

인공지능 기법(CNN)을 이용한 음성과 음악구분

한상일

한국기술교육대학교 산업경영학과

Speech-Music Discrimination Using Deep Learning

Sang-Il Han

Department of Industrial Management, Korea University of Technology and Education

요약 음성, 음악 구분 방법은 전통적으로 푸리에 분석치의 특성치를 직접 이용하는 방법이 사용되어 왔다. 하지만 딥러닝을 이용하면 end-to-end 방식으로 특성치를 별도로 부여하는 과정을 거치지 않고 구분이 가능하다. 본고에 사용된 자료는 국내 음악방송(FM 89.8MHz)에서 추출된 약 389만개의 음성, 음악 파일이다. 송출된 소리를 5초 단위로 구분하여 음성 2,401,040개 파일, 음악 1,489,168 총 3,890,208개의 파일을 구성했다. 5초 단위로 소리를 끊는 것은 일반적으로 음악 핑거프린팅이 5초 단위 소리를 사용하여 구분하기 때문이다. 이러한 자료에 대해 딥러닝 분석 기법인 CNN분석을 적용하되 기존의 이미지 학습을 이용한 전이학습(transfer learning)을 적용했다. 실험결과 모형의 복잡성을 높이지 않고도 기존 학습 모형을 응용해 약 89.6% 전후 정확도가 나왔다. 또한 혼동 행렬을 이용하면 음성을 음악으로 판단하는 오류는 12.3%, 음악을 음성으로 판단하는 오류는 12.2%를 보였다.

Abstract Traditionally, a method of directly using the characteristic value of the Fourier analysis has been used to distinguish voice and music. However, if deep learning is used, it is possible to distinguish voice and music with an end-to-end method without the need for the process of characteristic featuring. The data used in this paper are about 3.89 million voice and music files extracted from a domestic music broadcasting (FM 89.8MHz). The transmitted sound is divided into 5-second units. Breaking down sound into 5-second increments is applied because music fingerprinting generally uses 5-second increments to distinguish sounds. The CNN analysis, a deep learning analysis technique, was applied to these data, and transfer learning was performed using existing image learning models. An accuracy of about 89.6% was obtained in the existing learning model without increasing the complexity of the deep learning model. In addition, while using the confusion matrix, the error of judging voice as music was 12.3%, and the error in judging music as voice was 12.2%.

Keywords : Speech-Music Discrimination, Fingerprint, Deep Learning, CNN, Mel-Spectrum, Transfer Learning

1. 서론

음성 구분 기술(audio finger printing)은 크게 Shazam 알고리즘(Wang(2003))을 이용하는 방법과 DTW(dynamic time warping, Lijffijt et al(2010))을 이용하여 발전되

어 왔다[6,9]. 이중 Shazam 방법은 푸리에(Fourier) 변환인 MFC(mel-frequency cepstrum) 등으로 음악의 극치(peak)를 파악한 후 해싱함수를 이용한 인덱싱으로 점수를 부여하는 방식이다. Yun et al.(2016)은 mel 함수로 전처리한 음악을 CNN으로 처리해 음악인식 및 분

이 논문은 2019년도 한국기술대학교 교수 교육연구진흥과제 지원에 의해 연구되었음.

*Corresponding Author : Sang-Il Han(Korea Univ. of Technology.)

email: sihan@koreatech.ac.kr

Received June 17, 2021

Revised September 1, 2021

Accepted October 1, 2021

Published October 31, 2021

류 문제를 다루고 있다[2]. 반면 DTW도 극치를 이용하나 두 음악의 극치간 경로 탐색을 통해 원음악을 파악해낸다. 그런데 이미 Shazam, Bmat, ARCCloud 및 Naver 등에서 상용 제품이 나와 2천만 곡이 넘는 세계 음악중에서 소비자의 음악취향에 맞는 음악 선곡이 정교하게 가능해지고 있다. 동 분야는 MIR(music information retrieval)으로 칭해지며 Muller(2015) 등에 의해 연구가 집대성 되었다[8]. 우리나라는 물론 세계적으로 인터넷을 통한 음악 방송 송출도 활발하며, 공중파, 인터넷 상에 송출된 음악에 대해 저작권에 따른 징수를 위해 MIR을 이용하여 원음악을 파악하는 것이 매우 중요하다. 이러한 MIR을 정교하게 적용하기 위해서는 먼저 방송된 원음을 저장하여, 데이터베이스를 구축하고 여기에 저장된 음악이외 음성과 음악을 구분하고 음악에 대해서만 퍼거프린팅이 수행되어야 한다. 이때 음악이외 음성에는 진행자의 발언, 광고방송 등 다양한 소리가 섞여있다. 이 중 특히 진행자의 발언과

음악을 구분하는 게 중요하며 일반적으로 진행자의 발언은 악기와 섞여 송출되지 않고 순수한 인간음성에 가까운 소리이다.

