발성크기에 따른 Glottal Spectrum 성분 분석

이윤주, 조남수, 배명진
숭실대학교 정보통신공학과

Glottal Spectrum Analysis According to Speaking volume

Yoonjoo Lee, Namsu Cho, Myungjin Bae
Dept. of Telecommunication, Soongsil University
E-mail : mjbae@saint.soongsil.ac.kr

요약문

사람은 연령, 성별등에 따라 성도(vocal tract), 성대(vocal cord, 혹은 vocal fold), 비강(nasal tract) 등 발성기관의 차이가 있고, 이는 음성의 음색, 높낮이 등 음향 특성에 영향을 미치며, 시간이 지나감에 따라 변하는 특성을 가지고 있다. 예를 들어, 발성기관의 차이가 큰 남성과 여성은 동일한 단어를 발성하더라도 음향학적으로 매우 큰 차이를 보이며, 이러한 특성은 다른 문장 발성 시에도 음향학적으로 일정한 영향을 미치게 되므로 정적특성이라 한다. 본 논문에서는 이러한 정적특성 중 음성의 발성크기에 따른 Glottal Spectrum을 비교·분석 하고자 한다.

1. 서 론

음성에 의한 의사전달은 말하는 이가 전달하고자 하는 생각에서 출발하여 말하는 이의 생각을 일련의 신경과정과 근육의 움직임을 통해 음향 압력파로 듣는 이의 청각 시스템에 나른다. 이러한 인간의 의사소통 수단인 음성을 기계가 인지할 수 있게 하는 것을 음성인식 시스템이라 하고 이러한 음성 인식 알고리즘 개발은 현재 활발히 진행중이다. 개인이나 특정 단체의 정보의 보안을 위해서는 사용자의 확인 과정이 필요하다. 이때 확인 절차는 사용자에게 사용이 용이해야 하며 확인 내용은 정확해야 한다. 이러한 점을 고려하여 근래에 들어 사용자의 음성특성을 이용한 사용자 확인 방법이 고안되었다. 즉, 사용자가 특정 패스워드(Password) 또는 임의의 음성을 발성한 뒤 발성된 음성을 바탕으로 사용자를 확인하는 방법이다. 이러한 방법에는 화자가 발성한 음성으로부터 스펙트럼의 특성을 나타내는 특징벡터를 추출하여 저장된 각각의 기준패턴과 패턴매칭(Pattern Matching)을 통해 화자를 인식하는 방법이 있다.

이러한 패턴매칭 과정에서 어느정도로 고차 포만트 성분을 보상해 주느냐에 따라 인식률이 좌우될 수 있다. 고차 포만트 성분의 보상을 위해서는 현재 프리엠퍼시스 필터를 사용하고 있다. 본 논문에서는 이 필터를 이용해 보다 적응적으로 보상하기 위하여 음성 인식 과정시 차이가 날수 있는 음성발성 크기에 따른 Glottal Spectrum의 변화도 구하여 프리엠퍼시스 필터에 사용되는 계수 a값을 비교·분석하였다.[9]

2. 음성생성모델

2.1. 음성신호의 생성

음성신호의 구조에 대한 연구는 음성정보를 추출하거나 강조할 수 있다. 따라서 음서신호의 생성에 대한 수학적 모델은 음성을 처리하는데 있어서 매우 중요한 영역이다. 성도(vocal tract)는 성대(vocal cord)와 입술 끝까지를 말한다. 따라서 성도는 인두와 입 또는 구강으로 구성된다. 남성의 성도 길이는 17cm정도이다. 성도의 단면적은 혀, 입술, 턱 그리고 0㎠(완전히 닫혔을 때)에서 약 20㎠까지 변화하는 연구개의 위치에 의해 결정된다. 비도는 연구개에서 시작하여 콧구멍에서 끝난다. 연구개가 낮아질 때 비도는 비음을 생성하기 위해 음향학적으로 성도에 연결된다. 음성생성의 과정을 연구함에 있어서 수학적 모델로서 물리적 시스템을 표현하는 것은 매우 중요하다. 그림 2-1은 성문의 구조적 그림을 보여주고 있다. 블록도는 허파, 기관지와 호흡기관으로 구성된 하부-성문 시스템을 표현한다. 이 하부 성문 시스템은 음성을 생성하는 에너지원이다. 음성은 간단히 공기가 허파로부터 방출되고 결과적으로 성도에 있는 협착점에 의해 공기가 동요될 때 이 시스템으로부터

방사되는 음향학적 파형이다.

