http://dx.doi.org/10.14400/JDC.2015.13.10.313

음성 에너지 분포 처리와 에너지 파라미터를 융합한 음성 인식 성능 향상

오상엽
가천대학교 IT대학 컴퓨터공학과

Voice Recognition Performance Improvement using a convergence of Voice Energy Distribution Process and Parameter

Sang-Yeob Oh
Dept. of Computer Engineering, Gachon University

요 약 전통적인 음성 향상 방법은 잘못된 잡음의 추정에 따라 남아있는 잡음이 발생하여 음성 스펙트럼을 왜곡하거나 음성 프레임을 찾지 못하여 음성 인식 성능을 저하시키는 문제가 발생된다. 본 논문에서는 음성 에너지 분포 처리와 음성 에너지 파라미터를 융합한 음성 검출 방법을 제안하였다. 제안한 방법은 음성 에너지를 최대화시켜 잡음의 영향을 적게 받는 특성을 이용하였다. 또한, 음성 신호의 특징 파라미터 중에서 작은 값을 가지는 로그에너지 특징의 구간에서는 큰 에너지를 가지는 구간에 비해 상대적으로 로그에너지 값을 더 많이 키워서 잡음이 포함한 음성신호의 로그에너지 특징의 크기와 비슷하게 하여 훈련과 인식 환경의 불일치를 융합으로 인해 줄여준다. 인식 실험 결과 기존 방법에 비해 향상된 인식 성능을 확인할 수 있었으며, car 잡음 환경의 음성 구간 적중률은 낮은 SNR 구간인 0dB과 5dB에서는 97.1%와 97.3%의 정확도를 보였으며, 높은 SNR구간인 10dB와 15dB에서는 98.3%, 98.6%의 정확도를 보였다.

주제어 : 음성 인식, 음성 분포, 음성 에너지 파라미터, 음성 검출

Abstract A traditional speech enhancement methods distort the sound spectrum generated according to estimation of the remaining noise, or invalid noise is a problem of lowering the speech recognition performance. In this paper, we propose a speech detection method that convergence the sound energy distribution process and sound energy parameters. The proposed method was used to receive properties reduce the influence of noise to maximize voice energy. In addition, the smaller value from the feature parameters of the speech signal The log energy features of the interval having a more of the log energy value relative to the region having a large energy similar to the log energy feature of the size of the voice signal containing the noise which reducing the mismatch of the training and the recognition environment recognition experiments Results confirmed that the improved recognition performance are checked compared to the conventional method. Car noise environment of Pause Hit Rate is in the 0dB and 5dB lower SNR region showed an accuracy of 97.1% and 97.3% in the high SNR region 10dB and 15dB 98.3%, showed an accuracy of 98.6%.

Key Words : Voice recognition, Voice distribution, Voice energy parameter, voice detectin

Received 2 August 2015, Revised 7 September 2015
Accepted 20 October 2015
Corresponding Author: SangYeob Oh
(Dept. of Computer Engineering, Gachon University)
Email: syoh1234@gmail.com

