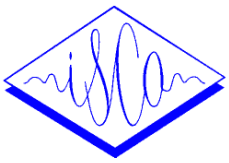

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Segmental Power Control for Japanese Speech Synthesis

Kenzo ITOH, Tomohisa HIROKAWA and Hirokazu SATO

Speech and Acoustics Laboratory
NTT Human Interface Laboratories
Take 1-2356, Yokosuka-shi, Kanagawa 238-03, Japan

ABSTRACT

This paper proposes a segmental power control method for speech synthesis by rule. The innovation of this method lies in its use of the phoneme environment characteristics and the relationship between speech power and pitch frequency. First, the Permissible Threshold (PT) for power modification is measured by subjective experiments using phoneme power manipulated speech material. As a result, it is concluded that the PT of phoneme power modification is 4.1 dB. This experimental result is significant when discussing power control and gives a criterion for power control accuracy. Next, the relationship between speech power and pitch frequency is analyzed using a very large speech data base. The results show that the relationship between phoneme segmental power and pitch frequency is affected by the kind of phoneme, the adjoining phonemes, rising or falling pitch conditions, and initial or final position of sentence. Finally, we propose that the segmental speech power should be controlled by the pitch and phoneme environment. This new method yields an averaged root mean square error between real and estimated speech power of 2.17 dB. This value indicates that 94% of the estimated power values are within the Permissible Threshold of human perception.

10.21437/ICSLP.1992-136

1. INTRODUCTION

Various new technologies have recently been proposed and studied for text-to-speech synthesis systems. Examples include the automatic synthesis unit generation method using phoneme environmental clustering[1], high quality synthesis systems using waveform-compilation synthesis techniques[2]-[4] and a new synthesis technique with non-uniform length synthesis units[5]. As many synthesis techniques are continually being improved, the demand for more accurate rules is more acute. In particular, prosody information control requires more advanced rules such as those generated from the two-stage pitch frequency control model[6], or speech power control rules which have rarely been studied up to now[7].

We have been researching high quality synthesis systems based on waveform type synthesis techniques using phoneme segments as synthesis units. Very natural and smooth synthesis speech can be generated with the following techniques, (a)synthesis unit: the center phoneme segment in a triphone configuration, (b)pitch and duration control: pitch synchronous over-lap add method[8], (c)phoneme duration: averaged center phoneme value of selected triphone. However, synthesized speech quality is fails to realize natural speech quality. To generate higher quality synthesis speech, the power control rule is important, as well as mention character to sound symbol conversion and pitch pattern generation.

This paper describes first, to discuss power control and to give a criterion for power control accuracy, the Permissible Threshold (PT) for power modification is measured by subjective experiments using phoneme power manipulated speech material. Next, the relationship between power and pitch frequency is analyzed using a very large speech data-base. Finally, we propose a phoneme power control method that considers the pitch frequency and the phoneme environment information.

2. INFLUENCE OF POWER MODIFICATION

A criterion for speech power control accuracy can be determined by confirming the influence of power fluctuations on human speech perception. In this section, the subjective experimental procedure and results are described.

2-1. Segmental power modification process

The segmental speech power of hand labeled speech material was modified. This modification process weighted every phoneme by a modification factor Pm (dB). In other words, the values of +Pm or -Pm were added to the average value of the target phoneme. Two following power modification process were used in the experiment. (1)Alternative Modification (ALT): +Pm and -Pm were added alternately to input phoneme sequence. (2)Random Modification (RAN): +Pm or -Pm were added randomly to input phoneme sequence. In the ALT condition, modified power difference between phonemes was 2Pm at every phoneme concatenation point. Under RAN, on the other hand, the average difference was Pm. Therefore, the influence of power modification for the human perception seems the ALT larger than the RAN condition.

The Pm value ranged from 0 dB to 9 dB in 1 dB steps, and the 0 dB value generated the reference signal, i.e., original speech signal. To study the influence of just the power modification process and to avoid contamination with click-noise, a smoothing technique with a 5 ms length window was used at each phoneme concatenation point. Figure 1 shows an example of a modified speech waveform [ALT, Pm=3dB] with the original speech signal. Two kinds of speech materials were used in the subjective experiment. (I)AD/DA: analog to digital conversion using 16 bit quantization and a 12 kHz sampling frequency. (II)LSP: speech synthesized by the Line Spectrum Pair (LSP) analysis technique[9]. Two sentences with different contents provided the speech material, and one male and one female speaker were used.

