Journal of the Korea Convergence Society
Vol. 8. No. 5, pp. 13-17, 2017 https://doi.org/10.15207/JKCS.2017.8.5.013

# 음성 신호 특징과 셉스트럼 특징 분포에서 묵음 특징 정규화를 융합한 음성 인식 성능 향상

황재천
가천대학교 컴퓨터공학과

# Voice Recognition Performance Improvement using the Convergence of Voice signal Feature and Silence Feature Normalization in Cepstrum Feature Distribution

Jae-Cheon Hwang
Department of Computer Engineering, Gachon University

**요 약** 음성 인식에서 기존의 음성 특징 추출 방법은 명확하지 않은 스레숄드 값으로 인해 부정확한 음성 인식률을 가진다. 본 연구에서는 음성과 비음성에 대한 특징 추출을 묵음 특징 정규화를 융합한 음성 인식 성능 향상을 위한 방법을 모델링 한다. 제안한 방법에서는 잡음의 영향을 최소화하여 모델을 구성하였고, 각 음성 프레임에 대해 음성 신호 특징을 추출하여 음성 인식 모델을 구성하였고, 이를 묵음 특징 정규화를 융합하여 에너지 스펙트럼을 엔트로피와 유사하게 표현하여 원래의 음성 신호를 생성하고 음성의 특징이 잡음을 적게 받도록 하였다. 셉스트럼에서 음성과 비음성 분류의 기준 값을 정하여 신호 대 잡음 비율이 낮은 신호에서 묵음 특징 정규화로 성능을 향상하였다. 논문에서 제시하는 방법의 성능 분석은 HMM과 CHMM을 비교하여 결과를 보였으며, 기존의 HMM과 CHMM을 비교한 결과 음성 종속 단계에서는 2.1%p의 인식률 향상이 있었으며, 음성 독립 단계에서는 0.7%p 만큼의 인식률 향상이 있었다.

• 주제어 : 음성 인식, 특징 추출, 묵음 특징 정규화, 음성 특징, 잡음

**Abstract** Existing Speech feature extracting method in speech Signal, there are incorrect recognition rates due to incorrect speech which is not clear threshold value. In this article, the modeling method for improving speech recognition performance that combines the feature extraction for speech and silence characteristics normalized to the non-speech. The proposed method is minimized the noise affect, and speech recognition model are convergence of speech signal feature extraction to each speech frame and the silence feature normalization. Also, this method create the original speech signal with energy spectrum similar to entropy, therefore speech noise effects are to receive less of the noise. the performance values are improved in signal to noise ration by the silence feature normalization. We fixed speech and non speech classification standard value in cepstrum For th Performance analysis of the method presented in this paper is showed by comparing the results with CHMM HMM, the recognition rate was improved 2.7%p in the speech dependent and advanced 0.7%p in the speech independent.

• Key Words : Voice recognition, feature extract, silence feature normalization, voice feature, noise

# 1. 서론

음성 인식 처리를 위해서는 기본적인 음성 신호에 추가 및 혼합되는 잡음을 효율적으로 삭제하고, 기본 음성 신호만으로 정확한 인식을 하도록 처리하는 방법이 매우 중요하다[1,2,5,6,7]. 음성 인식 처리를 위한 처리 방법은 기본 주파수를 가지고 음성 신호를 프로세스하는 잡음 제거 방법과 통계적 모델을 사용한 필터를 가지고 음성 신호를 처리하는 방법, 그리고 임의 공간(subspace)을 사용하여 처리하는 방법을 이용한다[3,11,12]. 주파수의 기본 영역을 이용하여 음성 신호를 처리하기 위한 방법으로는 주파수 차감법 (spectral subtraction)이 일반적으로 이용되고, 통계적 모델을 사용하여 음성 신호를 추출하는 방법은 Wiener 방법론을 이용한다. 그리고 임의 공간을 사용하여 음성을 처리하는 방법으로는 마스크 필터링(mask filtering) 기법이 이용되고 있다. 이들 방법 중에서 임의 공간을 이용한 방법은 잡음 제거 처리에서 우수한 처리 능력을 가지지만 특정 잡음 신호에 대해서는 잡음 제거 능력이 낮은 단점을 가진다. 이러한 특정 잡음 신호에는 음성 신호와 유사한 신호 특징을 가지는 잡음이나 시간이 경과함에 따라 음성 신호 특성이 변경되는 잡음 등이 해당되며, 이들은 불안정한(unstable) 잡음에 해당된다. 일반적인 음성 신호 처리 방법에서는 표준 샘플 데이터베이스들을 구축하고, 음성 인식 실험으로 추출한 음성에 대한 결정 값을 기본으로 구축하고, 이들에 대한 음성 구간을 가지고 음성 신호를 처리한다. 기존 방법에서 음성 인식 추출 방법은 음성 신호에서 잡음을 추정하여 음성과 비음성 구간을 결정하여 처리한다. 이 경우 음성 신호에 대한 잡음이 명확하게 구분되지 않는 경우에는 신호대 잡음이 낮은 환경에서 일정한 값을 가지지 않는 스레숄드(threshold) 값으로 인하여 정확성이 낮은 음성 인식 추출 결과를 가진다.

