Vowel Training Method Using Formant Space Information

Ilsuh Bak* · Cheolwoo Jo**

ABSTRACT

In this paper, we develop a vowel training assistant method using vowel formant statistics. Formant statistics were obtained from a PBW set consisting of 452 words from 8 persons. Then we calculated distance from input formants to each center of vowel formant space. Based on the distance, directions could be given to correct the speaker's manner of articulations, i.e. position of jaw and tongue.

Keyword: Formant, Training, Vowel

1. Introduction

Importance of communication through correct pronunciation in this information-age is rapidly increasing. Correction of disabled people's pronunciation is being performed by a small number of speech pathologists. Moreover, to be such an expert requires much time and extensive training.

Accordingly using training software is one method to overcome such barriers. Training by computer software can give visual feedback and can give the trainee detailed goals and increased motivation. Such software is being developed in various ways and types. [1][2][3][4]

The purpose of this paper is the development of assistant software for automatic correction and training of speech disabilities. It focuses on vowels using speech signal processing methods. We used formant statistics from a large speech database. Then we suggested a method to give information to guide vowel articulation movements in the right way based on the measured distance from the center of the formant map of vowels.

* SASPL, School of Mechatronics

** Changwon National University, Korea

2. Overview of Proposed Methodology

The proposed method in this paper is based on the formant space of vowels. Figure 1 shows the vowel articulation map of Korean vowels. It is well known that formant (F1, F2) space is similar to the articulatory vowel map. If we can detect the formant differences between the speaker's vowels and normal vowels, it is possible to compute the relative difference of location of tongue and jaw. [1]

Figure 2 shows a concept of the proposed method. First we computed the formant statistics from a large database, which consists of normal pronunciation. To obtain good results, a large-size database has to be compiled. Such databases have to be clustered in sex and age groups. From the formant statistics, the mean and deviations of each vowel is computed. So the mean of F1 and F2 in each vowel within the F1-F2 space becomes the numerical center of the vowel. Such centers are marked as o in figure 2.

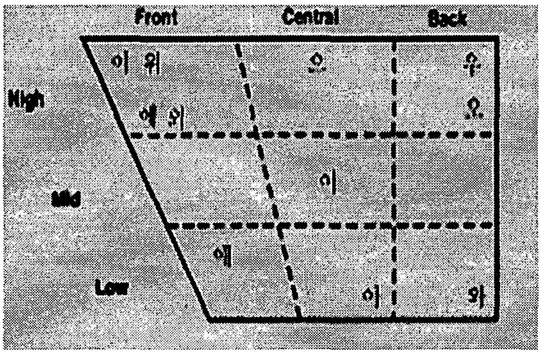


Figure 1. Articulation Map of Korean Vowel

If unknown speech is given, the formant values are computed and the distance to each vowel's mean formants are computed. If the calculated distance between input and target vowel's mean formant is the smallest, the input is considered to be coincide with target vowel. So no further action is required.

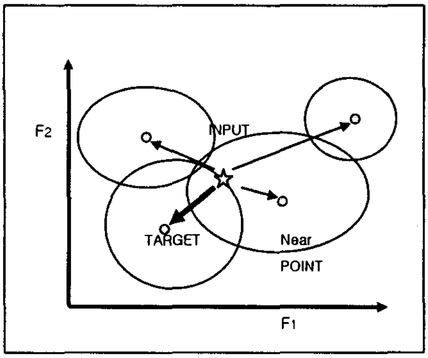


Figure 2. Tracking of Target Voice from Input

If the computed distance between input and the target is not the minimum, a distance between input and target is computed and the direction to the formant mean of target vowel is measured. Then the speaker is notified how his or her articulation can be changed toward the target by moving the speaker's mouth and jaw. The degree of the movement is computed by the difference of values of Fl and F2.[5]

The value of Fl is inversely proportional to the tongue position, while the value of F2 is proportional to jaw position and height of tongue. F2 tends to have a larger value for front vowels and a smaller value for back vowels. Fl tends to have larger values for lower jaw positions. Based on such knowledge, we can suggest directions correlating to correct pronunciation by referencing formant positions between the input vowel and the target vowel. Generally Fl-F2 spaces have wide variances depending on age and gender. Because of this, some kind of normalization procedure is usually required. To account for this, this experiment is confined to male announcers between the ages of twenty and thirty years old.

