J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

대화 영상 생성을 위한 한국어 감정음성 및 얼굴 표정 데이터베이스☆

Korean Emotional Speech and Facial Expression Database for Emotional Audio-Visual Speech Generation

백 지 영[1] 김 세 라[1] 이 석 필[1]*
Ji-Young Baek Sera Kim Seok-Pil Lee

요 약

본 연구에서는 음성 합성 모델을 감정에 따라 음성을 합성하는 모델로 확장하고 감정에 따른 얼굴 표정을 생성하기 위한 데이터베이스를 수집한다. 데이터베이스는 남성과 여성의 데이터가 구분되며 감정이 담긴 발화와 얼굴 표정으로 구성되어 있다. 성별이 다른 2명의 전문 연기자가 한국어로 문장을 발음한다. 각 문장은 anger, happiness, neutrality, sadness의 4가지 감정으로 구분된다. 각 연기자들은 한 가지의 감정 당 약 3300개의 문장을 연기한다. 이를 촬영하여 수집한 전체 26468개의 문장은 중복되지 않으며 해당하는 감정과 유사한 내용을 담고 있다. 양질의 데이터베이스를 구축하는 것이 향후 연구의 성능에 중요한 역할을 하므로 데이터베이스를 감정의 범주, 강도, 진정성의 3가지 항목에 대해 평가한다. 데이터의 종류에 따른 정확도를 알아보기 위해 구축된 데이터베이스를 음성-영상 데이터, 음성 데이터, 영상 데이터로 나누어 평가를 진행하고 비교한다.

☞ 주제어 : 음성합성, 감정음성, 데이터베이스, 멀티모달

ABSTRACT

In this paper, a database is collected for extending the speech synthesis model to a model that synthesizes speech according to emotions and generating facial expressions. The database is divided into male and female data, and consists of emotional speech and facial expressions. Two professional actors of different genders speak sentences in Korean. Sentences are divided into four emotions: happiness, sadness, anger, and neutrality. Each actor plays about 3300 sentences per emotion. A total of 26468 sentences collected by filming this are not overlap and contain expression similar to the corresponding emotion. Since building a high-quality database is important for the performance of future research, the database is assessed on emotional category, intensity, and genuineness. In order to find out the accuracy according to the modality of data, the database is divided into audio-video data, audio data, and video data.

☞ keyword : Speech Synthesis, Speech Emotion, Database, Multi Modal

1. 서 론

최근 인공지능 기법들의 발전과 함께 많은 연구들에서 딥러닝 기반의 음성 합성 모델들을 제안하였다.[1][2] 이를 통해 사람과 컴퓨터의 상호작용을 이용한 서비스가 상용화되고 있고 컴퓨터가 사람의 감정을 표현하려는 연구가 이어지고 있다.[3] 이러한 연구들의 성능을 높이기 위해서는 합성 모델, 데이터베이스 등 다양한 요소들이 중요하다.

음성만으로는 감정 표현에 몇 가지의 어려움이 따르는데, 음성에는 감정을 구분하는 명확한 특징이 없고, 화자마다 발화 스타일이 다를 수 있다는 어려움이 있다.[4] 기존의 연구들에서는 감정 표현을 위하여 음성 이외에 얼굴 표정도 사용한다. 사람들은 얼굴 표정을 통해 감정 상태를 즉각적으로 인지할 수 있으므로 얼굴 표정은 음성과 더불어 감정을 표현하는 데 중요한 역할을 한다.[5] 따라서 컴퓨터가 감정에 따라 발화하고 얼굴 표정을 표현한다면 사람과의 더 자연스러운 상호작용으로 발전할 수 있다.

본 논문은 한국어를 기반으로 하는 감정 음성 및 얼굴

1 Dept. of Electronic Engineering, Sangmyung University., Seoul, 03016, Korea.
* Corresponding author (esprit@smu.ac.kr)
[Received 19 January 2022, Reviewed 13 February 2022, Accepted 25 February 2022]
☆ 본 논문은 2021년도 한국인터넷정보학회 추계학술대회 우수논문 추천에 따라 확장 및 수정된 논문임.

