화자의 발음에대한 통계적 모델의 적용에관한 연구

김 대 식*, 배 명 진**, 윤 계 강*

* 숭실대학교 전자공학과 ** 호서대학 전자공학과

A study on application of the statistic model about an utterance of the speaker

Daesik Kim Myungjin Bae Jaegang Yoon

* Dept.of Electronic Eng. SoongSil University
** Dept.of Electronic Eng. HoSeo College

ABSTRACT

A speech that play a part of important mediation in the man's conversation is the sound of representa-tion to man's emotion and thought, then voice sound could be verified and identified a speaker's speech by individual property. This study indicates as distr-ibution of pitch in searching for sample number of each pitch with eye in the sound waveform of speaker. We propose the algorithm that judge speaker's emotion state, personality, regional gruop, age, sex distinct-ion,e.t.c., according to the deviation degree.

I. 서 론

우리는 어떤 사람의 목소리만 듣고도 감정의 상태가 어떻고 누구인가를 판별한다. 이것은 상대방 음성 내부에 언어와 내용 정보와 개인성 정보를 가지고 있기때문이다.

예로부터 음성을 판단하는 수단은 귀를 사용하는 방법 뿐이었기 때문에 음성에 관한 학문은 주관적이었다. 그러나 1945년 음성을 눈으로 볼 수 있게 분석하는 기계가 Bell 연구소에 의해 개발되어, 1947년 이 음성 분석을 사용한 음성 연구가 Potter 등에의해 최초로 행해졌으며, 1950년 이후 Potter,Peterson 등에의해 미국어 모음의 포먼트 주파수가 측정되었고, Fant는 스웨덴어 모음에 관한 포먼트 주파수를, 그리고 服部와 鳥井은 일본어 모음의 포먼트 주파수를 측정하였다.

1962년 Kersta가 성문(voice print)의 여러 특징을 관찰함으로서 발성자의 개인 식별이 가능함을 발표하였다

그뒤 선형 예측 계수에 의한 음성 인식에 관한 연구와 PARCOR법에 의한 음성 인식에 관한 연구가 진행되어 왔다. 1980년대에는 한국어 숫자음에 대한 연구가 발표되었으나 아직은 한국어 화자 인식에 관한 연구는 적기때문에 본 연구는 화자의 음성 파형에서 각 피치 내에서 샘플수를 눈으로 찾아 샘플수에 대한 피치의 빈도수를 분포도로 나타내었다. 가장 높은 빈도수를 중심으로 그 중심으로부터 이탈 정도에 따라 화자의 감정 상태, 성격, 지역, 나이, 성별 등의 판별을 제안 하였다.

II. 화자 인식법

사람의 물리적 특징(눈, 코, 입, 머리, 얼굴형태)은 지문, 필체, 등을 통하여 구별 할수 있듯이 음성을 분석하여 사람을 식별 할수있다.

성문(voice print)는 우리말로 "목소리의 무늬" 라한다. 음성학상의 성문(Glottis)은 손가락의 지문과 같은 역활을 하는데 폐에서 나오는 공기가 통과하는 성대(vocal cord) 사이의 공간이다. 목소리는 이 성대를 지나는 공기가 진동해 만들어진다. 사람마다 성문, 성대의 구조와 특징이 서로 달라 화자를 인식하는데 도움이 된다. 음성에서 개개인의 차이와 특징은 목소리의 음질(굵고 가늠, 맑고 탁함, 높낮이, 비음정도)과 발음상의 특징(속도, 사투리, 모음음가, 소리의 길이, 세기, 리듬, 억양)으로 나누어지고, 나이에 따른 음성, 학력과 출신에 따른 음성 특성은 그들이 사용하는 어휘, 날말에 주로 나타 난다. 그리고 화자 간의 변

화 즉, 발성자 내부변화(intraspeaker variability)와 발성자 간의 변화(interspeaker variability)가 있는데 전자는 어떤 사람이 같은 단어를 여러번 발음하면 이들은 완전히 같지 않은데, 이런 변화를 사람들은 잘 판별하지 못한다는 것이며 후자는 다른 사람이 같은 단어를 발음하면 듣는 사람은 그 변화를 잘 알수 있다는 것이다.

