음성 스펙트럼 분석에 의한 한국어 단모음 실시간 인식

김엄준°, 성미영
인천대학교 전자계산학과

Real-Time Recognition of the Korean Single Vowels Using the Speech Spectrum Analysis

Eom-Jun Kim°, Mee-Young Sung
Dept. of Computer Science, University of Inchon

요 약

본 연구에서는 짧은 시간에 계산이 가능하며, 음성을 특징 지울 수 있는 파라미터로서 영 교차율(zero crossing rate), 단 구간 에너지(short-term energy) 그리고 포만트(formant)를 사용하였다. 특정 화자의 음성을 입력받아서 단모음인 'ㅏ, ㅐ, ㅓ, ㅔ, ㅗ, ㅜ, ㅡ, ㅣ'에 대한 인식을 위해 위의 세 가지 파라미터를 측정하였다. 영 교차율과 단 구간 에너지 파라미터는 유성음과 무성음의 구별과 음성인지 아닌지를 판별하는데 사용하였다. 포만트 파라미터는 10차 켑스트럼(cepstrum)을 이용하여 구하였으며, 각 단모음을 판별하기 위해서 사용하였다. 하나의 단모음을 입력받아 처리하여 텍스트로 출력하는데 평균 0.065sec에 처리하며, 각각의 단모음에 대해 93%, 10개의 테스트 문장에 대해 72%의 인식률을 보이고 있다.

1. 서론

음성은 사람과 사람 상호간에 의사 전달을 하기 위해 인간이 사용하는 여러 가지 전달 수단 중 가장 기본적인 전달 수단이다. 인간에게 있어서 기본적인 의사 소통의 수단인 음성은 편리성과 경제성의 측면에서 다른 방법에 비해 우수한 특성을 가지고 있으며, 이러한 음성을 컴퓨터가 인지할 수 있도록 오래 전부터 많은 연구가 진행되어 왔다. 음성은 특성상 언어학, 음성학, 음운학, 생리학, 해부학 등의 여러 가지의 학문과 결합되어 연구되어 왔으며, 이러한 학문의 발전으로 인해서 컴퓨터의 음성 인식은 좋은 성과를 거두고 있다[1].

기존의 연속음 인식 시스템의 문제점으로 대부분의 음성 분석은 화자가 주의 깊게 정확히 발음한 데이터에 대한 것으로서 자연 발화시의 음성과는 다소 차이를 보인다[2]. 또한 발화 속도의 부분 변화나 강세 같은 음성의 길이 변화는 대부분의 은닉 마르코프 모델 시스템에서 무시되어 왔으며, 이로 인한 인식률의 저하는 피할 수 없었다[3][4]. 현재의 음성인식 시스템은 제한된 문법을 사용함으로써 대용량 단어인식 시스템의 인식률을 높이고 있다.

한국어의 음성 인식을 위해서 좋은 성능을 보이는 여러 가지 방식의 알고리즘이 연구되었다. 이를 바탕으로 본 연구에서는 마이크를 통하여 음성을 받아

들여 A/D converter를 통하여 디지털 신호로 변환한 후 음원 파형을 주파수 대역으로 변환하여 분석을 통한 음성의 특징 파라미터인 영 교차율, 단 구간 에너지 그리고 포만트를 구하여 단모음을 인식하고, 그 결과를 텍스트로 변환하여 출력하였다. 음성 인식을 위한 파라미터중 실시간으로 처리하여 결과를 보일 수 있는 파라미터에 대해서 알아본다. 음성 인식 단위로는 커다란 데이터베이스를 요구하지 않는 음소 단위로 하였으며, 음소중 공명 주파수의 특징을 나타내고 있는 한국어 단모음에 대해 연구를 하였다.

2. 시스템 구성

본 연구를 위한 시스템 구성, 음성 인식 순서도 및 처리 과정은 다음과 같다.