2-2. Experimental procedure

A subjective test with 5 categories {Excellent(E), Good(G), Fair(F), Poor(P), Bad(B)} was employed. The corresponding mean opinion scores(MOS) were set at 4, 3, 2, 1, 0 for each stimulus. An ARAEN receiving system with a DR-305 as the reference receiver was used. The system gain was set at + 5 dB(OTR) to ensure good listening conditions. Four female subjects were used and they listened to each source five times in random order.

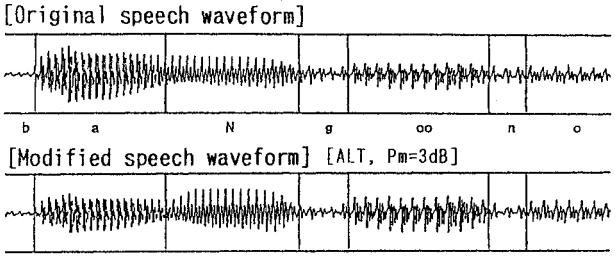

Fig.1 An example of modified speech waveform used in subjective experiment

2-3. Experimental results

Figure 2 shows the results of subjective experiment as expressed by the Detectability Threshold (DT) and the Permissible Threshold (PT) for phoneme power modification. DT and PT were derived using the following procedure from MOSs. First, DT was derived from the relationship between the MOSs and the power modification factor Pm as a deterioration function using the standard deviation of reference signal MOS. Separate tests based on the pair comparison method confirmed that the DT values agree well with the detectability threshold[10]. Next, PT was derived from the relationship between MOS values and the cumulative occurrence of (E+G+F), i.e., the quality is more than "Fair". When 90% of the subjects judge the speech quality to be better than "Fair", the MOS value is 2.5. This MOS value equivalents 1.1 to deterioration from reference signal MOS value. The PT of the each conditions were determined by the Pm which is same as this deterioration values.

One observation can be made from figure 2 and the results of analyzing the variance. The influence of phoneme power modification is larger for the male voice than the female voice, and the ALT than the RAN influence is larger. The difference between speech materials is small. These results indicate the following important features.

(1)The average DT value of LSP speech analysis/synthesis speech is 3.7dB, but this values decreases to 2.6dB with waveform processing. It is seen from this result that higher power control accuracy is needed for high quality synthesized speech.

(2)The PT value of LSP synthesized speech is 4.5dB, while it is 4.1dB for waveform processing speech. These values contain very important information about the design of the phoneme power control rules. That is, 4.1dB is the target when designing high quality waveform type speech synthesis systems.

3. ANALYSIS OF PHONEME POWER

The development of phoneme power control rules should start with an analysis of speech power characteristics. In-depth studies have not been made on phoneme power characteristics except for Sacia's study on English speech[11]. This section introduces a few examples that show the results of the phoneme power characteristics.

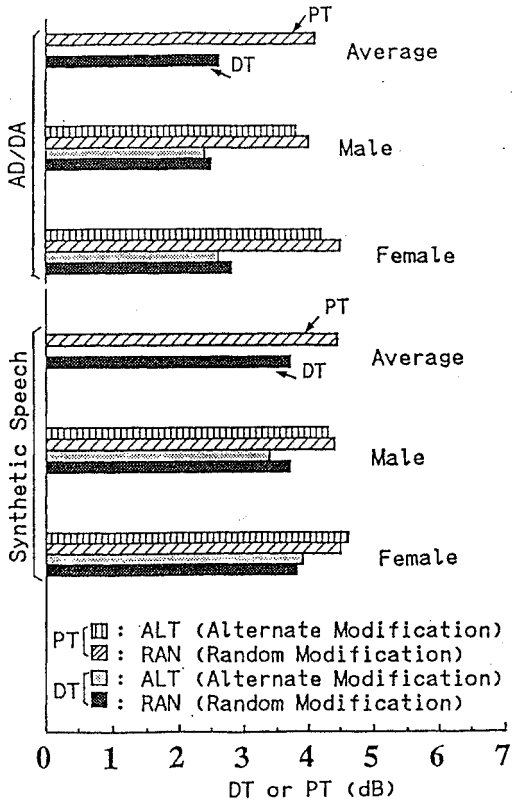

Fig.2 Detectability Threshold (DT) and Permissible Threshold (PT) for phoneme power modified speech

