한국정보전자통신기술학회논문지(jkiiect)'21-4, Vol.14 No.2
논문 21-14-02-134 http://dx.doi.org/10.17661/jkiiect.2021.14.2.134

자율차량 안전을 위한 긴급상황 알림 및 운전자 반응 확인 시스템 설계

손수락*, 정이나**

A Design of the Emergency-notification and Driver-response Confirmation System(EDCS) for an autonomous vehicle safety

Su-Rak Son*, Yi-Na Jeong**

요 약 현재 자율주행차량 시장은 3레벨 자율주행차량을 상용화하고 있으나, 여전히 운전자의 주의를 필요로 한다. 3레벨 자율주행 이후 4레벨 자율주행차량에서 가장 주목되는 부분은 차량의 안정성이다. 3레벨과 다르게 4레벨 이후의 자율주행차량은 운전자의 부주의까지 포함하여 자율주행을 실시해야 하기 때문이다. 따라서 본 논문에서는 운전자가 부주의한 상황에서 긴급상황을 알리고 운전자의 반응을 인식하는 자율차량 안전을 위한 긴급상황 알림 및 운전자 반응 확인 시스템을 제안한다. 긴급상황 알림 및 운전자 반응 확인 시스템은 긴급상황 전달 모듈을 사용하여 긴급상황을 텍스트화하여 운전자에게 음성으로 전달하며 운전자 반응 확인 모듈을 사용하여 긴급상황에 대한 운전자의 반응을 인식하고 운전 권한을 운전자에게 넘길지 결정한다. 실험 결과, 긴급상황 전달 모듈의 HMM은 RNN보다 25%, LSTM보다 42.86% 빠른 속도로 음성을 학습했다. 운전자 반응 확인 모듈의 Tacotron2는 deep voice보다 약 20ms, deep mind 보다 약 50ms 더 빨리 텍스트를 음성으로 변환했다. 따라서 긴급상황 알림 및 운전자 반응 확인 시스템은 효율적으로 신경망 모델을 학습시키고, 실시간으로 운전자의 반응을 확인할 수 있다.

Abstract Currently, the autonomous vehicle market is commercializing a level 3 autonomous vehicle, but it still requires the attention of the driver. After the level 3 autonomous driving, the most notable aspect of level 4 autonomous vehicles is vehicle stability. This is because, unlike Level 3, autonomous vehicles after level 4 must perform autonomous driving, including the driver's carelessness. Therefore, in this paper, we propose the Emergency-notification and Driver-response Confirmation System(EDCS) for an autonomousvehicle safety that notifies the driver of an emergency situation and recognizes the driver's reaction in a situation where the driver is careless. The EDCS uses the emergency situation delivery module to make the emergency situation to text and transmits it to the driver by voice, and the driver response confirmation module recognizes the driver's reaction to the emergency situation and gives the driver permission Decide whether to pass. As a result of the experiment, the HMM of the emergency delivery module learned speech at 25% faster than RNN and 42.86% faster than LSTM. The Tacotron2 of the driver's response confirmation module converted text to speech about 20ms faster than deep voice and 50ms faster than deep mind. Therefore, the emergency notification and driver response confirmation system can efficiently learn the neural network model and check the driver's response in real time.

Key Words : Autonomous vehicles, Hidden Markov Model, Tacotron2, Speech recognition, Speech to Text, Text to Speech

*Catholic Kwandong University, Department of Computer Engineering
**Catholic Kwandong University, Department of Software
Received March 02, 2021 Revised March 09, 2021 Accepted March 23, 2021