- 운영체제 : 윈도우 95
- 속 도 : 펜티움 200MHz
- 램 : 64M
- 입 력 부 : Sound Blaster PCI64
- 개 발 툴 : Visual C++ 5.0
- 샘 플 링 : 22050bit/s
- 비 트 수 : 8bit/sample

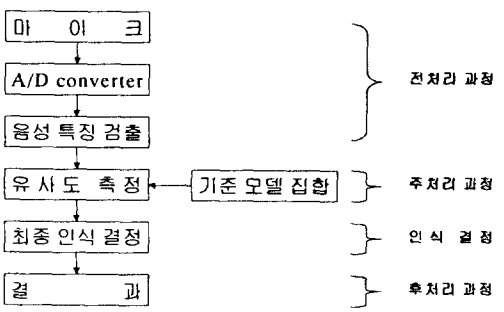

그림 1. 음성 인식 순서도

마이크를 통하여 실시간으로 입력된 음성은 A/D converter를 통하여 디지털 신호로 변환한다. 변환된 신호를 가지고 음성 구간을 검출하기 위하여 단구간 에너지와 영 교차율의 평균값과 표준 편차를 구하여 경계 잘림(threshold)을 행한 후 단구간 에너지의 경계치와 비교하여 대략적인 음성 구간을 구한다. 실시간으로 입력되는 음성 신호를 분석하여 음소를 기본 단위로 한 음성 특징을 검출한다. 검출된 음성 특징과 기준 모델 집합과의 유사도를 측정하여 하나의 자음이나 또는 모음으로 결정한다. 본 연구에서는 성대의 공명 주파수를 나타내는 파라미터인 포만트 계수의 특징값을 가지고서 모음에 대한 인식을 한다.

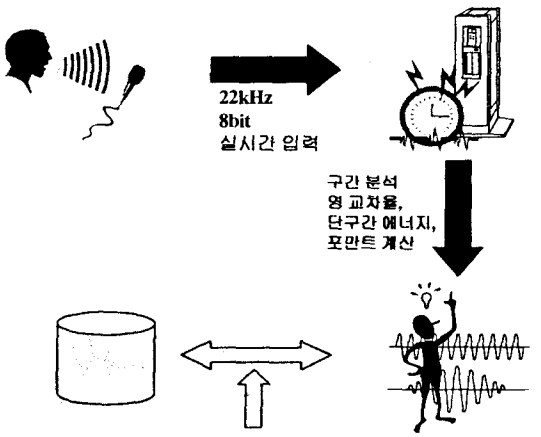

그림 2. 처리 과정

3. 처리 과정

음성 파형은 시간에 따른 데이터 양이 많은 단점을 가지고 있어서 음성 파형을 주파수 영역으로 변환하여 특징을 추출하는 방식을 주로 사용하고 있다[5]. 본 연구에서는 여러 가지의 음성 특징 파라미터 중에서 실시간으로 처리할 수 있는 파라미터를 이용하여 인식을 하였다. 음성을 특징 지울 수 있는 여러 가지 파라미터 중에서 빠른 시간 내에 처리를 할 수 있고, 각 단모음의 특성을 특징 지울 수 있는 단구간 에너지, 영 교차율 그리고 포만트를 가지고 음성을 인식하였다. 일반적인 음성 인식에서의 처리 과정은 다음과 같은 순서와 같다.

- 마이크를 통하여 보드에 입력된 음성은 8비트에 22kHz의 주파수로 샘플링 되어 디지털 신호로 변환된다.
- 전체 분석할 구간을 정한다.
- 각 구간에서의 음성 특징 파라미터를 구한다.
- 구해진 파라미터와 데이터베이스의 기준집합 파라미터와의 비교를 통하여 음소를 구별한다.

음성의 시작점은 대부분의 음성이 발음될 때 최고

점으로 상승하여 주기가 되는 샘플까지 떨어지는 점을 이용하여 최고 상승 점에서 이전 샘플이 가장 먼저 영점이 되는 지점을 시작점으로 한다. 끝점은 주기가 반복되는 지점으로 정한다. 끝점 검출은 평균차 함수법(AMDF)을 사용하여 피치가 되는 지점을 이용하여 구한다. 분석할 시작점과 끝점이 구해지면 이에 대한 영 교차율과 단구간 에너지를 구한다. 영 교차율은 식 (1)과 같이 구해지며, 단구간 에너지는 식 (2)와 같이 구해진다[6][7].

$$Z(i)=\sum_{n=0}^{N}|A|,\quad A=sign(s(N*i+n))-sign(s(N*i+n-1)),\quad i=0,1,\cdots,4 \qquad (1)$$

$$sign(s(n))=\begin{cases}1, & s(n)\geq UM\\ -1, & s(n)\leq LM\end{cases}$$

$$E(i)=\sum_{n=0}^{N}|s(N*i+n)|,\quad i=0,1,\cdots,4 \qquad (2)$$

식 (1)과 식 (2)에 의해서 구해진 영 교차율과 단구간 에너지는 유성음과 무성음, 그리고 음성인지 아닌지를 판별하는데 사용된다. 음성이 아닐 경우 높은 값을 가지게 되며, 유성음과 무성음일 경우 그에 맞는 영 교차율과 단구간 에너지 값을 가지게 된다. 본 연구에서 계산된 각 모음에서의 영 교차율과 단구간 에너지 값은 표 1과 같다.

