인지 LPC cepstrum의 새로운 구현 및 음성인식에의 적용

A new Implementation of Perceptual LPC Cepstrum and its Application to Speech Recognition

김 진 영*, 최 승 호**
(Jinyoung Kim*, Seongho choi**)

*본 연구는 1995년도 전남대학교 자동차연구소의 공모과제 연구비에 의하여 연구되었음

요 약

본 논문에서는 귀의 주요한 특징인 주파수가중특성과 Bark-scale이라는 비선형주파수특성을 선형주파수축상에서 고려한 거리함수를 정의하고, 이 거리함수로부터 새로운 LPC cepstrum 계수를 제안한다. 귀의 특성은 선형주파수축에서 로그 스펙트럼에 대한 가중함수로서 표현되며, 이 가중함수는 cepstrum 영역에서 콘볼루션으로 표현되어 콘볼루션적으로 가중되는 LPC cepstrum을 정의하게 된다. 제안된 cepstrum 계수에서 정의된 가중함수는 A-weighting의 영향과 비선형주파수축의 영향을 하나의 가중함수로 통합하여 사용된 것이다. 제안된 파라미터의 성능을 음성인식 실험을 통하여 검증하였다.

ABSTRACT

To improve the performance of a recognition system, namely the recognition rate, we propose a hew implementation of perceptual distance using LPC cepstrum(perceptual cepstrum, PLC). The PLC is caculated by convolution of a usual LPC cepstrum and a perceptual lifter(PL). To caculate PL, we define a new weighting function in the linear frequency domain considering the frequency scale(Bark-scale) characteristics. The PL is the inverse Fourier transform of the exponents of the weighting function. We verified our method through the speech recognition experiments. The performance of PLC was compared with that of the rasied sine liftering method.

I. 서 론

음성인식에 있어 일반적으로 사용되고 있는 파라미터로는 filter bank를 사용한 방법과 LPC를 사용한 방법으로 크게 나누어 볼 수 있다[1-5]. 이들 파라미터는 신호처리적 입장에서 정의되는 대로 사용되지 않고 귀의 인지적의 특성을 고려하여 사용하는 것이 음성인식에 있어서 일반적인 추세이다[6-9]. 예를 들어 filterbank를 사용한 파라미터로는 mel-cepstrum이 있는데, 이는 filter bank를 귀의 비선형주파수축 특성의 근사인 mel-scale 상에서 구현하고 IFFT를 사용하여 정의되는 파라미터이다. 한편 LPC 계통의 음성인식파라미터로는 PLP(Perceptual linear prediction)과 이선형변환(bilinear transform)을 이용한 cepstrum 계수가 있다. 그러나 PLP의 경우에는 입력된 신호를 물리적으로 귀의 특성에 맞는 주파축으로 변환한 후 다시 이를 역푸리에 변환을 통해 자기상관계수를 구하여 LPC 분석을 하는 관계로 인하여 많은 계산량을 필요로 하는 단점이 있다. 또한 이선형변환(bilinear transfrom)을 사용하는 cepstrum 분석에서는 이선형변환을 이용하기 때문에 귀의 비선형 주파수축 특성을 근사적으로 반영한다는 단점이 있으며, A-weighting과 같은 가중특성을 동시에 구현할 수 없다는 단점을 지니고 있다.

본 연구에서는 지금까지 귀의 특성을 고려한 음성인식용 파라미터로서 사용되고 있는 LPC계 분석의 단점을 보완하기 위한 새로운 구현 방법으로서 cepstrum영역에서 콘볼루션의 형태로 가중되는 인지 LPC cepstrum 계수를 제안하고자 한다. 이 cepstrum 계수는 A-weighting과 Bark-scale이라는 비선형 주파수축 특성이 동시에 한번의 계산을 통하여 반영되므로 매우 편리한 방법이라고 할 수 있다. 본 논문에서는 HMM을 사용한 화자독립 고립

*전남대학교 공과대학 전자공학과
Chonnam National University, Dept. of Electronics Eng.
**동신대학교 공과대학 전자공학과
Dongshin University, Dept. of Information & Communication Eng.
접수일자 : 1996년 6월 5일

단어 인식실험을 통하여 제안된 파라미터의 타당성을 검증하고자 한다.