1. 서론

자율주행차량은 이제 상용화를 앞두고 있다. 해외의 자율주행차량은 이미 사용화 되어 판매되고 있으나, 안정성의 문제로 완전 자율주행이 불가능하며 운전자는 기존 수동 운전과 동일하게 운전 중 집중상태를 유지해야 한다. 실제로 2020년 9월 18일에는 캐나다의 한 남성이 140km/h로 주행하는 자율주행차량에서 잠들어 경찰에 연행되기도 했다[1]. 즉, 현재 자율주행 차량은 완벽히 안전하지 않지만, 운전자는 쉽게 부주의해질 수 있다는 뜻이다. 따라서 본 논문에서는 차량에서 긴급 상황이 발생하면 음성으로 운전자에게 알리고, 운전자의 반응에 따라 자율주행 지속 여부를 결정하는 자율차량 안전을 위한 긴급상황 알림 및 운전자 반응 확인 시스템을 설계한다. 본 논문에서 제시하는 긴급상황 알림 및 운전자 반응 확인 시스템은 주어진 긴급상황을 텍스트 문장으로 생성하고 TTS(Text to Speech)를 통해 음성으로 출력하는 긴급 상황 전달 모듈과 STT(Speech to Text)를 통해 운전자의 음성을 텍스트 문장으로 생성하여 운전자의 수동 주행 여부를 인식하는 운전자 반응 확인 모듈로 구성된다. 본 논문의 2절은 해당 연구가 직, 간접적으로 영향을 받은 STT와 TTS에 대한 연구들을 소개하고, 3절은 본 논문에서 제안하는 긴급상황 알림 및 운전자 반응 확인 시스템에 대해 상세히 기술한다. 4절은 긴급상황 알림 및 운전자 반응 확인 시스템의 유효성을 검증하기 위한 실험을 진행하고 5절에서는 해당 논문을 전체적으로 정리하고 향후 연구에 대해 설명한다.

2. 관련 연구

수원대학교의 정보통신공학과에서는 car navigation용 '한국어 무제한 어휘 음성합성기'를 저가의 DSP chip(ADSP-2185)과 저용량의 4M bits ROM을 사용하여 low-cost system으로 하드웨어를 구성하였다[2].

버스 어플리케이션은 마켓에 이미 존재하지만, 오히려 다양한 기능을 가진 어플리케이션은 사용법을 더 어렵게 만들거나 글씨가 작아 시각 장애인 및 저시력자들에게는 사용하는 데에 여전히 불편한 상황이 발생한다. 해당 연구는 어플리케이션이 한 화면으로만 구성되며 검은 바탕에 흰 글씨를 사용하여 필요한 정보만 띄움으로써 보기 쉽게 만들고, 사용법은 더욱 간단해진 버스 안내 어플리케이션을 제안했다. 또한 해당 어플리케이션은 정보를 표시할 때 TTS를 사용하여 표시한 정보를 음성으로 안내한다[3].

국민 대학교에서는 시각장애인 보행 보조를 위한 스마트 폰 케이스를 설계하고 구현했다. 해당 폰 케이스는 조도 센서와 스마트폰 카메라 플래시를 이용하여 어두운 장소에서 자신의 위치를 알려주는 자기 위치 알림 시스템과 초음파 센서를 이용하여 장애물을 감지하고 시각장애인들에게 음성으로 경고를 해주는 음성 경고 시스템을 제공했다[4].

서일 대학교에서는 한국어의 특성에 기반한 STT엔진 변환 성능 평가에 대한 가이드를 제안했다. 제안한 가이드를 사용하면 엔진 제작사는 한국어 특성에 기반한 STT 변환을 수행할 수 있으며, 수요처에서는 더 정확한 평가를 수행할 수 있다. 해당 연구에서는 비교적 짧은 문장에 대한 테스트와 제한된 인력을 통한 평가를 수행했다. 또한 사투리나 외래어에 대한 충분히 고려되지 않았다[5].

연세대학교에서는 차량환경에서 잔향과 근접장 효과에 의해 발생하는 목적 음성 신호의 왜곡을 감소시킬 수 있는 마이크로폰 어레이 빔형성 기법을 제안하였다. 온라인으로 추정하기 어려운 소스와 마이크간의 전달함수 대신 상대적으로 추정이 용이한 기준 마이크와 다른 마이크 간의 상대전달함수를 조향 벡터로 이용함으로써, 원격장 모델의 조향 벡터를 이용한 빔형성기에 비해 목적 음성 신호의 왜곡을 감소시킬 수 있는 준최적 빔형성 기법을 제안하였다[6].

3. 음성 전달 및 인식 시스템 설계

3.1 시스템 개요

본 논문에서 제안하는 음성 전달 및 인식 시스템은 두 가지 모듈로 구성된다. 첫 번째 모듈인 긴급상황 전달 모듈은 긴급상황을 텍스트화하여 운전자에게 음성으로 전달한다. 두 번째 모듈인 운전자 반응 확인 모듈

은 긴급상황에 대한 운전자의 반응을 인식하여 운전 권한을 운전자에게 넘길지 결정한다. 그림 1은 음성 전달 및 인식 시스템의 구성을 나타낸다.