음성과 음악을 구분하는 기존 방법에는 푸리에 변환인 스펙트럼을 이용한 기법이 많이 사용된다. 즉 일정한 시계열적 특성을 띠는 소리에 대해 일정 구간을 정해 푸리에 변환을 가함으로써 시간-주파수 공간으로 소리를 표현하는게 가능하며 이때 주파수상의 극치를 이용하여 음성과 음악의 구분이 가능하다[1]. 이러한 접근법은 수학적 정교성에도 불구하고 그 성능이 제한되는 것으로 평가된다. 이외에도 Yun et. al(2016)은 심층 신경망을 이용하여 장르 분석을 연구했다[2]. 본고는 음성, 음악 구분 문제를 소리에 대한 분류(classification) 문제로 보고 심층 신경망 기법인 end-to-end 방식으로 접근하고자 한다. 즉 기존 MIR 방법과 유사하게 푸리에 변환을 하면 시간-주파수 정보가 얻어지며 이를 이미지로 간주하여 딥러닝 기법인 CNN을 적용할 수 있다.

음성 스트리밍 작업을 효과적으로 처리해주는 Gs-streamer 등을 이용하면 인터넷 스트리밍 음악에 대해 직접 일정 시간 간격으로 소리를 저장하는 것이 가능한데 본고는 자바로 직접 분리코드를 작성 녹음된 음향을 분리했다. 국내 음악방송(FM 89.8MHz)에서 송출된 소리를 5초 단위로 구분하여 음성 2,401,040개 파일, 음악 1,489,168 총 3,890,208개의 파일을 구성하여 딥러닝 분석 기법인 CNN분석을 사용했다. 일반적으로 음악 퍼거 프린팅을 하는데 5~10초 길이의 음악을 사용한다.

물론 이보다 짧은 3초 정도를 가지고 하는 경우도 있다. 본고에서는 5초 길이의 음성, 음악 파일을 사용하겠다. 실험 결과 모형의 정교성을 높이지 않고도 약 89.6% 전후 정확도를 보였다. 또한 혼동 행렬을 이용하면 음성을 음악으로 판단하는 오류는 12.3%, 음악을 음성으로 판단하는 오류는 12.2% 를 보였다. 모형의 정확도는 학습 횟수 등을 늘리면 90% 이상 달성될 수 있다.

본 논문은 다음과 같이 구성된다. 제 2장에서 음악과 주파수의 특성을 먼저 보고, 이를 인공지능에 적용하기 위한 자료변환 방법 그리고 인공지능중 CNN에 적용방법 및 실험 결과를 제시한다. 그리고 제 3장에서 결론을 제시한다.

2. 음성-음악에 대한 인공지능 학습

2.1 음악, 음성의 주파수 특성

음악에 대한 분석 방법은 크게 음악구조 분석(structure analysis), 코드 인식(chord recognition), 템포-비트 추적 및 음악 추출(audio retrieval) 등으로 구분된다. 이에 대한 간단한 소개는 www.musicinformationretrieval.com 에 잘 정리되어 있다. 음악도 이미지나 기타 음성과 유사하게 특색(feature)을 파악하여 학습하는 과정을 거치게 된다. 동 사이트의 signal analysis and feature extraction 부분을 보면 librosa 패키지를 이용하여 특색을 추출하는데 이에는 스펙트럼, 제로크로싱율(zero crossing rate), 에너지, 푸리에 분석, 음색(chroma) 등이 제시되고 있다. 따라서 딥러닝을 이용함에 있어 다양한 음색기법이 적용 가능하나 본고는 Shazam 알고리즘을 통해 성능이 검증된 스펙트럼 기법으로 특색치를 파악한다.

이렇게 특색을 분석하는 기법을 정하고 음악과 음성간 특색을 구분하면 사람의 소리는 그 특성치의 변화가 연속적으로 이루어지는 것에 반해 악기의 경우 특성치의 변화가 단절적이다[1]. 음악은 화음에 기반에 작곡되는 것이 태반이므로 특정 화음에서 연속성을 보인 후 타 화음으로 이동하는 순간 주파수 변화가 크다. 반면 인간의 음성은 주파수 지속시간이 짧고 변화가 점진적으로 이루어진다[1]. 따라서 동 특성을 이용하여 음성, 음악 구분이 가능하다. 즉 특정 시점-주파수 점에서 피크가 발생하는 정도를 이용하여 주파수 변화량을 측정할 수 있다.

