정보학연구 제3권 제4호 2000년 12월(pp.1-12)
The Journal of Information Technology, Vol.3, No.4

음성특성 학습 모델을 이용한 음성인식 시스템의 성능 향상

송점동*

요 약

음성은 특성에 따라 고음성분이 강한 음성과 저음성분이 강한 음성으로 구분할 수 있다. 그러나 이제까지 음성인식의 연구에 있어서는 이러한 특성을 고려하지 않고, 인식기를 구성함으로써 상대적으로 낮은 인식률과 인식모델을 구성할 때 많은 데이터를 필요로 하고 있다. 본 논문에서는 화자의 이러한 특성을 포만트 주파수를 이용하여 구분할 수 있는 방법을 제안하고, 화자음성의 고음과 저음특성을 반영하여 인식모델을 구성한 후 인식하는 방법을 제안한다. 한국어에서 가능한 47개의 모노폰을 이용하여 인식모델을 구성하였으며, 여성과 남성 각각 20명의 음성을 이용하여 인식모델을 학습시켰다. 포만트 주파수를 추출하여 구성한 포만트 주파수 테이블과 피치정보값을 이용하여 음성의 특성을 구분한 후, 음성특성에 따라 학습된 인식모델을 이용하여 인식을 수행하였다. 본 논문에서 제안한 시스템을 이용하여 실험한 결과 기존의 방법보다 인식률이 향상됨을 보였다.

Ⅰ. 서론

인간은 여러 가지 방법으로 의사를 교환하는데 그 중에서 가장 편리하게 사용하는 것이 음성을 이용한 대화이다. 음성의 편리함 때문에 기계가 음성을 인식할 수 있다면 인간과 기계의 인터페이스로 이용하여 여러 사용자에게 편리함을 줄 수 있다. 이러한 이유로 음성인식은 과거 수십 년 동안 많은 연구가 있었고 현재에도 많은 연구가 계속되고 있다[11]. 인간의 발성하는 음성에는 많은 정보가 포함되어 있다. 관습적으로 또는 문화적으로 이미 약속된 언어를 통해서 발화된 음성은 우리가 학습된 지식에 의해서 이해를 하게 되는데 우리는 듣는 음성에는 화자가 말을 하는 내용 즉 의미에 관한 정보를 가지고 있으며 음성의 특성에 따라 화자의 특성을 짐작할 수 있는 정보도 가지고 있다. 즉 화자의 음성이 고음성분을 가지고 있는지 저음 성분을 가지고 있는지를 알 수 있고 이것으로 화자의 성별을 어느 정도 구분할 수 있다[6]. 음성은 이와 같이 화자에 따라 고음특성과 저음특성으로 구분할 수 있음에도 불구하고 이제까지의 음성인식의 연구에 있어서는 이러한 것을 고려하지 않았다.

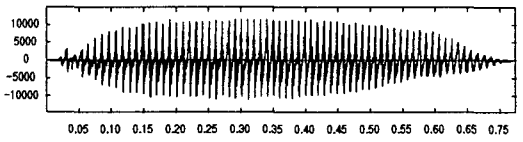
음성인식의 여러 가지 방법 중에서 가장 많이 사용하는 방법 중에 하나인 HMM(Hidden Markov Model)을 이용한 방법에서도 화자의 특성에 따라 구분할 수 있음에도 불구하고 이를 구별하지 않고 하나의 인식모델로 구성함으로써 이에 따라 상대적으로 낮은 인식률과 학습과정에 많은 데이터를 필요로 하고 있다. 본 논문에서는 화자의 고음과, 저음특성을 구분할 수 있는 방법으로 포만트 주파수와 피치정보를 이용하는 방법을 제안한다.

* 경문대학 전산정보과 교수

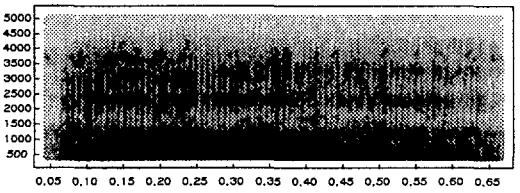
포만트 주파수란 모음의 주파수 중에서 에너지가 집중적으로 나타나는 영역 말하며 모음의 주파수 성분에는 3~4개의 포만트가 있다[3]. 고음성분음성과 저음성분음성의 차이는 주파수 대역에서 나타나는데 일반적으로 남성음성은 같은 소리를 내더라도 저음성분이 강하고 여성음성은 고음성분이 강하다. 이와 같은 특성은 모음의 포만트 주파수와 피치 값의 차이로 나타난다. 이러한 정보 값으로 음성의 특성을 구분하여 HMM을 구성할 때 음성의 고음과 저음에 따라 구분하여 학습한 후 인식과정에서 음성특성에 따라 고음의 특성이 강한 음성은 고음성분HMM과 저음의 특성이 강한 음성은 저음성분HMM과 인식을 수행하는 방법을 적용함으로써 적은 양의 학습데이타로 높은 인식률을 얻을 수 있었다. 본 논문에서는 포만트 주파수를 구하기 위해서 LPC로부터 Root solving과 Peak picking 방법을 이용했다[1][2][3]. 단모음에 대한 포만트 주파수 테이블을 구성하고 포만트 주파수 테이블과 자기상관계수로 구한 피치정보를 이용해서 음성의 특성을 구별하는 고음저음 결정 알고리즘을 적용한다. 이렇게 구별된 특성에 따라 47개의 모노폰으로 구성된 인식모델과 인식을 수행한다.