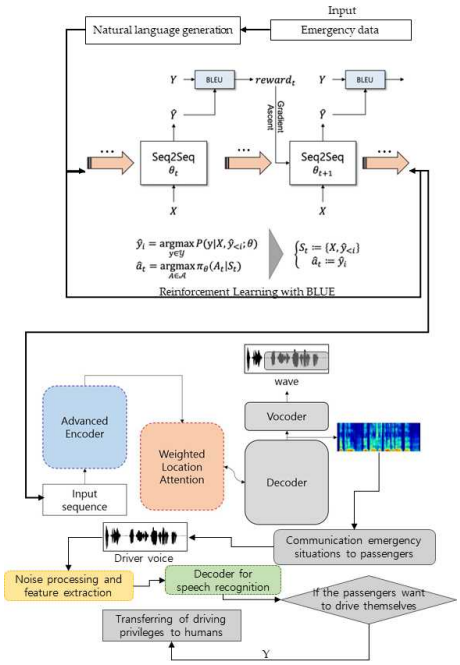


그림 1. 음성 전달 및 인식 시스템의 전체 구성도
Fig. 1. Overall configuration diagram of the Emergency-notification and Driver-response Confirmation System (EDCS)

3.2 긴급상황 전달 모듈

긴급상황 전달 모듈은 TTS를 이용하여 운전자에게 차량이 처한 긴급상황을 음성으로 전달한다. 긴급상황 전달 모듈은 Seq2Seq 모델을 통하여 긴급상황을 텍스트 문장으로 변환하고 Google의 타코트론2[7]를 이용하여 긴급상황 텍스트를 음성으로 변환한다. 우선 긴급상황 전달 모듈은 BLEU Score (Bilingual Evaluation Understudy Score)를 사용하여 Seq2Seq 모델을 학습시킨다.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{4} = w_n \log(p_n)\right) \qquad (1)$$

수식 1은 BLEU Score를 계산하는 방법을 나타낸다. 수식 1에서 BP는 트레이닝 데이터보다 생성된 텍스트 문장이 짧을경우 패널티를 부여하는 Brevity Penalty를 의미한다. 수식 2는 BP의 계산 방식을 보여준다.

$$BP = \begin{cases} 1 & \text{if}(c > r) \\ e^{(1-r/c)} & \text{if}(c \leqq r) \end{cases} \qquad (2)$$

수식 2에서 r은 트레이닝 데이터 문장의 길이를, c는 생성된 문장의 길이를 의미한다. 다음으로 수식 1의 n은 트레이닝 데이터와 생성된 텍스트 문장의 단어 단위를 결정한다. 즉, n이 1이라면, 트레이닝 데이터와 생성된 문장을 비교할 때 1개의 단어를 하나의 묶음으로 취급하여 비교하고, 4라면 4개의 단어를 하나의 묶음으로 취급하여 비교한다. 긴급상황 전달 모듈의 Seq2Seq 모델은 1~4의 n을 사용한다. w_n은 각 n에 대한 가중치를 의미한다. 본 논문에서 제안하는 모델은 가중치로 {0.16, 0.25, 0.33, 0.26}을 사용한다. 마지막으로 $\log(p_n)$은 실제로 트레이닝 데이터와 생성된 문장을 비교한 값이다. 수식 3은 p_n을 계산하는 과정을 나타낸다.

$$p_n = \frac{Count_{dip}(n.gram)}{Count(n.gram)} \qquad (3)$$

수식 3에서 $Count(n.gram)$ 생성된 문장의 n-gram의 수를 의미한다. n-gram이란 n개의 연속된 단어묶음을 의미한다. 다음으로 $Count_{dip}(n.gram)$은 트레이닝 데이터와 생성된 문장에 동시에 존재하는 n-gram의 수를 의미한다. 이때, 중복된 n-gram을 반복해서 인식하는 것을 막기 위해 $Count_{dip}(n.gram)$은 비교된 n-gram을 카운트하면서 계산된다. 학습을 완료한 Seq2Seq 모델은 차량에 적용되어 차량의 긴급상황을 텍스트로 출력한다. Seq2Seq 모델이 긴급상황을 텍스트 문장으로 생성하면, 긴급상황 전달 모듈의 타코트론 2 모델[7]은 Seq2Seq 모델이 생성한 텍스트 문장을 음성으로 변환하여 운전자에게 전달한다. 이때, 생성된 문장을 음성으로 변환하기 위하여 긴급상황 전

달 모듈은 LJ Speech Dataset[8]을 사용하여 타코트론 2 모델을 학습한다.

