포만트 통계치를 이용한 장애모음 발음 훈련 보조 방법에 관한 연구

조철우, 박일서, 정은태
창원대학교 제어계측공학과 음성 및 음향 신호처리 실험실
e-mail : cwjo@sarim.changwon.ac.kr, ilsuh@korea.com

Development of Vowel Training Assistant Method Using Formant Statistics

Cheol-Woo Jo, Il-Suh Bak, Euntae Jung
SASPL, Changwon National University

Abstract

In this paper, we tried to develop a vowel training assistant method using vowel formant statistics. Formant statistics were obtained from PBW set consists 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 to correct the speaker's manner of articulaltion, i.e. position of jaw and tongue.

I. 서론

정보화 시대에 있어서 장애인의 여러 장애 활동 중에서 올바른 발화를 통한 의사소통의 중요성은 날로 증가하고 있는 실정이다. 여러 가지 원인으로 인해 방성 장애가 있는 사람의 발성을 교정하는 과정은 극소수의 언어 치료 전문가에 의해서만 수행되고 있는 형편이며 이 과정에서 전문가가 되기 위해서는 많은 시간과 훈련이 필요하다.

기존의 장애음성 교정을 위한 훈련 방법에서는 교정용 소프트웨어를 이용한 방법이 여러 가지 측면에서 접근이 시도되고 있다. 소프트웨어에 의한 훈련방법은 시각적인 피드백이 가능하며 교정대상자에게 구체적인 목표치와 동기를 부여할 수 있다는 점 때문에 효과적인 방법으로 다양한 측면에서 개발되고 있다.

본 논문에서는 음성 처리 기술을 이용한 자동 언어장애 교정 및 훈련을 위한 보조장치의 개발을 목적으로 입력된 대상 음성의 포먼트를 분석하여 통계분포를 구하고 데이터베이스에서 분석된 정상 발음의 포먼트 분포의 중심값(평균)과 비교한 결과를 이용하여 입력된 발음의 교정을 위한 혀와 턱을 위치에 대한 교정 정보를 구하고자 하였다. 이렇게 구해진 정보를 이용하여 발음자의 발화 태도를 수정 할 수 있는 방법을 제시하였다.

II. 제안된 방법의 개요

본 논문에서는 장애 음성을 훈련하기 위해서 장애음성을 가지고 있는 환자가 발화한 음성으로부터 파라미터를 분석하고 분석된 특징을 토대로 발화하고자하는 음성으로 유도해 줄 수 있는 방법을 제안고자 한다.[4][5]

그림 1에서는 제안된 방법에 의한 교정원리를 설명한다. 먼저 각 모음의 포만트 분포도를 구한다. 통계치가 일반성을 갖게 하기 위하여 우선 충분한 양의 데이터베이스가 확보되어야 한다. 또한 데이터베이스는 성별, 연령별에 따라 독립적으로 데이터가 확보되어야 한다. 이렇게 구한 통계분포를 기준으로 각 표준모음의 중심점이 추출된다. 그림 1에서는 o으로 표시되어 있다.

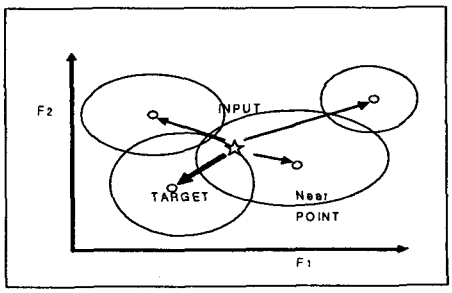

그림 1.입력음성에 대한 목표음성의 추정

그 다음 미지의 음성이 입력되면 역시 포만트 값을 계산하고, 이 값에 의해 각 표준모음과 입력음성의 포만트 중심값과의 거리를 계산한다. 이때 계산된 거리가 가장 작은 것부터 나열해 보면 입력음성과 표준모음간의 유사도 순위를 계산할 수 있다.

