실시간 고차통계 정규화를 이용한 강인한 음성인식*

정주현(부산대), 송화전(부산대), 김형순(부산대)

<차 례>

<Abstract>

Robust Speech Recognition Using Real-Time Higher Order Statistics Normalization

Ju-Hyun Jeong, Hwa-Jeon Song, Hyung Soon Kim

The performance of speech recognition system is degraded by the mismatch between training and test environments. Many studies have been presented to compensate for noise components in the cepstral domain. Recently, higher order cepstral moment normalization method has been introduced to improve recognition accuracy. In this paper, we present real-time high order moment normalization method with post-processing smoothing filter to reduce the parameter estimation error in higher order moment computation. In experiments using Aurora2 database, we obtained error rate reduction of 44.7% with proposed algorithm in comparison with baseline system.

* Keywords: Robust speech recognition, Cepstrum, Higher order statistics, Real-time processing.

* 이 논문은 산업자원부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활 지원 지능로봇 기술개발사업)의 일환으로 수행됨.

1. 서 론

주변 잡음과 채널 특성으로 인한 훈련환경과 인식환경 사이의 불일치는 음성 인식기의 성능을 저하시킨다. 이러한 불일치를 극복하기 위해 다양한 전처리 방법이 시도되고 있으며, Cepstral Mean Subtraction(CMS)를 비롯한 켑스트럼 영역의 정규화 방법이 널리 사용되고 있다. CMS 방법은 캡스트럼 영역에서 채널 특성을 포함하는 바이어스를 제거해줌으로서 채널 특성에 덜 민감하고 주변잡음에 강인한 특성을 가지게 하며, Cepstral Variance Normalization (CVN)은 분산을 정규화 함으로서 잡음음성의 확률분포를 원음성의 확률분포에 가깝게 해준다[1]-[4]. 그러나 캡스트럼 특징벡터에 대한 환경잡음에 의한 영향은 비선형적이기 때문에 선형적인 모델로서 잡음영향을 제거하기에는 어려움이 있다. 비선형 정규화 방법으로 CDF matching 방법[5]이 있으며, 최근에는 3차 모멘트나 그보다 높은 고차 모멘트를 정규화해주는 방법이 제안되었다[6][7]. 그러나 기존의 켑스트럼 영역 정규화 방법들은 입력음성이 모두 들어온 뒤에 특징벡터의 통계특성들을 추정하기 때문에 실시간 처리가 어려운 단점이 있다. 실시간 처리를 위한 CMS 계열의 방법으로는 Local CMS (LCMS) 방법과 Sequential CMS (SCMS) 방법이 있다. LCMS 방법과 SCMS 방법은 입력음성 전체를 이용하는 CMS 방법, 즉, Global CMS (GCMS) 방법에 비해 성능이 떨어진다. 본 논문에서는 LCMS 방법의 단점을 보완하고 성능향상을 위해 고차통계를 이용한 실시간 정규화방법을 도입하고, 또한 고차통계의 추정오차 문제를 완화시키기 위해 smoothing filter를 추가적으로 적용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 켑스트럼 영역에서의 여러 가지 정규화 방법들에 대해 기술한다. 3장에서는 본 논문에서 도입한 실시간 고차통계 정규화 방법에 대해 설명하고 4장에서 실험 및 결과를 서술한다. 마지막으로 5장에서 결론을 맺는다.