Ⅱ. 포만트 주파수의 특성

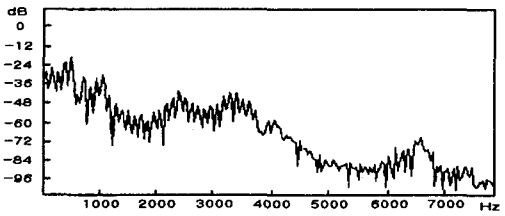
자음과 모음은 각각 무성음과 유성음에 해당되는데 음성의 무성음과 유성음 중에서 유성음은 성대의 진동에 의한 기본주파수성분을 가지고 있다[10][12]. 이 기본주파수가 성도의 공명강을 거치면서 음성의 통과시점에서의 공명강의 형태에 따라서 특정주파수 성분이 강조되는 대역이 나타나게 된다. 이와 같이 모음성분 중에서 강조되는 특정주파수를 포만트 주파수라고 부른다[4][5]. 보통 모음에서 5kHz이하에 3~4개의 포만트주파수를 가지고 있고 주파수가 낮은 쪽으로부터 제 1 포만트(F1), 제 2 포만트(F2), 제 3 포만트(F3)라고 부른다[8]. 포만트 주파수는 입의 모양에 따라 다르게 나타난다. 그러므로 포만트 주파수를 이용하면 모음의 종류를 구별할 수 있다. 그림4~6에서 모음 '아'에 대한 waveform, 스펙트럼, FFT변환 형태를 보여주고 있는데 여기에서 특정 주파수대역이 강조되는 포만트의 특성을 확인할 수 있다. 포만트 주파수는 모음의 종류를 결정하는 한편 음성의 또 다른 특성을 반영하고 있는데, F1, F2, F3의 전체적인 위치 이동에 의해서 음성의 높고 낮음이



〔그림 1〕 '아'에 대한 waveform



〔그림 2〕 '아'에 대한 스펙트럼



〔그림 3〕 '아'에 대한 FFT 변환

결정된다. 보통 여성음성이 고음에 가깝고 남성음성이 저음에 가깝다. 남성과 여성의 음성이 차이가 보이는 원인은 남성은 사춘기를 거치면서 변성(變聲)과정을 거치기 때문이다. 변성의 과정은 보통 12, 13세경에 나타나며 남성의 경우 성대가 있는 부분이 후두부(後頭部)가 급격히 발육하여 돌출하게 되어 성대의 길이가 길어지게 된다[10]. 이와 같은 성대 길이의 차이에 의해서 같은 음성을 발음하더라도 음성의 높낮이가 달라진다. 남성의 경우 여성보다 성대의 길이가 길기 때문에 기본진동수가 낮아지고 그래서, 저음이 강한 음성을 발성하게 되고, 여성은 이와 반대로 고음이 강한 음성을 발성한다. 이것은 포만트 주파수의 차이로 나타나게 된다.

〈표 1〉 성대의 길이 변화

	여 성	남 성
6세경의 어린이	9~10mm	9~10mm
사춘기 전	9~101/2mm	9~101/2mm
사춘기 종류후	12~15mm	14~21mm
성 인	12~16mm	13~24mm

포만트주파수를 구하는 방법은 역필터 $A(z)$의 근을 계산하는 Root solving 방법과 FFT로 역필터의 보간된 스펙트럼을 계산하고 $1/|A(e^{j\theta})|^2$의 스펙트럼의 peak값을 찾는 Peak picking 방법이 있다[3][7]. Root solving 방법은 모든 포만트의 값을 얻을 수 있으나 계산량이 많고 Peak picking은 계산량은 적지만 두 개의 포만트가 인접해 있는 경우 잘못된 위치를 지정하게 되는 단점이 있다[8]. Root solving에서 방법은 복소근 z에 대한 대역폭 $\hat{B}$와 주파수 $\hat{F}$는 s-평면에서 z-평면으로의 변환에 의해 얻어진다.

$$z=e^{sT} \qquad (2\text{-}1)$$

여기서 $s=-\pi\hat{B}\pm j2\pi\hat{F}$이고

$z=R_e(z)\pm jI_m(z)$는 복소근의 실수부와 허수부로 정의된다.

$$\hat{F}=(f_s/2\pi)\tan^{-1}[I_m(z)/R_e(z)]\ (Hz)$$

$$\hat{B}=-(f_s/\pi)\ln|z|\ (Hz) \qquad (2\text{-}2)$$

Peak picking방법은 N-point FFT를 사용하여 이산적 개수의 샘플에서 Peak picking과 포물선 보간을 적용하는 것이다. 포물선의 형태는 다음과 같다.

