SPEECH SYNTHESIS USING A NONLINEAR ENERGY DAMPING MODEL FOR THE VOCAL FOLDS VIBRATION EFFECT

Hiroshi OHMURA *Kazuyo TANAKA*

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305, Japan
E-mail: ohmura@etl.go.jp, ktanaka@etl.go.jp

ABSTRACT

From a theoretical viewpoint, the vocal folds vibration affects the vocal tract transfer characteristics through nonlinear time-varying interaction between the glottis and vocal tract. Therefore, it is crucial to investigate and model such effects in order to improve voice quality in parametric rule-based speech synthesis systems. In this paper, we first conducted analytic experiments on the vocal folds vibration effects on the appeared in formant energy damping patterns and then modeled them by a nonlinear 2nd order differential equation for speech synthesis. At last, we confirmed the feasibility of the nonlinear model by speech wave reconstruction experiments.

1. INTRODUCTION

Strictly speaking, the vocal folds vibration affects the vocal tract transfer characteristics through nonlinear time-varying interaction between the glottis and vocal tract. However, ordinary speech processing systems have neglected those effects under the assumption of linearity of the production systems. It was successful at a certain level, but if we try to improve the voice quality to be more natural or human-like, it becomes crucial to investigate and model such effects. An experimental system of processing such voiced speech segments as amplitude/frequency modulated signals have been presented [1]. We considered the subject from the aspect of nonlinear speech production process.

For this purpose, we first conduct analytic experiments on the vocal folds vibration effects as they appeared in formant energy damping patterns. From these experimental results , we propose two types of speech synthesis models: one is a wave function model in which formant energy damping is given by a time window function. The other is a 2nd order nonlinear differential equation in which the formant energy damping pattern is controlled by its friction term. We confirm its performance by the experiments on speech waveform reconstruction from extracted formant energy damping patterns.

2. ESTIMATION OF FORMANT ENERGY DAMPING PATTERNS

A system for estimating the formant energy damping patterns is shown in Figure 1, which consists of three stages: fine pitch contour extraction[2], formant extraction using a macro-adaptive vocal tract filter, and estimation of the energy damping patterns by using the TK-energy operator[3].

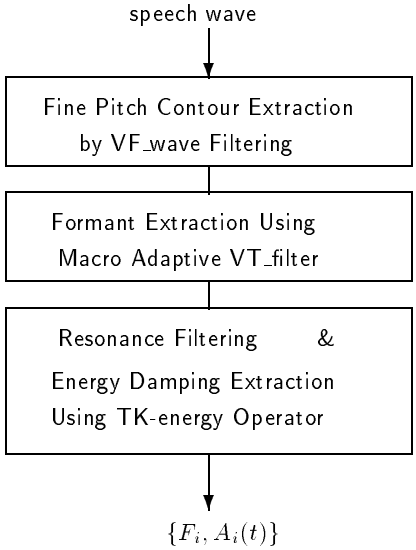

Figure 1: Block diagram for the analysis system, where F_i is the i-th formant and $A_i(t)$ is the envelope of resonance.

2.1. Formant Extraction Using Macro-Adaptive Vocal Tract Filter

One of the most important concerns in the automatic formant extraction is the automatic determination of length M_0 of the vocal tract characteristic function (vocal tract filter). Several approaches have been proposed for this problem[4][5]. We approached this matter from the macro-structure existing in the relation between pitch and vocal tract length[6] and proposed a new formant extraction method[7] based on the functional model of vocal tract length(L) with respect to

averaged fundamental frequency($\overline{f}_0$) represented by equation (1).

$$L(\overline{f}_0) = \alpha_1 \overline{f}_0 + \alpha_0 \quad (1)$$

The length M_0 is obtained by the equation (2).

$$M_0 = \frac{2F_s L(\overline{f}_0)}{c} \quad (2)$$

Where F_s is sampling frequency, c is sound velocity. Formants are given as poles of the vocal tract filter by means of polynomial root solving.