표정 생성 연구에 활용하기 위한 한국어 음성과 얼굴 표정으로 구성된 데이터베이스를 구축한다. 연구의 정확도를 높이기 위해서는 양질의 데이터를 사용하는 것이 중요하다. 따라서 수집한 데이터를 평가하는 과정이 필요하다. Steven R. Livingstone가 진행한 연구에서는 North American English를 기반으로 angry, happy, neutral, sad, fearful, calm, surprise, disgusted의 8가지 감정으로 촬영하여 4320개의 음성 및 영상 데이터를 수집하였다. 데이터 평가 결과 음성-영상 데이터는 80%, 영상 데이터는 75%, 음성 데이터 60%의 정확도를 얻었다.[6]

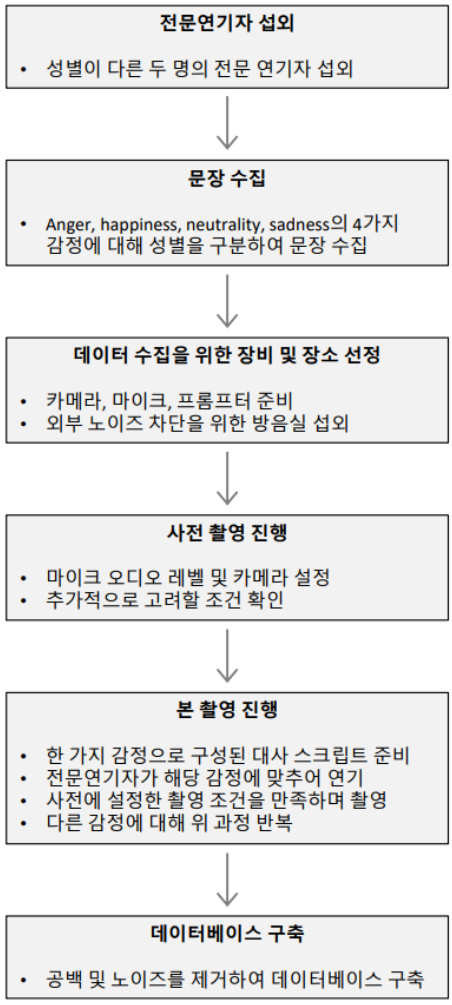
본 연구에서도 데이터의 종류에 따른 결과를 비교하기 위해 구축된 데이터베이스를 음성-영상 데이터, 음성 데이터, 영상 데이터로 나누어 평가를 진행한다.

2. 한국어 감정 음성 및 얼굴 표정 데이터베이스

본 연구에서 구축하는 한국어 기반의 감정 음성 및 얼굴 표정 데이터베이스는 남성 화자와 여성 화자의 데이터가 구분되며 감정은 anger, happiness, neutrality, sadness로 구분한다. 4가지 감정에 대하여 해당 감정과 유사한 내용을 담고 있는 문장을 수집한 뒤, 성별이 다른 2명의 전문 연기자가 주어진 문장을 감정에 따라 연기하는 것을 촬영한다. 문장을 수집하는 과정과 연기자의 촬영 과정을 그림 1에 나타내었다.

2.1 데이터베이스 구축을 위한 문장 수집

데이터베이스를 구축하기 위하여 전문 연기자들이 4가지의 감정별로 연기할 문장이 필요하다. 문장은 발화 시 5초 이내의 시간이 걸리는 것으로 드라마, 영화, 만화의 대사를 수집하였다. 대사 앞에 오는 '아', '오', '음' 과 같은 감탄사는 제거하였다. 화자의 성별에 따라 단어의 쓰임이나 어투의 차이가 존재하므로 남성 발화 문장과 여성 발화 문장을 구분하였다. 수집한 대사 형식의 문장을 감정별로 분류하기 위하여 고려한 사항은 다음과 같다. 해당 감정을 느낄 때 보편적으로 사용되는 단어가 포함됐는지 또는, 해당 감정을 느낄 수 있는 환경과 상황 내에서의 대사인지 판단하였다. 표 1 에 감정별 키워드와 상황의 예시를 나타내었다.