만약 거리가 최소인 음성이 목표음성과 일치 하지 않을 경우 F1과 F2의 차를 계산하고 이를 바탕으로 조음방법을 조절하도록 유도하여 입력된 음성이 목표값에 근사하도록 피드백을 시켜준다. 일반적으로 F1, F2값은 각 모음에서 구강내에서의 혀의 위치과 밀접한 상관관계가 있다. 포먼트의 F1에 해당하는 값은 혀의 위치 조절을 통해서 F2에 해당하는 값은 턱의 위치와 혀의 위치를 이용하여 조절이 가능하기 때문이다. 그림2는 혀의 위치에 따른 모음의 조음위치를 나타낸다. 혀의 위치가 낮을 수록 F2가 낮아지고 높을수록 F2가 높아지는 경향이 있다. F1은 전설모음의 경우 값이 작고, 후설모음일수록 값이 커진다. 턱의 위치는 혀의 위치가 낮을수록, F2의 값이 적을수록 낮아진다. 이러한 현상을 바탕으로 현재의 발성과 목표발성간의 조음방법의 차이를 추정하고 개선법을 제시해 줄 수 있다. 이런 방법을 사용할 경우 각 화자의 특성, 성별, 연령 등에 따라 포만트의 분포가 다르기 때문에 포만트 공간의 정규화 과정이 필요하나 본 실험에서는 별도의 정규화 과정을 거치지 않고 20~30대 남성 아나운서의 목소리로 제한하여 실험하였다.

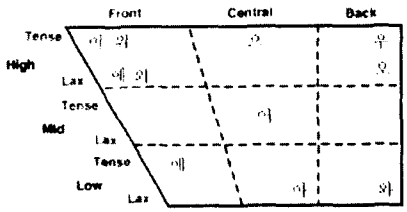

그림 2.혀에 따른 모음의 조음 위치

수집된 음성을 분석하기 위하여 선형예측계수(Linear Predictive Coefficient)를 이용해서 각 모음마다 성도의 포먼트를 분석하였다.

시스템의 전달 함수 $H(z)=\frac{1}{A(z)}$와 A(z)의 근을 갖는 전극 시스템은 H(z)의 극점을 표현한다. 그러므로 선형 예측계수 (∂_i)가 분석에 의하여 구해질 때, 다음의 복소수 방정식의 극점들을 결정한다.[1][2][6]

$$z^p + a_1 z^{p-1} + a_2 z^{p-2} + \cdots + a_{p-1} z + a_p = 0 \qquad (1)$$

식 (1)은 실수 계수 값을 가지는 P차의 방정식이며 일반적으로 p/2개의 공액 복소수를 가진다. 만약 한쌍의 근이 각각 $z_i = r_i e^{-j\theta}$, $z_i = r_i e^{j\theta}$ 라면 포먼트 주파수는 다음과 같다.

$$f_i = \frac{w_i}{2\pi} = \frac{1}{2\pi}\frac{1}{T} arg(z_i) \qquad (2)$$

여기에서 표본화 주기는 T이며, 분석차수인 p가 0에서 $\frac{1}{2T}$의 주파수 범위에 있는 포먼트 개수의 두배라면 윗 식의 근은 포먼트들과 일치하게 된다. 만약 p가 포먼트 개수보다 크다면 그 근은 포먼트 뿐만 아니라. 스펙트럼의 작은 봉우리까지 나타내게 된다. 이 실수 극점은 스펙트럼 포락의 기울기를 나타낸다. 본 논문에서는 12차의 LPC 계수의 이용하였다.[1][3]

III. 실험 및 결과

우선 발성자의 음성을 구분하는 기준을 정하기 위하여 정상 음성을 가지고 있는 직업인인 20~30대의 남자 아나운서 8명으로부터 녹음된 452개의 고빈도의 고립단어나 문장을 발음한 SITEC의 PBW DB중의 일부를 사용하였다. 사용된 데이터베이스는 HMM을 이용한 자동 세그멘테이션에 의해 각 음소를 분할한 뒤 모음 부분만을 추출하여 사용하였다.