2. 켑스트럼 영역에서의 정규화 방법

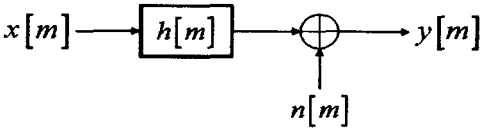
<그림 1>은 음성신호의 왜곡과정에 대한 모델이다. 잡음이 섞이지 않은 원음성 $x[m]$이 임펄스 응답 $h[m]$을 가지는 채널을 거쳐서 부가잡음 $n[m]$이 섞이면 왜곡된 음성 $y[m]$을 생성시킨다. 시간 영역에서 표현하면 다음과 같다.

$$y[m] = x[m] * h[m] + n[m] \tag{1}$$

식 (1)을 캡스트럼 영역에서 나타내면 다음과 같은 식으로 표현할 수 있다 [8].

$$y = x + h + C\ \ln(1 + \exp(C^{-1}(n - x - h))) \quad (2)$$

여기서, y는 관측 캡스트럼 벡터, h는 채널 성분의 캡스트럼 벡터, x는 입력음성의 캡스트럼 벡터, 그리고 C는 DCT 행렬이다. 부가잡음과 채널로 인한 바이어스를 제거해주면 인식성능의 향상을 기대할 수 있다.



<그림 1> 음성신호의 왜곡과정에 대한 모델

2.1. CMS (Cepstral Mean Subtraction)

식 (2)에서 부가잡음에 의한 영향을 무시하면 채널로 인한 바이어스인 h를 제거해 줌으로써 훈련환경과 인식환경의 차이를 보상할 수 있다. 순수한 음성의 캡스트럼 벡터의 평균이 0이라고 가정하면 바이어스의 추정치는 다음과 같이 구할 수 있다.

$$b_{CMS} = \frac{1}{N} \sum_{n=1}^{N} y\ (n) \quad (3)$$

여기서, $y(n)$은 n번째 프레임의 관측벡터이고, N은 관측벡터의 전체 프레임 수이다. CMS 방법에 의해 채널왜곡이 보상된 벡터 $x_{CMS}(n)$은 다음과 같다.

$$x_{CMS}(n) = x(n) - b_{CMS}(n) \quad (4)$$

2.2. CVN (Cepstral Variance Normalization)

음성인식 시스템에서 주변 잡음 환경의 변화에 따라 특징벡터의 평균 이외에 다른 통계적 특성도 달라진다. 이러한 통계적 특성을 보상해주는 한가지 방법으로 CVN 방법이 있다. CVN 방법은 CMS에 비해 원음성과 잡음음성사이의 확률밀도 함수의 차이를 더 줄이는 효과가 있다. 만약 각각의 벡터차원이 서로 독립이라고 가정한다면 CVN 방법은 식 (5)와 같다. CVN 방법은 CMS 방법을 통해 평균을 0으로 만든 후, 각 차원별 특징벡터의 표준편차로 나눠줌으로써 2차 모멘트인 분산

을 1로 정규화 해준다. $x(n)$의 N차 모멘트는 $E[x(n)^N]$으로 정의된다.

$$x_{CVN}(n) = x_{CMS}(n)/\sqrt{E[x^2_{CMS}(n)]} \tag{5}$$

2.3. CTN (Cepstral Third-order Normalization)

특징벡터에 대한 부가잡음의 영향은 평균이나 분산뿐만 아니라 확률분포의 고차 모멘트까지 영향을 준다. 따라서 3차 이상의 고차 모멘트에 대해서도 정규화가 필요하며, CTN 방법은 평균과 분산뿐만 아니라 3차 모멘트인 왜도(skewness)를 정규화 해준다[6]. 그러나 [6]에서 제시된 방법은 3차 방정식의 정확한 근을 찾기가 어려운 문제가 있어 단순히 근사식만을 이용하는 CTN 방법을 도입하였다[7]. 본 논문에서 사용한 CTN 방법은 CVN 과정을 거친 특징벡터를 이용해 식 (6)과 같이 정의된 비선형 변환으로 표현한다.

$$x_{CTN}(n) = a\,x^2_{CVN}(n) + x_{CVN}(n) + c \tag{6}$$

식 (6)에서 $x_{CTN}(n)$의 평균이 0이고 1인 분산을 가지며 3차 모멘트가 0이 되도록 a와 c를 정해야 한다. $x_{CVN}(n)$의 평균은 0이고 분산은 1이므로 a와 c는 다음의 관계를 가진다.