(그림 1) 데이터 수집 과정
(Figure 1) Data collection process

(표 1) 감정별 키워드 및 상황
(Table 1) Keywords and situations by emotion

	키워드	상황
anger	거짓말, 의심, 싫어	이혼, 경찰서, 싸움
happiness	기뻐, 좋아, 사랑	생일, 여행, 취업
neutrality	이 외	내레이션, 설명
sadness	지쳐, 힘들어, 포기	죽음, 이별, 병원

만약 두 가지 이상의 감정으로 사용될 수 있는 문장이라면 대화의 상황을 파악하여 감정을 정하였다. 양질의 데이터를 얻기 위하여 각 감정이 담긴 문장들이 중복되

지 않도록 수집하였다. 이러한 과정을 통해 anger, happiness, neutrality, sadness의 4가지 감정에 대해 성별을 구분하여 약 3300개씩, 총 26468개의 문장을 수집하였다. 표 2에 수집한 대사의 예시를 나타내었다.

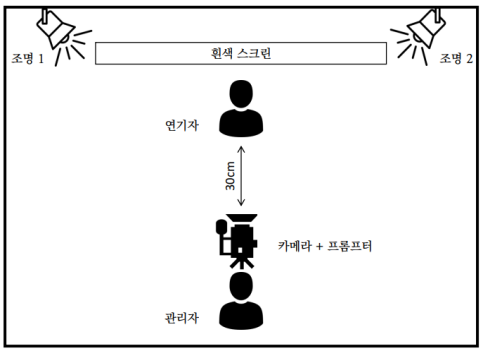
(표 2) 성별, 감정별로 수집된 대사
(Table 2) Sentences collected by gender and emotion

성별	감정	문장
남	anger	못난 형은 잘난 동생 일에 간섭 좀 하면 어디가 덧나냐?
	happiness	그래서 한 여자는 제가 데리고 살잖아요
	neutrality	아버지는 혼자서 나와 형을 키우셨다.
	sadness	내가 욕심이 나서.. 혼자 사는 게 싫어서..
여	anger	오빠 그럴 때마다 사람 무시하는 거 같아서 진짜 기분 나빠, 알아?
	happiness	우리 아들 손이 벌써 이만큼 컸네~
	neutrality	엄마 상 차리는 것 좀 도와라.
	sadness	우리 준희.. 엄마가 얼마나 보고싶었는데..

2.2 한국어 감정 음성 및 얼굴 표정 데이터베이스 구축

수집한 문장을 1회 당 감정별로 200개씩 분배하여 각 연기자는 16회의 촬영을 진행하였다.

촬영 시점에 관계없이 전체 데이터가 균일해야하므로 몇 가지의 촬영 조건을 설정하였다. 먼저, 외부의 노이즈가 차단되는 동일한 방음실에서 촬영하였다. 방음실의 조명은 일정한 밝기를 유지하였다. 카메라에 방송용 프롬프터를 장착하여 대본을 보면서도 연기자의 눈이 항상 카메라를 응시하도록 연출하였다. 카메라 및 마이크와 연기자의 거리는 약 30cm로 유지하며 흰 색의 벽을 배경으로 촬영하였다. 감정별로 균일한 오디오를 수집하기 위해 감정에 따라 마이크의 오디오 레벨을 조정하였다. happiness와 neutrality는 마이크 오디오 레벨 5.5로 설정하고 anger는 레벨2, sadness는 레벨 8로 설정하였다. 이후 연기자는 준비된 검은색 반팔 상의를 착용하고 동일하게 단장하였다. 그림 2는 촬영 환경을 도식화하였고 그림 3은 전문연기자가 위의 조건을 만족하면서 촬영하는 모습을 나타내었다.

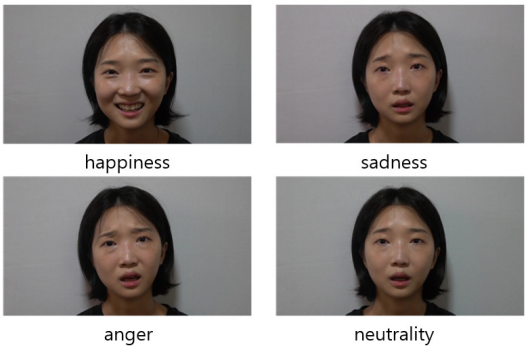


(그림 2) 촬영 환경
(Figure 2) Filming environment

(그림 3) 촬영하고 있는 두 명의 연기자
(Figure 3) Two actors filming

이러한 과정을 통해 그림 4와 그림 5와 같이 남성 연기자와 여성연기자의 감정에 따른 얼굴 표정 데이터를 수집했다.