2.2 대상 음악, 음성의 특성 분석

다음 Fig 1은 본고에서 사용된 음악중 하나인 선인장(이재경 작사, 이상준 작곡)의 스펙트럼 특성을 보여주고 있다. 음성 신호는 단일 채널 측면에서 보면 일종의 일차원 시계열로 볼 수 있다. 이를 스펙트럼 처리하는 경우 Fig 1에서 보듯 2차원 이미지로 전환할 수 있다. 이외에도 이미 언급한 제로크로싱율, 에너지, 푸리에 분석, 음색 등 다양한 특성치가 추출될 수 있다.

소리 자체는 3차원 파동이 시공간 상에서 발생하는 현상이다. 우리 인간은 파동에 의해 전달된 에너지를 이차원과 유사한 형태를 띠는 고막을 이용하여 인식한다. 인간이 외부 물질을 인식하는 방법은 빛 및 소리 등의 파동의 에너지를 통해 인식하게 되는데 매체들은 측정된 물리적 에너지를 특정 형태로 저장하는 것이라 할 수 있다. 음악을 매체를 이용하여 저장하는 경우 실수 값들이 벡터 형식으로 저장된다. 즉 시간 축 t 에 에너지 강도를 나타내는 실수값 $f(t)$ 이 저장되는 형태이다. 파동함수 $f(t)$ 는 푸리에 변환인

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt$$

을 이용하여 정의되며 상기 정의는 복소수 값을 가지므로 주파수, 각도 및 길이를 갖는 3차원 공간으로 변환이 가능하다. 이때 소리 값에서 적분이 가능하게 y 축에 대칭을 가정하면 된다. 이렇게 정의한 푸리에 변환에 오일러 공식을 이용해 주파수당 에너지에 대응하는 스펙트럼을 정의할 수 있다. 보통 소리에 대한 푸리에 분석은 이 스펙트럼을 이용해 수행되는데 상기 식을 이산식으로 전환한 후 적분 공간을 제약하면 각 시점에서 푸리에 분석이 가능하다. 따라서 스펙트럼 값을 y 축에 나타내면 2차원 이미지가 생성된다.

본고는 이중 스펙트럼을 이용한 특성치만을 이용해 음성, 음악을 신경망을 이용해 분석하고자 한다.

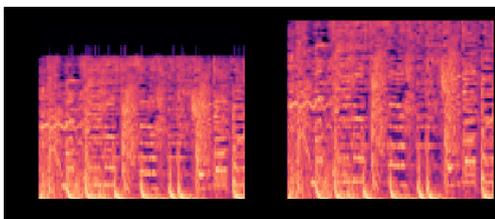


Fig. 1. Mel spectrum of Cactus(선인장 in korean)

2.3 CNN 기반 음성, 음악 구분

머신러닝은 크게 분류(classification) 및 군집(clustering) 등으로 구성된다. 일반적으로 분류는 이진 분류인 경우 종속변수 y 가 0, 1에 대응하게 된다. 고전적으로 분류는 로짓분석 등을 통해 수행될 수 있다. 이때 푸리에 분석을 통해 만들어진 2차원 정보 x 에서 특성치를 추출하여 로짓 분석이 가능하다. 하지만 이미지와 같은 다차원 정보가 내재된 경우 즉 독립변수 x 가 이미지인 경우 잡음 등에 대한 특정 가정을 통해 수행되는 로짓 모형은 한계를 띠는 것으로 판단된다. 그런데 머신 러닝의 일종인 딥러닝은 신경망(neural network) 기법으로 Hinton et al.(2006)에 의해 그 수렴원리가 규명된 이후 크게 발전하여 왔다[5]. 신경망은 인간의 뉴런을 모방한 모형으로 입력 변수에 대해 다층의 노드를 구성한 후 노드간 가중치를 에러-교정 역전파(back-propagation) 학습법을 이용하여 분류 등을 수행한다. 다층 신경망인 딥러닝은 복잡한 함수를 추정해 내는 능력을 보이며 이미지 구분 등에서 탁월한 효과를 보이고 있다[4]. 신경망 모형중 컨벌류션을 이용해 분류를 하는 CNN(convolutionary neural network)은 이미지 분석에서 그 활용도가 가장 높은 모형으로 평가된다.