2.2. Estimation of Formant Energy Damping

A resonance $s_i(t)$ is extracted by filtering input speech wave using a FIR-filter of which the center frequency corresponds to formant frequency F_i and its time window is the Blackman type[8]. Formant energy damping parameter $A_i(t)$ is calculated by TK-energy operator as in equation (3).

$$\begin{aligned} e_n &= s_n^2 - s_{n-1}s_{n+1} \\ A_n &= F_s \frac{\sqrt{e_n}}{\omega_i} \end{aligned} \quad (3)$$

Where s_n is a sample of wave $s_i(t)$ at $t = nT$, $\omega_i = 2\pi F_i$, e_n is TK-energy parameter,and A_n is a sample of $A_i(t)$.

Figure 2 shows outputs from the analysis system in Figure 1 . The speech sample is vowel /a/ uttered by a male whose averaged fundamental frequency is $f_0 = 107$Hz. In the left, s is input speech , s_0 is a glottal wave component , s_i is filtered resonant wave at frequency F_i. Their logarithmic energy damping pattern $\log A_i$ are shown in the right. The essential point here is the existence of a nonlinearity in $logA_i$ patterns.

3. MODELING OF FORMANT ENERGY DAMPING

In the beginning, we considered a second order differential equation with time-varying friction term as a simplified simulation of nonlinear interaction between vocal tract and time varying glottal impedance.

3.1. Second Order Differential Equation With Time Varying Friction Term

The differential equation is

$$\ddot{x} + K\dot{x} + \omega_0^2 x = 0 \quad (4)$$

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 $K \equiv 0$. Now,

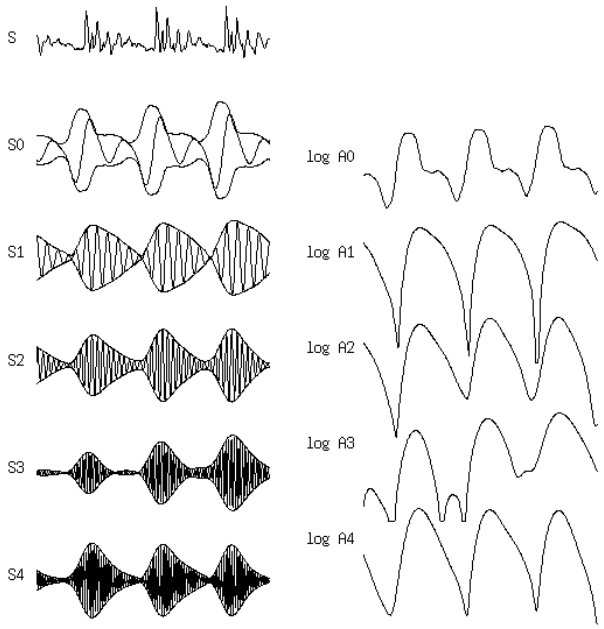

Figure 2: Extracted energy damping patterns $\log A_i$ for resonances S_i for $i = 0, 1, 2, 3, 4$.

we introduce $K \equiv K(t)$ representing time-varying energy loss within a pitch period.

Under the assumption of slow-changing $K(t)$, equation (4) will be approximated by a second order difference equation as follows.

$$x_{j+1} = (2 - K_j\Delta T - \omega_0^2 \Delta T^2)x_j - (1 - K_j \Delta T)x_{j-1} \quad (5)$$

Where x_j and K_j are samples of $x(t)$ and $K(t)$ respectively, at $t = j\Delta T$ and $\Delta T = 1/F_s$.

Figure 3 shows responses of equation (5) and its 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.

3.2. Resonant Wave Function Model

The wave function model is in the form of equation (6) which is the most simplest one for describing the nonlinear resonance, but this expression is not a strict solution of equation (4) .

$$x_i(t) = a_i \exp(K_i(t)) \sin(2\pi F_i t + \theta_i) \quad (6)$$

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 $i = 1, 2, ..., n$.

