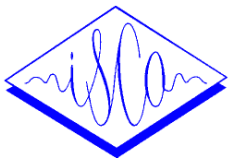

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Correlation Analysis Between Speech Power and Pitch Frequency for Twenty Spoken Languages

Kenzo Itoh

NTT Human Interface Laboratories
1-2356 Take, Yokosuka-shi, Kanagawa, 238-03 Japan
E-mail; itoh@nttspch.ntt.jp, Tel;+81-468-59-3289, Fax;+81-468-55-1054

ABSTRACT

This paper describes the relationship between speech signal power and pitch frequency for twenty main languages. The goal is to confirm of the applicability of our earlier proposed power control rule using the relationship. First, an overall analysis is conducted for each language. Second, the short term correlation is analyzed to study the relationship between high correlation values and other characteristics of the speech signal. Last, in order to get information for developing an English Text-To-Speech (TTS) system, averaged phoneme power and pitch frequency is analyzed using American English speech signals with phoneme labeled data. Main results are shown below. (1)The average correlation coefficient ranged from +0.72 to +0.44 for the twenty languages. (2)The short term analysis found that high level speech was accompanied by high correlation values. The reason for this assumed to be the relationship between accentuation or stress. (3)The prospect for phoneme power control in an American English TTS system is excellent given the strong relationship between power and pitch, Those results strongly suggest that the relationship between speech power and pitch frequency can be used in speech processing systems for all spoken languages.

1. INTRODUCTION

It is generally accepted that speech signal power and fundamental frequency (pitch frequency) are related in all spoken languages[1][2]. For example, Kobayashi[3] analyzed the relationship between pitch contour and signal intensity pattern using isolated words. Hiki[4] also studied the correlation between increments of pitch and glottal sound intensity. Suzuki[5] conducted an application study for transmission data compression for a speech coder(vocoder). In general, vocoders use both pitch frequency and signal power information as excitation source signals. In Suzuki's system, only one side information (pitch or power) is used because the other side information is estimated by the relationship between power and pitch.

Our earlier paper examined the correlation between speech signal power and pitch frequency for Japanese, and introduced a phoneme power control rule for a very high quality TTS system[6]. In this system, the rule considered not only the relationship but also the phoneme environment. The proposed power control rule yielded an average root mean square error between real and estimated power of 2.17 dB. This value indicates that for vowels, 94% of the estimated power did not exceed the permissible threshold value[7].

The above mentioned relationship between speech signal power and pitch frequency can be used in other speech processing systems. However, the speech materials were Japanese only. As described before, it is generally accepted that the relationship applies to all spoken languages. Therefore, our power control rule might be applicable to other languages. This paper analyzed twenty spoken languages to confirm the general applicability of the power control rule. First, an overall analysis is conducted for each language. Second, short term correlation was analyzed to study the relationship between high correlation values and other of speech signal information. Finally, in order to get the information for developing an English TTS system, averaged phoneme power and pitch frequency were analyzed for American English.

Table 1 Information Concerning the Speech Samples

Languages		Laboratory/Country	Number of Speakers	Number of Sentences
1, AE	American English	American Telephone & Telegraph AT&T/U.S.A.	4	4
2, AR	Arabic	CCITT Laboratory	2	4
3, CH	Chinese (Mandarin)	Inst. of Telecommunications Transmission of PTT/China	4	8
4, DA	Danish	Telecomm. Administration Research & Developement /Denmark	4	2
5, DU	Dutch	PTT D.N. Laboratories /Netherland	4	2
6, EN	English	British Telecom Research Laboratories /United Kingdom	4	8
7, FI	Finnish	Posts & Telecommunication /Finland	4	8
8, FR	French	Centre Nationale d'Etude de Telecommunications /France	4	2
9, GE	German	Fernmel de technishes Zentramlant /F.R.G.	4	8
10, GR	Greek	CCITT Laboratory	2	4
11, HI	Hindi	CCITT Laboratory	2	3
12, HU	Hungarian	Central Administration of Hungarian PTT /Hungary	4	8
13, IT	Italian	Centro Studie Laboratori Telecommunicazioni /Italy	4	8
14, JA	Japanese	Nippon Telegraph & Telephone Corp. /Japan	4	2
15, NO	Norwegian	Telecommunication Administration /Norway	4	2
16, PL	Polish	CCITT Laboratory	4	8
17, PR	Portuguese	Telecommunication of Reo de Janeiro /Brazil	4	8
18, RU	Russian	Ministry of Posts & Telecommunications /USSR	4	8
19, SP	Spanish (Castelian)	Telephonica /Spain	4	8
20, SW	Swedish	Telecommunication Administration /Sweden	4	8