컨벌류션 연산자는 기준점을 중심으로 주변 값을 이용해 가중 평균을 하는 방법으로 볼 수 있다. 두 함수 $f(x), g(x)$ 에 대해 컨벌류션은

$$(f * g)(x) = \int_0^t f(\tau)g(x - \tau)d\tau$$

으로 정의된다. 즉 모형에 따라 주어진 함수 g 를 x 에서 고정하고 f 를 가중치로 보고 적분한 것으로 해석된다. CNN은 이러한 컨벌류션을 신경망 모형으로 구현한 것으로 가중치인 f 를 역전파로 학습으로 잡아내는 방법이다. 이에 대한 직관적 설명은 Li가 강의하고 있는 스탠포드 대학의 CNN 강의인 <http://cs231n.stanford.edu/>를 참조하면 된다[7]. 가중 평균을 사용함에 따라 잡음에 강건한 모형이 되며 음성, 음악을 0, 1에 대응시키는 경우 감독학습을 통해 분류가 이루어 질 수 있다. 다음 실제 CNN은 컨벌류션 연산자 부분과 풀링(pooling) 부분으로 구성되는데 이를 다층으로 구성하여 입력 값에 숨겨진 복잡한 정보를 잡아내는 것이다. Fig 2는 CNN 모형을 나타내고 있는데 $28 \times 28 \times 1$ 차원의 이미지 정보가 행렬로 주어지면 컨벌류션을 이용하여 $24 \times 24 \times n_1$

차원의 행렬로 변환하고 이를 풀링 함수에 적용해 $12 \times 12 \times n_1$ 행렬로 변환하는 것을 보여주며 마지막 단계에서 벡터를 만들어 판단을 함을 보여준다. 이때 층의 개수, 풀링 방법에 따라 다양한 CNN 모형이 구성될 수 있다. 본고에서 사용되는 목적함수는 음악, 음성간 구분이고 범주형 변수이므로 로짓분석과 동일하게 교차 엔트로피(cross entropy)로 구성할 수 있다. 또한 입력변수는 소리를 스펙트럼 처리한 2차원 이미지가 된다. 따라서 음악의 파편을 듣고 음원을 구분하기 위해 점수 또는 순위를 사용하는 음악구분(finger printing)과는 다른 특색을 보인다.

구체적으로 CNN 모형을 적용하려면 사용 컴퓨터 언어, 딥러닝 툴 등이 먼저 정해져야 한다. 본고는 파이썬(Python) 언어를 사용하고 이에 기반한 파이토치(Pytorch) 툴을 사용하여 딥러닝에 적용하고자 한다. 음악을 분석하는데 있어 파이썬 및 파이토치만으로는 분석 능력이 제한되므로 음악 분석 툴로 Torchaudio와 Fastai audio v1을 추가로 사용했다. Fastai는 Pytorch 초창기에 제시된 툴로 최근에는 v3 버전까지 나왔으며 파이토치에 그 기능이 대부분 흡수되고 있다. 본고는 Fastai v1에서 제공하는 툴로 음악 및 음성의 스펙트럼을 구했으며 이를 입력 변수로 사용했다. 이때 잡음을 제거하지 않은 상태로 스펙트럼을 구해 사용했다. 음악과 음성을 별도로 구분한 디렉토리에 저장한 후 정규식 함수인 re(regular expression) 함수를 이용해 종속변수인 y 값에 대한 라벨링을 0, 1로 했다. 종속변수가 0, 1이므로 이는 Logit 회귀분석을 이용한 분류와 유사한 측면이 강하다. 로짓 분석의 경우 모수가 제한적이므로 적합도가 낮은 측면이 있으나 표본외 샘플에서 판단력이 좋을 수 있다. 일반적으로 독립변수인 X의 구조가 단순하면 로짓 분석도 유용한 것으로 판단된다. 그런데 음성 자료의 경우 구조의 복잡성으로 인해 로짓 분석보다 신경망에 의한 분석이 유용하다 본다. 자료에 대해 난수적

으로 80%를 추출해 학습용으로 쓰고 나머지 20%를 검증용으로 사용했다. CNN 모형으로 토치에서 제공하는 Conv2d 모형을 적용했으며 컨벌루션 연산을 위한 스트라이더(strider)로 (2,2) 그리고 패딩(padding)으로 (3,3)을 사용했다.

