제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

넥마이크로 입력된 음성 신호에 대한 인식 연구

이연철*, 이상운**, 홍훈섭**, 한문성**, 마평수**
*(주)휴먼미디어테크
**한국전자통신연구원
E-mail : yclee@e-human.co.kr

A Study on Speech Recognition for Neck-Microphone Input Signal

Yeon-Chul Lee*, Sahng-Woon Lee**, Hun-Sop Hong**,
Mun-Sung Han**, Pyong-Soo Ma**
*Human Media Tech. Inc.
**Electronics and Telecommunications Research Institute

요 약

본 논문에서는 일반적으로 사용되는 마이크가 잡음에 민감하여 음성인식기의 성능을 저하시키기 때문에 잡음의 영향을 받지 않는 고지향성을 가지는 넥마이크로 입력되는 음성신호에 대한 특성을 고찰하고 기존의 일반마이크 입력 음성을 이용하는 인식시스템에서의 인식성능을 살펴본다. 넥마이크는 일반마이크와 동일한 원리로 음성을 채집하며 목부위에 장착된다. 실험에서 넥마이크에 의한 음성은 일반마이크 입력 음성에 비해 인식 성능이 저하되는 결과를 보여주어 앞으로 새로운 인터페이스의 연구 대상으로 여겨진다.

1. 서론

음성인식시스템은 훈련단계에 사용된 음성데이터베이스와 동일한 환경과 잡음이 없는 환경의 입력 음성에서는 상당한 인식성능을 보인다. 하지만 입력 마이크가 달라지던지 또는 실제환경의 잡음이 존재하는 경우에 인식성능은 현저히 저하된다. 이러한 상황에 대처하기 위한 여러 가지 기술들이 연구되고 있다. 전처리 과정에서의 잡음제거 기술에는 음질향상법[1]과 신호원분리법, 강인한 특징벡터 추출[3], 채널잡음 제거법[4] 등이 있고 모델 차원에서의 보상시키는 기술로는 PMC[5]가 있고 또한 실제환경에 대한 모델 파라미터 적응법[6]이 있다.

본 논문에서는 새로운 입력원으로 넥마이크를 사용하는데 이는 휴대장착이 간편하고 특히 작업장에서 헬멧 등을 쓰고 있을 때 음성 입력원으로 이용될 수 있고 또한 외부 잡음원으로부터 영향이 없기 때문이다. 넥마이크는 목부분에 접촉장착되어 발성이 이루어질 때 목부분의 피부조직의 울림을 센싱하여 음성파형을 얻게 된다. 넥마이크에 의한 음성파형은 일반 마이크에 의해 공기를 울림을 센싱한 음성파형과 유사한 형태를 나타내고 있으나 재생할 때 음색은 다르게 나타나며 스펙트로그램 상에서도 대역별 에너지 및 포먼트의 위치들이 다르게 나타난다.

인식 실험을 위해서 넥마이크로 입력된 신호에 대해 일반 마이크 입력신호에 많이 적용되는 LPC계수[7] 및 LPC-켑스트럼[8], 멜-켑스트럼(MFCC)[8] 등을 추출하여 특징벡터로 사용하였다. 채널잡음을 제거하기 위해 켑스트럼 정규화(CMN)를 수행하였다. 인식 성능은 일반 마이크의 것보다 50~25%정도의 저하를 나타내었다. 실험을 통하여 넥마이크에 의한 입력 음성의 특성을 고려하는 새로운 특징벡터를 추출 또는 음질 향상법이 연구된다면 새로운 인터페이스로 활용이 될 수 있다는 가능성을 확인하였다.

2. 음성의 특징벡터

음성은 비정상적인 특징을 나타내기 때문에 프레임 단위로 나누어 해석을 하면 해당 프레임 내에서는 음성은 정상적인 특성을 나타낸다. 해당 프레임 내의 음성을 대표적으로 표현하기 위해서 특징벡터를 추출한다. 이러한 특징벡터 표현법 중에서 널리 쓰이는 것으로는 LPC계수와 LPC-켑스트럼과 멜-켑스트럼 등이 있다.

