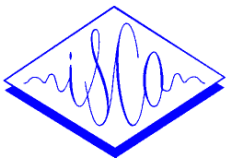

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

PROSODY GENERATION MODELS CONSTRUCTED BY CONSIDERING SPEECH TEMPO INFLUENCE ON PROSODY

Kazuhiko IWATA and Yukio MITOME

C & C Information Technology Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, 216 JAPAN

ABSTRACT

A duration model and a pitch pattern generation model are proposed, in which speech tempo influences on the duration and pitch are considered. Analysis results for the speech tempo influences are described. In the analysis, it was discovered that of all the factors affecting speech rates for individual phrases, at fast tempo, the phrase position within the sentence is the most influential factor, while, at normal and slow tempos, the most important factor is whether or not a pause exists after the phrase. The analysis also revealed that, while pitch frequency values may differ at different tempos, their normalized pitch patterns for a given sentence are quite similar. On the basis of these results, a duration model has been constructed, which determines a suitable tempo for a given sentence or paragraph, and estimates durations for the individual phrases that constitute the sentence or paragraph. The durations for the phonemes within the phrases are estimated according to the phoneme environment. A pitch pattern generation model has been also constructed, which determines the normalized pitch pattern that is little affected by changes in speech tempo. The model then calculates the pitch frequencies which would actually be produced at various tempos. These models have speech tempo parameters, and can generate adequate durations and pitch contours according to the tempo.

10.21437/ICSLP.1992-139

1. INTRODUCTION

Text-to-speech synthesis can be utilized for many applications, and is expected to be a natural human-machine interface. Prosody generation is one of the most important techniques for the text-to-speech synthesis, in order to produce high quality synthetic speech. Due to recent researches on prosody, synthetic speech is gradually being improved in naturalness. Several statistical methods for prosody generation, phoneme duration generation and pitch contour generation, have been proposed [1]–[4]. The prosodies generated by the conventional methods, however, while not unnatural, are rather monotonous and boring so far. This is because most previous studies used the speech samples uttered in a reading style, at a normal tempo, and without any context. The prosodies, even in the reading style, may vary due to reflecting the speech tempo, the context and so forth. The prosodies in spontaneous speech can be considered to vary still more. Various speaking styles and the synthetic prosodies that are suited to the speaking styles will be needed in order to make the synthetic speech closer to human speech and to widen the application field for synthetic speech. Therefore, it is necessary to generate the prosodies while considering not only the meaning of a given sentence, the context, and the intention of a speaker, but also the purpose and situation where the synthetic speech will used.

For the first step in such studies, this paper describes the speech tempo influence on the prosody. It is important to discriminate between prosodic features that are influenced by the speech tempo and those which are not, because the synthetic speech tempo should be positively controlled to realize various speech styles. Moreover, the comparison of the prosody among different tempos will certainly reveal an underlying prosodic structure.

Section 2 introduces the speech data and analysis method. Section 3 introduces the basic idea of the proposed duration model, and describes the analysis results for the speech tempo influence on the duration. Previous researches disclosed the relationship between the speech tempo and the durations for various speech segments, such as syllables and phonemes [5],[6]. This paper deals with the durations for speech segments longer than phonemes, namely phrases. Section 4 discusses the proposed pitch pattern generation model and the relationship between the speech tempo and the pitch contour.

2. SPEECH DATA AND ANALYSIS METHOD

To analyze the speech tempo influence on the prosody, speech samples composed of 100 sentences have been collected, which were read by a professional male announcer at three tempos (*fast*, *normal* and *slow*). The phoneme boundaries for the speech samples have been determined manually, and the pitch contours have been extracted. For the linguistic information, individual sentences have been segregated into morphemes, and also into accentual phrases, which resemble phrases. The average speech rates ([mora/sec]) for the 100 sentences at fast, normal and slow tempos are 11.13, 6.80 and 5.62, respectively.

For the prosody generation model, quantification theory (type one) [7], which is based on a linear regressive model for categorical values, has been applied. It can formulate the relationship between categorical and numerical values as:

$$\hat{y}_i = \bar{y} + \sum_f \sum_c x_{fc}\delta_{fc}(i), \qquad i = 1, 2, \ldots, N,$$

where $\hat{y}_i$ is the estimated value for the ith sample, $\bar{y}$ is the average for all samples, N is the total sample number, and δ_{fc} is the characteristic function, defined as:

$$\delta_{fc}(i) = \begin{cases} 1, & i\text{th sample falls into category c of factor f} \\ 0, & \text{otherwise.} \end{cases}$$

x_{fc} is obtained by minimizing

$$\sum_{i=1}^{N}(\hat{y}_i - y_i)^2.$$

The model is considered precise, as the multiple correlation coefficient is closer to 1.0. The partial correlation coefficients for the factors represent the importances of the factors in estimating y_i. The speech tempo influences on the prosody were analyzed using this mathematical model.