2.4 실험 및 결과분석

본고에 사용된 자료는 국내 음악방송(FM 89.8MHz)에서 추출된 약 389만개의 음성, 음악 파일이다. 송출된 소리를 5초 단위로 구분하여 음성 2,401,040개 파일, 음악 1,489,168 총 3,890,208개의 파일을 구성했다. 5초 단위로 소리를 끊는 것은 일반적으로 음악 핑거프린팅이 5초 단위 소리를 사용하여 구분하기 때문이다. 전세계적으로 음악관련 수만개의 인터넷 스트리밍이 이루어지고 있으며 이런 방식의 음악송출에 대해서는 음향 처리가 공개 소프트웨어로 용이하게 처리될 수 있다. 공중파로 전송되는 경우 음향을 수신한 후 컴퓨터가 가공 가능한 상태로 변환이 필요하다. 이렇게 구성된 자료에 대해 인공신경망 기법인 CNN을 적용하였다. 이때 학습 가중치를 난수로 주고 학습하는 경우 비효율성이 발생한다. 음성을 이미지 정보로 전환하여 CNN 학습을 적용하려면 이에 상당기간이 소요되며 학습 효율성을 담보하기 어렵다. 이에 본고는 전이학습(transfer learning) 기법을 적용하기 위해 Resnet18을 사용했고 첫 번째 층으로 2차원 자료에 대해 CNN을 구성해주는 Con2d 함수를 사용했으며 구체적으로 con2d(1, 64, kernel_size(7,7))로 설정했다. Resnet 18은 사전 nsfus된 신경망으로 18개의 계층으로 구성되며 보통 백만개 정도의 영상으로 기 훈련된 모형이다. 즉 본고는 기존 이미지 자료에 대해 이루어진 학습 결과를 이용하여 추가적인 학습을 하는 것이다, 따라서 본고는 기 훈련된 Resnet 18에 음성 신호에 대해 퓨리에 변환 처리된 2차원 행렬을 적용한 것으로 볼 수 있다. 역전파 학습을 함에 있어 학습속도는 Fastai에서 적절히 최적화하여 조절한다. 실험은 인텔 i9 cpu, Nvidia Titan-X GPU 컴퓨터로 수행했다. Fastai의 fit_one_cycle 함수를 이용하여 일회 학습을 한 후 재 학습을 한 결과인 학습자료 및 검증 자료에 대한 손실치 및 정확도가 Table 1에서 제시되고 있다. 일반적으로 정확도가 89.6% 정도에서 달성되고 있는 것을 알 수 있다. 신경망 모형을 좀 더 보완하고 학습 횟수를 늘리는 경우 93~94% 까지 달성이 가능하다. 또한 Fig 3의 왼쪽 그림은 학습 횟수에 따른 오류를 보여주고 있으며 학습

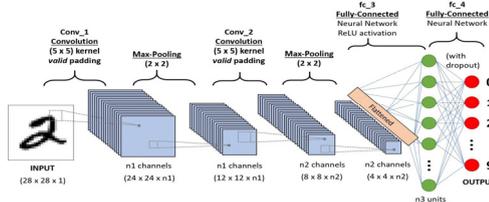


Fig. 2. Data flow in CNN
Source : Toward data science

횟수가 늘어남에 따라 오류가 안정적으로 줄고 있어 모형이 비교적 정확하게 작동하고 있음을 보여준다. Fig 3의 오른쪽은 혼동행렬을 보여주고 있는데 제 1조 및 2종 오류가 비슷한 수준을 보여주고 있어 모형의 정확도가 높다는 것을 보여준다. Fastai의 fit_one-cycle 함수를 이용하여 일회 학습을 한 후 재학습을 하므로 초기부터 학습 정확도가 매우 높음을 알 수 있다. 필요시 학습 횟수를 대폭 증가시킬 수 있으나 이보다 정확도를 높이기 위해선 전처리 등에서 음성, 음악 구분 특성치를 부여하고 자료를 늘리는 것이 더 유용하리라 본다.