$$E[x_{CTN}(n)] = E[a\,x^2_{CVN}(n) + x_{CVN}(n) + c] = a + c = 0 \tag{7}$$

c와 $-a$는 같으므로 식 (6)을 다음과 같이 다시 쓸 수 있다.

$$x_{CTN}(n) = a\left(x^2_{CVN}(n) - 1\right) + x_{CVN}(n) \tag{8}$$

식 (7)을 이용해 x_{CTN}의 3차 모멘트를 정리하면 식 (9)과 같다.

$$\begin{aligned} E[x^3_{CTN}(n)] &= E[a(x^2_{CVN}(n) - 1) + x_{CVN}(n)]^3 \\ &= a^3 E[x^2_{CVN}(n) - 1]^3 + 3a^2 E[x^2_{CVN}(n) - 1]^2 x_{CVN}(n) \\ &\quad + 3aE[x^2_{CVN}(n) - 1]x^2_{CVN}(n) + E[x^3_{CVN}(n)] = 0 \end{aligned} \tag{9}$$

a는 분산을 1로 정규화한 후에 왜도(skewness)만을 보상해주므로 매우 작은 값을 가진다. a의 고차부분을 무시하면 a는 식 (10)과 같이 근사적으로 구할 수 있다.

$$a \approx \frac{-E[x^3_{CVN}(n)]}{3E[x^4_{CVN}(n) - x^2_{CVN}(n)]} \tag{10}$$

식 (10)은 근사화된 식이기 때문에 정확한 a를 추정하기 위해서는 반복된 계산과정이 필요하다.

3. 실시간 켑스트럼 영역의 정규화 방법

3.1. LCMS 및 SCMS

일반적인 CMS 방법은 전체 음성이 시스템에 입력된 후에 동작하므로 실시간 처리의 어려움이 있다. CMS 방법의 실시간 처리를 위해 제안된 방법 중 하나가 Local CMS (LCMS) 방법이며, 이는 입력음성 전체에 대해 평균을 취하는 대신에 적당한 길이의 구간에 대해 moving average를 취함으로써 채널성분을 추정한다. N_L이 moving average를 취하는 구간이면, 추정된 채널 성분은 다음과 같다.

$$b_{LCMS} = \frac{1}{N_L} \sum_{n'=0}^{N_L} y(n-n') \tag{11}$$

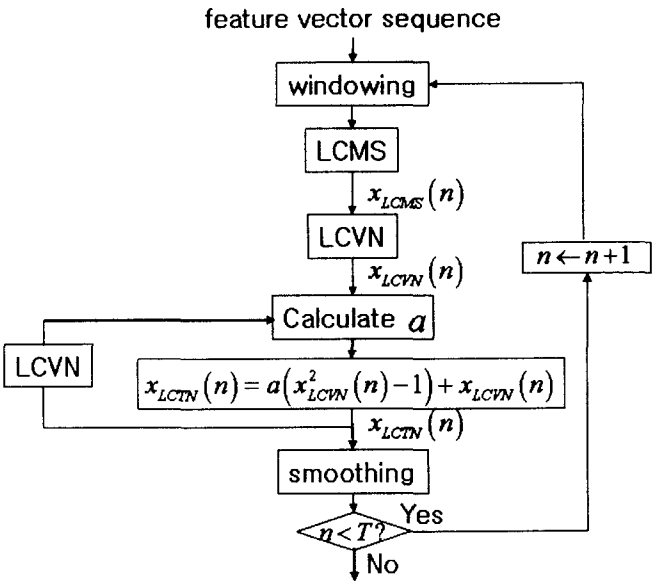
LCMS 방법의 변형된 형태인 Sequential CMS (SCMS) 방법은 식 (12)와 같이 n번째 켑스트럼 벡터와 $n-1$번째의 바이어스 추정벡터의 가중합으로 채널성분을 추정하는 방법이다.

$$b_{SCMS}(n) = \alpha\, b_{SCMS}(n-1) + (1-\alpha)\, y(n) \quad 0<\alpha<1 \tag{12}$$

여기서, $b_{SCMS}(n)$은 SCMS 방법에 따른 바이어스 추정벡터이고, $y(n)$은 관측 캡스트럼 벡터이다. 윈도우의 크기는 a에 따라 결정된다.

