대화체 음성에서의 한국어 연결 숫자음 인식

김중철[U] 고종철 이정현
인하대학교 전자계산공학과
[jchkim, jchko]@nlsun.inha.ac.kr jhlee@inha.ac.kr

Recognition of Korean Connected Digits in a Natural Spoken Dialog

Joong-Cheol Kim[U] Jong-Cheol Ko Jung-hyun Lee
Dept. of Computer Science and Engineering, Inha University

요 약

대화체 음성의 인식을 위해서는 음성 파형에 관한 음향학적인 연구뿐만 아니라 인식하려는 언어자체에 대한 언어학적인 연구를 필요로 한다.

본 논문에서는 숫자음의 언어학적인 요소를 고려하고, 포만트 주파수를 숫자음 검출과 숫자음 인식에 적용하는 방식을 제안한다. 시스템의 입력은 특정 질의에 대한 응답으로 대화체 문장이며, 끝점 추출 기술을 이용하여 고립단어로 분류한 후, 숫자음만을 검출해 내고, 검출된 숫자음을 인식하기 위해 포만트 주파수를 이용한다. 한국어 연결 숫자음 인식은 한국어 숫자음이 단음절로 구성된다는 점과 발음상의 조음효과 등으로 한계를 가지고 있다. 본 논문에서는 숫자음의 발성에 필요한 음소들을 추출하고, 숫자들을 모음에 따라 6개의 그룹으로 분류하여 인식의 범위를 좁히고, 포만트 주파수 정보와 음소 HMM모델에 의한 두 단계에 걸친 인식을 수행함으로써 연결 숫자음 인식에 대한 성능을 향상시킨다.

1. 서론

음성인식의 기술이 빠르게 발전함에 따라 그 응용도 다양해지고 있으며 사용자도 보다 자연스러운 인터페이스를 요구하고 있다. 이런 요구에 따라 음성인식의 응용분야도 점차 확대되고 있다. 음성인식의 응용분야 중 숫자음 인식에 대한 필요성은 텔레 뱅킹, 주식 시세, 각종 ARS 서비스 등 여러 분야에서 급증하고 있다.

본 논문에서는 자연스러운 대화체 내에서 한국어 연결 숫자음만을 추출하여 인식하는 시스템을 구현하였다. 대화체 내에 포함된 연결 숫자음을 추출해 내기 위해서 핵심어 인식기술을 변형하여, 비핵심어 모델인 가비지 모델 대신 포만트 주파수를 이용하여 비핵심어를 거절하고, 숫자음 검출을 위해 음소모델에 포만트 주파수 가중치를 두어 연결된 숫자음만을 핵심어로 추출하도록 설계하였다. 또한 한국어 숫자음이 단음절로 이루어졌다는 점과 한국어 숫자음에 포함된 모음이 한정되어 있다는 점에 착안하여 숫자음에 포함된 모음의 포만트 주파수와 포만트 주파수 전이율을 이용하여 숫자음 인식에 대한 인식률을 향상시켰다.

2. 관련 연구

2.1 LPC-Cepstrum

LPC 분석은 음성의 특징을 과거 몇 샘플들의 선형적인 조합으로 현재의 샘플을 예측하는 방법으로 음성 분석에 적용되어 효과적으로 음성을 모델링 할 수 있는 음성 특징 추출 방법이다. 시간 t에서의 예측 신호를 $s(n)$이라고 할 때, $s(n)$은 과거 p개의 음성샘플의 선형조합으로 근사할 수 있다. 이를 잔차신호 $Gu(n)$을 포함하여 식(1)로 표현할 수 있다.

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \qquad (1)$$

여기서, a_i가 LPC계수가 되며, p는 계수의 차수이다.