2. SPEECH DATA

Table-1 shows the information concerning the analyzed speech samples. This multi-lingual speech data base (MLSDB) was recorded at ITUT (CCITT) laboratories of each country. All speech samples were recorded in a quiet both, and sentences of about 10 seconds in length were chosen. A half-inch condenser microphone or the equivalent having flat sensitivity/frequency response from 80 Hz to 8 kHz was used. The distance between microphone and speaker (lip position) was 15 cm to avoid puffing noise from explosive stops. Most laboratories used conventional analog tape recorders, but a few used PCM recorders. Signal to noise ratios for the languages are from 38 to 42 dB. More details of the MLSDB are described in reference [8]. In the MLSDB, English consisted of American English and British English. Chinese was Mandarin and Spanish was Castelian. All speech . materials were normal sentences.

3. CORRELATION ANALYSIS

Basic sampling frequency of the MLSDB was set to 32 kHz. The analysis conditions of the paper were; a sampling frequency of 12 kHz (low pass filter set to 6.0 kHz and re-sampling by sample-down technique), 16 bits linear quantization including sign bit. The pitch frequency was extracted by the modified autocorrelation method[9]. Pitch frequency and signal power were calculated using 48 ms length windows and the frame interval was set to 12 ms. Pitch and power values were expressed in logarithmic values. For example, pitch frequency PT(t) at frame t was as follows,

$$PT(t) = \log10\{\frac{1}{N}\sum_{i=t-4,t+5} P(i)\} \quad \ldots\ldots \quad (1)$$

where, i is frame number and N is number of moving average, in this paper, N was set to 10. Figure 1 shows an example of analysis results of pitch and power contour for Japanese, French and German. The STC in the figure shows short term correlation as will be describe later. When analyzing the relationship between pitch and power, only voiced speech parts were analyzed. This figure show that the relationship between pitch and power of French or German differs from that of Japanese, especially German has negative correlation speech part compared to Japanese.

4. ANALYSIS RESULTS

4-1. Overall Characteristic

Figure 2 shows averaged correlation coefficient (R) for each language which were rearranged. The R values ranged from +0.72 to +0.44 for the twenty languages. Japanese, Hindi, Dutch and English have high correlation coefficients. Conversely, German, Russian, Norwegian and Portuguese have relatively low coefficients. The confidence interval of Japanese on the one hundred samples condition is from +0.59 to +0.79. Therefore, low correlation languages(German, Russian or Norwegian) were different from Japanese at the relationship between pitch frequency and speech power.

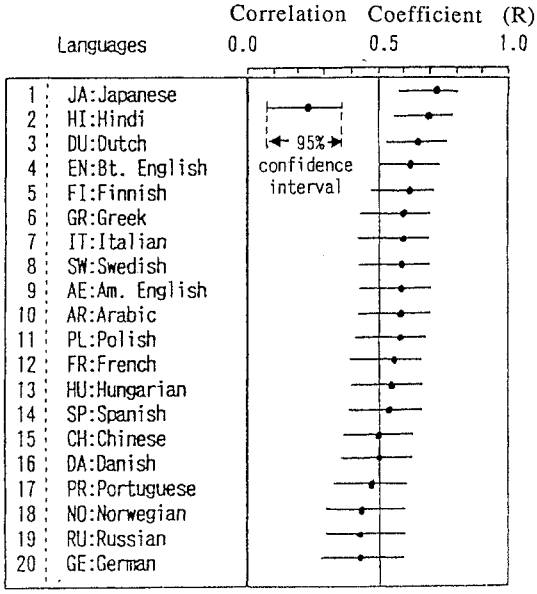

Fig. 2 Result of Correlation Analysis on Twenty Spoken Languages

4-2. Short Term Characteristic (STC)

STC(t) at time t is as following,

$$STC(t)=\sum_{i=t-4,t+5}(xi-xa)(yi-ua)\Big/\mathrm{sqrt}\{\sum_{i=t-4,t+5}(xi-xa)^2(yi-ya)^2\} \quad \ldots(2)$$

