Journal of Convergence for Information Technology
Vol. 10. No. 8, pp. 35-39, 2020
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2020.10.08.035

GMM 음소 단위 파라미터와 어휘 클러스터링을 융합한 음성 인식 성능 향상

오상엽
가천대학교 컴퓨터공학과 교수

Speech Recognition Performance Improvement using a convergence of GMM Phoneme Unit Parameter and Vocabulary Clustering

SangYeob Oh
Professor, Division of Computer Engineering, Gachon University, Professor

요 약 DNN은 기존의 음성 인식 시스템에 비해 에러가 적으나 병렬 훈련이 어렵고, 계산의 양이 많으며, 많은 양의 데이터 확보를 필요로 한다. 본 논문에서는 이러한 문제를 효율적으로 해결하기 위해 GMM에서 모델 파라메터를 가지고 음소별 GMM 파라메터를 추정하여 음소 단위를 생성한다. 그리고 이를 효율적으로 적용하기 위해 특정 어휘에 대한 클러스터링을 통해 성능을 향상시키기 위한 방법을 제안한다. 이를 위해 3가지 종류의 단어 음성 데이터베이스를 이용하여 DB를 가지고 어휘 모델을 구축하였고, 잡음 처리는 워너필터를 사용한 특징을 추출하여 음성 인식실험에 사용하였다. 본 논문에서 제안한 방법을 사용한 결과 음성 인식률에서 97.9%의 인식률을 나타내었다. 본 연구에서 개선된 오버피팅의 문제점을 향상시킬 수 있는 추가적인 연구를 필요로 한다.

주제어 : DNN, GMM, 어휘 클러스터링, 음소 단위, 음소 추출

Abstract DNN error is small compared to the conventional speech recognition system, DNN is difficult to parallel training, often the amount of calculations, and requires a large amount of data obtained. In this paper, we generate a phoneme unit to estimate the GMM parameters with each phoneme model parameters from the GMM to solve the problem efficiently. And it suggests ways to improve performance through clustering for a specific vocabulary to effectively apply them. To this end, using three types of word speech database was to have a DB build vocabulary model, the noise processing to extract feature with Warner filters were used in the speech recognition experiments. Results using the proposed method showed a 97.9% recognition rate in speech recognition. In this paper, additional studies are needed to improve the problems of improved over fitting.

Key Words : DNN, GMM, Vocabulary Clustering, Phoneme unit, Phoneme character extract

*This research was supported by Global Infrastructure Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT(NRF-2018K1A3A1A20026485)
*Corresponding Author : Sang Yeob Oh (syoh1234@gmail.com)
Received June 19, 2020 Revised July 14, 2020
Accepted August 20, 2020 Published August 28, 2020

1. 서론

음성 인식의 대중화와 음성 인식 처리에 대한 다양한 방법의 발전으로 컴퓨터 사용 환경에서 음성 인식이 이전보다 많이 사용되고 있으며, 인터넷의 대중화와 음성 인식 기술의 향상으로 컴퓨터 사용 환경에서 음성 인식에 대한 많은 변화가 발생되고 있다. 음성 인식 기술은 HMM(Hidden Markov Model), CHMM(Continuous Hidden Markov Model), GMM(Gaussian mixture model), DNN 등의 발전으로[1-4] 사용자들이 대중적으로 사용하는데 있어 많은 진보가 이루어 졌으나 대중화에 있어 가장 큰 문제점은 휴대용 단말기에서 많은 다양한 음성 데이터에 대한 인식과 음성 인식의 효율적 최적화 처리[5-6]를 위한 잡음의 제거 문제가 가장 주요한 요소이다. 또한, 음성 인식의 특정 대상에 제한되고, 이의 처리 환경에 대한 하드웨어적인 환경과 특정 단어에 대한 인식률 문제와 데이터베이스의 제한된 용량의 인식 수행으로 나타난다.