주로 발성자 내부 변화보다 발성자 간의 변화가 크기때문에 화자를 인식할수 있는 요소가 되어 청각으로, 소나그램을 보고, 기계에 의해 화자를 구별한다. 기계에의한 방법은 컴퓨터로 음성을 분석하고, 특징 파라미터를 추출하여 선형예측(LPC), PARCOR계수, 포먼트, 피치, 기본 주파수 등을 사용한다.

III. 기본 주파수 특성

음성 생성에서 기본 주파수(fundamental frequency) Fo는 음성의 유성음 부분동안 성대의 열고 닫는 비율로서 정의되는데 이 비율을 Hertz(Hz)로 나타낸다. 여기서 중요한 특성 차이는 Fo와 "피치"로 알려진 Fo의 지각적 상관(perceptual correlate)사이에 만들어져야만 한다. 그래서 성문의 단속하는 주기는 사람마다 다르기때문에 개개인의 특성을 결정해 준다. Fo와 달리 피치는 성대 진동 비율 뿐만 아니라 음성 강세(intensity)와 같은 요소에 의존한다.

음성파(sound wave)의 진폭 대시간은 신호의 유성음 segment에서 "준 주기(semiperiod)" 성질을 나타낸다. 반복하는 파형의 가장 긴 부분(portion)은 성대의 개폐에 상응하므로 "성대 싸이클(glottal cycle)"이라한다. 성문의 여닫는 주기가 8ms라하면 성문의 기본 주파수Fo는 1/8(ms)(=125Hz)로 된다.

그림1은 음원(sound source)측면을 고려한 것으로 성대의 특성을 G(z)로 놓고 역수를 취하면 성문 특성이 결정된다.

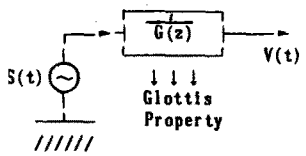

그림1 음원(sound source)의 모델화

IV. 음성에 대한 통계적 모델의 고찰

디지탈 파형 표현을 논하는 대에서 흔히 음성 파형이 ergodic random 처리에 의해 표현될 수 있다고 가정하는 것이 편리하다. 통계적 관점에서 이런 모델을 사용하면 유용한 결과를 얻게 된다.

아날로그 신호 x(t)를 샘프링하면 이산시간 랜덤 프로세서의 샘플 sequence가 된다. 통신 시스템 분석에서 아날로그 신호의 적당한 특성화는 1개확률 밀도p(x)로 존재하고 랜덤 프로세스의 자기상관 함수는 다음 식으로 정의 된다.

$$\phi_a(\tau)=E[x_a(t)x_a(t+\tau)] \qquad (1)$$

여기서, E[ ]은 괄호 내에서 기대값을 나타낸다. 아날로그 전력 스펙트럼은 $\phi_a(\tau)$의 fourier transform이다; 즉,

$$\Phi_a(\tau)=\int_{-\infty}^{\infty}\phi_a(\tau)e^{-j\Omega\tau}\,d\tau \qquad (2)$$

랜덤 신호$x_a(t)$를 샘플링에 의해 얻어진 이산 시간 신호는 다음 식의 자기상관 함수를 가진다.

$$\begin{aligned}\phi(m)&=E[x(n)x(n+m)]\\&=E[x_a(nT)x_a(nT+mT)]\\&=\phi_a[mT]\end{aligned} \qquad (3)$$

이래서 $\phi(m)$은 바로 $\phi_a(\tau)$의 샘플된 version이기 때문에, $\phi(m)$의 전력 스펙트럼은 다음식으로 주어진다.

$$\begin{aligned}\Phi(e^{j\Omega T})&=\sum_{m=-\infty}^{\infty}\phi(m)e^{-j\Omega Tm}\\&=\frac{1}{T}\sum_{k=-\infty}^{\infty}\Phi_a(\Omega+Tk)\end{aligned} \qquad (4)$$

위 식은 음성의 랜덤 프로세스 모델에서 샘플된 신호의 전력 스펙트럼은 원래 아날로그 신호의 전력 스펙트럼에 대한 aliased version이다. Amplitudes x(n)에 대한 확률밀도 함수는 x(n)=x(nT)이기 때문에 amplitudes x(t)에서와 같다는 것이다. 이래서 mean과 variance와 같은 평균은 원래 아날로그 신호에서처럼 그 샘플들에 대해 같은 것이다.

