학사학위청구논문

2020학년도

심층신경망 학습을 위한

음성신호 특징 추출 방법론

Audio Signal Feature Extracting Process

for Automatic Speech Recognition based on Deep Nueral Network

광운대학교

전자통신공학과

원 철 황

심층신경망 학습을 위한

음성신호 특징 추출 방법론

Audio Signal Feature Extracting Process

for Automatic Speech Recognition based on Deep Nueral Network

지도교수 박 수 원

이 논문을 공학사 학위논문으로 제출함

2020년 11월 일

광운대학교
전자통신공학과
원 철 황

원철황의 공학사 학위논문을 인준함

지 도 교 수 _____ 인

광운대학교

전자통신공학과

2020년 11월 일

국문 요약

심층신경망 학습을 위한
음성신호 특징 추출 방법론

음성은 사람 간의 가장 자연스러운 의사소통 방식이다. 최근 컴퓨팅 파워의 상승과 빅데이터, 심층신경망 기반 기술 발달의 조화로 이러한 의사소통 방식을 기계와 사람으로의 영역으로 확장하려는 연구가 각광을 받고 있다. 이러한 기술은 사용자가 발생한 내용을 받아 적거나, 일상에서 발생하는 위협 감지, 기계와 사람간의 새로운 명령 전달 인터페이스 등에 활용될 수 있다.

본 연구에서는 음성인식에 활용되는 여러 음성신호 특징을 추출하는 방법과 그 특징을 소개한다. 또한 음성신호처리는 인간 발화 주파수 영역에 초점을 맞추어지는 이유를 이론적 배경과 특징 텐서를 추출하는 과정을 통해 설명하고자 한다. 이 특징 텐서는 인간의 청각이론을 바탕으로 하고 있으며, 주파수 영역에서 스펙트로그램 형태로 나타나거나 특정 영역의 에너지와 관계있는 계수 값으로 변환되어 각 주파수 영역당 하나의 대표값을 취하는 형태로 보여진다.

끝으로 해당 과정을 통해 얻은 Mel-Spectrogram, Spectrogram, MFCC, Mel-Filterbank Coefficient를 모델의 입력으로 하였을 때 결과를 음절오차율(CER, Character Error Rate)으로 나타낸다.

핵심어: 음성신호처리, 심층신경망, 음성인식

Abstract

Audio Signal Feature Extracting Process for Automatic Speech Recognition based on Deep Nueral Network

CheolHwang, Won

Dept. of Electronics and Communications Eng.

College of Electronics & Information Engineering

Kwangwoon University

Voice is the most natural mode of communication between people. With the recent rise in computing power and the combination of big data and deep neural network-based technology development, research to expand this communication method into the realm of machines and people is in the spotlight. These technologies can be utilized by users to write down what they speak, to detect threats in their daily lives, and to communicate new commands between machines and people.

This study introduces the process and how to extract the various speech signal features used in automatic speech recognition. In addition, would like to explain why speech signal processing focuses on the human ignition frequency domain through the theoretical background and process of extracting feature tensor. Thses feature tensors are based

on human auditory theory and apears in the form of spectrogram in the frequency domain or is converted into a coefficient value related to the energy of a particular region to take one representative value for each frequency domain.

Finally, when the Mel-Spectrogram, Spectrogram, MFCC, and Mel-Filterbank Coefficent are inputted to the model, the results are expressed as a CER(Character Error Rate).

Keywords: Speech Signal Processing, Deep Neural Network, Speech Recognition

차 례

그림 차례

코드 차례

표 차례

제1장 서론

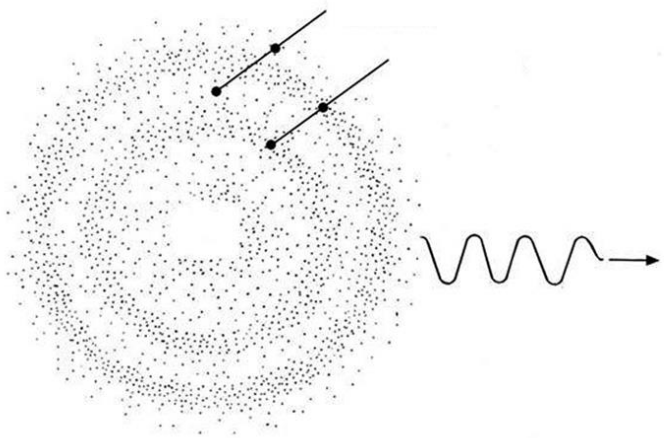
음성, 오디오, 소리는 방송과 대화, 여러 엔터테인먼트의 주류 요소로서 소위 킬러 어플리케이션이라 정의되는 기술 또는 서비스의 핵심요소이다. 아날로그 통신에서 디지털 통신으로의 전환에서 MMS, Video steraming이 차지했던 위치를 현재 심층신경망과 빅데이터 기반 기술의 도래 속, 음성인식이 그 자리를 차지하고있다. 특히 음성 AI 시장은 스마트 스피커 등에 내린 음성 명령으로 기기를 제어하거나 음악 스트리밍, 검색, 온라인 쇼핑 등에 활용하는 시장이며, 딥러닝 기술 발달, 다양한 고객 접점 단말 확대, 유저의 인식 및 행동 변화와 함께 부상하고 있다 [1].

그러나 AI 음성 엔지니어는 영상처리, 자연어처리 전문 엔지니어의 수와 비교하면 비교적 부족한 실정이다. 따라서 해당 연구에서는 청각이론을 바탕으로 한 음성인식에 사용되는 여러 특징들을 추출하는 과정을 소개하고, 그 예시와 결과를 소개한다. 따라서 AI 음성연구로의 연구 진입장벽을 낮추고, 해당 논문을 통해 처음 음성 인식을 접하는 많은 연구자들에게 기초적인 이론을 제공할 수 있기를 희망하는 바이다.

본 논문에서는 음성인식의 기본이 되는 인간 청각이론을 다룬다. 이후 특징 추출을 위한 신호처리 이론을 간략하게 소개하고, 이를 예시 샘플에 적용하여 나타나는 형태를 스펙트로그램 형태로 확인한다. 끝으로 각 특징 값들을 모델에 적용하였을 때, 나타는 결과를 나열하고 해당 결과를 해석해본다.