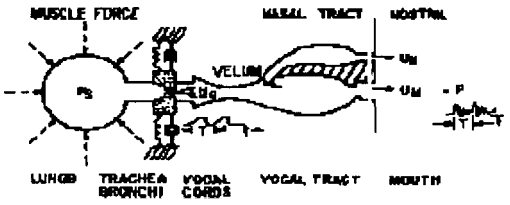


그림 2-1. 음성발성기관의 블록도

유성음은 조정된 성대의 팽창과 함께 성문을 통한 공기의 힘에 의해 생성되어 감쇄 진동하여 성도를 자극하는 공기의 준 주기적인 펄스를 만든다. 마찰음 또는 무성음은 성도에 있는 어떤 점에서 협착을 형성하고, 동요를 만들기 위해 고속으로 협착점을 통과하는 공기의 힘에 의해 발생된다. 파열음은 완전히 입을 폐쇄하고, 이 폐쇄 뒤에서 압력을 만들어 갑자기 느슨하게 함으로써 생성된다.

2.2. 포만트 주파수

성도의 단면적은 입술까지 이르는 각 부분에서 다양하게 변하게 되지만 평균적으로 5㎠로 볼 수 있다. 만약 성도가 구부러지지 않고 일직선으로 되어 있고 실린더로 생각한다면 실린더의 길이는 17cm이고 단면적이 5㎠, 그 지름은 2.5cm로 생각할 수 있다. 이러한 관에서 첫 번째 공명은 관의 길이가 파장의 1/4과 같게 되는 주파수이다. 그러므로 파장은 68cm가 된다. 음파의 속도는 340m/s이므로 그 주파수는 500Hz가 된다. 즉, 주파수축에서 보면 500Hz에서 뾰족한 부분이 있다는 것을 나타내고, 이 뾰족한 정도는 그 관이 얼마나 딱딱한가에 의해 결정된다. 공명이 500Hz에서만 일어나는 것은 아니고 두 번째 공명주파수는 파장의 3/4에서, 세 번째 공명은 파장의 1/4에서 일어난다. 이러한 공명주파수는 또한 어떠한 음성을 발음하느냐에 따라 혀의 움직임에 의해 그리고 목안에서의 관의 움직임에 의해 그 공명주파수가 달라지게 될 것이다. 역으로 우리가 만약 입술에서 나온 음성파형의 공명주파수를 알면 그 사람이 어떤 음성을 발음했는지를 알아낼 수 있을 것이다. 또한 스펙트럼에서 생각해 보면 이러한 공명주파수는 스펙트로그램의 봉우리를 나타내게 된다. 이것을 음성의 포만트 주파수라고 한다.[6]

2.3. 포만트 크기

포만트 크기라는 것은 스펙트로그램에서의 포만트의 높이를 말한다. 모음과 같은 경우에는 직렬합성기에 의하여 합성되기 때문에 F1, F2, F3, F4, F5의 개개 포만트 크기를 직접 조정할 수 없다 . 다만 대역폭의 크기를 조정함에 의해서 간접적으로 조정할 수 있다. 다만 대역폭의 크기를 조정함에 의해서 간접적으로 조정할 수 있을 따름이다.[1][6]

2.4 포만트 대역폭

포만트 대역폭이라는 것은 포만트 주파수에서의 포만트 주파수에서의 포만트 크기의 3dB아래의 대역폭이다. 포만트 대역폭은 음소에 따라 포만트 주파수의 위치에 따라 큰 차이를 보인다. 또한 하나의 음소 내에서도 발음상태에 따라 변동이 심하다. 일반적으로 말하면 포만트 대역폭은 포만트 주파수가 높을수록 커진다고 말 할 수 있다.[1][6]

