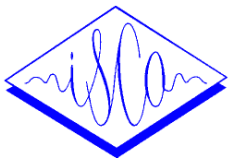


The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

A NEW SYNTHESIZER MODEL FOR HIGH QUALITY SYNTHETIC SPEECH

Tomoki Hamagami, Shinichiro Hashimoto

SECOM Intelligent Systems Laboratory, SECOM Co.,Ltd.,
6-11-23 Shimorenjyaku, Mitaka, Tokyo, 181 Japan

Abstract

We describe a new speech production model for improving the quality of synthetic vowel speech. The strong point of this model is that the source model has a continuous harmonic structure in the time domain and the frequency domain. This report provides a comparative study of the ordinary source model, such as impulse and Rosenberg one, and the new model. By listening test, it is confirmed that our model produces high quality synthetic speech which is better as compared with synthetic ones using ordinary models. Thus, we understand that the really continuous harmonic structure in speech spectrum significantly contributes to synthetic speech quality.

INTRODUCTION

Pulse trains and white noise model has been widely used to simulate periodic vocal cord vibration and turbulence noise made by constrictions in the oral cavity. Nevertheless, there is a strong suspicion that the ordinary procedure has limitation in synthesizing high quality speech which is hardly distinguishable from the human speech. Thus, in order to remove this defect, many voice source model have been studied using the glottal waveform model; such as Rosenberg[5], LF-model[2]. However, as many ordinary source models like so, the model is defined within a pitch period interval. Therefore, the fine harmonic discontinuities cannot be avoided in synthesized speech. For example, as with the Sona-Graph outputs, these discontinuities can be the origin of degrading and changing synthesized speech quality. In order to remove these discontinuities and these effects, a new source generating algorithm is produced with operates in the time domain and frequency domain. This paper shows the effect of a continuous fine harmonic structure on the quality of voiced synthetic speech. At the same time, we demonstrate that our model has the ability to produce the sound of dynamic consonants.

THE IMPROVEMENT OF SOURCE MODEL

In general, a source model is shown in [Fig.1][4].The impulse generator produces a sequence of unit impulses which are spaced by the analyzed fundamental period. This signal in turn excites a glottal pulse model whose impulse response has the desired glottal wave shape. [Fig.2] denotes our $PIFM$ (Pulse source Interpolated by Frequency Modulation) model. The glottal wave shape equivalent to one pitch period is defined by three successive impulse sequences. That is, the fundamental frequencies for each sample point

are estimated by using linear interpolation of two fundamental frequencies. These two fundamental frequencies, $F_0(k), F_0(k+1)$ are the reciprocals of two fundamental periods, $T(k), T(k+1)$. The simplier linear interpolation method was used. That is,

$$f_0(nT_s) = \frac{2nT_s}{T(k)+T(k+1)}(F_0(k+1) - F_0(k)) + F_0(k) \tag{1}$$

$$t(k) + \frac{T(k)}{2} \leq nT_s \leq t(k+1) + \frac{T(k+1)}{2}$$
$$T(k) = t(k+1) - t(k)$$

where, T_s = sampling rate (sec), n = sampling number,

The glottal wave shape is generated as follows. The frequency modulated fundamental wave , u^0, and its corresponding harmonic waves,u^h are summed with their zero phases synchronized, that is

$$U_{-6db/oct}(\tau(nT_s, \Delta)) = \sum_{h=0}^{CutOff} u^h(nT_s) \tag{2}$$

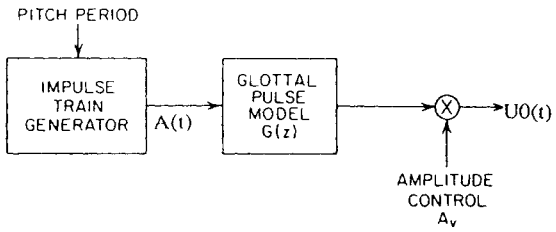
$$u^h(nT_s) = \frac{\varepsilon(nT_s, h)}{2\pi h f_0(nT_s)} \int_{T_k}^{nT_s} (sin(h\theta(nT_s)) + \zeta(nT_s, \Delta))dt \tag{3}$$

Where $\tau()$ is the warping function for synthesizing dynamic consonant, $\zeta()$ is noise source function, and Δ is distinctive feature parameter for controlling $\tau()$, $\zeta()$.

