Journal of Digital Convergence
Vol. 16. No. 11, pp. 335-340, 2018

ISSN 1738-1916
https://doi.org/10.14400/JDC.2018.16.11.335

가우시안 분포에서 Maximum Log Likelihood를 이용한 벡터 양자화 기반 음성 인식 성능 향상

정경용[1], 오상엽[2*]
[1]경기대학교 컴퓨터공학부 교수, [2]가천대학교 컴퓨터공학과 교수

Vector Quantization based Speech Recognition Performance Improvement using Maximum Log Likelihood in Gaussian Distribution

Kyungyong Chung[1], SangYeob Oh[2*]
[1]Division of Computer Science and Engineering, Kyonggi University, Professor
[2]Division of Computer Engineering, Gachon University, Professor

요 약 정확한 인식률을 보이고 있는 상업적인 음성인식 시스템은 화자종속 고립데이터로부터 학습 모델을 사용한다. 그러나 잡음 환경에서 데이터양에 따라 음성인식의 성능이 저하되는 문제점이 있다. 본 논문에서는 가우시안 분포에서 Maximum Log Likelihood를 이용한 벡터 양자화 기반 음성 인식 성능 향상을 제안한다. 제안하는 방법은 음성에 대한 특징을 가지고 벡터 양자화와 Maximum Log Likelihood 음성 특징 추출 방법을 이용하여 유사 음성에 대한 음성 인식의 정확성을 높이는 최적 학습 모델 구성 방법이다. 이를 위해 HMM을 기반으로 음성 특징을 추출하는 방법을 사용한다. 제안하는 방법을 사용하여 기존 시스템에서 생성되어 사용되는 음성 모델에 대한 부정확한 음성 모델에 대한 정확성을 향상시킬 수 있으므로 음성 인식에 강인한 모델을 구성할 수 있다. 제안하는 방법은 음성 인식 시스템에서 향상된 인식의 정확도를 보인다.

주제어 : 음성 인식, HMM, 특징 추출, 음성 모델, 가우시안 분포

Abstract Commercialized speech recognition systems that have an accuracy recognition rates are used a learning model from a type of speaker dependent isolated data. However, it has a problem that shows a decrease in the speech recognition performance according to the quantity of data in noise environments. In this paper, we proposed the vector quantization based speech recognition performance improvement using maximum log likelihood in Gaussian distribution. The proposed method is the best learning model configuration method for increasing the accuracy of speech recognition for similar speech using the vector quantization and Maximum Log Likelihood with speech characteristic extraction method. It is used a method of extracting a speech feature based on the hidden markov model. It can improve the accuracy of inaccurate speech model for speech models been produced at the existing system with the use of the proposed system may constitute a robust model for speech recognition. The proposed method shows the improved recognition accuracy in a speech recognition system.

Key Words : Speech Recognition, HMM, Feature Extraction, Speech Model, Gaussian Distribution

*This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW(R7015-16-1003) supervised by the IITP(Institute for Information & communications Technology Promotion"(R7015-16-1003)
*Corresponding Author : Sang Yeob Oh (syoh1234@gmail.com)
Received October 10, 2018 Revised November 5, 2018
Accepted November 20, 2018 Published November 28, 2018

1. 서론

컴퓨터에서 사용되는 음성 인식 연구는 언어학, 음성학, 음운학, 자연어처리와 같은 여러 분야의 내용을 근거로 하고 있다. 컴퓨터 분야의 하드웨어와 소프트웨어 테크놀로지의 발전과 음성 데이터의 분산 프로세스 및 분석 기술의 발달, 그리고 사물 인터넷의 발전에 의해 음성 인식의 고도화 기술이 개발되고 있다[1-3].

