

Automatic Evaluation of Pronunciation in Language Classes

Yoshiyuki Kawazoe[†] and Hiroshi Kanai[‡]

[†]Education Center for Information Processing,
and

[‡]Department of Electrical Engineering, Faculty of Engineering,
Tohoku University, 980 Sendai, Japan
KAWAZOE@JPNTOHOK and KANAI@JPNTOHOK

An experimental system is proposed for language courses to teach correct pronunciation to foreigners. Dynamic programming technique is applied to compare the student's voice with model pronunciation. Several important features are automatically extracted and accordingly the student's pronunciation is evaluated.

Introduction

Due to general trend in international society, to obtain multiple language skill is becoming necessary. Especially, language skill focused on conversation is now thought to be fundamentally important. Up to the present, most frequently used method to learn pronunciation is to listen to recorded materials. Language laboratory is an advanced version of this lesson. These methods, however, are not satisfactory for the evaluation of achievement level of each student. Everywhere there are absolute lack of good trainers. Moreover, even well trained teachers sometimes cannot explain the difference in pronunciation explicitly to their students. Their classical explanation is qualitative and trainees are only forced to follow the model pronunciation.

On the other hand, recent progress in speech recognition study is dramatic. We planned to apply the result of the speech recognition study to language education[1]. Although the system was originally aimed at the Japanese Training Course for foreign students of Tohoku University, it has been designed and developed to fit general purpose.

Model pronunciation is presented to trainee. The trainee follows the given example and his/her voice is recorded. The voice is analyzed by linear prediction. The Cepstrum coefficients are numerically deduced and the time scale is altered to fit to the model pronunciation by dynamic programming (DP) method[2]. The DP path accordingly obtained is a measure of global and detailed speed differences of the trainee's pronunciation.

The obtained non-linearly matched result gives a good measure of characteristic features of trainee's pronunciation. The most important features we have extracted for the evaluation of the voice are (1) timing (speed), (2) power (stress), (3) pitch (accent, intonation), and (4) formants (tone).

The next section introduces the principal concept of the present study. Following the method indicated, we have performed experimental application of the developed system. Several typical examples of pronunciations are tested and the extracted features are discussed in the following.

Principal Concept of the Present Study

Recorded voice is A/D converted with 12bits accuracy and the sampling rate of 10kHz. Linear prediction is applied to the digitized data with the frame length of 25.6ms and the frame interval of 10ms. After these procedures, the following characteristic features of the voice are extracted.

(1) Normalization of the time scale by dynamic programming

For each frame, the 1st to 15th order LPC Cepstrum coefficients are calculated by the linear prediction, and the voices of trainee and trainer are matched by dynamic programming according to the Euclid distance (distance between smoothed logarithmic spectrum). The time scale of the trainee's voice is thus fitted to the reference by the trainer. The trainee's voice altered according to this procedure is named "normalized voice"

(2) Evaluation of the characteristic voice features

Since, due to the previous procedure, the time scale of the voice of student is matched to the model pronunciation, it is possible to compare and evaluate the student's voice in each frame.

a. Tone difference by Cepstrum coefficients

The normalized voice is evaluated for each frame compared with the model pronunciation in the measure of the Euclid distances of the 1st to 15th Cepstrum coefficients. The total distance is referred to as E_c .

b. Pitch frequency difference

After the linear prediction analysis, pitch frequency is extracted by the clear peaks of autocorrelation of the obtained wave shape of residue. The sum of differences of the logarithmic pitch frequencies for each frame is referred as E_p .

Especially, since change in pitch frequency is very important in Japanese, we should spend a time for the student training concerning this feature. On the contrary, for most of the European languages, intonation is defined according to power accent. In Japanese, even the same sound (in written form) expresses completely different meanings according to pitch. (Exactly speaking, it depends on dialects; standard Tokyo pronunciation and the other pronunciations are sometimes different; as a special case, in Tohoku area we have originally almost no pitch accent!, and the meanings of the words are recognized only with context.) In this study, all the vowels of the model pronunciation are labeled in advance, and clear comparison of change in pitch is thus realized.

c. Power difference

The sum of the differences of powers in each frame is referred as E_a . Power change in speech (strss) is only useful to express emotion in Japanese, and categorized as an advanced subject. The power information itself, however, is very important to check the DP result.

