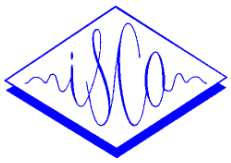

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

POSSIBILITY OF SPEECH SYNTHESIS BY COMMON VOICE SOURCE

Satoshi YUMOTO, Jouji SUZUKI, Tetsuya SHIMAMURA

Dept.of Information and Computer Sciences, Saitama University,
Shimo-okubo 255, Urawa, 338 Japan

ABSTRACT

In order to improve naturalness of synthesized speech, this paper intend to utilize common residual wave as voice source of Japanese typical sentences which have the same intonation or accent pattern. An analysis-synthesis with 14 cosfficients are amployed in this experiment. The synthetic experiments are conducted on some words. At first, a residual wave extracted from a word, are used as excitation source waves for other words. Secondly, fundamental frequency and intensity extracted by the original words are used to generate source waves. This synthesized sounds are less noisy than aforesaid sounds. In the last, fundamental frequency is generated by Fujisaki-model. This paper suggests that there are some possibilities to use common voice source for synthesis.

I. INTRODUCTION

In speech synthesis from text by use of vocoder type synthesizer, voiced source wave is generated from contour of fundamental frequency (F_0) and intensity pattern. Usually, naturalness of synthesized speech is poorer than that of human speech. Contour of F_0 and waveform of source wave intensity pattern affect quality of synthesized speech.

Residual source wave extracted by the inverse characteristics of voval tract keeps prosodic features of voice.[1] In Japanese, contour of fundamental frequency is a very important factor, because accents of Japanese are mainly ruled by of fundamental frequency. This paper intend to use residual wave as voice source of Japanese typical sentences which have the same intonation or accent pattern to improve naturalness of synthesized speech. This paper shows experiments on Japanese words whose accent type is common as a preliminary experiment.

II. SPEECH MATERIAL

Fourteen Japanese words (shown in Table 1) which have the same accent types and consist of voiced sounds are prepared. All words consist of three voiced syllables, and have two types of accent. Accent patterns are"$\underline{O}\overline{O}\overline{O}$" and "$\underline{O}\overline{O}\overline{O}$".

These words are utteranced by two males(speaker A,B) who speak Tokyo dialect.[2] These speech signals are quantizated in 16 bits at the sampling frequency of 9450Hz after band-limitation to 3.4kHz. Then, 16 of PARCOR coefficients are calculated at every 10ms.

Table 1: Accent Pattern and meaning of Japanese words

word	$\underline{O}\overline{O}\overline{O}$	$\underline{O}\overline{O}\overline{O}$
aoi	blue	name of the plant
ueru	be hungry	plant(*verb*)
noroi	slow	curse
nobiru	extend	name of the plant
omori	baby-sitting	weight
iwaba	so to speak	the rocks
jowai	weak	age

Table 2: Scales and comments

5:It is heard like the original speech.
4:It is good as synthesized speech.
3:It is acceptable.
2:It is hard to hear.
1:It is anacceptable, or heard as another word.

III. EXPERIMENTS AND DISCUSSIONS

First, all words are synthesized by residual waves which are extracted from "*aoi*" of speaker A, and PARCOR coefficients of each words.[3] Synthesized words

are arranged in random, and presented to 10 students through a headphone in a sound treating room. Quality of synthesized sounds are evaluated by MOS(mean opinion score) of 5 scales. These scales are shown in Table 2. If it is heard as another word, its evaluation scale is 1. The MOS are shown in Table 3. At "====" of Table 3, original speeches are reproduced, because they are synthesized by residual wave of speaker A. For example, if the residual wave of the original speech is "*aoi*($o\bar{o}o$)" of speaker A, and PARCOR coefficients of original speech is "*ueru*($o\bar{o}\bar{o}$)" of speaker A, MOS is 3.3 as shown in a frame in Table 3. Hearing test shows possibility to use common residual wave for words which have the same accent. Even if residual wave is exchanged for other man's, quality of sounds are almost the same. If the original speech of residual wave and that of PARCOR coefficients have different accent patterns, it has little effects on quality of sounds. Residual wave is extracted from words consisting vowels, therefore "*aoi*" and "*jowai*" which all syllables consist of vowels and semivowels are evaluated by high MOS.