(그림 4) 남성 연기자의 얼굴 표정 데이터 예시
(Figure 4) Example of facial expression data for a male actor



(그림 5) 여성 연기자의 얼굴 표정 데이터 예시
(Figure 5) Example of facial expression data for a female actor

두 과정을 거쳐 anger, happiness, neutrality, sadness의 4가지 감정에 대한 데이터베이스를 구축하였다. 표 3에 각 감정에 대해 성별을 구분하여 수집한 데이터의 수를 나타내었다.

(표 3) 분류별 데이터의 수

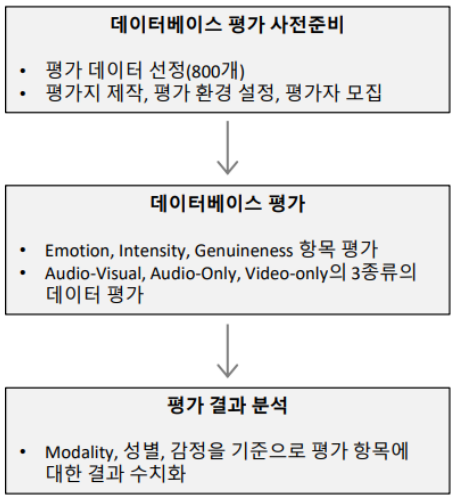
(Table 3) Number of data by gender and emotion

	남	여
anger	3173	3426
happiness	3154	3432
neutrality	3181	3436
sadness	3269	3397
합계	12777	13691

위와 같이 분류별로 약 3300개씩 데이터를 수집하여 전체 26468개의 데이터베이스를 구축하였다.

3. 데이터베이스 평가

감정 음성 및 얼굴표정 생성 연구의 정확도를 높이기 위해서는 양질의 데이터를 사용하는 것이 중요하다. 따라서 본 연구에서 구축한 데이터베이스가 양질의 데이터로 구성되었는지 평가하는 과정이 필요하다. 감정이 뚜렷하게 담겨있고 발화가 명확히 이루어진 것을 양질의 데이터로 판단한다. 각 감정 당 약 100개의 데이터를 선별하여 평가를 진행한다. 평가 과정을 그림 6에 나타내었다.



(그림 6) 데이터베이스 평가 과정
(Figure 6) Database assessment process

3.1 평가 방법

구축된 데이터베이스를 음성-영상 데이터, 음성 데이터, 영상 데이터로 구분하여 평가 결과를 비교해보기 위해, 선별된 800개 데이터를 audio-visual(AV), video-only(VO), audio-only(AO)의 3가지 modality에 대해 모두 평가하였다. 평가 항목은 감정의 범주, 강도, 진정성의 3가지로 구성하였다. 감정의 범주는 평가할 데이터가 어떤 감정에 해당하는가를 의미하며 평가자는 anger, happiness, neutrality, sadness 중에 선택하였다. 강도는 감정의 세기가 얼마나 강한가를 의미하며 평가자는 '매우 약하다, 약하다, 보통이다, 강하다, 매우 강하다'의 5가지 척도를 가지고 판단하였다. 진정성은 연기자가 정신적, 신체적으로 감정을 느끼며 표현하고 있다고 생각하는지 여부를 의미하며 평가자는 '매우 아니다, 거의 아니다, 보통이다, 약간 그렇다, 매우 그렇다'의 5가지 척도를 가지고 판단하였다. 평가에는 3명이 참여하였으며 균일한 평가 환경에서 평가를 진행하였다. 평가자들이 외부의 방해를 받지 않는 공간에서 개별적으로 평가가 이루어졌고 음성 및 영상 재생에는 동일한 헤드폰과 모니터가 사용되었다. 음량 크기 및 평가자와 모니터 사이 거리는 일정하게 유지되었다.

