A VOWEL TRAJECTORY DISPLAY FOR SPEECH TRAINING

Ken'iti KIDO, Kenji TANAHASHI, Yasuhiro OHUCHI

Dep. of Computer Science
CHIBA INSTITUTE OF TECHNOLOGY
2-17-1 Tsudanuma, Narashino 275, Japan

ABSTRACT

A speech display system is developed for the evaluation and the training of speech utterance. The speech is analyzed by linear predictive technique every 5 ms and the frequencies of the lowest two spectral local peaks P1 and P2 are extracted. The vowel trajectory is displayed using those frequencies on the P1-P2 plane. In most cases, P1 and P2 correspond to the first and the second formants, but in the case of indistinct utterance, the correspondence between the local spectral peaks and the formants tends to fall into disorder. And the system is considered to be useful for the evaluation of speech quality. The examples of some words uttered by normal speakers and some patients with difficulty in utterance are compared each other for the discussion of the effectiveness of the system.

1. INTRODUCTION

This study was begun by a request of dentist who engaged in the orthodontics. He said that the speech uttered by patients of opposite clench was usually improved by the surgical operation of orthodontic treatment. But, there was no subjective way to assert it. And we begun to develop a method for the evaluation of speech quality.

We tried to display the movement of formant on the first and second formant plane. But we use the local spectral peaks instead of the formants according to the assumption that the local spectral peaks coincide with the formants in the normal and good utterance and there will be some discrepancy between them in the abnormally uttered speech.

The improvement of utterance is not yet subjectively proved by the method. This is a very difficult task. But some interesting results are obtained. And this paper will show some of them.

2. ANALYSIS

The speech sample is first A/D converted at 16 bits accuracy and 16 kHz sampling rate. The speech samples are cut out every 5 ms using the Hamming window of 30 ms length. The cut out sequence is analyzed by 18-degrees LPC after taking the difference of every successive data. From the result, the speech spectrum is made using 1,024 points FFT.

The local spectral peaks are extracted from the speech spectrum. And,we use from the peak of the lowest frequency to express the frequencies of the spectral peaks P1, P2, P3, ,,,.

3. DISPLAY METHOD

To display the transition of speech spectrum, so far mainly used is so called sound spectrogram or sonagram in which the abscissa stands for time axis and the ordinate the frequency. The spectrum of every time slot is drawn by the light and shade and sometimes the contour lines or colors are used to express the intensity of the frequency components. And the frequency axis is usually linear.

Such the expression is very familiar for us speech scientists. But, it is not easy to get useful information on the speech quality from those spectrograms without rich experience. To use the speech display system in the training of utterance, easier method to get useful information for the evaluation of speech quality is strongly required.

3-1. T-P pattern

It is considered that the detailed spectral information is not necessary to use the spectrogram for the visual feedback to the speech trainee. The transition of frequencies of the formants or the local spectral peaks is considered to be more important and effective.

According to the above mentioned consideration, the transition of the frequencies of spectral peaks is displayed as shown in Fig.1 which is called the T-P pattern and uses up to 6 peaks from the peak of the lowest frequency. The frequency scale in the ordinate is logarithmic. Using the logarithmic frequency scale, it becomes easier to observe the transition of significant spectral peaks.

3-2. P1-P2 pattern

The points are dotted every 5 ms on the P1-P2 plane using P1 and P2 as the values of coordinates of abscissa and ordinate, respectively. The scales of abscissa and ordinate are the logarithmic frequency. To draw the trajectory, the averaged value of successive 8 values of P1 and P2 are computed every 5 ms and the averaged points are connected by line as shown in Fig.2. Taking such the averaged values, the irregularity of the trajectory due to the absence and the addition of the local peaks in the LPC spectra are much reduced.

The beginning point of trajectory is shown by a small circle and the ending point a rectangle in P1-P2 pattern. The elapsed time is indicated by the small triangle every 50 ms.

4. COMPARISON OF T-P AND P1-P2 PATTERNS BY NORMAL UTTERANCE

Figure 1 shows a T-P pattern of a continuous speech /a,i,u,e,o/ uttered by a normal young male speaker. As seen in the figure, T-P pattern gives better perspective than the conventional spectrogram as the transition of the frequency of spectral peaks are clearly displayed and the significant movements of the spectral peaks can be easily observed.