음성 인식 처리를 위해서는 음성 인식에 대한 음성이 지정되면, 인식하고자 하는 여러 사람의 음성에 대한 발성을 구축하고, 이를 데이터베이스에 저장하여[4] 인식하고자 하는 인식 대상 어휘에 대한 음성과 유사 음성의 단위 모델과 인식 작업을 수행한다. 이미 생성된 단위 모델에서 인식되지만 사용 중에 표현되지 않는 음성이 나타날 수 있다. 이러한 결과로 음성에 대한 부수적인 변경 작업 및 전처리 작업이 수행될 수 있다. 여기서 전처리 작업은 전체 시스템의 오버헤드 문제를 발생시킨다. 인식하고자 하는 대상 음성이 삽입 및 수정이 되면, 기존의 음성 데이터베이스에 대해 추가적인 음성의 수집으로 인해 많은 시간과 비용에 대한 오버헤드가 발생한다[9,10]. 음성 인식 시스템에서 사용되는 유사한 음성 모델의 처리 및 인식 문제와 사용자의 부정확한 음성에 대한 인식 오류가 발생한다. 입력되는 유사한 음성은 전체 시스템에서 인식률의 저하 및 성능 문제가 있다. 음성 인식은 파형에 대한 패턴 인식을 기반으로 하고 있다. 여기서 벡터 거리의 개념은 패턴이 공간상에서 상호간에 서로 분리된 정도를 통해 패턴들 사이의 비슷한 정도를 계산한다. 서로 인접되어 있는 형태는 유사한 특징을 가지며, 이 경우에 유사도가 높은 특징을 가진다[13-18]. 이러한 거리 측정 방법을 위해 바타챠랴 거리 측정 방법이 일반적으로 사용되며, 유클리디안과 DTW 방법도 사용되고 있다[5,6,8].

유클리디안 알고리즘은 다양한 속성 변수의 거리를 측정하기 위해 이용하며, 개체에 대한 속성이 다양한 경우에는 각 속성 변수 사이의 값들에 대한 유사도를 구한다. DTW 알고리즘은 비선형 시간 정규화에 대한 패턴 정합 알고리즘을 사용하며, 공통적이고 균일한 샘플간격을 가지는 음성패턴을 대상으로 시간에 대한 샘플을 가지고 처리한다. 바타챠랴 거리 측정 방법은 에러율을 가지고 거리를 측정하는 방법이며, 단순한 거리 계산을 구하는 방법으로 동적 프로그램을 사용하며, 포워드 및 백워드 확률 연산을 이용하는 경우에는 인식률이 높으나 추정 계산량과 복잡도는 증가한다. 비터비 코딩을 사용하는 경우에는 인식률이 저하되지만 연산의 부하가 상대적으로 적으므로 인식과정에서 많이 사용된다.

본 논문에서는 음성이 갖는 특징을 기반으로 학습 데이터의 음성에 가우시안 특징 추출 방법을 사용하고, 모노폰으로 훈련시킨 훈련 음성 데이터의 벡터 양자화와 Maximum Log Likelihood을 이용한 특징 추출로 음성 인식을 향상시켰다. 이를 확인하기 위해 성능 평가에서는 다른 방법과의 결과를 비교하여 하였으며, 비교한 시스템의 성능 평가 결과 향상된 인식률을 나타내었다.

본 논문의 구성은 다음과 같다. 2장에서는 은닉 마르코프 모델((HMM, hidden markov model)을 이용한 음성 인식과 가우시안 모델에 대해 기술하고, 3장에서는 음성 신호의 단위를 위해 벡터 단위를 사용하며, 이를 Maximum Log Likelihood로 사용하여 학습된 데이터와 유사도가 가장 높은 것을 탐색한다. 4장에서는 제안한 시스템의 음성 인식 성능 평가를 제시한다.