제2장 소리의 기본 개념과 요소

소리를 물리적 측면에서 정의하면, 음원으로부터 방사되는 압력파가 매질내에서 전달되는 것이라 정의할 수 있고, 이러한 소리가 일으키는 파동을 음파라고 한다[2]. 여기서 매질은 어떤 압력 등에 의해서 위치가 변화 되었을 때 가능한 빨리 자기 위치로 돌아오려는 성질을 가지고 있으며, 어떤 파동 또는 물리적 작용을 한 곳에서 다른 곳으로 옮겨주는 매개물 역할을 하는 것을 뜻한다. 그 예시로는 공기나 물 등이 있다.

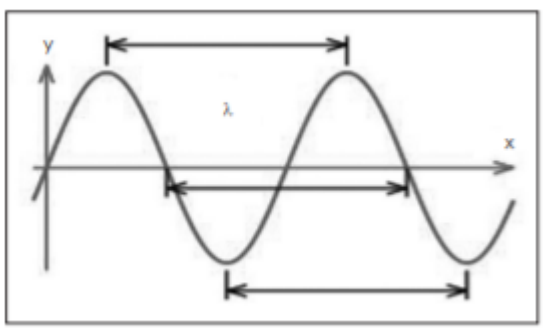


[그림 1] 음원으로부터 퍼져나가는 파동.

[그림 1]은 음원으로부터 퍼지는 파동이 입자를 진동시켜 매질의 밀도가 변화하는 것을 그림으로 나타낸 것이다. 주파수는 주기파에서 반복적으로 관찰되는 한 단위의 패턴이 일정한 시간 내에 얼마나 반복되는 가를 나타낸다[2]. 보통 알파벳 소문자 f로 표시하며, 단위는 보통 Hz, 헤르츠로

읽는다. 1Hz는 파동이 1초에 한 번 진동하는 것으로 같은 패턴이 1초 안에 한 번 존재한다고 이해할 수 있다. 주파수는 청각과 밀접한 연관을 가진다. 우리가 높은 음이라고 느끼는 것은 높은 주파수를 가진 음이다. 주파수에 대한 척도를 알아보는 방법으로 옥타브가 있다. 이것은 2를 밑으로 하는 로그 척도로 표현되며, 한 옥타브가 증가한다는 것은 주파수가 2배로 증가한 것으로 해석할 수 있다.

파장은 단순한 파장에서 동일한 패턴을 그리는 파동에서 한 주기에 해당하는 거리를 뜻한다. 이러한 파장은 왜곡 또는 감쇄와 같이 파동의 요소를 변화시키는 외부 요인의 영향을 제외하면, 파동 내에서는 항상 같다.



[그림 2] 단일 파동에서의 파장

[그림 2]는 임의의 지점에서 한 주기에 해당하는 지점까지의 거리를 이은 것으로 모두 같은 파장을 가짐을 보여준다. 이러한 파장은 그리스 문자 λ로 표시한다. 반면 주기는 시간의 개념으로 하나의 완전한 진동을 완성하는데 걸리는 시간을 의미하며 알파벳 대문자 T로 표시한다. 이 주기는 주파수와 반비례의 관계에 있다.

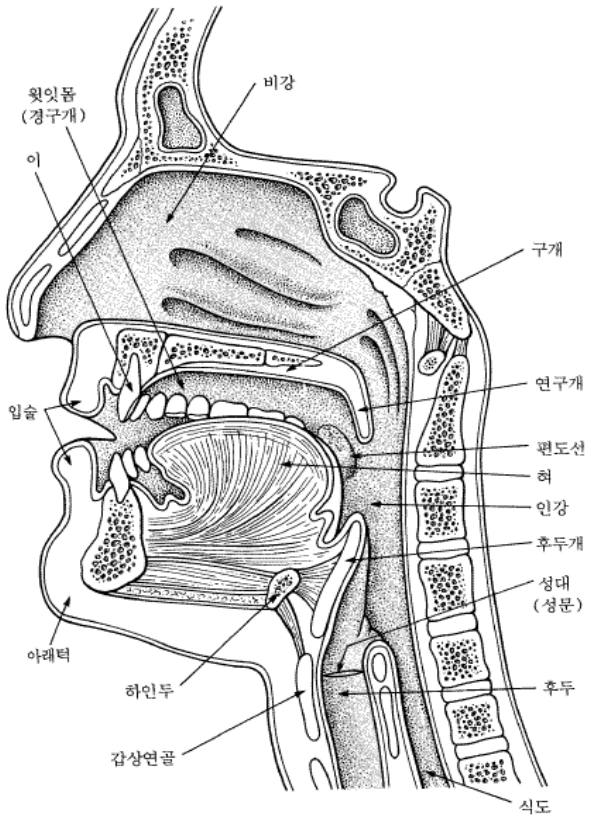
물리학적 측면에서, 진폭은 정적인 상태의 입자가 얼마나 많이 이동했는가라고 할 수 있다. 이는 그래프에서 표현될 때 y축에 위치하며, 공기분자가 진동에 의해 인접 공기분자를 미는 힘을 의미한다. 측정 스케일은 데시벨(dB)을 사용하며, 측정 단위는 파스칼(Pa)이다. 이러한 음압이 방향을 따라 전달된 힘을 강도(Intensity)라 하고, $Watt/m^2$의 단위를 사용한다. 강도는 음압의 제곱에 비례하는 관계를 가지고 있으며, 사람이 들을 수 있는 가장 작은 소리의 강도인 절대가청입계값 $10^{-16}\ Watt/cm^2$을 음압으로 환산하면 20μPa이 된다. 이렇게 언급된 강도와 음압의 단위는 소리의 절대적인 힘을 나타내는 단위이며, 소리의 상대적인 힘을 나타내는 단위는 데시벨(dB)이라 한다.

데시벨(dB)은 인간의 청각기관이 음의 강도를 감지하는 방식을 연구한 결과를 바탕으로 산출된 단위로 로그 단위를 바탕으로 하며 $dB = 20log_{10}(P_a/P_b)$의 관계를 가진다. 우리가 그래프에서 보는 진폭(Amplitude)의 기준 값은 위에서 제시한 절대가청임계값을 음압으로 환산한 20μPa을 기준으로 한다. 이값을 기준으로 계산된 강도의 수준을 음압수준이라 하고, 0 dB SPL이라 함은 우리가 들을 수 있는 최소한의 소리를 의미한다.