Ⅱ. 인지 LPC cepstrum

인간의 귀는 사람에 따라 조금씩 다르지만 약 20Hz-20kHz영역의 소리를 들을 수 있으며 신경 시스템과 관련되어 주파수분석기와 같은 동작을 한다. 사람이 소리를 듣는 과정은 귓불이 소리를 모아 이도(auditory canal)로 전달해서 고막을 진동시키게 되고 이 진동이 중이의 세 뼈를 차례로 거쳐 내이의 달팽이관에 이르게 된다. 달팽이관의 기저막에는 신경세포가 연결되어 있어 소리가 뇌에 전달되게 된다. 주파수 분석기로서 귀는 첫째, 비선형 스케일의 주파수축 둘째, 주파수가중 특성 그리고 마지막으로 배스킹 효과(masking effect)의 특성을 가지고 있다. 본 절에서는 이러한 귀의 특성을 고려한 인지거리함수를 귀의 비선형 스케일 축상에서 정의하고 이를 다시 선형주파수축으로 변환하여 로그스펙트럼에 대한 가중함수를 정의하고 정의된 가중함수로부터 cepstrum 영역에서 효율적으로 구현하기 위한 방안에 대하여 논한다. 물론 도입한 가중함수는 주파수가중 특성과 비선형스케일의 주파수 축 특성을 함께 고려한 것이다.

Ⅱ-1. 인지거리함수

일반적으로 음성신호처리에 있어, 두 스펙트럼의 차이는 로그 스펙트럼의 차이로서 정의되는데, 두 스펙트럼 $A(f)$와 $B(f)$의 거리는 다음과 같다.

$$DIST(A, B)=\int(\log A(f)-\log B(f))^2\, d2\pi f \qquad (1)$$

그런데, 인간의 귀는 소리를 분석함에 있어, 선형주파수축에서 분석하는 것이 아니라 비선형주파수축상에서 분석하는데, 이를 나타내는 것이 일반적으로 알려진 Bark-scale이다. Bark-scale의 축을 z이라고 하면, z과 선형주파수축 f와는 다음과 같은 관계를 갖는다[10].

$$z=7\cdot\log((f/650)+((f/650)^2+1)^{0.5}) \qquad (2)$$

한편, 귀는 모든 주파수축에서 같은 정도의 감도를 느끼는 것이 아닌데, 이를 귀의 주파수 가중특성이라고 한다. 지금까지 사용되는 주파수 가중방법은 A, B, C, D의 네 가지 방법이 제안되어 사용되고 있는데 B, C, D는 특수한 용도(예를 들면 음향기기 성능 테스트)에 사용되는 특성이며, A-weighting이 일반적인 귀의 특성을 나타내는 것이다. A-weighting은 40 phon의 소리를 기준으로 하여 구한 것이며 phon은 loudness level의 단위로 주파수 1kHz에서의 intensity level과 같다. 이 가중특성을 $W(f)$라 하자. 그러면 귀의 특성을 고려한 두 스펙트럼의 차이함수(distance function)는 다음과 같이 정의되어야 한다.

$$DIST(A, B)=\int(\log A(z)-\log B(z))^2\, W(z)dz \qquad (3)$$

단, $A(z)$과 $B(z)$은 z상에서의 스펙트럼
$W(z)$는 z상에서의 A-weighting

그런데, $dz=df/f'(z)$이므로 z을 f축으로 변환하면, 다음과 같은 식을 얻을 수 있다.