2. 관련연구

2.1 은닉 마르코프 모델을 이용한 음성 인식

음성 신호가 가지는 특정 신호에 대한 유한개의 상태와 상태전이를 사용하는 은닉 마르코프 모델((HMM, hidden markov model)은 스펙트럼의 시간에 대한 변경 내용을 추적이 가능하다[8]. HMM 인식 단계에서는 주어진 음성 신호 데이터를 이용하여 음성 신호에 대한 파라미터를 가지고 입력된 음성신호에 대해 가장 적합한 모델을 구한다. HMM에서는 음성 신호의 연속적인 일련의 상태를 가지고 이산 신호를 산출하는 확률을 사용하며, 이러한 모델은 음성이 전이되는 확률을 가지고 음성 신호의 상태를 변경하게 된다. 특정한 상태에 따라 출력 확률이 정해지며, 이에 의한 음성 신호의 관측을 수행하게 된다.

HMM 모델에 대한 작업은 음성 신호에 대한 파라미터를 사용하며, 음성신호에 대한 분류자를 저장한 음성 데이터베이스를 이용하므로 음성 신호에 대한 다양한 특징이 존재하는 경우에는 실제 음성이 가지는 차이를 나타낼 수 있는 유용한 방법이다[7]. Fig. 1은 HMM의 3상태를 나타낸다. HMM의 음성 인식 처리는 특징 벡터 x에

대한 확률을 가지고 전향 알고리즘과 후향 알고리즘으로 분류하여 처리할 수 있다. 특징 벡터 x에 대한 은닉 상태 열을 찾는 관측 열 $O=\{o_1,o_2,\cdots,o_T\}$과 모델 $\lambda=(A,B,\pi)$에서 최적의 상태 열 $Q=\{q_1,q_2,\cdots,q_T\}$을 탐색하기 위한 방법으로는 비터비 알고리즘을 이용한다[8]. HMM 모델 학습 $P(O|\lambda)$를 최대화하기 위한 방법으로 모델 매개 변수 $\lambda=(A,B,\pi)$를 조정한 바움-웰치 알고리즘을 사용하며, A와 B는 확률매개변수이고, 다음 그림에서 특징 벡터 x에 대한 은닉 상태 열을 찾는 관측 열 부분이 a_{11}부터 a_{33}이 된다.

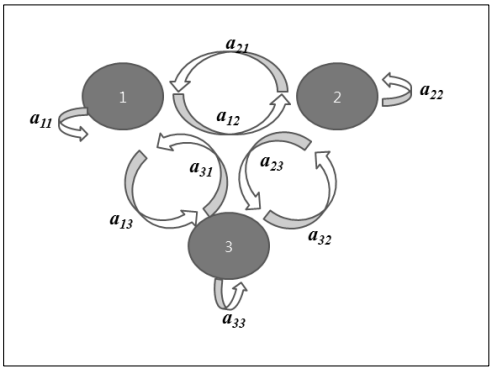

Fig. 1. 3-State of HMM

2.2 가우시안 모델

음성 데이터는 다양한 환경에 따라 수집이 되어 활용 가능한 형태로 변환하는 모델링 과정이 필요하다. 이는 수집한 음성 데이터를 활용하기 형태로 정제하는 모델링 과정이다. 확률 밀도 함수를 사용하여 3상태의 음성 모델을 생성한 훈련 모델을 사용한다. 모델에서 사용하는 음성 모델에 대해 대응되는 3상태 모델에 대한 출력 분포를 가지고, 분포의 집합에서 선택되는 특정 상태들에 대해 모델의 정확도를 향상한다[7]. 모델을 대표하는 각 가우시안 성분은 정방 행렬의 형태에서 음성에 대한 학습 데이터 집합을 가지고 사용하며, 이 경우 음성에 대한 크기와 가우시안의 혼합 성분의 개수를 변경하여 가우시안 혼합 모델을 구성하여 사용한다. 가우시안 혼합 모델을 가지고 음성에 대한 데이터의 분포를 모델링하면 모델에서 사용하는 음성 데이터가 충분히 생성되고 음성에 대한 파리미터 값의 수정을 통해 연속적인 분포 추정 모델로 사용한다[11,12].