$$y(\lambda)=a\lambda^2+b\lambda+c \qquad (2\text{-}3)$$

만일 $y(0)$이 이산적 peak 값을 정의하고 $y(-1)$이 왼쪽 $y(1)$이 오른쪽 샘플 값이 된다면 이 세 점을 지나는 포물선의 계수는 다음과 같다.

$$c=y(0) \qquad (2\text{-}4)$$

$$b=[y(1)-y(-1)]/2 \qquad (2\text{-}5)$$

$$a=[y(1)+y(-1)]/2-y(0) \qquad (2\text{-}6)$$

$dy(\lambda)/d\lambda=0$의 미분계산에 의해 peak 위치 즉 극대 값은 $\lambda_p=-b/2a$가 된다. 이 위치가 포만트의 위치이다. 만일 이산적 peak가 n_p에 위치한다면 보간된 포만트와 대역폭의 값은 다음과 같이 나타낼 수 있다.

$$\hat{F}=(n_p+\lambda_p)f_s/(2N) \qquad (2\text{-}7)$$

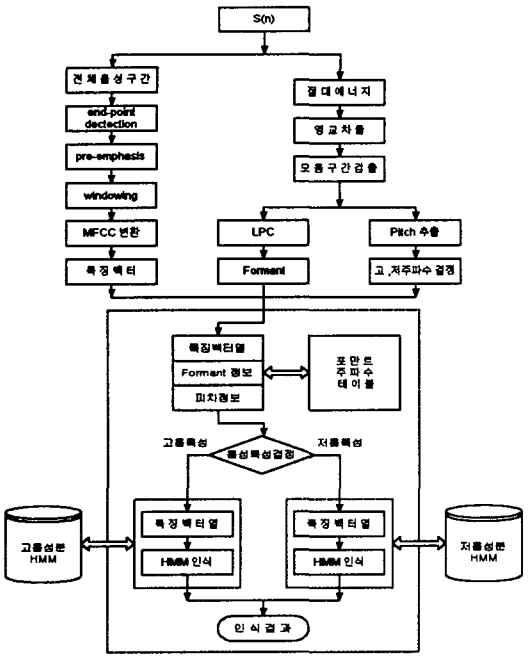
$$\hat{B}=\frac{-\{b^2-4a[c-0.5y(\lambda_p)]\}^{1/2}f_s}{aN} \qquad (2\text{-}8)$$

Ⅲ. 전체시스템의 설계

이제까지 음성의 고음과 저음 특성을 구별할 수 있는 포만트 주파수에 대해서 알아보았다. 이러한 내용을 적용하여, 본 논문에서 설계하고 구현한 시스템 구성도는 그림 4와 같으며 세부

단계의 작업과정은 다음과 같다.

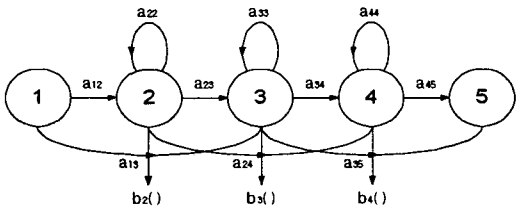
제안된 전체 시스템은 크게 두 단계로 나뉘어지는데 입력 음성에 대하여 인식을 하기 위한 처리과정과 음성특성을 분석하기 위한 과정으로 구분된다. 인식을 위한 과정에서는 전체음성 구간을 MFCC변환을 하여 특징 벡터열을 추출한다. 이와 동시에 음성특성 분석을 위한 과정에서는 입력음성에 대하여 절대에너지와 영교차율을 이용해서 모음구간을 결정하고 다시 두 과정으로 분리되는데 모음구간에 대한 포만트 주파수를 구하는 과정과 피치정보를 추출하는 과정으로 나뉜다. 여기서 구한 포만트 주파수 값과 피치정보는 인식을 위한 특징벡터열에에 의해서 음소가 변하는 조음화 현상을 모델링할 수 있지만 인식대상의 수가 많아져서 훈련 데이타의 확보가 어렵고 인식모델의 크기가 커지는 단점이 있다. 트라이폰의 경우 36,584개 다이폰의 경우 514개의 경우의 수가 나온다[13].



〔그림 4〕 전체 시스템의 설계

본 연구에서는 인식모델의 수가 적고 확장성도 우수한 모노폰을 이용한 인식단위를 이용하였다. 모노폰의 인식단위를 이용할 경우 47개의 경우의 수가 나온다.

표 2에서 한국어의 가능한 47개의 모노폰과 이에 대한 PLU를 정의하였는데, 모노폰 중에서 발음상 유사한 'ㅚ,ㅙ,ㅞ', 'ㅐ,ㅖ', 'ㅐ,ㅔ'는 각각 같은 PLU를 적용한다.



〔그림 5〕 Simple Left-Right HMM의 구조

각 모노폰에 대한 모델은 그림 5와 같이 5개의 상태를 갖는 Simple Left-Right HMM으로 구성하였고, 인식을 하기 위한 음성의 특징벡터로 MFCC(Mel-Scale Frequency Cepstrum Coefficient)를 추가된다. 인식모델과 비교를 하기 위해 처리된 특징 벡터열은 포만트 주파수 값과 피치 정보에 의해서 입력된 음성의 특성이 고음인지 저음인지 결정된 후 각각 분리해서 학습시킨 HMM과 인식을 수행한다.