ISSN: 1738-1916

1. 서론

컴퓨터 시스템의 하드웨어와 소프트웨어 기술의 발전으로 다양한 스마트 기기 들이 제공되고 있으며, 이들 스마트 기기에서의 컨텐츠 시장은 급속하게 발전하고 있다. 이러한 기기들에서 사용되는 음성 인식 시스템도 하드웨어에서의 음성 신호 처리 기술과 이의 지원을 위한 소프트웨어의 발전으로 인해 시스템의 성능이 향상되고 적용 분야도 확대되고 있지만, 음성 인식 시스템의 실용적인 측면에서의 방해 요인으로 기기에서 발생하는 잡음과 네트워크의 신호에서 발생되는 잡음의 환경적인 변화가 음성 인식 시스템의 성능 저하를 야기 시키고 있으며, 인식 모델 시스템을 사용할 때 정의되지 않은 어휘나 인식 모델을 사용하면서 추가되어진 어휘, 그리고 어휘에 대한 모델이 부족하여 새로이 모델링하여 생성된 인식 어휘들은 인식률 저하의 원인이 된다. 음성인식 알고리즘 개발에 있어 기존 연구들에서는 퍼지로직, Neural Network 및 HMM(Hidden Markov Model) 방법들을 주로 사용하고 있으며[1,14], 이를 기반으로 한 GMM(Gaussian Mixture Model)[5], 그리고 CHMM(Continuous Hidden Markov Model)[2,3,4,6] 모델들을 사용한다. HMM은 시공간적인 정보를 통한 모델링과 학습 및 인식을 위한 효과적이고 우수한 알고리즘을 가지고 있어 여러 분야에서 응용되고 있으며[9,12,13], 음성인식에 있어 가장 널리 사용되어지고 있다. 음성 인식을 위한 모델 파라미터들로부터 데이터 부족 문제가 발생하게 되는데[15] 이는 인식을 위한 모델별 훈련용 데이터의 량이 일정하지 않기 때문이다. 음성구간의 검출은 음성코딩, 음성향상, 음성인식 분야에서 인식 성능 향상과 밀접한 관련이 있다. 음성 인식 향상 시스템에서 음성 구간을 검출하는 것은 상당히 중요하며 잡음 구간의 정확한 추정에 영향을 주기 때문이다. 음성 향상 시스템에서 사용하는 전통적인 잡음 제거 방법으로는 주파수 공간에서 추정된 잡음 신호를 이용하여 음성 신호와 섞여 있는 잡음 신호를 차감하는 주파수 차감법이 주로 사용된다. 일반적인 잡음 추정 방법으로는 음성검출기(VAD, Voice Activity Detection)에 의존하여 음성 부재 구간에서 잡음의 평균을 구하는 방법이 있다.

음성 검출 알고리즘은 음성과 비음성 신호를 판별하기 위한 특징 파라미터를 구하여 적절한 문턱(theshold) 값을 특징 파라미터에 적용하는 결정식(decision rule)의 형태로 음성과 비음성을 구분한다. 음성과 비음성 검출을 위하여 음성 에너지 파라미터 방법을 사용한다. 본 논문에서는 음성 에너지 분포 처리와 음성 에너지 파라미터를 이용한 음성 검출 방법을 제안하였다. 제안한 방법은 음성 에너지 분포 처리를 이용하여 잡음의 편차를 줄이고, 음성 에너지 파라미터를 사용하여 잡음이 포함한 음성 신호의 로그 에너지 값을 조정하여 룬련과 인식 환경의 불일치를 감소하였다, 인식 실험 결과 기존 방법에 비해 향상된 인식 성능을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 언급하고 3장에서는 음성 에너지 분포 처리와 음성 에너지 파라미터를 이용한 음성 검출 방법에 대해 설명하며, 4장에서는 시스템 평가를 수행하고 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

2.1 HMM, CHMM, GMM 모델

HMM 알고리즘에서 처리되는 음성은 Markov 프로세스로 표준화 되어 처리되며, 이를 위한 음성의 학습 과정에서 Markov 모델의 변수를 가지고 기준 Markov 모델을 작성한 다음 입력된 음성과 저장된 기준 Morkov 모델을 비교하여 유사도가 가장 높은 기준 Morkov 모델을 인식된 어휘로 결정한다[7].

GMM은 출력 확률밀도함수가 가우시안 밀도 혼합인 1개의 상태만으로 구성된 CHMM의 한 형태로서, GMM은 다음과 같은 특징을 가지고 있다.

첫째, GMM은 음향학적 어느 공통 특성을 가진 집합을 모델링할 수 있다. 음성에 대한 발성에 대응되는 음향공간은 모음이나 비음, 파찰음과 같은 음소를 표현하는 음향학적 클래스의 집합으로 잘 표현된다.

둘째, 단봉 가우시안 음소모델은 음소분포를 표현하기 위해 평균 벡터의 특징벡터와 공분산으로 각 음소의 특징벡터의 이산집합으로 표현한다. 이와 같은 점을 고려하여 구성된 GMM은 가우시안 함수의 이산집합을 사용하여, 각각의 평균과 공분산을 가지게 함으로써 이들 두 모델의 특징을 혼합한 형태이다.

