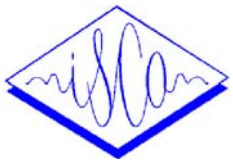

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

A Study of Clarity Control of Synthesized Speech with Prosodic Features and Phonemic Features

Noriki Fujiwara[*], *Makoto Hiroshige*[*], *Kenji Araki*[*] *and Koji Tochinai*[**]

[*]Graduate school of Engineering, Hokkaido University, Japan
[**]Graduate school of Business Administration, Hokkai Gakuen University, Japan
{fujiwara, hiro, araki}@media.eng.hokudai.ac.jp, tochinai@econ.hokkai-s-u.ac.jp

Abstract

In spontaneous conversational speech, all portions of speech do not always have high clarity. For example, the portions not having important information or the end of a sentence are not very clear. We consider that clarity of speech is controlled by F0, power, speech rate, place of articulation and so on. We consider that the clarity changes continuously, and change of clarity of speech produce a fluent rhythm in human speech. The purpose of our research is introducing the change of clarity into synthesized speech. In this paper, we try to control clarity of synthesized speech by post-processing of F0, power and formants. We evaluate the synthesized speech by auditory tests using SD method. The synthesized speech with control of clarity is better than the synthesized speech without control of clarity in several speech properties, e.g., calmness and smoothness.

10.21437/SpeechProsody.2004-125

1. Introduction

Speech synthesis is prospective method for natural human–computer communication. The studies on speech synthesis have been carried out for many years [1]. There are already high quality speech synthesis methods now [2]. Those systems aim to make more clear speech. However, in spontaneous conversational speech, human do not always speak very clearly. For example, the portions without important information or the end of a sentence are not very clear.

The contents of utterance or the speaking style effect prosodic factors and phonemic factors. Then the clarity of speech changes continuously. This change of clarity may produce a fluent rhythm or a tempo of speech. We aim to introduce this change of clarity into synthesized speech. In this paper, we try to control clarity of synthesized speech by post-processing of F0, power and formants. In Chapter 2, we explain our concept and the behavior of F0, power and formants in several speaking style. In Chapter 3, we explain about post-processing of F0, power and formants. In Chapter 4 and Chapter 5, we describe our auditory tests using re-synthesized speech by post-processing.

2. Our basic concept

2.1. Concepts relevant to clarity control and utterance generation

We have two basic concepts in this research. One concept is relevant to clarity control. In natural speech, prosodic factors and phonemic factors change with influence of the contents of utterance, the speaking style, and so on. We consider that the change of those factors effect the clarity of speech. We consider that this change of clarity may produce a fluent rhythm or a tempo of speech. The clarity of speech also changes continuously in speech. In other words, we think prosodic high-tension speech is loud, high F0, large change of speech rate and with clear phoneme utterance. We think prosodic low-tension speech is quiet, low F0, little change of speech rate and unclear articulation. We consider that the transition of prosodic tension produce a fluent rhythm or a tempo of speech.

The other concept is relevant to utterance generation. Conventionally, segmental (phonemic) features and supra-segmental (prosodic) features are separately investigated. However, we consider that some of phonemic features depend upon prosodic conditions, e.g., a word may be differently pronounced whether there is an emphasis or not. The prosodic features are F0, power, speech rate and so on. We consider that smooth utterance with the intelligibility is realized by adding the prosodic features in phonemic factor. We consider that the introduction of the prosodic features in phonemic factor realizes synthesized speech like natural speech.

In this study, we aim to introduce transition of prosodic tension and the prosodic features in phonemic factor into speech synthesis.

2.2. A preliminary study

As a preliminary study, we researched prosodic tension of natural speech in order to confirm existence of transition of prosodic tension.

2.2.1. Speech recording

We prepare five short sentences. We define one portion of a sentence as topic portion. We define other portions as non-topic portion. The speakers utter sentences in three speaking style. One speaking style is speaking whole sentence clearly. This style is called "whole clear" style. The second style is speaking topic-portion clearly and speaking non-topic portions relaxing as usual conversation. This style is called "topic only clear" style. The third style is speaking whole sentence relaxing as usual conversation. This style is called "whole relax" style. In these short sentences, a topic portion is defined always between two non-topic portions, i.e., a sentence is divided into 3 portions in order of a non-topic portion, a topic portion, and a non-topic portion. The each portion is called "preceding non-topic portion", "central topic portion" and "following non-topic portion". The speakers are two native Japanese who belong to student theatrical circle.