표 1. 단모음에서의 영 교차율과 단구간 에너지

	영 교차율	단구간 에너지
ㅏ	16 ~ 24	1303 ~ 5916
ㅐ	12 ~ 20	
ㅓ	14 ~ 20	
ㅔ	12 ~ 24	
ㅗ	8 ~ 16	
ㅜ	8 ~ 20	
ㅡ	8 ~ 16	
ㅣ	8 ~ 12	

표 1에서는 각 모음을 오전, 오후 그리고 저녁으로 나누어 각각 14번 발음하여 허용한도를 나타내었다. 연속하여 발음되는 음성을 빠른 시간 내에 처리하고, 측정단어의 연속발음에 의하여 합성이 되어 있으므로 이의 영향을 최소한으로 줄이기 위해서 구간의 시작점으로부터 한 주기에서 측정하였다. 처리 과정을 알고리즘으로 표현하면 다음과 같다.

```
index = 1024;              //입력 음성 크기
Pitch( 1, index);          //피치 추출
region_s = pitch[0];       //구간의 시작
region_e = pitch[1];       //구간의  끝
region = 256;              //분석 구간의 크기
ZCR( region_s, region_e);          //영 교차율
ShortTermEnerge(region_s, region_e);
                           //단구간 에너지
for( i=0; i<index; ++i) {
   fft_r[i] = 0.;
   fft_i[i] = 0.;
}
for( i=region_s; i<region_s+region; i++) {
   fftdata[i-region_s].re = save_signal[i];
   fftdata[i-region_s].im = 0.;
}
FFT( fftdata, region_s, region, 1); //푸리에 변
환
for( i=region_s; i<region_s+region; i++) {
   PW[i]=sqrt(fft_r[i]*fft_r[i]+ fft_i[i]*fft_i[i]);
}
for( i=region_s; i<region_s+region; ++i) {
   fftdata[i-region_s].re = PW[i];
   fftdata[i-region_s].im = 0.;
}
FFT( fftdata, region_s, region, -1);//역변환
Cepstrum( region_s, region, fft_r, 10);//10차 캡
스트럼
Lifter( region_s, region, 512);    //10차 필터
```

처리 과정에서 측정된 파라미터를 기준 음소 모델의 값과 비교한다. 비교 과정에서 입력된 음성은 사람에 따라서 또는 환경에 따라서 차이가 나므로 이를 감안한 편차를 적용하여 범위 내에 위치하게 되면 그에 해당되는 출력을 하게 된다. 표 2는 본 연구에서 측정된 단모음에서의 공명 주파수와 공명 주파수의 편차를 보여주고 있으며, 포만트 값인 F1, F2 그리고 F3의 값의 분포를 그래프로 표현하여 그림 5, 6, 7에 각각 나타내었다.

표 2. 단모음의 공명 주파수

	F1	F2	F3
ㅏ	657± 19	916± 19	2397±674
ㅐ	493± 19	1884± 29	2777±107
ㅓ	625± 39	792± 39	2918±127
ㅔ	303± 58	571± 39	1904± 97
ㅗ	441± 29	638± 19	2716±156
ㅜ	263± 59	724± 68	1785±205
ㅡ	349± 9	1030± 88	2331± 97
ㅣ	249± 9	2094± 68	3124± 58