가우시안 혼합 밀도는 M성분 밀도의 가중합계로서

식 (1)에 의해 얻어진다.

HMM 모델에서 상태열 q에 대한 관측열의 확률은 다음과 같이 표현된다.

$$p(x|\lambda)=\sum_{i=1}^{M} c_i b_i(x) \quad (1)$$

x는 d-차원 랜덤 벡터이며, $b_i(x),\ i=1,...,M$는 i번째의 성분 밀도이고, $c_i,\ i=1,2,...,M$는 i번째 혼합 밀도 가중치이다. 각 혼합 밀도의 가중치는 다음과 같이 제한된다.

$$\sum_{i=1}^{M} c_i = 1 \quad (2)$$

각 성분 밀도는 평균 μ_i과 공분산 Σ_i를 가지는 d-변량 가우시안 함수이다.

CHMM은 가우시안 확률 밀도 함수를 가장 많이 사용하며, 1차원의 특징 벡터는 두 개의 파라미터인 평균 μ과 표준편차 σ를 구하여 가우시안 확률 밀도 함수로 다음과 같이 표현한다[8].

$$f_x(x)=\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

또한, 가우시안 확률 밀도 함수는 2차 지수 함수의 형태를 가지므로 2차원 이상의 다차원 식으로 표현이 가능하다.

2.2 주파수 마스킹

주파수 마스킹은 주파수 대역 값에 의해서 마스킹 처리되는 음성의 범위가 변경되며, 같은 마스킹 값을 가지는 주파수의 대역 폭을 크리티컬 밴드(critical band)라고 부른다. 크리티컬 밴드의 범위는 1kHz 아래의 주파수에서는 100Hz의 범위로 대부분 동일하며, 1kHz 이상의 값에서는 주파수가 가지는 값에 비례하여 변경되는 특징을 가지고 있다[10]. 바크스케일 주파수는 음성의 주요한 값에 따른 대역에 대한 표현이 우수하며[11], 입력신호에 대해 FFT를 결정된 값의 프레임내의 각 주파수 빈(Frequency bin)에 대해 바크스케일로 변경한다. 또한, 밴드 특성은 음성에너지를 표현하기에 적합하게 구성되므로 음성에너지 최대화에 도움이 된다. 최대화된 음성은 잡음에 대해 상대적으로 SNR이 높아지고 음성과 비음성을 구분하기가 편리한 장점을 가진다. 최대화된 음성을 기반으로 문턱값에 의해 음성과 비음성 구간을 결정한다.

3. 시스템 모델

3.1 음성 에너지 분포와 음성 검출

음성 검출 및 음성 강화 알고리듬들은 대부분 잡음을 추정하고 잡음의 변화를 시간의 변화에 따라 추적하는 학습 알고리듬을 사용하였다[11]. 하지만 이와 같은 방법들은 잡음의 변화량을 계산하기 위해 현재 프레임과 이전 프레임을 비교거나 100ms 이상의 긴 시간을 분석 구간으로 사용하여 실제 환경에 응용하기에 적합하지 않은 점이 있다. 주어진 프레임 내에서 음성 에너지를 최대화시키고 잡음을 억제하는 방법을 제안한다. 제안된 방법은 주어진 프레임에서 모든 것을 처리하기 때문에 분석 구간이 상대적으로 적어 실제 응용에서 유리하다[12]. 제안된 알고리듬을 처리하기 위해 입력 신호에 대한 단 구간 퓨리에 분석은 음성 신호 x(t)를 정의하고 주파수 영역에서 처리하기 위하여 DFT(Discrete fourier Transform) 처리하여 다음 식과 같이 주파수 성분을 X_i로 나타낸다.