3.1 인식모델의 구성

인식모델을 구성하는데 있어서 인식단위의 결정이 선행되어야 한다. 음성인식을 하기 위한 인식 단위는 단어모델로 구성하는 방법과 모노폰, 다이폰, 트라이폰 단위로 구성하는 방법이 있다

〔표 2〕 모노폰의 구성

	phoneme	PLU		phoneme	PLU
자음 (19)	ㄱ	k		ㅂ	b0
	ㄲ	kk		ㅇ	ng
	ㄴ	n	단모음 (8)	ㅏ	a
	ㄷ	d		ㅓ	au
	ㄸ	dd		ㅐ	e
	ㄹ	r		ㅔ	e
	ㅁ	m		ㅗ	o
	ㅂ	b		ㅜ	u
	ㅃ	bb		ㅡ	eu
	ㅅ	s		ㅣ	i
	ㅆ	ss	이중 모음 (13)	ㅚ	oi
	ㅇ	-		ㅟ	ui
	ㅈ	j		ㅑ	aa
	ㅉ	jj		ㅕ	auu
	ㅊ	c		ㅒ	ye
	ㅋ	q		ㅖ	ye
	ㅌ	t		ㅘ	oa
	ㅍ	p		ㅙ	oi
	ㅎ	h		ㅛ	oo
대표 발침 소리 (7)	ㄱ	k0		ㅝ	uau
	ㄴ	n0		ㅞ	oi
	ㄷ	d0		ㅠ	uu
	ㄹ	r0		ㅢ	ei
	ㅁ	m0			

단어모델인 경우는 초기 음성인식기에서 이용하는 방법으로 음성DB를 구성하기는 쉽지만 인식모델의 확장성이 나쁘다. 모노폰은 음소각각을 기준으로 모델을 구성하고 다이폰이나 트라이폰은 음소의 앞뒤의 연관성을 고려하여 모델링을 한다. 음성의 특징 벡터인 MFCC는 사람이 주관적으로 인지하는 주파수특성을 반영하는 mel-scale로 신축한 critical band 필터뱅크를 씌워 추출하는 것으로 추출과정은 다음과 같은 과정으로 수행하였다. 먼저 8kHz로 샘플링된 음성신호를 전달함수가 $1-0.97z^{-1}$인 1차 디지털 필터로 선강조를 한 후 길이가 20msec이고 10msec 씩 중첩되는 프레임 단위로 나눈 다음, 각각의 프레임에 해밍창을 씌운다. 매 프레임을 고속퓨리에 변환(FFT)을 하여 주파수 영역에서 파워 스펙트럼을 구하고, 그 스펙트럼에 mel-scaled 삼각 필터링을 수행하여 26개의 필터 뱅크 출력을 얻는다. 그리고 각각의 필터 뱅크 출력의 log값을 Discrete Cosine Transform(DCT)하여 12차의 MFCC를 구한다.

$$MFCC_k = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} \{\log(x[i]) \cos\left(\frac{2\pi k}{N}(i-0.5)\right)\} \quad (3\text{-}1)$$

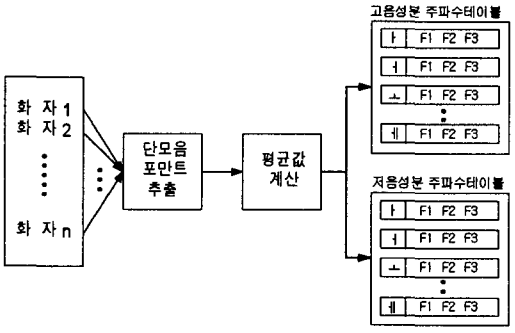
$x[i]$: filter bank출력 값

N : filter bank의 개수

k : cepstrum 차수의 index

MFCC를 표 2에서 정의한 PLU로 HMM을 구성하는데 그림 6에서와 같이 수집한 음성을 남성음성과 여성음성으로 나누어서 HMM을 구성한다.

포만트 주파수 테이블을 작성하기 위해 변성과정을 거친 20대 이상의 화자 중에서 저음특성을 가지고 있는 남성과 고음특성을 가지고 있는 여성 각각 20명이 단모음에 대해 5번씩 발음한 음성을 이용해서 포만트 주파수 테이블을 작성한다.



〔그림 6〕 포만트 주파수 테이블 작성과정

각 화자의 단모음에 대한 입력을 받아 다음과 같은 식을 이용해서 포만트 주파수 값의 평균을

낸다.

$$(F_{m,1}+F_{m,2}+\cdots\cdots+F_{m,n})/n=F_m' \quad (3\text{-}2)$$

여기서 m(m=1,2,3)은 제 1 포만트부터 제 3 포만트까지의 값이고 n(n=1,2,···10)은 전체 화자의 수가된다.