LPC스펙트럼으로부터 Cepstrum계수를 추출해 낼 수 있으며, LPC계수로부터 얻어진 Cepstrum계수를 LPC-Cepstrum계수라고 한다. LPC-Cepstrum계수는 식(2)에 의해 LPC계수 a_i로부터 구할 수 있다.

$$c_0 = \ln \sigma^2$$

$$c_m = a_m + \sum_{k=1}^{m-1}(\frac{k}{m})c_k a_{m-k}, \quad 1 \le m \le p \qquad (2)$$

$$c_m = \sum_{k=1}^{m-1}(\frac{k}{m})c_k a_{m-k}, \quad m > p$$

여기서 c가 Cepstrum계수이며, σ^2은 LPC모델의 이득항이다. LPC-Cepstrum계수는 Cepstrum의 성질을 충분히 나타내면서 적은 계산량으로 구할 수 있다는 장점이 있다[1].

본 논문에서는 16차 LPC-Cepstrum 계수를 음성 특징 파라미터로 사용하였다.

2.2 포만트 주파수

모음구간에서 추출되는 포만트 주파수는 음소의 식별에 있어서 중요한 정보를 담고 있다. 특히 포만트 주파수를 청각 모델인 바크 스케일링으로 정규화하고 채널별 바크 스케일링 주파수의 차를 구함으로써 문맥과 화자의 변화에 의해 발생하는 변이를 줄일 수 있다. 기본 주파수 F_0와 1,2,3차 포만트 주파수 F_1, F_2, F_3로부터 바크 스케일링을 하기 위해 식(3)을 사용한다[2][3][6].

$$B_i = 13\arctan(0.76F_i) + 3.5\arctan(F_i/7.5)^2 \qquad (3)$$

2.3 한국어 숫자음에 대한 고찰

한국어 연결 숫자음은 조음효과로 인해 하나의 음을 여러 가지 소리로 발음한다. 예를 들어, 6의 발음은 그 위치에 따라 '육', '륙', '뉵', '융' 등의 여러 가지 소리로 발음한다. 이것은 한국어 숫자음이 단음절로 이루어져서 각각의 구별된 특징을 추출하기 어렵다는 점과 함께 숫자음 인식의 저해 요인이 된다.

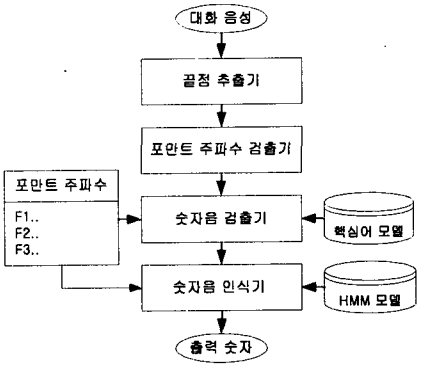
[표1]은 한국어 숫자음에 대한 종류와 그에 따르는 다양한 발음들을 보이고 있다. [표1]에서 보는 바와 같이 전후 숫자음의 영향으로 변경된 숫자음의 발음은 1과 2에서 발생되는 '이'라는 발음과 3과 4사이에 발생되는 '사'라는 발음은 서로 같은 발음을 낸다. 그러나 3과 4의 발음 파형을 조사한 결과 3뒤에 모음이 올 경우 '삼'의 받침이 뒤의 모음에 영향을 줄 뿐 2의 경우에서처럼 확연하게 분리되지 않음을 알 수 있었다.

[표1] 한국어 숫자음의 종류와 발음

종류	발음	예제
일	일, 길, 이, 릴, 밀	(19)(61)(15)(71)(31)
이	이, 기, 리, 미	(23)(62)(72)(32)
삼	삼, 쌈, (사)	(37)(73)(35)
사	사, 싸	(34)(74)
오	오, 고, 로, 모	(55)(65)(75)(35)
육	육, 뉵, 륙, 융, 유	(63)(36)(76)(66)(62)
칠	칠, 치	(17)(72)
팔	팔, 파	(18)(82)
구	구, 꾸	(96)(69)
영	영, 넝, 령	(00)(70)
공	공, 퐁	(60)

3. 대화음성 내의 숫자음 검출 인식을 위한 시스템

본 논문에서 제안하는 시스템은 [그림1]과 같으며, 크게 끝점 추출기, 포만트 주파수 검출기, 숫자음 검출기, 숫자음 인식기로 나눌 수 있다. [그림1]의 숫자음 검출기는 기존의 핵심어 검출 기술에서 비핵심어 모델 부분을 포만트 주파수를 이용하는 것으로 대체하고, 숫자음 인식기 부분에서 포만트 주파수를 검사한 후에 HMM모델을 적용하는 2단계의 인식을 수행한다.