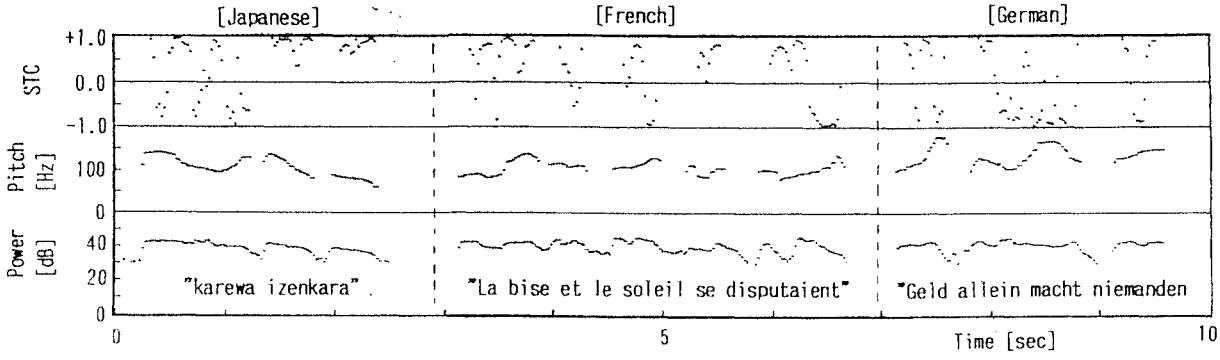

Fig. 1 Examples of Speech Power, Pitch Frequency and STC Time Pattern (STC: Short Time Correlation Coefficient)

where, xa and ya are averaged pitch frequency and power in the short term, xi and yi are pitch and power at frame i-th. Therefore, STC(i) shows the relationship between pitch and power patterns at neighborhood of i-th frame. The STC value fluctuated with time as shown in figure 1. In normal speech signal, there are high or low correlation parts. To discuss further in this section, the ratio of high or low correlation speech parts was analyzed. Figure 3 shows an example of the result. In this figure, four special ranges are plotted; range-I: very high correlation part (STC = +1.0 ~ +0.8), range-II: high correlation part (STC = +0.8 ~ +0.2), range-III: no correlation part (STC = +0.2 ~ -0.2), range-IV:negative correlation part (STC = + -0.2 ~ -1.0). Horizontal axis (%) shows the ratio accounting for special ranges present in the total speech length. This figure shows that the correlation between pitch and power of all languages was strong because, high correlation ranges (range-I and range-II) occupy the majority, over 70%, of the speech length. However, range-I as a very high correlation part of Japanese or English is not the same in German or Norwegian.

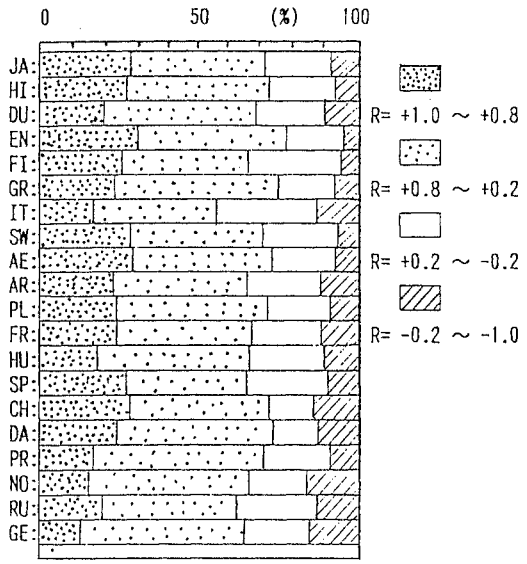

Fig. 3 Ratios for Special Ranges of Correlation Coefficients in Test Speech Materials (R Shows Correlation Coefficients by STC)

Next, both pitch frequency vale and signal power level were analyzed at range-I. to clarify the reason of high correlation vale. Figure 4 shows an example of the analysis result. Vertical axis shows relative value, o dB means averaged value. This result shown that both pitch frequency and signal power strengthen in the very high correlation speech part, the pitch frequency increases by 15 Hz and signal power increases by 6 dB from average values. Against this result, those values decrease at range-IV as negative correlation part. Range-I, exhibited stress or accentuation and this is not surprising.

