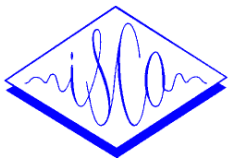

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

# ENGLISH SPEECH TRAINING USING VOICE CONVERSION

*Keiko Nagano and Kazunori Ozawa*

C&C Information Technology Research Labs., NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 213 Japan

## ABSTRACT

This paper proposes an English prosody training method using voice conversion technique. The unique point of this proposed method is that voice converted synthetic speech is used to train English prosody pronunciation. The synthetic speech is produced by converting important prosodic parameters in the student's speech into corresponding native English speaker's speech, while the student's voice characteristics except prosody are preserved. By using the proposed method, pronunciation problem can be easily found out, and training efficiency is improved. The comparative evaluations for training efficiency of the proposed training method and the conventional method show that the proposed method, using the voice converted synthetic speech, is more effective in English prosody training than the conventional method, using the native English speaker's original speech.

## 1. INTRODUCTION

In the conventional English speech training method[1], a native English speaker's speech is usually presented as the teacher's speech. The teacher's voice quality and the student's voice quality are usually very different. Accordingly, when students practice English pronunciation by using the conventional training method, it is not easy for them to focus on their pronunciation problems by just listening to the teacher's speech, whereupon the training efficiency would be low. If the teacher's voice quality is as the same as the student's one, it is presumed for students to train English pronunciation easier.

The authors propose a new English prosody speech training method, using voice converted synthetic speech. The proposed method uses the synthetic speech as the teacher's speech, whose voice characteristics except prosody are the same as the student's. This method can reveal the pronunciation problems clearly as the difference between the target speech and the student's speech, and the students could imagine the target of their practice easily. Further, they can train English prosody pronunciation by the proposed method more effectively than by the conventional method.

The next section explains an overview of the proposed English training method and the procedure for producing synthetic speech. In Section 3, important prosodic parameters for English pronunciation are examined. Finally, the comparative evaluation of the proposed training method and the conventional method is carried out in Section 4.

## 2. ENGLISH SPEECH TRAINING METHOD BY VOICE CONVERSION

Figure 1 shows a block diagram for the proposed English prosody training method. Voice converted synthetic speech is used as the teacher's speech as shown in Fig. 1. The native speaker's speech and the student's speech are analyzed and prosodic parameters are calculated. The student's prosodic parameters are converted into the native English speaker's prosodic parameters in the conversion process. In the synthesis process, the converted prosodic parameters, the student's LPC parameters and others are used to produce the voice converted synthetic speech. For the speech training, the students can listen to both the teacher's speech and the student's speech, and catch up the difference and correct their pronunciation.

Figure 2 shows the procedure of the voice converted synthetic speech. Voice detection is used to find the start and end points of student's speech. Then, Dynamic Programming (DP) matching[2] is used to correspond phonetic segments in the student's speech to the native English speaker's speech. The native English speaker's speech has already been analyzed and phonetic segmentation has been carried out. After the DP matching, the student's speech is analyzed by LPC analysis. In this analysis, the residual signal and the LPC parameters are calculated. The analysis order is 10 and the sampling frequency is 10 kHz. Prosodic parameters are calculated from the residual signal in each phonetic segment. Prosodic parameters in the student's residual signal are converted into native English speaker's prosodic parameters after the analysis. The residual wave excitation synthesis method [3] is used to produce the voice converted synthetic speech.

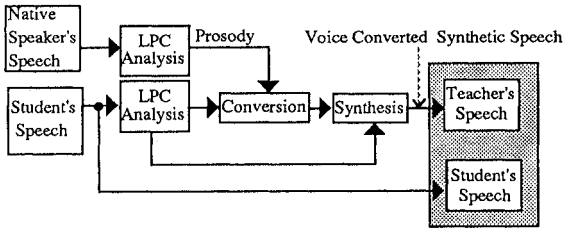


Fig. 1 Proposed English speech training method.