본 연구에서는 음성 인식 추출을 효율적으로 수행하기 위해 잡음 환경에서 효율적으로 음성 특징 추출하는 방법을 제안한다. 이를 위해 제안한 방법은 음성 인식을 위한 신호대 잡음비가 큰 값에서는 음성과 비음성에 대한 분류를 위한 특징 값을 추출하여 음성 인식 모델을 구성하 여 음성의 특징이 잡음을 최소화하여 모델을 구성하였다. 이를 위해 각 음성 프레임마다 음성 신호 특징 값을 추출하였다. 또한, 생성되는 원 음성 신호의 각 프레임마다 음성 신호에 대한 음성 특징 값을 추출하는 방법을 사용하며, 셉스트럼에서 음성과 비음성 분류의 기준 값을 정하여 신호 대 잡음 비율이 낮은 신호에서 묵음 특징 정규화로 성능을 향상하고, 정규화를 위해 작은 Log-energy을 갖는 비음성 특징을 찾고, 이를 IIR(Infinite Impulse Response) filter를 통과시켜서 사용한다. 논문에서 제시하는 방법의 성능 비교 분석을 위해 HMM[9,14,16]과 CHMM[10, 13]을 비교하여 결과를 보였으며, 음성 종속 단계에서는 각 음성 인식 평균의 차이가 2.3과 3.2의 차이를 가지며, 음성 독립 단계에서는 1.2와 3.1의 음성 인식률 차이가 존재하며, 기존의 HMM과 CHMM을 비교한 차이에 대한 전체 평균에서 약 2.45%의 인식률 향상이 있었다.

본 논문의 2장에서 기본적인 음성 인식을 위해 사용되는 음성 구간과 특징 검출에 대한 내용을 다룬다. 3장 부분은 설명하고, 베이시안의 기본적인 내용을 언급한다. 3장에서는 silence feature normalization의 단점을 해결하기 위해 엔트로피는 DFT(Discrete Fourier Transform) 내용을 다루며, 4장은 성능 분석을 위한 음성 인식 실험을 HMM과 CHMM을 비교하여 수행하였으며, 5장에서 결론을 맺는다.

# 2. 관련 연구

## 2.1 음성 구간 검출

음성에 대한 구간의 추출 방법으로는 ZCR(Zero Crossing Rate)을 이용하며, ZCR은 시간 도메인에서 구간의 샘플이 +1과 -1로 나타나는 2개의 샘플 값을 가지고 처리한다. 계산된 결과를 가지고 음수의 개수에 따라 음성과 비음성 구간을 분류하므로 음수의 개수가 많이 발생하지만 처리가 간단하고 성능이 우수한 특징을 가진다[12]. ZCR은 지정된 구간에서 각각 제곱하고 모두 더해진 값을 Log-energy로 사용하며 수식은 다음과 같다.

$$\log E = 10\log\left[\sum_{k=1}^{n}\left|S_k^2\right|\right] \quad (1)$$

E는 이산 신호 시간(discrete time signal,)의 로그 에너지이며, k는 데이터 시퀀스(data sequence), S는 제한된 타임 윈도우 시퀀스(limited time window sequence)이다.