To analyze the collected voices, we computed formant frequencies of vowels in the database using linear predictive analysis. 17th order LPC analysis was performed. To compute the Fl and F2 values, the parabolic interpolation method was used. The Euclidian distance measure was used to compute distance between target and input.

3. Experimental results

The data used in this experiment is the subset of SITEC's PBW DB consist of 452 isolated words recorded from 8 male speakers whose ages are between 20~30. The speakers have normal voices considered to be standard.

We extracted only vowel parts after segmentation of each phoneme by automatic segmentater using HMM.

Recording equipments and conditions were as follows:

-Microphone: Senheizer HMD224X

-Place: soundproof booth

-Sampling: 16KHz, 16bit

After being recorded and stored on digital audio tape, the data was stored in Sun Workstation Sparc 20, using the AD/DA Module of DAT-Link+.

Table 1 shows speakers' geographical background and ages

Table 1. Speakers' information

speaker	ago	Born area	Living area	a living period	Patient's born area	
					father	mother
lyj	40	kwangju	daejyun	15	kwangju	kwangju
yol	27	seoul	jyunbuk	5	seoul	seoul
kbw	30	jeju	jyunbuk	6	jeju	jeju
sch	25	kyungki	jyunbuk	2	kyungki	kyungki
ssh	26	seoul	jyunbuk	2	gangwon	chungnam
lgh	25	kyungki	jyunbuk	3	kyungki	jyunnam
lgh	25	inchyun	chungnam	3	chyungbuk	incchyun
shj	28	junnam	jyunbuk	9	jyunnam	jyunnam

Table 2 shows means and variance distributions of each vowel formant using linear prediction coefficients from the database.

Table 2. Mean and variance distribution of vowel formant

Vowel		F1		F2		Sample Frequency
Eng	Kor	mean	variance	mean	variance	
aec	ㅐ ㅔ	468	89	1895	157	1441
axc	ㅏ	662	115	1440	168	1357
eoc	ㅓ	502	90	1193	218	1090
euc	ㅡ	409	144	1604	309	829
euic	ㅢ	337	42	2166	321	336
ixc	ㅣ	347	112	2174	234	1198
jac	ㅑ	665	93	1496	175	266
jec	ㅖ	404	72	2047	177	105
jeoc	ㅕ	493	79	1449	237	609
joc	ㅛ	395	64	1375	265	277
juc	ㅠ	334	73	1850	263	217
oxc	ㅗ	388	75	1087	261	932
uxc	ㅜ	372	151	1360	401	708
wac	ㅘ	651	103	1292	191	343
wec	ㅞ	445	70	1799	162	468
wic	ㅟ	332	59	2108	195	259
woc	ㅝ	465	51	1062	153	217

From the database we only used Korean single vowels which are /axc/, /eoc/, /oxc/, /uxc/, /ixc/, /euc/, /aec/.

Figure 3 shows the formant distribution from the 8 male speakers' utterance database. In the figure, separate regions of each vowel are represented in different gray levels. From this distribution, the center values of F1 and F2 is computed. These values are marked 'o' in figure 4. The F1 value is the smallest in vowel /uxc/ and the F2 value is the smallest in vowel /oxc/.

To implement the suggested method, separate input vowels were spoken. One was the correct vowel while the other was a simulated wrong vowel. These vowels were recorded in the same sampling rate and resolution as the experimental database.

Then, we calculated formants from the recorded voice and computed the distance from the mean values of each vowel.