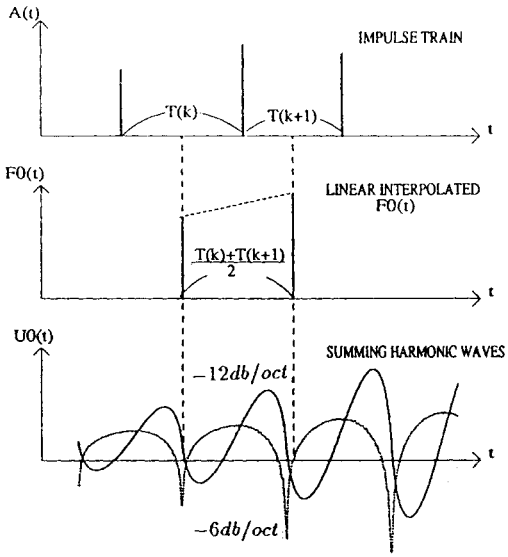
By contorolling these sub-function, $\tau()$, $\zeta()$, Δ, the dynamic consonant variations and the characteristic changes can be obtained.

$U_{-6db/oct}()$ has a frequency quality of $-6db/oct$ which is used as a source of synthesized speech wave at the case of general pre-emphasis analyzing. On the other hand, natural glottal wave has a frequency quality of $-12db/oct$ which is obtained by integrating of $U_{-6db/oct}()$ $PIFM$ wave. Using the method described above, the amplitude variation is defined by interpolating between the two points which defined an interval. The frequency quality of the two type $PIFM$ waves,$U_{-6db/ocr}, U_{-12db/oct}$ is shown [Fig.2].

As mentioned above, a source wave which has continuous harmonic structure like nature ones is obtained by using $PIFM$ model.



[Fig.1 : Generation of the excitation signal for voiced speech.]



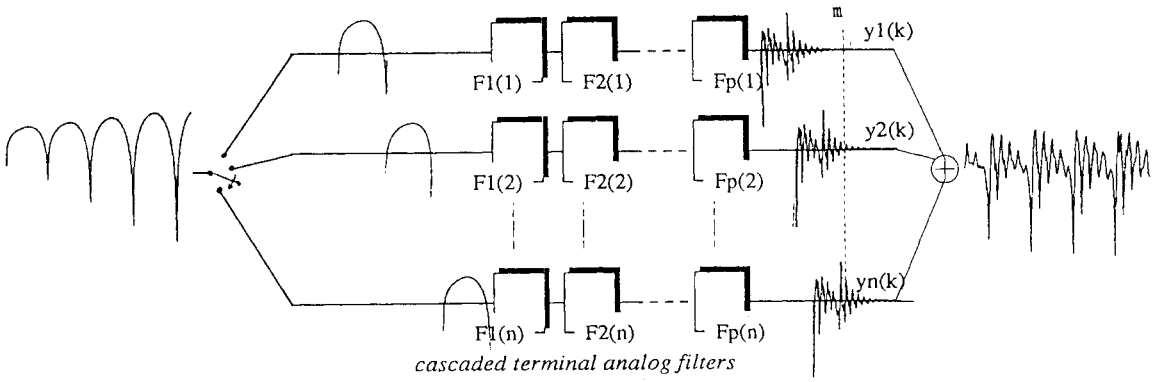
[Fig.2 : Procedure of PIFM algorithm]

THE VOCAL TRACT MODEL

The vocal tract model is an improved version of a typical cascaded formant speech synthesizer[3]. One of the features of this model is to make use of multi cascaded filters in parallel. This method is called " Alternating Reset Multi Terminal Analog Speech Synthesizer".

[Fig.3] denotes this vocal tract model. The input signal ($PIFM_{-6db}$ wave) is alternately fed into each filter for every pitch period as if these cascaded terminal analog filters are provided for its pitch period. Each analog filter's response is excited by the filter's source piece. These responses are repeatedly summed in the time domain until each response perfectly converges. That is, the output, $y(m)$, is obtained by summing all filter's responses,$\sum_{i=0}^{\infty} y_i(m)$.

The advantage of this vocal tract model is that the distortions which are caused by parameter change reduced.



[Fig.3 : Alternating reset multi terminal analog speech synthesizer]

THE LISTENING EXPERIMENT AND RESULTS

We have produced synthesized vowel speech by using the glottal source and vocal tract models which are mentioned above. At the same time, in order to show to exhibit superior performance of continued harmonic structure , formal listening experiments were conducted.