[그림1] 전체시스템 개념도

3.1 끝점 추출기

대화체 음성에서 숫자음만을 추출해 내기 위한 전처리 부분으로 대화문장을 단어로 잘라내는 부분이다. 전체 음성에 대하여 20ms씩 진행하고 10ms간격으로 중첩하여 선강조 및 창함수를 적용한 후 절대에너지를 구하고, 영교차율을 계산하면 그 경계를 구분해 낼 수 있다. 각각의 상위임계값과 하위임계값은 정규화된 음성에 대해 실험을 통해 조절하였다.

3.2 포만트 주파수 검출기

본 논문에서는 한국어 숫자음이 [표1]과 같이 모음에 따라 분류할 수 있다는 점에 착안하여 숫자음의 각 음소에 대한 포만트 주파수 정보를 숫자음 검출 부분과 숫자음 인식 부분에 적용하여 그 성능을 향상시켰다.

구해진 LPC계수를 이용하여 Root-Solving방법과 Peak-Picking방법을 적용하여, 포만트 주파수와 대역폭을 추출하였다[4]. 훈련을 위한 음성 데이타를 통해 단모음에 대한 포만트 주파수를 구하고 바크 스케일링을 하여 [표2]와 같은 테이블로 저장하였다.

[표2] 포만트 주파수 테이블

INDEX	F1	F2	F3	B01	B23
0(아)	730±50	998±50	2440±50	6.62±0.5	2.78±0.5
1(오)	544±50	828±50	2787±50	3.36±0.5	6.40±0.5
2(우)	322±50	712±50	2344±50	3.47±0.5	3.66±0.5
3(이)	274±50	2216±50	3060±50	2.18±0.5	2.23±0.5
4(어)	432±50	835±50	2762±50	4.38±0.5	4.79±0.5
5(으)	360±50	1180±50	2408±50	2.37±0.5	3.57±0.5
6(애)	409±50	1613±50	2506±50	3.28±0.5	2.69±0.5

3.3 숫자음 검출기

숫자음 검출기 부분에서는 포만트 주파수 검출기에서 검출된 포만트 주파수 정보를 이용하여 숫자음에 포함된 모음 이외의 모음이 발생하면 인식의 대상에서 제외한다. 이중모음의 경우 전반부와 후반부의 발음이 틀려지는데 일반적으로 이중 모음 음성 신호의 많은 부분을 차지하는 후반부가 미치는 영향을 고려하여, 이중 모음으로 부터 지속시간을 측정 한 후 가장 긴 단모음 구간의 36% 이하의 지속 시간을 가지는 경우 이중모음 특징을 부여하도록 정의하였다. 포만트 주파수의 전이는 음성에 지배적인 영향을 미치는 F_3 까지만을 가지고 측정하는데 F_1, F_2, F_3의 세 포만트가 모두 변화되었을 경우 모음의 변화가 있는 것으로 간주한다.

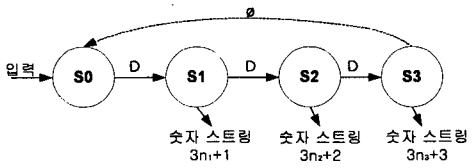
3.4 숫자음 인식기

한국어 숫자음은 단음절로 이루어져 있을 뿐만 아니라, 발음의 변화가 심하며, 숫자음의 조합을 모두 모델링 한다는 것은 비효율적이다. 본 논문에서는 인식모델을 음소 HMM모델로 하고, 포만트 주파수 추출기에서 생성된 포만트 주파수 테이블을 이용하여 2단계의 인식을 시도하여 인식의 성능을 향상시켰다. [표3]은 숫자음의 모음에 따른 분류이다.

