ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 24-02-07
46
j.inst.Korean.electr.electron.eng.Vol.28,No.2,168~173,June 2024
https://doi.org/10.7471/ikeee.2024.28.2.168

잔차 신경망을 활용한 펫 로봇용 화자인식 경량화

Lightweight Speaker Recognition for Pet Robots using Residuals Neural Network

강 성 현*, 이 태 희**, 최 명 렬*★

Seong-Hyun Kang*, Tae-Hee Lee**, Myung-Ryul Choi*★

Abstract

Speaker recognition refers to a technology that analyzes voice frequencies that are different for each individual and compares them with pre-stored voices to determine the identity of the person. Deep learning-based speaker recognition is being applied to many fields, and pet robots are one of them. However, the hardware performance of pet robots is very limited in terms of the large memory space and calculations of deep learning technology. This is an important problem that pet robots must solve in real-time interaction with users. Lightening deep learning models has become an important way to solve the above problems, and a lot of research is being done recently. In this paper, we describe the results of research on lightweight speaker recognition for pet robots by constructing a voice data set for pet robots, which is a specific command type, and comparing the results of models using residuals. In the conclusion, we present the results of the proposed method and Future research plans are described.

요 약

화자인식은 개개인마다 다른 음성 주파수를 분석하여 미리 저장된 음성과 비교해 본인 여부를 판단하는 하나의 기술을 의미한다. 딥러닝 기반의 화자인식은 여러 분야에 적용되고 있으며, 펫 로봇도 그 중 하나이다. 하지만 펫 로봇의 하드웨어 성능은 딥러닝 기술의 많은 메모리 공간과 연산에 있어 매우 제한적인 상황이다. 이는 펫 로봇이 사용자와 실시간 상호작용에 있어 해결해야 할 중요한 문제점이다. 딥러닝 모델의 경량화는 위와 같은 문제를 해결하기 위한 하나의 중요한 방법으로 자리하였으며, 최근 많은 연구가 진행되고 있다. 이 논문에서는 특정한 명령어 형태인 펫 로봇용 음성 데이터 세트를 구축하고 잔차(Residual)를 활용한 모델들의 결과를 비교해 펫 로봇용 화자인식의 경량화 연구의 결과를 서술하며, 결론에서는 제안한 방법에 대한 결과와 향후 연구방안에 대해 서술한다.

Key words : Speaker recognition, CNN, Resnet, Lightweight, Classification

* Graduate Student, Dept. of Applied Artificial Intelligence, Hanyang University.
Professor, Dept. of Electrical Engineering in Hanyang University.
★ Corresponding author
E-mail : kangsh0820@naver.com, Tel : +82) 31-400-4036
※ Acknowledgment
Manuscript received May. 20, 2024; revised Jun. 13, 2024; accepted Jun. 19, 2024.

Ⅰ. 서론

빅데이터, 컴퓨팅 파워의 향상, 알고리즘의 발전 등으로 딥러닝은 최근 몇 년간 비약적으로 발전하여 다양한 분야에서 혁신적인 성과를 보였다.

그 중 딥러닝 기반의 화자인식은 보안 및 맞춤형 서비스 제공에 의미있는 성과를 보여주며, 최근 펫 로봇 시장에서 맞춤형 서비스 제공을 위한 기술로 사용되고 있다. 하지만 딥러닝 기술을 활용함에 있어 펫 로봇의 하드웨어 성능은 매우 제약적이며, 이를 고려한 딥러닝 모델 경량화 연구가 활발하게 진행되고 있다.

딥러닝 기반의 화자인식은 학습 음성 데이터의 발화하는 형태나 문맥을 고려해 문장독립[1] 또는 문장종속[1] 방식으로 나뉜다. 문장독립은 형태나 문맥에 제한이 없는 방식을 의미하며, 문장종속은 형태나 문맥에 제한이 있는 방식을 의미한다. 펫 로봇용 화자인식의 명령어는 대부분 형태나 문맥에 제한이 있는 문장종속 방식을 따른다.

