정규화된 Mel-cepstrum을 이용한 숫자음 인식성능 향상에 관한 연구

이기철°, 최승호, 홍형진, 김기철, 이황수
한국 과학 기술원 정보및 통신 공학과

An Improved Digit Recognition using Normalized mel-cepstrum

Kee-Cheol Lee°, Seung-Ho Choi, Hyong-Jin Hong, Ki-Chul Kim, and Hwang-Soo Lee

요약

음성은 화자의 상태 및 주변 환경에 따라 그 특징이 다양하게 변화한다. 본 논문에서는 음성신호의 특징 파라미터로 널리 쓰이고 있는 mel-cepstrum에 대해, 단어내에서의 변화를 정규화함으로써 인식성능을 향상시키고자 하였다. mel-cepstrum이란 단어 전체에 대한 mel-cepstrum의 평균 값으로 normalize 시킨것이다.

한국어 숫자음에 대한 인식 실험결과, 본 논문에서 제안한 정규화된 mel-cepstrum이 정규화되지않은 mel-cepstrum에 비해 우수한 인식 성능을 나타내었다. 또한 잡음 환경하에서 비교 실험한 결과에서도 상대적으로 우수한 인식률을 보였다.

I. 서론

최근 과학기술 분야에서의 눈부신 발전과 정보화 시대의 도래에 따라 인간과 기계간의 자유로운 의사 소통이 무엇보다도 절실히 요구되고 있다. 음성은 인간이 의사 전달을 하기위해 쓰여지는 방법중에서 가장 자연스러운 매개 수단이라 할수 있다. 따라서 음성을 이용한 기계와의 대화는 키보드나 마우스와 같은 번거로운 수단을 사용하지 않고도 자연스럽고 빠른 일 처리를 가능하게 해준다. 이를 위해서는 음성인식이 무엇보다도 가장 먼저 해결해야할 과제임을 부인할 수 없다. 그러나 많은 기술적인 진보에도 불구하고 현재까지 완전한 음성인식기를 개발할 수 없었던 것은 음성자체에 내재된 독특한 특징때문이라고 할수 있다. 그 특징을 나열하면 다음과 같다.

첫째로 음성은 공간적인 영향을 받는 특징을 지닌다. 말하는 화자의 장소에 따라 같은 사람의 말일지라도 다른 특징벡터열로 표시될 수 있다. 이것은 주위에서 발생하는 잡음에 기인한 결과이다.

둘째로, 같은 화자라 할지라도 그날의 기분과 상태에 따라 같은 단어에 대한 발음의 억양과 액센트가 변하게 된다. 그리고 같은 화자의 발음에서도 매 발음의 속도와 방법이 항상 일정하다고 할수 없다. 이와같이 음성은 단일 화자에 대해서도 많은 변화요소를 지니고 있는데, 임의의 화자가 발음하는 음성을 고려한다면 연령과 성별, 방언 등에 따라 더욱 크게 변화한다는 난이점을 가지게 된다. 따라서 인식률의 저조성을 발생시키는 위의 요소들에 대처할수 있는 방법의 개발이 시급한 실정이다.

본 논문에서는, 첫번째로 예를 든 실제적인 잡음 환경하에서도 인식률이 저하되지 않는 특징 파라미터를 추출하는 방법에 대해 주안점을 두었다. 이를 위해 현재 음성인식에서 특징벡터를 추출하기 위해 널리 사용되는 mel-cepstrum을 이용하여 잡음에 따른 인식률의 변화상태를 살펴보았고, 기존의 mel-cepstrum을 변형한 normalized mel-cepstrum으로 인식률의 변이를 비교·평가하는 실험을 수행하였다.

mel-cepstrum은 음성신호에서 추출된 선형 예측(Linear Predictive Coding, LPC) 계수를 이용해서 cepstrum을 구한후, 귀의 비선형적인 인식특성을 고려하여 mel-scale로 Warping시킨 mel-scaled cepstrum이라 할수 있다.