3.2 입력음성에 대한 처리과정

앞에서 구성한 인식모델과 포만트 주파수 테이블은 인식처리를 하는 과정에서 이용하는데, 기존의 인식모델과의 차이점은 같은 단어에 대해서 화자독립을 실현하기 위해서 다양한 사람에 대한 음성을 수집하여 하나의 HMM으로 학습시키는 과정에서 발생하는 오차를 줄이기 위해서 화자특성별로 인식대상을 구성하였다는 데에 있다. 인식을 하기 위해서 불특정 화자가 발화한 음성에 대해 인식처리를 위한 과정은 그림 4의 전체시스템에서와 같이 화자가 발화한 음성을 이용하여 인식을 하기 위한 전체 음성구간의 전처리 단계와 화자의 특성을 추출하기 위해서 피치와 포만트 주파수를 구하는 두 단계로 구분되어 처리된다. 인식을 위한 특징 추출과정의 첫 번째 단계는 전체음성구간에서 정확한 음성구간을 결정하기 위한 과정이다. 음성인식에서 발화한 음성 중 정확한 음성구간을 판단하는 것은 인식률에 큰 영향을 미치는데, 잡음을 고려하지 않는 환경에서는 음성구간검출이 쉽지만, 주변잡음의 영향이 클 때는 힘들어 진다. 보통 음성구간검출 즉 end-point detection에는 절대에너지와 영교차율을 사용한다. 절대에너지는 음성의 진폭의 크기에 비례하는 값으로 다음 수식으로 구한다.

$$E = \sum_{i=0}^{N}|X(i)| \quad (3\text{-}3)$$

이 값이 에너지 상한 값을 넘으면 음성이 있다고 간주한다. 그러나 무성음의 경우 에너지가 작으므로 다음의 영교차율을 이용한다.

$$ZCR = \sum_{i=0}^{N-1}|sgn(x(i)sgn(x(i+1))|*(1/2) \quad (3\text{-}4)$$

$$sgn(x)=1\ (x\geq 0) \quad sgn(x)=-1\ (x<0) \quad (3\text{-}5)$$

본 시스템에서는 절대에너지와 영교차율을 같이 이용해서 음성구간을 검출한다. 음성신호의 주파수 스펙트럼은 일정하지 않고, 주파수 값이 높을수록 그 성분이 작아지게 되어 주파수가 2배가 되면 약 6(dB)의 기울기로 그 파워의 진폭특성이 작아진다. 그러므로 음성신호 분석 전에 6(dB)기울기를 갖는 고역강조 필터를 통과시켜 음성신호의 스펙트럼이 저역부터 고역까지 같은 S/N비를 갖게 하는 과정이 pre-emphasis이며, 전달함수는 다음과 같다.

$$H(Z) = 1-aZ^{-1} \quad (a: 0.9\ or\ 0.875) \quad (3\text{-}6)$$

다음과정으로 음성을 프레임 단위로 나눌 때 양끝 단에서는 날카로운 잡음성분을 방지하기 위해서 창함수를 씌운다. 창함수의 종류에는 사각창(Rectangular window), 바틀레트창(Bartlet window), 해밍창(Hamming window), 블래크만창(Black window)가 있는데 본 연구에서는 해밍창을 이용한다. 해밍창은 분석구간의 가장자리로 갈수록 신호자료의 크기를 점차 줄여나감으로서 양끝 부분의 날카로운 잡음 성분의 발생을 방지시켜 준다[9].

$$W(i) = 0.54-0.46*\cos(2\pi i*i/N)$$
$$0\leq i\leq N \quad (3\text{-}7)$$

또한 창함수를 씌울 때 가장자리 부분의 신호가 손실되는 것을 방지하기 위해서 보통 창함수의 길이의 1/4의 구간을 중복시켜 추출한 데이

터에서 MFCC를 구해서 인식을 위한 특징벡터로 이용한다. 음성특성 구성을 위한 단계에서는 앞에서 음성구간을 결정할 때 이용했던 절대에너지와 영교차율을 가지고 모음구간을 검출하는데 이용한다. 일반적으로 모음구간은 자음구간보다 에너지 값이 크고 영교차율이 낮기 때문에 실험을 통한 한계 값을 정해서 모음구간을 검출에 이용한다. 모음구간이 검출되면 포만트 주파수 추출을 위한 과정과 피치추출을 위한 과정을 거친다. 포만트 주파수 추출법은 1장에서 언급한 것과 같이 Root solving과 Peak picking 방법을 이용해서 추출한다. 이와 동시에 피치추출 과정을 수행하는데 음성신호에서 피치를 이용하면 같은 파형이 반복되는 주기를 결정할 수 있다. 피치의 값이 작으면 높은 소리가 되고 피치의 값이 크면 주기가 기본주기가 길어져서 낮은 소리가 된다. 피치값을 구하기 위한 방법으로 다음과 같은 자기상관계수법을 적용한다. 자기상관계수법은 신호에서 어느 시점과 일정간격에 있는 신호와의 유사도를 구해서 같은 파형이 반복되는 위치를 계산하는 방법으로 피치를 구하는 방법이다. 표본단위로 k만큼 떨어져 있는 신호표본 끼리의 유사도를 나타내는 식은 다음과 같다[9].