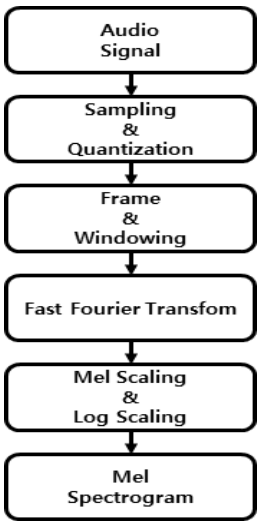


Fig. 1. Speech feature Extraction framework.
그림 1. 음성 특징 추출 프레임워크

딥러닝 모델의 사용에 있어 데이터의 전처리는 학습 효과를 높인다. 화자인식 데이터인 음성 데이터는 연속적인 흐름을 가지는 아날로그 신호로 그림 1의 과정을 통해 디지털 신호로 변환되어, 학습에 사용된다. 그림 1 특징 추출 프레임워크는 오랜 시간의 발전과 변화를 통해 음성의 주요 특징을 추출하는 것을 검증 받았으며, 추출된 데이터인 Mel Spectrogram 및 MFCC(Mel-Frequency Cepstral Cofficient)[2]는 오디오 신호 처리 분야 분야에서 널리 쓰인다. Mel Spectrogram 및 MFCC는 시간-주파수 대역으로 구성된 2차원 특징이며, 최근 음성 데이터의 가용성이 좋아지면서 풍부한 음성 특징을 더 담고 있는 Mel Spectrogram을 특징으로 모델의 학습에 사용하는 추세이다.

잔차 신경망 모델(Residual Neural Network, ResNet)은 2015년 ILSVRC(ImageNet Large Scale Visual Recognition Competition)에 처음 소개되었으며, 깊은 신경망에서 발생될 수 있는 가중치 소실 문제를 [3]에서 제안하는 잔차 블록(Residual Block) 구조를 통해 해결하였다. 잔차 블록 구조는 그림 2와 같은 구조이며, 여기서 x는 정해진 값으로 F(x)를 0으로 만드는 것을 통해 학습을 진행한다. 이러한 방향성 있는 전제조건을 가지는 학습은 좋은 학습 방향을 보일 것이라 가정했으며, [3]의 결과를 통해 그를 입증했다. 또한 단순히 x가 출력값에 합산되는 연산을 제외하면 추가적인 연산량 증가가 없다.

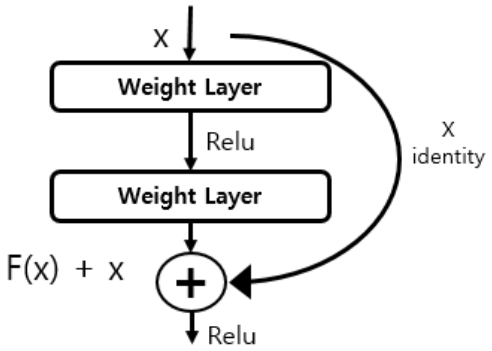


Fig. 2. Residual block.
그림 2. 잔차(Residual) 블록

딥러닝 모델은 높은 정확도를 위해 복잡한 구조가 필요하지만, 경량화가 되기 위해 간단한 구조가 필요하다. 경량화된 모델 설계에 있어 정확도와 최적화 성능 간의 트레이드 오프(Trade-off)를 이해하고 이를 최소화 하는 전략이 중요하다.

따라서 본 논문에서는 보편적으로 실제 반려견 명령어(ex. 이리와, 저리가...)가 정해진 형태나 문맥을 갖추고 있는 것을 참고해 문장종속 형태의 펫 로봇용 화자인식 음성 데이터 세트를 구축하고 특징 추출 프레임워크를 통한 전처리 과정을 진행한다. 또한 잔차 신경망을 활용한 모델들의 학습 결과인 추론 시간, MAC ,파라미터, 학습 시간, 정확도를 비교해 펫 로봇용 화자인식 모델의 경량화 연구를 진행한다.