3-1. Speech material

The analysis conditions were; a sampling frequency of 12 kHz (low-pass filter set to 5.1 kHz), 16bit quantization including sign bit, and the total length of speech samples is about 30 minutes. A male professional speaker was used, and a normal speaking speed was used. The speech power was calculated as the moving average of 32 ms window for all hand labeled phonemes.

3-2. The power of part of sentence

Figure 3 shows phoneme power distribution of "Initial", "Final" and "medial" in sentence. The vertical axis shows relative frequency (%) and sigma is the standard deviation. From this figure, it is seen that the scatter of phoneme power (initial) is small at 3.0 dB, while the scatter (final) is very large at 6.1 dB. Thus it is reasonable to employ different power control methods for different parts of the sentence.

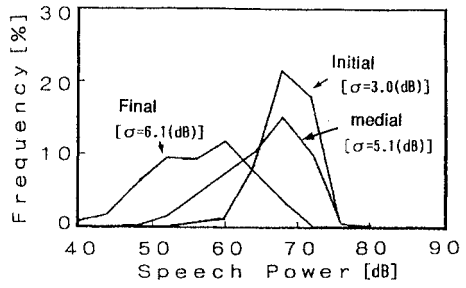

Fig.3 Speech power distribution of phoneme

3-3. Influence of phoneme environment

The phoneme power is influenced with not only intra-sentence as described before but also phoneme environment[7]. Therefore, the scatter of phoneme power can be decreased by restricting the phoneme environment. Figure 4 shows an example of phoneme power /a/ distribution in the phoneme environments /k/-/a/-/N/ and /s/-/a/-/N/ and for the all phonemes /a/. This figure shows that the power scatter of /a/ in a restricted phoneme environment is smaller seen with the all phonemes /a/. The same results were showed for other phonemes or phoneme environments.

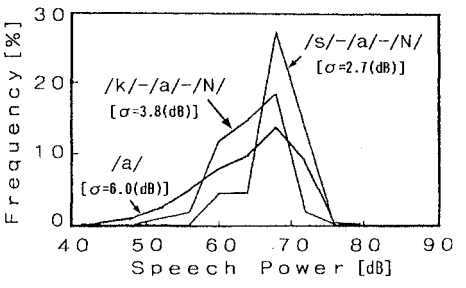

Fig.4 Power distribution influenced by phoneme environment

4. CORRELATION ANALYSIS BETWEEN PITCH AND POWER

Using a new approach to power control, this paper uses the relationship between pitch frequency and power. In other words, in general, the higher the pitch frequency, the greater the measured speech power. The correlation between power and pitch frequency have been studied by Hiki[12] and Suzuki[13].

The speech material described in section 3-1 was used in the correlation analysis. The pitch frequency was manually corrected after using the automatic extraction method. The pitch frequency of each phoneme was calculated as the moving average of the short term window (32ms length).

4-1. Overall characteristics

Figure 5 shows the relationship between pitch frequency and power for all voiced phonemes. Each point in the figure corresponds to the averaged power and pitch frequency of one phoneme. This figure also plots the correlation coefficients (r) and root mean square error (err) as a first-order regression line. The two lines paralleling the regression line indicate the permissible threshold (PT=4.1dB). It should be clear from this figure that strong correlation is clearly noticeable. However, this relationship can not be used in the phoneme power estimation rules because, only 65% of all points lie within permissible threshold conditions.

