EVALUATION OF A SPEECH SYNTHESIS METHOD FOR NONLINEAR MODELING OF VOCAL FOLDS VIBRATION EFFECT

Hiroshi Ohmura *Kazuyo Tanaka*

Electrotechnical Laboratory 1-1-4 Umezono, Tsukuba, Ibaraki 305, Japan

ABSTRACT

In this paper, we present a new speech synthesis method for improving voice quality in parametric rule-based speech synthesis systems. We also describe the results of a preference test on speech wave reconstruction to confirm the performance of the proposed method. The method is based on the functional approximation of vocal tract resonance produced by nonlinear interaction between the glottis and the vocal tract. In the performance test, evaluators listen to two kinds of reconstructed speech samples: one is synthesized by the proposed method and the other is by an ordinary LPC(Linear predictive coding)-based method. The speech sample set used in this test contains 60 sentences uttered by four speakers. Results show that the proposed method is superior in its quality.

1. INTRODUCTION

It is well known that the vocal folds play a crucial role in the speech production because of their ability to modulate the flow of air. From a theoretical viewpoint, the vocal folds vibration affects the vocal tract transfer characteristics through a nonlinear time-varying interaction between the glottis and vocal tract. However, ordinary speech processing system have neglected this effect by assuming the producing system is linear. But if we try to improve the voice quality to be more natural or human-like, it becomes crucial to investigate and model such nonlinear effect.

For this purpose, we conducted analytic and synthetic experiments on this nonlinear effect where smooth movement of the vocal folds vibration is assumed for computational feasibility of the modeling. From this we proposed a new speech synthesis model, called the Nonlinear-Energy-Damping(NED) wave function model, in which formant energy damping is given by a time window function[1].

The following sections describe the principle of our method, the procedures used for evaluating the experiment, and the experiment's results.

2. SPEECH PRODUCTION MODEL

2.1. The second order nonlinear differential eqation

There are several approaches[2,3] for the modeling of self-excited speech production in which a dynamic model of vocal folds vibration is implemented for generating the synthetic speech with a high level of naturalness. As a model for integrating speech analysis and synthesis systems, the higher order nonlinear model is very attractive but has difficulty in estimating a glottal area waveform from real speech. The second order model Eq.(1) is suited for considering a practical model from a viewpoint of easily controlling each formant magnitude.

The differential equation is

$$\ddot{x} + K\dot{x} + \omega_0^2 x = 0 \qquad (1)$$

where $K\dot{x}$ is regarded as characterizing a nonlinear friction term. Angular frequency ω_0 is given by $2\pi F_0$ in which parameter F_0 is a resonant frequency on condition $K \equiv 0$.

Next, we introduce $K \equiv K(t)$ representing time-varying energy loss within a pitch period. Under the assumption of slow-changing $K(t)$, equation (1) will be approximated by a difference equation[1].

Figure 1 shows its responses and the resultants of second order LP analysis. The figure indicates a great difference between A and $A^{'}$. It will be maintained that a nonlinear energy damping model is required for more precise description of such phenomenon.

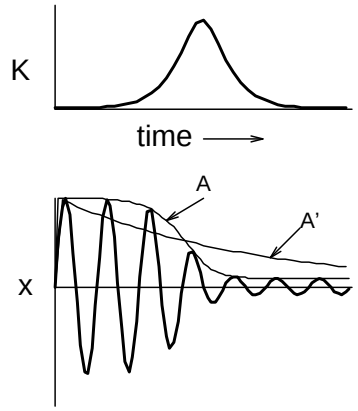

Figure 1. Responses of the second order nonlinear system (1). K is given, X is the nonlinear response at $F = 700$Hz, A is its energy envelope, and $A^{'}$ is the second order linear filter energy envelope estimated by LP analysis.

