signal to generate synthetic speech.

B. Analysis Parameters

For the basic system a fixed frame size of 20 msec (200 samples at 10kHz sampling rate) and a frame rate of 100 frames per second are used. For each frame a set fo 14 LPCs are extracted using the autocorrelation method [2]. Pitch period and voice/unvoiced decisions are determined using the SIFT algorithm [2]. The glottal pulse information is not extracted in the basic system. The gain for each analysis frame is computed from the linear prediction residual. The residual energy for an interval corresponding to only one pitch period is computed and the energy is divided by the period in number of samples. This method of computation of squared gain per sample avoids the incorrect computation of the gain due to arbitrary location of analysis frame relative to glottal closure.

C. Synthesis

Synthesis consists of two steps: Generation of the excitation signal and synthesis of speech. Separation of the synthesis procedure into these two steps helps when modifying the voice characteristics as will be evident in the following sections. The excitation parameters are used to generate the excitation signal as follows: The pitch period and gain contours as a function of analysis frame number (i) are first nonlinearly smoothed using a 3-point median smoothing. Two arrays (called Q and H for convenience) are created as illustrated in Figure 1. The smoothed pitch contour P(i) is used to generate a Q-array using the value of the pitch period at any point to determine the next point on the pitch contour. Since the pitch period is given in number of samples and the interframe interval is known, say N samples, the value of the pitch period at the end of the current pitch period is determined using suitable interpolation of P(i) for points in between two frame indicies. The values of the pitch period as read from the pitch contour are stored in the Q-array. The entry in the Q-array is the value of the pitch period for that frame. For nonvoiced frames the number of samples to be skipped along the horizontal axis is N, although on the pitch contour the value is zero. The entry in the Q-array for unvoiced frames is zero. For each entry in the Q-array the corresponding squared gain per sample can be computed from the gain contour using suitable interpolation between two frame indices. The squared gain per sample corresponding to each element in the Q-array is stored in the H-array.

From the Q and H arrays an excitation signal is generated as follows. For each nonvoiced segment, identified by an entry zero in the Q-array, N_s samples of random noise are generated. The average energy per sample of the noise is adjusted to be equal to the entry in the H-array corresponding to that segment. For a voiced segment identified by a nonzero value in the Q-array, the required number of excitation samples are generated using any desired excitation model. In the initial experiments only one of the five exctiation models shown in Figure 2 were considered. The model parameters were fixed apriori and they were not derived from the speech signal. Note that the total number of excitation samples generated in this way are equal to the number of desired synthetic speech samples.

Once the excitation signal is obtained, the synthetic speech is generated by exciting the vocal tract system with the excitation samples. The system parameters are updated every N samples. We are not using pitch synchronous updating of the parameters, as is normally done in LPC synthesis. Therefore, interpolation of parameters is not necessary. Thus, the instability problems arising out of the interpolated system parameters are avoiced. We still obtain a very smooth synthetic speech.

III. STUDIES USING THE BASIS SYSTEM

Two sentences spoken by a male speaker were used in our studies with the system:

S1: WE WERE AWAY A YEAR AGO

S2: SHOULD WE CHASE THOSE COWBOYS

Speech data sampled at 10kHz was analyzed under the following conditions:

Frame size: 200 samples
Frame rate: 100 frames/sec
Each frame was preemphasized and windowed
Number of LPC's: 14
Pitch contour: (SIFT algorithm)
Gain contour: (from LP residual)
3-point median smoothing of pitch and gain contour

The excitation signal was generated using the smoothed pitch and gain contours with the non-overlapping samples per frame being N=200. The excitation model-3 (Fig. 2) was used throughout the initial studies. This model was a simple impulse excitation normally used in most LPC synthesizers. Synthesis was performed by using the excitation signal with the all-pole system. The system parameters were updated every 100 samples.

We conducted the following studies using this system.

A. Time expansion/compression with spectrum and excitation characteristics preserved.
B. Pitch period expansion/compression with spectrum and other excitation characteristics preserved.
C. Spectral expansion/compression with all the excitation characteristics preserved.
D. Modification of voice characteristics (both pitch and spectrum).

The list of recordings made from these studies is given in Appendix.

The synthetic speech is highly intelligible and devoid of clicks, noise, etc. The speech quality is distinctly synthetic. The issues of quality or naturalness will be addressed in Section IV.