2.2.2. Analysis

We calculate "F0 maximum value", "RMS power average" and "the distance from neutral vowel" on preceding non-topic portion, central topic portion and following non-topic portion. The "RMS power average" is the value dividing the sum of RMS power by the number of the analysis frames. "The distance from neutral vowel" is defined a Euclidean distance

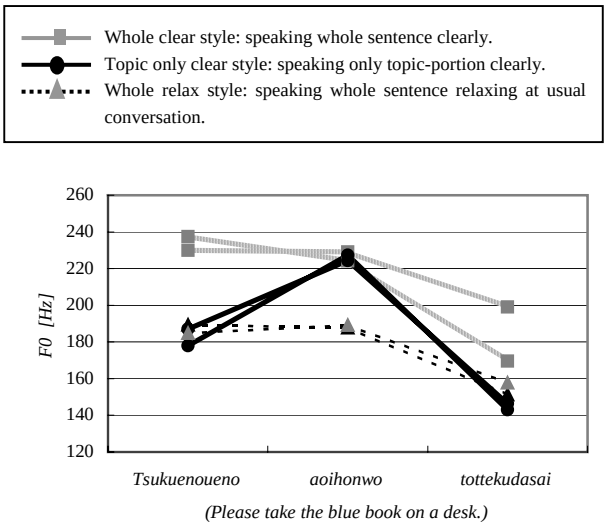

Figure 1: *F0 maximum value.*

between two spectra, i.e., a spectrum of a "neutral vowel" and that of current speech data. Actually, these spectra are expressed by sets of frequency value of lower 4 formants, i.e., F1, F2, F3 and F4. A "neutral vowel" is similar to schwa with a frequency spectrum. Furthermore, the formants of the "neutral vowel" are similar to the frequency response of uniform tube with yielding walls, friction and thermal loss shown (in Fiugre 2 [3]). From this frequency response and some unclear speech spoke by the native Japanese, we define that the formant frequencies of "neutral vowel" are F1=481, F2=1342, F3=2203 and F4=3316 [Hz].

2.2.3. Results and discussion

An example of the results is shown in Figure 1, Figure 2 and Figure 3. We pay attention to features of "topic only clear style" speech. In the "preceding non-topic portion" and the "following non-topic portion", the F0 maximum values in the "topic only clear style" are approximate to the F0 maximum values of "whole relax style". In the "central topic portion", the F0 maximum values in the "topic only clear style" are approximate to the values of "whole clear style". The features of RMS power average are same tendency of F0 maximum value. In the "preceding non-topic portion" and the "following non-topic portion", the values of RMS power average in the "topic only clear style" is also approximate to the values of "whole relax style". In the "central topic portion", the value of RMS power average in the "topic only clear style" is approximate to the values of "whole clear style". In the "preceding non-topic portion", the values of the distance from neutral vowel are approximate to the values of "whole relax style". In the "central topic portion" and "following non-topic portion", the values of the distance from neutral vowel are values approximate to the "whole clear style". The speaking style like "topic only clear style" is approximate to "whole clear style" in the topic portion, and approximate to "whole relax style" in the non-topic portion. The clarity of speech change in topic portion and non-topic portion. The speaking styles have different tendencies of prosodic features and phonemic features. Speaking style changes prosodic factors and phonemic factors. According to this preliminary study, we consider that the transition of prosodic tension exists in natural speech.