3. 화자인식 시스템

3.1 화자 인식의 분류

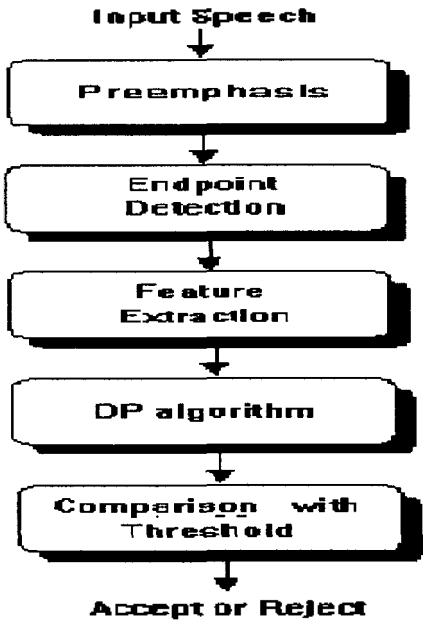
일반적으로 화자 인식은 크게 두 가지로 나누어 처리되고 있다. 첫째로 화자식별(Speaker Identification)은 등록된 화자집단에 지금 요청중인 화자의 발성이 등록되어 있는지를 결정하는 과정이다. 둘째로 화자확인(Speaker Verification)은 지금 발성중인 화자가 인식시스템이 요청한 그 사람인지 아닌지(Yes-no task)를 결정하는 과정이다.

또한 화자인식은 인식 방법에 따라 4가지로 구분할 수 있다. 첫째로 패턴정합법(Pattern Matching)에 의한 동적 정합(Dynamic Time Warping)은 입력패턴을 미리 정해진 기준 패턴과 비교하여 최적화된 유사성을 판단하는 방법이다. 둘째로 신경회로망을 이용한 방법은 각 화자별로 신경회로망을 구성하고 화자간의 변별력을 갖도록 학습을 수행하는 인식 방법이다. 그러나 이 방법은 새로운 화자의 추가시 인식 시스템을 다시 학습시켜야 하고 고도의 병렬계산 능력이 요구되기 때문에 실제 응용시에는 적합하지 않다는 단점이 있다. 세 번째 방법인 백터양자화 방법은 입력 패턴과 양자화 코드북(Codebook) 사이의 거리로 유사성을 판단하는 방법이지만 많은 학습자료가 필요하고 화자간의 동적인 변화 특성을 이용하지 못하기 때문에 인식률에 한계가 있다.

마지막으로 은닉마코프모델(Hidden Markov Model-HMM)은 학습기능을 이용하여 화자내의 변이를 흡수 할 수 있으며, 입력패턴의 비선형 정합을 수행하는 특성이 있다. 화자인식 시스템은 인식에 사용하는 문장의 종속여부에 따라 정해지지 않는 어휘로 인식을 수행하는 텍스트 독립형(Text Independant)과 정해진 어휘만을 발성해야 하는 텍스트 종속형(Text Dependant)으로 나눌 수 있다.

3.2 화자 인식 과정

일반적으로 패턴매칭을 이용한 화자 인식 과정은 다음과 같다. 먼저 발성된 음성신호로부터 음성구간을 검출한다. 검출된 음성신호를 창함수를 이용하여 단구간으로 나눈다. 이렇게 단구간으로 나누어진 음성 데이터에서 화자의 특징벡터를 추출하여 기준패턴으로 사용한다.



3-1. 일반적인 화자인식 과정

이러한 방법으로 저장된 기준패턴들과 음성입력단에서 들어온 비교패턴을 DTW 방법을 이용하여 화자인식을 수행한다.[2][3]

4. 제안한 알고리즘

본 논문에서는 프리엠퍼시스 필터계수를 성문특성에 따라 적응적으로 구하기 위하여 음성을 발성할 때 크기의 차이를 두어 스펙트럼을 비교하였다. 필터계수를 구하는 방법은 다음과 같다.[2][5] 단구간 자기상관 함수는 (식4.1)로 표현 가능하다.

$$\Phi_n(i,j)=\sum_{m=0}^{N-1-(i-j)} s_n(m)s_n(m+i-j),\ 1\le i\le p,\ 0\le j\le p \quad (식4.1)$$

