SPEECH SYNTHESIS BY RULE USING SYNTHESIS UNITS CONSIDERING PROSODIC FEATURES

Yasushi Ishikawa* and Kunio Nakajima*

*Computer and Information Systems Laboratory
Mitsubishi Electric Corporation

ABSTRACT - This paper describes on synthesis units for text-to-speech synthesis. A kind of synthesis unit and extraction of units are very important problems in speech synthesis by rule. Many kinds of units have been proposed, VCV, CVC, demi-syllable, tri-phone are typical units for Japanese text-to-speech system. Recently, non-uniform units or context-dependent units have been proposed, and good results were reported. However such a kind of unit is considered only phonetic context as a factor of spectral variation. Our basic idea is introducing prosodic features into control of spectral features in order to realize natural sounded synthetic speech. In this paper, we report results of basic analytic experiments. The results show that there is obvious relation between prosodic features and spectral features, and that spectral control method considered not only phonetic context but also prosodic feature is able to improve quality of synthetic speech in text-to-speech system.

Introduction

In a text-to-speech system, it is very important to control spectra for realization of coarticulation and generation natural sounded speech. A simple method for this problem is to provide longer synthesis units that include coarticulation pattern. VCV syllable, CVC syllable, demi-syllable, tri-phone, quadri-phone are typical synthetic units for a text-to-speech system (Sagisaka, Sato 1986) (Ichikawa, *et.al.* 1987) (Salza, *et.al,* 1987). In general, the quality of synthetic speech based on such units is better than synthesized speech from short units. And recently non-uniform units or context dependent units have been proposed, and good results were reported (Sagisaka, 1988) (Hakoda *et.al.* 1990). On the other hand, an approach to control spectra at concatenation of synthetic units has been proposed in order to generate natural coarticulation and realize a variety of pattern which depends on speaking rate or manner of speaking (Ishikawa, Nakajima 1991).

However in these works, only realization of coarticulation is focused in spectral control. In other words, only phonetic context is considered as a factor of spectral variation in natural speech. It is fact that phonetic context is one of the most important factors, but, other factors, for example, pitch period, stress, speaking rate, should be considered in order to improve the quality of synthetic speech.

Our basic idea is introducing prosodic features into spectral control as factors of spectral variation. As spectral control method based on prosodic features two types of methods are considered. One is a method based on phonetic models which generate or modify spectral parameters with a pitch period or energy generated by rules in a text-to-speech system. Another is a method that provides a set of synthetic units which includes units differ with pitch, energy or speaking rate, and selects an optimal unit according to generated prosodic parameters. Although our final goal is the former method. as a first experimental work, we have carried out analytic experiments in order to consider a method based on synthetic units considering prosodic features.

We have reported results of basic experiments on relation between spectral features and prosodic features using clustering of CV syllables which were extracted from natural speech utterances (Ishikawa, Nakajima 1994), In this paper we discuss on synthetic units for Japanese Text-to-Speech system in detail.

Clustering of CV Syllables

Since CV syllable is one of the typical synthesis unit for a Japanese text-to-speech system (Ishikawa, Nakajima 1985), we use the syllable as a unit of analysis. For evaluation the relation between prosodic parameters and spectra, clustering of CV syllables which are detected from natural speech utterances has been done. To reduce an effect of pitch period in analysis, improved cepstral analysis is used. In clustering we used Mel-log spectral distance in 0.6 - 5.0 kHz as spectral distortion, it also reduce an effect of pitch frequency and vocal source component (Figure 1), and linear matching with constraint of CV boundary as distortion between CV syllables (Figure 2).

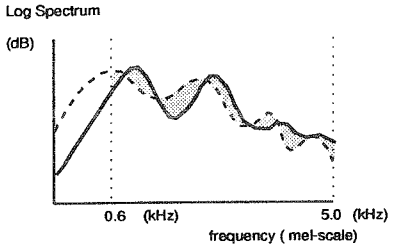


Figure 1. Spectral Distortion

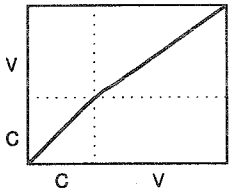


Figure 2. Matching of CV Syllables

At first, an initial codebook is generated by averaging all syllables extracted from utterances for each CV syllable. This codebook means the synthesis units which have only one segment each CV syllable. The average distance between a codeword and members of a cluster is calculated, and a cluster with maximum distortion is selected. Then classification of the cluster is performed using methods.

1) Previous Phoneme
A cluster is divided into two clusters under constraint of previous phoneme. CV syllables with same privious phoneme will be members of one of the new clusters. Namely, a set of CV syllables with previous phonemes set V is divided into new sets with previous phoneme set V1 and with phonme set V2 where V1 and V2 have no common element Sets V1 and V2 which give minimum average distortion between new codewords and each member are selected

2) following Phoneme
A cluster is divided into two clusters under constraint that CV syllables with same following phoneme will be memers of one of the new clusters.