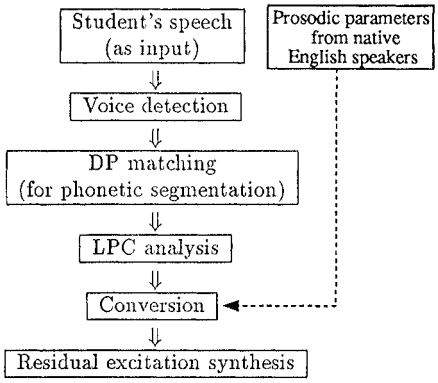


Fig. 2 Procedure of the voice converted synthetic speech.

The voice converted synthetic speech has the same voice characterisitcs except prosody as the student's one, and has correct prosody pronunciation, which is obtained from the native speaker's speech. Therefore, it is expected that students can easily find what part of their pronunciation is incorrect, by comparing the synthetic speech and the student's speech. Moreover, students can more easily imagine the target of their practice by hearing the correct pronunciation, presented in the synthetic speech which preserves their own voice characteristics except prosody.

## 3. EXPERIMENT 1: EVALUATION FOR IMPORTANT PROSODIC PARAMETERS IN ENGLISH PROSODY PRONUNCIATION

Before starting English prosody training, the first evaluation experiment [4] was carried out to find the most important prosodic parameters for the English pronunciation.

### 3.1 Data

An American male was chosen for the native English speaker's example, and a Japanese male was selected as the student's example. The utterances were "ambassador", "embassy", and "imbalance".

### 3.2 Prosodic Parameters

The prosodic parameters discussed here are pitch pattern and duration of sonorants. The authors used two kinds of parameters for the pitch pattern, called F0 contour A and F0 contour B. F0 contour A has the same F0 and pitch contour as that for the native English speaker. F0 contour B has the same pitch contour as those for the native English speaker, but has different F0. The F0 in F0 contour B is as the same as the student's one. F0 contour B is calculated by the following equations:

$$\Delta P_T(i) = P_T(i) - \bar{P}_T \quad (1)$$

$$P'_S(i) = \Delta P_T(i) + \bar{P}_S \quad (2)$$

where $\Delta P_T(i)$ is the pitch contour of the native English speaker. $P_T(i)$ and $\bar{P}_T$ are the $i$-th pitch period and average pitch period of the native speakers, respectively. $\bar{P}_S$ is average pitch period of the student.
In the evaluation, the synthetic speech was produced by converting one or more of these prosodic parameters in student's speech into native English speakers' speech. There were 6 samples ( 5 kinds of synthetic speech and 1 original speech ) for each utterance. The combinations were as follows:

1. F0 contour A(FA)
2. F0 contour B(FB)
3. Segmental duration of sonorants(D)
4. F0 contour A and segmental duration of sonorants(FA&D)
5. F0 contour B and segmental duration of sonorants(FB&D)
6. Student's original speech(Org)

### 3.3 Subjects and Evaluation Method

The subjects of this evaluation test were 12 ( 10 male, 2 female ) native English speakers. They were all teaching English in Japan. A paired comparison method was used for this evaluation. The subjects were asked to choose which speech was closer to native English prosody pronunciation. The subjects evaluated 15 pairs for each word.

### 3.4 Results

Figure 3 shows the result of synthetic speech evaluation. In Fig. 3, the lowest average score is plotted on 0, and the highest average score is plotted on 100. Others are plotted between these two points, according to their scores. The scores of F0 contour A (FA) and F0 contour B (FB) were not significantly different from that of the original speech. The result showed that F0 contour A and duration (FA&D) were the two important parameters for changing Japanese English prosody to English like prosody. Duration is more important than F0 contour A. When one of the prosodic parameters was converted into the native English speaker's prosodic parameters, the evaluation score was not as high as the two parameters simultaneous conversion. The three most important parameters evaluated by the native speakers were:

1. F0 contour A and segmental duration of sonorants(FA&D)
2. F0 contour B and segmental duration of sonorants(FB&D)
3. Segmental duration of sonorants(D)

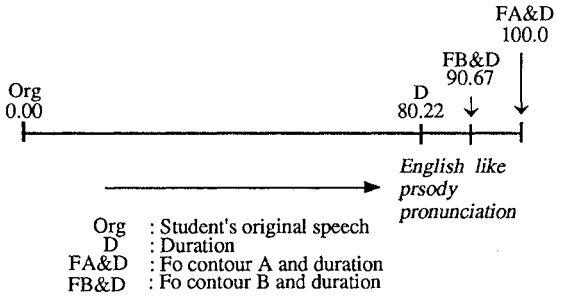


Fig.3 Evaluation result of experiment 1.