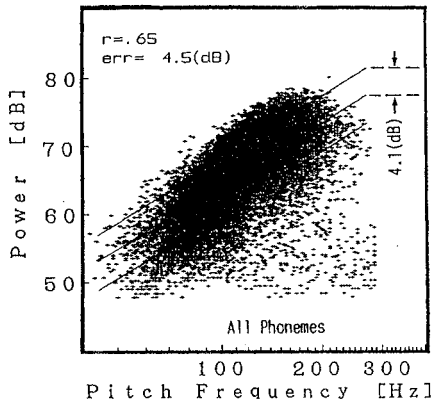

Fig.5 Relationship between phoneme segment power and pitch frequency (All phoneme conditions except voiceless sounds)

4-2. Characteristics with different phoneme conditions

Section 3-3 proved that phoneme power scattering decreased if the phoneme environment was restricted. Therefore, a higher correlation between pitch and power was extracted in the restrict phoneme environment conditions.

Figure 6 shows relationship between pitch frequency and power for each vowel. The high correlation coefficients 0.77(/a/), 0.80(/e/) and 0.76(/o/) were obtained except /i/ and /u/ conditions. Figure 7 shows the same relationship for vowel /a/ when it is bracketed by voiced phonemes. It is clear from these figures that a very high correlation coefficient can be achieved by phoneme or the phoneme environment. Therefore, when the pitch information is used for power estimation, the phoneme environment must be specified.

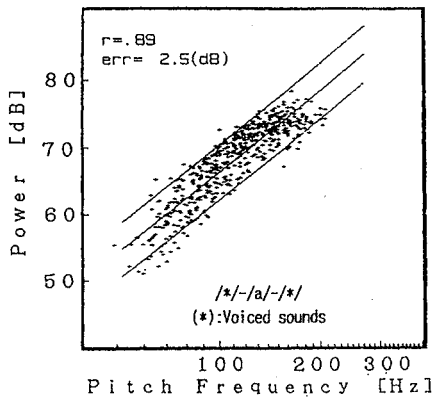

Fig.7 Relationship between power and pitch frequency (/a/ within voiced phoneme environment)

4-3. Initial/Final syllable and Rising/Falling pitch

Figure 8 (a) and (b) show the relationship between pitch and power of the phoneme in the initial and final syllables. The respective correlation coefficients are 0.56 and 0.06. Therefore, the relationship between pitch frequency and power can not be used for power control in the initial and final syllables. Figure 8 (c) and (d) show the relationship between pitch frequency and power of the phonemes with rising and falling pitch, respectively. The dynamics of pitch frequency was determined at the center of phoneme. The flat pitch condition was contended to falling pitch condition. It is seen from these figures that the correlation coefficient is higher with falling pitch than with rising pitch. This situation may be explained by the following reason. Rising pitch demands greater effort to vocalize such as greater tension of the vocal folds and higher subglottal air pressure. Therefore, pitch frequency and speech power values are more unstable with rising pitch; more study is necessary to confirm the reasons.

5. POWER CONTROL RULE

The phoneme power characteristics and the correlation analysis show that the relationship between pitch frequency and power is affected by phoneme, phoneme environment, pitch changes, and syllable position. From those viewpoints, phoneme power is best estimated by equations (1) and (2).

$$PP = Pop*wj + Pot*(1-wj) \quad (1)$$
$$Pop = aj*F0 + bj \quad (2)$$

In these equations, Pop is the power estimated by pitch frequency, Pot is the power estimated by the triphone information table, wj is the weighting factor for pitch frequency, F0 is the synthesized pitch frequency, and aj and bj are regression coefficients expressing the relationship between pitch and power. The j indicates phoneme or phoneme group. Grouping is determined by the kind of phoneme such as /a/, /i/, /u/,.., kind of proceeding