3.2. LCTN

본 논문에서는 고차통계 정규화 방법의 실시간 구현을 위해 기존 LCMS 방법의 아이디어를 CTN 방법에 적용한 Local CTN (LCTN) 방식을 도입하였다. <그림 2>에 실시간 고차통계 정규화방법의 흐름도를 나타내었다.



<그림 2> 실시간 고차통계 정규화방법(LCTN)의 흐름도

실시간으로 3차 모멘트를 정규화 할 경우에 추정에 사용되는 데이터의 양이 작기 때문에 정확한 추정이 이루어지지 않아 정규화한 값들 중에서 outlier가 발생하는 문제점이 있다. 이러한 outlier를 제거하기 위해 LCTN 방법을 사용하여 나온 값들에 대해 smoothing filter를 적용함으로써 시간 축 상에서의 바이어스 변화량이 크지 않도록 하였다. 여러 가지 smoothing 필터 중 식 (13)과 같은 moving average를 이용한 경우가 성능이 가장 우수하였다.

$$\hat{x}(n) = \frac{1}{2M+1} \sum_{m=-M}^{M} x(n+m) \tag{13}$$

본 논문에서 M의 크기는 4를 사용하였다.

또한 제안된 LCTN 방법은 매 프레임마다 고차 모멘트 항을 구해야 하기 때문에 계산량이 많은 문제가 있다. 그래서 계산량을 줄이기 위해 식 (14)와 같이 이전 프레임에서 구한 모멘트 항에 현재 프레임의 모멘트 항과 윈도우 크기 N_L 이전 프레임에 대한 모멘트 항을 이용해 업데이트하는 방법을 사용하였다.

$$E[x^N(n)] = E[x^N(n-1)] + \frac{1}{N_L}\{x^N(n) - x^N(n-N_L)\} \tag{14}$$

여기서 $x^N(n)$은 특징벡터의 N 차승을 의미한다.