d. Timing

The result of DP, the most fitted path $G(n)$, is used to check the local time difference. If the trainee's voice is completely the same as the model pronunciation, $G(n) = n + b_0$, where b_0 is the difference in frame at the starting point. If the student speaks fast or slow uniformly, $G(n) = a_0 \times n + b_0$. Using the least square fit, these simple functions are fitted to the obtained $G(n)$'s. The differences of the fits are referred as E_{d1} and E_{d2} , and they correspond to starting timing mismatch of voice and to speed difference, respectively. An important factor of Japanese pronunciation is to speak it in a fixed timing; for example, even if the sound is a double consonant. Figure 1 shows the DP matching procedure and the two simple fitting functions.

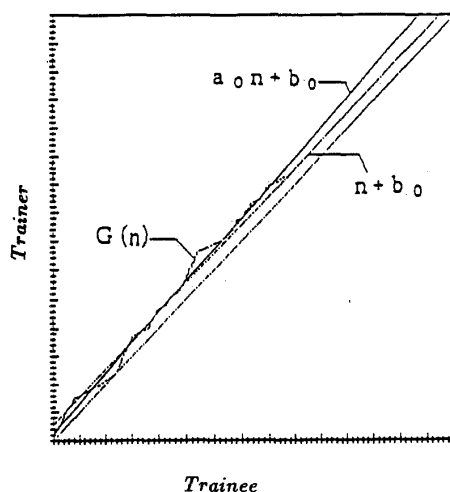


Fig.1. DP path $G(n)$ and two fitting functions.

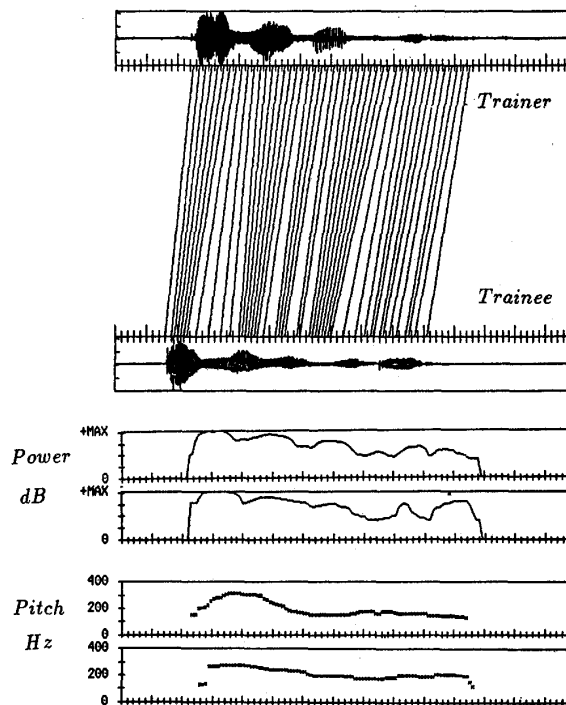
Experiments

Using the developed system described in the previous section, we have evaluated the pronunciations of the foreign students in Tohoku University. The model pronunciation of female was selected from the pronunciation guide tape by NHK[3], and a trainee was a Chinese female student. For male, the model pronunciation is given by a professor of the Japanese Training Course, and an Indonesian male student was tested.

We have recorded 84 short phrases spoken by the 4 people. These phrases are selected to form 42 pairs of characteristic pronunciations. As mentioned above, the two most important features of Japanese pronunciation are its pitch and timing. We have selected the pairs to correspond to these important features; as an example, for pitch, (a) ascending pitch of "watashi ga iku" for a question (Shall I go?), and (b) the same sound in flat pitch for normal expression (I will go.). These specific combinations are analyzed and the results are shown below.

Figure 2 shows the comparison of the extracted characteristic features after the non-linear fitting by DP. This example gives a typical Japanese pitch feature. The same sound of "ame" is used to express two completely different meanings; (a) descending pitch of "ame" is for "rain", ("ame ga furu" = Rain falls.), and (b) ascending pitch of it is for "candy", ("ame wo nameru" = Lick a candy.). It is difficult for new comers to recognize this pitch control. Accordingly, the pitch differences of the students are very poor compared to the model pronunciations as is seen in Fig.2.