끝으로 위상이란 반복되는 파형의 한 주기에서 첫 시작점의 각도 또는 어느 한 순간의 위치를 말한다. 단주기 파동은 진폭에 해당하는 값과 위상 값만을 이용한 표현으로 나타낼 수 있으며, 이를 페이져 표현이라 한다. 위상차란 따라서 사인 곡선간의 시작점의 차이를 나타낸다. 사인 곡선은 주기적인 모습을 보이기 때문에 원과 연관하여 각도로 표현할 수 있다.

제3장 발성 이론

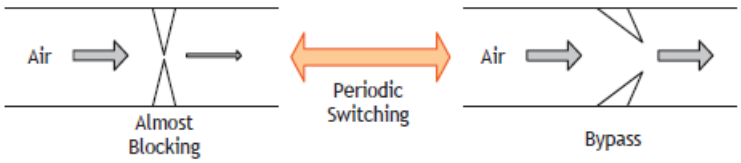
음성은 인간의 특별한 신체 기관을 이용하여 만드는 특이한 신호라 할 수 있다. 폐는 음의 세기를 결정하며, 성대(vocal cord)는 진동 장치로서 발성되는 소리의 높낮이, 즉 진동수를 결정한다. 끝으로 성도(vocal tract)는 주파수 변형 장치로서 발음을 결정한다.



[그림 3] 인간 발성 기관.

[그림 3]은 입술부터 인강까지 이어지는 내부 공간인 성도와 음의 높낮이를 결정하는 성대의 위치를 보여준다[3]. 개인의 목소리는 [그림 3]에 표시된

다양한 요소에 따라 달라진다. 성대의 단면 모양을 성문이라 하며, 이 모양은 사람마다 각기 달라 개인의 특징이 반영된다. 연구개란 소리에 비음이 섞이는 정도를 결정하는 기관으로 이 역시 개인 특성이 반영되는 영역이다. 또한 구강구조 상태, 혀의 위치 등의 요소 역시 영향을 끼친다.



[그림 4] 성대의 동작과 그 결과

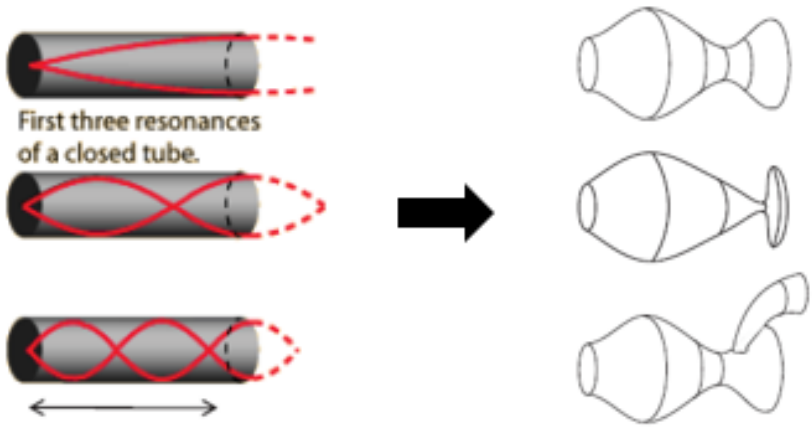
[그림 4]는 성대의 동작을 설명한다. 폐에서 형성된 공기의 흐름은 성대의 닫힘과 열림에 따라 일정한 주기를 가지게 된다. 성대는 벨브와 같은 역할로 관을 지나는 공기의 흐름을 막게 되며, 압력이 높아지게 되면 공기를 지나게 한다. 이같은 과정이 반복되며 주기를 가지게 되며, 남자가 압력을 버티는 정도가 더 높기 때문에 그 주기가 길어지게 되어, 대체로 낮은 주파수를 가지는 목소리를 보인다.

[그림 5] 성대 동작에 따른 결과.

[그림5]는 성대에 의해 형성되는 음을 유성음이라 한다. 성대의 진동에 의해 주기성을 가지게 된 신호로서 음의 높낮이가 존재한다. 유성음은 /b/, /d/, /g/, /v/, /z/ 등의 소리를 나타내며, 성대에 손을 대고 발음할 때, 진동을 느낄 수 있다.

이때, 주기성을 가지게 된 신호의 진폭의 일정함 정도를 나타내는 지표를 Shimmer, 주기의 일정함 정도를 나타내는 지표를 Jitter라고 한다. 이 값들이 소위 '듣기 좋은 목소리'를 결정하는 요소가 된다.

성대가 열린 상태로 지속된 발화 신호는 주기를 띄지 않으며, 음의 높낮이를 가지지 않는다. 이러한 신호를 무성음이라 한다. 무성음은 /p/, /t/, /k/, /f/, /s/ 등의 소리를 나타내며, 성대에 손을 대고 발음할 때, 진동을 느낄 수 없다. 이렇게 성도를 지나기 이전 성대를 통과하여 주파수 변형을 겪기 전 모든 신호를 Excitation이라 한다. 이렇게 형성된 음은 성도를 통과하며 주파수 변형을 겪게 된다.

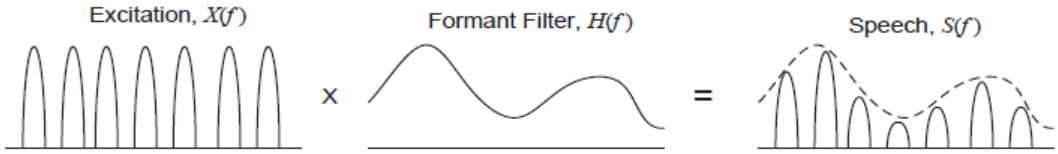


[그림 6] 단순 공명관과 변형된 공명관

[그림 6]은 평평한 파이프가 가진 길이에 의해 형성되는 정수배의 공명주파수를 보여준다. 우리 발성기관 내부 성도는 [그림 6]과 같이 다양한 모양을 가지는 형태의 튜브라 할 수 있다. 따라서 관의 모양, 길이, 굴곡에 따라 공명주파수가 결정된다고 가정할 때, 단순히 정수배가 아닌 다양한 공명 주파수 영역이 결정된다.