The peaks corresponding to the first formant disappear in some period in /a/ and some discontinuity of peaks corresponding to the second formant are observed at the phoneme boundaries. The peaks higher than P3 seem to give no significant information.

Figure 2 shows the P1-P2 pattern of the same speech. This figure looks like the formant transition pattern on the first and second formant plane. At the ending of the utterance, trajectory goes to the point of /e/ instead of the final phoneme /o/. The power of this part is very low as seen in Fig.1 and no strange sound is felt in hearing.

The frequencies of the spectral peaks corresponding to the higher formants are not displayed on this figure. We investigated also the displays of the speech transition on P1-P3 and P2-P3 planes. But we could not find useful feature from those displays.

5. COMPARISON OF T-P AND P1-P2 PATTERNS BY ABNORMAL UTTERANCE

Figures 3 and 4 shows the T-P pattern and the P1-P2 pattern of the speech /a,i,u,e,o/ ventriloquized by the same speaker. In the ventriloquism, the lips are kept almost closed during the utterance and the speech quality is unfamiliar, but the phonemic information is correctly transmitted.

As seen in Fig.3, the changes in P1 is small but T-P pattern does not express clearly the difference from the normal utterance shown in Fig.1.

The deference is well demonstrated by P1-P2 pattern as shown in Fig.4. It is observed that /e/ approaches to /u/ in this case.

Comparing those figures, it can be said that P1-P2 pattern is superior than T-P pattern for the visual feedback of utterance training.