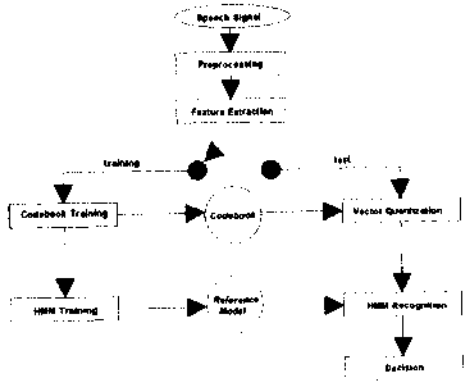
그런데 mel-cepstrum은 주위의 잡음환경에 따라 매우 민감한 인식률의 변화를 일으키는 단점을 가지고 있으므로, 이를 개선하고자 하는 노력이 최근에 많은 논문에서 발표되고 있다. 예를 들면 IMELDA(Integrated MEl-scale Linear Discriminant Analysis)와 같은 특징 파라미터는 modified contrast normalization이라는 방법을 이용하여 cepstrum값을 변환시킴으로써 잡음에 좀더 강력한 인식성능을 발휘하고 있다. 본 연구에서는 단어내에서의 변화를 정규화시켜 인식률의 향상을 꾀하고자 하는 방법을 시도하였다.

논문의 구성은, 2 절에서 인식시스템의 개요 및 구성을 먼

처 살펴보고, 3 절에서는 기존의 mel-cepstrum 추출과정과 정규화된 mel-cepstrum을 이용한 특징 추출방법을 설명하고, 4 절에서는 두가지 특징 파라미터를 통해 실험된 결과를 가지고 성능 비교를 하였다. 마지막 5 절에서는 결론 및 추후 연구 방향을 정리하였다.

2. 인식 시스템의 개요 및 구성

음성인식을 하기 위하여 사용된 시스템의 처리과정은 그림 1 과 같다.



<그림 1> 음성인식 시스템의 처리과정

첫번째로, 인간의 음성이 microphone을 통해서 전기적 신호로 바뀌어진 연속파형을 인식시스템에 적합한 신호로 변환하기 위한 전처리 과정을 거치게 된다. 여기에는 LPF(Low-Pass Filtering), Sampling, A/D변환과 음성신호를 묵음구간으로부터 분리해내는 끝점검출 과정등을 포함하고 있다.

본 시스템은 음성신호를 10 kHz, 12 bit로 샘플링하여 고주파 영역을 강조하기 위한 preemphasis를 거친후 30 msec Hamming window를 사용하여 분석구간을 정하였고 10 msec씩 전이하였다.

두번째로, 전처리 과정에 의해 변형된 음성 신호는 특징추출 과정(feature extraction)을 통과한다. 여기서 분석 구간별로 음성의 특징을 표현하는 특징 파라미터를 구하게 된다. 음성의 특징 파라미터 종류는 무척 다양하나, 본 논문에서는 가장 널리 쓰이는 mel-cepstrum을 사용하였고, 아와 함께 제안한 normalized mel-cepstrum으로 또다른 특징 파라미터를 추출하였다.

세번째로, 학습단계에서는 위에서 구한 특징 파라미터를 가지고 modified K-means clustering algorithm을 이용하여 64 개의 대표 패턴으로 양자화한 코드북(codebook)을 만들었다. 단어모델은 Baum-Welch reestimation algorithm으로 학습시킨 descrete HMM을 이용하였다.

네번째로, 인식단계에서는 각 모델들에 대한 관측열의 생성확률을 viterbi algorithm을 이용하여 구하였다.

끝으로, 최종단계에서는 학습에 의해 얻어진 기준 모델과의 비교를 통해 주어진 결정법칙에 따라 인식된 단어를 출력하게 된다.