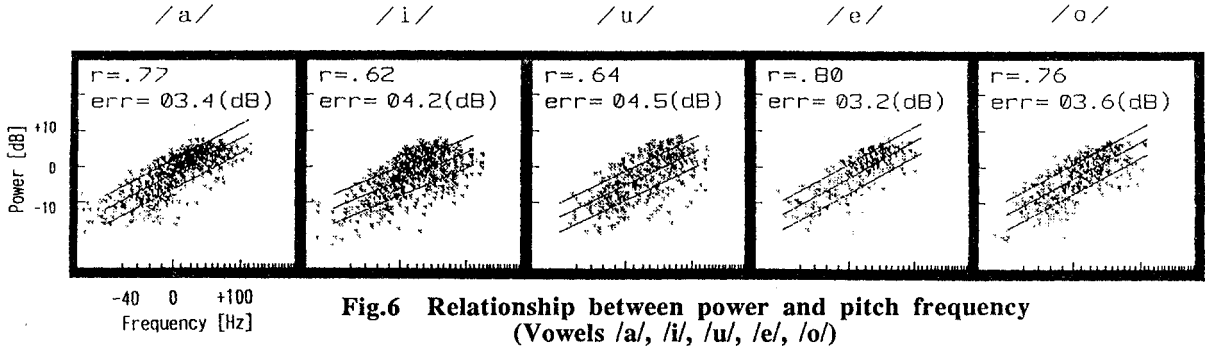

Fig.6 Relationship between power and pitch frequency (Vowels /a/, /i/, /u/, /e/, /o/)

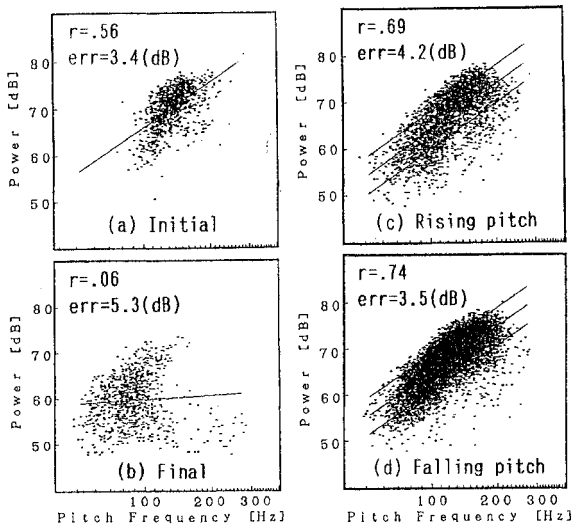

Fig.8 Relationship between power and pitch frequency
(a): Initial syllable, (b): Final syllable,
(c): Rising pitch, (d): Falling pitch

and succeeding phonemes, rising or falling pitch, and syllable position. For example, j=1 group includes vowel /a/ (rising pitch) that is bracketed by voiced phonemes such as /m/-/a/-/g/ or /z/-/a/-/n/.

The weighting factor wj is determined by the following process. First, regression coefficients aj and bj are determined using wj fixed condition wj=1.0 for the phoneme group j. Next, minimum error condition between real and estimated power is selected for wj values from 0.0 to 1.0. Figure 9 shows an example of the weighting factor and root mean square error. It should be clear from this figure that the estimation error decreases approximately 1.0 dB when using optimum pitch and phoneme environment condition.

Figure 10 shows phoneme estimation results using all Wj factors for vowels (a) and voiced consonants (b). The proposed method gives 2.17 dB and 2.48 dB as the averaged root mean square error between real and estimated speech power for vowels and voiced consonants, respectively. This value shows the good estimation because about 94% and 90% of the estimated powers are within the Permissible Threshold (PT) value for vowels and voiced consonants.

6. CONCLUSION

A new segmental phoneme power control technique was proposed to generate high quality synthesized speech. First, the Permissible Threshold (PT) for power modification was measured by a subjective experiment. It was concluded that the PT of phoneme power modification is 4.1 dB. This PT value is significant when discussing the power control and provides a criterion for power control accuracy. Next, the relationship between speech power and pitch frequency was analyzed using a very large speech data base. The analysis showed that the relationship between pitch and power is influenced by phoneme, phoneme environment, pitch changes, and syllable position. Finally, considering the above viewpoints, we determined that phoneme power should be controlled by information about pitch and phoneme environment. The proposed method yielded an average root mean square error between real and estimated speech power of 2.17 dB. This value indicates that for vowels, 94% of the estimated powers do not exceed the Permissible Threshold value. In the near future, this new proposed power control method will be implemented in a speech synthesis system based on concatenation of phoneme segment wavelets[14].