Table 1. Accuracy trend according to the number of training

epoch	훈련 오류	검증 오류	정확도
0	0.311237	0.290136	0.884785
1	0.301692	0.288071	0.883131
2	0.297452	0.283172	0.888093
3	0.286648	0.282540	0.890298
4	0.264098	0.281176	0.890298
5	0.270169	0.280967	0.893054
6	0.244241	0.275080	0.895259
7	0.243506	0.278776	0.895810
8	0.241784	0.276585	0.896913
9	0.245924	0.271669	0.896362

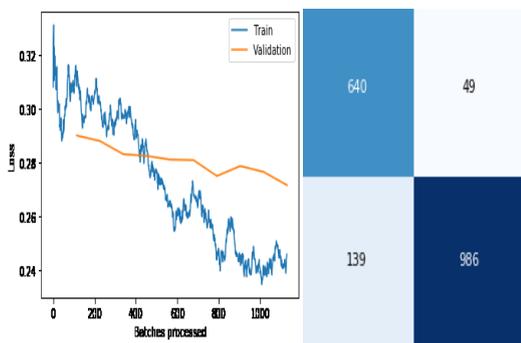


Fig. 3. Learning time and confusion matrix

3. 결론

기존의 스펙트럼을 이용한 극치분석을 통해 음악, 음성을 구분하는 경우 푸리에 변환 등을 이용해 특성치를 도출한 정보를 이용해 구분을 한다. 반면 심층 신경망은 이러한 극치에 대한 사전 정보에 기초한 통계량을 사용

하지 않고 자료에 의존한 end-to-end 감독학습이 가능한 방법론이다. 또한 기존에 타 자료에 대한 학습 모형인 Resnet 등을 이용하는 경우 학습 시간을 크게 단축하면서도 높은 성능을 보이는 장점을 갖는다.

본고는 국내 음악방송(FM 89.8MHz)에서 송출된 소리를 5초 단위로 구분하여 음성 2,401,040개 파일, 음악 1,489,168 총 3,890,208개의 파일을 구성하여 딥러닝 분석 기법인 CNN분석을 사용했으며 모형의 정교성을 높이지 않고도 약 89.6% 전후 정확도를 보였다. 또한 혼동 행렬을 이용하면 음성을 음악으로 판단하는 오류는 12.3%, 음악을 음성으로 판단하는 오류는 12.2% 를 보였다. 모형의 정확도는 학습 횟수 등을 늘리면 90% 이상 달성될 수 있다. 본고는 특성치를 산정함에 있어 스펙트럼 기법만을 적용했으며 다양한 방법을 시도하지는 않았다. 또한 Resnet 모형도 18층에 기초한 방법만을 사용했다. 향후 동 방법이 상업적으로 사용되기 위해서는 다양한 방법의 적용이 가능하며 이럴 경우 본고보다 매우 높은 수준의 정확도가 기대된다.

References

- [1] Kyong-Chul Yang, Yong-Chan Bang, Sun-Ho Cho, Dongsuk Yook, "Speech and Music Discrimination Using Spectral Transition Rate", *The Journal of Acoustic Society of Korea*, vol 28, 2009, pp 273-278
- [2] Ho-Won Yun, Seong-Hyeon Shin, Woo-Jin Jang, and Hochong Park, "On-Line Audio Genre Classification using Spectrogram and Deep Neural Network", *Korean Institute of Broadcast and Media Engineers*, Vol 21-6, 2016, pp. 977-985
- [3] Byeong-Yong Jang, Woon-Haeng Heo, Jung-Hyun Kim & Oh-Wook Kwon, "Music detection from broadcast contents using convolutional neural networks with a Mel-scale kernel", *EURASIP Journal on Audio, Speech, and Music Processing*, 2019
- [4] Bengio, Y., *Learning Deep Architectures for AI*, Now Publishers, 2009
- [5] Hinton, G., Osindero, S., and The, Y, "A Fast Learning Algorithm for Deep Belief Nets", *Neural Computation*, 18, 2006, pp. 1527-1554
- [6] Lijffijt, J., Papapetrou, P. and Hollmen, J, "Benchmarking Dynamic Time Warping for Music Retrieval", *PETRA*, 2010
- [7] Li, Fei-Fei, <http://cs231n.stanford.edu> , 2020
- [8] Muller, Meinard, *Fundamentals of Music Processing*, Springer, 2015

[9] Wang, A., "An Industrial-Strength Audio Search Algorithm", *ISMIR*, 2003

한 상 일(Sang-II Han)

[정회원]



- 1992년 2월 : 포항공과대학교 수학과 (이학석사)
- 1998년 8월 : 서강대학교 경영학과 (재무학박사)
- 1999년 2월 ~ 2005년 2월 : 한국금융연구원 부연구위원
- 2005년 3월 ~ 현재 : 한국기술교육대학교 산업경영학과 교수

<관심분야>

경영학 전반, 인공지능 응용