포만트 주파수를 이용한 음성인식 전처리 시스템의 설계 및 구현

김태욱*, 한승진*, 김민성**, 이정현*
**안산1대학 전자계산과
*인하대학교 전자계산공학과
*E-mail : kimtw@nlsun.inha.ac.kr

A Design and Implementation of Speech Recognition Preprocessing System using Formant Frequency

Tae-Wook Kim*, Seung-Jin Han*, Min-Sung Kim**, Jung-Hyun Lee*
Dept. of Computer Science, AnSan 1 College, AnSan, Korea
Dept. of Computer Science & Engineering, Inha University, Inchon, Korea

요 약

인간이 발성하는 음성에는 의미에 대한 정보 뿐만 아니라 화자의 성별에 따라 고유한 특성을 가지고 있다. 즉 음성은 고음이 강한 여성음성과 저음이 강한 남성음성으로 분류할 수 있다. 그러나, 기존의 HMM을 이용한 음성인식시스템에서는 남성과 여성음성의 이러한 특성이 있음에도 불구하고 이를 고려하지 않고, 하나의 HMM으로 구성하고 있다. 본 논문에서 제시하는 알고리즘으로 실험한 결과 남성과 여성의 포만트 주파수가 100~300Hz차이가 나는 것을 알 수 있었고, 이러한 특성을 고려하여 남성과 여성의 음성을 구별할 수 있는 방법을 제안한다. 또한 남성과 여성음성을 각각 구분하여 HMM을 훈련시킨 후 인식과정에서 입력된 음성의 포만트 특성에 따라 남성음성이면 남성HMM으로 여성음성이면 여성HMM으로 인식을 수행함으로써 기존의 인식방법보다 남성음성은 5.2% 여성음성은 4.4% 향상된 결과를 얻었다.

1. 서론

음성의 가장 기본적인 파라메타중에 하나가 포만트 주파수이다[1]. 포만트 주파수란 모음의 주파수 중에서 에너지가 집중적으로 나타나는 영역을 말한다. 보통 모음중에는 3~4개의 포만트가 있으며 모음의 종류마다 포만트 주파수의 값이 다르다. 또한 인간의 성대는 사춘기를 지나면서 변성과정을 거치고 이로 인해서 남성의 성대길이는 여성보다 평균 4.5mm정도 길어진다[2]. 이것이 남성의 목소리가 여성보다 저음성분이 강하다는 것을 의미한다. 기존의 HMM을 이용한 음성인식시스템에서는 남성과 여성음성의 고유한 특성이 있음에도 불구하고 이를 고려하지 않고 하나의 HMM으로 구성하고 있다. 본 논문에서는 포만트 주파수와 피치정보를 이용해서 남성과 여성의 음성을 구분할 수 있는 방법을 제안하고, HMM을 각각 구성하여 처리하는 SRPF(Speech Recognition Preprocessing system using Formant frequency)를 제안하여 인식률이 향상됨을 보인다.

2. 포만트 주파수의 추출

포만트 주파수는 음성신호의 LPC(Linear predictive coefficient)를 이용해서 구할 수 있다[3][4]. LPC는 현재의 출력음성신호를 과거의 입력신호와 과거의 출력신호와의 선형적 결합에 의해 예측할 수 있다[5]. 이것은 음성발생모델과 연관이 있어서 음성에 관한 특징을 적은수의 파라메터만으로 표현할 수 있고, 정확도와 계산속도면에서도 좋은 성능을 보이고 있다. 음성의 일정구간을 N개의 표본으로 나누면 음성신호 $s(1),\cdots,s(N)$에서 한 시점의 신호 $s(n)$을 그 이전 M ($M<N$)개의 신호 $s(n-1),\cdots,s(n-M)$에 의해 (1)과 같이 표현할 수 있다.

$$s(n)=\sum_{i=1}^{M}a_i s(n-i)+e(n) \qquad M+1\le n\le N \tag{1}$$

$s(n)$:음성신호 a_i:예측계수 M:예측차수 $e(n)$:예측오차

$$e(n)=\sum_{i=0}^{M}a_i s(n-i)=s(n)+\sum_{i=1}^{M}a_i s(n-i) \tag{2}$$

이고, 예측된 샘플 $\hat{s}(n)$은 다음과 같이 정의 할 수 있다.

$$e(n)=s(n)-\hat{s}(n) \tag{3}$$

$$\hat{s}(n)=-\sum_{i=1}^{M}a_i s(n-i) \tag{4}$$

$e(n)$은 샘플 $s(n)$과 예측값 $\hat{s}(n)$ 사이의 오차로 정의되므로 $e(n)$이 최소가 되도록 계수 a_i ($i=1,2,\cdots,M$)값