[그림 7] 공명에 따른 주파수 변형

[그림 7]은 위에서 설명한 특정 부분의 주파수 영역이 증폭되는 현상을 보여준다. 아래 결과와 같이 증폭되는 주파수를 포먼트(Formant)라고 하며, 해당 포먼트 주파수의 위치에 따라 발음이 결정된다.

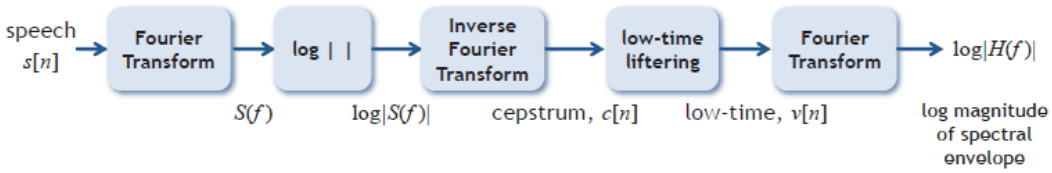


[그림 8] 포먼트 필터의 동작

[그림 8]은 이러한 공명이 주파수 영역에서 어떤 형태로 나타날 수 있는지를 보인다. 포먼트 필터의 역할은 주파수 영역에서 특정 주파수는 증폭하고, 다른 영역은 감쇄시키는 것이다. 익사이테이션은 포먼트 필터와 곱해져 끝으로 발음 정보를 가지는 하나의 음절로 나타나게 된다. 같은 형태로 나타낼 수 있다. 포먼트 필터의 역할은 주파수 영역에서 특정 주파수는 증폭하고, 다른 영역은 감쇄시키는 것이다. 익사이테이션은 포먼트 필터와 곱해져 끝으로 발

음 정보를 가지는 하나의 음절로 나타나게 된다.

우리가 발음을 구분하는 데 필요한 정보는 포먼트 필터이다. 이는 주파수 영역에서 느리게 변하는 성분이며, 이를 우리는 low time 이라 정의한다. 같은 개념으로 익사이테이션은 high time 이라 할 수 있다. 포먼트 정보는 X(f)와 H(f)를 분리하여 필터링하고, H(f)만을 취함으로써 얻을 수 있다. 따라서 로그 연산을 통해 high-time 성분과 low-time 성분으로 신호를 분리한 후 high-time 성분을 제거하여 원하는 정보 H(f)를 얻을 수 있다. 이때, log|S(f)|의 IDFT 값 c[n]을 캡스트럼(Cepstrum)이라 한다. 이는 스펙트럼의 앞 부분 Spec 을 거꾸로 돌린 뒤 붙인 것으로 새롭게 정의되는 값이다. 동일한 작명 원리로 필터링(Fil-tering)은 리프터링(Lif-tering)이 된다. 끝으로 변환된 c[n]에서 n을 프리퀀시(Frequ-ency)에서 변형된 큐프런시(Que-frency)라 한다 [3].



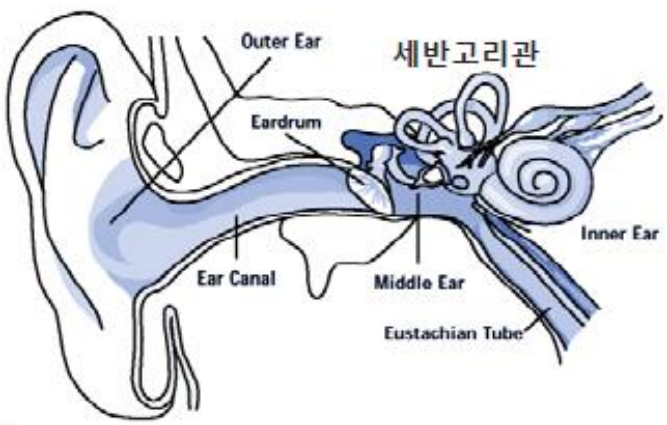
[그림 9] Low-time 성분을 추출하는 과정

[그림 9]는 주파수 축에서 발음 정보를 가진 H(f)를 필터링하는 과정을 나타낸다. 주어진 캡스트럼으로부터 high-time 성분을 필터링함으로써 느리게 변화하는 성분의 envelope을 얻어낼 수 있다.

제4장 청각 이론

인간의 귀는 단순히 듣는 행위에 그치지 않고 부가적인 정보를 인식한다. 우리는 2개의 귀를 이용해 소리를 통해 전달되는 내용적 정보 외 소리가 발생한 고도, 거리, 방위각 등을 동시에 인지할 수 있다. 해당 장에서는 이러한 인간의 청각 이론을 살펴보고, 소리의 특성과 연관한다.

청각 기관은 2개이기 때문에 정면에서 발생하지 않은 소리는 시간차를 두고 각각의 청각 기관에 도달하게 된다. 이 시간 차이를 이용해 우리는 방위각을 인지하게 된다. 이때, 방위각은 추론할 수 있으나 앞과 뒤를 구분할 수 없다. 해당 문제는 인간 귀의 형태를 관찰함으로써 해결점을 찾을 수 있다 [3].