4-3. Scatter of Pitch Frequency and Correlation Coefficient

As was pointed out, the overall correlation coefficient between pitch and power shows a positive value, however, German, Russian or Norwegian have lower correlation than Japanese or Hindi. In this section, to clarify the reason for

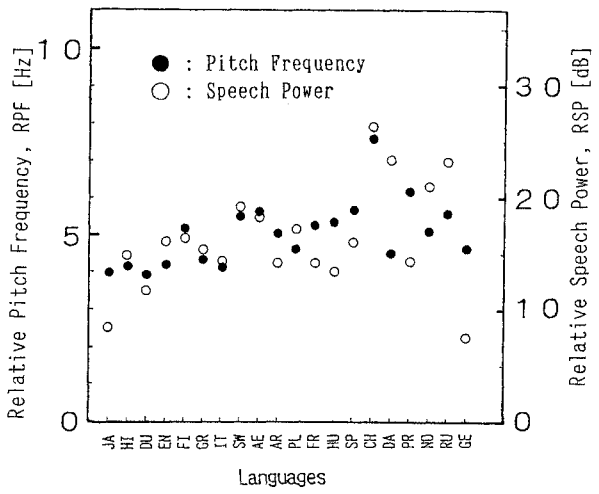

Fig. 4 Speech Power and Pitch Frequency in High Correlation Parts of Speech Signal (RPF & RSP are expressed in Values Relative to Average Values)

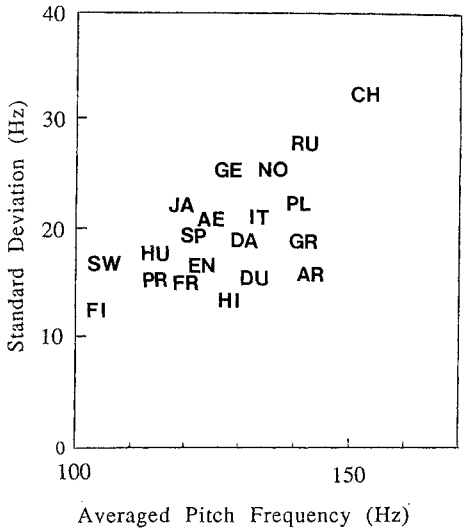

Fig. 5 Relationship Between Averaged Pitch Frequency and Standard Deviation of Pitch Frequency

this low correlation, the scatter of pitch frequency was analyzed. Figure 5 shows relationship between averaged pitch frequency (Pa) and its scatter of frame by frame (standard deviation :Ps) for each language. It should be clear from this figure that a strong correlation is clearly noticeable. This relationship shows a similar tendency for female speech. From this figure, standard deviations of CH, RU, NO or GE are large. I discuss a few points about this result using Fig.2 and Fig.5. Considering the languages whose scatter of pitch frequency is large, correlation coefficient between pitch and power is small. Those languages have the following characteristic, for example, Russian has grave accentuation and Chinese has four tones. In other words, those languages transmit meaningful information through the pitch contour, therefore, the relationship between pitch and power, which showed a normal relationship based on speech mechanism, was distorted However, more understanding is not possible with just the above experimental results.

4-4. Influence of Frequency Band

Figure 6 shows an example of the influence of signal frequency band on the correlation coefficient between pitch and power. Three languages (FR, JA and RU) were used. Horizontal axis shows filter types, for example, "low pass" and "cutoff=500 Hz" conditions, correlation coefficient was calculated using the speech signal passed through the low pass filter with 500 Hz cutoff frequency. Results at "all pass" condition are the as same as given in section 4-1. This figure shows that low frequency signal components give higher correlation than high frequency components. In other words, relationship between excitation source signal power and pitch frequency gave strong correlation. This conforms well to Hiki's experimental results[4].

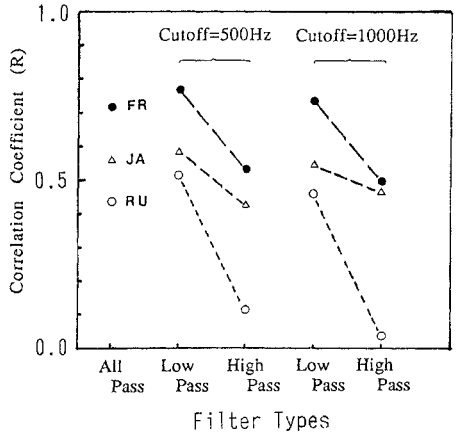

Fig. 6 Influence of Speech Frequency Component for Relationship Between Pitch Frequency and Speech Power

5. RELATIONSHIP BETWEEN PITCH and POWER in ENGLISH PHONEME

In order to get information for developing an English TTS systems, averaged phoneme power and pitch frequency were analyzed using American English speech signal with phoneme labeled data. One male speaker was used, not the same as the AE speaker in the MLSDB. Words and short sentences were selected for speech materials. Total number of phonemes was 901. Pitch frequency extraction method was the same as described in section 3. Both pitch and power values were calculated as moving averages for each phoneme.