II. 본론

2.1. 펫 로봇용 음성 데이터 세트

본 논문의 음성 데이터는 1초~2초 사이의 발화시간을 가지는 '이리와', '저리가', '기다려', '앉아', '엎드려' 5가지의 문장종속 방식의 실제 음성을 녹음을 진행한다. 각기 다른 화자 20명이 각 발화당 50번의 발화를 통해 총 5000개의 데이터 수로 구성된다. 음성 데이터 세트는 모델의 학습을 위해 각 명령어 당 8:2의 비율로 학습 데이터와 테스트 세트로 구분하며 이는 표 1와 같다.

Table 1. Voice data set for pet robots.
표 1. 펫 로봇용 음성 데이터 세트

Person		1	2	···	20	Total
		A	B	···	T	
Come	Train	40	40	···	40	800
	Test	10	10	···	10	200
Go	Train	40	40	···	40	800
	Test	10	10	···	10	200
Wait	Train	40	40	···	40	800
	Test	10	10	···	10	200
Sit	Train	40	40	···	40	800
	Test	10	10	···	10	200
Down	Train	40	40	···	40	800
	Test	10	10	···	10	200
Total		250	250	···	250	5,000

2.2. 음성 특징 추출 프레임워크

음성 신호는 연속적인 아날로그 신호로 딥러닝의 입력 값으로 사용되기 위해 샘플링(Sampling)과 양자화(Quantization) 두 Step에 걸쳐 이산적인 디지털 신호로 변환한다. 디지털 신호로 변환된 음성신호는 그림 1에 따라 음성 특징을 함축한 Mel Spectrogram으로 표현된다. 샘플링과 양자화를 거친 신호는 일정한 길이의 프레임으로 분할하는 과정을 거친다. 이를 'Framing'이라고 하며, 일반적으로 20~40ms 단위 정도로 분할되고, 프레임끼리 서로 떨어지지 않게 시간적으로 겹쳐 프레임을 구성한다. 그림 3은 Framing의 간단한 예로 프레임을 25ms간격으로 분할하며, 프레임 간 10ms를 겹쳐 구성하는 것을 보여준다.

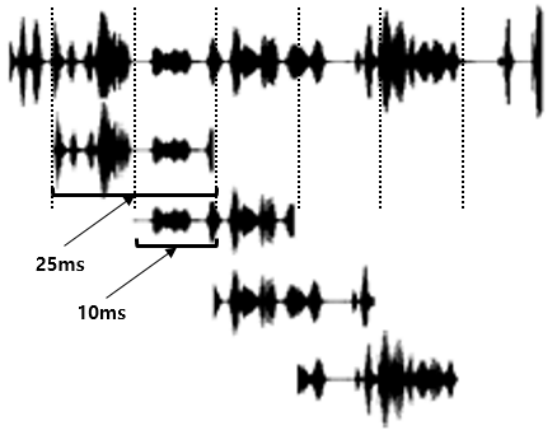


Fig. 3. Framing of voice signals.
그림 3. 음성 신호의 Framing

하지만 프레임은 잘려진 음성 신호로 양 끝에 불연속성이 발생되며, 이러한 불연속성은 노이즈(Noise)처럼 동작하게 되는 주파수 누설(Spectral Leakage) 문제가 발생한다. 주파수 누설 문제를 해결하기 위해 그림 4[4]과 같은 윈도우 함수를 각 프레임마다 적용해 양 끝의 불연속성을 줄여주는 'Windowing' 과정을 거친다.