(a) a me ga fu ru



(b) a me wo na me ru

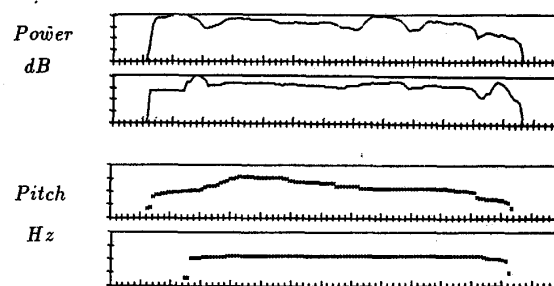


Fig.2. Comparison of pronunciations by trainer and trainee featuring pitch difference.

Another example is the timing. A long sound of "so'oko ga mieru" (I can see a warehouse.), and a short one of "soko ga mieru" (I can see the bottom.) are compared in Fig.3 (a) and (b). In each figure, digitized wave forms, powers (with difference), and pitches (with difference) are shown. The upper sections of them are for the model pronunciation and the lower ones for the trainee's. The obtained values of the coefficients in $G(n)$, E_c , E_p , E_a , and E_d are listed in Table I. Another example is shown in Fig.4, for which a typical abbreviation of two consonant sounds (kaku) into a double consonant (kakku) is indicated. The word "ryokakuki" (means aircraft) is normally pronounced as "ryokakki" in standard Japanese. The rule we are applying is simply understandable to native Japanese speakers but not so for foreigners. Although the value of E_{d2} is a good measure of the timing, it is very difficult to draw the conclusive results. One of the reason is that we evaluate the whole sentence and not the part specifically interested in. It, however, is also important to evaluate the total phrase to extract the difference in actual situation. We are now accumulating analyzed data to establish an evaluation standard for Japanese pronunciation.

(a) so o ko ga mi e ru

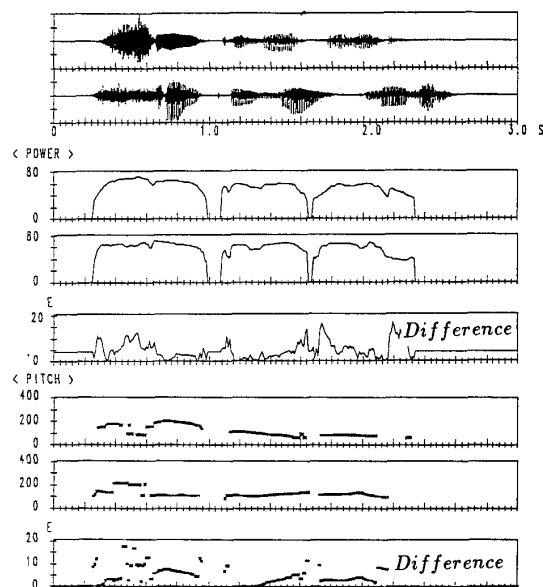


Fig.3(a). Comparison of pronunciations by trainer and trainee featuring a long sound of "o". Upper parts of paired figures are for the trainer and the lower for trainee. Power and pitch differences are also shown.

case	E_c	E_p	E_a	E_{d1}	E_{d2}
3(a)	35.1	3.2	6.9	14.3	4.8
3(b)	34.7	2.8	8.0	29.0	5.5

Table 1. Obtained values of differences in characteristic features between trainee and trainer for Fig.3.

(b) so ko ga mi e ru

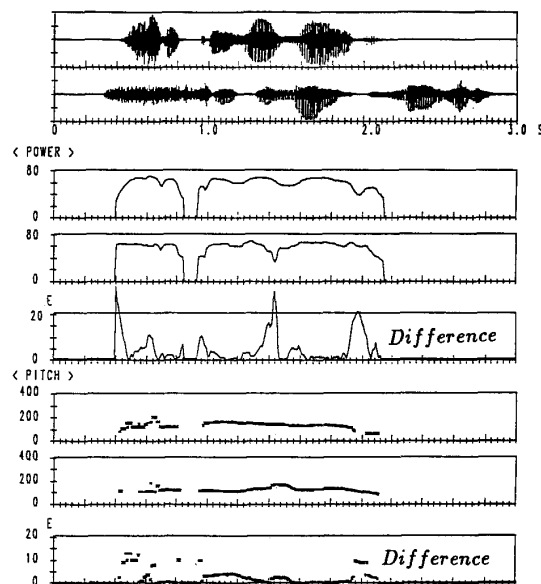


Fig.3(b). Same as Fig.3(a) except for featuring a short sound "o".