편차가 50보다 작은 값을 보이는 단모음은 거의 일정한 포만트 값을 가지고 있으며, 만약 편차에 의한 값보다 높은 값이 나올 경우의 값은 그 다음 포만트의 값을 나타낸다. '아'의 경우 F1과 F2의 값이 거의 일정하게 나오고 있으며, F3의 경우 측정한 값의 절반 정도는 F4의 값을 나타내고 있다. 이와 마찬가지로 다른 단 모음도 비슷한 경향을 나타내고 있다. 포만트의 F1, F2, F3의 분포를 보여주는 그림 5, 6, 7에서와 같이 F1의 값은 모든 단모음에서 오차범위를 벗어나지 않는다. 그러나 F2와 F3의 값은 중간에 오차범위를 벗어나는 수치가 나오고 있다. 오차범위를 벗어나는 원인으로 포만트 중에 인접한 값들이 이전의 값에 동화되어서 나타나지 않는 것으로 여겨진다. 그러나 오차범위를 벗어나는 값은 그 다음의 포만트 값을 가지고 있다. 그러므로 단지 정확한 값이 나타나지 않았을 뿐이지 음성의 특징 값은 가지고 있다. 그림 6에서 단모음 '어'의 분포를 보게되면 많은 부분에서 오차범위를 벗어나는 수치가 나오고 있다. 그러나 그 수치는 그림 7에서 보는 바와 같이 F3의 범위에 속한다. 그림 8에서부터 그림 15까지는 측정한 단모음의 F1, F2, F3의 분포를 그래프로 나타내었다. 그림 8에서 단모음 'ㅏ'에 대한 F1과 F2는 좁은 오차 범위 내에서 정확한 값이 나오는 반면에 F3은 넓은 분포를 보이고 있다. F3이 넓은 분포를 보이는 이유는 특정 사람의 음성 특성이 나타나는 것으로 여겨진다. 그러나 결과적으로 F1과 F2는 명확한 특징을 나타내고 있으므로 인식을 위한 특징 값으로 사용할 수 있다.

4. 실험 및 측정

처리 과정에서 나온 결과를 가지고 간단한 음성-텍스트 변환기를 제작하였다. 마이크를 통하여 실시간으로 들어오는 음성을 빠른 시간 내에 계산하여 인식을 할 수 있으며, 화자 종속으로 측정하였다. 먼저 각각의 단모음에 대해 낮은음, 보통음, 높은음으로 구별하여 측정하였고, 아침, 점심, 저녁으로 구분하여 10개의 문장을 읽어 인식 결과를 측정하였다.

- 측정한 문장
  1. 안녕하세요
  2. 멀티미디어 시스템
  3. 꼬리에 꼬리를 무는 영어
  4. 아메리카 생활 영어
  5. 사회를 본다 사람이 보인다
  6. 나는 활동 범위를 넓혔다
  7. 어서 오십시오
  8. 무엇으로부터 안전하다는 것이지요
  9. 스스로를 잊어버릴 정도였습니다
  10. 다음과 같이 제안합니다

- 측정 시간
  오전 9시,
  오후 15시,
  저녁 21시

음성을 발음한 이후 약 2초간 음성을 받아들였으며, 이 때 각 구간마다 파라미터를 계산하여 인식을 행한 후 결과를 텍스트로 출력하였다. 그림 3은 '아' 음성에 대해 텍스트로의 출력 결과를 보여주고 있다.

문장에서의 단모음 인식은 단모음의 앞과 뒤에 연속해서 들어오는 자음이나 모음의 영향을 받는다. 그러므로 다른 자음이나 모음에 대해서 정의를 하고, 자음이나 모음과의 여러 가지 상관 관계를 알아보아야 할 것으로 보인다. 그림 4는 측정한 문장중 '안녕하세요'에 대한 인식 결과를 출력한 결과이다.

결과적으로 현재까지 연구되어온 음성 인식을 위한 포만트 파라미터의 측정은 실시간으로 처리하기 위해서는 표 2와 같이 약간의 편차를 가지고 측정하여야 한다. 각 단모음에 대한 인식률은 평균 93%의 인식률을 보이고 있으며, 문장에서의 단모음 인식률은 평균 72%의 인식률을 보였다.

