A STUDY ON THE SPEECH SYNTHESIS-BY-RULE SYSTEM APPLIED MULTIBAND EXCITATION SIGNAL

Younjeong Kyung*, Geesoon Kim**, Hwangsoo Lee*, Yanghee Lee**

* Department of Information and Communication Engineering, Korea Advanced Institute of Science and Technology, P.O.Box 201, Cheongryang, Seoul, 130-650, KOREA

** Department of Computer Science, Dong-Duck Women's Univ. Wolgokdong 23-1, Sungbukgu, Seoul, 136-714, KOREA

ABSTRACT

In this paper, we design and implement the Korean speech synthesis by rule system. This system is applied the multiband excitation signal on voiced sounds. The multiband excitation signal is obtained by mixing impluse spectrum and white noise spectrum. We find that the quality of synthesized speech is improved using this application. Also, we classify the voiced sounds by cepstral euclidian distance measure for reducing overhead memory. The representative excitation signal of the same group's voiced sounds is used as excitation signal on synthesis. This method does not affect the quality of synthesized speech. As the result of experiment, this method eliminates the "buzziness" of synthesized speech and reduces the spectral distortion of synthesized speech.

1. INTRODUCTION

Generally, the speech wave production mechanism can be devided into two stages[1]. These stages consist of sound source production, and articulation by vocal tract. Speech information processing techniques are based on the linear separable equivalent circuit model of the speech production mechanism. In the basic vocoder scheme, speech signals are generated from a simple source-filter model as illustrated in figure 1.

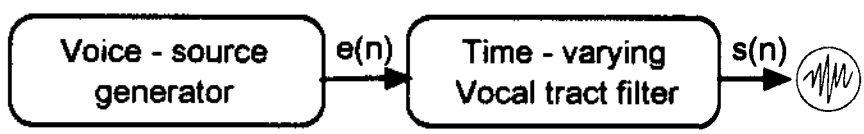

Fig.1 Basic speech production mechanism. The e(n) corresponds to excitation signal and the s(n) corresponds to synthesized speech signal.

The filter characteristics in this model are obtained from an analysis procedure and approximate the short-time power spectral envelope of the input speech. The source signal therefore has a flat spectral envelope and

contributes only fine spectral detail and phase information. In typical excited vocoders the source signal is either a periodic impulse train for generation voiced sounds or a random sequence for generation unvoiced sounds[1]. This simple excitation signal causes the poor quality of synthetic speech such as the buzziness of voiced sounds and so forth. In this paper, we use the multiband excitation signal which is proposed in speech analysis/synthesis part on voiced sounds of TTS. Also we classify the voiced sounds by cepstral distance for reducing overhead that is required for multiband excitation signal.

In section 2, we briefly summarize our TTS system. In section 3, our grouping result is described. In section 4, our experimentation result and estimation is discussed. Section 5 concludes the paper with a discussion.

2. IMPROVE THE QUALITY OF SYNTHESIZED SPEECH BY MULTIBAND EXCITATION SIGNAL

Typically, in order to synthesize speech, the excitation signal consists of a periodic impulse train in voiced regions or random noise in unvoiced regions. This excitation signal is then filtered using the estimated system parameters. Even though vocoders based on this above-mentioned speech synthesis system have been quite successful in synthesizing intelligible speech, the vocoders have not been successful in synthesizing high quality speech. The example of poor quality of synthesized speech is 'buzziness'. Many researchers have noted that buzziness of synthetic speech could be caused by the use of a purely periodic excitation for segments of natural speech that have turbulent noise-like spectral regions.

Fig.2 shows the spectrum of voiced original speech. We can see the harmonics and noise-like spectrum component in this figure. We ought not to disregard this noise-like component for high quality synthesized speech[2].