(a) ryo ka ku ki

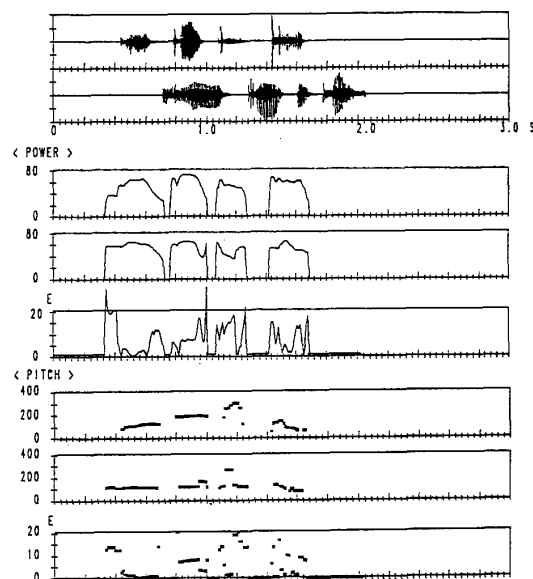


Fig.4(a). Comparison of pronunciations by trainer and trainee featuring a full sound of "kaku".

(b) ryo kakku ki

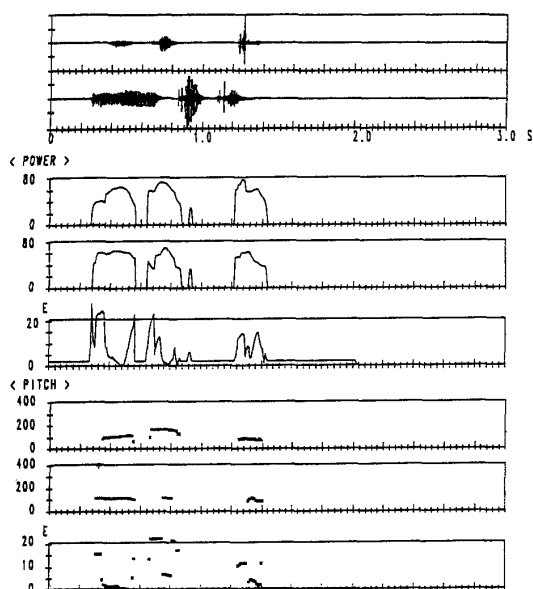


Fig.4(b). Same as Fig.4(a) except for featuring a double consonant sound "kakku".

case	E_c	E_p	E_a	E_{d1}	E_{d2}
4(a)	20.3	12.5	11.6	11.3	1.6
4(b)	22.8	8.9	10.8	3.1	7.4

Table 2. The same as Table 1, except for the values for Fig.4.

Conclusion

Based on the dynamic programming, time scale of the trainee's voice is altered non-linearly to fit to the model. According to the results, we can automatically evaluate the trainee's pronunciation. This method is applied to the students in Japanese Training Course of Tohoku University. This study has just started and the results given here are preliminary. We plan to test the system to large number of students and refine it. To extract the progress of trainee's pronunciation in the course, it is essential to gather basic data and establish the evaluation system. Another important subject is of course to brush-up the curriculum from the lowest to the advanced levels of students. According to the information, we should reconstruct the hardware and software systems to realize the response in real time. The final system is expected to be really useful to assist language training courses.

Acknowledgements

The authors are grateful to Professor Izumi Saita and Professor Hiroshi Hondo with the Japanese Training Course of Tohoku University for their useful discussions. They are also thankful to Mr. Shin Hoshi and Miss Etsuko Hibino for their contributions to this study when they were students with us.

References

- [1] E. Hibino et al., "Automatization of Evaluation of Foreign Student Pronunciation in Japanese Training System", Tohoku Sect. Joint Conv. of IEAIE, Japan, 2B-15 (1988).
- [2] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. ASSP-26, pp.43-49 (1978).
- [3] Ed. by NHK, "Accent Dictionary for Japanese Pronunciation", (1986)