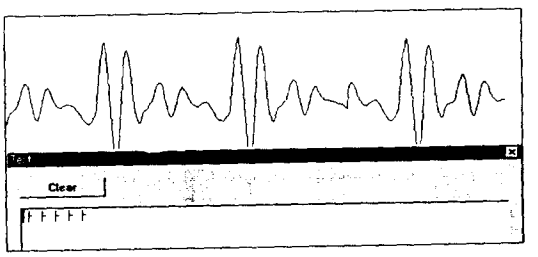

그림 3. 단모음 'ㅏ'의 인식 결과

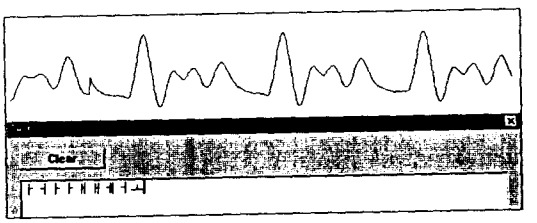

그림 4. 문장 '안녕하세요'의 인식 결과

5. 결론 및 향후 연구 방향

본 연구에서는 멀티미디어 요소 중의 하나인 음성을 인식하여 텍스트로 변환하는데 필요한 기술상의 방법, 문제점 및 해결방안을 연구하였다. 미디어간의 변환을 위해서는 각 미디어에서의 처리기술이 발달해 있어야만 한다. 사람에게 있어서 중요한 의사전달 수단중의 하나인 음성은 오래 전부터 진행되어온 연구를 바탕으로 실시간으로 입력되는 음성 신호의 스펙트럼 분석을 통하여 특징 파라미터를 구하고, 음성을 인식하여 그 결과를 텍스트로 변환하였다. 음성 신호를 분석하기 위해서는 음소를 기본 단위로 하여 데이터베이스를 최소화하고, 음성 특징을 검출하기 위해 처리 시간이 빠르면서 각 음소를 특징 지울 수 있는 파라미터를 사용하였다. 음성을 특징 지울 수 있는 파라미터로서 영 교차율, 단 구간 에너지 그리고 포만트 값을 사용하였다. 영 교차율과 단 구간 에너지 파라미터는 유성음과 무성음의 구별을 하는데 사용하였고 음성인지 아닌지를 판별하는데도 사용하였다. 유성음일 경우 영 교차율은 높은 값이 검출되고, 음성일 경우 단 구간 에너지는 높은 값이 검출된다. 포만트 파라미터는 10차 켑스트럼을 사용하여 구하였으며, 본 연구에서 측정한 단모음의 구별을 위하여 사용하였다. 이 때, 켑스트럼은 시간 영역의 음성 파형을 주파수 영역으로 변환한 후 로그 대수를 취하고 다시 역 변환하여 계산하였다. 이 연구에서는 특정 화자에 대하여 한국어의 단모음인 'ㅏ, ㅐ, ㅓ, ㅔ, ㅗ, ㅜ, ㅡ, ㅣ'에 대한 포만트 값을 측정하였다. 각 포만트 값은 오차범위 한도 내에서 측정되었으며, 범위를 벗어나는 값이 나올 경우에는 그 다음의 포만트 값을 취하였다. 예를 들면 'ㅓ'의 F1값과 F2의 포만트 값은 200Hz이내에 인접해 있기 때문에 하나의 포만트로 합쳐지는 경우가 발생한다. 이 때 구하게 되는 F2의 값은 정의해 놓은 값의 오차범위를 벗어나게 된다. 그러나 벗어나는 F2의 값은 F3의 범위에 속하게 되어 단모음 'ㅓ'로서 인식이 가능하다. 각각의 단모음에 대한 음성-텍스트 변환의 실험에서는 평균 93%의 인식률을 얻을 수 있었고, 10개의 문장을 오전, 오후 그리고 저녁으로 구분하여 단모음에 대한 음성-텍스트 변환을 실험한 결과는 평균 72%의 인식률을 얻을 수 있었다.

본 연구를 바탕으로 복모음과 자음을 인식하여 텍스트로 변환하는 연구가 진행되어야 할 것이며, 모든 음소를 인식할 수 있는 빠르고, 간단한 알고리즘의 연구가 필요하다.

참 고 문 헌

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

[2] 김재범, 이홍규, 이정행, "한국어 연속음 인식을 위한 발화 속도 측정기의 설계 및 구현", 한국정보처리학회 제 3권 제 2호, pp. 755-758, 1996.

[3] David Burshtein, "Robust Parametric Modelling of Durations in Hidden Markov Models," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Vol. 1, pp. 683-686, 1995.

[4] D. O'shaughnessy, "Timing Patterns in Fluent and Disfluent Spontaneous Speech," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Vol. 1, pp. 600-603, 1995.

[5] 허만택, 김재창, "확산필터뱅크를 전처리기로 사용한 한국어 단모음인식", 한국음향학회지, 제16호 제 1권, pp. 81-87, 1997.