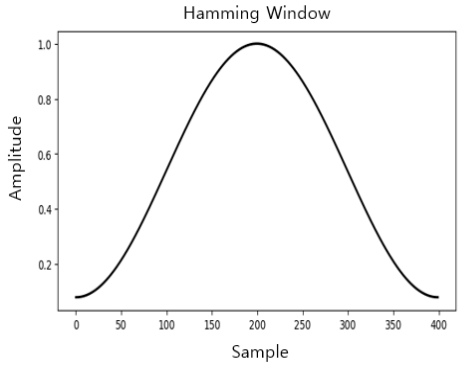
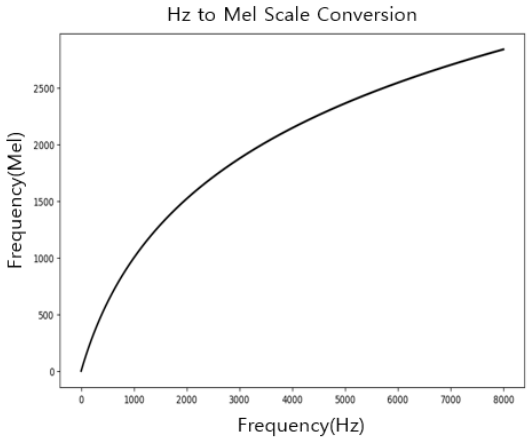
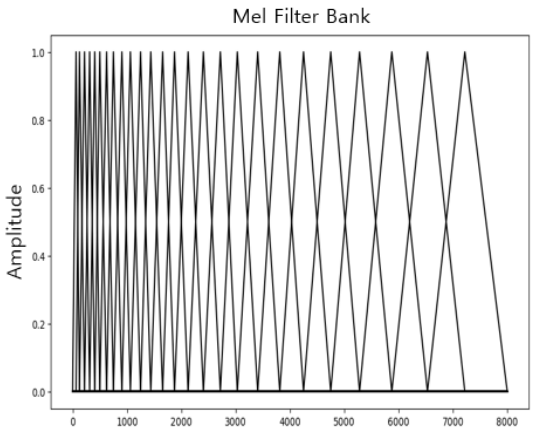


Fig. 4. Hamming Window function.
그림 4. Hamming Window 함수



(a) Hz to Mel Scale



(b) Mel Filter Bank

Fig. 5. Mel Scale.
그림 5. 멜 스케일

프레임 분할 과정을 마치면, 각 프레임마다 고속 푸리에 변환(Fast Fourier Transform, FFT)[5]를 적용한다. 고속 푸리에 변환은 이산 푸리에 변환(Discrete Fourier Transform, DFT)와 그 역변환을 빠르게 수행하는 효율적인 알고리즘으로 신호에 대한 주파수 정보를 제공한다.

사람의 청력은 주파수가 낮은 대역의 변화는 민감하게 반응하지만, 주파수가 높은 대역의 변화는 둔감하게 반응하는 특성을 가지고 있다. 그림 (a)는 이와 같은 특성을 반영해 기존 주파수와 변환된 주파수의 관계[6]를 보여주며, (a)를 기반으로 (b)의 Filter를 생성한다. 앞서 고속 푸리에 변환으로 변환된 데이터들은 그림 5 (b) Filter를 통해 스케일링을 진행한 후, Log연산을 취한다.

이후 모든 연산을 거친 데이터들을 일련의 순서로 나열을 통해, 그림 6과 같은 시간 대역과 주파수 대역으로 구성된 음성 신호의 특징인 Mel Spectrogram을 얻는다.