DNN(Deep Neural Network)를 이용한 음성인식은 기존의 음성 인식 시스템에 비해 에러가 적으나 병렬 훈련이 어렵고, 계산의 양이 많으며, 음성 인식에 대한 데이터가 작으면 오버피팅(overfitting) 되기 때문에 많은 양의 데이터 확보를 필요로 한다. 또한, 상용화에 최소 1000시간 데이터를 훈련해야 한다. HMM을 이용하는 음성 인식 방법에서는 음성 인식에 대한 특정 모델을 작성하고, 이 모델이 가지는 음성 모델의 이산적인 분포를 사용하여 음성 인식에 대한 계산양이 적지만, 데이터베이스에 구성된 음성 데이터만을 처리하여 인식률이 낮은 단점을 가진다.[7-10] 이와 같은 문제를 해결하기 위해 본 논문에서는 DNN의 데이터 확보와 훈련 시간의 문제점을 해소하고, HMM의 인식률 개선을 위해 GMM의 음소 단위 파라미터와 어휘 클러스터링을 융합한 방법을 제안한다.

음성 인식에서 GMM 모델은 한 상태로 구성되는 특성을 가지므로 모음과 자음에 대한 확률 분포가 적어 데이터베이스 용량의 차이를 최소화하는 장점을 가지므로 음소를 이용한 연속 인식 작업에 적합한 특성을 이용한다, 또한, 주어진 음성을 가지고 음성의 유사도에 대한 모델 파라미터를 추정하고, 이를 효율적으로 적용하기 위해 특정 어휘에 대한 클러스터링을 통해 성능을 향상시키기 위한 방법을 적용하였으며, 이의 효율을 높이기 위하여 무음 구간과 음성구간 사이의 큰 에너지 변화를 이용하여 끝점을 검출하여 음성 인식의 성능을 높이도록 하였다. 어휘 구성을 위한 GMM 음소 단위 파라미터와 어휘 클러스터링을 융합한 모델 방법을 적용한 결과 어휘 인식률에서 잡음이 없는 환경에서는 97.9%, 잡음 환경에서의 인식률은 84.8의 인식률을 나타내었으며, 이는 기존의 방법보다 향상된 성능으로 DNN의 적은 데이터 처리 시에 발생되는 오버피팅에 대한 문제점을 해결할 있음을 확인하였다.

본 논문의 2장에서는 기존의 관련 연구인 HMM과 GMM에 대해 언급하고 3장에서는 본 논문에서 제안하는 GMM 파라미터 추정 방법과 음소 단위 어휘 특징 추출과 어휘 클러스터링 에 대해 설명한다. 4장에서는 기존의 방법들과 비교한 실험 분석을 수행하고 5장에서 결론을 맺는다.

2. 관련연구

2.1 HMM

HMM은 음성에 대한 Markov process 모델링 작업으로 음성 인식 과정에서 Markov 모델에서 사용하는 파라미터를 이용한다. 표준 Markov 모델을 작성하고 입력으로 사용되는 음성과 기억장치에 기억된 표준 Morkov 모델에 대한 유사도를 비교하여 가장 근접한 유사도 기준의 Markov 모델을 가지고 인식된 어휘로 결정하며, HMM 알고리즘에서는 인식 가능한 표준 음성 패턴을 음소와 음절 단위로 구분하여 추출 모델을 구성하고 인식하는 방법을 사용한다.[11-14].

기존 HMM 방법에서 사용하는 가우시안 확률 밀도 함수는 평균과 표준편차를 사용하여 확률에 대한 분포를 처리하며, 평균 μ과 표준편차 σ를 구하기 위하여 가우시안 확률 밀도 함수로 다음과 같이 표현한다.

$$b_i(y) = f(y; \theta_i) \quad (1)$$

가우시안 확률 밀도 함수는 2차원 이상의 다차원 식으로 표현이 가능하므로, n-차원에 대한 특징 벡터 x를 확률 변수로 한 가우시안 확률 밀도 함수를 다음과 같이 나타내며, μ와 Σ는 가우시안 분포에서 사용되는 주요 파라미터로 사용된다.

$$g_i,k(y)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_{i,k}|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}(y-\mu_i,k)^t\sum_{i,k}^{-1}(y-\mu_{i,k})\right) \quad (2)$$

식 (2)에서 n차원에 대한 가우시안 확률 밀도 함수는 n차원 공간의 한 점인 중심 μ으로 표현되고, Σ는 $n\times n$의 가역적인 양의 정부호 대칭 행렬로 나타내므로 $X^T\Sigma^{-1}X$의 형태로 표현된다,