2.1 LPC 계수

LPC(Linear Prediction Coefficient)는 인간의 발성 구조를 모델링하는 과정에서 성도(vocal tract)를 특성을 반영하는 것이다. 기본적인 개념은 현재 n번째의 음성 샘플 $s(n)$은 과거의 p개의 샘플들의 선형조합으로 예측할 수 있다는 것이다. 즉,

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) \quad (1)$$

로 표현할 수 있으며, 여기신호 $Gu(n)$을 포함하면

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \quad (2)$$

이 된다. z-도메인으로 표현하면 전달특성은

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-1}} \quad (3)$$

이 되며 전극(all-pole)모델로써 성도의 특성을 근사화한다. 계수 $a_i,\ i = 1, \cdots, p$는 Levison- Durbin 알고리즘[7]에 의해 구할 수 있다. LPC 계수는 성도의 주파수 특성의 포락선을 근사화하여 표현하며 여기신호와 성도 특성을 분리시키는 성질을 가져 널리 쓰인다.

2.2 LPC-켑스트럼

켑스트럼 계수는 로그스펙트럼의 퓨리에 변환으로 얻을 수 있으며 LPC 계수보다 인식 성능에 나은 특징벡터 역할을 한다. LPC 계수로부터 변환되어 얻을 수 있다[8].

2.3 멜-켑스트럼(MFCC)

멜 밴드는 인간의 귀의 특성과 유사한 주파수 분해능을 가진다. 그림 1에서와 같이 인간의 귀는 낮은 주파수대에서는 분해능이 높으나 높은 주파수대에서는 분해능이 떨어지며 1kHz이하에서는 선형적이고 이상에서는 대수적인 스케일의 분해능을 가진다. 이와 같은 청각신경을 모델링한 특징벡터가 일반적으로 인식시스템에 사용된다. 멜-켑스트럼은 프레임 신호를 퓨리에 변환한 후 멜-대역으로 주파수 워핑하여 대역에너지의 대수값을 DCT변환하여 얻는다.[9]

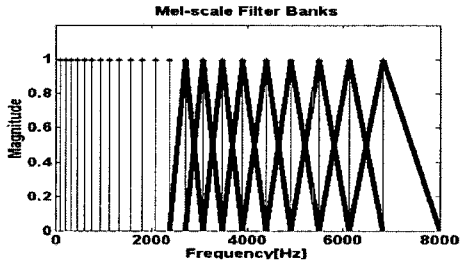


그림 1. Mel-scale Filter Banks

2.4 켑스트럼 정규화

채널의 특성은 신호적인 측면에서 입력 음성에 컨볼루션되어 나타나지만 로그 스펙트럼 상에서는 부가적(additive)으로 나타나므로 켑스트럼상에서 평균을 구하여 차감함으로써 제거될 수 있다.

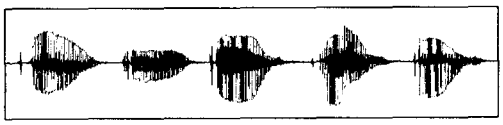
3. 넥마이크폰으로 입력된 음성신호의 특성

넥마이크로 입력된 신호의 음성파형을 그림 2의 (b)에 나타내었다. 일반마이크의 음성파형 (a)와 비교하여 비슷한 형태를 보여주어 넥마이크에서 채집된 신호의 이용 가능성을 보여준다. 하지만 실제로 음성을 들어보면 넥마이크 음성은 명료도가 낮아 분별력을 떨어뜨리는 경우도 있다. 여기서 주목할 만한 점은 음성의 발성과정이 생략된 상황인데도 입을 통해서 발성된 음파를 신호원으로 하는 경우와 목에서의 진동을 신호원으로 하는 경우가 비슷한 파형을 보여준다는 것이다. 음성의 발성과정을 살펴보면 모음의 경우 성대의 주기적인 떨림에 의한 여기신호에 성도(vocal track)를 거치면서 혀의 위치나 비강 등의 영향을 받아 특정 주파수 성분이 공진이 되어 음의 분별력을 제공하고 입을 통하여 음파로서 전달된다. 성도를 거치진 않은 신호가 비슷한 파형으로 관측되는 것은 피부조직을 통해서 전달되는 것으로 여겨지는데 더욱 자세한 고찰과 규명이 필요하다.