$$X_i = |\sum_{k=0}^{N-1} x_k e^{-j(\frac{2\pi mk}{N})} \quad (4)$$

단, 구간 퓨리에 변환된 는 음성에너지 최대화를 위해 선형 주파수를 인간의 청각모델에 기반 한 비선형 주파수 크기로 변환한다. 음성은 모음에서 피치 주파수(pitch frequency)를 가지며 피치 주파수를 기본 주파수라고 한다. 기본 주파수는 음성 영역의 전 대역에 걸쳐 에너지가 가장 큰 특징을 가지고 있어 최대 에너지로 나타나며 최소 에너지는 음성 신호와 무관한 잡음 신호로 나타난다. 각 프레임에서 표준 편차는 음성 에너지의 분포를 표현하며 음성 에너지의 중요한 성분들은 100Hz ~ 600Hz 대역에 집중되어 있는 특성이 있어 모든 가청 영역에 걸쳐 에너지 크기의 편차가 크게 나타난다. 잡음의 경우는 에너지 크기가 모든 가청 영역에 비교적 고르게 분포되어 편차가 작게 나타난다. 따라서 음성이 존재하는 프레임

과 존재하지 않는 프레임의 표준편차 값은 차이가 나타나고 음성에너지를 최대화한 상태에서의 평균값 역시 차이가 나타난다. 본 논문에서는 비음성 구간에 대한 구분을 더욱 명확히 하기 위해 주어진 프레임에 표준편차와 평균값을 더한다. 따라서 음성 검출을 위한 문턱 값의 기준이 되며 음성의 존재 여부를 결정하게 된다.

3.2 음성 에너지 파라미터

기존의 음성 검출 및 음성 강화 알고리즘들은 대부분 잡음을 평가하고 잡음의 변화를 시간의 변화에 따라 평가하는 학습 알고리즘을 사용하였다. 그러나 이와 같은 방법들은 잡음의 변화폭(variable breadth)을 계산하기 위해 현재 프레임과 이전 프레임을 비교거나 100ms 이상의 긴 시간을 분석 구간으로 사용하여 실제 환경에 응용하기에 적합하지 않은 점이 있다. 주어진 프레임 내에서 음성 에너지를 최대화시키고 잡음을 억제하기 위해서 음성 에너지 파라미터 방법을 제안한다. 음성 에너지 파라미터 처리를 위해 잡음 환경에서 발성된 음성 신호는 큰 에너지 값을 가지는 음성 구간에서는 부가 잡음의 영향을 거의 받지 않으나 작은 에너지 값을 가지는 구간에서는 큰 영향을 받는다. 음성 신호의 특징 파라미터 중에서 에너지 특징을 [Fig. 3]과 같이 처리한다. 작은 값을 가지는 로그에너지 특징의 구간에서는 큰 에너지를 가지는 구간에 비해 상대적으로 로그에너지 값을 더 많이 키워서 잡음이 포함한 음성신호의 로그에너지 특징의 크기와 비슷하게 하여 훈련과 인식 환경의 불일치를 줄여준다. 이를 위하여 각 음성 신호의 로그에너지가 동일한 변화 범위를 갖기 위한 DR(dynamic range) 함수는 다음과 같이 정의 한다.

$$DR(dB) = 10 \times \frac{Max(logE_n)}{Min(logE_n)} \quad (5)$$

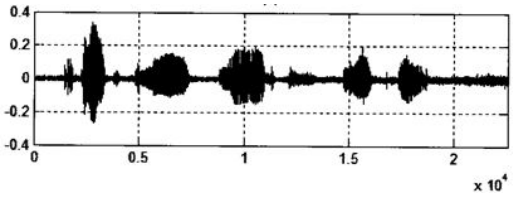
Max($logE_n$)는 N개 프레임에서 최대의 로그에너지 값, Min($logE_n$)는 최소의 로그에너지 값을 나타낸다. DR의 값이 정해지면 각 음성 신호에 대한 로그에너지 특징의 최소값을 계산한다.

4. 실험 결과

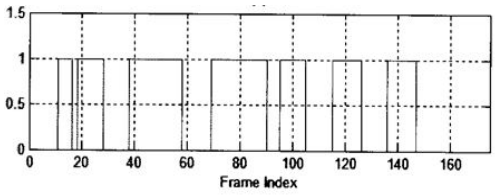
본 논문에서 제안한 음성 에너지 분포 처리와 음성 에너지 파라미터를 융합한 음성 인식 성능을 실험하였다.