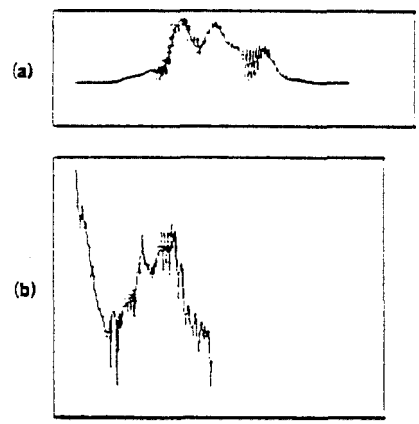

Fig. 2 This figure shows the original data : (a)Waveform, (b) spectrum

Therefore we use the multiband excitation sinal. Fig.3 describes the procedure to obtain the multiband excitation signal. This processing follows the proposed method in speech analysis/synthesis fields. And fig.4 shows the multiband excitation signal which is used in our system. The general spectrum

of synthesized speech on voiced, $|Sp'(\omega)|$ is obtained by $|Sp'(\omega)| = |Ip(\omega)| \cdot |Se(\omega)|$ where $|Ip(\omega)|$ is impulse spectrum and $|Se(\omega)|$ is spectral envelope. But we have the specturm of syntehsized speech on voiced, $|Sp'(\omega)| = |$(e) in fig.3$| \cdot |Se(\omega)|$ where (e) is obtained by mixing (c) and (d) in accordance with (b) - V/UV information of (a).

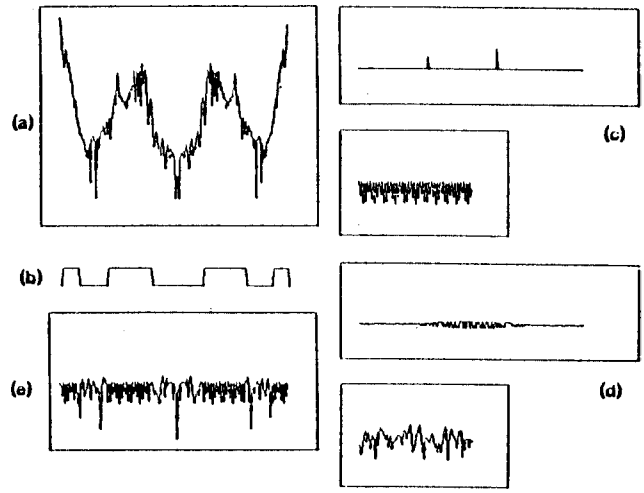

Fig. 3 The procedure of obtaining the mixed excitation signal (a)spectrum (b)V/UV info. (c)impulse spectrum (d)noise specturm (e)mixed spectrum

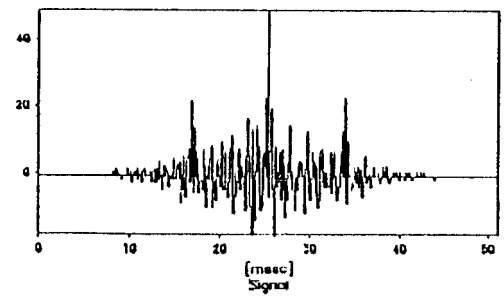

Fig.4 Multiband Excitation Signal

The summary of our system is as follows :

- A/D Conversion : 5KHz LPF, 10KHz Sampling, 12bit Quantization
- Analysis Parameter : Improved Cepstrum
- Analysis(Vocal Tract Approximation)
  - Frame period : 10ms for male, 5ms for female
  - Window function :
    - W - 25.6ms Blackman window for short-term predictor
    - Wp - 40ms Blackman window for long-term predictor
  - Cepstrum order : 30 for male, 25 for female
  - Spectral envelope : by Improved Cepstral method[3]
  - V/UV determinant : by Spectrum Envelope parameter
  - Pitch detection : by Cepstrum peak picking[4]
- Synthesis filter : LMA(Log Magnitude Approximation) filter[5]
- Excitation signal :
  - White noise for unvoiced
  - Mixed signal(Impulse + Noise) for voiced

3. GROUPING OF VOICED SOUNDS FOR THE REPRESENTATIVE EXCITATION SIGNAL

Preliminary results indicate that high quality reproduction can be obtained with this speech synthesis system for both clean and noisy speech without the "buzziness".