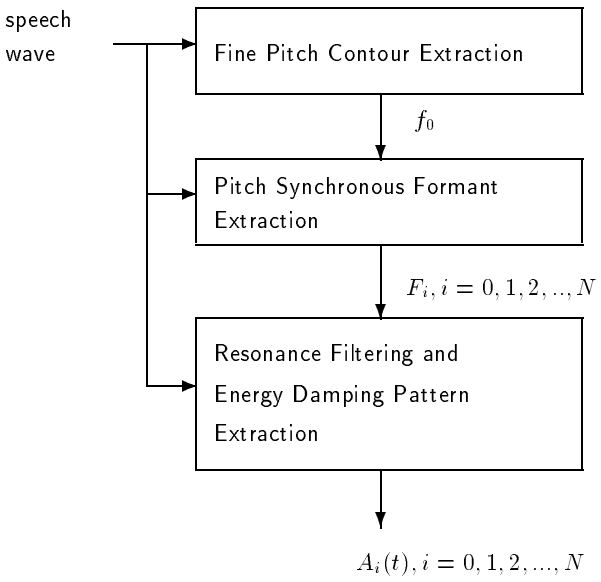

Figure 2. Block diagram for the speech parameter extraction system.

2.2. NED Wave Function Model

Based on these experimental results, we propose a speech synthesis technique called Nonlinear Energy Damping(NED) wave function model in which formant energy damping is given by a time window function. The model is in the form of equation (2) which is the most simple for describing the nonlinear response, this expression, however is not a strict solution of equation (1) .

$$x_i(t) = a_i \exp(K_i(t)) \sin(2\pi F_i t + \theta_i) \tag{2}$$

Where a_i is amplitude constant ,$\exp(K_i(t))$ is energy damping function, F_i is $i-$th formant, and θ_i is phase constant.

The function $\exp(K_i(t))$ can be interpreted as a kind of time window function.

Speech wave $s(t)$ is the sum of $x_i(t)$ for $i = 0, 1, 2, \cdots, N$, that is,

$$s(t) = \sum_{i=0}^{N} x_i(t) \tag{3}$$

where N is the number of formants.

3. SPEECH ANALYSIS-SYNTHESIS EXPERIMENTS

3.1. Estimation of Formant Energy Damping Patterns

The knowledge of the vocal folds vibration effect obtained from our analysis is that the effect appears in formant energy damping patterns. We estimate these damping patterns by applying Teager-Kaiser(TK)-energy operator[5,6], that is, the TK-energy operator $e_{i,j}$ is

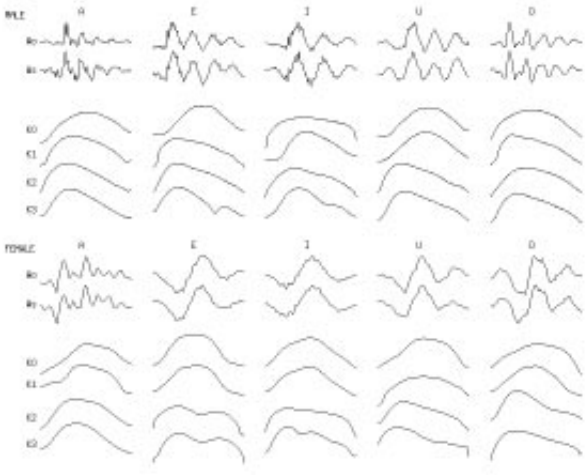

Figure 3. Reconstructed waveform for five Japanese vowels, uttered by a male and a female speakers. W_o: Original waveforms. W_s:Reconstructed waveforms. K_i: Log energy damping patterns of resonant frequency F_i (K_0 is a glottal wave component) and ignored their phase differences for plotting.

$$e_{i,j} = s_{i,j}^2 - s_{i,j-1}\, s_{i,j+1} \tag{4}$$

$$A_{i,j} = f_s \frac{\sqrt{e_{i,j}}}{\omega_i} \tag{5}$$

where $s_{i,j}$ is the output wave corresponding to i-th resonant frequency F_i, $A_{i,j}$ is a sample sequence of the damping patterns, f_s is sampling frequency, and $\omega_i = \sin(2\pi F_i / f_s)$. where $x_0(t)$ is a voice fundamental wave component.

Figure 2 is the speech parameter extraction system, where f_0 is fundamental frequency, F_i is i-th resonant frequency, and $A_i(t) = a_i \exp\{K_i(t)\}$ is the envelope function within a pitch period of resonance F_i.

3.2. Speech Synthesis Experiments

We then conducted speech synthesis experiments using speech parameters extracted by the analysis system in Figure 2. Figure 3 shows reconstructed waveforms of five Japanese vowels for a male and a female speakers their averaged fundamental frequencies are 111Hz and 213Hz respectively. The values of cross-correlation coefficient between original and synthesized waves become 0.8 and up.

From these experimental results, NED wave function model(Eq.2) confirmed its performance for application to real speech.