$$R(k) = \sum_{n=-\infty}^{\infty} x(n)x(n+k) \qquad (3\text{-}8)$$

$R(k)$는 어느 한시점 n에서 표본 $x(n)$의 값과 그로부터 k 만큼 떨어져 있는 표본의 값을 서로 곱한 것을 모든 n에 대해서 합한 값이다. $x(n)$과 $x(n+k)$가 서로 비슷한 위상의 값을 가진다면 $R(k)$의 값은 최대가 된다. 최대값을 갖는 위치에서 다시 최대 값을 갖는 위치를 찾으면 피치 값을 구할 수 있다. 인식을 위해 입력된 음성의 특징 벡터열과 피치정보 포만트 주파수 값을 구했다.

고음과 저음의 음성특성을 결정하는 방법은 알고리즘 1과 같다. 입력된 음성에서 모음구간의 각 프레임에서 구한 포만트 값의 합을 계산한 후 전체 프레임수(n)로 나누면 입력음성중 모음구간의 평균 포만트 값 (F_1, F_2, F_3)이 구해진다. 이 값과 미리 작성된 포만트 주파수 테이블을 비교하기 위해서, 입력음성의 포만트 값과 테이블 값의 차를 계산한 후 가장 작은 값을 가지는 테이블의 번호를 선택한다. 포만트 주파수 테이블은 0~7번까지 고음성분음성, 8~15까지는 저음성분음성의 포만트 주파수로 구성하였다.

피치정보는 실험과정 수집한 음성에서 구한 실험 값에 의해서 피치가 나뉘어지는 평균값을 결정하여 기준 값으로 이용한다. 음성의 특성이 결정되면 그 결과에 의해서 고음성분으로 학습된 HMM과 저음성분으로 학습된 HMM으로 각각 구분하여 인식을 수행함으로써 인식률을 높일 수 있다.

[알고리즘 1] 음성특성결정 알고리즘

```
Begin
  while(zero < zero_threshold){
     while(energy > energy_threshold){
       Formant();
       Form_sum[i] += Form_temp[n];
       n++;
       Picth();
      }
  }
  Form_res[k] = Form_sum[i]/n
  while(Form_table[m]-Form_res[k]<Threshold)
  m++;
  result m; // Formant table number
  If(m<8 && pitch_result < pitch_threshold)
   then
       Speech_recog_Low_model();
   else
       Speech_recog_High_model();
End
```

Ⅳ. 실험결과

음성데이터 수집을 위해서 펜티엄 300MHz 휴대용 컴퓨터를 이용하여 수집하였다. 샘플링 비율은 16000 bit, 양자화 해상도는 16bit로 하였고, 266MHz 펜티엄 pc와 Ultra-10 sparcⅡ 333MHz을 이용하여 인식을 수행하였다. 포만트 주파수 추출을 위해서 20대 이상으로 구성된 남성20명과 여성20이 단모음에 대해서 5번씩 발화한 음성을 이용했다. 음성인식 대상으로는 임의로 50명의 이름을 작성하여 인식대상으로 사용하였다.

실험은 먼저 포만트 주파수 테이블 구성을 위해서 남녀 40명이 8개의 단모음 "ㅏ,ㅓ,ㅗ,ㅜ,ㅡ,ㅣ,ㅐ,ㅔ"를 조용한 실험실 환경에서 정확한 구강 구조를 유지하면서 3초 동안 발음하게 하였고, 각 모음별로 5번씩 발음하게 하였다. 1600개의 단모음에 대해서 정확한 모음구간만을 샘플링 하여 포만트 주파수 추출 알고리즘을 적용하여 주파수를 추출하였다. 포만트 주파수 테이블 작성을 위해서 추출한 포만트 주파수의 값은 표 3과 같다. 위에서와 같이 구한 포만트 값을 이용해서 포만트 주파수 테이블을 표 4와 같이 구하였다. 이 포만트 주파수 테이블은 음성특성 결정 시에 이용하게 된다.