을 선택할 수 있다. 여기서 구한 LPC값을 이용해서 포만트주파수값을 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. root solving은 포만트주파수와 대역폭의 모든 후보를 뽑을 수 있기 때문에 root solving방법을 이용했다. root solving에서 어떤 복소근 z에 대한 대역폭 $\hat{B}$와 주파수 $\hat{F}$는 s-평면에서 z-평면으로의 변환에 의해 얻어진다.

$$z = e^{sT} \tag{5}$$

여기서 $s = -\pi\hat{B} \pm j2\pi\hat{F}$이고 $z = R_e(z) \pm jI_m(z)$는 복소근의 실수부와 허수부로 정의 된다. 그러면 포만트와 대역폭은 다음과 같은 식으로 구할 수 있다.

$$\hat{F} = (f_s/2\pi)\tan^{-1}[I_m(z)/R_e(z)] \ (Hz) \tag{6}$$

$$\hat{B} = -(f_s/\pi)\ln|z| \ \ (Hz) \tag{7}$$

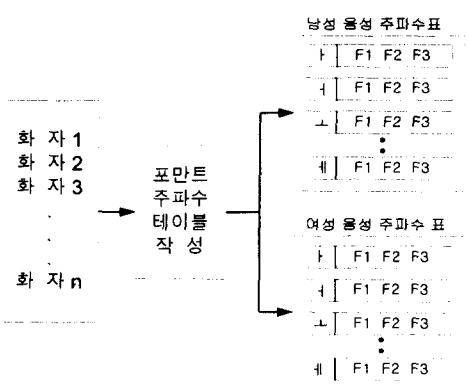
3. SRPF 시스템

본 논문이 제시하는 SRPF시스템은 두 단계로 구성되는데, 첫 번째 과정은 포만트 주파수를 추출해서 여러화자의 음성에서 추출한 포만트값으로 테이블을 구성하는 과정이고 두 번째 과정은 전체인식과정에서 포만트 주파수를 추출하여 포만트 주파수 테이블과 비교하는 전처리 과정이 포함된 음성인식 과정으로 구분한다.

3.1 포만트 주파수 테이블의 구성

포만트 주파수 테이블을 작성하기 위해 변성과정을 거친 20대 이상 남성과 여성 20명의 단모음에 대해 5번씩 발음한 음성을 이용해서 포만트 주파수 테이블을 작성한다. 각 화자의 단모음에 대한 입력을 받아 다음과 같은 식을 이용해서 포만트 주파수 값의 평균을 낸다.

$$(F_{m,1} + F_{m,2} + \cdots\cdots + F_{m,n})/n = F_m' \tag{8}$$

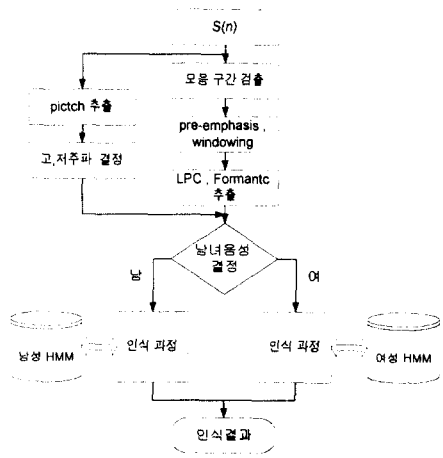


[그림 1] 포만트주파수 테이블 작성과정

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

3.2 SRPF시스템의 처리과정

앞절에서 제시한 남성과 여성의 포만트 주파수 테이블을 이용해서 음성의 특성을 판단해내는 FPSR시스템은 그림2와 같다. 먼저 입력된 음성신호에 모음구간을 검출하기 위해서 영교차율과 절대에너지 값을 이용해서 모음구간을 결정한다[4].



[그림 2] SRPF 시스템

음성신호의 주파수 스펙트럼은 일정하지 않고, 주파수 값이 높을수록 그 성분이 작아지게 되어 주파수가 2배가 되면 약 6dB의 기울기로 파워의 진폭 특성이 작아지므로, 음성신호 분석전에 6dB기울기를 갖는 고역강조 필터를 통과시켜 음성신호의 스펙트럼이 저역부터 고역까지 같은 S/N비를 갖게 하는 선 강조 과정을 거친다[1][6].

$$H(Z) = 1 - aZ^{-1} \ \ (a : 0.9 \text{ 또는 } 0.875) \tag{9}$$

음성 양끝 단에서는 날카로운 잡음성분을 방지하기 위한 창함수로는 해밍창을 이용한다[6].

$$W(i) = 0.54 - 0.46 * \cos(2\pi i * i/N) \quad 0 \le i \le N \tag{10}$$

```
begin
  while(zero < zero_threshold){
    while(energy > energy_threshold){
      Formant();
      Form_sum[i] += Form_temp[n];
      n++;
    }
  }
  Form_res[k] = Form_sum[i]/n
  while(Form_table[m]-Form_res[k]<Threshold)
  m++;
  result m; // Formant table number
end
```