## 2.2 음성 특징 추출

음성 특징 추출의 일반적인 방법은 인간의 청각 기관

에서 처리되는 단계를 모델링한다. 청각 기관 모델링은 기본적으로 달팽이관 지도를 사용하며, 주기성 특징을 찾는 기본적인 처리 단계는 방법은 자기 상관 계수(autocorrelation)를 가지고 다음 식과 같이 표현한다[4].

$$ACF = \frac{1}{t-\tau}\int_{0}^{t-\tau} x(t)x(t+\tau)dt \quad (2)$$

T-F 단위가 있는 $c$번째 채널의 지연(lag) $\tau$에 윈도우 함수 $\omega(k)$를 곱한 연산을 수행한다[8].

사각 윈도우 함수를 사용하며 주기성을 표현하는 상관 지도를 가지고 프레임에서 상관도를 표현된다. 식 (4)는 채널의 중심 주파수 $f_1, f_2$의 상관도를 표시한다[15].

$$CCF = \int_{-\infty}^{\infty} x(t-\tau)y(\tau)d\tau \quad (3)$$

$x(t-\tau)$는 자기 상관 계수이고 계산된 계수는 정규화 작업을 수행하며 0~1의 값을 가진다[5, 6].

## 3. 음성신호특징과 셉스트럼 특징 분포를 이용한 묵음특징 정규화

기존의 음성 인식 추출 방법들은 음성 인식 측정을 통해 구해진 결정 값들을 가지고 음성 구간을 결정하며, 기존 연구에서 음성 인식 추출 값에서 추가된 음성 신호의 잡음을 추정하여 음성과 비음성 구간을 정하는 방법을 사용하였다. 만일 음성 신호에 대한 잡음이 명확하게 측정되지 않는 경우에는 신호대 잡음 비율이 낮은 환경, 또는 여러 가지 잡음 환경에서 일정한 값을 가지지 않는 스레숄드(threshold) 값으로 인해 부정확한 음성 인식 추출 결과를 가진다. 또한, 음성 인식 처리를 위한 모델 훈련 환경이 다양하며, 이와 같은 환경의 차이로 발생하며, 이들에 대한 인식률 향상 방법으로 묵음 특징 정규화(silence feature normalization)을 사용한다. 묵음 특징 정규화 방법에서는 낮은 신호 대 잡음에서 묵음 구간의 에너지 레벨이 증가되므로 음성과 비음성에 대한 명확하지 않은 경계 분류 때문에 음성 인식 성능이 저하되는 문제점을 가지고 있다. 본 연구에서는 음성과 비음성에 대한 특징 추출을 묵음 특징 정규화를 융합한 음성 인식 성능 향상을 위한 방법을 모델링 한다. 제안한 방법은 높은 신호대 잡음비에서는 음성과 비음성에 대한 특징을 추출하기 위한 모델을 구성하여 음성의 특징이 잡음을 줄여 모델을 구성하였고, 낮은 신호대 잡음비에서는 음성과 비음성의 셉스트럼(cepstrum) 특징 분포 특성을 이용한 방법을 사용하여 음성 인식률을 높이는 방법을 사용한다. 음성 에너지 스펙트럼을 엔트로피와 유사하게 표현하므로 스펙트럴 에너지 영역을 엔트로피로 정의하여 사용한다. Shannon이 표현한 entropy 에 의해 계산되어 지는 spectral energy 확률의 변형된 수식은 다음과 같다.

$$P(|Y(k,l)^2|) = \frac{|Y(k.l)^2|}{\sum_{k=1}^{N/2}|Y(k.l)^2|} \quad (4)$$

k 값은 frequency bin index를 dmlalgkau, l은 프레임에 위한 색인 값을 의미한다. 프레임 l에서 주파수 빈 k에 대한 스펙트럴 에너지 확률을 계산하며, 계산된 각 주파수 빈의 확률은 엔트로피로 표현된다. 엔트로피는 DFT(Discrete Fourier Transform)를 가지고 처리하며, 이를 가지고 이산 스펙트럴 파워 에너지를 구하며, 에너지 스펙트럼 엔트로피 확률 수식은 다음과 같다.

$$S(|Y(k,l)|^2) = -\sum_{k=1}^{N/2}\{P(|Y(k,l)|^2)\cdot\log_2(P(|Y(k,l)|^2))\} \quad (5)$$

k는 주파수 빈(Frequency bin) 인덱스를 나타내고, l은 프레임 인덱스를 나타낸다. N은 심볼의 전체를 나타내고, Y(i)는 i의 심볼을 나타내며 P(i)는 심볼 i에 대한 사후 확률을 나타낸다.