3.3 운전자 반응 확인 모듈

긴급상황이 TTS를 통해 운전자에게 전달되면, 운전자 반응 확인 모듈은 운전자의 음성을 인식하여 운전자의 의사를 파악한다. 운전자 반응 확인 모듈은 HMM기반 신경망을 사용하여 운전자의 음성을 인식하고, 자율주행을 그만둘지, 자율주행을 계속할지 결정한다. 그림 3은 운전자 반응 확인 모듈에서 사용되는 신경망 모델의 구성을 나타낸다.

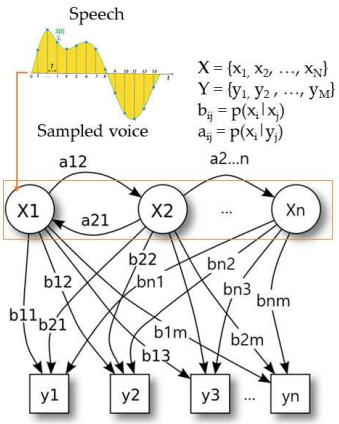


그림 3. 운전자 반응 확인 모듈의 음성인식 신경망
Fig. 3. The voice recognition neural network of the driver response confirmation module

그림 3의 신경망은 크게 4가지로 구성된다. 첫째, 그림 3의 X는 HMM의 은닉층을 나타낸다. 본 논문에서 X는 운전자의 음성을 샘플링한 값이 저장된 배열을 의미한다. 둘째, Y는 X로 인하여 발생할 수 있는 관측값들의 집합을 의미한다. 본 논문에서 Y는 X생성된 텍스트 단어들의 집합을 의미하며, Y 집합 전체는 문장을 의미한다. 셋째, a_{ij}는 X 집합에서 x_i에서 x_j로 이동할 확률을 나타내는 전이 확률 $p(x_j \mid x_i)$을 의미한다.

$$p(Y \mid \theta) = \sum_{i}^{1}\sum_{i}^{2}\cdots\sum_{i}^{m} p(q_{i\ 1})p(y_1 \mid q_{i\ 1}) \quad (4)$$
$$p(q_{i\ 2} \mid q_{i\ 1})p(y_2 \mid q_{i\ 2})\cdots p(q_{i\ m} \mid q_{i\ m-1})$$
$$p(y_m \mid q_{i\ m})$$

넷째, b_{ij}는 X집합의 x_i에서 y_j가 발생할 확률 $p(y_i \mid x_i)$을 의미한다. 운전자 반응 확인 모듈은 a_{ij}와 b_{ij}를 파라미터 θ로 신경망에 입력하고, 신경망은 입력된 θ를 통하여 $p(Y \mid \theta)$를 계산한다. 수식 4는 $p(Y \mid \theta)$의 계산을 나타낸다. 운전자 반응 확인 모듈은 Y집합의 텍스트를 이용하여 운전자가 운전에 대한 의사를 드러내는지 검사하고, 만약 운전자가 직접 운전하길 원한다면 자율주행을 중단한다.

4. 실험

본 논문은 긴급상황 알림 및 운전자 반응 확인 시스템의 유효성을 판단하기 위하여 가상환경의 긴급상황 데이터를 사용하여 긴급상황 전달 모듈과 운전자 반응 확인 모듈에 대한 실험을 진행했다. 긴급상황 전달 모듈의 경우 많은 긴급상황을 빠르게 학습하는지 검증하기 위하여 다른 음성인식 모델들을 동일 환경에서 학습시켜 학습 시간을 비교하였다. 운전자 반응 확인 모듈의 경우 실시간성을 입증하기 위하여 다른 TTS 모델들과 동일 환경에서 연산 시간을 비교하였다. 실험에 사용된 모델들은 LJ Speech Dataset을 사용하여 학습되고 테스트되었다.