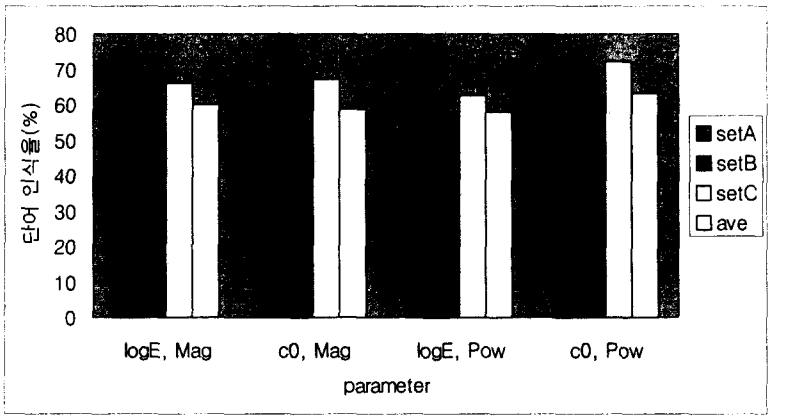
4. 실험 및 결과

4.1. 음성 데이터베이스

제안된 방법의 평가를 위해서 Aurora2 데이터베이스[9]가 사용되었다. Aurora2 데이터베이스는 1자리에서 7자리까지의 영어 연결숫자로 구성된 TI Digit에 다양한 잡음을 인위적으로 더한 것이다. Aurora2 데이터베이스는 훈련 데이터와 테스트 데이터로 구분되어 있으며 테스트 데이터는 채널 특성은 동일하고 서로 다른 잡음이 더해진 두 개의 subset(set A, set B)과 채널특성이 다른 subset(set C)으로 총 3개의 subset으로 구성되어 있다. 잡음환경은 8가지의 잡음종류(subway, babble, car, exhibition, restaurant, street, airport, station)와 각각 5가지 잡음 레벨(clean, 20dB, 15dB, 10dB, 5dB)로 구성되어 있다. 성능평가는 각 잡음의 종류에 대해서 20dB에서 0dB까지의 잡음 레벨에 대해 수행된다. 본 논문에서는 잡음이 섞이지 않은 깨끗한 음성에 대해서만 훈련을 하는 clean condition에 대해 실험을 수행하였다.

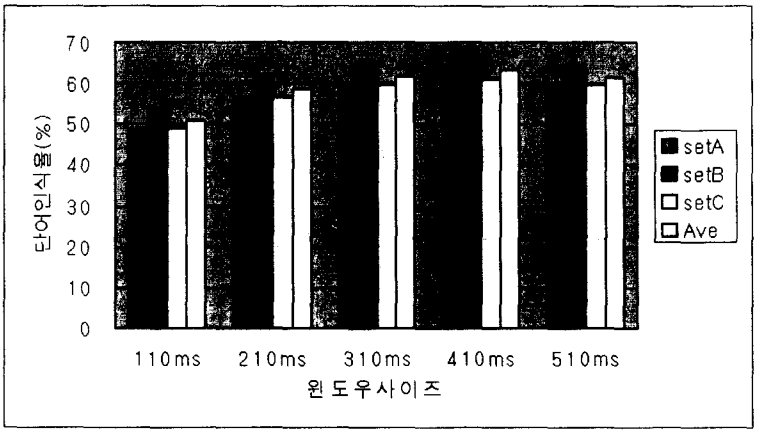
4.2. 실험결과

Aurora project는 다른 기관에서 제안된 방법을 비교할 수 있도록 baseline system의 성능을 제공한다. baseline system의 특징벡터는 23개의 mel frequency 삼각 필터로부터 추출된 MFCC (Mel Frequency Cepstral Coefficient) 계수를 사용한다. MFCC 계수는 magnitude spectrum으로부터 추출된 계수와 로그 에너지를 사용한다. 음성모델은 16개의 state를 사용하고 각 state는 3개의 diagonal Gaussian mixture를 가진다[9]. <그림 3>은 로그 에너지와 c0, 그리고 magnitude spectrum에서 구한 MFCC 계수와 power spectrum에서 구한 MFCC 계수의 조합에 따른 인식성능을 보여준다. 실험결과에서 c0와 power spectrum을 이용한 MFCC 계수를 사용한 경우가 인식률이 가장 좋음을 알 수 있다. 이후 실험에서는 성능이 가장 좋은 c0와 power spectrum을 이용해 구한 MFCC 계수를 사용하였다.



<그림 3> 에너지와 spectrum 종류에 따른 성능비교
(logE - 로그 에너지, Mag - magnitude spectrum, Pow - power spectrum)

<그림 4>는 윈도우 사이즈에 따른 LCMS의 인식률이다. LCMS 방법의 경우에 윈도우 사이즈가 커질수록 성능이 올라가는 특성을 보여주고 있다. 윈도우 사이즈를 410ms를 사용했을 경우의 성능이 가장 우수하므로 이후 실험에서는 410ms의 윈도우를 사용했다. 410ms의 윈도우 사이즈에서 SCMS 방법과 LCMS 방법을 비교한 경우 SCMS 방법이 LCMS 방법에 비해 성능이 우수하다. 하지만 SCMS는 고차통계정규화 방법을 적용하기가 어렵기 때문에, LCTN의 실시간 구현은 LCMS 방법을 기반으로 결과를 얻었으며 <표 1>에 실험결과를 나타내었다.