〔표 3〕 포만트 주파수

모음	프레임	저음특성화자1(Hz)			고음특성화자1(Hz)		
		F1	F2	F3	F1	F2	F3
ㅜ	1	321.05	715.22	2313.15	457.48	823.92	2408.72
	2	326.18	716.13	2332.07	459.12	821.72	2406.41
	3	321.85	709.77	2307.08	449.55	820.59	2396.24
	4	318.36	710.58	2399.76	451.55	818.99	2410.77
	5	319.17	708.00	2347.47	453.27	823.45	2408.12
ㅡ	1	363.21	1150.42	2403.30	491.51	1399.17	2558.79
	2	356.23	1141.58	2389.21	491.86	1404.51	2557.47
	3	370.70	1164.63	2406.80	492.92	1397.02	2553.65
	4	365.71	1207.88	2439.75	493.17	1399.83	2544.95
	5	363.43	1128.34	2411.62	488.81	1388.45	2557.39
ㅣ	1	270.84	2273.16	3069.54	393.89	2744.51	3275.70
	2	277.34	2299.60	3117.14	391.50	2748.93	3265.23
	3	268.71	2181.96	3091.24	397.78	2753.78	3274.00
	4	281.73	2129.71	3065.56	395.06	2757.94	3278.41
	5	276.74	2213.99	3091.29	390.58	2740.47	3281.97
ㅐ	1	415.62	1613.72	2510.66	484.01	1877.06	2832.88
	2	413.28	1605.27	2496.76	483.78	1862.27	2334.39
	3	419.12	1612.23	2497.41	483.91	1879.57	2363.99
	4	412.11	1608.63	2514.41	485.45	1875.98	2832.26
	5	411.89	1625.42	2495.99	484.74	1876.21	2836.67
ㅔ	1	432.53	1773.70	2633.60	488.18	1921.61	2897.58
	2	433.39	1769.92	2618.56	490.30	1922.53	2890.57
	3	428.92	1779.11	2620.34	491.54	1921.52	2894.62
	4	432.30	1770.68	2619.95	486.96	1929.06	2891.07
	5	434.23	1763.86	2622.12	487.36	1926.47	2895.62

〔표 4〕 포만트 주파수 테이블

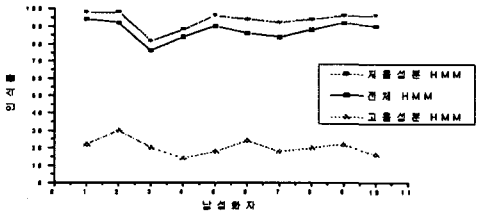
모음	저음특성화자(Hz)			고음특성화자(Hz)		
	F1	F2	F3	F1	F2	F3
ㅏ	784	1230	2777	944	1403	2947
ㅓ	432	835	2762	536	930	2868
ㅗ	459	781	2696	544	828	2787
ㅜ	322	712	2344	452	820	2406
ㅡ	360	1180	2408	492	1397	2553
ㅣ	274	2216	3063	393	2340	3273
ㅐ	413	1613	2500	505	1876	2735
ㅔ	431	1775	2622	515	1926	2893

포만트 주파수는 실험결과 중설 저모음'ㅏ'의 경우 제 1 포만트가 700~900Hz로 가장 높게 나타났고, 전설 고모음인 'ㅣ'의 제 1 포만트가 가장 낮은 결과를 보였다. 또한 'ㅓ'와 'ㅗ'를 비교해 볼 때 제 1 포만트는 'ㅗ' 가 높지만 제 2 포만트와 제 3 포만트는 'ㅓ'가 높은 수치를 보였다. 전체적으로 저음특성화자의 포만트와 고음특성화자의 포만트는 100~200Hz 정도의 차이를 보였다.

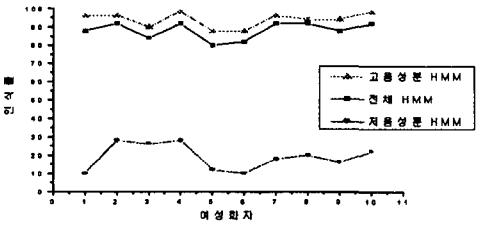
〔표 5〕 인식 결과

		인 식 결 과			
		전체 HMM	저음성분 HMM	고음성분 HMM	향상률
남성	화자1	94.0%	98.0%	22.0%	4.0%
	화자2	92.0%	98.0%	30.0%	4.0%
	화자3	76.0%	82.0%	20.0%	6.0%
	화자4	84.0%	88.0%	14.0%	4.0%
	화자5	90.0%	96.0%	18.0%	6.0%
	화자6	86.0%	94.0%	24.0%	8.0%
	화자7	84.0%	92.0%	18.0%	8.0%
	화자8	88.0%	94.0%	20.0%	6.0%
	화자9	92.0%	96.0%	22.0%	4.0%
	화자10	90.0%	96.0%	16.0%	6.0%
	평 균	87.6%	93.4%	19.4%	5.6%
여성	화자1	88.0%	10.0%	96.0%	8.0%
	화자2	92.0%	28.0%	96.0%	4.0%
	화자3	84.0%	26.0%	90.0%	6.0%
	화자4	92.0%	28.0%	98.0%	6.0%
	화자5	80.0%	12.0%	88.0%	8.0%
	화자6	82.0%	10.0%	88.0%	4.0%
	화자7	92.0%	18.0%	96.0%	4.0%
	화자8	92.0%	20.0%	94.0%	2.0%
	화자9	88.0%	16.0%	94.0%	6.0%
	화자10	92.0%	22.0%	98.0%	6.0%
	평 균	88.2%	19%	93.8%	5.4%

다음으로 표 5는 제안된 전체시스템을 이용하여 실험한 인식결과의 백분율을 보여 주고 있다. 실험은 음성의 저음특성을 주로 가지고 있는 남성화자와 고음특성을 가지고 있는 여성화자 각각 20명이 50개의 실험대상에 대해 발화한 음성을 이용하여 인식률을 계산하였다. 실험방법은 화자의 음성에 대해서 기존의 인식방법인 전체HMM으로 구성하였을 때 와 각각 화자의 특성에 따라 구분하여 HMM을 구성하였을 때를 실험하였다.