여기서 $R_n(j)=\sum_{m=0}^{N-1-j} s_n(m)s_n(m+j)$ (식4.2)

$$\sum_{j=1}^{p} a_j\Phi_n(i,j)=\Phi_n(i,0),\ \text{for } i=1,\ldots,p \quad (식4.3)$$

자기상관법(Auto-correlation Method)을 이용하여 (식4.3) 를 풀면 다음과 같이 표현된다.

$$\begin{vmatrix} R_n(0) & R_n(1) & \cdot & R_n(p-1) \\ R_n(1) & \cdot & \cdot & R_n(p-2) \\ \cdot & \cdot & \cdot & \cdot \\ R_n(p-1) & \cdot & \cdot & R_n(0) \end{vmatrix} \begin{vmatrix} a_1 \\ a_2 \\ \cdot \\ a_p \end{vmatrix} = \begin{vmatrix} R_n(1) \\ R_n(2) \\ \cdot \\ R_n(p) \end{vmatrix} \quad (식4.4)$$

p=1에 대하여 위의 식을 정리하면 다음과 같은 식으로 표현 가능하다.

$$a_1=\frac{R_n(1)}{R_n(0)} \quad (식4.5)$$

그림 4-1, 그림 4-2는 음성신호와 자기상관법을 이용하여 측정한 음성신호의 기울기를 나타낸 것이다.

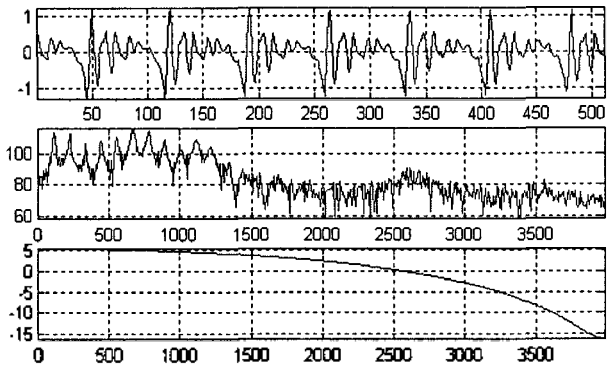


그림 4-1. 유성음 (a) 음성신호,
(b) Glottal Spectrum, (c) 측정된 기울기

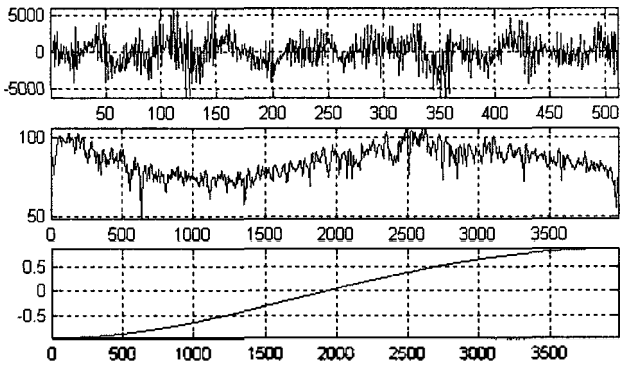


그림 4-2. 무성음 (a) 음성신호,
(b)Glottal Spectrum, (c) 측정된 기울기

위의 특성은 유성음과 무성음간의 평탄화를 위한 필터 계수의 기울기를 보인 것이고 이러한 특성의 차이를 음성발성의 크기에 따라 어떻게 변하는가를 실험하고자 한다.

5. 실험 및 결과

본 논문의 모의실험을 하기 위해 IBM PC에 마이크가 장치된 16비트 A/D변환기를 인터페이스 시켰다. 실험은 일반 실험실 환경에서 20대 남녀각각 4명과 중년의 남녀 각각 4명이 유성음을 약 1초간 발성하였다. 음성 시료를 8khz로 샘플링하고 16비트로 양자화하여 사용하였다. 한 프레임의 길이는 512샘플이다.