표 1. HMM, RNN, LSTM의 학습 시간
Table 1. Learning time for HMM, RNN, LSTM

the number of sentences	HMM	RNN	LSTM
5000	12	14	23
5500	15	15	22
6000	14	19	24
7000	18	21	25
8000	16	23	25
9000	19	24	26
10000	20	28	29
11000	21	31	32
12000	26	32	38
13000	28	35	40

그림 4는 음성인식 데이터를 학습할 때 HMM, RNN과 LSTM의 학습 시간을 나타낸다. RNN과 LSTM은 기존 구조에 HMM의 입출력과 같은 노드를 설계하여 같은 테스트 데이터 세트를 입력했다[9, 10]. 테스트 데이터 세트의 문장을 500개부터 1,300개까지 증가시키면서 각 모델의 학습 시간이 측정되었다.

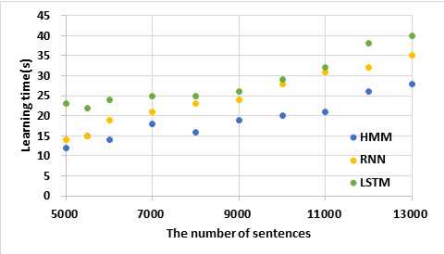


그림 4. HMM, RNN, LSTM의 학습 시간
Fig. 4. Learning time for HMM, RNN and LSTM

그림 5는 TTS 엔진인 Tacotron2, deep voice, deep mind가 텍스트를 음성으로 변환하는데 걸리는 연산 시간을 나타낸다. deep voice와 deep mind는 기존 구조에 Tacotron2의 입출력과 같은 노드를 설계하여 같은 테스트 데이터 세트를 입력했다[11, 12].문장은 1개부터 10개까지 증가하였다. 표 2는 Tacotron2, deep voice, deep mind의 연산 결과를 나타낸다.

표 2. Tacotron2, deep voice, deep mind의 연산 시간
Table 2. calculation time of the Tacotron2, deep voice and deep mind

the number of sentences	taco tron2	deep v oice	deep mi nd
1	52ms	60ms	92ms
2	52ms	61ms	92ms
3	53ms	63ms	93ms
4	53ms	65ms	94ms
5	53ms	65ms	94ms
6	54ms	70ms	95ms
7	56ms	72ms	96ms
8	56ms	74ms	98ms
9	57ms	75ms	97ms
10	58ms	76ms	101ms

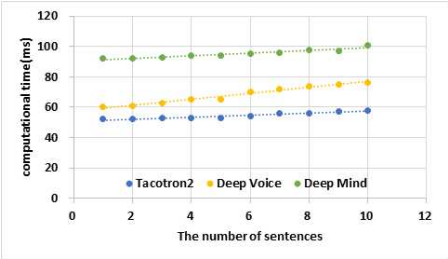


그림 5. 타코트론2, 딥 보이스, 딥 마인드의 연산 시간
Fig. 5. Calculation time of the Tacotron 2, Deep Voice and Deep Mind

5. 결론

본 논문에서는 긴급상황을 텍스트 문장으로 생성하여 TTS를 통해 운전자에게 음성으로 긴급상황을 전달하고 운전자의 음성을 입력받아 자율주행 지속 여부를 결정하는 자율차량 안전을 위한 긴급상황 알림 및 운전자 반응 확인 시스템을 제안했다. 긴급상황 알림 및 운전자 반응 확인 시스템은 주어진 긴급상황을 텍스트 문장으로 생성하고 TTS(Text to Speech)를 통해 음성으로 출력하는 긴급상황 전달 모듈과 STT(Speech to Text)를 통해 운전자의 음성을 텍스트 문장으로 생성하여 운전자의 수동 주행 여부를 인식하는 운전자 반응 확인 모듈로 구성되었다. 본 논문에서는 긴급상황 알림 및 운전자 반응 확인 시스템의 효율성을 검증하기 위해 서로 다른 음성인식 모델의 학습 시간을 비교하고, 서로 다른 TTS 모델의 연산 시간을 비교했다. 실험 결과, HMM은 RNN보다 25% LSTM보다 42.86% 빠른 속도로 음성을 학습했고 Tacotron2는 deep voice보다 약 20ms, deep mind 보다 약 50ms 더 빨리 텍스트를 음성으로 변환했다. 그러나 실제 긴급상황이 아닌 가상환경에서 임의로 설정된 긴급상황이었으며, 긴급상황 알림 및 운전자 반응 확인 시스템은 스스로 긴급상황을 탐지할 수 없다. 따라서 향후 차량 센서 데이터를 이용하여 긴급상황을 스스로 탐지하기 위한 연구가 진행되어야 하며, 해당 연구 이후 실제 차량에서 긴급상황을 탐지하고 음성으로 알림을 전달하는 실험이 이루어져야 할 것이다.