3.2 평가 결과

평가를 마친 후 평가자들의 모든 응답을 수치화하여 분석하였다. 평가 항목 중 감정의 범주에서는 데이터의 실제 감정 범주와 평가자의 응답이 일치하면 1, 그렇지 않으면 0으로 나타내었다. 감정의 강도와 진정성 항목에서는 각각 '매우 약하다', '매우 아니다'를 1, '매우 강하다', '매우 그렇다'를 5로 나타내는 것을 기준으로 하여 5가지 척도 응답을 1부터 5까지 수치화하였다. 그 후 데이터의 modality를 구분하여 각 평가 항목의 평균과 표준편차를 구하는 방식으로 분석을 진행하였다. 이 때, 감정의 범주 항목에서 평가자 응답이 일치하는 비율을 proportion correct라고 표기한다.[7] 표 4를 보면 음성-영상 데이터는 약 95%, 영상 데이터는 약 93%, 음성 데이터는 약 88%의

(표 4) 데이터 종류별 평가 결과
(Table 4) Assessment results by data modality

Modality	Gender	N	Mean (SD) Proportion correct	Mean (SD) Intensity	Mean (SD) Genuineness
AV	Male	400	0.96 (0.20)	3.52 (1.25)	4.40 (0.74)
	Female	400	0.93 (0.26)	3.69 (1.21)	4.48 (0.70)
VO	Male	400	0.94 (0.24)	3.49 (1.24)	4.50 (0.63)
	Female	400	0.92 (0.27)	3.95 (1.13)	4.50 (0.72)
AO	Male	400	0.87 (0.34)	3.55 (1.21)	4.36 (0.77)
	Female	400	0.88 (0.32)	3.58 (1.18)	4.44 (0.72)

(표 5) 음성-영상 데이터 평가 결과
(Table 5) Audio-video data assessment result

Modality	Gender	Emotion	N	Mean (SD) Proportion correct	Mean (SD) Intensity	Mean (SD) Genuineness
AV	Male	Anger	107	1.00 (0)	3.50 (1.26)	4.21 (0.75)
		Happiness	106	0.95 (0.21)	3.31 (1.05)	4.36 (0.81)
		Neutrality	96	0.93 (0.26)	4.26 (1.16)	4.69 (0.53)
		Sadness	91	0.96 (0.20)	3.03 (1.21)	4.35 (0.75)
	Female	Anger	89	0.96 (0.21)	3.65 (1.20)	4.67 (0.51)
		Happiness	124	0.90 (0.30)	3.19 (1.26)	4.40 (0.73)
		Neutrality	94	0.91 (0.28)	4.31 (1.07)	4.56 (0.68)
		Sadness	93	1.00 (0)	3.90 (0.96)	4.32 (0.81)

(표 6) 음성 데이터 평가 결과
(Table 6) Audio data assessment result

Modality	Gender	Emotion	N	Mean (SD) Proportion correct	Mean (SD) Intensity	Mean (SD) Genuineness
AO	Male	Anger	107	0.91 (0.29)	3.36 (1.25)	4.21 (0.91)
		Happiness	106	0.80 (0.40)	3.13 (1.12)	4.15 (0.88)
		Neutrality	96	0.89 (0.32)	4.02 (1.31)	4.53 (0.82)
		Sadness	91	0.89 (0.31)	3.66 (1.08)	4.46 (0.75)
	Female	Anger	89	0.99 (0.11)	3.98 (1.20)	4.54 (0.67)
		Happiness	124	0.90 (0.31)	3.53 (1.11)	4.44 (0.76)
		Neutrality	94	0.94 (0.24)	4.10 (1.05)	4.46 (0.75)
		Sadness	93	1.00 (0)	4.10 (0.93)	4.25 (0.77)

(표 7) 영상 데이터 평가 결과
(Table 7) Video data assessment result

Modality	Gender	Emotion	N	Mean (SD) Proportion correct	Mean (SD) Intensity	Mean (SD) Genuineness
VO	Male	Anger	107	0.96 (0.19)	3.31 (1.28)	4.47 (0.59)
		Happiness	106	0.98 (0.14)	3.12 (1.22)	4.53 (0.59)
		Neutrality	96	0.88 (0.33)	4.05 (1.15)	4.58 (0.66)
		Sadness	91	0.92 (0.27)	3.52 (1.15)	4.36 (0.83)
	Female	Anger	89	0.98 (0.15)	3.72 (1.18)	4.73 (0.51)
		Happiness	124	0.92 (0.27)	3.44 (1.19)	4.46 (0.64)
		Neutrality	94	0.73 (0.44)	3.78 (1.30)	4.40 (0.82)
		Sadness	93	0.90 (0.30)	3.42 (1.01)	4.18 (0.80)

정확도를 갖는다. 선행연구인 북미의 RAVDESS 데이터 평가 결과가 순서대로 80%, 75%, 60%를 도출한 것과 비교한다면 높은 정확도를 갖는 것임을 확인할 수 있다. 데이터의 modality 별 평가 결과를 표 5, 표 6와 표 7에 나타내었다.