통계적 개념을 음성신호에 적용시킬때, 음성 파형으로부터 확률 밀도와 상관 함수(또는 전력 스펙트럼)를 평가할 필요가 있다. 확률 밀도는 많은 샘플들 - 즉, 긴 시간에 걸쳐 - 에 대한 amplitude의 histogram을 결정하므로서 측정된다.

V. 실험및 결과

실험에 사용된 음성 데이타는 10개, "영 ~ 구", 독립된 한국어 디지트음을 사용했다. 데이타는 12비트 A/D 변환기를 사용했으며, 이들 모든 데이타는 16비트 IBM PC XT 컴퓨터에서 처리했다.

한 사람의 화자가 "영"에서부터 "구"까지 발음한 것

을 A/D 변환기를 거쳐 각 샘플된 데이타를 디스켓에 저장하였다. 이 음성 신호의 데이타에 대한 피치 주기를 눈으로 일일이 찾아 각 피치의 샘플수를 check해 간다. 이렇게 "영"에대한 것을 찾고 차례로 "구"까지 찾아보면 한 피치내의 샘플수가 같은 것이 여러개 나오게 된다. 그 같은 샘플수가 나온 것을 모두 더하면 화자가 "영"에서 "구"까지의 모든 샘플에대한 피치 빈도수가 결정된다. 표1은 한 피치 내의 샘플수에 대한 피치의 빈도수를 나타내었다. 물론 한 두번이 아닌 여러 차례 해보아 평균을 취해 표 1에 나타난 빈도수를 그래프로 표시해보면 그림2와 같은 곡선이 그려 진다.

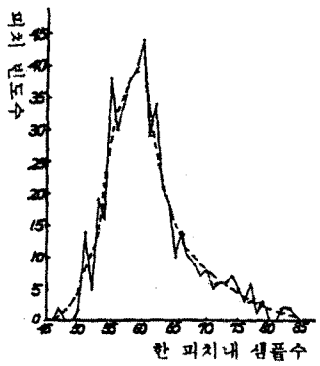

그림2 deviation에 따른 확률 분포 (화자 A)

x축 상에는 한 피치 내의 샘플수를 놓고 y축 상에는 같은 샘플수를 갖는 피치의 갯수 (샘플에 대한 빈도수)를 놓고 나타낸것이다. 그래프에서 보듯이 가장 높이 나타나있는 60개의 샘플에 대한 피치 빈도수는 44개로 표시되어 있다. 이 60을 이탈률의 기준으로 정하여 그 점을 0으로 놓는다. 그림(A),(B) 각각에대한 전체 곡선 상에서 한 피치내 샘플수에 대한 피치 빈도수를 평균하면 높이가 나온다. 다음과 같은 절차를 거쳐 화자에 대한 음성을 분석, 정규화 했다.

(1) 전체 곡선 상에서 높이, 길이, 면적 비 계산
(A) 총 피치갯수: 488개
각각의 총 샘플수: 84-47=37
평균 피치갯수: 488/37=13.2

13.2는 평균 높이이므로 최고치 44에 대한 높이비를 계산하면

13.2/44=0.3

그리고 13.2와 곡선과 교차하는 두 점이있는데 그 점에서 x축 상으로 선을 내려보면 각각 만나는 점이 있을 것이다. 이 같은 52.7과 65.5인데 중심점(x축상)으로부터 거리를 구해보면

60-52.7=7.3
65.5-60=5.5

즉, 52.7 60 65.5
←-7.3 — 0 — +5.5→

같이 나온다. 그런데 13.2는 높이가 되며, 7.3+5.5=12.8 한같은 밑변 길이가 된다. x축의 총 길이는 84-47=37이므로 길이의 비를 구해보면

12.8/37=0.346

이 된다. 또, 면적을 구하면

13.2*12.8=168.96Cm²

이다.