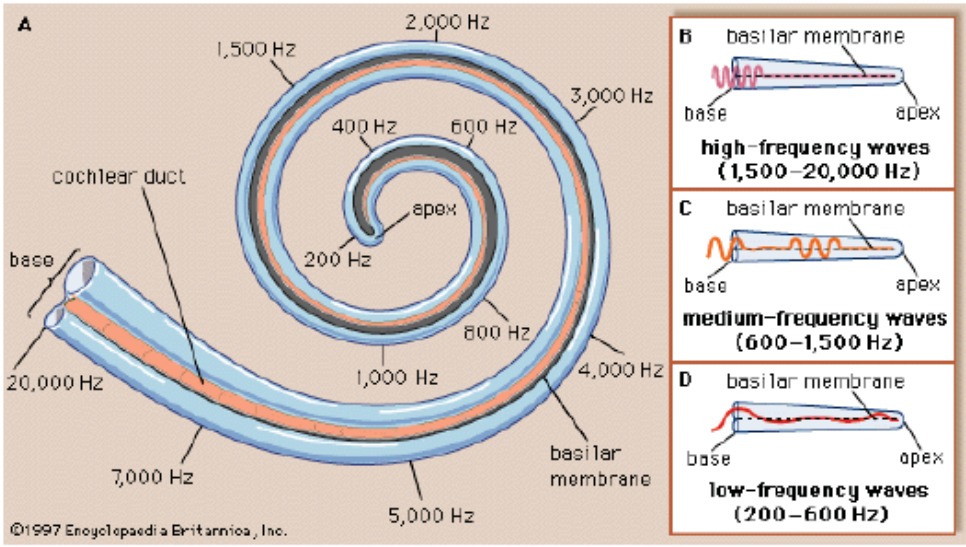


[그림 10] 청각 기관 해부도

[그림 10]은 인간의 청각기관을 나타낸다. 외이는 음파를 모으는 기능을 한다. 이때, 귓바퀴는 앞과 뒤, 위와 아래로 비대칭을 이룬다. 이는 같은 소리라도 소리의 발생 위치에 따라 서로 다른 방향전달함수(HRTF, Head Related

Transfer Function)를 갖는다. 따라서 사고를 통해 귓바퀴 모양이 바뀌게 되면 소리의 위치를 감지하는 기능을 일시적으로 상실하게 된다.

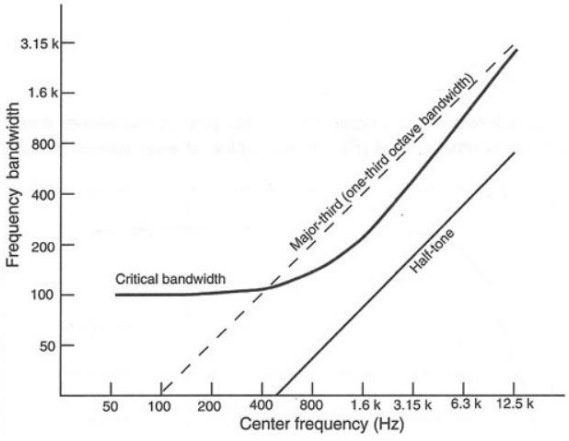
외이도(Ear Canal)를 따라 모인 소리는 고막(Ear Drum)으로 전달된다. 이 과정에서 3~4kHz 주파수 영역의 공명이 발생하며 12~15dB의 증폭이 발생한다. 이렇게 발생한 진동은 중이(Middle Ear)의 청소골에서 지렛대의 원리를 이용해 약 30dB의 증폭을 통해 달팽이관으로 전달된다.



[그림 11] 달팽이관과 주파수에 따른 진동 위치

[그림 11]은 달팽이관과 주파수에 따른 진동 위치를 개략적으로 보여준다. 전달된 음파의 주파수에 따라 달팽이관 내부에 있는 Basilar Membrane의 서로 다른 위치에 큰 파형이 위치하여 Membrane을 자극한다. [그림 11(B)], [그림 11(C)], [그림 11(D)]는 말려있는 달팽이관을 직선으로 폈을 때, 주어지는 입력 주파수에 따라 서로 다른 곳이 진동하는 모습을 보여준다.

[그림 11]을 통해 우리는 서로 다른 주파수가 어떻게 인지되는 지 확인할 수 있다. 그러나 인간의 달팽이관은 인지할 수 있는 주파수 범위가 선형적이지 않다. 임계대역(Critical Band)은 싱글톤(Single-tone) 음성과 노이즈를 함께 인지할 때, 노이즈의 대역폭에 의해 싱글톤 신호가 가려지는 최소 범위를 말한다. 이 임계대역이 좁다는 것은 해당 싱글톤 신호를 중심으로 임계대역의 반만큼 떨어진 주파수 영역의 신호는 인지할 수 있다는 것이다. 즉, 임계 대역이 좁을 수록 정교하게 동작한다는 것을 알 수 있다.



[그림 12] 중심 주파수에 따른 임계대역 크기

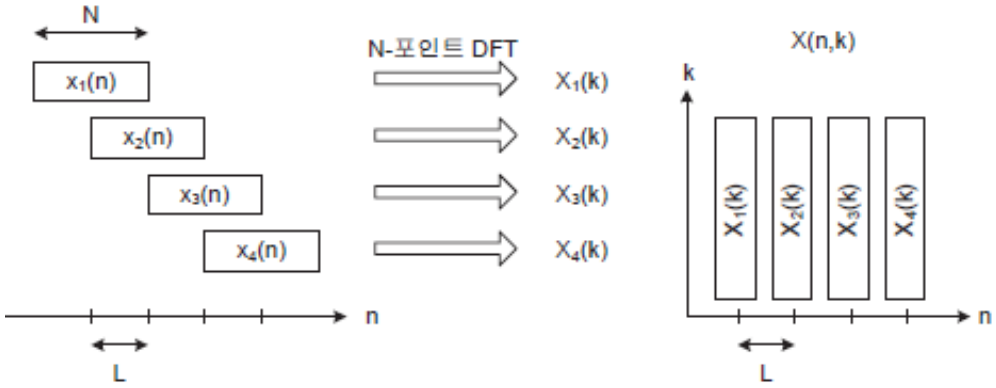
[그림 12]는 임계대역이 주파수가 높아질 수록 넓어지는 것을 보인다. 이렇게 인간이 주파수를 인지하는 민감도를 반영한 것을 멜 스케일(Mel-Scale)이라 한다. 이는 식 (1)과 같이 표현할 수 있다.

$$Mel(f) = 2595 \times log_{10}(1 + \frac{f}{700}) \quad \cdots\cdots (1)$$

끝으로 에너지 역시 log 연산을 적용하여 인간이 소리의 강도를 느끼는 정도를 dB 스케일링하면 청각 이론을 충분히 적용한 특징 그래프가 나타난다. 이를 Mel-band Log Energy라 한다.

제5장 단시간 푸리에 변환

푸리에 변환은 무한대의 시간 영역을 적분하는 작업이다. 이러한 푸리에 변환은 전체 시간 영역의 평균적인 주파수 특성을 보여주며, 페이즈가 시간적 특성 변화 정보를 제공한다. 그러나 음성, 오디오 신호는 시간 진행에 따른 주파수 성분 변화가 핵심적인 정보이다. 따라서 시간 진행에 따른 스펙트럼 정보 변화 과정을 담은 분석이 필요하기에 짧은 시간 영역에 한정된 푸리에 변환을 적용하고 이를 시간의 순서로 나타낸다. 이를 스펙트로그램이라 한다.