3) Pitch Frequency
A cluster is devided with a pitch frequency (Figure 3) .

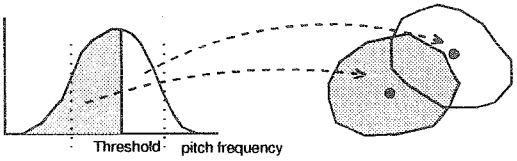


Figure 3. Classification with pitch frequency

A pitch frequency at center of vowel segment is detected. CV syllables with pitch frequency which is lower than a threshold value will make a new cluster, and others will make another cluster. The

threshold which gives minimum distortion is selected in the range that a number of cluster members of one of the new cluster will be greater than one fourth and less than three forth of original cluster size.

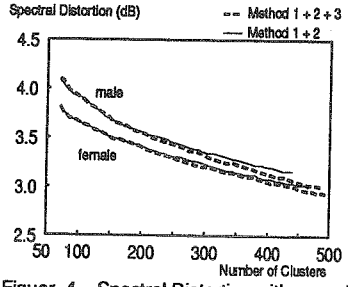
4) Energy
A cluster is devided with a cepstral 0th coefficient at center of vowel segment that means energy of speech. Classification method is same as classification with pitch frequency.

Classifications by these methods are performed and classification that minimized distortion is selected, then new clusters and codewords is added. If we use classifications 1) and 2), a generated codebook is an optitimum synthesis units considered on phonetic context, and if 3) or 4) is used adding 1) and 2), the codebook means synthesis units also considered on a pitch period or energy.

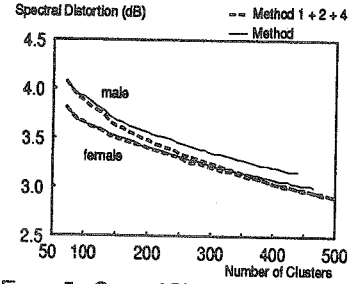
Experiments 1

About 3,000 CV syllables that extracted from 115 sentence utterances are used as test data. Two sets by a male speaker and a female speaker have been analyzed Clustering Experiments were carrid out with three conbination of clustering methods.
1) + 2) .... context dependent.
1) + 2) + 3) .... context and pitch frequency dependent.
1) + 2) + 4) .... context and energy dependent.

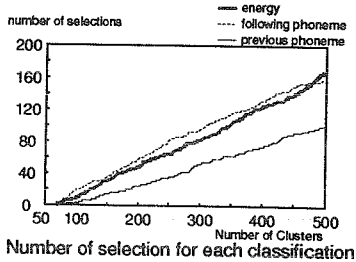
The spectral distortion with a number of clusters are shown in Figure 4 and 5. These figures show that the clustering method considered on pitch frequency or energy is better than the context dependent method, and that best performance is obtained by the context and energy dependent clustering method. The numbers of selections for each classification with a number of clusters are shown as Figure 8 in the case of phonetic context and energy dependent clustering on male utterances. Figure 7 and 8 show spectral distortion with a codebook size grouped with vowels.



Figuer. 4 Spectral Distortion with a number of cluster (context dependent, context and pitch frequency dependent methods)



Figuer. 5 Spectral Distortion with a number of cluster (context dependent, context and energy dependent methods)



Figuer. 6 Number of selection for each classification method.

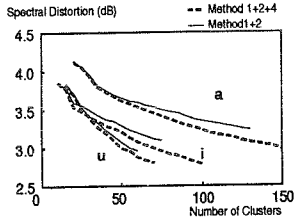


Figure 7. Spectral Distortion with a number of codebook (/C-a/, /C-i/, /C-u/)

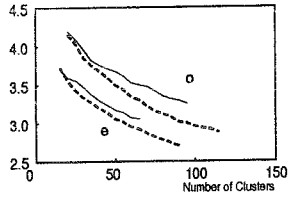


Figure 8. Spectral Distortion with a number of codebook (/C-e/, /C-o/)

Discussion

Figure 4 and 5 show that there is an obvious relation between spectral feature and a pitch frequency, energy. Although it is not simple to decide which is a major factor of spectral variation, since a pitch frequency and energy of speech correlate to each other especially in Japanese, the stronger relation was observed between energy and spectra in this experiement. In this experiement, the energy dependent clustering is more effective in a male speaker. It is thought that one of the reason is speaking rate, this male speaker uttered faster and conversationally in comparison with the female speaker.

As shown in Figure 7 and 8, it is clear that the spectral variation with energy for each syllables is different. From the experiment, vowels, /a/ and /o/ vary with energy obviously, rather than /i/, /e/ and /u/.