주파수 영역에서의 특성을 살펴보기 위해 그림 3에 스펙트로그램을 나타내었다. (a)의 일반 마이크로 입력 받은 음성에 대한 스펙트로그램에서 각 모음 음소의 포먼트들이 뚜렷이 나타나며 포먼트의 위치 또한 각 음소들을 구별할 수 있도록 각각 다른 위치에 존재한다. 하지만 (b)의 넥 마이크로 입력된 파형의

스펙트로그램에서는 포먼트의 위치가 더 이상 음소의 분별력에 도움을 주지 못하고 또한 높은 주파수 대역 쪽에서 신호가 거의 나타나지 않아 넥마이크 특성 또는 목부분의 피부조직의 전달특성에서 대역제한 특성이 있는 것으로 여겨진다. 모음 /a/에 대한 로그-스펙트럼과 LPC계수의 포락선을 그림 4에 나타내었다. 일반마이크에 의한 신호의 로그 스펙트럼에서는 성도의 특성을 나타내는 포먼트 성분들이 뚜렷히 나타나지만 넥마이크에 의한 신호의 로그스펙트럼에서는 저주파수 대역에서 포먼트가 발생하며 고주파 대역으로 갈수록 감쇄가 심하게 일어난다. 이러한 현상은 다른 모음 음소에서 관찰되었으며 연속음성인식에서 음소모델의 분별력을 약화시킬 것이다. 앞으로 넥마이크의 특성이 규명된다면 특징벡터 추출과정에 반영하여 넥마이크에 의한 음원에 적합한 특징벡터를 얻을 수 있을 것이다.

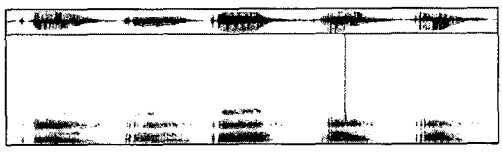
(a)



(b)

그림 2. 모음 /a/-/e/-/i/-/o/-/u/를 발성했을 때의 음성 파형 : (a) 일반 마이크 (b) 넥 마이크

(a)



(b)

그림 3. 모음 /a/-/e/-/i/-/o/-/u/의 스펙트로그램; (a) 일반 마이크 (b) 넥 마이크

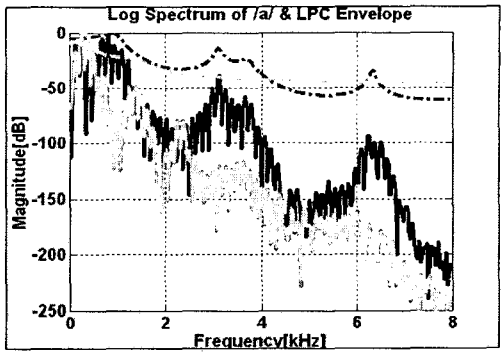


그림 4. 모음 /a/의 로그 스펙트럼 및 LPC계수의 포락선; 진한 실선: 일반 마이크의 로그스펙트럼, 연한 실선: 넥 마이크의 로그스펙트럼, 일점쇄선: 일반마이크의 LPC포락선, 파선: 넥마이크의 LPC 포락선

4. 넥마이크폰으로 입력된 음성신호에 대한 인식실험

넥마이크로 입력된 음성신호는 3장의 설명처럼 일반 마이크의 음성에 대한 특성과 다른 점을 가지는데 2장에서 설명된 특징벡터 추출법들을 적용하여 그 적합성을 알아본다.

먼저 기존 일반마이크의 음성에 의해 만들어진 음소모델에 대해 넥마이크의 음성을 테스트해 보았다. 음소 모델의 훈련과정에서 초기화는 ETRI의 PBW611 데이터베이스로 이루어졌고 훈련모델은 ETRI의 PBW445 데이터베이스로 embedded 훈련방법에 의해 생성되었다. 음소 모델은 CV-decision tree 방법에 의한 트라이폰 모델이며 39차의 MFCC 특징벡터에 의한 것이다. 인식대상은 단어발성문장이고 테스트 음성은 넥마이크로 채집된 신호이며 인식어휘 445개를 발음한 것이다. 테스트 음성이 훈련과정에 전혀 참여하지 않은 환경에서의 인식률은 15.51%로 상당히 저조히 나타났는데 이는 넥마이크의 특징벡터가 모델 파라미터 값과 불일치한데 기인한다. 다음으로 넥마이크로 발성된 한 명의 화자가 445개의 단어를 6번 발성한 30분의 데이터로 embedded 훈련하고 훈련에 참여한 1번의 발성을 테스트 음성으로 하였을 때, 인식률은 34.72%이었다. 보통 훈련에 참여한 데이터로 테스트할 경우 인식률은 95%정도를 나타내는데 넥마이크 음성에 의한 낮은 인식률은 embedded 훈련과정에서 특징벡터의 상

이함에 의해 모델 파라미터의 업데이트가 제대로 수행되지 않는다는 원인과 그림2의 스펙트로그램에서와 같이 음소간의 분별력이 없는 원인 때문이다.