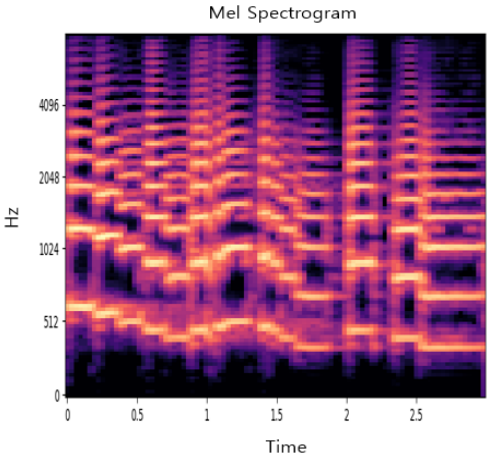


Fig. 6. Mel Spectrogram.
그림 6. 멜 스펙트로그램

2.3. ResCNN 모델

본 논문에서는 [3]에서 제시하는 가장 작은 모델인 ResNet-18모델과 더불어 Residual Block을 활용해 ResCNN-6, ResCNN-8. ResCNN-10을 설계 및 구현한다. ResCNN-6, ResCNN-8. ResCNN-10 모델 모두 처음 계층은 64 채널, 7×7 커널, 1 스트라이드, 3 패딩을 가지는 CNN 계층이며, 끝 계층은 AdaptiveAvgPool 계층으로 시간과 주파수 대역을 1×1 크기로 고정시킨다. 표 2와 같이 중간 계층은 CNN 기반 Residual Block 총 4개로 구성된다. 각 채널의 수는 64, 128, 256, 512로 설정되며, 채널 128, 256, 512 블록은 2 스트라이드로 설정한다. ResCNN-6은 2개의 Residual block(1 · 2), ResCNN-8은 3개의 Resblock(1 · 2 · 3), ResCNN-10은 4개의 Resblock(1 · 2 · 3 · 4)로 구성된다.

Table 2. Four types of residual blocks.
표 2. 4가지의 Residual Block

Resblock1	Resblock2	Resblock3	Resblock4
3×3, 64 3×3, 64	3×3, 128 3×3, 128	3×3, 256 3×3, 256	3×3, 512 3×3, 512

2.4. ResCNN 모델 학습 및 결과

본 논문에서는 표 1의 음성 데이터 세트를 사용하며 2.2절 음성 전처리 과정을 통해 음성 신호를 Mel Spectrogram으로 추출한다. Mel Spectrogram은 2초의 시간, 64개의 Filter 수로 설정해 200×64의 크기를 가지며, 2.3절에서 설계한 ResCNN-6, ResCNN-8. ResCNN-10, ResNet-18 모델의 입력값으로 사용되어 학습을 진행한다. 학습은 0.001의 학습률, 128개의 배치 사이즈로 100 Epoch 동안 진행하였으며, L2규제 1e-5인 Adam 옵티마이저[7]를 사용한다.

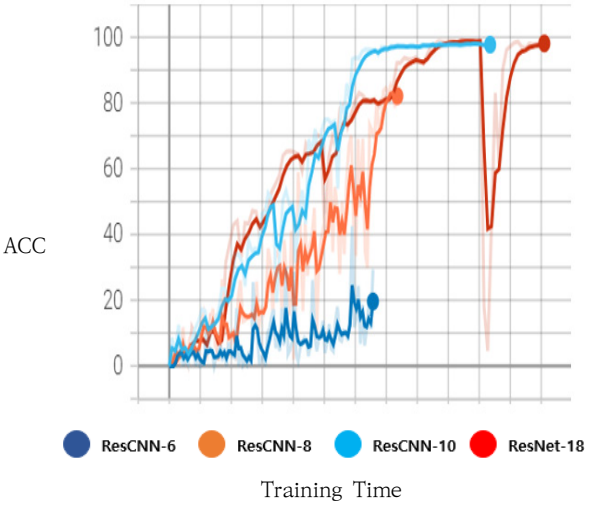


Fig. 7. Training result of 4 Models.
그림 7. Training result of 4 Models

그림 7은 GPU RTX3090 환경에서 100 Epoch에 따른 학습 결과를 보여주는 그래프이며, 가로의 길이는 학습 시간의 길이를 나타내 모델들 간의 학습시간 차이를 시각적으로 보여준다. 표 3을 통해 ResCNN-10과 ResNet-18의 레이어 수가 많이 차이나지만, 비슷한 정확도의 성능을 보인다. 또한 모델의 레이어 수가 적어질수록 Inference Time, MAC, Param는 작아짐을 보여준다. 따라서 레이어의 수가 적어짐에 따라 모델이 경량화됨을 알 수 있다.