IV. FACTORS FOR UNNATURAL QUALITY OF SYNTHETIC SPEECH

It appears that the quality of the overall speech depends on the quality of reproduction of voiced segments. To determine the factors responsible for synthetic quality of speech, a systematic investigation was performed. The first part of the investigation consisted of determining which of the three factors namely, the vocal tract response, pitch period contour, and glottal pulse shape contributed significantly to the unnatural quality. Each of these factors was varied over a wide range of alternatives to determine whether a significant improvement in quality can be achieved. We have found that glottal pulse approximation contributes to the voice quality more than the vocal tract system model and pitch period errors.

Different excitation models were investigated to determine the one which contributes most significantly to naturalness. If we replace the glottal pulse characteristics with the LP residual itself, we get the original speech. If we can model the excitation suitably and determine the parameters of the model from speech, then we can generate high quality synthetic speech. But it is not clear how to model the excitation. Several artificial pulse shapes with their parameters arbitrarily fixed, are used in our studies (Fig. 2).

Excitation Model-1: Impulse excitation
Excitation Model-2: Two impulse excitation
Excitation Model-3: Three impulse excitation
Excitation Model-4: Hilbert transform of an impulse
Excitation Model-5: First derivative of Fant's model [6]

Out of all these, Model-5 seems to produce the best quality speech. However, the most important problem to be addressed is how to determine the model parameters from speech.

The studies on excitation models indicate that the shape of the excitation pulse is critical and it should be close to the original pulse if naturalness is to be obtained in the synthetic speech. Another way of viewing this is that the phase function of the excitation plays a prominent role in determining the quality. None of the simplified models approximate the phase properly. So it is necessary to model the phase of the original signal and incorporate it in the synthesis. Flanagan's phase vocoder studies [7] also suggest the need for incorporating phase of the signal in synthesis.

V. SIGNAL-DEPENDENT ANALYSIS-SYNTHESIS SCHEME

The quality of synthetic speech depends mostly on the reproduction of voiced speech, whereas, we conjecture that intelligibility of speech depends on how different segments are reproduced. It is known [8] that analysis frame size, frame rate, number of LPCs, pre-emphasis factor, glottal pulse shape, should be different for different classes of segments in an utterance. In many cases unnecessary preemphasis of data, or high order LPCs can produce undesirable effects. Human listeners perform the analysis dynamically depending on the nature of the input segment. So it is necessary to incorproate a signal dependent analysis-synthesis feature into the system.

There are several ways of implementing the signal dependent analysis ideas. One way is to have a fixed size window whose shape changes depending on the desired effective size of the frame. We use the signal knowledge embodied in the pitch contour to guide the analysis. For example, the shape of the window could be a Gaussian function, whose width can be controlled by the pitch contour. The frame rate is kept as high as possible during the analysis stage. Unnecessary frames can be discarded, thus reducing the storage requirement and synthesis effort.

The signal dependent analysis can be taken to any level of sophistication, with consequent advantages of improvement in intelligibility, bandwidth compression and probably quality also.

VI. DISCUSSION

We have presented in this paper a discussion of an analysis-synthesis system which is convenient to study various aspects of the speech signal such as the importance of different parameters of features and their effect on naturalness and intelligibility. Once the characteristics of the speech signal are well understood, it is possible to transform the voice characteristics of an utterance in any desired manner. It is to be noted that modelling both the excitation signal and the vocal tract system are crucial for any studies on speech. Significant success has been achieved in modelling the vocal tract system accurately for purposes of synthesis. But on the other hand we have not yet found a convenient way of modelling the excitation source. It is to be noted that the solution to the source modelling problem does not lie in preserving the entire LP residual or its Fourier transform or parts of the residual information in either domain. Because any such

approach limits the manipulative capability in synthesis especially for changing voice characteristics.

APPENDIX A: LIST OF RECORDINGS

1. Basic system
   Utterance of Speaker 1: (a) original (b) synthetic (c) original
   Utterance of Speaker 2: (a) original (b) synthetic (c) original
   Utterance of Speaker 3: (a) original (b) synthetic (c) original
2. Time expansion/compression
   (a) original (b) 1½ times normal speaking rate (c) normal speaking rate (d) ½ the normal speaking rate (e) original
3. Pitch period expansion/compression
   (a) original (b) twice the normal pitch frequency (c) normal pitch frequency (d) half the normal pitch frequency (e) original
4. Spectral expansion/compression
   (a) original (b) spectran expansion factor 1.1 (c) normal spectrum (d) spectral compression factor 0.9 (e) original
5. Conversion of one voice to another
   (a) male to female voice:
   original male voice - artificial female voice - original female voice
   (b) male to child voice:
   original male voice - artificial child voice - original child voice
   (c) child to male voice:
   original child voice - artificial male voice - original male voice
6. Effect of excitation models
   (a) original (b) single impulse excitation (c) two impulses excitation (d) three impulses excitation (e) Hilbert transform of an impulse (f) first derivative of Fant's model of glottal pulse