〔그림 7〕 남성 화자별 인식결과



〔그림 8〕 여성 화자별 인식결과

남성화자의 경우 저음특성을 가지고 있기 때문에 저음성분HMM 에 인식을 하였을 때 평균 5.6% 인식률 향상되었고 여성화자의 경우는 기존시스템 보다 5.4% 향상되었다. 또한, 저음특성

이 강한 남성화자의 음성을 고음성분HMM에 인식하였을 때 와 고음특성이 강한 여성화자의 음성을 저음성분HMM에 인식을 할 경우에는 평균 19.2%의 저조한 인식률을 보인다. 이제 까지 실험으로 화자의 특성을 구분하여 인식모델을 구성하면 인식률을 향상시킬 수 있다는 것이 확인되었다.

V. 결론

본 논문에서는 포만트 주파수를 이용하여 화자의 특성을 구분할 수 있는 방법을 제안하고, 화자음성의 고음과 저음특성을 반영하여 인식모델을 구성한 후 인식하는 방법을 제안하였다. 실험결과 남성음성이 저음성분이 강하고, 여성음성이 고음성분이 강하다는 것이 확인되었다. 이러한 특성은 제 1, 제 2, 제 3포만트 주파수로 구분할 수 있는데, 여성음성인 경우 남성음성보다 평균 100~200Hz 정도 높은 값을 보였다.

본 논문에서 제안한 시스템을 이용하여 실험한 결과 저음특성을 가지고 있는 남성화자의 경우 저음성분 HMM에 인식을 수행하였을 경우 인식률이 평균 5.6% 향상되었고, 고음특성을 가지고 있는 여성화자의 경우 고음성분HMM에 인식을 수행하였을 때 평균 5.4%향상된 결과를 보였다.

향후 연구 과제는 화자의 특성을 보다 세분하여 구분할 수 있는 방법이 필요하다. 본 논문에서는 주로 1차 변성기를 지난 남성과 여성음성을 저음특성과 고음특성으로 구분하였는데, 사람의 목소리는 성대의 퇴화로 인하여 나이별로 다른 특성을 보이고 있다. 이런 특성을 반영하여 성별과 연령별 특성을 구분할 수 있다면 보다 높은 인식률을 얻을 수 있고, 화자 인식 분야에도 응용할 수 있으리라 생각된다.

참고문헌

[1] A. V. Oppenheim, R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 1989.

[2] H. Wakita, "Direct Estimation of the vocal Track shape by Inverse Filtering of Acoustic Speech waveforms," IEEE Trans. A&E, vol.50, No2, pp. 637-655, Aug.,1971.

[3] J. D. Markel and A.H.Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, 1976.

[4] L. R. Rabiner, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood cliffs, N.J., 1993.

[5] Lutz Welling and Hermann Ney, "Formant Estimation for Speech Recognition," IEEE Trans. on Speech and Audio Processing, vol.6, pp.1063-1076, 1998.

[6] P. A. Busby and G. L. Plant, "Formant frequency values of vowels procuced by presdolescent boys and girls," J. Acoust. Soc. Am. 97 (4), April., 1995.

[7] Schafer, R.W. and Rabiner, L. R, "Speech for automatic formant analysis of voiced speech," J.Acoust. Soc., 1970.

[8] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans.

Acoust., Speech, Signal Processing, vol. ASSP-22, pp. 135-141, 1974.

[9] V. K. Madisetti, D. B. Williams, *The Digital Signal Processing handbook*, CRC Press, 1996.

[10] 문영일, *기초음성학과 발성기법*, 청우, 1987.

[11] 오영환, "음성언어 정보 처리 연구의 동향," 정보과학회지, 제16권,제2호, pp. 5-11, 1998.

[12] 이호영, *국어음성학*, 태학사, 1996.

[13] 김재범, *발화속도 측정 및 보상에 의한 한국어 연속음 인식기*, 인하대학교 공학석사 학위 논문, 1997.

Improvement of Speech Recognition System Using the Trained Model of Speech Feature

Jeom-Dong, Song*

Abstract

We can devide the speech into high frequency speech and low frequency speech according to the feature of the speech. However so far the construction of the recognizer without concerning this feature causes low recognition rate relatively and the needs of an amount of data in the research on the speech recognition. In this paper, we propose the method that can devide this feature of speaker's speech using the Formant frequency, and the method that can recognize the speech after constructing the recognizer model reflecting the feature of the high and low frequency of the speaker's speech. For the experiment, we constructed the recognizer model using 47 mono-phone of Korean and trained the recognizer model using 20 women's and men's speech respectively. We divided the feature of speech using the Formant frequency Table, that had been consisted of the Formant frequency, and the value of pitch, and then We performed recognition using the trained model according to the feature of speech. The proposed system outperformed the existing method in the recognition rate, as the result.

* Dept. of Computer Information, Kyungmoon College.