Table 3. Accuracy and lightweight performance comparison between used models
표 3. 사용된 모델 간의 정확도 및 경량화 성능 비교

Model	ResCNN 6	ResCNN 8	ResCNN 10	ResNet 18
Acc	42.52	85.27	98.21	99.11
Training Time	3m 32s	4m 16s	5m 13s	5m 39s
Inference Time	0.002 Sec	0.004 Sec	0.006 Sec	0.012 Sec
MAC	467.05 M	650.91 M	841.94 M	1.8 G
Param	307.9 k	1.23 M	4.9 M	11.17 M

Table 4. clustering results between models.
표 4. 모델들 간의 화자 클러스터링 결과

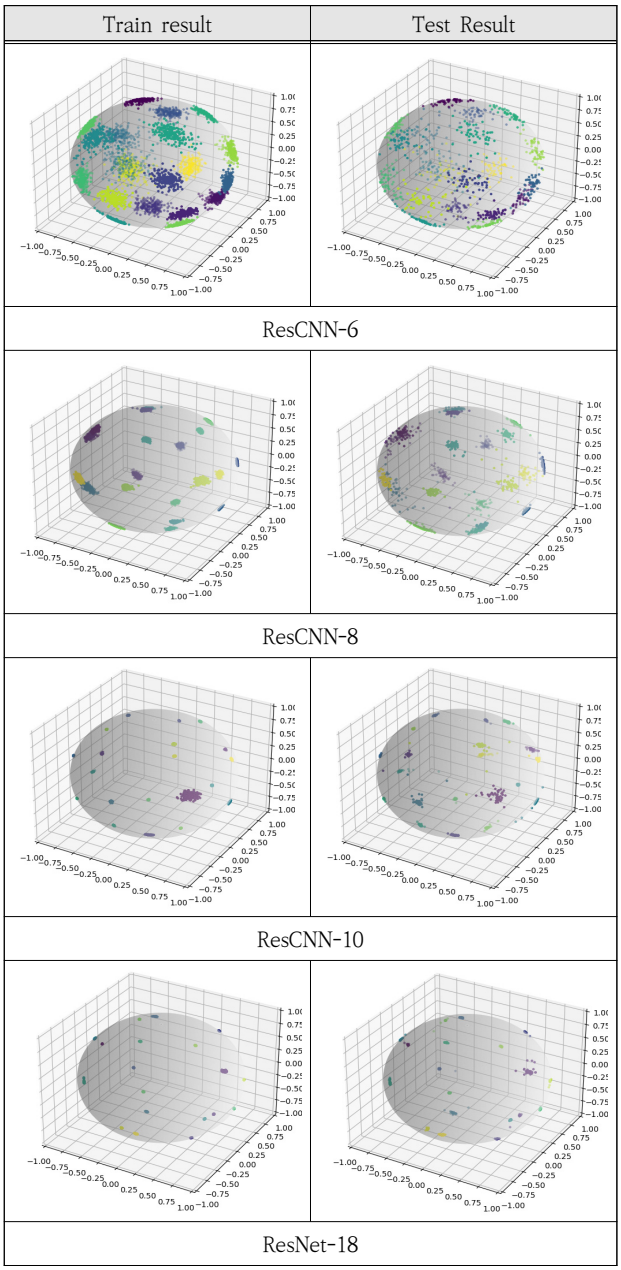


표 4는 학습 및 테스트 데이터에 대해 각 모델들의 클러스터링 결과를 보여주며, 구에 표시된 각 색상마다 각기 다른 화자들을 의미한다. ResCNN-6 모델의 각 화자 클래스들은 클러스터가 잘 형성되지 않아 분류가 잘 이루어지지 않는 것을 확인 가능하며, ResCNN-10 및 ResNet-18 모델은 클러스터가 잘 형성되어 화자들 간 분류가 잘 이루어지는 것을 확인 할 수 있다.