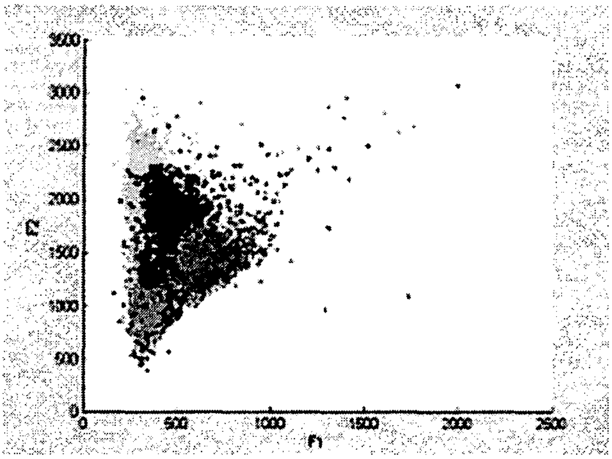


Figure 3. Formant Distribution of Utterance Database

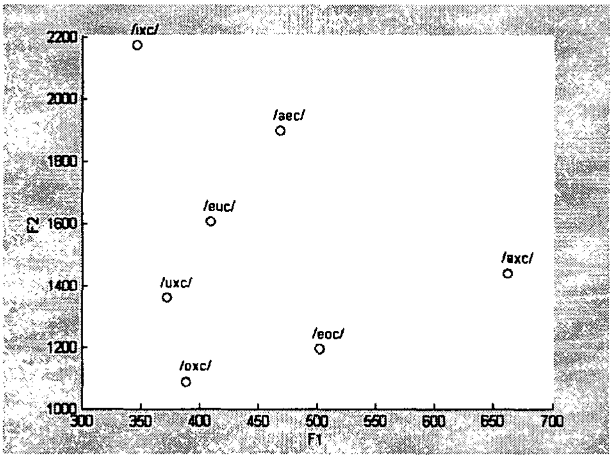


Figure 4. Statistical Vowel Mean Map

From the computed distance of F1, the direction of a tongue's movement is decided as high or low. Those rules are as follows:

If the distance of F1 is positive, the tongue's position has to be lowered. If the distance of the F1 value is negative, the tongue's position has to be moved higher.

If the distance of the F1 value is smaller than 20, the tongue has to be moved only a small amount. If it is bigger than 20, it has to be moved much more.

Distance of F2 decides the direction of a tongue's movement into the front or back of the mouth. If the distance of the F2 value is positive, the tongue's position has to move back. If the distance of the F2 value is negative, the tongue's position has to move to the front. As the distance range decreases below 200, the movements become smaller.

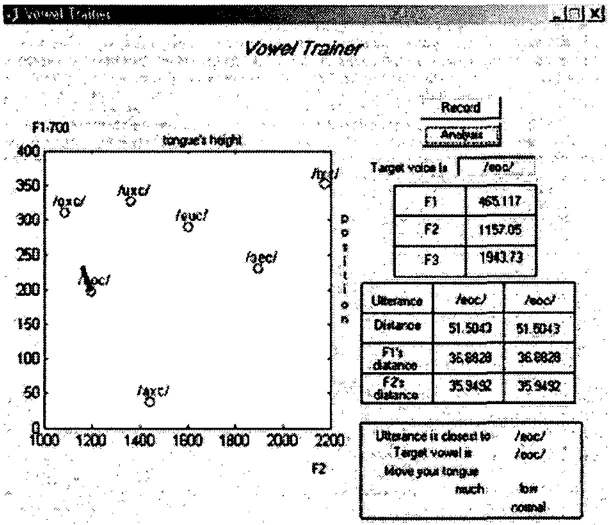
Two different cases are demonstrated. One is when the input is a similar vowel and the other is in the case of a different vowel (the /eoc/ vowel).

First, in the similar vowel's case, we can confirm that the input does indeed have the closest distance with /eoc/ as in table 3.

Table 3. Distance for identical vowel input

	F1	F2	Distance
axc	197	283	345
eoc	37	36	52
oxc	-77	-70	104
uxc	-93	203	223
euc	-56	447	450
ixc	-118	1017	1024
aec	3	738	738

The screen shot of an actual trainer in this case is shown in figure 5. It shows the speaker's utterance is the closest to the target vowel and it is in the normal range.