The above-stated system has many multiband excitation signals correspondent to each voiced sound. In this case, large memory size is required for excitation signals. Thus we classify the voiced sounds by level of spectral distortion to reduce overhead memory.

Spectral distortion measure is cepstral distance measure. It is defined as follows :

$$Cd = 10\ln 10 \sqrt{2 \sum_{i=1}^{M} (c^{v1}[i] - c^{v2}[i])^2}$$

v1 ≠ v2 : voiced sound

It is known as good measure of spectral distortion. Table 1 display the cepstral euclidian distance between two voiced sounds. Table 2 shows the groups of voiced sounds. The classification threshold value is about 2.5dB. Experimentally, the fact that the representative excitation signal of the same group's voiced sounds is used as excitation signal on synthesis, does not affect the quality of synthesized speech.

Table 1. The Cepstral distance of the voiced sounds

	a	ɐ	o	u	i	ɨ	ɵ	ɛ	j	ɳ	ɲ	ɯ	ɟ	ɰ
a	-	2.350	5.596	14.227	9.792	15.113	15.852	15.380	13.177	14.406	7.124	14.141	6.699	14.213
ɐ		-	1.760	11.327	7.286	12.956	13.687	13.857	12.432	12.155	4.174	14.591	4.189	12.784
o			-	11.741	6.029	14.288	14.955	15.344	13.035	14.061	4.321	15.757	4.707	13.727
u				-	7.460	14.119	14.509	14.774	11.827	12.522	3.364	13.360	4.850	12.915
i					-	13.993	13.321	13.924	7.006	7.264	5.181	8.683	2.286	10.521
ɨ						-	0.183	0.171	3.145	2.294	2.547	4.430	0.917	6.690
ɵ							-	0.099	3.577	3.104	2.315	4.935	0.621	7.579
ɛ								-	3.492	3.067	3.211	4.834	0.966	7.962
j									-	2.348	2.968	2.253	4.354	2.904
ɳ										-	1.491	3.180	3.644	2.657
ɲ											-	3.829	2.619	3.575
ɯ												-	5.738	4.934
ɟ													-	7.749
ɰ														-

Table 2. The Groups of the voiced sounds

Group 1	j, ɛ, ɸ, e
Group 2	u, ɐ, o
Group 3	ɨ, ɳ, ɰ
Group 4	n
Group 5	m
Group 6	a
Group 7	i

4. IMPLEMENTATION AND ESTIMATION

Fig.5 shows the waveform of TTS using the representative excitation signal of each group. (a) is the original, (b) is the synthesized speech by impulse train, (c) is the synthesized speech by mixed excitation signal. We can find the (c) is better than (b) in this figure. The methode of using the multiband excitation signal is so useful for variation of the pitch in TTS. Fig.6 shows the waveform of synthesized speech which is obtained by multiplication of original pitch value by 1.5. (a) is the synthesized speech by impulse train, (b) is by multiband excitation signal. In the case of (a), as pitch period is longer, synthesized speech's quality is poorer. But in the case of (b), although pitch period is longer, synthesized speech's quality is not especially distorted.

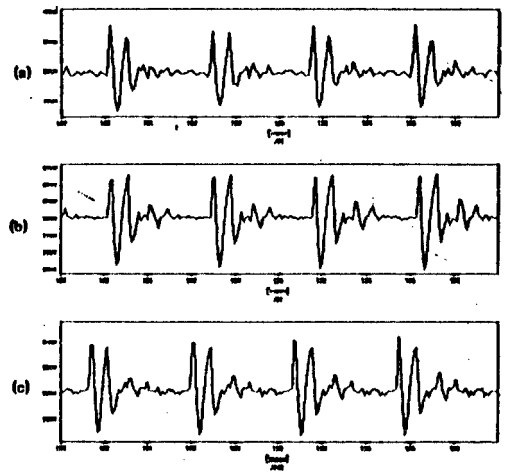

Fig. 5 Synthesized speech waveform (a)original data (b)by impulse (c)by multiband excitation signal