[그림 13] 단시간 푸리에 변환을 이용한 스펙트로그램 생성

[그림 13]은 그 과정을 개략적으로 보이고 있다 [4]. 이때, 잘리는 끝 부분에서 발생하는 급격한 신호 변화가 스펙트럼에 반영되는 것을 막기 위해 window를 적용한다. 이러한 윈도우는 해밍, 해닝, 블랙맨까지 다양한 종류를 띄고 있다.

```
        // hanning window
        for (int n =0; n < N; n++) {
                signal[n] = signal[n] * (0,5 -0,5*cos(2 * PI * n / (float)(N -1)));
        }
```

[코드 1] 해닝 윈도우 (Hanning Window)

[코드 1]은 C언어로 작성된 해닝 윈도우 작업이다. 이렇게 윈도우가 적용된 샘플에 대하여 N-포인트 DFT를 적용하면 한 샘플에 해당하는 스펙트럼을 얻을 수 있다. 이렇게 한정된 샘플에 대해서 DFT를 취하는 것을 단시간 푸리에 변환이라 나타낸다.

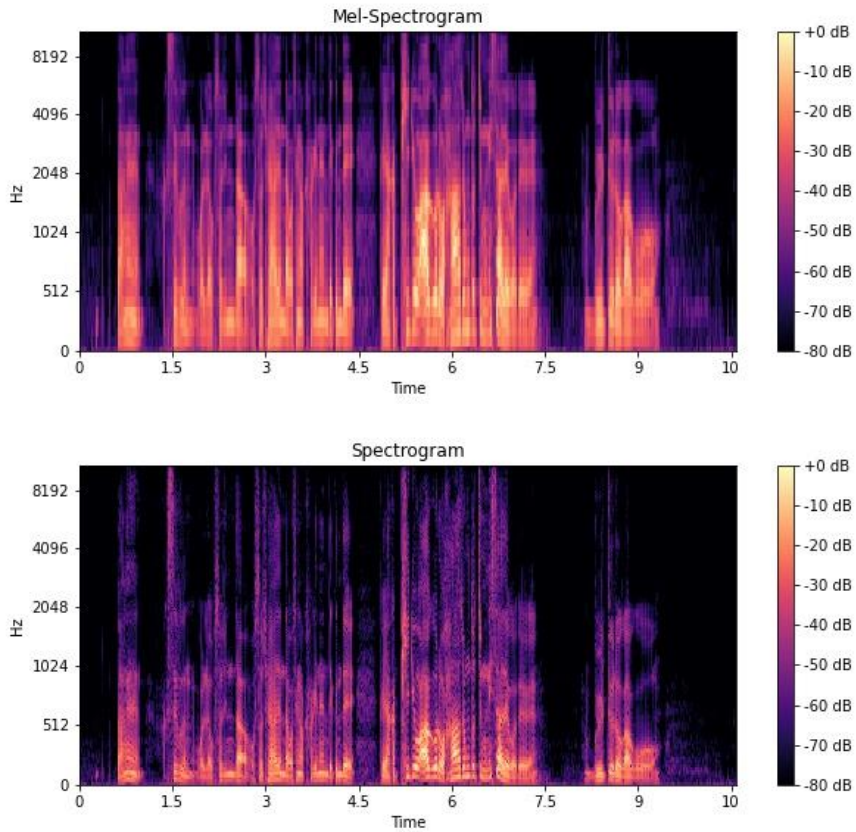
```
// N-point DFT
        for (int k =0; k < N; k++) {
                spec_real[k] = spec_imag[k] =0,0;
                for (int n =0; n < N; n++) {
                        spec_real[k] = spec_real[k] + signal[n] * cos(2 * PI * k * n / (float)N); // DFT Real
                        spec_imag[k] = spec_imag[k] - signal[n] * sin(2 * PI * k * n / (float)N); // DFT Imai
nary
                }
                spec_mag[k] = sqrt(pow(spec_real[k], 2) + pow(spec_imag[k], 2));                // |spec_mag[k]
|^2 = |spec_real[k]|^2 + |spec_imag[k]|^2
        }
```

[코드 2] N-포인트 DFT 코드

[코드 2]는 구해진 한 샘플에 대한 스펙트럼의 실수 크기를 구하는 과정을 보여준다. C언어에 자료형은 복소수를 지원하지 않으므로 실수항과 복소수항을 따로 계산하여 크기를 계산한다.

제6장 Mel-Spectrogram

멜스펙트로그램은 위의 과정을 거쳐 얻어낸 스펙트로그램의 주파수 축을 4장 청각이론에서 다룬 멜 단위로 변환한 스펙트로그램이다. 이는 단순 스펙트로그램보다 더욱 유사한 형태의 스펙트로그램에 가까운 정보량을 가진다.



[그림 14] 같은 오디오 파일에 대한 멜스펙트로그램과 스펙트로그램.

[그림 14]는 같은 샘플의 주파수 영역을 멜 스케일로 변환한 멜 스펙트로

그램과 단순 스펙트로그램을 비교한 것이다.

```python
import librosa
import librosa.display
import matplotlib.pyplot as plt
import numpy as np

%matplotlib inline

y, sr = librosa.load('./sample.wav', sr=16000)

# frame length (window) = 20ms
frame_length = 0.02

# -> 16000 * 0.02 = 320 : nfft
nfft = 320

# hop length (stride) = 10ms
hop_n = 160

# Mel-Spectrogram
Mel = librosa.feature.melspectrogram(y=y, n_mels=40, n_fft=nfft, hop_length=hop_n)

# Spectrogram
Spec = np.abs(librosa.stft(y))**2

plt.figure(figsize=(10, 4))
librosa.display.specshow(librosa.power_to_db(Mel, ref=np.max), y_axis='mel', sr=sr, hop_length=hop_n, x_axis='time')
plt.colorbar(format='%+2.0f dB')
plt.title('Mel-Spectrogram')
plt.savefig('mel_spectrogram.jpg')

plt.figure(figsize=(10, 4))
librosa.display.specshow(librosa.power_to_db(Spec, ref=np.max), y_axis='mel', sr=sr, hop_length=hop_n, x_axis='time')
plt.colorbar(format='%+2.0f dB')
plt.title('Spectrogram')
plt.savefig('spectrogram.jpg')
```