Table 3: MOS of sounds synthesized from original residual waves

Residual wave	Speaker A			
Accent Pattern	*aoi* $O\bar{O}O$	*aoi* $O\bar{O}\bar{O}$	*aoi* $O\bar{O}O$	*aoi* $O\bar{O}\bar{O}$
PARCOR coefficients	Speaker A		Speaker B	
aoi($o\bar{o}o$)	====	3.60	3.30	2.70
aoi($o\bar{o}\bar{o}$)	3.70	====	3.80	3.60
ueru($o\bar{o}o$)	2.90	2.90	2.60	2.50
ueru($o\bar{o}\bar{o}$)	3.30	2.80	3.10	3.30
noroi($o\bar{o}o$)	3.40	3.40	3.30	2.50
noroi($o\bar{o}\bar{o}$)	3.40	3.30	3.30	3.20
nobiru($o\bar{o}o$)	2.60	2.60	3.40	2.00
nobiru($o\bar{o}\bar{o}$)	3.00	2.50	2.60	2.50
omori($o\bar{o}o$)	2.10	2.70	2.20	2.70
omori($o\bar{o}\bar{o}$)	2.40	3.00	2.60	2.60
iwaba($o\bar{o}o$)	2.10	2.00	2.50	1.20
iwaba($o\bar{o}\bar{o}$)	2.10	1.80	2.20	2.00
jowai($o\bar{o}o$)	3.40	3.20	3.50	2.90
jowai($o\bar{o}\bar{o}$)	3.60	3.20	3.80	2.90
Average	2.92	2.85	3.01	2.61

Table 4: Confusion Patterns

iwaba	→	*iwama,iwawa*
omori	→	*omoi,omoni*
nobiru	→	*noiru,noiu*
noroi	→	*nooi*
ueru	→	*ueu*

In case of confusion, accent errors haven't appeared, but consonants are confused, because intensity pattern is heard that is different from the original speech. For instance, amplitude of residual wave of "*a*" is larger than that of "*i*" which is heard as same intensity. This difference effects difference of intensity. Principal confusion patterns are shown in Table 4. It is shown that voiced plosive "*b*" is often confused. Therefore, in Table 3, MOS of "*nobiru*" and "*iwaba*" are low. Source amplitude is the same of original "*aoi*" of speaker A, then characteristics of amplitude of plosives can not be reproduced. Liquids are confused by the same reason.

Figure 1 is narrow band spectrogram of original "*nobiru* ($o\bar{o}o$)" of speaker A, and Figure 2 is that of synthesized speech from residual wave of "*aoi*($o\bar{o}o$)" of speaker A and PARCOR coefficients of "*nobiru*($o\bar{o}o$)" of speaker A. Synthesized waveforms indicate that intensity pattern is not reproduced in synthesized speech. Spectrograms show that harmonics are reproduced, but the plosive "*b*" and the liquid "*r*" are not reproduced.

In this experiment, synthesized sound is heard noisy, because the length of source wave is different from the length of coefficients of other words or the lack in whitening the residual signal. Then, frame period of PARCOR coefficients must be varied to make adjustment of length and to decrease noise, and whitening of residual signal should be done. In the other way, this problem will be reduced by a pitch excited analysis-synthesis systems.