3. SPEECH TEMPO AND DURATION

3.1 Duration Model

Figure 1 shows the flow for the proposed duration model. The model explicitly controls the rhythm in speech segments longer than the phoneme.

Conventional models, on the other hand, determine phoneme durations according to the phoneme environment, that is, the preceding and following phonemes, the position within the phrase, and so forth. The durations for speech segments longer than phonemes, such as phrases, are automatically determined as the sums of the phoneme durations. Consequently, the prediction errors for the phoneme durations accumulate, and often cause unnatural rhythms in the phrases and sentences. Moreover, it is difficult to control the rhythm appropriately for the intended tempo.

To solve these problems, first of all, the proposed model determines a suitable tempo for a given sentence or paragraph, by taking into account the meaning and the context. According to the tempo, the model estimates the durations for the individual phrases that constitute the sentence or paragraph. The durations for the segments that correspond to speech phenomena, such as the individual phonemes, are determined by dividing the phrase durations in the proportion estimated by taking into account the phoneme environment.

Thus, the rhythm in the longer speech segment can be explicitly controlled.

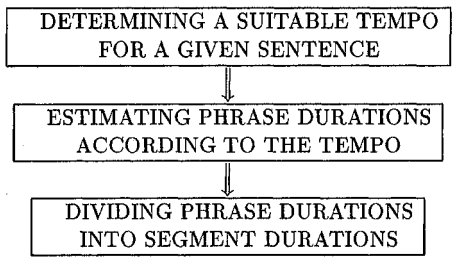

Figure 1. Flow for duration control, in which the speech tempo is taken into account.

3.2 Analysis for Speech Tempo Influence on Phrase Duration

For the first step to realize the proposed model, the relationship between the speech tempo and the rhythm in a speech segment longer than the phoneme was analyzed. An accentual phrase, which is one of the linguistic units, was employed as the longer-term speech segment. The rhythm in the accentual phrase was measured by the ratio of the speech rate for the sentence to that for each accentual phrase within the sentence.

$$\text{accentual phrase tempo} = \frac{\text{speech rate for the accentual phrase [mora/sec]}}{\text{speech rate for the sentence [mora/sec]}}.$$

The accentual phrase rhythms were analyzed by the quantification theory, using the factors shown in Table 1. The combinations for the parts of speech at the accentual phrase boundaries are utilized instead of the syntactic structure, because it is difficult to correctly obtain the syntactic structure.

Table 1. Factors for quantification analysis.

Factors	Number of categories
Position within a sentence	7
Preceding pause	2
Following pause	2
Number of morae	
For preceding accentual phrase	6
For current accentual phrase	6
For following accentual phrase	6
Accent type	
For preceding accentual phrase	3
For current accentual phrase	3
For following accentual phrase	3
Parts of speech combination	
At preceding accentual phrase boundary	40
At following accentual phrase boundary	40

Table 2 shows the multiple and partial correlation coefficients for predicting accentual phrase rhythms. The factors that strongly affect the accentual phrase rhythms differ from each other among the three speech tempos. At the fast tempo, the phrase position within the sentence has the strongest correlation, while, at normal and slow tempos, the most important factor is whether or not a pause exists after the phrase. The result indicates that it is necessary to control the phrase rhythm by taking into account the speech tempo.

4. SPEECH TEMPO AND PITCH CONTOUR

4.1 Pitch Patterns at Different Speech Tempos

Figure 2 shows the observed pitch contours for a sentence that is composed of 6 accentual phrases, uttered at three tempos. The time-axis is normalized, for convenience of comparison. Table 3 shows several average pitch frequency values for the 100 sentences. The average pitch frequency is highest and the pitch frequency range is widest at the fast tempo, among the three tempos. Though it was found that the pitch frequency values at different speech tempos differ from each other, the intonations are quite natural at any tempo.

Then, a normalized pitch pattern is proposed, in which the differences in the actual pitch frequencies can be ignored. Three pitch frequencies for the individual accentual phrases are extracted, which correspond to the center of the first vowel, the peak position and the center of the last vowel for the accentual phrases. They are normalized by the peak pitch frequency for

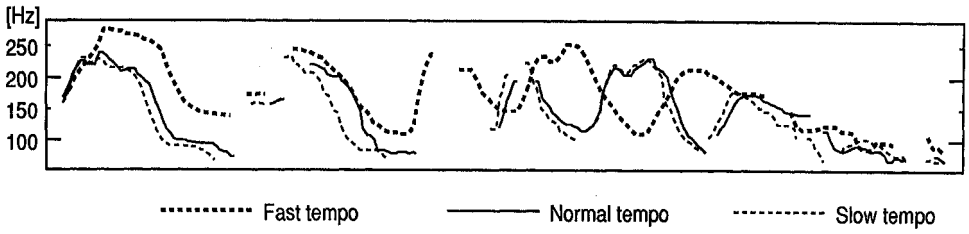

Figure 2. Observed pitch contours at three tempos.