EXperiment 2

The results of the experiments described above show strong correlation between spectra and prosodic features. However, it is difficult to apply the obtained codebook as synthetic units, since classification with energy or pitch frequency is performed many times, the codebook is not always optimal for open data. And in a synthesis by rule method, selection rules that decide a segment based on phonetic context and prosodic features must be simple. We have carried out other experiments in order to consider selection rules. In this experiements, we used a clustering method: 1) clustering with energy is applyed on an initial CV syllable codebook, and in muxmum N codowords are obtained from each codeword. (N equals one means clustering with energy is not performed). If clustering with energy increases average distortion, clustering is not performed, 2) on the codebooks obtained by clustering with energy clustering with phonetic context are performed. The codebook which obtained by this clustering method can be considered as synthetic units to be used in a synthetic method which selects CV segment from N segments based on phonetic context and energy thresholds that are common to CV syllables. The results of this experiment distortion are shown in Figure 9.

The results show once or twice clastering with energy decreases spetcral distortion, but performance is not better than without clustering with energy at the same codebook size. One of the reasons is considered that the basis of this clustering method which performes clustering with energy at first is thought that energy is the major factor of spectral variation. And other reason is spectral variation with prosodic features is not distinct in a consonant. Then the experiment in which average distance only in a vowel segment is used as distortion between two CV syllables was carried out. The results is shown in Figure 10. In this figure a solid line, which shows the distortion when N equals one, stops at a codebook size of 120. It means that the distortion of vowel segments does not decrease with any context dependent clustering any more. Figure 11 shows the distribution of energy thresholds with energy of CV syllables when N equals two.

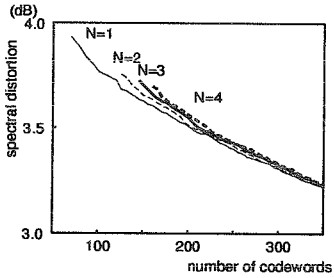


Figure 9. Spectral distortion with a number of cod book

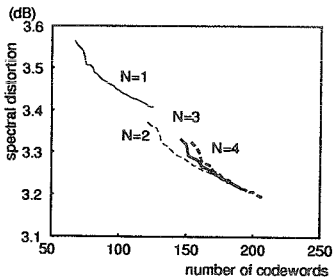


Figure 10. Spectral distortion of vowel segment with a number of cod book

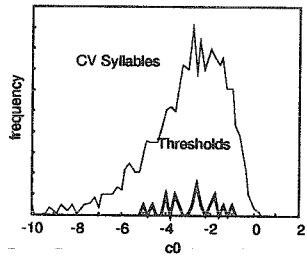


Figure 11. Distribution of C(0) threshold and C(0) of test CV syllables

It is obvious that clustering with energy improves greately the distortion of vowels, and that only once clustering is sufficiently effective.

Conclusion

Using clustering of CV syllables that detected from sentence utterances, we analyzed the relation between spectra and prosodic features. From experiments the obvious correlation was observed between spectral features and energy of speech. The experiments that were carried out in order to consider unit selection rules in synthesis by rule show that clustering with energy decreases spectral distortion of vowel segments dramatically. It strongly suggests that spectral control based on prosodic parameters, for example, synthesis method which provides two or three segment with energy as a synthesis unit will give good quality of synthetic speech in a text-to-speech system.

Further studies need to be done in the future: to analyze the detailed relation between spectra and prosodic features including speaking rate, to develop a synthetic scheme with controlling spectra with prosodic features., and to do auditory evaluation.

REFERENCES

Hakoda,H., *et.al* .(1990) *"A New Japanese Text-to-Speech Synthesizer based on COC Synthesis Method"*, Proc. ICSLP '90, pp. 809 - 812

Ichikawa,M. *et.al.* (1987), *"Phoneme Sequence Unit for Speech Synthesis by Rule"*, Trans. Comm. on Speech Res. Acoust. Soc. Jpn. Sp-87-6-1987 (in Japanese)

Ishikawa,Y. and Nakajima,K., (1985) *"Text-to-Speech Synthesis of Japanese based on Cepstrum-CV Method"*, autumn meeting of ASA, 1985

Ishikawa,Y. and Nakajima,K., (1991) *"Neural Network Based Spectral Interpolation Method for Speech Synthesis by Rule"*, Proc. EUROSPEECH 91, pp. 47 - 50

Ishikawa,Y. and Nakajima,K., (1994) *"On Synthesis Units for Japanese Text-to-Speech Synthesis"*, Proc. ICSLP '94, pp. 1751-1754

Sagisaka,Y., and Sato,H., (1986) *"Composite Phoneme Units for the Speech Synthesis of Japanese"*, Speech Communication Vol. 5, pp. 217 - 223.

Sagisaka,Y. (1988) *"Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units"*, Proc. of ICASSP'88 pp. 679-682

Salza,P.L.,, Sandri,S., and Forte,E., (1987) *"Evaluation of Experimental Diphones for Text-to-Speech Synthesis"*, Proc. of Euro. Conf. on Sp. Tech. Vol. I pp.63-66