평가를 위해 사용한 Aurora 2.0 데이터베이스를 사용하였으며, Aurora 2.0에는 잡음 환경과 각각의 잡음 레벨로 구성되어 white Gaussian noise, babble noise 등을 포함하며 street, airport, car noise 등 잡음 환경별로 구분되어 voice improvement algorithm의 성능 검증용으로 사용한다.[16].

음원은 8kHz sampling rate, 16bit를 사용하였으며 FFT 크기는 256 샘플, 1/2 오버레핑(overlapping) 구간을 이용하였고 해밍 윈도우(Hamming Window)를 사용하였다[12]. 실험 결과 [Fig. 1]은 15dB에서 입력 신호 값에 따른 에너지 분포 처리와 파라미터가 적용된 음성검출 결과를 [Fig. 2]에 표현하였다. 또한, 잡음 환경(15dB, 10dB, 5dB, 0dB)을 구분하여 실험하였다.



[Fig. 1] Input Signal of SNR 15dB



[Fig. 2] Results of Speech Detection

다음의 <Table 1>은 음성 검출 성능을 평가한 결과이며, car 잡음 환경의 음성 구간 적중률(Pause Hit Rate)은 낮은 SNR구간인 0dB과 5dB에서는 97.1%와 97.3%의 정확도를 보였으며, 높은 SNR구간인 10dB와 15dB에서는 98.3%, 98.6%의 정확도를 보였다. 또한, 음성구간에 대한 비음성 구간 오보율(False Alarm Rate)은 SNR 10dB와 15dB에서 1.7%%, 1.4%로 좋은 성능을 보였으나 낮은 SNR구간인 0dB와 5dB에서는 각각 3.1%, 6.7%의 성능을 보였다.

〈Table 1〉 PHR and FAR for the SNR

Noise	SNR	VAD Result (%)	
		PHR	FAR
Car	0	97.1	3.1
	5	97.3	6.7
	10	98.3	1.7
	15	98.6	1.4

5. 결론

본 논문은 음성 에너지 분포 처리와 음성 에너지 파라미터를 융합한 방법을 이용한 음성 검출과 인식 성능을 평가하였다. 실제 자동차의 잡음 환경에서는 주변의 차소리와 실내 소음 등으로 인해 신호 대 잡음비가 낮은 음성 신호에 대해서는 피쳐 파라미터들이 잡음 신호에 민감하기 때문에 음성 검출의 성능이 저하되는 원인이 된다.

따라서 음성 에너지 분포 처리와 파라미터를 융합한 음성 검출 방법을 제안하였으며, 제안한 방법에서는 음성 에너지 분포 처리를 이용하여 잡음의 편차를 줄이고, 음성 에너지 파라미터를 사용하여 잡음이 포함한 음성 신호의 로그 에너지 값을 조정하여 룬련과 인식 환경의 불일치를 감소하였다, 인식 실험 결과 향상된 인식 성능으로 car 잡음 환경의 음성 구간 적중률(Pause Hit Rate)은 낮은 SNR구간인 0dB과 5dB에서는 97.1%와 97.3%의 정확도를 보였으며, 음성구간에 대한 비음성 구간 오보율(False Alarm Rate)은 SNR 10dB와 15dB에서 1.7%%, 1.4%로 좋은 성능을 보였다. 기존의 연구 방법은 100ms 이상의 긴 시간을 분석 구간으로 사용하여 실제 환경에 응용하기에 적합하지 않은 점이 있기에 향후 연구 과제로는 음성 에너지 최대화에 대한 성능을 보다 향상시키기 위한 프레임 단위 짧은 시간의 적용 및 이에 대한 구체적인 방법을 필요로 한다.

REFERENCES

[1] Chan-Shik Ahn, Sang-Yeob Oh. Gaussian Model Optimization using Configuration Thread Control In CHMM Vocabulary Recognition. The Journal of Digital Policy and Management. Vol. 10, No. 7, pp. 167-172, 2012.