조음성분		F1		F2		빈도
영문	한글	평균	분산	평균	분산	수
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

표.1 모음별 포먼트의 평균과 분산

이 DB의 녹음 조건은 방음 부스에서 Senheizer HMD224X를 사용하여 녹음했으며, 디지털 오디오 테이프에 저장된 뒤 A/D는 PC환경에서 실시하였으며, AD/DA Module은 KAY CSL 4300B를 사용하였다. 그리고 16 kHz로 샘플링하고 16 Bits로 양자화 되어 있다.

위의 데이터베이스를 선형 예측 계수을 이용해서 포먼트를 구한 결과를 각 음소 정보에 따른 평균값과 분산값은 표 .1 과 같다.

이때 대표음성 '아', '에', '이', '오', '우' 에 대한 8명의 남자 아나운서들의 포먼트 분포는 그림.4와 같다.

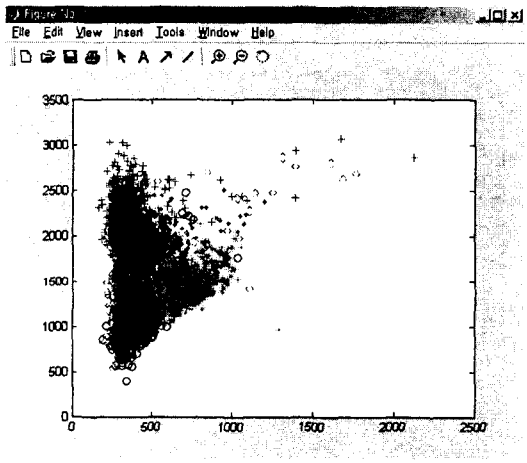

그림4. 발성 데이터베이스의 포먼트 분포

여기에서 발화된 음성이 일정한 포먼트 분포를 가짐을 확인 할 수 있었다.

이러한 포먼트 기준을 이용하여 발화된 음성 정보를 분석해보기 위하여 임의의 대상에 대한 음성 데이터를 채집하였다. 사용된 음성은 일반을 대상으로 '아'발음에 대하여 방음실에서 DAT 녹음기를 이용하여 녹음하였다 16khz의 샘플링 주파수를 가지며 16비트로 양자화 하였다. 임의의 장애음성은 성대암을 가지고 있는 환자의 음성을 이용하였다.

녹음된 음성으로부터 포먼트를 구하고, 구해진 포먼트와 구해진 기준값과의 거리를 환산하였다. 이때, 가장 적은 거리값을 가지는 음성을 선택하였고, 동일한 음성에 대해서는 편차값의 차이로 그 음성을 추론하였다.

다음은 분석을 위해 사용된 음성 '아'와 기준음성의 F1, F2 값의 차이값과 이 차이를 거리값으로 환산한 결과이다.

표2는 입력음성 '아'를 동일 조음으로 판단한 경우이며 표3은 입력음성 '아' 를 유사 위치의 다른 조음으로 판단한 경우이다. 표4는 상대적으로 발음이 명확하지 않은 장애음성 '아'를 '어'로 식별한 경우이다.

조음 위치가 비슷한 다른 조음으로 판단한 경우에는 , F1의 차에 의해 혀를 앞쪽으로 움직이도록 유도함으로써 모음발음을 보정 할 수 있을 것이다.

마지막으로 유사도가 떨어지는 음성으로 판단하는 경우로 실험에서는 성대암 환자 음성의 경우이다. 테스트에 사용된 음성의 경우 '어'로 판단한 경우로 이 경우 재활 치료를 행한다면 혀를 좀더 안쪽으로 넣어 발음을 교정하도록 유도할 수 있다