[6] J. G. Proakis, D. G. Manolakis, "Digital Signal Processing", Prentice Hall, 1996.

[7] J. R. Deller, J. G. Proakis, J. H. L. Hansen, "Discrete-Time Processing of Speech Signals", Prentice Hall, 1987.

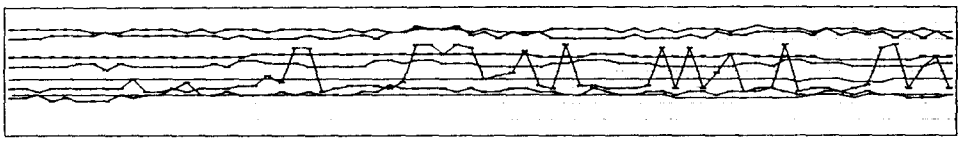

그림 5. 포만트 F1의 분포 그래프

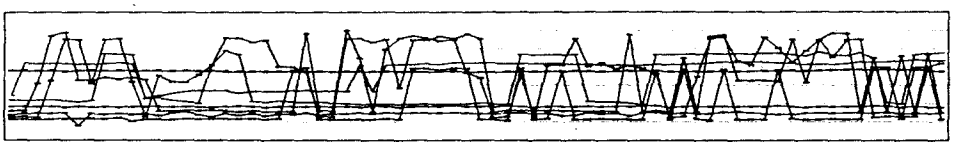

그림 7. 포만트 F2의 분포도

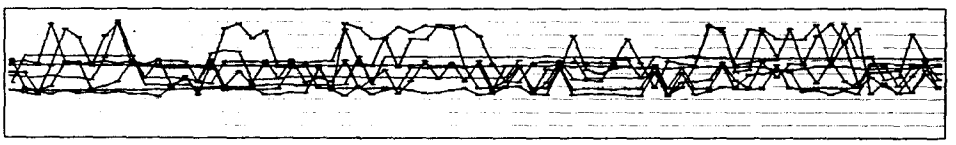

그림 8. 포만트 F3의 분포도

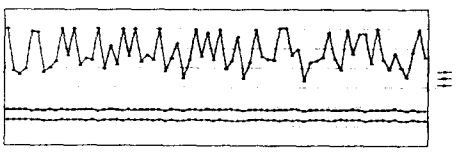

그림 6. '아'의 포만트 분포도

그림 12. '오'의 포만트 분포도

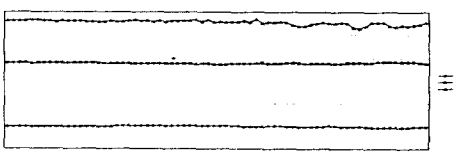

그림 9. '애'의 포만트 분포도

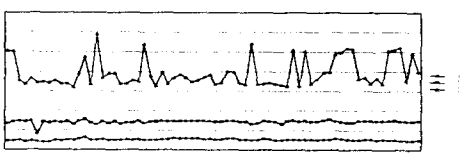

그림 13. '우'의 포만트 분포도

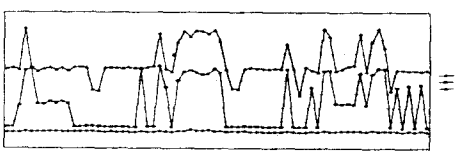

그림 10. '어'의 포만트 분포도

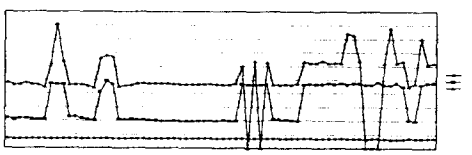

그림 14. '으'의 포만트 분포도

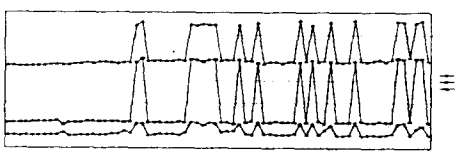

그림 11. '에'의 포만트 분포도

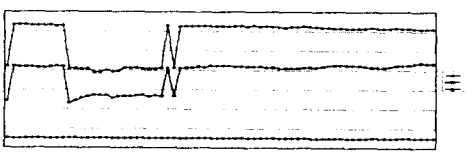

그림 15. '이'의 포만트 분포도