2.2. GMM

GMM은 출력 확률밀도함수로 가우시안 밀도혼합 1개의 상태만을 가지고 작성된 CHMM의 한 형태로 사용된다. 이와 같은 GMM은 음향학적 클래스의 음성 집합을 표현할 수 있으며, 발성과 연관된 모음, 비음, 그리고 파찰음과 같은 음소를 나타내는 음성에 대한 분류의 집합으로 나타낸다. 또한, GMM에서 사용되는 단봉 가우시안 음소모델은 평균 벡터에 대한 특징벡터와 공분산을 가지고 각 음소의 특징을 나타내는 벡터 값을 가지고 음소에 대한 분포를 나타낸다. 이 방법의 장점은 가우시안 함수에 대한 이산집합을 이용하여, 각 평균과 공분산을 구하며, 가우시안 혼합 밀도는 M성분 밀도에 대한 가중합계이며, 식 (3)과 같다[5-6].

$$p(x|\lambda)=\sum_{i=1}^{M}c_ib_i(x) \quad (3)$$

x는 d-차원 랜덤 값을 가지는 벡터이며, $b_i(x),\ i=1,\ ...,\ M$는 i번째의 성분 밀도이고, $c_i,\ i=1,2,\ ...,\ M$는 i번째에 대한 혼합 밀도 가중치를 나타낸다.

3. 시스템 모델

시스템 모델은 GMM을 기반으로 하며, 이 모델은 표본 데이터 집합에 대한 분포를 확률밀도함수로 사용하는 가우시안 확률밀도함수로서 M개에 대한 가우시안 확률밀도함수는 다음과 같이 표현한다.

$$p(x|\theta)=\sum_{i=1}^{M}p(x|\omega_i,\theta_i)P(\omega_i) \quad (4)$$

$p(x|\omega_i,\theta_i)$는 데이터 x, ω_i번째 성분에 대한 파라미터 θ_i를 가지며, $P(\omega_i)$는 가중치 값으로서 사용되는 확률밀도함수에 대한 중요도를 나타낸다.

모델 학습을 위해 실험에 사용되는 음성에 대한 학습 벡터를 고려해야 하며, 이 학습벡터에 대한 파라미터를 추정해야 한다. 이를 위해 MLE(Maximum Likelihood Estimation)를 사용하여 주어진 음성 데이터에서 GMM의 유사도를 고려한 파라미터를 찾기 위해 이용되며, GMM 유사도에 대한 식은 다음과 같이 정의한다.

$$P(X|\lambda)=\prod_{t=1}^{T}p(x_t|\lambda) \quad (5)$$

T는 음성 학습 벡터이며, $X=x_1,x_2,\ ...,x_T$에 대한 열의 값을 로그영역에서 나타내어 GMM에서 사용한 식은 다음과 같다.

$$L(X|\lambda)=\sum_{i=1}^{T}\log p(x_t|\lambda) \quad (6)$$

GMM에서 사용하는 특정 파라미터의 값은 기본적으로 가우시안 분포를 사용하며, Σ는 가우시안 모델의 공분산 매트릭스를 나타낸다.

GMM 음소 학습단계를 위해 학습된 음성 모델에 대한 연속적인 자동 음소 모델 구분을 위해 CHMM 모델에 의한 자동 음소분할로 구한 라벨 정보를 사용하여 음소 단위 데이터베이스를 구축하였으며, 이를 각각의 음소 단위에 대한 GMM 파라미터를 추정하여 사용한다. 음소 인식을 위해 음소에 대한 GMM의 평균, 공분산, 그리고 CHMM에 대한 중간 상태 천이 확률을 사용하여 연속 음소 인식 작업을 수행한다. GMM은 1 상태로 구성되는 상대적 특성으로 모음과 자음의 데이터베이스 용량 차이에서 발생하는 확률 분포의 차이를 최소화하는 장점을 가지므로 음소를 이용한 연속 인식 작업에 적합하다.

음소에 대한 클러스터링을 위한 대상 객체들은 클러스터링 작업을 수행하여 특정 군집에 포함되며, 각 군집에서는 포함된 객체들의 속성에 대한 정보를 가진다. 음소 객체에 대한 클러스터링 결과 분석을 위해

k-means 기법을 사용한다. 이를 위해 가장 근사한 거리에 있는 중심점을 가지는 군집에 대상 음소를 배분하는 단계를 반복으로 사용하여 k개의 군집에 대한 내용들로 분배한다. 이를 위해 거리에 기반을 가지는 클러스터링 방법을 사용하여 선호도를 다차원 공간에 존재하는 점을 사용하여 나타내고, 이들 거리를 계산하여 음소에 대한 집합을 k개의 군집으로 표현할 수 있다. 음소 모델 a와 k에 대한 거리는 다음 식과 같이 적용한다.

$$d_{a,k} = \sqrt{\sum_{t}(a_t - k_t)^2} \qquad (7)$$

a_j는 모델 a의 속성 i에 대한 선호도 값을 의미한다.