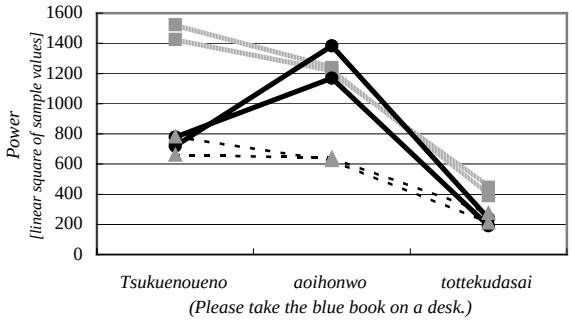

Figure 2: *RMS power average.*

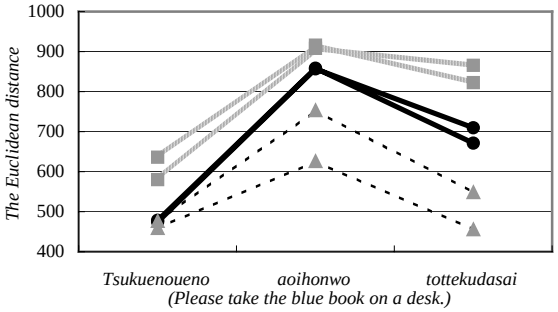

Figure 3: *The distance from neutral vowel.*

3. Processing synthesized speech

In this chapter, we describe processing method with the transition of prosodic tension and the prosodic features in phonemic factor. First, we investigate the prosodic factors and phonemic factors of natural speech controlled clarity [4]. Second, we describe processing method of the prosodic factors and phonemic factors of synthesized speech to control clarity.

3.1. Investigation of natural speech

We investigated features of natural speech spoken in several degrees of clarity. We prepare some words and some short sentences. One native Japanese utters these words and short sentences in several degrees of clarity. We investigated the speech on prosodic features and phonemic features. The remarkable changes between high clarity and low clarity are "planarization of F0 contour", "reduction of power" and "neutralization of spectrum". "Planarization of F0 contour" denotes that F0 contour becomes flat in whole speech. "Reduction of power" denotes that speech power become small. "Neutralization of spectrum" denotes that the formants of a speech become similar to the formants of a "neutral vowel".

3.2. Processing synthesized speech

We synthesized five short speech using "Visual Speech Creator ver. 1.21" made by NTT-IT Corporation [5]. The synthesized speech are analyzed and re-synthesized using ESPS/waves+ made by Entropic Research Laboratory, Inc. [6]. In re-synthesizing, we modify synthesized speech on F0, power and formants independently. We define these modified synthesized speech as "processed-speech". Also synthesized speech without modification on F0, power and formants are defined as "original-speech". The processing of power and

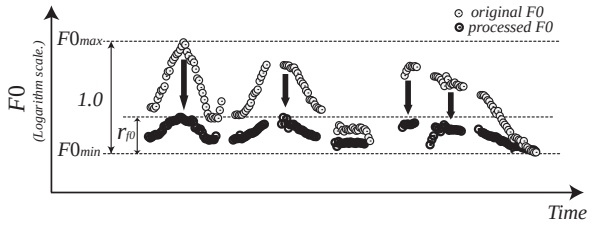

Figure 4: *The method of planarization of F0.*

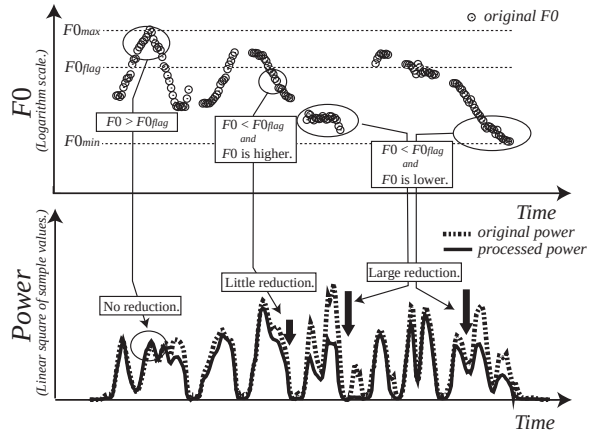

Figure 5: *The method of reduction of power.*

formants is synchronized with the F0 contour of original-speech when F0 of original-speech is lower than a constant threshold. We call the threshold "$F0_{flag}$". The "$F0_{flag}$" is defined by equation (1).

$$\log F0_{flag} = (\log F0_{\max} - \log F0_{\min}) \times 2/3 + \log F0_{\min} \quad (1)$$

where "log$F0_{max}$" denotes maximum value of F0 in logarithm, and "log$F0_{min}$" denotes minimum value of F0 in logarithm.

3.2.1. F0 control

Figure 4 shows an overview of the processing of F0 control. Processing of F0 control is performed by narrowing of the range of F0 (i.e., the range between the maximum and minimum on logarithm scale). In processing of F0 control, we do not modify the minimum value of F0. The range is narrowed by multiplying a constant value, r_{f0}.

3.2.2. Power control

Figure 5 shows an overview of the processing of power control. Processing of power control is performed by multiplying the power of original-speech by a variable. We make the transition of the variable correspond to the movement of F0 in original-speech. The variable is called r_{pow} The r_{pow} takes 1.0 as a maximum value and takes a constant as a minimum value. The constant is called r_{pow_min}.