Target voice is /eoc/

F1	465.117
F2	1157.05
F3	1943.73

Utterance	/eoc/	/eoc/
Distance	51.5043	51.5043
F1's distance	36.8828	36.8828
F2's distance	35.9492	35.9492

Utterance is closed to /eoc/
Target vowel is /eoc/
Move your tongue
much low
normal

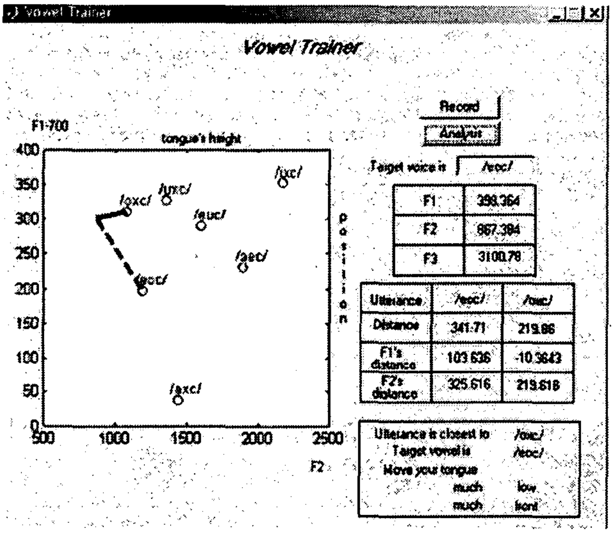
Figure 5. Example of correct pronunciation of vowel /eoc/

The next case is when the input is different from the target vowel /eoc/. Table 4 shows the computed distance values and figure 6 shows the screen shot of the trainer.

Table 4. Distances for wrong vowel input

	F1	F2	Distance
axc	264	573	630
eoc	104	326	342
oxc	-10	220	220
uxc	-26	493	493
euc	11	737	737
ixc	-51	1307	1308
aec	70	1028	1030

In this case, the trainer suggests to move the tongue lower and to the front.



Target voice is /eoc/

F1	398.364
F2	867.384
F3	3100.78

Utterance	/eoc/	/oxc/
Distance	341.71	219.86
F1's distance	103.636	-10.3643
F2's distance	325.616	219.618

Utterance is closest to /oxc/
Target vowel is /eoc/
Move your tongue
much low
much front

Figure 6. Example of wrong pronunciation of vowel /eoc/

4. Conclusions

In this paper we calculated the statistical characteristic of formant values from a normal voice database. Based on that, we proposed a method to suggest directions for correcting pronunciation using the distances of F1 and F2 frequencies. A vowel trainer was constructed and the usefulness of the tool was demonstrated.

The suggested method can be used to teach and train vowel pronunciation for disabled people. However, the current limit of this system is that current formant distribution is not in the normalized form, nor is it variable for different age groups and genders. To reduce such drawbacks separate statistical values have to be computed from a larger data set and proper normalization methods have to be derived.

References

[1] Jo, C. W., Bak, I. S., Jung, E. T. 2003. 'Development of Vowel Training Assistant Method Using Formant Statistics.' *Proceeding of the 2003 Korean Signal Processing Conference*, vol. 16, No.1, pp. 325-328.

[2] Blair, A. D., Ingram, J. 2003. 'Learning to Predict the Phonological Structure of English Loanwords in Japanese.' *Applied Intelligence 19*, pp. 101-108.

[3] Vicsi, K., Roach, P. 2000. 'A Multimedia, Multilingual Teaching and Training System for Children with Speech Disorders.' *International Journal of Speech Technology 3*, pp. 289-300.

[4] Blamey, P. J., Sarant, J. Z., Paatsch, L. E. 2001. 'Effects of Articulation Training on the Production of Trained and Untrained Phonemes in Convert-sations and Formal Tests.' *Journal of Deaf Studies and Deaf Education,* vol. 6, no.1, pp. 32-42.

[5] Yang, B. 1992. 'An acoustical study of Korean monophthongs.' Journal of Acoustical Society of America 91, 4, pp. 2280-83.

Received: 2004. 11. 01
Aceepted: 2004. 11. 26

▲ Ilsuh Bak
#1 Sarimdong, Changwon, Gyeongnam, Korea
SASPL, School of Mechatronics, Changwon National University (641-773)
Tel: +82-55-279-7559, Fax: +82-55-262-5064
E-mail: ilsuh@korea.com

▲ Cheolwoo Jo
#1 Sarimdong, Changwon, Gyeongnam, Korea
SASPL, School of Mechatronics, Changwon National University (641-773)
Tel: +82-55-279-7552, Fax: +82-55-262-5064
E-mail: cwjo@changwon.ac.kr