Ⅲ. 결론

이 논문은 각기 다른 화자들로부터 녹음된 펫 로봇용 명령어 발화를 활용하여 특징 추출 프레임워크를 통해 음성 특징을 추출하고 Residual Block을 활용한 ResCNN-6, ResCNN-8, ResCNN-10, ResNet-18 모델을 학습 및 비교한다. ResCNN-6은 모델이 가장 가벼워 Inference Time, MAC, Param에서 좋은 결과를 보였지만 정확도의 결과에 있어 음성 특징을 잡아내지 못했으며, ResNet-18은 정확도의 결과에 있어 가장 좋은 성능을 보이지만 Inference Time, MAC, Param에서 좋지 않은 결과를 보인다. 정확도와 경량화 사이의 트레이드 오프를 고려해, ResCNN-10 모델이 ResNet-18 모델과 정확도가 가장 유사하며, Inference Time, MAC, Param도 ResNet-18 모델의 절반 이상으로 줄인다. 따라서 20명 정도의 적은 화자의 수와 짧은 발화의 형식을 가지는 펫 로봇용 화자인식에 있어서 이 논문에서 제시하는 ResCNN-10 모델은 충분한 정확도를 보이면서 경량된 모델로 사용될 수 있음을 보여준다.

향후 특징 추출 프레임워크에 추가적인 특징 추출을 통해 차원을 줄여 모델에 가용할 파라미터를 줄이고, 이를 활용할 다양한 모델들의 설계와 연구를 통해 모델의 경량화 고도화를 진행할 예정이다. 또한 이 논문에서는 문장종속 방식의 데이터 세트를 사용했지만, 문장독립 방식 데이터 세트를 활용해 실험을 진행할 예정이다.

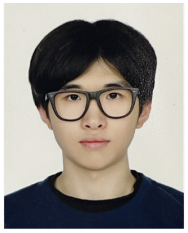
References

[1] Campbell, J. P. (1997). “Speaker recognition: A tutorial,” *Proceedings of the IEEE*, vol.85, no.9, pp.1437-1462, 1997. DOI: 10.1109/5.628714
[2] Logan, B, “Mel frequency cepstral coefficients for music modeling,” *In Ismir* Vol.270, No.1, pp. 11, 2000.

[3] He, K., Zhang, X., Ren, S., & Sun, J. "Deep residual learning for image recognition," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.770-778, 2016.
DOI: 10.1109/CVPR.2016.90
[4] Hannah, A. A., & Agordzo, G. K, "A Design of a low-pass FIR filter using Hamming Window Functions in Matlab," *Comput. Eng. Intell. Syst,* vol.11, no.2, pp.24-30, 2020.
DOI: 10.7176/CEIS/11-2-04
[5] Singleton, R. C., "An algorithm for computing the mixed radix fast Fourier transform," *IEEE Transactions on audio and electroacoustics*, vol.17, no.2, pp.93-103, 1969.
DOI: 10.1109/TAU.1969.1162042
[6] Muda, L., Begam, M., & Elamvazuthi, I. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *arXiv preprint,* 2010.
DOI: 10.48550/arXiv.1003.4083
[7] Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. "On empirical comparisons of optimizers for deep learning," *arXiv preprint,* 2019.
DOI: 10.48550/arXiv.1910.05446

BIOGRAPHY

Seong-Hyun Kang (Member)



2022 : BS degree in Information Security Engineering, Sangmyung University.
2022: MS candidate in Applied Artificial Intelligence, Hanyang University.

Tae-Hee Lee (Member)

2021 : BS degree in Electronics Engineering, Hanyang University.
2023 : MS degree in Electronics Engineering, Hanyang University.
2023 : Ph.D candidate in Electronics Engineering, Hanyang University.

Myung-Ryul Choi (Member)

1983 : BS degree in Electronics Engineering, Hanyang University.
1985 : MS degree in Computer Engineering, Michigan State University.
1991 : Ph.D degree in Computer Engineering, Michigan State University
1991~1992 : Research Engineer and Assistant Prof. KITECH
1992~ : Professor, Electrical Engineering in Hanyang University.