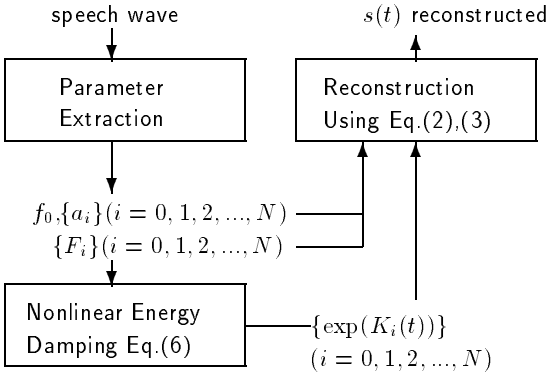

Figure 4. Speech wave reconstruction system.

4. EVALUATION EXPERIMENTS ON A SPEECH WAVE RECONSTRUCTION SYSTEM

4.1. Speech Wave Reconstruction System

We implemented a speech wave reconstruction system to evaluate the proposed synthesis method. Figure 4 is the speech wave reconstruction system in which $a_i, F_i (i = 0, 1, 2, \cdots, N)$ are extracted from real speech.

Energy damping term $\exp\{K_i(t)\}$ is represented by a function of time t and F_i as follows.

$$\exp\{K_i(t)\} = \exp\{k_{0,i}t + k_1(t)\} \tag{6}$$

The first term in the left $k_{0,i}t$ means linear energy damping and its constant is given by $k_{0,i} = -\pi B(F_i)/f_s$ where $B(F_i)$ is the bandwidth function derived from Dunn's formant-bandwidth data[7]. The second term for nonlinear energy damping is in the form of equation (7) and is commonly used for all resonances.

$$\begin{aligned}
\exp\{k_1(t)\} &= 0.5\{1 + \cos\frac{\pi(t - T_2)}{T_2}\} \text{ for } 0 \le t \le T_2 \\
&= 1 \text{ for } T_2 < t < T_1 \\
&= 0.5\{1 + \cos\frac{\pi(t - T_1)}{T_0 - T_1}\} \text{ for } T_1 \le t \le T_0 \\
&= 0 \text{ for } |t| > T_0
\end{aligned} \tag{7}$$

Time function $\exp\{k_1(t)\}$ consists of two ordinary windows (in this paper, the Hanning type is used). T_0 is a pitch period and T_1, T_2 are those window's center coordinates respectively. Example patterns of $\exp\{k_1(t)\}$ and $\exp\{K_i(t)\}$ are shown in Figure 5 for resonant frequencies 0.5, 1.5, 2.5, and 3.5kHz. Phase parameters in Eq.(2) $\theta_i (i = 0, 1, 2, \cdots, N)$ are all 0.

The evaluation is conducted by comparing this method with an ordinary LPC-based reconstruction method[8].

In the proposed method(NED-method), pitch contours[9], formant trajectories, and formant intensity parameters are estimated from original speech signals[1], and then they are reconstructed by the method in Figure 4. The sampling frequency of original speech is 16kHz, and the number of formants used is adaptively determined depending on its fundamental frequency(in this case, it ranges from 5 to 9). An impulse sequence is used as the voiced sound source and speech waves are generated in pitch synchronous. On the other hand, sound source for voiceless consonants are generated by using certain noises which were used in our previous speech synthesis system.

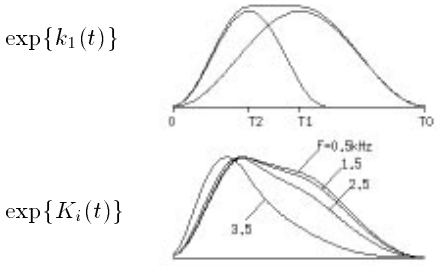

Figure 5. The window function models for nonlinear energy damping of resonance, Eq.(6) and Eq.(7).

The speech wave reconstructed by the LPC filter uses also the same impulse sequence and noises as its source and generated in the pitch synchronous. The length of the LPC filter is 20 for all samples. Basically the same equipment and tools are used in the implementation of both systems.

4.2. Speech Materials

The speech samples used for the preference test are 15 Japanese sentences uttered by two males and two females(60 samples in all) which were selected from ASJ-Continuous-Speech-Database[10]. The length of the sentences ranges about from fifteen to thirty moras.