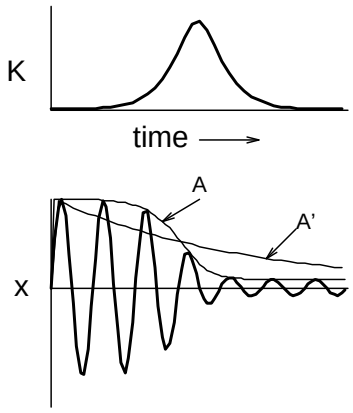

Figure 3: Responses of the second order nonlinear system (5). From the top to the bottom, K is the time pattern for $K(t)$, X is the nonlinear response at $F = 700$Hz, its energy envelope A and the second order linear filter energy envelope $A^{'}$ estimated by LP analysis.

3.3. Nonlinear Network Model

A network representation for the fourth order nonlinear equation is shown in figure 4. This network is known as a nonlinear vocal tract system with Norton equivalent model (inductance L_g ignored in this paper) for the glottal source[9]. In this circuit, it is assumed that glottal flow

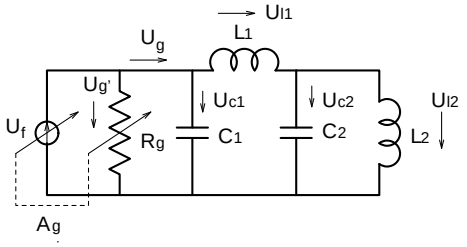

Figure 4: A nonlinear network model

$U_f(t)$ is proportional to glottal are $A_g(t)$ and glottal resistance $R_g(t)$ varies inversely with $A_g(t)$. If one replaces the volume velocity U_{12} by variable x , the differential equation with external periodic excitation is as follows:

$$\dddot{x} + \frac{1}{\tau_1}\ddot{x} + (\omega_1^2 + \omega_2^2)\dot{x} + \frac{1}{\tau_2}\dot{x} + \omega_1^2\omega_2^2 x = \frac{\omega_1^2\omega_2^2}{R_g}P_g \qquad (7)$$

where, $\tau_1 = C_1 R_g(t)$,$\tau_2 = \tau_1/(\omega_1^2 + \omega_2^2 - \omega_1\omega_2)$, $\omega_i = 2\pi F_i$, and $P_g/R_g(t) = U_f(t)$.

Figure 5 shows a glottal area wave $A_g(t)$ given, the nonlinear system response Wv, the responce Wc outputted from the linear system of which the glottal resistance $R_g(t)$ is a constant calculated by averaged $A_g(t)$, and their spectra Spv and Spc respectively. The differences between these two spectra appeared clearly on their formant magnitude and bandwidth. We expect that the nonlinear energy damping model is a useful and important approach for pliable speech synthesis. As a model for integrating speech analysis and synthesis system, the higher order nonlinear model is very attractive but its has a difficulty in estimating a glottal area wave from real speech. We think that a parallel circuit constructed by the second order model Eq.(4) is suited for practical purposes from the viewpoint of easily controlling each formant magnitude.

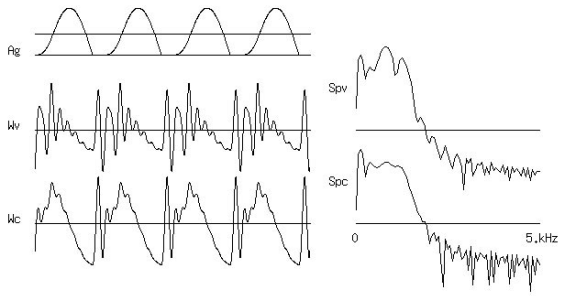

Figure 5: Time and frequency responses of the nonlinear network in Fig.4. Where Ag is a given glottal area wave, Wv is the time responce and Wc is the one for the glottal resistance kept in constant calculating by the averaged glottal area wave $\overline{A}_g$ shown by a horizontal line in the top.

4. SPEECH SYNTHESIS EXPERIMENTS

Now we conducted speech synthesis experiments using formants $\{F_i\}$ and their energy damping patterns $\{A_i(t)\}$ extracted by the analysis system in Fig.1 . The synthesis model is the preceding wave function equation (6) . Speech wave $s(t)$ is given by

$$s(t) = \sum_{i=0}^{n} x_i(t) \qquad (8)$$

Figure 6 shows synthesized wave forms of five 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 became 0.8 and up.