[코드 3] 멜스펙트로그램과 스벡트로그램 변환 코드

[코드3]은 그 과정을 보여주며, 해당 코드를 통해 멜스펙트로그램과 스펙트로그램을 얻을 수 있다. 이 과정은 5장에서 언급된 단시간 푸리에 변환(Short Time Fourier Transform)을 활용한다. 단시간 푸리에 변환을 통해 얻어진 스펙트로그램은 실수 값과 복소수 값을 가지고 있는데, 이는 np.abs()를 이용해 크기를 구하고, 제곱하여 에너지 크기 값을 취할 수 있다. 이 스펙트로그램, 멜스펙트로그램 텐서를 power_to_db() 함수를 사용하여 인간이 인지하는 청각 강도에 맞춘 값으로 조정한다. librosa.feature.melspecrogram은 내부적으로 mel scaling matrix를 곱하는 방식으로 연산을 실행한다. 사용자는

운영체제에 따라서 직접 매트릭스를 계산하는 방법을 구현할 수 있다.

제7장 Mel-Filterbank Coefficient

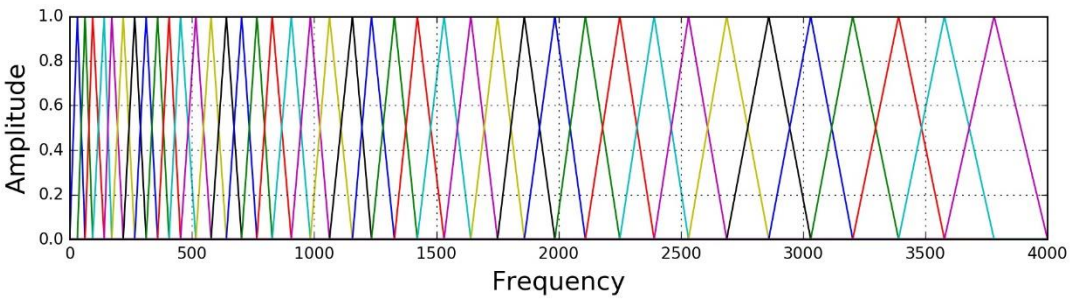
Index	Bark scale		Mel scale	
	Center freq (Hz)	BW (Hz)	Center freq (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6038	843
24	13500	3500	6954	969

[표 1] 멜-필터뱅크 표

[표 1]은 중심 주파수를 기준으로, 밑변을 대역폭으로하는 이등변 삼각형의 필터를 표로 나타낸다. 인간 청각 임계대역은 주파수가 커질 수록 높아진다.

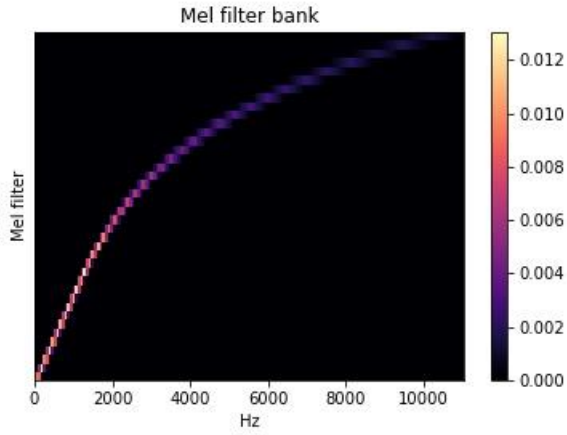
따라서 저주파에 민감하고, 고주파에 둔감한 인간의 청각 형태를 모방하여, 낮은 중심 주파수를 가진 필터뱅크는 작은 대역폭을, 높은 중심 주파수를 가진 필터뱅크는 높은 대역폭을 가지게 된다.

이러한 필터뱅크는 librosa.filters.mel 을 통해 구할 수 있다. 이때 활용되는 값으로 샘플링 주파수와 필터뱅크의 개수, 최소 주파수와 최대 주파수가 있다. 결과값의 형태는 (필터뱅크의 개수, 1 + n_fft/2) 를 갖는다. n_fft는 주파수 해상도를 나타내는 값이다.



[그림 15] 멜스케일로 나타낸 필터뱅크.

[그림 15]는 인간 청각 임계대역은 주파수가 커질 수록 높아지는 것을 하나의 그림에 나타낸다. 이러한 성질에 의해 높은 중심 주파수를 가진 필터뱅크일 수록 높은 대역폭을 가지게 된다.



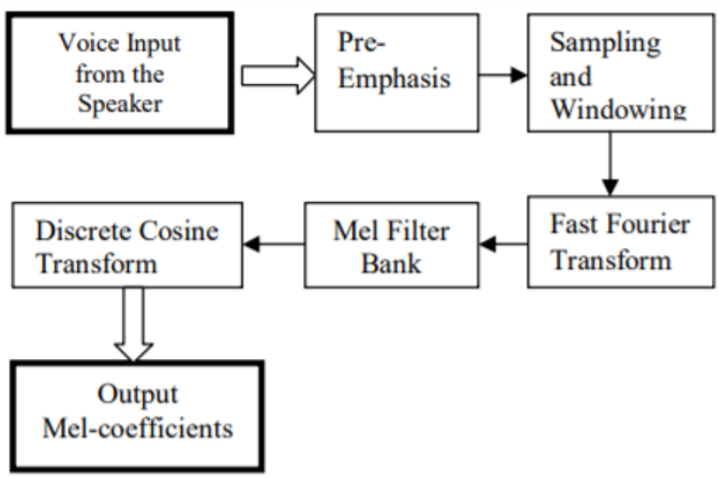
[그림 16] 40개의 필터뱅크 행렬

[그림 16]은 해당 과정을 거쳐 얻은 필터뱅크로 이를 각 시간축의 스펙트럼과 곱하여 멜필터뱅크 계수 값으로 변환할 수 있다. 이 동작은 내부적으로 행렬의 연산으로 구현되어있다.