3.2.3. Formant control

Processing of formant control is performed by making the formants similar to that of neutral vowel. The modification is performed when the F0 of original-speech is lower than $F0_{flag}$. Formant frequencies of neutral vowel are defined from [3] and the preliminary study with measuring of actual neutral vowel (F1=481, F2=1342, F3=2203, F4=3316 [Hz]). Processed formant frequency f_i' is derived form the following equation.

$$f' = f_{i \cdot lazy} + \left(f_i - f_{i \cdot lazy}\right) \times r_{for} \quad (2)$$

where the r_{for} denotes a variable to process formants to be closer to neutral vowel. We define that the r_{for} takes 1.0 as maximum value and takes a constant minimum value. The minimum is called r_{for_min}. The f_i denotes the i-th formant frequency of the original-speech, $f_{i \cdot lazy}$ denotes i-th formant frequency of neutral vowel.

In this paper, the r_{pow_min} set to be 0.5, r_{for_min} set to be 0.3 and r_{f0} set to be 0.5.

4. Auditory Test

We prepare two kinds of synthesized speech. One kind of synthesized speech has processed F0, power and formants by previous method in non-topic portions. These synthesized speech are called "processed speech". Another kind of synthesized speech has original F0, power and formants without modification. These synthesized speech are called "non-processed speech". Non-processed speech is also analyzed and re-synthesized by formant synthesizer. The reason of re-synthesizing is to give the degradation of speech quality to non-control speech, and we make that the conditions other than processing of clarity control are the same.

The auditory test using SD (Semantic Differential) method is carried out. The SD method is one of a method on evaluation of the impression of sound stimuli [7]. The subjects are 13 native Japanese who are graduate school students. The test is carried out with headphones in the quiet room. We use WWW browser with CGI program as the questionnaire. We prepare five non-processed speech stimuli and five processed speech stimuli. These ten speech stimuli are called a "stimuli-set". Speech stimuli are arranged at random order for every subject. In a rehearsal test, one stimuli-set is presented. In an actual test, three stimuli-sets are presented. An interval of speech stimuli is 60 seconds in the rehearsal test. In the actual test, an interval of speech stimuli is 50 seconds. We do not evaluate the results of the rehearsal test, since the aim of rehearsal test is that the subjects become familiar with operation. In SD method, the stimuli impressions are measured by opposite adjective pairs. Table 1 shows fifteen adjective pairs using in the auditory test. In the Table 1, "positive adjective" is an adjective with good impression generally, and "negative adjective" is an adjective with bad impression generally. The auditory test is carried out with Japanese adjectives. In the table, the upper word of a cell is Japanese adjective used in the auditory test, and the lower word of a cell is meaning of the adjective.

5. Results and discussion

We calculate the average of points selected by each subjects in three sets. We define the average as the subject's point about one adjective pair. Figure 6 shows the number of subjects whose point of processed speech is larger than point of non-processed speech. Figure 6 indicates that subjects feel "quit", "calm", "smooth" and "soft" impressions for processed speech. These adjectives express humanness. We consider that processing on F0, power and formants is able to bring synthesized speech close to natural speech.

Table 1: *Adjective pairs used in the auditory test.*

	positive	negative		positive	negative		positive	negative
1	ningenteki (humane)	kikaiteki (mechanical)	6	maruminoaru (smooth)	togetogeshii (sharp)	11	akarui (bright)	kurai (dark)
2	shizukana (quiet)	souzoushii (loud)	7	konomashii (desirable)	konomashikunai (undesirable)	12	utsukushii (beautiful)	kitanai (dirty)
3	ochitsuita (calm)	kandakai (shrill)	8	hakkirishita (clear)	bonyarishita (vague)	13	reiseina (cool)	kanjoutekina (emotional)
4	tokeatta (blended)	barabarana (scattered)	9	tyouwanotoreta (harmonious)	futyouwana (disharmonic)	14	wakariyasui (intelligible)	wakarinikui (unintelligible)
5	kokoroyoi (pleasant)	fukaina (unpleasant)	10	yawarakai (soft)	katai (hard)	15	hikishimatta (tight)	tarunda (loose)

(The upper word in a cell is Japanese adjective, the lower word in a cell is meaning of the adjective in English.)