음소단위의 모델링 접근법의 어려움 때문에 단어단위의 모델 생성으로 전환하였다. 인식어휘는 445개의 단어 중 임의로 44개를 선택하였고 단어모델은 한 명의 화자가 5번 발성한 것으로 생성되었고 5개의 상태와 1개의 가지를 가진다. 훈련은 Viterbi tranining방법, 인식은 Viterbi 알고리즘이 사용되었다. 모든 특징벡터의 차수는 13차이며 MFCC는 20개의 멜 밴드의 에너지 값을 DCT변환하였다. 테스트는 같은 화자의 1번 발성한 것으로 하였다. 인식결과를 표1에 나타내었다.

표 1. 화자 종속 단어모델에 대한 넥마이크로 입력된 음성의 특징벡터별 인식률[%]

	LPC Cepstrum	MFCC	MFCC+CMN
일반마이크	96.4	97.7	N
넥마이크	52.3	50.0	75.0

일반마이크의 특징벡터별 인식률은 훈련과정 및 인식과정과 특징벡터 추출법의 적합성을 보여준다. 하지만 넥마이크에 입력 음성에 대한 특징벡터는 그 접근법에 문제가 있음을 내포하고 있다. 분별력이 없는 음소 모델보다는 단어모델 접근법의 타당성을 보여준다. 넥마이크 입력 음성에 CMN을 적용하였을 때 인식률이 어느 정도 향상되었는데 이는 넥마이크의 채널잡음이 상당히 존재함을 의미한다.

5. 결론 및 향후

본 논문에서는 외부 잡음원의 영향을 받지 않으며 휴대장착이 편리하며 특수 상황에서 활용 가능한 넥마이크의 입력 음성에 대한 음성인식 성능을 평가하였다. 넥마이크의 입력 음성은 특징벡터의 추출에 많이 이용되는 주파수 영역의 정보가 일반 마이크와 다른 측면을 가지는데 고주파 대역의 신호가 존재하지 않으며 모음 음소의 경우 성도의 특성을 나타내는 포먼트의 위치 정보가 많이 소실되어 나타나고 대역별 에너지가 비슷하게 분포하여 음소단위의 모델링에 문제점을 보인다. 따라서 음소모델을 생성하기 위해서는 새로운 특징벡터 추출법이 연구되어야 하며 새로운 데이터베이스 구축도 병행되어야 할 것이다. 해당 단어의 전체 특징벡터를 모델링에 참여시키는 단어모델은 단어 레벨에서의 어느 정도의 분별력을 가지고 있지만 일반마이크에 의한 음성 단어보다는 그 성능이 떨어진다. 넥마이크에 의한 음성은 전화망 음성과 비슷한 대역이 제한된 특성을 보이며 넥마이크 자체 또는 전달 특성에서 채널잡음이 존재함을 확인하였다. 향후에는 넥마이크에서 음성이 샘플링되는 과정을 더욱 자세히 고찰하여 이것을 특징벡터 추출과정에 반영하여 일반마이크에 의한 음성의 특징벡터 추출과정과 차별화되는 특징벡터를 추출하여야 할 것이다.

참고문헌

[1] S. F. Boll, "Suppression of acousic noise in speech using spectral subtraction," IEEE trans. Acoust., Speech, Signal Processing, vol. ASSP-27, pp. 113-120, Apr. 1979.

[3] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Processing, vol. 2, pp. 578-589, Oct. 1994.

[4] F.-H. Liu, "Environment normalization for robust speech recognition using cepstral comparison," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 2, pp. 61-64, 1994.

[5] Gales, M.J.F.and Young, S.J., "Robust continuous speech recognition using parallel model combination," IEEE Trans. Speech and Audio Processing, vol. 4 pp. 352-359, Sep. 1996.

[6] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adapdation of continuous density hidden Markov models," Computer Speech and Language, vol. 9, pp. 171-185, 1995.

[7] J.D. Markel and A.H. Gray, "Linear Prediction of Speech,", Springer-Verla, 1976.

[8] HTK Book ver. 2.2, Cambridge University.

[9] Molau, S. and Pitz, M., "Computing Mel-frequency cepstral coefficients on the power spectrum," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 1, pp. 73-76, 2001.