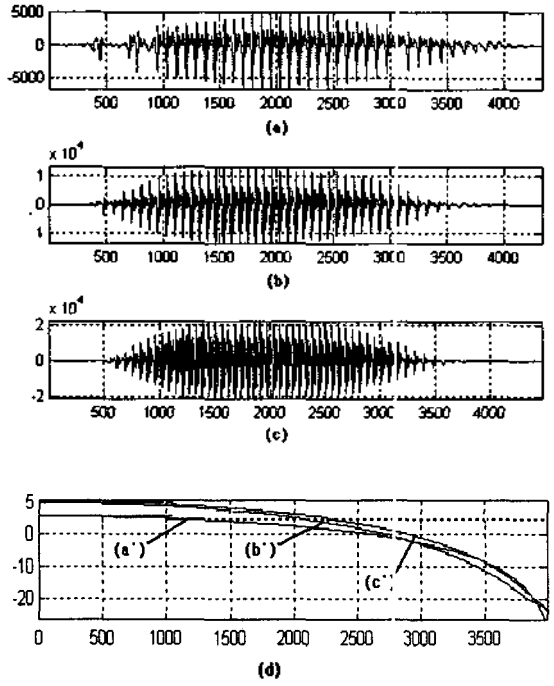


그림 5-1. 20대 남성의 발성크기에 따른 기울기 변화도
(a')는 (a)에 대한 평균기울기
(b')는 (b)에 대한 평균기울기
(c')는 (c)에 대한 평균기울기
(d)는 기울기 특성 비교

1초의 음성을 8프레임으로 나누어 구간별로 계수 a를 구하여 다시 그 값의 평균값을 취하였다. 아래의 표는 발성자에 따른 프레임별 계수 a의 평균값을 나타낸다.

표 5-1. 프레임별 계수 a의 평균값

발성크기 / 발성자	작게 발성	보통으로 발성	크게 발성
20대 남성	0.9327	0.8855	0.8704
20대 여성	0.8643	0.7892	0.6747
중년 남성	0.8775	0.8615	0.8418
중년 여성	0.8433	0.8207	0.8005

6. 결론

개인이나 특정 단체의 정보의 보완을 위해서는 사용자의 확인 과정이 필요하다. 이때 확인 절차는 사용자에게 사용이 용이해야 하며 확인 내용은 정확해야 한다. 이러한 점을 고려하여 근래에 들어 사용자의 음성특성을 이용한 사용자 확인 방법이 고안되었다.

본 논문에서는 프리엠퍼시스 필터 계수를 적응적으로 구하기 위하여 성문 특성을 발성크기에 따라 실험하였고 그 결과 발성크기가 커짐에 따라서 그 기울기의 작아짐을 알 수 있었다. 또한 표에서 보는 바와 같이 20대 발성자의 a값의 변화가 중년 발성자의 a값 변화보다 크고 여성이 남성보다 작은 것으로 나타났다.

7. 참 고 문 헌

[1] L. R. Rabiner & Biing-Hwang Juang, *Fundamentals Of Speech Recognition*, Prentice-Hall AT&T, U.S.A, 1993

[2] L. R. Rabiner & R.W.Schafer, *Digital Processing of Speech Signal*, Prentice-Hall, Englewood Cliffs, N.J., U.S.A., 1978

[3] A.M. Kondoz, *Digital Speech*, Jhon wiley & Sons, 1994

[4] Hiroaki Sakoe & Seibi Chiba, *"Dynamic Programming Algorithm Optimization for Spoken Word Recognition"*, IEEE Trans. on ASSP, vol.26, No.1, pp.43-49, Feb.1978.

[5] L R. Rabiner, R.W Schafer, " Digital Processing of Speech Signal", Prentice Hall, 1978.

[6] 배명진,"디지탈 음성분석", 동영출판사, 1998. 4.

[7] Oppenheim, Schafer,"Discrete Time Signal Processing", Prentice Hall, 1989.

[8] Emanuel C. Ifeachor,"Discrete Time Signal Processing", Addison Wesley,1993.

[9] 오영환, "음성언어정보처리",홍릉과학출판사,1998.

[10] Douglas O, shaughnessy,"Speech Communication",IEEE Press,1996.

[11] A. M. Kondoz,"Digital Speech", John Wiley & Sons Ltd, 1994.

[12] 배명진,"디지탈 음성합성", 동영출판사, 1998. 2.

[13] 민소연, 강은영, 배명진, "성문특성이 제거된 성도특성에 관한 연구", 대한전자공학회, 추계 종합학술대회, 2000년 11월 25일.