Figure 7 shows an example of an analysis result. In this figure, a dot corresponds to a phoneme and only seven vowels were analyzed. This figure also plots the correlation coefficient (R) and root mean square error (ER) as first-order regression lines. The two lines parallel the regression line and indicate the permissible threshold (4.1 dB) of power fluctuation by another subjective experiment[7]. It should be clear from this figure that strong correlation (R=0.56, ER=2.6 dB) is clearly noticeable. 86% of all points lie within permissible threshold conditions. A very high correlation coefficient can be achieved by phoneme or the phoneme environment. Therefore, when the pitch frequency information is used for English phoneme power estimation is as accurate as the Japanese TTS system.

6. CONCLUSION

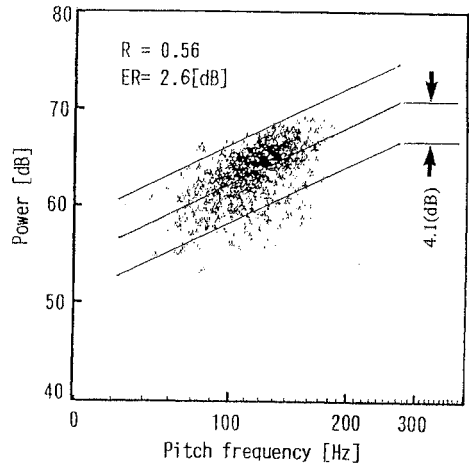

Fig. 7 Relationship Between Pitch Frequency and Speech Power on Averaged Phoneme Segments (Language: American English, Number of Phoneme: 901, Vowels Only)

The relationship between pitch frequency and signal power for twenty spoken languages was determined to ascertain the applicability of a previously proposed power control rule for TTS systems. First, overall characteristic was determined for each language. Second, the short term correlation was analyzed to study the relationship between high correlation values and other information in the speech signal. Last, in order to get information for developing an English TTS system, averaged phoneme power and pitch frequency was analyzed using American English speech signal with phoneme labeled data. Main results were shown below; (1) The average correlation coefficient ranged from +0.72 to +0.44 for the twenty languages. Japanese, Hindi, Dutch and English have high correlation coefficients. Conversely, German, Russian and Norwegian have relatively low correlation. (2) The short term analysis found that high level speech part was accompanied by high correlation values. The reason for this was assumed to be the relationship between accentuation and stress. (3) The prospect for phoneme power control in an American English TTS system was excellent given the strong relationship between pitch frequency and signal power. The results strongly suggest that the relationship between pitch frequency and signal power can be used in speech processing systems for all spoken languages.

ACKNOWLEDGEMENT

The author wish to thank Dr. N. Kitawaki, director of the Speech & Acous. Lab., Dr. N. Sugamura, group leader, for their encouragement during this work. The author also wish to thank Dr. H. Sato, executive manager, NTT Advanced Technologies Corp. for many helpful suggestions.

[References]

[1]C.F. Sacia and C.J. Beck, "The power of fundamental speech sounds," Bell Syst. Tech. J., Vol.5, p.393(1926), [2]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 Acous. Soc. of Japan, SP91-4(1991)(in Japanese), [3]R. Kobayashi and Y. Edano, "Correlation between pitch and intensity patterns in Japanese," Spilling Meeting of Acous. Soc. of Japan, p.127(1966), [4]S. Hiki, "Correlation between increments of voice pitch and glottal sound intensity," J. Acous. Soc., of Japan, Vol.23, No.1, p.2023(1967), [5]J. Suzuki and R. Tanaka, "An LPC vocoder excited by synthesized pitch," Fall Meeting of Acous. Soc. of Japan, p.511(1979)(in Japanese), [6]K. Itoh, T. Hirokawa and T. Sato, "Phoneme power control for speech synthesis," Trans of IEICE, Vol.E76-A, No.11, p.1911(1993), [7]K. Itoh, T. Hirokawa and H. Sato, "Segmental power control for Japanese speech synthesis," ICSLP92 in Banff, p.1143(1992), [8]H. Irii, K. Itoh and N. Kitawaki, "Multilingual speech data base for speech quality evaluation and statistical features," Trans. of IEICE, Vol.E74, No.1, p.33(1991), [9]F. Itakura and S. Saito, "Digital filtering technique for speech analysis and synthesis, " 7th Int. Cong. Acoust., 25-C-1, Budapest(1971)