The difference between FA&D and FB&D was about 10% in the average score. From the results of this evaluation, it was concluded that for English prosody training, F0 contour A and duration should be converted to produce voice converted synthetic speech for English prosody training.

### 3.5 Discussion

Ohyama [5] reported that the fundamental frequency contour and the segmental duration for phonemes were the two most important parameters for achieving accurate English prosody pronunciation. His statement was almost the same as the authors' result obtained from the experiment, which showed that the F0 contour A and duration are the two most important parameters in the English prosody. Therefore, the authors decided to use the synthetic speech, in which these two parameters are converted, as the teacher's speech in the proposed English speech training method.

## 4. EXPERIMENT 2: COMPARISON BETWEEN TWO TRAINING METHODS

Two separate groups of students were trained to determine which training method is more effective in the training efficiency for English prosody training. One group was trained by the proposed method using synthetic speech. The other group was trained by the conventional method using original speech spoken by native English speakers.

### 4.1 Data

An American male was used for an example of native English speaker's utterance. Ten Japanese male subjects were chosen as students for this experiment. Before the practice, they pronounced forty samples, to make a recording for pre-training speech. The pre-training speech utterances were evaluated in regard to prosody by two different male native Americans. They were asked to evaluate all samples with 5 point scores shown in Table 1.

### 4.2 Training Procedure

According to the score of pre-training speech, the subjects were divided into two groups. The average English speaking levels for each group were almost equivalent at this time. Group A practiced English prosody by the proposed method, using the synthetic speech. Group B practiced English prosody by the conventional method, using the native English speakers. Ten words were used for the experiment. The list of words is shown in Table 2.

In one training session, the total number of speaking and listening activities was controlled by giving instructions to the subjects. Figure 4 shows an example of the training procedure. Each training period was about 30 minutes. All subjects were required to practice in the scheduled time. All subjects practiced twice a day during two days. The second practice was carried out 6 hours after the first practice, to make the same span between the trainings for every subject. A personal computer was used for this practice. The personal computer was equipped with a means to automatically draw a pitch pattern for the intonation and intensity for the rhythm. The intonation and rhythm were calculated from teacher's speech and the subject's speech. The computer also had recording and sounding functions, so the students could hear their own speech, as well as the teacher's speech. After each training period, all subjects had their speech recorded. The obtained data were used for comparative evaluation of post training.

Table 1. Evaluation Scores and Corresponding Expressions.

| Score | Expressions |
|---|---|
| 1 | Not close to the native speaker's pronunciation at all. |
| 2 | Slightly close to the native speaker's pronunciation. |
| 3 | Fairly close to the native speaker's pronunciation. |
| 4 | Very close to the native speaker's pronunciation. |
| 5 | The same as the native speaker's pronunciation. |

Table 2. Word List.

| No. | Word |
|---|---|
| 1. | message |
| 2. | announce |
| 3. | incident |
| 4. | immediate |
| 5. | systematic |
| 6. | damage |
| 7. | imagine |
| 8. | ambassador |
| 9. | communication |
| 10. | question |

### 4.3 Evaluation Test for Pre/Post Training

The comparison test was carried out to evaluate the two training methods. Samples used in this test were the pre/post speech utterances for all students in two groups. Two male native Americans evaluated prosody pronunciation of each individual sample according to 7 points evaluation scores. The evaluation scores and corresponding expressions are shown in table 3.