[알고리즘 1] 남녀음성결정 알고리즘

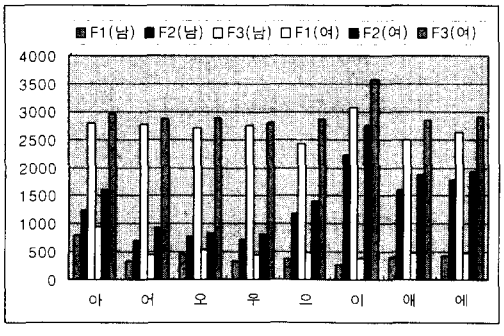
입력된 음성의 모음구간에 대한 포만트값을 추출한 후 남녀의 음성을 결정하는 알고리즘은 [알고리즘1]과 같다. 입력된 음성에서 모음구간의 각 프레임에서 구한 포만트 값의 합을 계산한 후 전체 프레임수(n)으로 나누면 입력 음성중 모음구간의 평균 포만트값 (F_1, F_2, F_3)이 구해진다. 이 값과 미리 작성된 포만트주파수 테이블을 비교하기 위해서, 입력음성의 포만트 값과 테이블 값의 차를 계산한 후 가장 작은 값을 가지는 테이블의 번호를 선택하면 남녀의 음성을 결정할 수 있다. 남녀의 음성이 결정되면 각각 남성음성인식과정과 여성음성인식과정으로 나뉘어져 처리가 된다. 인식모델은 무작위로 선출한 30명의 이름에 대해서 각각 3번씩 발음 한 후 MFCC(Mel Frequency Cepstral Coefficients)를 이용해서 모노폰으로 구성하였다. 이때 남성에 대한 발음은 남성HMM으로, 여성에 대한 발음은 여성HMM으로 구성하여 학습시킴으로서 전체를 하나의 HMM으로 구성할 때 보다 인식률을 향상시킨다.

4.실험결과

남녀 각각 10명이 단모음에 대해서 5번씩 발음한 800개의 음성을 이용해서 작성한 포만트 주파수 테이블과 분포도는 아래와 같다. 실험결과 여성의 포만트 분포는 남성보다 평균 100~300Hz 높은 결과가 나왔다. 이 결과

[표 1] 포만트 주파수 테이블

모음	남성			여성		
	F1	F2	F3	F1	F2	F3
ㅏ	784	1230	2777	944	1603	2947
ㅓ	332	705	2762	436	930	2868
ㅗ	459	781	2696	544	828	2887
ㅜ	322	712	2744	452	820	2806
ㅡ	360	1180	2408	492	1397	2853
ㅣ	274	2216	3063	393	2740	3573
ㅐ	413	1613	2500	487	1926	2893
ㅔ	431	1775	2622	485	1876	2835



[그림 3] 포만트 주파수 분포도

를 이용해서 남녀의 음성을 구별할 수 있었다. 인식과정에서 남녀로 구분하여 훈련한 경우가 전체HMM으로 구성한 것보다 남성화자의 경우 5.2% 여성화자의 경우 4.4%의 향상됨을 보였다. 반면에 남성과 여성음성의 특성에 의해서 남성화자의 음성을 여성음성으로 구성된 HMM으로 인식을 한 경우 26.7%의 인식률이 저하됨을 확인하였다.

[표 2] 인식결과

	인 식 결 과			향상률
	전체HMM	남성HMM	여성HMM	
남성화자	93.3%	98.5%	66.6%	5.2%
여성화자	92.9%	61.1%	97.3%	4.4%

5. 결론

일반적으로 우리가 발성하는 음성에는 화자의 특성을 추정할 수 있는 정보를 가지고 있고, 이러한 정보를 이용하여 화자의 성(性)을 측정할 수 있다. 음성인식시스템에서 화자의 특성에 따라 HMM을 클러스터링하여 훈련시킨다면 적은 양의 훈련 값을 가지고도 높은 인식률을 얻을 수 있음을 확인하였다. 포만트 주파수를 이용하여 제1차변성기를 거친 남성과 여성음성을 구별하여 인식률의 결과를 보았는데, 향후 연구 과제로는 이중모음에 대한 포만트 추출방법과 연령별로 음성특성을 구별하는 시스템의 연구가 필요하다. 또한 화자인식시스템에 화자의 특성을 구별할 수 있는 방법이 적용되면 화자의 성(性)과 연령, 상태 등을 추정할 수 있는 시스템을 구현할 수 있을 것이다.

참고문헌

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

[2] 문영일, *기초음성학과 발성기법*, 청우, 1987.

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

[4] 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.

[5] 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.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals, Prentice Hall*, 1978.