5. RESULTS AND DISCUSSION

Observing the analysis results, we found that the energy damping patterns estimated from vowel samples rapidly change depending on the opening and closing of the vocal folds, especially for those of the first and the second formants. Therefore, the effect of the vocal folds movement appears in not only the gross spectral envelope but also in the individual formant damping patterns.

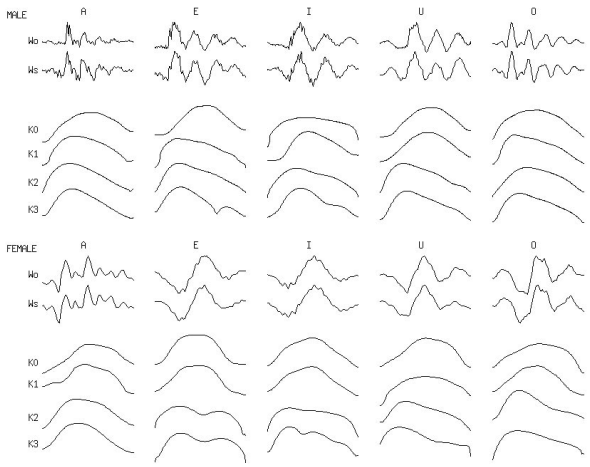

Figure 6: Synthesized wave forms for five vowels uttered a male (averaged pitch 111Hz) and a female (213Hz). Wo and Ws are original and re-synthesized waves respectively and K_i is a energy damping pattern of formant F_i for $i = 0, 1, 2, 3$,where K_0 is a glottal wave component.

On the results of simulating the acoustic dynamics of a single resonance, it was confirmation that, when the system had a time-varying friction term, the ability of 2nd order LP analysis rapidly deteriorated, especially in terms of formant energy damping. From these experimental results, two kinds of speech synthesis models were presented, one is a wave function model in which formant energy damping is presented by a time window function. The other is a 2nd order nonlinear differential equation in which the formant energy damping pattern is controlled by treating its friction term.

Through the experiments of restoring speech wave form from extracted formant energy damping patterns, the model confirmed its performance for application to real speech. Now we are developing a system of speech analysis-synthesis based on the formant energy damping model and evaluating the models using the restored speech.

6. ACKNOWLEDGMENT

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

7. REFERENCES

1. H.Hanson,P.Maragos,A.Potamianos,"Finding Speech Formants and Modulation via Energy Separation with Application to a Vocorder," IEEE Proc. ICASSP93,Vol2.pp716-719,1993.
2. H. Ohmura,"Fine Pitch Contour Extraction by Voice Fundamental Wave Filtering Method,"IEEE Proc. ICASSP94, Vol.2 pp189-192, 1994.
3. J.F. Kaiser,"On a Simple Algorithm to Calculate the Energy of a Signal, " IEEE Proc. ICASSP90 S7.3, 1990.
4. J.D.Markel,"Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation," IEEE Trans.,Vol.AU-20.N0.2,pp129-137,1972.
5. S.Ishizaki,T.Nakajima,"Estimation of Vocal Tract Length by Use of an Information Criterion," A.S.J. Autumn Meeting, 1-4-18,1975.
6. N.Umeda,R.Teranishi,"Phonemic Feature and Vocal Feature," J.A.S.J.,Vol22,No.4,pp195-203,1966.
7. H.Ohmura,"Analysis of the Relationship Between Fundamental Frequency and Vocal Tract Length," A.S.J. Autumn Meeting,1-7-14 ,1993.
8. H.Ohmura,"Intensity Envelope Controlled Speech Synthesis for Considering a Nonlinearity due to the Vocal Folds Vibration," Tech. Report of IEICE,SP95-78,1995.
9. M.Rothenberg, S.Zahorian,"Nonlinear inverse filtering technique for estimation the glottal area waveform," J.A.S.A. Vol.61, No.4 pp1063-1071, 1977.