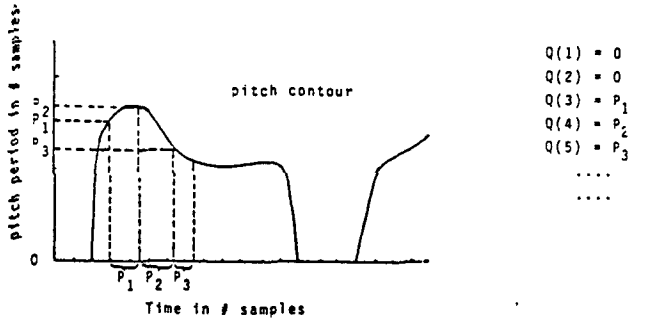

Fig 1a. Illustration of generating Q-Array from smoothed pitch contour

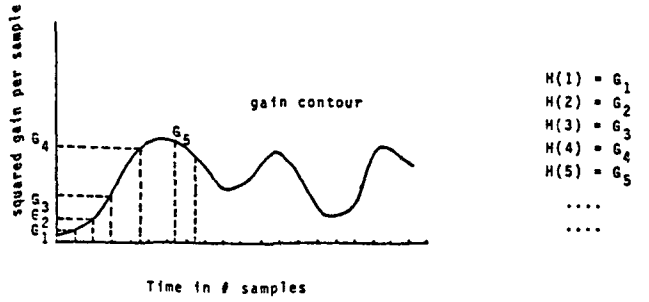

Fig 1b. Illustration of generating H-Array from smoothed pitch and gain contours

REFERENCES

[1] B.S. Atal and S.L. Hanauer, J. Acoust. Soc. Amer., vol. 50, pp. 637-655, 1971.

[2] J.D. Markel and A.H. Gray, Linear Prediction of Speech, Springer-Verlag, 1976.

[3] J.L. Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, 1972.

[4] S. Seneff, IEEE Trans. Acoust., Speech and Signal Processing, vol. ASSP-30, no. 4, pp. 566-577, August 1982.

[5] R.H. Cotton and J.A. Estrie, Elements of Voice Quality in Speech and Language, N.J. Lass (Ed.), Academic Press, 1975.

[6] G. Fant, "The Source Filter Concept in Voice Production," IV FASE Symposium on Acoustics and Speech, Venezia, April 21-24, 1981.

[7] J.L. Flanagan, J. Acoust. Soc. Amer., vol. 68, pp. 412-420, August 1980.

[8] C.R. Patisaul and J.C. Hammett, Jr., J. Acoust. Soc. Amer., vol. 58, pp. 1296-1307, December 1975.

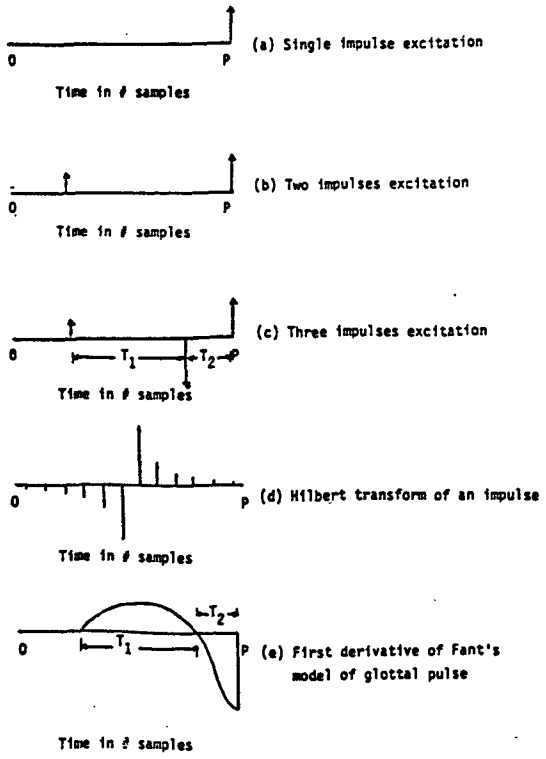

Fig 2. Different Models for excitation