3. 시스템 모델

음성의 성능 향상을 위한 처리 과정은 모노폰을 이용한 적용 데이터의 음성을 사용하여 특징을 추출하는 방법을 사용한다. 추출되는 음성의 특징에 대해 유사한 음성을 수집한 후에 대푯값을 추출한다. 음성 유사율은 2개의 음성 사이의 거리를 의미하며, 음성들에 대한 분리 상태를 나타내는 통계적 의미로 사용되며, 가우시안 분포 사이의 거리를 계산한다. 이를 사용하여 음성 사이의 계산 과정을 단순화하며, 오류 발생에 대한 경계 값을 처리하여 유연성을 제공할 수 있으며, 식(1)은 음성 사이에 대한 정의를 나타난다.

$$b=-\ln\int_{\Omega}\left[P(X|w_1)P(X|w_2)\right]^{\frac{1}{2}}dX \tag{1}$$

$P(X|w_i)$는 음성 $w_i(i=1,2)$에 대한 확률 밀도 함수를 나타내고, Ω는 확률 분포에서 처리되는 랜덤 함수 X의 분포이다. 음성에 대한 확률 분포를 정규 분포로 처리할 경우 식(2)과 같이 나타난다.

$$b=\frac{1}{8}(u_2-u_1)^T\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(u_2-u_1)+\frac{1}{2}ln\frac{\left|(\Sigma_1+\Sigma_2)/2\right|}{\left|\Sigma_1\right|^{1/2}\left|\Sigma_2\right|^{1/2}} \tag{2}$$

u_1와 Σ_i는 음성 w_i에 대한 평균적인 값과 분산에 대한 행렬 값을 의미하며, 식(3)을 이용하여 유사율을 측정하고, 식(4)의 S_p는 0과 1 사이 분포를 정규화한다. $b_{\max}$와 $b_{\min}$는 음성이 가지는 최대값과 최소값을 의미하며, b_{xy}는 두 개의 음성 x와 y에 대한 거리를 의미한다. 유사한 음성에 해당될수록 1에 가까운 값이 표현되고, 반대의 경우에는 0에 유사한 값을 가진다.

$$d(i,j)=\left[\frac{1}{n}\sum_{k=1}^{n}\frac{(\mu_{ik}-\mu_{jk})^2}{\sigma_{ik}\,\sigma_{jk}}\right]^{\frac{1}{2}} \tag{3}$$

$$S_p=\frac{b_{\max}-b_{xy}}{b_{\max}-b_{\min}} \tag{4}$$

음성 신호의 단위를 위해 벡터 단위를 사용하며, 이 단위는 음성 신호의 학습 과정에서 블록 단위로 구분하는 최소 단위가 된다. 이를 위해 음성의 특징 변화가 거의 발생하지 않는 20~30ms 구간에 대한 샘플 사용한다.

가우시안 모델에서 초기에 균일한 상태를 가지는 확률은 상태가 변화되는 전이확률로 나타내며, 관측 확률은 하나의 상태에서 다른 이벤트가 발생할 확률을 의미한다. 이를 위해, $A=a_{ij}$ 행렬, $B=b_j(o_t)$ 집합, $\pi=\pi_j$ 벡터로 표기를 정의하여 전체 모델을 $\lambda=(A,B,\pi)$로 나타낸다. 만약 $o=\{o_1, o_2, ..., o_T\}$, $q=\{q_1, q_2, ..., q_T\}$라고 표현 할 때 λ모델에서 $o=\{o_1, o_2, ..., o_T\}$가 차례대로 발생할 확률은 식(5)와 같이 표현된다.

$$P[o|\lambda]=\sum_{q=0}^{n} P[o|\lambda,\lambda]P[o|\lambda] \quad (5)$$

이와 같이 표현되는 상태 표현은 가우시안 분포에서 벡터 양자화를 이용하여 MFCC(Mel Frequency Cepstral Coefficient)와 같은 2차원의 다양한 값을 정량화하여 표현하기 위해 사용한다[19,20,21]. 각각의 주어진 음성 단어와 음절을 이용하여 적용하며, 학습 이후의 시험 과정을 각각의 음성에 대해 적용하여 이미 학습된 데이터와 유사도가 가장 높은 것을 탐색하며, 이를 Maximum Log Likelihood로 사용한다.