### 4.4 Results

The evaluation test results for pre/post training are shown

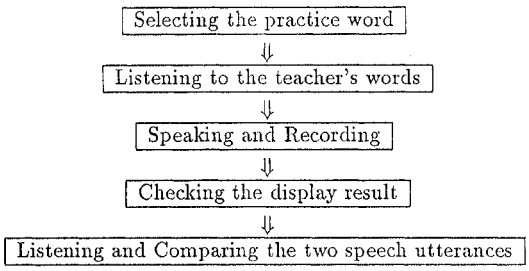


Fig. 4 Training procedure example.

Table 3. Evaluation Scores and Corresponding Expressions.

| Score | Expressions |
|---|---|
| -3 | Typical Japanese |
| -2 | Fairly Japanese |
| -1 | Slightly Japanese |
| 0 | Middle |
| +1 | Slightly English |
| +2 | Fairly English |
| +3 | Typical English |

in Fig. 5. In Fig. 5, average evaluation scores and standard deviations for each group are plotted. The two dimensional analysis of variance [6] is used to examine the interaction between two variables and to determine the training effect achieved by each group. The two variables are "pre/post training" and "group difference". In addition to this analysis, group differences for each training method were examined by the analysis of variance. The results of the two dimensional analysis of variance indicate that in both groups the evaluation scores for pre/post training are significantly different from each other. This means that both groups had significant improvement after the training.

The pre training scores were not significantly different between two groups. The post training scores were significantly different. In post training, the average score for Group A was 1.68, while the Group B score was 1.13. The score achieved by Group A was significantly higher than that by Group B. Therefore, Group A was considered to have been trained more effectively than Group B. From the result of the evaluation, it is concluded that the proposed method, using the voice converted synthetic speech, is more effective in the training efficiency for English prosody training than the conventional method, using the native English speaker's speech.

## 5. CONCLUSION

A new English prosody speech training method, using the voice converted synthetic speech, has been proposed. The synthetic speech was produced from a student's voice, whose important prosodic parameters were converted into those of native English speaker's. In the English prosody training,

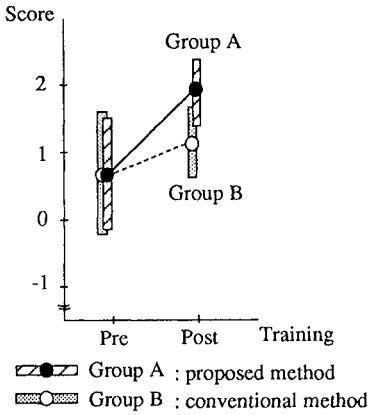


Fig. 5 Evaluation result for pre/post training.

the proposed method and the conventional training method were compared in view of training efficiency. Pre and post evaluation test results showed that the proposed method, using the voice converted synthetic speech, was more effective in English prosody training than the conventional method, using native English speech. Several problems still remain in producing the synthetic speech, such as automatic pitch detection and phonetic segmentation. Further study is necessary to solve these problems.

## ACKNOWLEDGMENTS

The authors wish to thank Dr. T. Yamazaki of the Applied Information Technology Research Laboratory for meaningful discussions on the training efficiency evaluation.

## REFERENCES

[1] Y. Uekawa, et al., "Development of English Speech Training System," *IEICE Technical Report*, ET86-12, pp.49-52, 1987 (in Japanese).

[2] H. Sakoe, et al., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. ASSP*, Vol. ASSP-26, No. 1, pp.43-49, FEBRUARY, 1978.

[3] K. Iwata, et al., "A Speech Synthesis System Using Pitch Controlled Residual Wave Excitation," *Proc. of Fall Meeting of ASJ*, 3-2-7, 1988 (in Japanese).

[4] K. Nagano and K. Ozawa, "English Speech Training Method Applying Voice Conversion," *Proc. of Spring Meeting of ASJ*, 2-5-8, 1990 (in Japanese).

[5] G. Ohyama, et al., "A Study on the Prosody of the English Spoken by Japanese Speakers," *Onsei gengo*, vol.3, pp.284-298, 1989 (in Japanese).

[6] S. Iwahara, "Kyouiku to shinri no tame no suikeigaku (Statistics for Education and Psychology)," *Nihonbunkakagaku sha*, pp.274-282, 1986 (in Japanese).