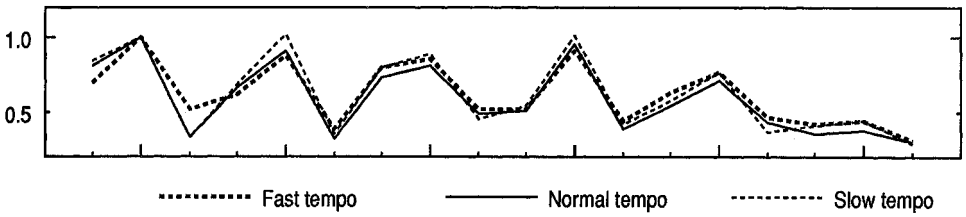

Figure 3. Normalized pitch patterns at three tempos.

Table 2. Multiple and partial correlation coefficients for predicting accentual phrase rhythms.

Factors			Speech tempos		
			Fast	Normal	Slow
Partial correlation coefficients	Position		0.562	0.364	0.306
	Pause	pre.	0.083	0.019	0.109
		fol.	0.061	0.565	0.403
	Number of morae	pre.	0.120	0.159	0.165
		cur.	0.246	0.298	0.157
		fol.	0.102	0.075	0.128
	Accent type	pre.	0.033	0.063	0.084
		cur.	0.043	0.141	0.207
		fol.	0.074	0.088	0.061
	Parts of speech combination	pre.	0.419	0.359	0.299
		fol.	0.366	0.335	0.421
Multiple correlation coefficient			0.629	0.675	0.621

the first accentual phrase within the sentence, in order to represent the relative pitch frequency levels for the accentual phrases. These values are referred to as P_s, P_p and P_e, respectively.

Figure 3, on the other hand, shows the normalized pitch patterns for the same sentence. The time-axis represents the position where the pitch frequencies are extracted. The normalized pitch patterns are quite similar regardless the speech tempo, while the observed pitch contours at the different speech tempos are different. Consequently, the normalized pitch pattern is significant and useful to construct a pitch pattern generation model.

4.2 Pitch Pattern Generation Model

Figure 4 shows the flow for the pitch pattern generation, according to the proposed model.

First, the model predicts the normalized pitch pattern for a given sentence, which is little affected by the speech tempo. The pitch patterns for the individual accentual phrase are modeled by considering the phoneme influence on the pitch contour [8], and are inserted into the normalized pitch pattern. Actual pitch frequencies are estimated by considering the speech tempo, according to the result shown in Table 3.

Table 3. Comparison between average pitch frequency values among three tempos.

Values	Fast		Normal		Slow	
	Average	S. D.	Average	S. D.	Average	S. D.
Average	166.5	12.16	136.9	7.28	132.4	5.97
Range	173.8	12.92	142.1	13.94	137.3	13.28
Beginning	183.0	36.42	158.0	28.09	150.4	26.10
Maximum	249.4	11.72	211.3	13.82	206.2	13.02
Ending	80.4	8.50	72.8	4.06	73.1	4.36

4.3 Analysis of Speech Tempo Influence on Pitch Pattern

The speech tempo influence on the pitch contour was analyzed by the quantification theory, using categorized P_p and P_e, in addition to the factors shown in Table 1.

Table 4 shows the multiple and partial correlation coefficients for predicting P_s, P_p and P_e. The P_s, P_p and P_e values at each tempo are closely predicted, because the multiple correlation coefficients indicate more than 0.86. For the P_s prediction, the accent type and P_p for the current accentual phrase, and the combination for the parts of speech are important. For the P_e, in addition to these factors, the number of morae for the current accentual phrase is also important. The position within the sentence, the number of morae for the current accentual phrase are most significant in predicting the P_p value.

There is a slight difference in the importance for the factors among the three tempos, except whether or not a pause precedes or follows the current accentual phrase. The result shows that the normalized pitch pattern is less affected by the speech tempo, and is effective for modeling the pitch contour.

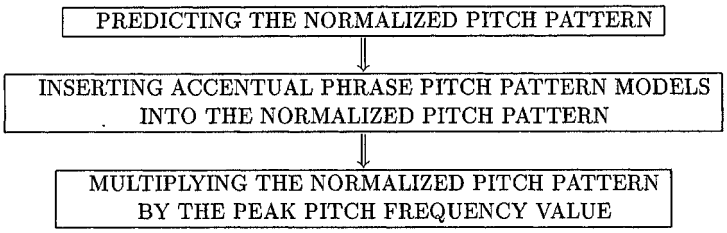

Figure 4. Flow for pitch pattern generation.