제8장 MFCC

MFCC(Mel-Frequency Cepstral Coefficient)는 음성/음악 등 음성신호처리 분야에서 사용되는 특징 값으로 소리의 고유한 특징을 나타내는 수치이다. 켑스트랄이란 스펙트랄(Spectral)의 켑스트럴형 단어로 스펙트럼을 나타내는 형용사이다 [5].

제 8장의 [그림 15]에서 필터뱅크는 각각 중심 주파수로부터 일정 대역폭을 밑으로 하는 삼각형을 가진다. 이때, 양 옆의 필터뱅크와 중첩되는 부분이 존재한다. 이는 구해진 필터뱅크의 에너지 대표 값들이 큰 상관관계를 가지게 됨을 뜻한다. 따라서 이러한 상관성을 상쇄하는 작업이 필요한데, 이를 위해 DCT(Discrete Cosine Transform) 연산을 사용한다.



[그림 17] MFCC 추출 과정

[그림 17]은 MFCC를 추출하는 모식도를 보여준다.

```python
import librosa
import librosa.display
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline

# load
y, sr = librosa.load('./sample.wav', sr=16000)

# frame length (window) = 20ms
frame_length = 0.02

# -> 16000 * 0.02 = 320 : nfft
nfft = 320

# hop length (stride) = 10ms
hop_n = 160
_mfcc = librosa.feature.mfcc(y=y, sr=16000, n_mfcc=40, dct_type=2)

_mfcc.shape
```

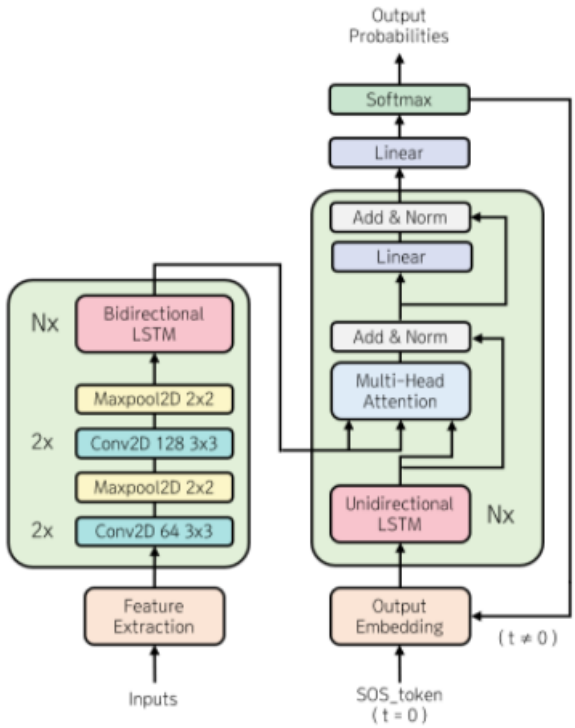
```
(40, 331)
```

[코드 4] MFCC 추출 코드

[코드 4]는 불러온 오디오 시그널을 이용해 MFCC 값을 계산하고, 그 텐서의 형태를 확인한다. (40, 331)의 결과 값에서 확인할 수 있듯, 40개의 MFCC 계수를 취했으며, 시간축 역시 331로 축소된 것을 확인할 수 있다. 끝으로 연산효율과 정확도 상승을 위해 40개 중 24개의 계수 값들만을 골라 모델을 학습하는 방법이 추천된다.

제9장 결론

KoSpeech 모델은 [그림 18]과 같은 모델 구조를 가진다[6]. Inputs로 표시된 곳에서 입력으로 주어지는 것들은 MFCC, Log mel spectrogram, Log spectrogram, Filter Bank 등의 음성신호로부터 추출할 수 있는 특징 벡터들이다.



[그림 18] KoSpeech 모델 구조

Feature	CER(%)
# MFCC (40)	17.31
# Log mel spectrogram (161)	15.79
# Log spectrogram (161)	10.72
# Filter Bank (80)	**10.31**

[표 2] 음성 특징 벡터에 따른 CER

[표 2]는 음성으로부터 추출한 특성 벡터들을 KoSpeech 모델의 입력으로 사용했을 때 결과를 보여준다. 모든 특징 벡터들은 정보량과 성능지표에 따라 적절한 n_mels를 선택할 때, 모델의 성능을 최대로 끌어올릴 수 있다. 현재 표에서는 필터뱅크의 크기를 80으로 지정하고, 필터링은 거치지 않았다. 그러나 음성인식이 아닌 높은 주파수의 음역대를 다루는 악기의 품질을 검증하는 모델이라면 주파수 범위를 높은 음역대로 조정하고, 필터뱅크의 크기를 늘려 더욱 세밀한 정보량을 포함할 수 있도록 수정하여야 할 것이다.

연산량의 관점에서 볼 때, MFCC는 시퀀스 열 축소 역시 발생하여 많은 정보 함축이 발생한다. 정보량 손실은 존재하지만 연산 속도는 그만큼 빨라질 수 있다. 위와는 반대로 간단한 소리 종류 구분을 목적으로 하는 모델은 MFCC를 사용하여 연산 속도를 향상시킬 수 있다.

끝으로, 최근 형태소를 벡터로 대응하는 word2vec 으로부터 파형을 벡터로 대응시키는 wav2vec 방식이 연구되고 있으며, 높은 정확도와 빠른 추론 속도를 보여주고 있다. 해당 논문의 음성신호들의 특징과 생성된 벡터와의 관계를 분석하여, 대규모의 학습을 진행하지 않아도 벡터를 생성하는 방안을 연구하는 것이 필요할 것이다.

참고문헌

[1] ISSU MONITOR 제 126호 음성 AI 시장의 동향과 비즈니스 기회, 삼정 KPMG 경제연구원, 2020-04-13

[2] 소리 물리학의 기본 개념, 대한음성언어의학회지 제22권 제2호, 성균관대학교 의과대학 강북삼성병원 이비인후과학교실, 진성민, 2011

[3] 인공지능과 음성신호처리, 광운대학교 전자공학과 박호종, 2019

[4] 디지털 신호처리, 광운대학교 전자공학과 박호종, 2009

[5] Speech Recognition - Feature Extraction MFCC & PLP, Jonathan Hui, 2019

[6] KoSpeech: Open-Source Toolkit for End-to-End Korean Speech Recognition, arXiv, Soohwan Kim, Seyoung Bae, Cheolhwang Won