F_0 is calculated by auto-correlation function of residual signal on two accent types "*aoi*". Speech synthesis is carried out by Rosenberg wave as glottal wave, and by employing these F_0 contours and intensity patterns. They are evaluated by the aforesaid way. The MOS of this experiment is shown in Table 5. In this experiment, synthesized words are less natural than that of the preceding experiment, and more easy to understand. These synthesized words are less noisy, but confusion often happens on consonants, because of difference of intensity level which depend on the difference of articu-

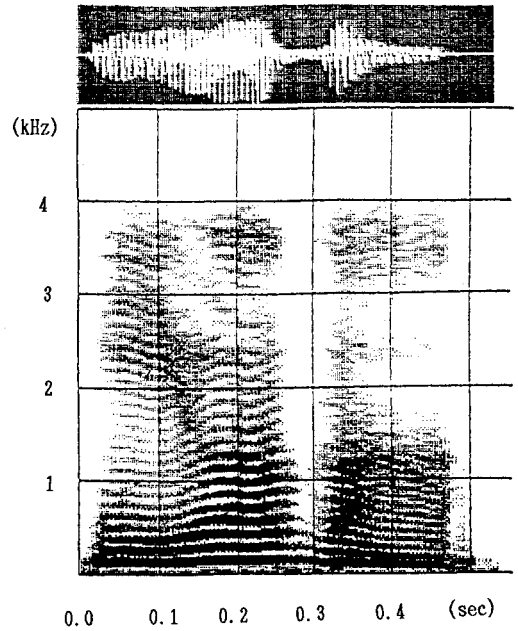

Figure 1: Spectrogram of original "*nobiru*($o\bar{o}o$)" of speaker A

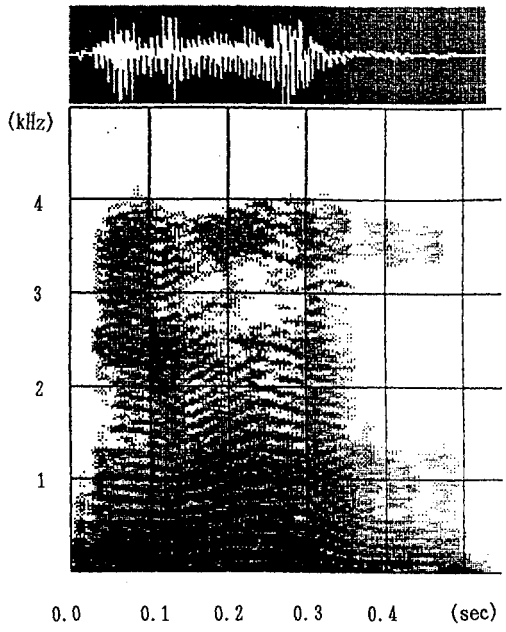

Figure 2: Spectrogram of synthesized "*nobiru*($o\bar{o}o$)" by residual wave of "*aoi*($o\bar{o}o$)" and PARCOR coefficient of "*nobiru*($o\bar{o}o$)"

Table 5: MOS of sounds synthesized by source wave generated by fundamental frequency and intensity level

Prosodic Features	Speaker A			
Accent Pattern	*aoi* $O\bar{O}O$	*aoi* $O\bar{O}\bar{O}$	*aoi* $O\bar{O}O$	*aoi* $O\bar{O}\bar{O}$
PARCOR coefficients	Speaker A		Speaker B	
aoi($o\bar{o}o$)	4.20	3.70	4.30	3.30
aoi($o\bar{o}\bar{o}$)	4.40	4.10	4.30	4.00
ueru($o\bar{o}o$)	2.20	3.20	3.00	2.70
ueru($o\bar{o}\bar{o}$)	2.30	3.10	3.70	3.40
noroi($o\bar{o}o$)	3.20	3.40	2.90	2.70
noroi($o\bar{o}\bar{o}$)	3.10	3.70	3.00	3.90
nobiru($o\bar{o}o$)	2.70	2.10	2.50	2.60
nobiru($o\bar{o}\bar{o}$)	3.30	2.30	2.40	3.00
omori($o\bar{o}o$)	2.80	3.40	3.30	2.90
omori($o\bar{o}\bar{o}$)	2.40	2.90	3.90	3.30
iwaba($o\bar{o}o$)	3.00	2.60	3.20	2.60
iwaba($o\bar{o}\bar{o}$)	2.40	2.70	3.20	3.00
jowai($o\bar{o}o$)	3.60	3.90	4.00	3.40
jowai($o\bar{o}\bar{o}$)	3.90	3.30	4.10	3.80
Average	3.11	3.17	3.41	3.19