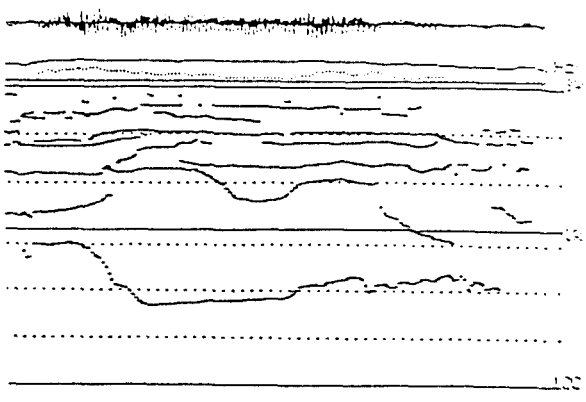

Figure 1

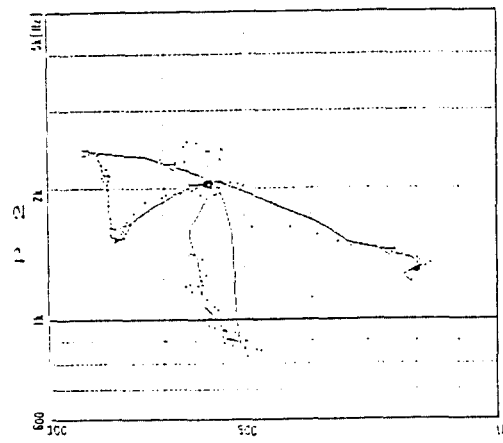

Figure 2

/a,i,u,e,o/ uttered by a normal young male speaker

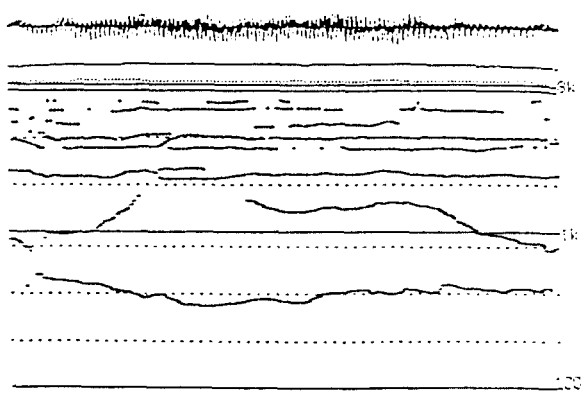

Figure 3

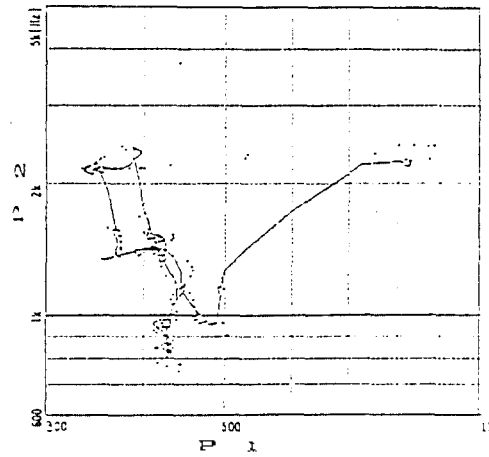

Figure 4

/a,i,u,e,o/ ventriloquized by a normal young male speaker

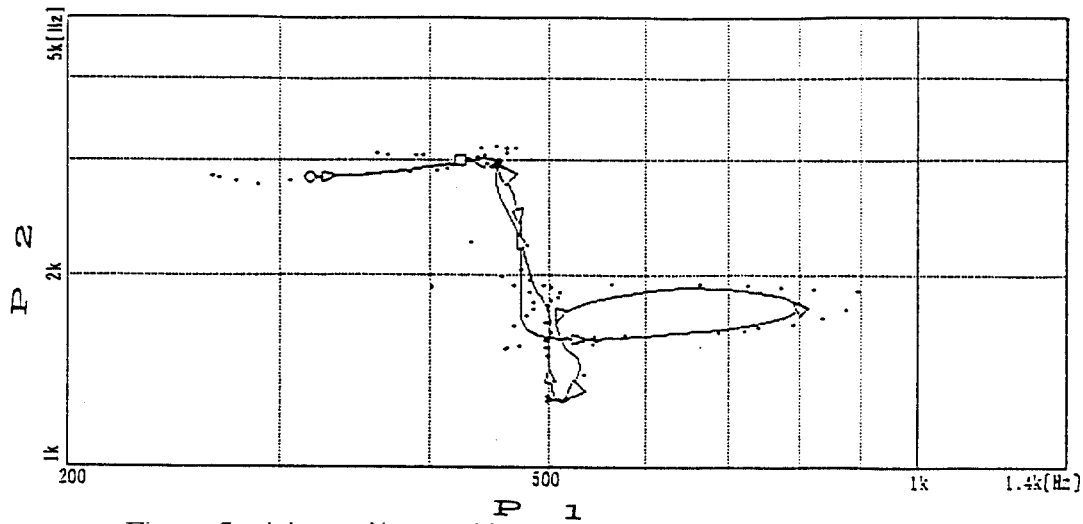

Figure 5 /niwatori/ uttered by a normal young female speaker

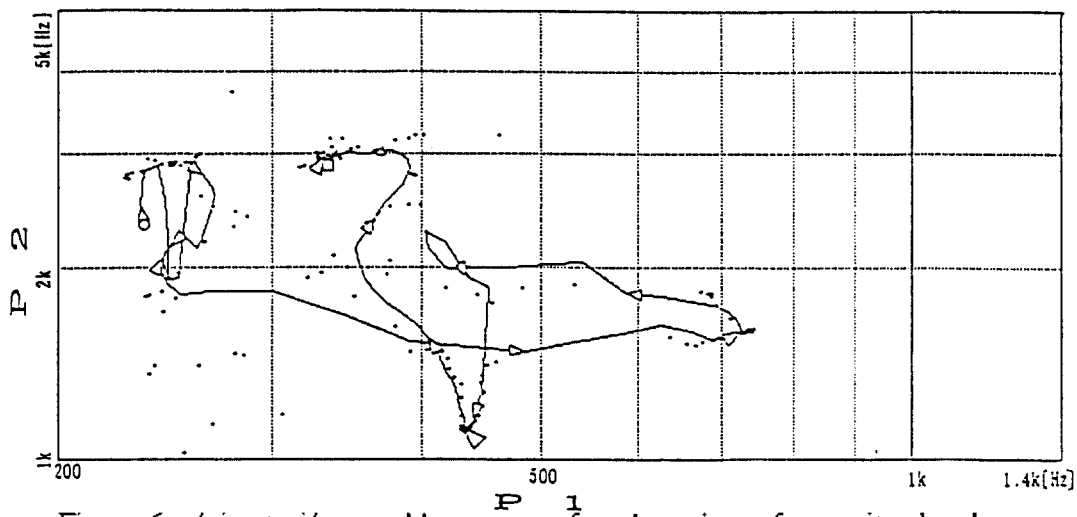

Figure 6 /niwatori/ uttered by a young female patient of opposite clench

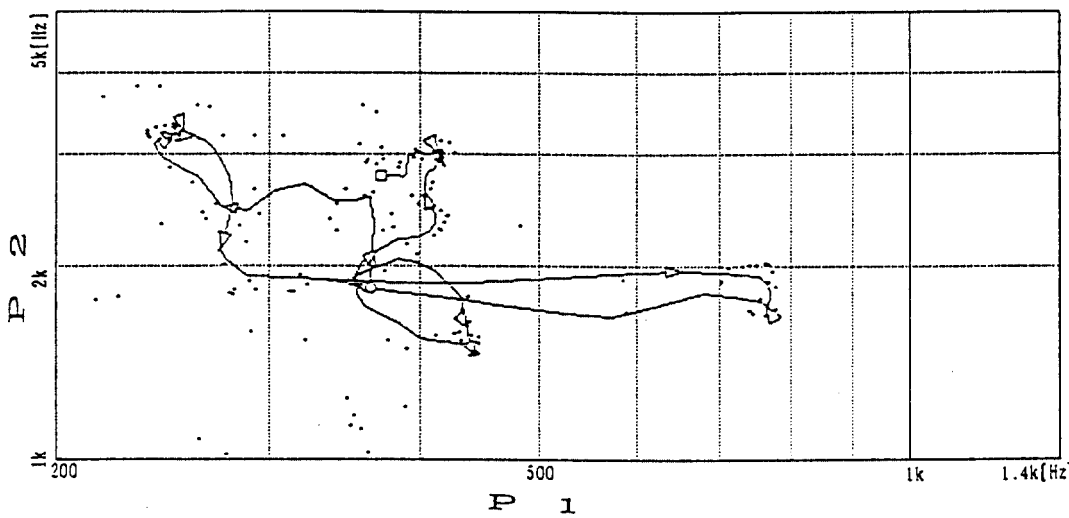

Figure 7 /niwatori/ uttered by a young female patient one year after the surgical operation

6. EXAMPLES OF P1-P2 PATTERNS OF SPOKEN WORDS

Figure 5 shows P1-P2 pattern of /niwatori/ uttered by a normal young female speaker. The trajectory has just the expected pattern as the first and second formant pattern except for that P2 is somewhat higher. But the speech is very clear in hearing. It is considered from the observation of many utterances that the higher P2 is a special feature of her utterance.

The same speech uttered by a young female patient of opposite clench is analyzed and displayed on Fig.6. The main difference of this figure from Fig.5 is lower P1 in the initial part /ni/ of the word. The quality of speech uttered by the patient is inferior than that of the normal speakers in hearing.

We could analyze the same speech uttered by the same speaker one year after the surgical operation. Figure 7 shows the P1-P2 pattern. The substantial difference from Fig.6 can not be observed in Fig.7.

The improvement of the utterance will be in the other points, for example in alveolar conconants and the easiness of utterance.

7. SUMMARY

A vowel trajectory display system is now under being developed for the purpose of the use as a visual feedback system in the utterance training and as a speech quality evaluation system.

The frequencies of the local spectral peaks are used as the main feature to display.

T-P pattern and P1-P2 pattern are compared and the superiority of P1-P2 pattern is demonstrated using the speech samples of normal and abnormal utterances.

The system is considered to be useful for the use in the speech training of the language which has many vowels as Japanese.

Acknowledgment:
The authors thanks to Professor Takayuki Kuroda, Ms. Rieko Ohashi and Professor Hiroshi Mimura for their kind offer of speech samples and helpful discussions and Dr. Takahiko Ono for his kind support of this work.