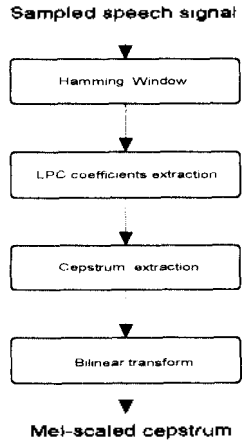
3. 정규화된 mel-cepstrum

3.1 mel-cepstrum

음성신호의 분석을 위해 사용되는 특징 파라미터는 무척 다양하다. 예를들면 에너지(energy), 영교차율(zero-crossing rate), 피치 주기(pitch period), 포먼트(formant), 단구간 스펙트럼(short-time spectrum), 선형예측 계수(LPC Coefficient), LSP 계수(Line Spectrum Pair Coefficient), PARCOR 계수등이 있다. 이중에서 현재 가장 널리 쓰이는 특징 파라미터로 mel-cepstrum을 꼽을 수 있다.

본 논문에서의 mel-cepstrum은 선형예측 계수로부터 구한 cepstrum값을 귀의 비선형적인 특성을 고려하여 mel-scale로 변형시킨 것이다.

그림 2 는 본 실험에서의 mel-cepstrum 추출과정을 나타내고 있다.



<그림 2> Mel-Cepstrum 추출 과정

첫번째로, 분석 구간별로 Sampling된 음성신호에서 LPC 계수를 추출하게 된다. 그 과정은 다음과 같다.

음성 신호의 스펙트럼은 all-pole filter (AR model)로 근사화한다 . 이때 LPC 계수는 Autocorrelation method를 이용하여 구한다.

두번째로, LPC 계수에서 Cepstrum을 추출해낸다. Cepstrum은 적당한 Liftering을 통하여 성도의 전달함수를 여기신호(excitation signal)로부터 분리 할수 있는 장점을 지니고 있다. 여기서 LPC계수는 아래와 같은 식을 사용하여 구해낼수 있다.

$$c_0 = E(0) = r(0),$$

$$c_1 = -a_1,$$

$$c_l = -a_l - \sum_{k=1}^{l-1} \frac{i-k}{i} c_{l-k} a_k, \quad i = 2,3,\cdots,p,$$

$$c_l = -\sum_{k=1}^{p} \frac{i-k}{i} c_{l-k} a_k, \quad i = p+1,\cdots$$

마지막으로 bilinear transform을 거치게 된다. LPC-cepstrum은 성도를 표현하는 우수한 특징 파라미터이지만 인간이 인식하게 되는 귀의 특성까지 고려하지는 않는다. 따라서 logarithmic scale로 왜곡시킨 특징 파라미터를 구함으로써 더 나은 인식성능을 나타낼 수 있다. 본 시스템에서 사용된 방법은 all-pass filter를 이용하여 frequency scale을 왜곡하는 bilinear transform을 채택하였는데 변환식은 다음과 같다.

$$z_{new}^{-1} = \frac{(z^{-1}-a)}{(1-az^{-1})}, \quad -1<a<1$$

$$w_{new} = w + 2\tan^{-1}\frac{(a\sin w)}{(1-a\cos w)}$$

where w : *normalized angular sampling frequency*

w_{new} : *converted frequency*

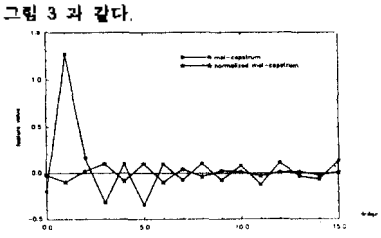
a : *frequency warping parameter*

주파수 변환 파라미터인 a 는 + 0.47 로 하여 bark scale(mel-Scale)이 되도록 하였다.