$$DIST(A, B)=\int(\log A(f)-\log B(f))^2\, W(f)/f'(z(f))\, d2\pi f \qquad (4)$$

여기서, $f'(z(f))=$

$$\frac{7}{650}\cdot\frac{1}{\cosh(\ln(f/650+((f/650)^2+1)^{0.5}))} \qquad (5)$$

이때, $U(f)=W(f)/f'(z(f))$라 하면, $U(f)$는 선형주파수축상에서 새로운 가중함수가 된다. 즉,

$$DIST(A, B)=\int(\log A(f)-\log B(f))^2\, U(f)\, d2\pi f \qquad (6)$$

과 같이 표현되는데, 이 새로운 가중함수는 귀의 비선형 특성과 A-weighting을 함께 고려한 새로운 가중함수가 된다. 한편 본 논문에서 A-weighting이라는 가중함수를 사용하였으며 따라서 음성의 분석시에는 pre-emphasis과정이 생략되어진다. 만약 A-weighting과정 대신에 pre-emphasis를 사용하고 싶다면, 이의 주파수특성을 $W(f)$대신에 대입하여 구하면 pre-emphasis과정을 통하지 않고 pre-emphasis와 비선형 주파수 스케일 특성이 통합된 가중함수를 구할 수 있다.

Ⅱ-2. 인지 LPC cepstrum

식 (6)으로 주어진 인지거리함수를 스펙트럼 영역에서 구현하기 위해서는 스펙트럼을 실제 계산해야 하는 문제가 있을 뿐 아니라 $U(f)$가 로그스펙트럼에 대한 가중함수이므로 스펙트럼에 로그를 취하는 과정이 필요하여 많은 계산량을 요구하게 된다. 또한 LPC 분석을 한 경우에는 FFT와 log를 취하는 과정이 필요하고 이를 다시 cepstrum 계수로 바꾸기 위해서는 IFFT의 과정이 필요하므로 매우 비효율적인 구현방법이 된다. 따라서 효율적인 구현 방법이 요구되는데 이를 위하여 식 (6)으로 주어진 인지거리함수를 위하여 다음의 식으로 반환하여 보자.

$$DIST(A, B)=\int(U(f)^{0.5}\log A(f)-U(f)^{0.5}\log B(f))^2\, d2\pi f \qquad (7)$$

위의 식에서 $U(f)^{0.5}\log A(f)$를 생각하여 보자. 이를 달리 표현하면 다음의 식과 같이 된다.

$$\log A'(f)=U(f)^{0.5}\cdot\log A(f)=\log(\exp(U(f)^{0.5})\cdot\log A(f) \qquad (8)$$

윗식의 $A'(f)$는 새로의 귀의 특성을 고려한 스펙트럼이 된다. 만약 스펙트럼 $A(f)$가 선형예측에 의하여 구한 스펙트럼이라고 하면, 로그스펙트럼 $\log A'(f)$의 푸리에 변환후의 cepstrum 계수는 다음과 같이 구할 수 있다.

$$C'_A(n) = C_A(n) \otimes C_U(n) \tag{9}$$

여기서, $C_A(n)$은 $\log A'(f)$의 푸리에 변환, $C_A(n)$은 $\log A(f)$의 푸리에 변환, 그리고 $C_U(n)$은 $U(f)^{0.5}$의 푸리에 변환이며, 모두 cepstrum 영역이고 n은 quefrency이며 ⊗은 콘볼루션이다.

식(9)를 고려하여 보면, 식(3)의 거리함수는 귀의 특성을 고려한 인지 cepstrum 계수들의 유클리디안 거리임을 알 수 있다. 그림 1은 $C_u(n)$을 보이고 있다.