표2. '아'와 동일 조음으로 판단한 경우

조음		F1차	F2차	거리
aec	ㅐ ㅔ	273	480	552
axc	ㅣ	79	25	83
eoc	ㅓ	239	222	326
euc	ㅡ	332	189	382
euic	ㅢ	404	751	853
ixc	ㅣ	394	759	855
jac	ㅑ	76	81	111
jec	ㅖ	337	632	716
jeoc	ㅕ	248	34	250
joc	ㅛ	346	40	348
juc	ㅠ	407	435	596
oxc	ㅗ	353	328	482
uxc	ㅜ	369	55	373
wac	ㅘ	90	123	152
wec	ㅞ	296	384	485
wic	ㅟ	409	693	805
woc	ㅝ	276	353	448

표3. '아' 를 유사 위치의 다른 조음으로 판단한 경우

조음		F1차	F2차	거리
aec	ㅐ ㅔ	188	551	582
axc	ㅏ	6	96	96
eoc	ㅓ	154	151	216
euc	ㅡ	247	260	359
euic	ㅢ	319	822	882
ixc	ㅣ	309	830	886
jac	ㅑ	9	152	152
jec	ㅖ	252	703	747
jeoc	ㅕ	163	105	194
joc	ㅛ	261	31	263
juc	ㅠ	322	506	600
oxc	ㅗ	268	257	371
uxc	ㅜ	284	16	284
wac	ㅚ	5	52	52
wec	ㅞ	211	455	502
wic	ㅟ	324	764	830
woc	ㅝ	191	282	341

조음		F1차	F2차	거리
aec	ㅐ ㅔ	197	803	827
axc	ㅏ	3	348	348
eoc	ㅣ	163	101	192
euc	ㅡ	256	512	572
euic	ㅢ	328	1074	1123
ixc	ㅣ	318	1082	1128
jac	ㅑ	0	404	404
jec	ㅖ	261	955	990
jeoc	ㅕ	172	357	396
joc	ㅛ	270	283	391
juc	ㅠ	331	758	827
oxc	ㅗ	277	5	277
uxc	ㅜ	293	268	397
wac	ㅘ	14	200	200
wec	ㅞ	220	707	740
wic	ㅟ	333	1016	1069
woc	ㅝ	200	30	202

표4. 장애음성에 '아'의 경우

VI. 결론

본 논문에서는 정상발화음성의 포만트 값의 통계적 특성을 구하고, 포만트 공간과 조음공간의 유사도를 이용하여 발화된 음성의 목표음성과의 조음유사도를 판단하여 장애음성을 교정하는 한가지 방법을 제안하였다.

실험에 사용된 음성에서는 입력값에서 기준 점까지의 거리가 최소인 지점이 목표값으로 선택 되거나 최소값에 근사한 값을 가지는 것을 확인 할 수 있었다.

그러나 포만트 값의 통계적 분포를 구하는 과정에서의 성별, 연령별 차이에 따른 정규화 문제는 향후 해결해야할 과제이다. 또한 도출된 조음현상의 차이를 시각적 형태로 효과적으로 표현해 줄 수 있는 방법의 개발도 필요하다.

참고문헌

[1]L.R. Rabiner, R.W.Schafer, Digital Processing of Speech Signals, Prentice-hall, 1978

[2]R.W.Schafer, L.R.Rabiner,"System for Automatic Formant Analysis of Voiced Speech", *J. Acoust .Soc. Am*, Vol47,No.2, pp634-648

[3]J.L.Flanagan, C.H.Coker, L.R.Rabinier, R.w.Schafer, N.Umeda, "synthetic Voices for Computers", IEEE *Spectrum*, Vol.7, No.10, pp 22-45, Oct 1970

[4]J.l.Flanagan, Speech Analysis, Synthesis and Perception, 2nd Ed, Springer-Verlag, NewYork ,1972

[5]W.Koenig, H.K.Dunn, L.Y.Lacy,"The Sound Spectrograph", *J.Acoust.Soc.Am*, Vol.17, pp.19-49, 1946.7

[6]L.R.Rabiner, R.W.Schafer, C.M.Rader, "The Chirp z-Transform Algorithm and Its Application",Bell System Tech. J.,Vol.48, pp1249-1292, 1969