Table 6: MOS of sounds synthesized by source wave from contaur of frequency by Fujisaki-model and original intensity level

Source Ampritude	Speaker A			
Accent Pattern	*aoi* $O\bar{O}O$	*aoi* $O\bar{O}\bar{O}$	*aoi* $O\bar{O}O$	*aoi* $O\bar{O}\bar{O}$
PARCOR coefficients	Speaker A		Speaker B	
aoi($o\bar{o}o$)	3.30	3.20	3.50	3.30
aoi($o\bar{o}\bar{o}$)	4.10	4.30	3.70	3.60
ueru($o\bar{o}o$)	2.60	3.10	3.10	2.20
ueru($o\bar{o}\bar{o}$)	2.60	2.50	3.10	2.50
noroi($o\bar{o}o$)	3.60	3.70	2.50	2.90
noroi($o\bar{o}\bar{o}$)	3.80	3.60	2.50	3.60
nobiru($o\bar{o}o$)	1.80	2.10	1.50	2.30
nobiru($o\bar{o}\bar{o}$)	2.00	2.40	1.40	2.30
omori($o\bar{o}o$)	3.10	2.80	3.30	3.10
omori($o\bar{o}\bar{o}$)	2.80	2.80	3.80	3.70
iwaba($o\bar{o}o$)	2.60	2.40	2.60	2.10
iwaba($o\bar{o}\bar{o}$)	2.80	2.70	2.90	2.70
jowai($o\bar{o}o$)	3.30	3.40	3.20	2.90
jowai($o\bar{o}\bar{o}$)	3.40	3.40	3.70	3.10
Average	2.98	3.03	2.91	2.88

lation.

In the last, contour of fundamental frequency of "$\underline{Q}\overline{O}\underline{Q}$" and "$\underline{Q}\overline{O}\overline{O}$" is generated by Fujisaki-model.[4] This frequency is adjusted to a mean of speaker A and B. Source amplitude is extracted from the original speech of "*aoi*" utteranced by speaker A. Rosenberg wave is used as glottal wave.

They are evaluated by the aforesaid way. The MOS of this experiment are shown in Table 6. In this experiment, opinion scales are lower than the second experiment. It is considered that the fluctuation of fundamental frequency is not reproduced in this way.

IV. CONCLUSION

This paper shows some possibility to use common voice source for synthesis of words. The method which uses fundamental frequency and intensity level extracted from residual waves is promising method. Less noisy sounds are synthesized by Rosenberg wave than by original residual wave. Synthesized sounds utilized prosodic feature of original speech are more natural than that by use of generated contour of fundamental frequency. And more natural sounds are synthesized by prosodic features of original speech than by synthesized glottal wave. Whitening of residual wave require further examination. Algorythm generating intensity pattern should be modified to synthesize more natural speech. Especially, if phonemes of original wave of PARCOR coefficients are plosives or liquids, modification of intensity pattern is needed. Furthermore, adjustment of length which correspond to syllables is needed.

We will apply this method to Japanese typical sentences like *haiku* or *tanka*.

REFERENCES

[1] T.Kamai, K.Matsui, S.Pearson, M.Murata, N.Morinaga: "Formant Synthesis System Using Natural Voicing Source", Full Meeting of Acoustic Society of Japan, 1-6-15, 1990/9 [in Japanese]

[2] Editer;NHK: "Japanese Accent Dictionary", Japan Broadcasting Corporation pub., 1985/6 [in Japanese]

[3] T.Aguin, M.Nakajima: "Voice Transaction with Computer", Akiba pub., pp56-70, 1980/6 [in Japanese]

[4] H.Fujisaki, H.Sudo: "A Model for the Generation of Fundamental Frequency Contours of Japanese Word Accent", Journal of Acoustic Society of Japan, pp445-453, 1971/9 [in Japanese]