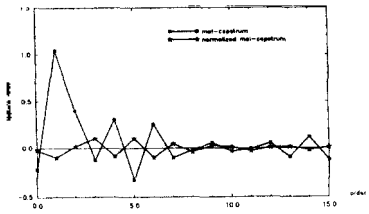
3.2 정규화된 mel-cepstrum

본 논문에서 제안한 정규화된 mel-cepstrum(normalized mel-cepstrum)이란 기존의 mel-cepstrum 값에서 단어 전체에 대한 mel-cepstrum의 평균값을 빼는 방법으로 normalize 시킨 것이다. 이를 실제 시스템에 적용시킨 절차를 살펴보면, 우선 기존의 cepstrum에서 분석 구간인 30 msec 별로 16 차의 melcepstrum을 구한후, 각 차수 별로 전체 프레임에 해당하는 Cepstrum의 종합을 얻어내어 이를 분석구간 횟수로 나눈다.

이렇게 얻어진 값은 차수별 Cepstrum의 총 분석 프레임당 평균값이 되므로, 이를 mel-scale 시킨후 기존의 mel-cepstrum 값들로 부터 감하게 되면 변형된 mel-cepstrum 값을 구하게 된다. mel-cepstrum과 변형된 mel-cepstrum의 변화를 단적으로 살펴보면 그림 3 과 같다.



(a) The first analysis frame



(b) The last analysis frame

〈그림 3〉 숫자 /이/ 에 대한 차수 별 feature value 비교(남성 화자 발음시)

위와 같이 변형된 mel-cepstrum을 normalized mel-cepstrum이라 칭하였고, 이것은 기존의 mel-cepstrum보다 동적 범위(Dynamic Range)가 줄어드는 효과를 나타낸다. 본 실험에서는 Distance measure로 Euclidean Distance를 사용하여 VQ(Vector Quantization) encoding을 하고있다.

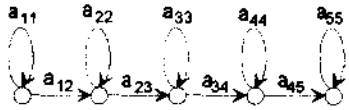
따라서 normalized mel-cepstrum은 각 차수별 mel-cepstrum 값을 변화시킴으로써 Euclidean distance 계산에 영향을 주게 된다. 즉 그림 3 에서처럼 각 차수별 크기의 차가 줄어들게 되므로 Euclidean distance 계산시 낮은 차수의 mel-cepstrum인 값에 편향적인 효과를 어느정도 제거할수가 있게 된다. 이는 distance measure에서 weighted cepstral distance를 사용하는 것과 유사한 작용을 하는 것으로 볼수 있다.

4. 실험 및 결과

4.1 실험 환경

본 연구에서는 남성화자 10 명, 여성화자 10 명이 10 회씩 발음한 10 개의 숫자음 / 일, 이, 삼, 사, 오, 육, 칠, 팔, 구, 공 /을 대상으로 인식실험을 수행하였다. 이중 남성화자 5 명, 여성화자 5 명이 각각 5 회씩 발음한 숫자음을 학습에 사용하였고, 학습에 참가하지 않은 남성화자 5 명, 여성화자 5 명이 10 회씩 발음한 숫자음으로 인식 테스트를 한 화자독립 방식을 채택했다.

그리고 단어모델로는 Left-to-Right 이산형 HMM을 사용하여 패턴을 모델링 했다.(그림 4)



〈그림 4〉 Left-to-Right HMM의 구조

4.2 실험 결과

3 절에서 설명한 방법으로 mel-cepstrum과 normalized mel-cepstrum을 구한 후 Clean Speech를 대상으로 첫 번째 인식 실험을 한 결과를 표 1 에 나타내었다. 그리고 표 2 는 오인식 숫자음 리스트를 나타낸다.