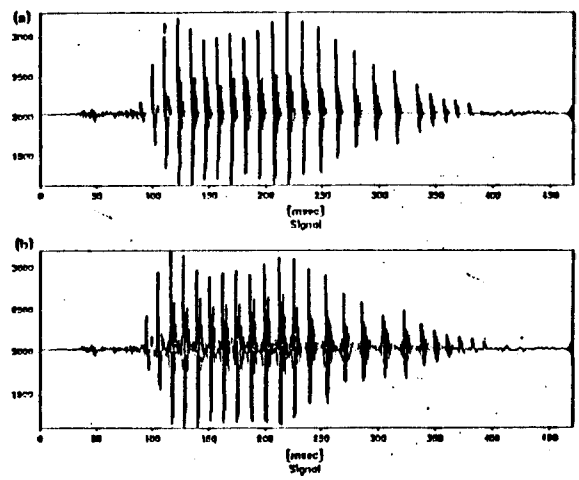

Fig. 6 Synthesized speech waveform is obtained by variation of the pitch (a)by impulse (b)by multiband excitation signal

Fig.7 is the example of TTS synthesized by the mixed excitation signal. (a) is the example of TTS by impulse, (b)is by the mixed excitation signal. It is the waveform of /annjəθhasim/- a part of synthetic speech /annjəθhasimnikʻa/. According to the observations on synthetic waveform and the result of subjective test, we conclude that the method using the multiband excitation signal is better than that of using the impulse train.

(a)

(b)

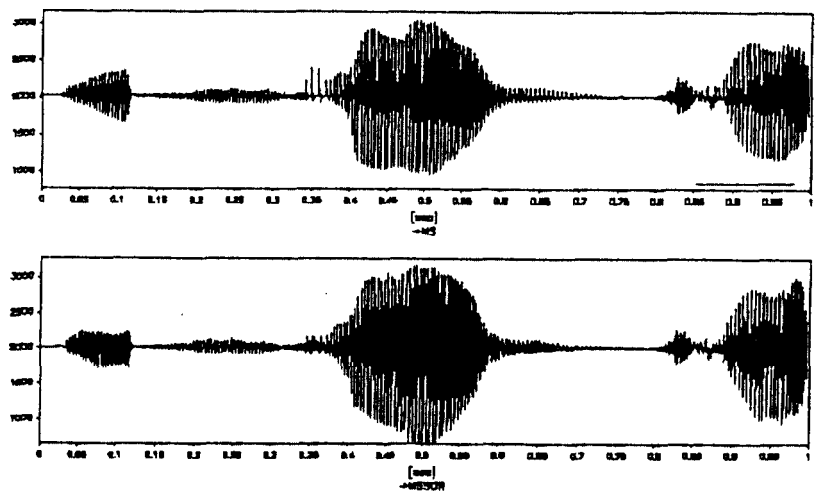

Fig. 7 The example of TTS synthesized speech (a)by impulse (b)by multiband excitation signal

5. CONCLUSION

In this paper, we inquire the problems of the common TTS and propose the solution. The solution is the application of the multiband excitation signal to TTS. As the result of experiment, it is found that the application improves the synthetic speech quality, eliminates the "buzziness". In addition, it reduces the spectral distortion that is caused by the variation of pitch period in TTS. Also, we classify the voiced sounds to reduce overhead memory by cepstral euclidian distance measure.

REFERENCES

[1] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, MARCEL DEKKER, 168-169 (1992).

[2] Denial W. Griffin, et al., "A High Quality 9.6kbps Speech Coding System", ICASSP, 125-128 (1986).

[3] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method.", Trans IECE Vol.62-A No.4, 217-223 (1979).

[4] A. Michael Noll, "Cepstrum Pitch Determination", J.Acoust.Soc.Amer., 179-195 (1967).

[5] S. Imai, et al., "Log Magnitude Approximation (LMA)filter", Trans. IECE Vol.J63-A No.12, 886-893 (1981).