엔트로피 확률 수식에 의해 분해된 음성 신호는 필터를 적용하여 음성밴드와 잡음밴드로 분리하고, 분리되어진 잡음 신호 프레임에서 음성 구간 밴드별 계수의 편차에 따라 특징 벡터를 추출한다. 잡음 신호의 셉스트럼 특징은 음성 신호에 비해 분산 값이 작게 나타나는 특성을 이용하였으며, 셉스트럼에서 음성과 비음성 분류의 기준 값을 정하여 신호 대 잡음 비율이 낮은 신호에서 묵음 특징 정규화로 성능을 향상시켰다. 정규화를 위해 작은 Log-energy을 갖는 비음성 특징을 찾고, 이를 IIR(Infinite Impulse Response) filter를 통과시켜서 사용한다. 음성 에너지 평균값을 계산하고, 음성과 비음성의 분류 기준 값을 T로 설정하여 계산된 음성 에너지 평균값이 T보다 크면 음성 구간으로 분류하고 로그에너지 값을 유지한다. 기준 값인 T 보다 작으면 비음성으로 분류하여 기준 값보다 작은 비음성 값으로 정규화를 시킨다. 음성 구간으로 구분된 프레임은 구해진 로그에너지 값을

유지하고, 묵음 구간으로 분류된 프레임은 작은 상수값으로 정규화하여 처리한다.

## 4. 실험 결과

음성 에너지 스펙트럼을 엔트로피와 유사하게 표현하여 잡음 환경에 효율적인 음성 인식 방법을 제안하였으며, 음성 인식에 대한 성능 분석 작업을 수행하기 위해 음성 인식에서 많이 사용되는 Aurora 2.0을 이용한 환경에서 실험을 수행하였다[13].

Aurora 2.0은 다양한 음성 인식 분석을 위한 비교를 위해 데이터베이스에 구축된 표준 음성과 음성에 대한 잡음을 포함하고, 이들에 대한 여러 단계의 발음을 내포하고 있으며, 포괄적으로 백색 가우시안 잡음과 바블 잡음 등을 포함한다. 본 연구의 음성 인식을 위한 음성 원본은 8kHz sampling rate, 16bit를 사용하며 FFT size는 256 표본, 1/2 오버래핑(overlapping) 영역을 사용하고 해민 윈도를 가지고 처리한다. 음성 인식 실험을 위해 여러 다양한 잡음 신호들로부터 생성된 잡음 음성을 포함하였으며, 이들 잡음 신호는 지하철, 자동차, 환경 잡음의 각 모델들로부터 특징 벡터를 추출하여 처리하였다. 본 연구에서 제안한 방법의 성능 분석을 위해 음성 인식 실험을 수행하였으며, 잡음 처리는 워너 필터를 사용하고, 어휘 인식 리스트는 서울 시내 지역명 30, 지하철역명 30 구성하였다. 또한, 음성 인식 실험을 위해 참가한 사람의 음성 인식 리스트를 세 번 발음하도록 하였으며, 음성 인식을 위한 50단어를 대상으로 하여 성능을 분석하였다.

<Table 1> Noise Environment Recognition Rate

| Speech | Recognition Rate (%) | | |
|---|---|---|---|
| | HMM | CHMM | Proposed Method |
| Speech Dependent | 82.3 | 84.3 | 85.3 |
| | 81.6 | 84.5 | 84.7 |
| | 82.7 | 84.2 | 85.5 |
| Average | 82.2 | 84.3 | 85.2 |
| Speech Independent | 81.3 | 82.3 | 84.3 |

논문에서 제시하는 방법의 성능 비교 분석을 위해 HMM과 CHMM을 비교하여 결과를 보였으며, <Table 1>은 잡음 환경에서의 실험 결과를 나타내며, 음성 종속 단계에서는 각 음성 인식 평균의 차이가 3%p와 0.9%p의 차이를 보이며, 음성 독립 단계에서는 2.8%p와 2.1%p의 음성 인식률 차이가 존재하므로 음성 종속 단계에서는 2.1%p의 인식률 향상이 있었으며, 음성 독립 단계에서는 0.7%p 만큼의 인식률 향상이 있었다.