[표3] 모음에 따른 한국어 숫자음

INDEX	GROUP
1(아)	삼(쌈), 사(싸), 팔(파)
2(이)	일(길,이,밀,릴), 이(기,리,미), 칠(치)
3(오)	오(고,로,모), 공(꽁)
4(우)	구(꾸)
5(유)	육(뉵,륙,융,유)
6(여)	영(녕,령)

인식을 위해 한 음절을 인식의 단위로 하며, 연결 숫자음에 대한 문법구조는 [그림2]를 적용하였다.



[그림2] 연결 숫자음 인식을 위한 문법 구조

4. 실험 및 평가

실험을 위한 문장을 얻기 위해 일정 질의를 정하여, 그 질의에 대한 응답으로 입력음성을 구성하였으며, 남여 각각 20명씩의 음성 중 남여 10명씩의 음성을 훈련을 위한 데이타로 사용하고, 훈련에 참여하지 않은 남여 10명씩의 음성을 인식을 위한 데이타로 사용하였다. 인식실험을 위한 문장의 구성은 다음과 같고, 사람이 들어도 구별할 수 없는 숫자음의 발음은 실험 대상에서 제외하였다.

문장1. 제 비밀번호는 OOOO 입니다.
문장2. 제 계좌번호는 OOO (다시) OOO (다시) OOOO 입니다.
문장3. 제 전화번호는 OOOOOOOOOO 입니다.

위의 문장은 발성한 사람에 따라 약간의 차이가 있으며, 되도록 또박또박 읽은 문장에 대해 실험을 했다. 다음의 표는 인식결과에 대한 결과이다.

[표4] 실험 결과 평가 [단위:%]

		가	나	향상률	오차의 원인
남	문장1	84.2	89.3	5.1	인식기 부분
	문장2	82.7	89.4	6.7	포만트 주파수 오차
	문장3	79.3	82.3	3.0	포만트 주파수 오차
여	문장1	85.0	89.0	4.0	인식기 부분
	문장2	81.0	88.9	7.9	인식기 부분
	문장3	77.3	80.8	3.5	인식기 부분
평균		81.6	86.6	5.0	

[표4]의 '가'는 가비지 모델을 사용하고, 포만트 주파수를 전혀 고려하지 않은 인식률이며, '나'는 포만트 주파수를 이용한 시스템의 인식률이다. 오차의 원인은 대부분 숫자음 인식기에서 나오고 있으며, 훈련 데이타의 부족이 가장 큰 원인으로 분석된다.

5. 결론 및 향후 연구과제

본 논문에서는 음성 파형의 음향 특징인 포만트 주파수와 한국어 숫자음의 언어학적인 연구를 바탕으로 대화체내에서 한국어 숫자음만을 검출하여 인식하는 시스템을 구현하고, 숫자음 인식에 대한 전체 성능을 평균 5.03%향상시켰다. 향후 연구 과제로는 숫자음의 검출 부분에서 발생되는 오류의 감소와 숫자음 인식이 주로 채널상의 음성을 대상으로 한다는 점을 감안하여 채널 잡음에 강인한 특징 추출 방식의 도입이 필요하다.

참고문헌

[1]. L. R. Rabiner, *Fundamentals of Speech Recognition*, pp.69-140, pp.390-433, Prentice Hall, Englewood cliffs, N.J., 1993.

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

[3]. Antonio Marcos de Lima Araujo, Fabio Violaro, "Formant Frequency Estimation Using a Mel Scale LPC Algorithm," SBT/IEEE Iternational Telecommunications Symposium, vol.1, pp.207-212, 1998.

[4]. L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, pp.38-115, Prentice Hall, Englewood cliffs, N.J., 1978.

[5]. J. D. Markel, A. H. Gray, Jr., *Linear Prediction of Speech*, pp.164-189, Springer-Verlag, Berlin Heidelberg N.Y. 1976.

[6]. 박찬규, 이홍규, 우정원, 이정현, "핵심어 검출을 위한 언어학적 지식에 기반한 음향적 음성 특징," 정보처리학회 추계 학술발표논문집 제4권 2호, pp.1217-1222, 1997.