4. 시스템 평가

본 논문에서는 음성이 갖는 특징을 기반으로 학습 데이터의 음성에 가우시안 특징 추출 방법을 사용하였고, 모노폰으로 훈련시킨 훈련 데이터의 음성에 벡터 양자화와 Maximum Log Likelihood 방법을 사용하여 특징 추출한다. 제안한 시스템의 음성 인식의 성능 평가를 위해 실험 환경은 워너 필터를 이용하고, 서울 시내 지역명 50개와 지하철역명 50개를 임의로 선택하여 음성 인식 목록으로 구성한다. 인식을 평가를 위해 [13]의 평가 내용을 기반으로 하였으며, 평가에 3명의 참여자가 음성 인식 단어 내용을 3회 이상 제시한 총 700단어를 대상으로 진행한다. 또한, 음성 인식 실험을 위해 캠브리지 대학의 HTK (Hidden Markov models toolKit)[12,21]를 사용하여 실험한다. 제안한 방법의 성능 분석을 처리하기 위해 유클리디안 방법, DTW 방법과 비교하여 인식률을 측정한다.

Table 1과 Table 2는 실내 환경과 실외 환경에서의 실험 결과를 나타낸다. 화자 종속(speech dependent)은 이리 정의된 특정 화자만을 대상으로 하며, 화자 독립(speech independent)은 불특정 다수를 대상으로 하며, 음성 특징을 찾기 위한 데이터베이스를 필요로 한다. 표 1은 잡음이 발생하지 않는 실내에서 인식률을 실험하였으며, 유클리디안 방법, DTW 방법, 그리고 제안한 음성 인식률은 평균 95.67%, 96.63%, 97.25%를 나타낸다. 인식률에 따른 학습 시간은 유클리디안 방법, DTW 방법, 그리고 제안한 음성 인식률은 평균 1.26초, 1.28초, 1.18초를 나타낸다. 표 2의 잡음 환경의 실험 결과에서 유클리디안 방법, DTW 방법, 그리고 제안한 음성 인식률은 평균 85.28%, 84.87%, 85.70%를 나타내었다. 인식률에 따른 학습 시간은 유클리디안 방법, DTW 방법, 그리고 제안한 음성 인식률은 평균 1.61초, 1.65초, 1.53초를 나타내어 인식률에서 다소 개선됨을 확인할 수 있다.

Table 1. Non-noise environment recognition rate

Speech	Euclidean		DTW		Proposed Method	
	Recog-nition Rate (%)	Recog-nition Time (sec)	Recog-nition Rate (%)	Recog-nition Time (sec)	Recog-nition Rate (%)	Recog-nition Time (sec)
Speech De-pendent	95.3	1.1	96.1	1.3	96.8	1.2
	95.7	1.3	97.7	1.2	97.1	1.1
	96.1	1.2	97.3	1.3	98.1	1.2
Speech Inde-pendent	95.3	1.3	96.1	1.4	97.3	1.2
	95.7	1.4	96.1	1.2	96.9	1.3
	95.9	1.3	96.5	1.3	97.3	1.1

Table 2. Noise environment recognition rate

Speech	Euclidean		DTW		Proposed Method	
	Recog-nition Rate (%)	Recog-nition Time (sec)	Recog-nition Rate (%)	Recog-nition Time (sec)	Recog-nition Rate (%)	Recog-nition Time (sec)
Speech De-pendent	86.3	1.4	85.6	1.4	86.1	1.5
	87.5	1.3	86.1	1.6	87.3	1.4
	86.7	1.5	86.0	1.5	86.2	1.3
Speech Inde-pendent	83.4	1.7	83.7	1.8	84.5	1.7
	83.7	1.9	83.5	1.7	85.2	1.6
	84.1	1.9	84.3	1.9	84.9	1.7