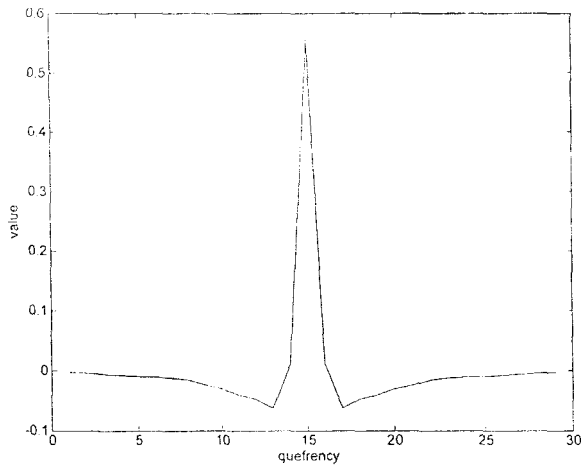


그림 1. 셉스트럼영역에서의 인지가중함수
Fig 1. Perceptual weighting function in cepstrum domain

Ⅱ. 실험 및 결과

제안된 cepstrum계수의 음성인식 상에서의 성능을 평가하기 위하여 본 연구에서는 벡터양자화기(VQ)와 hidden Markov model(HMM) 방법의 음성인식 시스템을 구현하였다[11]. 음성인식은 '공'부터 '구'까지의 한국어숫자음 10개를 대상으로 하였다. 인식용 DB로는 발성화자 총 24명으로부터 2회씩 녹음된 숫자음이 사용되었는데, 8kHz로 샘플링하였고 매 12.8ms 마다 12차의 LPC 분석을 25.6 ms의 프레임단위로 수행하였다. 인지 cepstrum계수는 20차까지 사용하였다. 한편 VQ를 위한 codebook은 128개의 코드워드로 이루어져 있으며 단어단위의 HMM 모델을 사용하였는데 3개의 상태를 사용하였다. 본 연구에서는 closed test와 open test의 두 방법을 사용하여 인식 실험을 수행하였다. Closed test의 경우, 학습은 2회씩 발음한 것 중 하나씩을 뽑아 단어당 총 24개를 사용하였으며 인식실험에서는 구축된 DB를 모두사용하여 인식실험을 수행하였으며, open test의 경우 10명의 화자를 사용하여 학습된 후 나머지 14명의 화자를 인식실험에 적용하였다. 본 논문에서 제안된 인지 LPC cepstrum의 성능을 평가하기 위한 비교 파라미터로서 인식률이 높다고 알려져 있는 liftering을 이용한 LPC cepstrum을 사용하였다. 물론 기준 인식실험에는 lifter를 사용하지 않는 LPC cepstrum를 사용하였다. Lifter로서는 raised sine 함수를 사용하였는데 다음과 같다.

$$w(n) = \begin{cases} 1 + h\sin(\frac{n\pi}{L}) & \text{for } n = 1, 2, \cdots, L \\ 0 & \text{for } n \le 0,\ n > L \end{cases} \tag{10}$$

한편, rasied sine lifter의 장점과 인지 cepstrum의 장점을 결합한 형태의 인식실험을 수행하였다. 이는 cepstrum영역에서 lifter를 통과시킨 후 convolutional weighting을 가하여 인지 cepstrum을 구하는 것이다. 표 1은 실험결과를 보여준다.

표 1. 음성인식 실험결과
Table 1. Speech recognition experiment results

	Closed test	Open Test
Baseline	95.42	79.29
Raised sine lifter (12차)	96.46	82.50
Convolutional Weighing(CW)	96.25	85.36
Lifter + CW	97.08	85.71

표 1의 실험결과를 살펴보면, closed test의 경우 lifter를 사용한 인식 결과와 인지 cepstrum을 사용한 실험결과가 비슷한 인식률을 보이고 있으며, 물론 LPC cepstum을 사용한 경우보다는 높은 인식률을 나타내었다. 최고의 인식률은 lifter와 인지 cepsteum을 결합한 경우가 가장 높은 인식률을 보인다. Open test의 경우도 closed test와 비슷한 현상을 보이고 있다. 그러나 LPC cepstum을 사용한 경우와 타방법과의 차이가 크게 나타남을 알 수 있다. 또한 lifter을 사용한 경우보다 인지 cepstrum을 사용한 경우 약 3%정도의 인식률 향상을 확인할 수 있었다. Open test의 경우에도 lifter와 인지 cepstrum을 사용한 경우가 가장 높은 인식률을 보였다. 위의 실험결과를 통하여 본 논문에서 제안한 파라미터의 타당성을 검증할 수 있다.