REFERENCES

[1] Tesla owner in Canada charged with 'sleeping' while driving over 90 mph, Andrew J. Hawkins, Available online: https://www.theverge.com/2020/9/18/21445168/tesla-driver-sleeping-police-charged-canada-autopilot

[2] Ji Hoon Na, Jung Mo Sung, Yoon Gi Yang, "Low-cost implementation of text to speech(TTS) system for car navigation", Journal of the Acoustical Society of Korea 2000 Summer Conference, The Acoustical Society of Korea, Vol.19, No.1, pp. 141-144, 2000

[3] Ji Young Yoon, Hyeon Ji Lee, Sung Soo Hwang, "Bus Arrival Guide Application for the Blind and Low Vision", PROCEEDINGS OF HCI KOREA 2018, The HCI Society of Korea, pp. 858-861, January, 2018

[4] Jin Woo Choi, Gu Min Jeong, "Development of Walking Assist Smartphone Case for Blind People", The Journal of Korea Institute of Information, Electronics, and Communication Technology, Korea Information Electronic Communication Technology, Vol.8, No.3, pp. 239-242, 2015

[5] So Yeon Min, Kwang Hyong Lee, Dong Seon Lee, Dong Yeop Ryu, "A Study on Quantitative Evaluation Method for STT Engine Accuracy based on Korean Characteristics", Journal of Korea Academia-Industrial cooperation Society, Korea Academy Industrial Cooperation Society, Vol.21, No.7, pp. 699-707, 2020

[6] Chul Hee Han, Hong Goo Kang, Young Soo Hwang, Dae Hee Youn, "A Microphone Array Beamformer for the Performance Enhancement of Speech Recognizer in Car", The journal of the acoustical society of Korea, The acoustical society of Korea, Vol.24, No.7, pp.423-430, 2015

[7] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis", Interspeech 2017, Stockholm, Sweden, August, 2017

[8] The LJ Speech Dataset. Available online: https://keithito.com/LJ-Speech-Dataset/ (2020-09-03).

[9] Won-jun Yoo, "Introduction to natural language processing using deep learning", 2020

[10] Understanding LSTM Networks, colah, Available online: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[11] Miles Anderson, Yadong Wang, Francois Leo, Stephane Coen, Miro Erkintalo, Stuart Murdoch, "Super cavity solitons and the coexistence of multiple nonlinear states in a tristable passive Kerr resonator", Phys. Rev., Vol. 7, 2017

[12] Jeff Donahue, Sander Dieleman, Mikołaj Binkowski, Erich Elsen, Karen Simonyan, "´END-TO-END ADVERSARIAL TEXT-TO-SPEECH", DeepMind, 2020

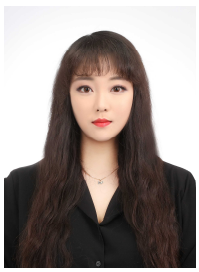
저자약력

손 수 락(Su-Rak Son) **[정회원]**

- 2018년 2월 : 가톨릭관동대학교 컴퓨터학과(공학사)
- 2019년 8월 : 가톨릭관동대학교 컴퓨터학과(공학석사)
- 2019년 8월 ~ 현재 : 가톨릭관동대학교 컴퓨터학과(공학박사 재학)

〈관심분야〉 빅데이터, 네트워크, 프로그래밍 언어

정 이 나(Yi-na Jeong) **[정회원]**



- 2011년 2월 : 가톨릭관동대학교 컴퓨터학과(공학사)
- 2013년 8월 : 가톨릭관동대학교 컴퓨터학과(공학석사)
- 2018년 8월 : 가톨릭관동대학교 컴퓨터학과(공학박사)
- 2017년 3월 ~ 현재 : 가톨릭관동대학교 컴퓨터학과 초빙교수

〈관심분야〉 빅데이터, AI, 프로그래밍 언어