## 5. 결론

음성 인식 처리를 위한 방법으로는 기본주파수를 이용한 방법, 통계적 모델을 사용한 필터를 가지고 음성 신호를 처리하는 방법, 그리고 임의 공간(subspace)을 사용하여 처리하는 방법을 이용한다. 이들 방법에서 음성 특징 추출은 음성 신호의 잡음을 추정하여 사용하므로 음성 신호의 잡음이 부정확한 경우에는 명확하지 않은 스레숄드 값으로 인해 부정확한 음성 인식률을 가진다. 본 연구에서는 음성과 비음성에 대한 특징 추출을 묵음 특징 정규화를 융합한 음성 인식 성능 향상을 위한 방법을 모델링 한다. 제안한 방법에서는 각 음성 프레임에 대해 음성 펄스 특징을 추출하여 음성 인식 모델을 구성하고 에너지 스펙트럼을 엔트로피와 유사하게 표현하여 원래의 음성 신호를 생성하여 음성의 특징이 잡음의 영향을 줄이도록 모델을 구성하였고, 낮은 신호대 잡음비에서는 음성과 비음성의 셉스트럼(cepstrum) 특징 분포 특성을 이용한 방법을 사용하여 음성 인식률을 높이는 방법을 제안하였다. 논문에서 제시하는 방법의 성능 비교 분석을 위해 HMM과 CHMM을 비교하여 잡음 환경에서의 실험 결과에서 음성 종속 단계에서는 각 음성 인식 평균의 차이가 3%p와 0.9%p의 차이를 보이며, 음성 독립 단계에서는 2.8%p와 2.1%p의 음성 인식률 차이가 존재하므로 음성 종속 단계에서는 2.1%p의 인식률 향상이 있었으며, 음성 독립 단계에서는 0.7%p 만큼의 인식률 향상이 있었다.

[1] J. C. Hwang. Voice Recognition Performance Improvement using the Convergence of bayesian

method and Selective Speech Feature Extraction. The Journal of the Korea Convergence Society. Vol. 7, No. 6, pp. 7-11, 2016.

[2] C. S. Ahn, S. Y. 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] C. S. Ahn, S. Y. Oh. Efficient Continuous Vocabulary Clustering Modeling for Tying Model Recognition Performance Improvement. Journal of the Korea Society of Computer and Information. Vol. 15, No. 1, pp. 177-183, 2010.

[4] C. S. Ahn, S. Y. 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.

[5] C. S. Ahn, S. Y. 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.

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

[7] S. M. Naqvi, M. Yu, J. A. Chamber. A Multimodal Approach to Blind Source Separation of Moving Sources. IEEE Trans. Signal Processing. Vol. 4, No. 5, pp. 895-910, 2010.

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

[9] S. Y. 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] Hirsch, H. G. & Pearce, D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in Proc. ICSLP. pp. 18-20. 2000.

[11] Young, S. HTK: Hidden Markov Model Toolkit V3.4.1. Cambridge University, Engineering Department, Speech Group. 1993.

[12] J. Y. Ahn, Sang-Bum Kim, Su-Hoon Kim, Kang-In Hur. A study on Voice Recognition using Model Adaptation HMM for Mobile Environment. The Journal of the Institute of Webcasting, Internet and Telecommunication. Vol. 11, No. 3, pp. 175-179, 2011.

[13] S. Y. Oh. Selective Speech Feature Extraction using Channel Similarity in CHMM Vocabulary Recognition. The Journal of digital policy and management. Vol. 11, No. 7, pp. 453-458, 2013.

[14] S. Y. 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] S. Y. 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] B. O. Kank, S. H. Lee, "Requirements Analysis in ID-based Future Internet," Journal of IT Convergence Society for SMB, Vol. 6, No. 3, pp. 43-48, 2016.

## 저자소개

황 재 천(Jae-CheonHwang) [정회원]

- 1981년 2월 : 동국대학교 수학과 (이학사)
- 1983년 8월 : 동국대학교 수학과 (이학석사)
- 2006년 3월 ~ 현재 : 가천대학교 IT대학 컴퓨터공학과 교수

<관심분야> : 음성/음향 신호 처리, 차량 통신