[2] Chan-Shik Ahn, Sang-Yeob Oh. Echo Noise Robust HMM Learning Model using Average Estimator LMS Algorithm. The Journal of Digital Policy and Management. Vol. 10, No. 10, pp. 277-282, 2012.

[3] Chan-Shik Ahn, Sang-Yeob Oh. CHMM Modeling using LMS Algorithm for Continuous Speech Recognition Improvement. The Journal of digital policy and management. Vol. 10, No. 11, pp. 377-382, 2012.

[4] Sang-Yeob Oh. Selective Speech Feature Extraction using Channel Similarity in CHMM Vocabulary Recognition. The Journal of digital policy and management. Vol. 11, No. 10, pp. 453-458, 2013.

[5] A. Srinivasan, Speech Recognition Using Hidden Markov Model, Applied Mathematical Sciences, vol. 5, no. 79, pp. 3943-3948, 2011.

[6] Campbell, W. M., Sturim, D. E., Reynolds, D. A., Solomonoff, A. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. Proc. ICASSP, No. 1, pp. 97-100, 2006.

[7] Zhang, Y., Xu, J., Yan, Z. J., & Huo, Q. An i-vector based approach to training data clustering for improved speech recognition. Proc. Interspeech, pp. 1247-1250. 2011.

[8] Beaufays, F., Vanhoucke, V., & Strope, B. Unsupervised discovery and training of maximally dissimilar cluster models. Proc. Interspeech, pp. 66-69, 2010.

[9] Sang-Yeob Oh. Improving Phoneme Recognition based on Gaussian Model using Bhattacharyya Distance Measurement Method. Journal of Korea Multimedia Society. Vol. 14, No. 1, pp. 85-93, 2011.

[10] Caban, A. Dolinska, B. Budzinski, G. Oczkowicz, G. Ostrozka-Cieslik, A. Cierpka, L. Ryszka, F. The Effect of HTK Solution Modification by Addition of Thyrotropin and Corticotropin on Biochemical Indices Reflecting Ischemic Damage to Porcine Kidney. Transplantation proceedings. Vol. 45, No. 5,

pp. 1720-1722, 2013

[11] Myoung-hwan Ahn, Joon-hee Kwon. Ontology based Context-Aware Recommendation System using Concept Hierarchy. Journal of Korean Society for Internet Information. Vol. 8, No. 5, pp. 81-89, 2007.

[12] Chan-Shik Ahn, Sang-Yeob Oh. Vocabulary Recognition Retrieval Optimized System using MLHF Model . Journal of the Korea Society of Computer and Information. Vol. 14, No. 10, pp. 217-223, 2009.

[13] Sang-Yeob Oh. Noise Removal using a Convergence of the posteriori probability of the Bayesian techniques vocabulary recognition model to solve the problems of the prior probability based on HMM, The Journal of digital policy and management. Vol. 13, No. 8 pp. 295-300, 2015

[14] Sang-Yeob Oh. Bayesian Method Improve Recognition Rates using HMM Vocabulary Recognition Model Optimization. The Journal of digital policy and management. Vol. 12, No. 7, pp. 273-278, 2014.

[15] Sang-Yeob Oh. Decision Tree State Tying Modeling Using Parameter Estimation of Bayesian Method The Journal of Digital Policy and Management. Vol. 13, No. 1, pp. 1243-248, 2015.

[16] C.-C. Wang, C.-A. Pan, and J.-W. Hung, “Silence Feature Normalization for Robust Speech Recognition in Additive Noise Environments,” Proc. ICSLP, pp. 1028-1031, 2008.

오 상 엽(Oh, Sang Yeob)

・1991년 2월 : 광운대학교 대학원 전자계산학과 (이학석사)
・1999년 2월 : 광운대학교 대학원 전자계산학과 (이학박사)
・2007년 2월 ~ 현재 : 가천대학교 IT대학 컴퓨터공학과 교수
・관심분야 : 버전관리, 형상관리, 음성/음향 신호 처리, 차량 통신
・E-Mail : syoh1234@gmail.com