4. 실험 결과

본 연구에서 제안한 시스템의 음성 인식 성능 향상 시스템은 세 가지 종류의 단어 음성 데이터베이스인 ETRI의 PBW445 데이터베이스, POW3848 데이터베이스, 국어공학연구소의 PBW452 데이터베이스를 사용하였으며, GMM의 구조에 적합한 3가지 상태와 Gaussian Mixture의 개수를 결정하여 Left-to-Right의 상태를 갖는 유사 음소단위에 대한 문맥 종속 트라이폰(triphone)을 기반으로 가변 어휘에 대한 인식을 위한 음소 모델 훈련을 데이터베이스에 적용하였다.

본 연구의 성능 평가를 수행하기 위해 기존 방식에서 사용되는 유클리디안 알고리즘, 가우시안 Maximum Log Likelihood등과 비교하여 제안한 방법과의 음성 인식률에 대한 비교를 한 결과 , DTW 알고리즘, 제안 방법에 대해 인식률을 측정하였다.

Table 1은 잡음이 없는 실내 환경에서 기존 방식인 유클리디안 알고리즘, Maximum Log Likelihood 방법, 그리고, 제안한 방법에 대한 실내 환경의 적용 결과를 나타낸다. Table 1에서 나타내는 것과 같이 유클리디안 알고리즘을 이용한 음성 인식률 평균 95.3%로 나타났으며 Maximum Log Likelihood 방법을 이용한 음성 인식률 평균 97.3%의 인식률 나타내었고 , 본 연구에서 제안한 방법에 대한 인식률 평균은 97.7%를 나타내었다.

Table 1. Non-Noise Environment Recognition Rate

Speech	Recognition Rate (%)		
	Euclidean	MLE	Proposed Method
Speech Dependent	95.2	96.8	97.1
	94.7	97.1	97.3
	95.6	98.1	98.3
Speech Independent	95.6	97.3	98.3
	95.1	96.9	97.6
	95.8	97.3	97.7

Table 2는 잡음이 있는 실내 환경에서 기존 방식인 유클리디안 알고리즘, Maximum Log Likelihood 방법, 그리고, 제안한 방법에 대한 실내 환경의 적용 결과를 나타낸다.

Table 2. Noise Environment Recognition Rate

Speech	Recognition Rate (%)		
	Euclidean	MLE	Proposed Method
Speech Dependent	82.5	85.1	85.7
	81.4	83.3	84.9
	82.1	85.2	85.3
Speech Independent	80.8	82.5	83.9
	81.1	83.2	83.7
	80.1	83.9	85.1

Table 2에서는 잡음 환경에서 실험한 결과 유클리디안 알고리즘을 이용한 평균 음성 인식률 평균은 81.3%로 나타났으며 Maximum Log Likelihood 방법을 이용한 음성 인식률 평균 83.9%의 인식률 나타내었고 ,제안방법의 인식률 평균은 84.8%를 나타내었다. Table 1과 Table 2에서 각각 기존의 방법들의 평균과 비교하여 각 각 04%, 0.9%의 향상을 보였으며, 특히 잡음 환경에서의 유의미한 0.9%의 향상은 기존의 방법보다 향상된 성능으로 DNN의 적은 데이터 처리 시에 발생되는 오버피팅 문제를 해결하는데 본 논문의 방법이 효율적임을 나타낸다.

5. 결론

본 연구에서는 기존의 DNN, HMM 등의 방법에서 사용되는 음성 인식의 성능을 향상시키기 위해 GMM에서 주어진 음성을 가지고 음성의 유사도에 대한 모델 파라미터를 추정하고, 이를 효율적으로 적용하기 위해 특정 어휘에 대한 클러스터링을 통해 성능을 향상시키기 위한 방법을 적용하였으며, 이의 효율을 높이기 위하여 무음 구간과 음성구간 사이의 큰 에너지 변화를 이용하여 끝점을 검출하여 음성 인식의 성능을 높이도록 하였다. 어휘 구성을 위한 GMM 음소 단위 파라미터와 어휘 클러스터링을 융합한 모델 방법을 적용하여 실험한 결과 어휘 인식률에서 잡음이 없는 환경에서는 97.9%, 잡음 환경에서의 인식률은 84.8의 인식률을 나타내었으며, 이는 기존의 방법보다 향상된 성능으로 DNN의 적은 데이터 처리 시에 발생되는 오버피팅의 문제점을 해결할 수 있음을 확인하였으며, 이를 개선할 방법에 대한 연구를 필요로 한다.