4.3. Testing Method

Pairs of reconstructed speech samples, i.e. stimuli, are sequentially presented. Each pair consists of one reconstructed by the NED method and the other by the LPC filter, and of course, their order is randomized(See Fig.6). But a sequence set consists of speech samples uttered by a single speaker, so that each evaluator listens to 4 sets of stimuli sequences, each of which contains 15 pairs. The stimuli(speech samples) are presented through an audio speaker. Evaluators listen to the sequences of the sample pairs and decide which one is more preferable as to the natural voice quality.

We selected six evaluators with normal hearing ability to do the judging. They are not experts in speech technology and have little experience in listening to synthesized speech. No preliminary training was carried out.

4.4. Results and Discussion

Results for the preference test are shown in Table 1, in which the score are sorted for speakers. Table 2 shows the scores given by individual evaluators. As a whole, the result indicates the superiority of the NED method. However, contrary scores appears for the stimuli of speaker W1(female). If we remove these W1 scores from the results, the superiority of NED method becomes clearer. We presume the cause is that these stimuli feel a little smoother due to unstable formant trajectory estimation in the analysis stage. It will

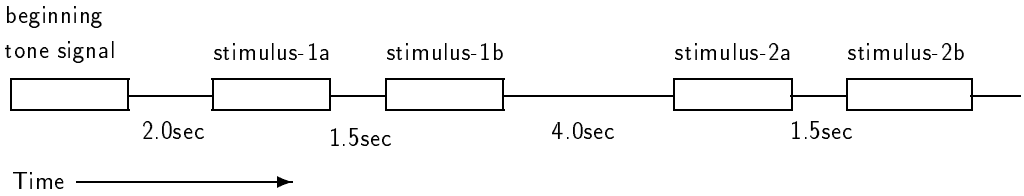

Figure 6. An example of the stimulus sequences.

be possible to resolve such a problem when applying the NED method to text-to-speech conversion.

Table 1. Preference scores for individual speakers.

	Spk	M1	M2	W1	W2	(Total)
NED		51	71	29	66	217
LPC		39	19	61	24	143

Table 2. Preference scores for individual evaluators.

	Evl	E1	E2	E3	E4	E5	E6	(Total)
NED		35	45	29	26	39	43	217
LPC		25	15	31	34	21	17	143

5. FUTURE WORK

We are now implementing a revised speech reconstruction system in which some improved techniques are employed into formant trajectory modeling, and we are also planning to construct a text-speech conversion system.

6. ACKNOWLEDGMENT

We wish to thank Dr. Nobuyuki Ohtsu, Director of the Machine Understanding Division and all the members of that section for their usual discussion and support.

REFERENCES

[1] H. Ohmura and K. Tanaka. "Speech synthesis using a nonlinear energy damping model for the vocal folds vibration effect". In *Proc. ICSLP96*, pages 1241–1244, 1996.

[2] K. Ishizaka and J. L. Flanagan. "A self-oscillating model of glottal sound source". *J. Acoust. Soc. Japan*, 34(3):122–131, 1978.

[3] T. Ikeda and Y. Matsuzaki. "Flow theory for analysis of phonation with a membrance model of vocal cord". In *Proc. ICSLP94*, pages 643–650, 1994.

[4] M. Rothenberg and S. Zahorian. "Nonlinear inverse filtering technique for estimation the glottal area waveform". *J. Acoust. Soc. Am.*, 61(4):1063–1071, 1977.

[5] J. F. Kaiser. "On a simple algorithm to calculate the energy of a signal". In *Proc. ICASSP90*, pages 381–384, 1990.

[6] P. Maragos, T. F. Quatiern, and F. Kaiser. "Speech nonlinearities, and energy operators". In *Proc. ICASSP91*, pages 421–424, 1991.

[7] J. L. Flanagan. *Speech Analysis Synthesis and Perception.* Springer-Verlag, New York, 1972.

[8] B. S. Atal and S. L. Hanauer. "Speech analysis and synthesis by linear prediction of the speech wave". *J. Acoust. Soc. Am.*, 50(2):637–655, 1971.

[9] H. Ohmura. "Fine pitch contour extraction by voice fundamental wave filtering method". In *Proc. ICASSP94*, pages 189–192, 1994.

[10] T. Kobayashi, S. Itahashi, S. Hayamizu, and T. Takezawa. "ASJ continuous speech corpus for research". *J. Acoust. Soc. Japan*, 48(12):888–893, 1992.