5. 결론

본 논문에서는 가우시안 분포에서 Maximum Log Likelihood를 이용한 벡터 양자화 기반 음성 인식 성능 향상을 제안하였다. 음성 인식의 성능 향상을 하기 위해 벡터 형태를 이용한 패턴 인식을 사용한다. 서로 인접되어 있는 2개의 형태는 거의 유사한 특징을 가지며, 이 경

우에 유사도가 높은 특징을 가진다. 이러한 거리 측정 방법을 시스템 평가에서 기존의 유클리디안과 DTW를 제안하는 방법과 성능평가를 진행하였고, 비교한 시스템 성능 평가 결과 향상된 인식률을 나타내었다. 음성이 갖는 특징 벡터를 기반으로 학습 데이터에 가우시안 분포에서 벡터 양자화 기반의 특징 추출에서 Maximum Log Likelihood 방법을 이용하였을 때 기존 방법과 비교하여 음성 인식의 성능이 향상된 것으로 나타났다. 이를 통해 기존 시스템에서 생성되어 사용하는 부정확한 음성 모델에 대한 정확성을 향상시킬 수 있다.

REFERENCES

[1] C. S. Ahn & S. Y. Oh. (2012). Gaussian model optimization using configuration thread control In CHMM vocabulary recognition. *Journal of Digital Policy and Management, 10(7),* 167-172.

[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.

[3] 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.

[4] W. Kim & H. T. Chou. (1988). Versions of schema for object-oriented databases. *In Proc. of 14th International Conference on Very Large Data Base.* 148-159.

[5] C. S. Ahn & S. Y. Oh. (2010). Phoneme similarity error correction system using bhattacharyya distance measurement method. *Journal of the Korea Society of Computer and Information, 15(6),* 73-80.

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

[7] W. Reichl & W. Chou. (1998). Decision tree state tying based on segmental clustering for acoustic modeling. Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing. 801-804.

[8] 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.

[9] R. Agrawal, S. J. Buroff, N. H. Gehani & D. Shasha. (1991). Object versioning in ode. In Proc. of 7th International Conference on Data Engineering, 446-455.

[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. E. Gorlen. (1987). An object-oriented class library for C++ program. s*oftware-practice and experience, 17(12),* 899-922.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, Valtcher & P. Woodland. (2002). The HTK Book, Cambridge University Engineering Department.

[13] 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.

[14] S. Y. Oh & K. Chung. (2014). Target speech feature extraction using non-parametric correlation coefficient. *Cluster Computing, 17(3),* 893-899.

[15] S. Y. Oh & K. Chung. (2014). Improvement of speech detection using ERB feature extraction. *Wireless Personal Communications, 79(4),* 2439-2451.

[16] K. Chung & S. Y. Oh. (2016). Vocabulary optimization process using similar phoneme recognition and feature extraction. *Cluster Computing, 19(3),* 1683-1690.

[17] 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.

[18] 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.

[19] H. Yoo & K. Chung. (2018). Mining-based lifecare recommendation using peer-to-peer dataset and adaptive decision feedback. *Peer-to-Peer Networking and Applications, 11(6),* 1309-1320.

[20] 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.

[21] 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.

정 경 용(Chung, Kyung Yong) [정회원]

- 2000년 2월 : 인하대학교 전자계산공학과 (공학사)
- 2002년 2월 : 인하대학교 전자계산공학과 (공학석사)
- 2005년 8월 : 인하대학교 컴퓨터정보공학부 (공학박사)
- 2006년 3월 ~ 2017년 2월 : 상지대학교 컴퓨터정보공학부 교수
- 2017년 3월 ~ 현재 : 경기대학교 컴퓨터공학부 교수
- 관심분야 : 데이터 마이닝, 헬스케어, 빅데이터, 지능시스템, 인공지능, HCI, 정보검색, 추천 시스템
- E-Mail : dragonhci@gmail.com

오 상 엽(Oh, Sang Yeob) [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