REFERENCES

[1] C. S. Ahn & S. Y. Oh. (2012). Gaussian model optimizationusing configuration thread control In CHMM vocabulary recognition. *Journal of Digital Policy and Management, 10(7),* 167-172. DOI : 10.14400/JDPM.2012.10.7.167

[2] C. S. Ahn & S. Y. Oh. (2012). Echo noise robust HMM learning model using average estimator LMS algorithm. *Journal of Digital Policy and Management, 10(10),* 277-282. DOI : 10.14400/JDPM.2012.10.10.277

[3] C. S. Ahn & S. Y. Oh. (2012). CHMM modeling using LMS algorithm for continuous speech recognition improvement. *Journal of Digital Policy and Management. 10(11),* 377-382. DOI : 10.14400/JDPM.2012.10.11.377

[4] S. Y. Oh & K. Chung. (2018). Performance evaluation of silence-feature normalization model using cepstrum features of noise signals. *Wireless Personal Communications, 98(4),* 3287-3297. DOI : 10.1007/s11277-017-4645-x

[5] K. Chung & S. Y. Oh. (2016). Vocabulary optimization process using similar phoneme recognition and feature extraction. *Cluster Computing, 19(3),* 1683-1690. DOI : 10.1007/s10586-016-0619-0

[6] K. Chung & S. Y. Oh. (2015). Improvement of speech signal extraction method using detection filter of energy spectrum entropy. *Cluster Computing, 18(2),* 629-635. DOI : 10.1007/s10586-015-0429-9

[7] C. S. Ahn & S. Y. Oh. (2010). Vocabulary recognition post-processing system using phoneme similarity error correction. *Journal of the Korea Society of Computer and Information. 15(7),* 83-90. DOI : 10.9708/jksci.2010.15.7.083

[8] M. F. Gales. (1995). *Model-based techniques for nosie robust speech recognition,* Ph. D. dissertation, University of Cambridge.

[9] A. S. Manos & V. W. Zue. (1996). *A study on out-of-vocabulary word modeling for a segment-based keyword spotting system.* Master Thesis, MIT.

[10] T. Jitsuhiro, S. Takatoshi & K. Aikawa. (1998). Rejection of out-of-vocabulary words using phoneme confidence likelihood. In Proc of the IEEE *International Conference on Acoustics, Speech and Signal Processing,* 217-220.

[11] K. Chung & S. Y. Oh. (2016). Voice activity detection using improvement unvoiced feature normalization process in noisy environment. *Wireless Personal Communications, 89(3),* 747-759. DOI : 10.1007/s11277-015-3169-5,

[12] S. Y. Oh & K. Chung. (2014). Target speech feature extraction using non-parametric correlation coefficient. *Cluster Computing, 17(3),* 893-899. DOI : 10.1007/s10586-013-0284-5

[13] S. Y. Oh & K. Chung. (2014). Improvement of speech detection using ERB feature extraction. *Wireless Personal Communications, 79(4),* 2439-2451. DOI : 10.1007/s11277-014-1752-9

[14] J. C. Kim & K. Chung. (2018). Mining health-risk factors using PHR similarity in a hybrid P2P network. *Peer-to-Peer Networking and Applications, 11(6),* 1278-1287. DOI : 10.1007/s12083-018-

오 상 엽(Sang Yeob Oh) [정회원]

· 1991년 2월 : 광운대학교 대학원 전자계산학과 (이학석사)
· 1999년 2월 : 광운대학교 대학원 전자계산학과 (이학박사)
· 2007년 2월 ~ 현재 : 가천대학교 IT대학 인터랙티브미디어학과 교수

· 관심분야 : 인공지능, HCI, 차량 통신, 형상관리, 음성 및 음향 신호처리, 정보검색, 추천 시스템, 기계학습
· E-Mail : syoh1234@gmail.com