4. 결 론

최근의 음성 합성 기술은 사람처럼 자연스럽게 말하는 것뿐만 아니라 감정까지 표현하는 방향으로 발전하고 있다. 이러한 기술의 연구를 위해서는 감정 데이터베이스의 구축이 중요하다. 음성과 얼굴 표정으로 사람의 감정을 나타낼 수 있으므로 본 연구에서는 anger, happiness, neutrality, sadness의 4가지 감정이 담긴 한국어 음성과 얼굴 표정 데이터베이스를 구축하였다. 데이터는 남성 화자와 여성 화자가 구분되어 있으며 한 성별에 대해 감정당 약 3300개씩, 전체 26468개의 데이터를 수집하였다. 데이터의 종류에 따른 결과를 비교하기 위해 구축된 데이터베이스를 음성-영상 데이터, 음성 데이터, 영상 데이터로 modality를 나누어 평가를 진행하였다. 평가 항목으로는 감정의 범주, 강도, 진정성이 있다. 평가 결과 음성-영상 데이터는 약 95%, 영상 데이터는 약 93%, 음성 데이터는 약 88%의 정확도를 갖는다.

이렇게 구축된 데이터베이스는 향후 한국어를 기반으로 감정에 따른 음성 합성 모델과 얼굴 표정 생성을 위한 연구에 활용될 계획이다.

감사의 글

"이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2019R1F1A1050052)."

참고문헌(Reference)

[1] A. H. Ali, M. Magdy, M. Alfawzy, M. Ghaly and H. Abbas, "Arabic Speech Synthesis using Deep Neural Networks," 2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA), pp. 1-6, 2021. https://doi.org/10.1109/ICCSPA49915.2021.9385731

[2] L. Liu et al., "Controllable Emphatic Speech Synthesis based on Forward Attention for Expressive Speech Synthesis," 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021. https://doi.org/10.1109/SLT48900.2021.9383537

[3] H. Choi, S. Park, J. Park, and M. Hahn, "Emotional Speech Synthesis for Multi-Speaker Emotional Dataset Using WaveNet Vocoder", 2019 IEEE International Conference on Consumer Electronics (ICCE), pp. 1-2, 2019. https://doi.org/10.1109/ICCE.2019.8661919

[4] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, Volume 44, Issue 3,

pp.572-587, 2011.
https://doi.org/10.1016/j.patcog.2010.09.020

[5] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, Remigiusz J. Rak, “Emotion recognition using facial expressions”, Procedia Computer Science, Volume 108, pp.1175-1184, 2017.
https://doi.org/10.1016/j.procs.2017.05.025

[6] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English,” PLoS ONE, vol. 13, no. 5, pp. e0196391, 2018.
https://doi.org/10.1371/journal.pone.0196391

[7] Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, et al. “The NimStim set of facial expressions: judgments from untrained research participants.”, Psychiatry Research, Volume 168, Issue 3, pp.242-249, 2009.
https://doi.org/10.1016/j.psychres.2008.05.006

◐ 저 자 소 개 ◑

백 지 영(Ji-young Baek)
2022년 상명대학교 융합전자공학과(학사)
2022년~현재 상명대학교 일반대학원 컴퓨터과학과(석사과정)
관심분야 : 신호처리, 음성 합성, 인공 지능
E-mail : b00217@naver.com

김 세 라(Sera Kim)
2022년 상명대학교 융합전자공학과(학사)
2022년~현재 상명대학교 일반대학원 컴퓨터과학과(석사과정)
E-mail : 98sera@gmail.com

이 석 필(Seok-Pil Lee)
1990년 연세대학교 전기공학과(공학사)
1992년 연세대학교 대학원 전기공학과(공학석사)
1997년 연세대학교 대학원 전기공학과(공학박사)
1997년~2002년 대우전자 영상연구소 선임연구원
2002년~2012년 KETI 디지털미디어연구센터 센터장
2010년~2011년 미국 Georgia Tech Faculty Member
2012년~현재 상명대학교 융합전자공학과 교수
관심분야 : 신호처리, 멀티미디어시스템, 방송통신시스템, 인공지능
E-mail : esprit@smu.ac.kr