Table 4. Multiple and partial correlation coefficients for predicting normalized pitch patterns.

Factors			P_s			P_p			P_e		
			Fast	Normal	Slow	Fast	Normal	Slow	Fast	Normal	Slow
Partial correlation coefficients	Position		0.437	0.247	0.127	0.633	0.692	0.621	0.357	0.159	0.210
	Pause	pre.	0.076	0.058	0.042	0.039	0.506	0.354	0.012	0.038	0.039
		fol.	0.041	0.088	0.086	0.038	0.275	0.334	0.086	0.506	0.316
	Number of morae	pre.	0.146	0.154	0.156	0.227	0.246	0.219	0.161	0.147	0.142
		cur.	0.160	0.096	0.105	0.492	0.473	0.501	0.628	0.571	0.571
		fol.	0.101	0.101	0.182	0.136	0.104	0.088	0.139	0.100	0.140
	Accent type	pre.	0.244	0.033	0.077	0.309	0.260	0.192	0.162	0.089	0.109
		cur.	0.573	0.609	0.682	0.148	0.180	0.160	0.763	0.770	0.782
		fol.	0.018	0.125	0.103	0.022	0.030	0.048	0.089	0.094	0.149
	Parts of speech combination	pre.	0.411	0.357	0.398	0.547	0.500	0.429	0.369	0.337	0.319
		fol.	0.470	0.338	0.396	0.458	0.323	0.306	0.416	0.277	0.294
	P_p	pre.	0.311	0.190	0.262	—	—	—	0.383	0.158	0.198
		cur.	0.569	0.629	0.611	—	—	—	0.721	0.608	0.586
		fol.	0.235	0.309	0.268	—	—	—	0.445	0.457	0.374
	P_e	pre.	0.440	0.360	0.348	—	—	—	—	—	—
		cur.	0.168	0.145	0.149	—	—	—	—	—	—
		fol.	0.215	0.203	0.198	—	—	—	—	—	—
Multiple correlation coefficient			0.874	0.895	0.887	0.865	0.895	0.878	0.936	0.925	0.910

5. CONCLUSION

The speech tempo influences on the prosody were analyzed using the quantification theory, and the results showed several significant prosodic features. On the basis of the results, the duration model and the pitch pattern generation model were proposed, in which the speech tempo influences on duration and pitch were taken into account. These models explicitly have speech tempo parameters, and generate adequate durations and pitch contours, according to the tempo. For further research, the speech tempo influences on the durations for shorter-term speech segments than accentual phrases, such as phonemes, will be investigated.

REFERENCES

[1] N. Kaiki, K. Takeda and Y. Sagisaka, "Statistical Analysis for Segmental Duration Rules in Japanese Speech Synthesis," *Proc. ICSLP*, pp.17–20, 1990.

[2] K. Iwata, Y. Mitome, J. Kametani, M. Akamatsu, S. Tomotake, K. Ozawa and T. Watanabe, "A Rule-Based Speech Synthesizer Using a Pitch Controlled Residual Wave Excitation Method," *Proc. ICSLP*, pp.185–188, 1990.

[3] Y. Sagisaka and N. Kaiki, "Optimization of Intonation Control Using Statistical F0 Resetting Characteristics," *Proc. ICASSP*, Vol. 2, pp.49–52, 1992.

[4] M. Abe and H. Sato, "Two-stage F0 Control Model Using Syllable Based F0 Units," *Proc. ICASSP*, Vol. 2, pp.53–56, 1992.

[5] S. Hiki, Y. Kanamori and J. Oizumi, "On the Duration of Phoneme in Running Speech," (in Japanese) *J. IECEJ*, Vol. 50, No. 5, pp.849–850, 1967.

[6] S. Hiki, "On the Duration of Various Segments in Sentence Speech," (in Japanese) *J. IECEJ*, Vol. 50, No. 8, pp.1465–1470, 1967.

[7] C. Hayashi, "On the Quantification of qualitative data from the mathematico-statistical point of view ," *Ann. Inst. Statist., Math 2*, 1950.

[8] K. Iwata, Y. Mitome, S. Tomotake, M. Akamatsu, J. Kametani, K. Ozawa and T. Watanabe, "Japanese Text-to-Speech Conversion System Using a Pitch Controlled Residual Wave Excitation Method," (in Japanese) *IEICE Technical Report*, Vol. 90, No. 335, pp.15–22, 1990.