(B) 총 피치갯수: 482개
각각의 총 샘플수: 87-41=46
평균 피치 갯수: 482/46=10.5
위와 같은 방법으로 차례대로 구해보면
54-47.5=7.5
64-54=10

47.5 54 64
← -7.5 — 0 — +10 →

그래프상에서 평균 피치 갯수 10.5는 높이가 되므로 최고치 51에 대한 높이 비를 구해보면

10.5/51=0.2059

이 된다. x축 상에서 길이의 비를 구하면

17.5/36=0.486

이다. 또, 면적은

10.5*17.5=183.75Cm²

이다.

(2) 양분된 곡선 상에서 높이, 길이, 면적 비 계산
(A-1) 좌측 상에서
총 피치갯수: 286개
각각의 총 샘플수: 60-47=13
평균 피치 갯수 : 286/13=22

높이 비: 22/44=0.5
길이 비: 22와 곡선과 만난 점에서 x축 상으로 내려 만난점 54이므로
60-54=6
6/13=0.46
면적: 22*6=132Cm²

(A-2) 우측 상에서
총 피치갯수: 246개
각각의 총 샘플수:84-60=24
평균 피치갯수: 246/24=10.25

높이 비: 10.25/44=0.233
길이 비: 66.7-60=6.7
6.7/24=0.279
면적:10.25*6.7=68.675Cm²

(B-1) 좌측 상에서
총 피치갯수:167개
각각의 총 샘플수:54-41=13
평균 피치갯수: 167/13=12.85

높이 비: 12.85/51=0.252
길이 비: 54-48.4=5.6
5.6/13=0.43
면적: 12.85*5.6=71.96Cm²

(B-2) 우측 상에서
총 피치갯수: 366개
각각의 총 샘플수: 87-54=33
평균 피치 갯수: 366/33=11.1

높이 비: 11.1/51=0.218
길이 비: 63.5-54=9.5
9.5/33=0.288
면적: 11.1*9.5=105.45Cm²

곡선의 최고치를 중심으로 x축상의 좌측을 -로 우측을 +로하여 이 기준점으로부터 곡선이 얼마만큼 이탈되느냐에 따라 그 사람의 성격, 감정 상태, 나이, 성별, 지역성 등을 알게 된다.

기준점으로부터 곡선 형태가 적게 이탈되었다면 그 사람은 일반적으로 차분한 성격이며 감정 상태가 안정되 있다고하는데, 목소리의 tone이 비교적 고르다고 봐야하

겠다. 위와같이 두 사람에 대해 실시했지만 다음 각 사항에 해당되는 여러 사람의 목소리 특성을 조사해 분포도의 곡선 형태에따라 분류하는 작업을 세우고 있다.

- •. 각 지역의 화자(방언)에 대한 음성 특성 조사
- •. 남. 여 성별에 따른 음성 특성 조사
- •. 연령에 따른 음성 특성 조사
- •. 학력에 따른 음성 특성 조사
- •. 직업에 따른 음성 특성 조사
- •. 체격에 따른 음성 특성 조사

VI. 결 론

사람의 목소리와 말씨는 개개인의 특성을 갖고 있어 특정인을 인식하는 단서가 되고 있다. 특정 파라미터를 추출하여 선형 예측 계수(LPC), PARCOR계수, 포먼트 등의 기법을 이용한 음성처리는 좋은 분석 결과를 얻으나, 시간이 많이 소요되거나 다른 제약 요건이 많다. 그러므로 위의 여러기법들을 사용하기 전 데이타를 Group별로 분류, 인식하게 하는 사전 분류 처리 단계로 시간과 노력을 절약할 수 있게 된다.