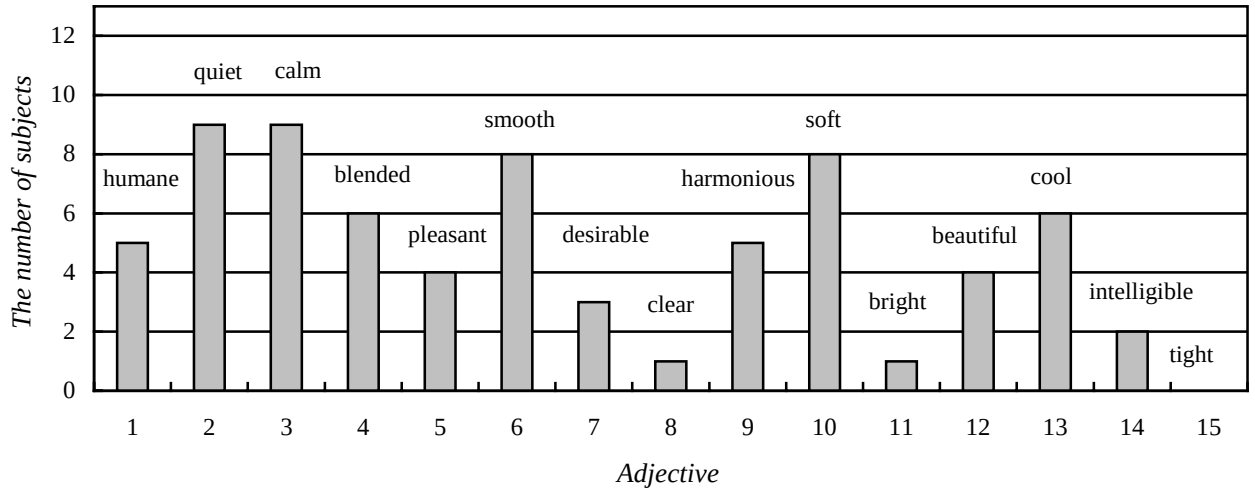

Figure 6: *The numbers of subjects whose point of processed speech is larger than point of non-processed speech.*

Conversely, Figure 6 indicates that subjects do not feel "clear", "bright" and "tight" impressions for processed speech. In other word, subjects have "vague", "dark" and "loose" impressions for processed speech. As one of reasons, we consider that the end of sentence has low clarity. According to the introspective reports of subjects after the auditory test, several subjects express the impression of whole speech by the impression of the end of sentence. The values of F0 at the end of sentence are the lowest in the speech. Since processing is synchronized with F0 contour, the clarity is low at the end of sentence. In the processing, the end of a sentence is processed too much unclear. Therefore, the impression of whole sentence may be too much unclear. To resolve this problem, we have to examine the amounts of processing and the improvement of processing method in future.

6. Conclusion

We have tried to control clarity of synthesized speech by post-processing of F0, power and formants. We have evaluated the synthesized speech by the auditory test using SD method. The synthesized speech with control of clarity have been better than the synthesized speech without control of clarity in several impressions expressed by several adjectives (e.g., calm, smooth). Since these adjectives express humanness, we have considered that post-processing of F0, power and formants are able to bring synthesized speech to be close to natural speech.

Future issues are improvement of processing method and refinement of the amounts of processing F0, power and formants.

7. References

[1] D.H.Klatt. 1980. Software for a cascade/parallel formant synthesizer, In *J.Acoust.Soc.Am.*67, 971-995.

[2] A.W.Black; P.Taylor, 1994. CHATR: a generic speech synthesis system. In *Proceedings of COLING'94*, volume II, 983-986.

[3] L.R.Rabiner; R.W.Schafer, 1978. *Digital processing of speech signals. Englewood Cliffs*. New Jersey: Pretice-Hall, Inc., 71-74.

[4] N.Fujiwara; M.Hiroshige; K.Araki; K.Tochinai. 2003. A proposal and fundamental study of clarity control in rule synthesis (in Japanese). *The spring meeting of the ASJ*. Kanagawa, 381-382.

[5] Software manuals of Visual Speech Creator ver. 1.21. 2000. NTT-IT.

[6] Software manuals of ESPS/waves+ with EnSigTM. 1997. Entropic Research Laboratory, Inc..

[7] N.Fujiwara; M.Hirosige; K.Araki; K.Tochinai, Clarity control of synthesized speech by post-processing using formant, power and F0 (in Japanese). Akita. *The autumn meeting of the ASJ*. 381-382.