Ⅳ. 결 론

본 논문에서는 주파수가중 특성과 비선형주파수 스케일 이라는 귀의 인지적 특성을 고려한 거리함수로부터 새로운 인지 LPC cepstrum를 제안하였으며 숫자음을 대상으로 한 화자독립 음성 인식 실험을 통하여 그 성능을

검증하였다. 비교 대상으로는 LPC cepstrum보다 높은 인식률을 보이는 것으로 알려진 liftering을 이용한 LPC cepstrum을 사용하였다. 인식실험결과로부터 인식률이 향상됨을 검증할 수 있었다. 그러나 앞으로 더 많은 화자 및 데이타를 통하여 검증하여야 하겠으며, 최적의 cepstrum 차수에 대한 연구도 진행되어야 하겠다.

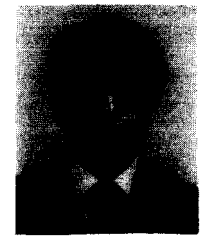
참 고 문 헌

1. B.A. Dautrich, L.R. Rabiner and T.B. Martine, "On the effects of varying filter bank parameters on isolated word recognition," IEEE Trans. on ASSP, ASSP-31(4), pp. 793-807, August, 1983.
2. G.M. White and R.B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," IEEE Trans. on ASSP, ASSP-24(2), pp.183-188, 1976.
3. J.D. Markel and A.H. Gray, Linear Prediction of Speech, Springer-Verlag, 1982.
4. S. Saito, Fundamentals of Speech Signal Processing, Academic Press, 1985.
5. J.L. Flanagan, Speech Analysis Synthesis and Perception, Springer-Verlag, 1972.
6. R.A. Bladon and B. Lindblom, "Modeling the judgement of vowel quality difference," J.A.S.A., 69(5), pp.1414-1422, May, 1981.
7. H. Harmanskey, "Perceptual Linear Predictive(PLP) analysis of speech," J.A.S.A., 87(4), pp.1738-1752, April, 1990.
8. N. Nocerino, F.K. Soong, L.R. Rabiner and D.H. Klatt, "Comparative study of several distortion measures for speech recognition," Speech Communication, 4, pp.317-331, 1985.
9. K.F. Lee, Automatic Speech Recognition : The Development of the SPHINX system, Kluwer Academic Publishers, Boston, 1989.
10. B.C. Moor and B.R. Glaberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation pattern," J.A.S.A., 74(3), pp.750-753, 1985.
11. S.E. Levison and L.R. Rabiner, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," BSTJ, vol.62, No.4, pp.1035-1073, April, 1983.

▲김 진 영(Jinyoung Kim) 1962년 4월 26일생

1986년 2월 : 서울대학교 전자공학과 졸업
1988년 2월 : 서울대학교 대학원 전자공학과 졸업(석사)
1994년 8월 : 서울대학교 대학원 전자공학과 졸업(박사)
1993년 1월~1994년 12월 : 한국통신 소프트웨어연구소 연구원
1995년 1월~현재 : 전남대학교 전자공학과 전임강사
※주관심분야 : 음성인식, 음성합성 등 음성신호처리

▲최 승 호(Seungho Choi) 1955년 8월 24일생



1981년 2월 : 전북대학교 물리학과 졸업(학사)
1984년 8월 : 명지대학교 대학원 전자공학과 졸업(석사)
1992년 2월 : 명지대학교 대학원 전자공학과 졸업(박사)
1992년 2월~현재 : 동신대학교 정보통신공학과
※주관심분야 : 디지탈 신호처리, 음성인식 및 합성