이런 맥락에서 본 연구는 각 화자에대한 음성 특성을 통계적으로 고찰, 확률 분포도상에서 여러 방법(면적, 길이 비교)을 실행하여 특징화하는데 중점을 두었다.

참 고 문 헌

[1] L.R. Rabiner and R.W. Shafer, "Digital processing of speech signals, ";Prentice Hall, Ic., 1978.

[2] A.E. Rosenberg, "Automatic speaker verification: A review"; Proc. IEEE, 64, 475-487, 1976.

[3] A.E. Rosenberg, "Evaluation of an automatic speak-er verification system over telephone lin-es"; B.S.T.J., 55, 6, 723-743, 1976.

[4] M.R. Sambur, "Selection of acoustic features for speaker identification"; IEEE Trans. ASSP-24, 176-182, 1975.

[5] L.G. kersta, "Voiceprint identification"; J.A.S.A., 34, 725. 1962

[6] H. Levin and W. Lord, "Speech pitch frequency as emotional state indicator";IEEE Trans. SMC, 5, 2, 259, 1975.

[7] R. Demori, P. Laface, V.A. Makhonine, M. Mezzalama. "A syntatic procedure for the recognition of glottal pulses in continuous speech"; Pat-tern Recognition 9, 181-189, 1977.

[8] G. Fant, "Speech analysis and synthesis"; Technical (final) reort. Stocholm. Speech Transmission Laboratory. 1962.

[9] Myungjin BAE, "A study onthe fundamental frequency extracting of speech signals using second order rundown method." Seoul National Uni-versity, MA paper , Jan. 1983.

[10] 김영일, 차일환, 산업기술연구소 논문집 제1집 제1권 (22), 1985.

표1 화자A에 대한 음성 분석

한 피치 내의 샘플수	샘플수에 대한 피치 갯수	발음 숫자 0	1	2	3	4	5	6	7	8	9
47	2									2	
50	2								2		
51	14							1	13		
52	5	1							2		2
53	19							4	4		11
54	16		4		1				4		7
55	38			11		1		5	2	11	8
56	30		2	9		1	3	5		6	4
57	34			3			12	2	6	7	4
58	38		4	7		6	9	4	3	3	2
59	39		7	2	8	5	6	3	1	4	3
60	44		3		15	9	6	2	2	1	
61	29	3	10	3	1	5			1	3	3
62	34	8	6	1	4	5	4	2	2	2	
63	21	5			6	3	3			1	3
64	18	9	4		2				1	2	
65	10	2		3		3	1			1	
66	14	4			1		2	1		4	2
67	10	1	2	2		1	1		2		1
68	9	5	1			2				1	
69	5			1		3			1		
70	10		2	2	2		2			2	
71	4	1			1		1			1	
72	7	1		1	1					3	1
73	6	2		2	2						
74	7			1		4	2				
75	5	1		2	1					1	
76	3				2			1			
77	6		1	1	1		1		2		
78	1										1
79	3	1			2						
82	2				1		1				
83	2				1						1
84	1				1						

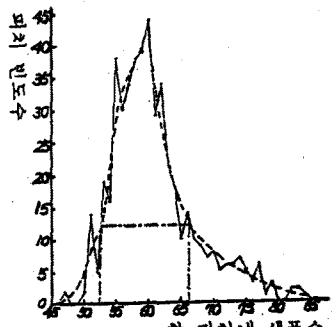

그림3
(1) 전체 곡선 상에서 높이, 길이, 면적

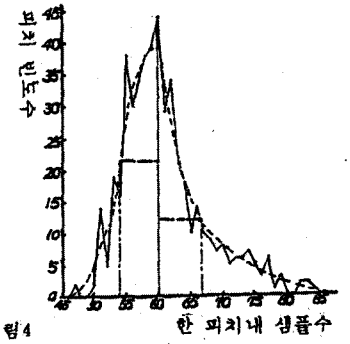

그림4
(2) 양분된 곡선 상에서 높이, 길이, 면적