Feature 종류 / 숫자음	mel-cepstrum (%)	normalized mel-cepstrum (%)
/일/	80	96
/이/	94	88
/삼/	100	100
/사/	82	100
/오/	99	85
/육/	100	99
/칠/	99	99
/팔/	98	93
/구/	97	96
/공/	86	99
총 인식률	93.5 (%)	95.5 (%)

<표 1> mel-cespstum과 normalized mel-cepstrum간의 Clean Speech 인식 테스트

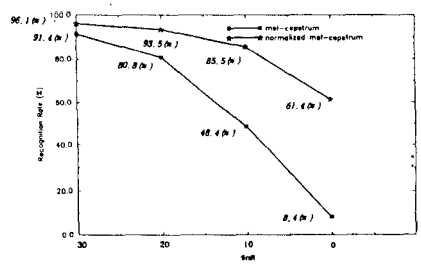
Result / Test	/일/	/이/	/삼/	/사/	/오/	/육/	/칠/	/팔/	/구/	/공/
/일/		1 (3)			(5)	9	8	3		
/이/	6 (5)				(7)					
/삼/										
/사/			18							
/오/	(2)	(11)							(2)	1
/육/				(4)						
/칠/	1		(1)							
/팔/			2 (4)				(2)		(9)	
/구/	1	(2)			2					(2)
/공/	(1)		1		4	4			5	

<표 2> mel-cepstrum 오인식 숫자음 리스트 (괄호안은 normalized mel-cepstrum 오인식 숫자음 갯수임)

표 에서 보듯이 숫자음 별로 인식률의 변화가 발생했지만 전체적으로 볼때 인식 성능이 향상되었음을 알 수 있다.

두 번째 실험은 Clean Speech에 White Noise를 첨가한 SNR 별 Noisy Speech를 가지고 인식률을 비교 평가해 보았다. 이때 코드북은 Clean Speech를 가지고 만든 것을 이용하였고 학습에서도 역시 Clean Speech를 사용하였다.

그 결과는 그림 5 와 같다.



<그림 5> mel-cepstrum과 normalized mel-cepstrum의 SNR 별 Noisy Speech의 인식 테스트

위의 결과처럼 잡음이 첨가된 입력음성에 대해서도 SNR 30-SNR0 까지 normalized mel-cepstrum이 기존의 mel-cepstrum보다 더 나은 인식률을 나타냄을 알 수가 있다.

따라서 잡음 환경하에서도 normalized mel-cepstrum이 더 좋은 인식성능을 보임을 실험결과를 통해 증명할수 있었다..

V. 결론 및 검토

본 논문에서는 단어 전체에 대한 mel-cepstrum의 평균값으로 정규화 시킴으로써 한국어 숫자음에 대한 인식 성능의 향상을 가져 왔다. 그리고 잡음이 첨가된 음성 신호에 대해서도 기존의 mel-cepstrum보다 우수한 인식률을 나타냈다.

앞으로의 연구방향은 백색 잡음뿐만이 아니라 여러 잡음 환경하에서도 인식 성능이 우수한지를 비교해보고, 더 나은 성능 향상을 위해서 특징 파라미터를 개선해 나가는데 중점을 두고자 한다.

[참고 문헌]

[1] Yifan Gong and William C. Treurniet, *"Speech Recognition in Noisy Envirnments : A Survey"*, CRC-TN 93-002, pp. 3-12, June, 1993

[2] D. C. Bateman, D. K. Bye, and M.J. Hunt, *"Spectral Contrast Normalization and other techniques for Speech Recognition in Noise"*, Proc. of the ICASSP, pp. 241-244, 1991

[3] Melvyn J. Hunt and Claude Lefebvre, *"A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech"*, Proc. of the ICASSP, pp. 262-265, 1989

[4] 최승호 외, *"음성 인식 다이얼링 시스템 개발연구"*, 한국과학 기술원, 한국 이동통신 위탁연구 과제 최종 보고서, pp. 3-24, Aug. 1993

[5] Alejandro Acero, *"Acoustrical and Envirnmental Robustness in Automatic Speech Recognition"*, Department of Electrical and Computer Engineering, Carnegie Mellon Univ., pp.35-59, September 13, 1990