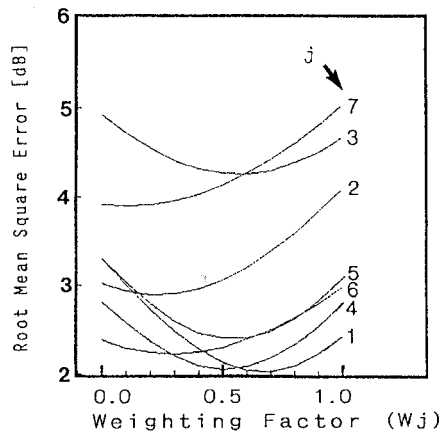

Fig.9 Effects of weighting factor Wj on estimated phoneme power accuracy
(j=1:/a/, 2:/i/, 3:/u/, 4:/e/, 5:/o/, 6:Initial, 7:Final)

[Acknowledgments]
We are grateful to the members of the Speech and Acoustics Department for their helpful discussions. We also thank Dr. Yukio Kobayashi, Dr. Sadaoki Furui and Dr. Noboru Sugamura for their continuous support of this work.

[References]

[1] S. Nakajima and H. Hamada, "Automatic generation of synthesis units based on context oriented clustering," ICASSP'88, pp.659-662, New York, (Apr. 1988)
[2] T. Hirokawa and K. Hakoda, "Segment selection and pitch modification for high quality speech synthesis using waveform segments," ICSLP'90, pp.337-340, Kobe, (Nov. 1990)
[3] K. Itoh and H. Sato, "Excitation waveform extraction for pitch control in residual-excited LPC speech synthesis," 118th Fall Meeting of ASA, FF22, S79, St. Louis, (Dec. 1989)
[4] K. Itoh, H. Mizuno, T. Nomura and H. Sato, "Phoneme segment concatenation and excitation control based on spectral distortion criterion for speech synthesis," ICSLP'90, pp.189-192, Kobe, (Nov. 1990)
[5] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," ICASSP'88, S14.8, New York, (Apr. 1988)
[6] M. Abe and H. Sato, "Two-stage F0 control model using by syllable based F0 units," ICASSP'92, II-53, (1992)
[7] K. Mimura, N. Kaiki and Y. Sagisaka, "Analysis and control of temporal patterns of speech power using statistical methods," Trans. Committee on Auditory Research, The Acoust. Soc. of Japan, SP91-4, (1991) (in Japanese)
[8] F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using Diphones," Proc. of Eurospeech'89, (1989)
[9] N. Sugamura and N. Farvardin, "Quantizer design in LSP speech analysis-synthesis," IEEE J. Select. Areas Commun., Vol.6, No.2, pp432-440, (1988)
[10]K. Itoh and N. Kitawaki and K. Kakehi, "Objective quality measures for speech waveform coding systems," Review of the ECL, Vol.32, No.2, pp.220-228, (1984)
[11]C. F. Sacia and C. J. Beck, "The power of fundamental speech sounds," Bell Syst. Tech. J., Vol.5, p.393 (July, 1926)
[12]S. Hiki, "Correlation between increments of voice pitch and glottal sound intensity," J. Acous. Soc. of Japan, Vol.23, No.1, pp.20-23, (1967)
[13]J. Suzuki and R. Tanaka, "An LPC vocoder excited by synthesized pitch," Fall Meeting of Acous. Soc. of Japan, p.511, (Oct. 1979) (in Japanese)
[14]T. Hirokawa, K. Itoh and H. Sato, "High quality speech synthesis based on wavelet compilation of phoneme segments," ICSLP'92, Synthesis-II, Alberta, (Oct. 1992)

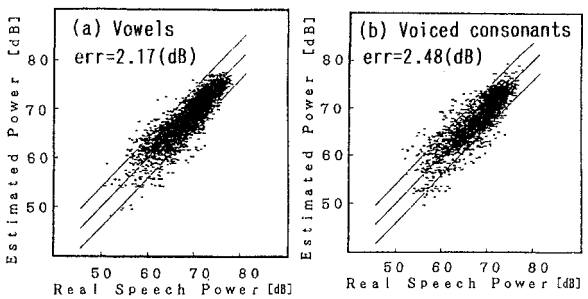

Fig.10 Relationship between real and estimated phoneme power (a):Vowels, (b):Voiced consonants