We provided a comparative study of ordinary source models, such as impulse and Rosenberg models, and our model. [TABLE I] shows the analyzing and listening environment during the study. Four types of original(natural) vowel speech samples were provided. In addition, three types of synthesized speech which used the glottal source model, Pulse, Rosenberg, $PIFM$ model were produced from each of the four original samples. Listeners were requested to chose the better speech from a pair of samples. The pair of samples were taken among the four speech samples. (one original and three synthesized)

The result of this experiment is shown in [TABLE II]. We confirmed that the speech quality, measured by the psychological distance of **Thurston case IV**[6], was improved by as much as 0.8 in our model, as compared to the quality obtained by the conventional Impulse-model, and it is improved by as much as 0.4, as compared to the quality obtained

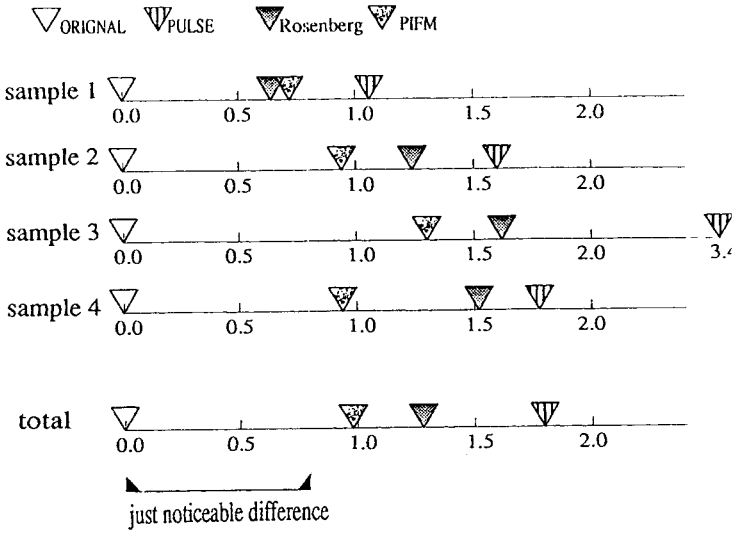
by the results from Rosenberg-model, and the distance between our synthetic speech and original one is 0.9.

In case of this measurement, the jnd (just noticeable difference) is equivalent to 0.75. Thus, It is probable that the synthesized speech which used $PIFM$ model is distinguishable from conventional Impulse-model. In addition, regardless of each original speech, the distance between original and the synthesized speech is nearly constant, $0.7 \sim 1.3$. These distances are more stabiler than other source's distanses.

A/D data	10 kHz 12bit
Analysis Method	10 orders LPC analysis useing a crude pitch-synchronouse
Sampled Data	4 types male voised speech※
Listeners	9 males and 1 female
Method of Distance measure	Thurston case IV
Just Noticeable Difference	0.75

※ 1. /iaeuoai/
2. /aoiueoa/
3. /aiueo/
4. [We were away a year ago]

[TABLE.I : Test condition of analyzing and listening]



[TABLE.II : Results of the experiment]

CONCLUSION

The improved sound source model for high-quality synthetic speech sound is proposed. This model provides the continuous harmonic structure which had not been taken into account by other models. By listening test, it has been shown that the continuous harmonic structure is contributes to synthetic speech quality. Examples of synthetic speech using the new source model will be played during the conference presentation.

References

[1] S Furui. *Digital Speech Processing,Synthesis,and Recognition.* Dekker,INC, 1989.

[2] G.Fant and Q.Lin. Frequency domain interpretation and derivation of glottal parameters. *STL-QPSR,* 2-3:1–21, 1988.

[3] Shinichiro Hashimoto. Alternating reset dual terminal analog speech synthesizer. *Proc.Spring Meet. Acoust.Soc.Jpn.*, pages 367–368, 1974.

[4] L.R.Rabiner/R.W.Schafer. *DIGITAL PROCESSING OF SPEECH SIGNALS.* Prentice-Hall, 1978.

[5] A.E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *J.Acoust.Soc.Am.*, 49(2):583–590, 1971.

[6] Thurston. Psychophysical analysis. *Amer. Jour. Phychol.*, (38), 1927.