<그림 4> 윈도우 크기에 따른 LCMS 성능비교 (c0와 power spectrum 사용)

LCTN 방법을 사용하였을 때 baseline에 비해 38.06%의 오류감소율을 얻었다. 또한 outlier에 의한 왜곡을 감소시키기 위해 smoothing 필터를 사용하여 (<표 1>에서 LCTN [smoothing]) 44.70%의 오류감소성능을 얻었다. 이는 전체 음성에 대한

보상을 실시하는 GCMS 방법과 Global CVN (GCVN) 방법보다 우수한 결과이며, Global CTN (GCTN)보다는 성능이 떨어지지만, 전체 음성이 다 들어온 후에야 처리가 가능한 GCTN에 비해서 실시간 처리가 가능하다는 장점이 있다.

<표 1> 정규화 방법에 따른 성능비교 (c0와 power spectrum 사용)

실시간 전처리 방법	clean condition				
	set A	set B	set C	Ave	ERR
Baseline	61.34	55.75	66.14	60.06	0.00%
SCMS	61.41	66.68	62.13	63.66	9.01%
LCMS	60.77	66.14	61.23	63.01	7.38%
LCTN	73.90	76.43	75.68	75.26	38.06%
LCTN [smoothing]	76.89	78.73	78.33	77.91	44.70%
GCMS	66.36	71.43	67.20	68.55	21.26%
GCVN	75.08	75.92	76.38	75.68	39.09%
GCTN	80.71	82.32	81.32	81.48	53.62%

5. 결 론

본 논문에서는 강인한 음성인식을 위한 켑스트럼 영역에서의 정규화 방법들을 검토한 다음, 잡음보상의 실시간 처리를 위해 LCMS 방법의 아이디어를 고차통계 기반의 정규화 방법인 CTN 방법에 적용한 LCTN 방법을 도입하였다. 또한 LCTN 방법에서의 고차통계 추정오차 문제를 극복하기 위하여 smoothing filter를 추가적으로 사용하였다. Aurora 2 데이터베이스의 clean condition 환경에 대해 실험을 해본 결과 실시간 처리가 가능하면서도 Aurora 2 baseline 시스템에 비해 44.70%의 성능 향상율을 나타내었다.

참 고 문 헌

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *in Proc. of JASA*, Vol. 65, no. 6, pp. 1304-1312, 1974.

[2] O. Viikki, D. Bye and K. Laurila, "A recursive feature vector normalization approach for robust speec recognition in noise", *in Proc. of ICASSP*, Vol. 2, pp. 733-736, May 1998.

[3] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, Vol. 25, pp. 133-147, Aug. 1998.

[4] 김민성, 정성윤, 손종목 외, "채널보상기법을 사용한 전화 음성 연속숫자음의 인식 성능 향상", 말소리, Vol. 44, pp. 73-82, Dec. 2002.

[5] G. A. Saon and J. M. Huerta, "Improvement to the IBM Aurora 2 multi-condition system", *in Proc. of ICSLP*, pp. 469-472, Sep. 2002.

[6] Y. H. Suk, S. H. Choi and H. S. Lee, "Cepstrum third order normalization method for noisy speech recognition," *Electronic Letters*, Vol. 35, no. 7, pp. 527-528, Apr. 1999.

[7] C.-W. Hsu and L.-S. Lee, "Higher order cepstral moment normalization for robust speech recognition," *in Proc. of ICASSP*, Vol. 1, pp. 197-200, May 2004.

[8] X. Huang and A. Acero, H.-W. Hon, Spoken Language Processing : A Guide to theroy, Algorithm, and System Development, Prentice-Hall, 2001.

[9] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000*, Paris, Sep. 2000.

접수일자 : 2005년 5월 15일
게재결정 : 2005년 6월 20일

▶ 정주현(Ju-Hyun Jeong)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: jeongju78@pusan.ac.kr

▶ 송화전(Hwa-Jeon Song)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: hwajeon@pusan.ac.kr

▶ 김형순(Hyung Soon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr