Audio
for
Deep Learning

Audio

- 인공지능 스피커의 대중화로 음성 분리, 인식분야 연구 활발
- 음성 합성 분야 연구 활발

  - 글자를 입력하면 유인나, 손석희 목소리로 읽어주는 서비스(Naver)

  - 2시간 음성 데이터가 있으면 목소리 생성 가능

[책 읽어 주는 유인나]

[영어 하는 김정은]

- Tacotron2+ Waveglow(Google, Nvidia, Baidu)
- 분간하기 힘든 사람 목소리를 실시간 생성(2초 이내)
- 국내 기업들은(네이버, 카카오) 구글 기술을 한국어에 접목시켜 사업화 진행중
- 그러나 기술적으론 구글, Nvidia의 기술을 겨우겨우 구현하여 한글에 적용하는 수준

Audio task

- Speech Separation(음성분리)
  - 간섭하는 배경음(background interference)에서 target speech 를 분리해내는 작업. Speech Segregation 이라고도 한다.
- Speech Recognition(음성인식)
  - Speech의 내용을 인식하여 문자 데이터로 전환하는 작업. STT(Speech-To-Text)라고도 한다.
- Speech Synthesis(음성합성)
  - Speech의 음파를 자동으로 만드는 작업. TTS(Text-To-Speech)라고도 한다.

** 전반적인 '소리'는 Sound, Audio 또는 Source(음원), 그 중 '음성'이면 Speech

Contents

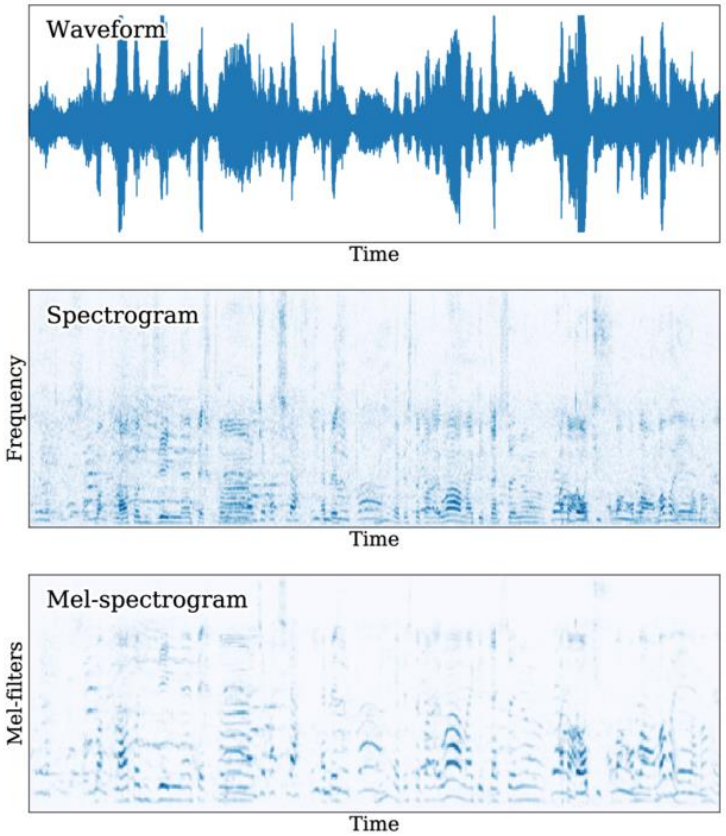
- **Neural networks for Speech Processing**
  - **Speech Emotion Recognition(음성 감정 인식)**
  - **Language Identification(언어 식별)**
  - **Speaker Recognition(화자 인식)**
  - **Disease Detection(질병 감지)**
  - **Speaker Verification(화자 인증)**

  - **Speech Recognition(STT)**
  - **Voice Conversion(목소리 변조)**
    - **source speaker의 입력 음성을 말의 의미 (linguistic content) 변화 없이 마치 target speaker가 말하는 것처럼 만드는 task**

  - **Speech Separation: 음성 분리**
    - **여러 화자 음성중 특정 화자 음성 분리**
  - **Voice Activity Detection(음성 구간 검출)**
    - **잡음 신호로부터 음성을 검출**

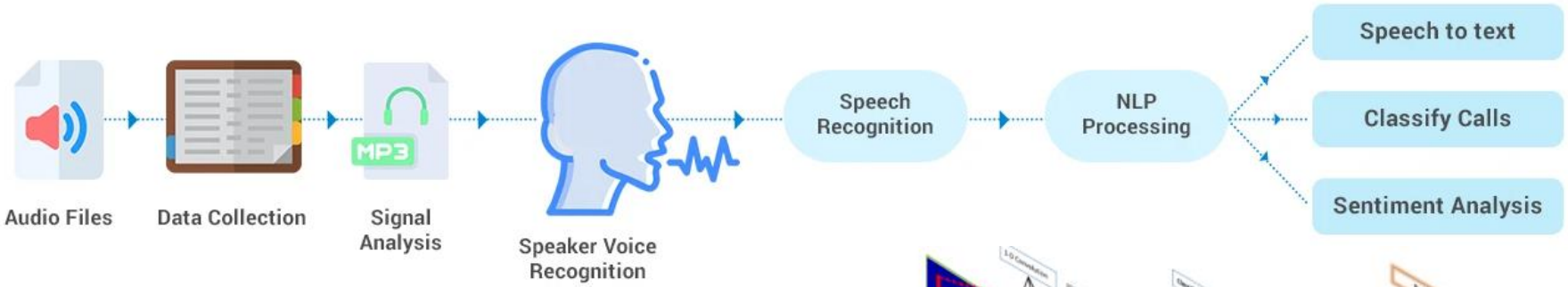
Contents

- **Neural networks for Speech Processing**
  - Speech Emotion Recognition(음성 감정 인식)
  - Language Identification(언어 식별)
  - Speaker Recognition(화자 인식)
  - Disease Detection(질병 감지)
  - Speaker Verification(화자 인증)

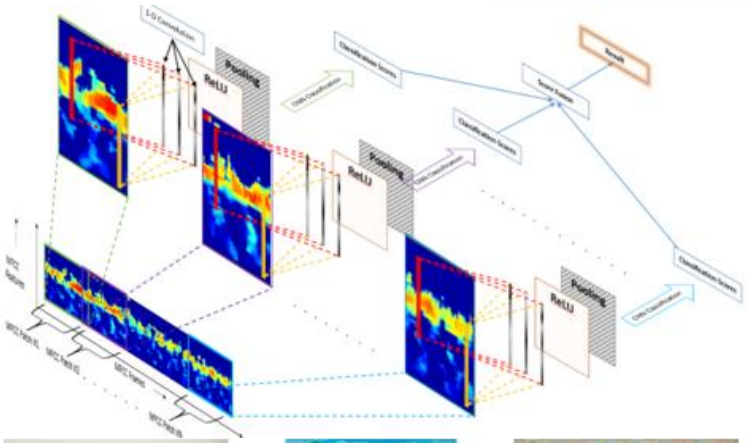


  - Speech Recognition(STT)
  - Voice Conversion(목소리 변조)
    - source speaker의 입력 음성을 말의 의미 (linguistic content) 변화 없이 마치 target speaker가 말하는 것처럼 만드는 task
  - Speech Separation: 음성 분리
    - 여러 화자 음성중 특정 화자 음성 분리
  - Voice Activity Detection(음성 구간 검출)
    - 잡음 신호로부터 음성을 검출

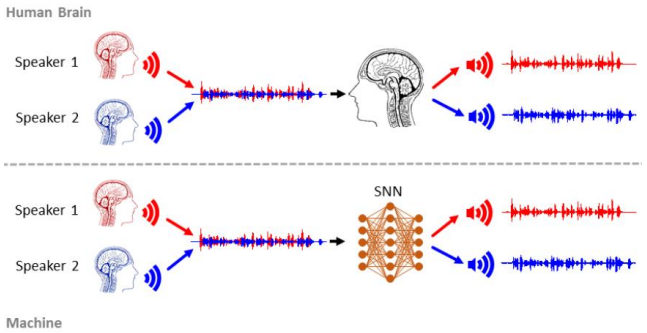
https://medium.com/@stanleydukor/neural-style-transfer-part-1-e82c2e774443
https://xuchenglin28.github.io/files/iccbc2019/index.html
https://iprobe.cse.msu.edu/project_detail.php?id=11&?title=Speaker%20recognition%20from%20degraded%20audio%20sample



- Speech Emotion Recognition(음성 감정 인식)
- Language Identification(언어 식별)
- Speaker Recognition(화자 인식)
- Disease Detection(질병 감지)
- Speaker Verification(화자 인증)



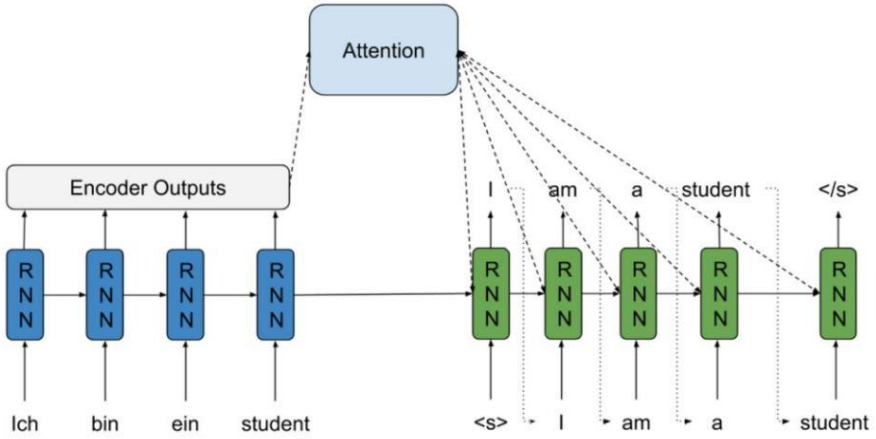
- Speech Recognition(STT)
- Voice Conversion(목소리 변조)
  - source speaker의 입력 음성을 말의 의미 (linguistic content) 만을 보존하여 마치 target speaker가 말하는 것처럼 만드는 task



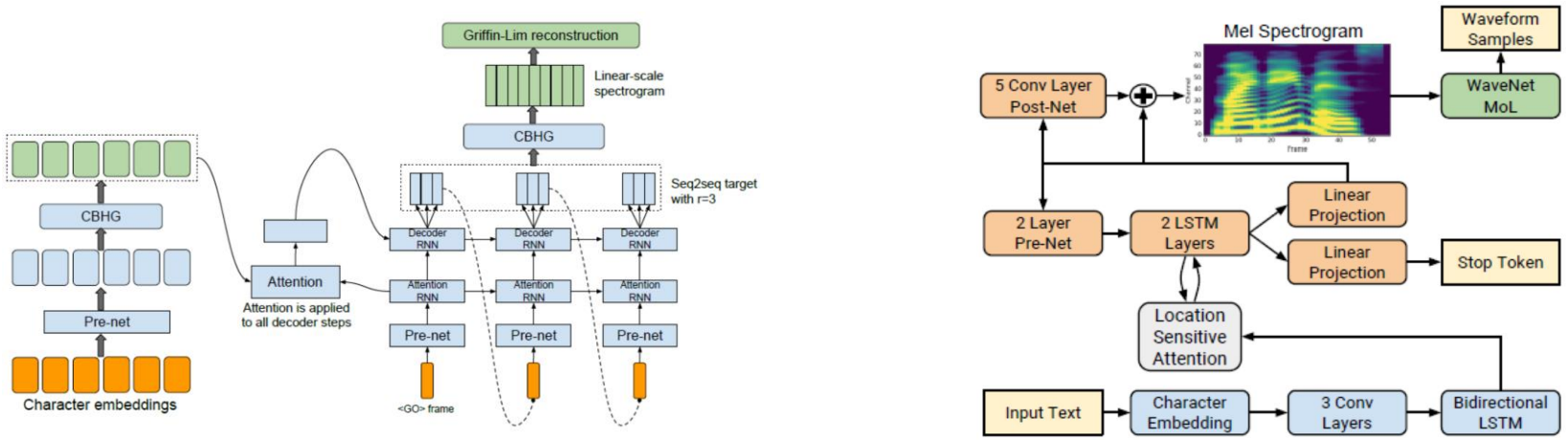
- Speech Separation: 음성 분리
  - 여러 화자 음성중 특정 화자 음성 분리
- Voice Activity Detection(음성 구간 검출)
  - 잡음 신호로부터 음성을 검출

https://medium.com/@stanleydukor/neural-style-transfer-part-1-e82c2e774443
https://xuchenglin28.github.io/files/iccbc2019/index.html
https://iprobe.cse.msu.edu/project_detail.php?id=11&?title=Speaker%20recognition%20from%20degraded%20audio%20sample

Contents



NLP Sequence2Sequence

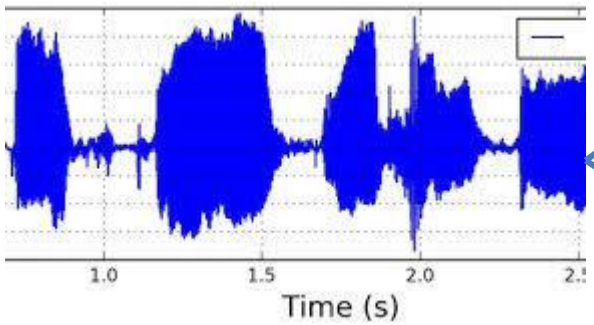


TTS Sequence2Sequence

https://chldkato.tistory.com/143?category=839554
https://www.osplabs.com/ai-speech-recognition/
https://www.researchgate.net/figure/Waveform-spectrogram-and-mel-spectrogram-of-a-10-s-speech-segment-obtained-from-Google_fig1_333834541

Contents

- **언어모델**
  - 문장의 흐름상 올바른 정도를 나타내는 확률 모델
  - P_LM(점심 뭐 먹을까?) > P_LM(점심 뭐 먹을자!)



점심 뭐 먹을자!

0.0013%

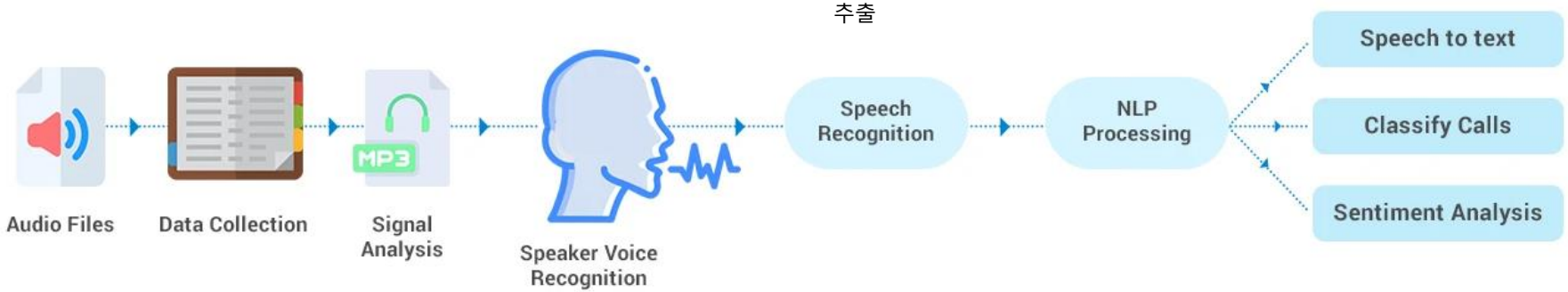
점심 뭐 먹을까?

0.043%

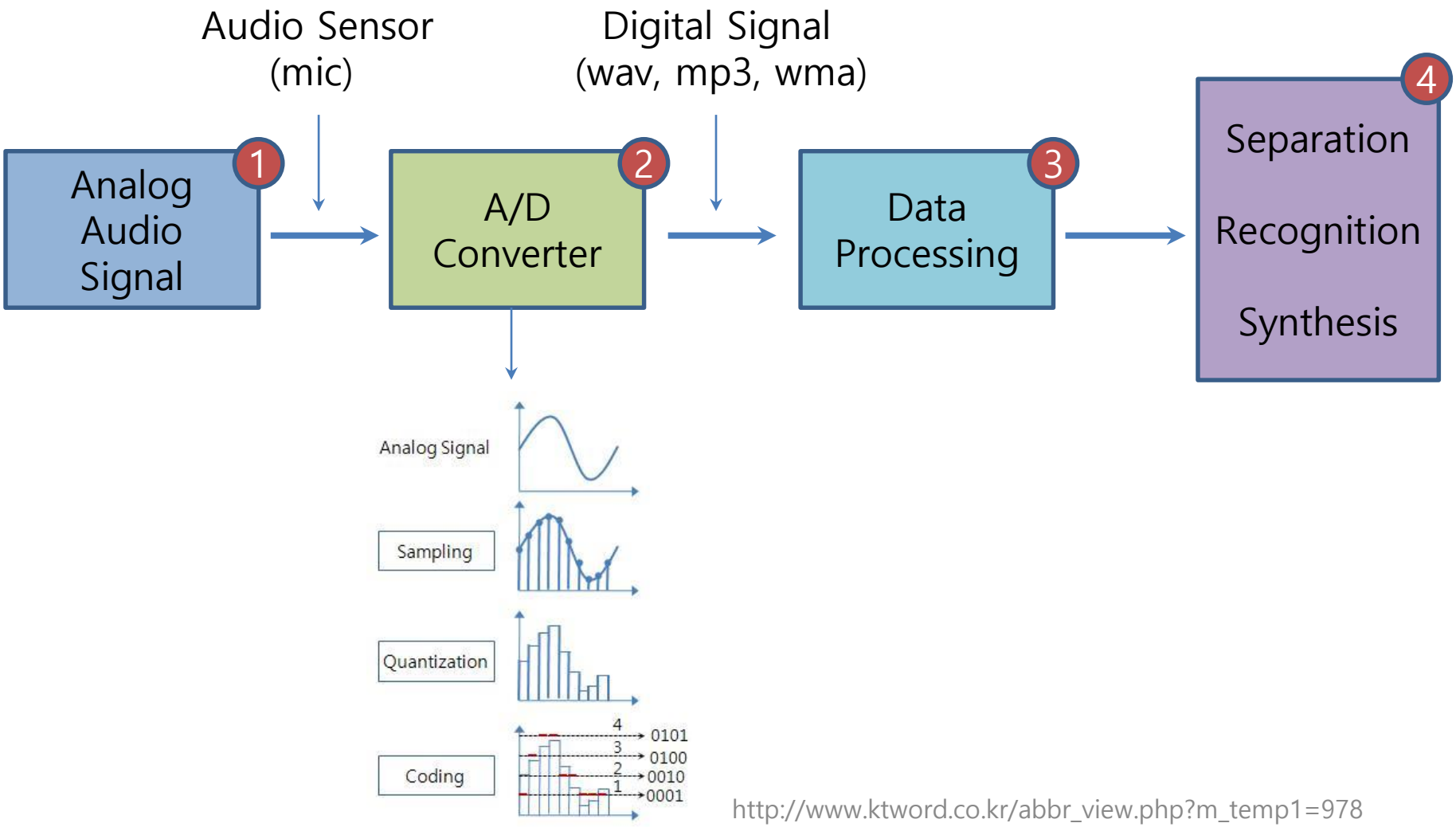
$$\hat{Y} = \underset{Y}{\operatorname{argmax}} P(Y|X)$$

$$\hat{Y} = \underset{Y}{\operatorname{argmax}} \frac{P(X|Y)P(Y)}{P(X)}$$

X: 입력 음성 신호
Y: 가장 그럴듯한 음소/단어 시퀀스
P(Y): 언어모델
P(X|Y): 음향모델
- 음소/단어 시퀀스와 입력 음성 신호가 어느 정도 관계를 맺고 있는지 추출



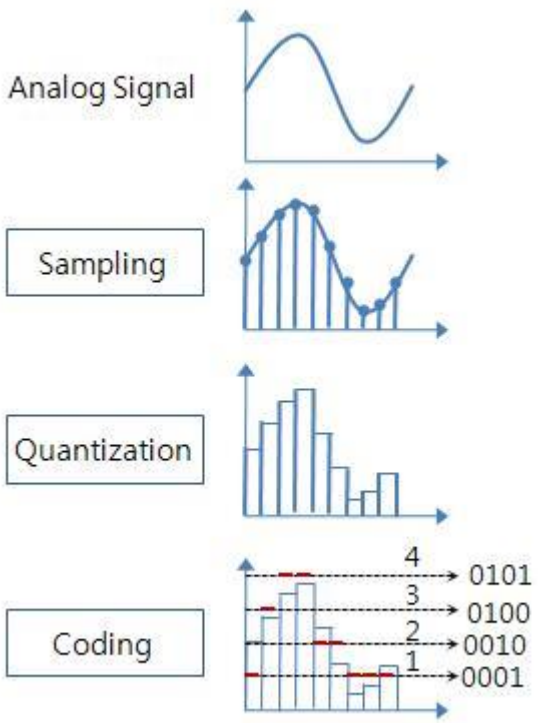
Audio Data



http://www.ktword.co.kr/abbr_view.php?m_temp1=978

A/D Converter

- **연속적인 아날로그 신호**를 **표본화(Sampling), 양자화(Quantizing), 부호화(Encoding)**을 거쳐 **이진 디지털 신호(Binary Digital Signal)**로 변화시키는 과정



샘플링 단계에서 초당 샘플링 횟수를 정하는데, 이를 Sampling rate 또는 Sampling frequency라고 한다.

나이키스트-섀넌 표본화 정리에 따르면 신호의 완전한 재구성은 표본화 주파수가 표본화된 신호의 최대 주파수의 두 배보다 더 클 때 가능하다고 한다.

인간의 최대 가청 주파수가 20KHz 이기 때문에 이론상 40KHz의 Sampling frequency 를 기준으로 잡는다. 실제로는 추가적인 이유로 오디오 레코드에선 44.1KHz 를 사용한다.

우리가 접하는 오디오 파일들은 다 이러한 과정을 거친 오디오 데이터!

http://www.ktword.co.kr/abbr_view.php?m_temp1=978

소리(Sound)

- 소리의 세기
  - 소리의 세기는 물체가 진동하는 폭(진폭, Amplitude)에 의해 정해진다. 진폭이 크면 소리가 세지고, 진폭이 작으면 소리가 약해진다. 주파수(frequency) 와는 상관이 없고, 세기 단위로는 dB(데시벨)을 사용한다

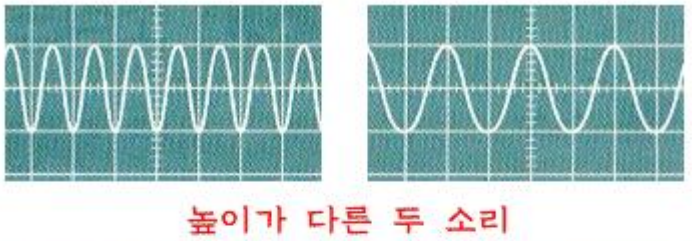
- 소리의 높이(고음/저음)
  - 소리의 높낮이는 음원의 주파수에 의해 정해지며, 주파수가 높으면 높은 소리가, 낮으면 낮은 소리가 난다. 단위는 Hz.

- 소리의 맵시(음색)
  - 소리의 맵시는 파형(파동의 모양)에 따라 구분된다.

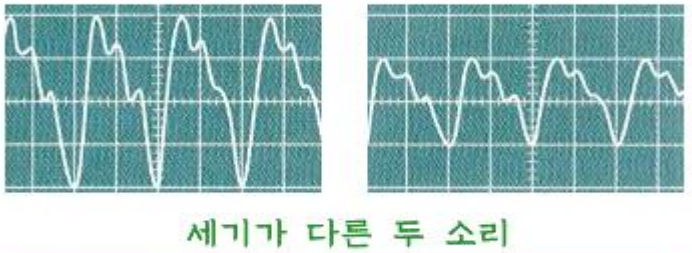
물리량 (physical quantity)	심리량(subjective quantity)		
	소리의 크기 (loudness)	소리의 높이 (pitch)	음색 (timbre)
음압(pressure)	◎◎◎	◎	◎
주파수(frequency)	◎	◎◎◎	◎◎
스펙트럼(spectrum)	◎	◎	◎◎◎
파형(waveform)	◎	◎	◎◎
지속시간(duration)	◎	◎	◎

(◎◎◎ : 관계 높음, ◎◎ : 관계 중간, ◎ : 관계 낮음)

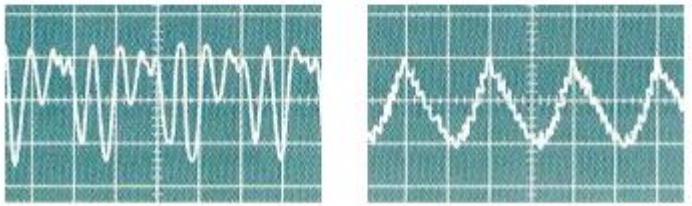
https://blog.naver.com/hearmall/220957997531



높이가 다른 두 소리



세기가 다른 두 소리



맵시가 다른 두 소리

http://lovescience.pe.kr/ms/chapter12/12_2_study2.htm

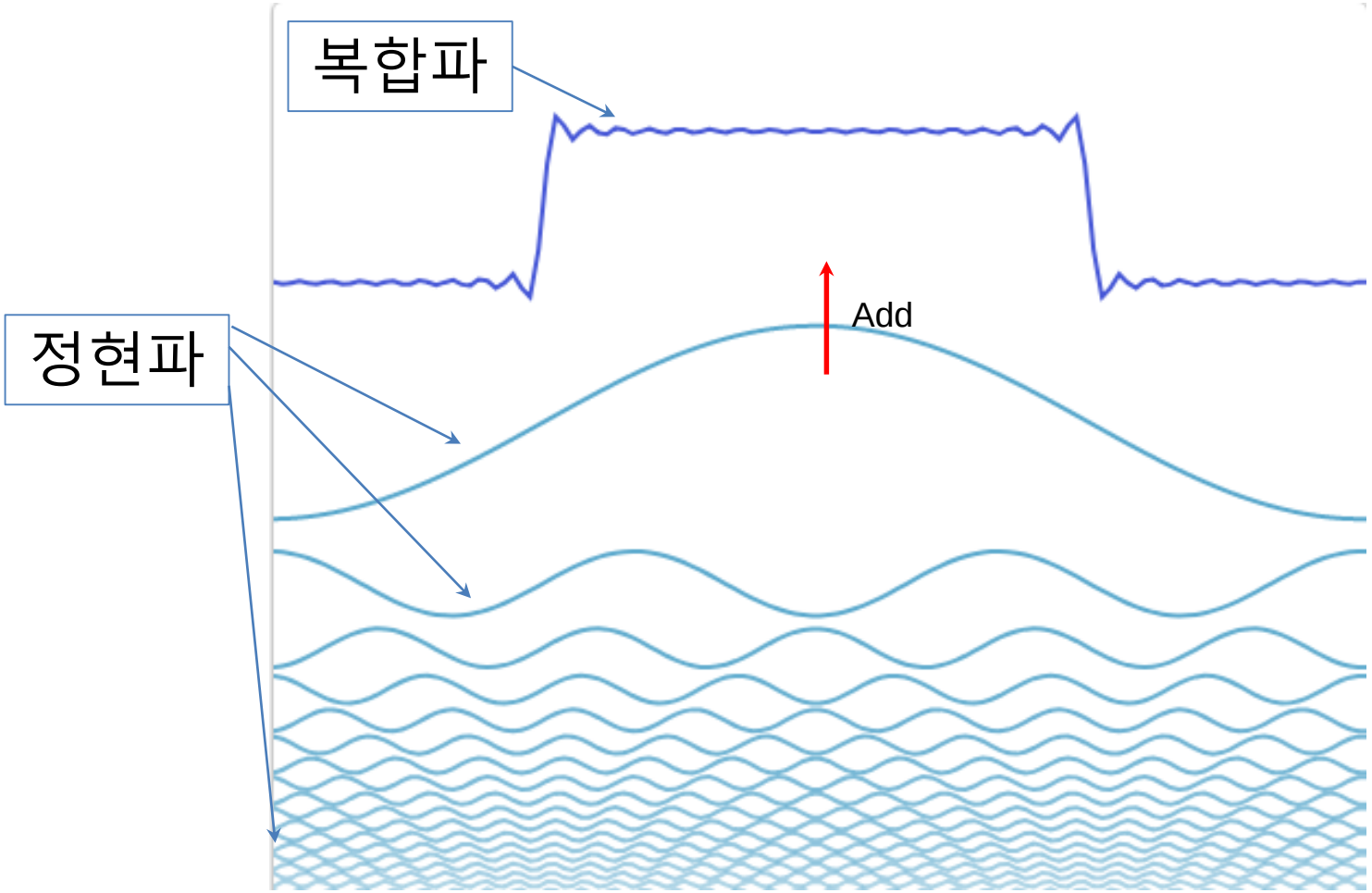
Wave

- 소리는 파형(wave) 신호이다. 어떤 소리냐에 따라 파형의 형태는 매우 다양해진다.

- 삼각함수로 된 파형인 사인파(sine wave)와 코사인파(cosine wave)가 있는데 주기적(Periodicity)인 특징이 있다. 이 주기 신호들의 파형을 정현파(Sinusoidal wave) 라고 한다.

- 복합파(Complex wave)는 복수의 서로 다른 정현파들의 합으로 이루어진 파형이다. 우리가 흔히 듣는 소음, 성대의 발성(음성) 등은 다 복합파다. 자연에서 듣는 대부분의 소리는 복합파라고 생각해도 무방하다.

Wave



Fourier Series

- 복합파(Complex wave)가 복수의 서로 다른 정현파들의 합으로 이루어진 파형이라는 것은 다시 말해서 복합파인 파형을 정현파들로 다시 표현할 수 있다는 얘기이다.

- 푸리에 급수(Fourier Series)는 주기 함수 sin과 cos로 표현한 무한급수인데, 이를 이용해서 주기가 있는 임의의 주기 함수를 sine wave와 cosine wave의 합으로 표현 할 수 있다.

$$f(t) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos n\omega t + b_n \sin n\omega t \right), \; -\infty < t < \infty$$

- 계수의 공식은 아래와 같다. 소리에서 계수는 진폭(Amplitude)을 의미한다.

$$a_0 = \frac{1}{T} \int_0^T f(t)\,dt$$

$$a_n = \frac{2}{T} \int_0^T f(t) \cos n\omega t \, dt$$

$$b_n = \frac{2}{T} \int_0^T f(t) \sin n\omega t \, dt$$

https://kkhipp.tistory.com/3

Fourier Series

- 오일러 공식을 이용하면 cos, sin의 합을 복소지수로 표현이 가능해진다.

$$e^{i\pi} = \cos\pi + i\sin\pi$$

- 오일러 공식을 이용해 다시 푸리에 급수를 정리하면 아래와 같이 나타난다.

$$f(t) = \sum_{n=-\infty}^{\infty} C_n e^{inwt} \qquad C_n = \frac{1}{T}\int_0^T f(t)e^{-inwt}dt$$

- 계수를 구하는 식은 두 벡터의 직교성(orthogonal) 관계를 이용해서 전개한 것이다.

- 또한, 급수의 주기 T를 무한으로 표현하면 비주기 함수에 대해서도 적용이 가능해진다. 이를 푸리에 적분 혹은 푸리에 변환(Fourier Transform) 이라고 한다. 푸리에 적분과 푸리에 변환은 같은 수식의 서로 다른 관점일 뿐이다.

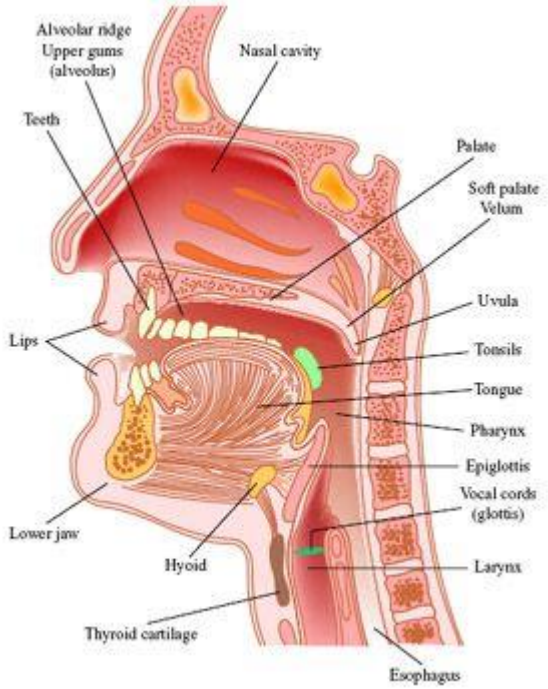
https://kkhipp.tistory.com/3

음성(Speech)

- Speech는 사람의 발음 기관을 통해 나타나는 목소리(voice) 중에서 의사 소통(communication)의 목소리이다.

- Voice는 크게 성도(Voice tract)과 성문(Glottis)를 통해 생성이 된다.
  - 성문에서 성대가 떨리기(열림과 닫힘) 시작하면서 공기가 끊어지게 되는데 이 성대를 통해 소리가 발생한다. 성대의 열고 닫히는 최저 속도를 기본 주파수(Fundamental Frequency)라고 하고 또는 가장 낮은 주파수라고한다. 성대를 통해서 발생하는 소리는 기본주파수를 갖는 톱니파형으로 나타나는데 이 톱니파형은 여러 주파수 성분을 갖는 정현파의 합(즉, 복합파)으로 이루어져있다.

  - 이렇게 성문에서 나온 소리가 성도를 지나가게 되는데 성도를 지나가면서 이 소리가 공명현상이 일어나게 된다. 즉, 성도는 소리에 대한 공명 공간인데 이 성도가 하는 역할이 일종의 필터링을 하는 것이다. 성대에서 발생한 소리의 일부는 강하게, 또 일부는 약하게 필터링을 한다고 생각하면 된다...



https://soryro.tistory.com/71

Data Processing

Audio 관련 작업에 대표적으로 많이 사용되는 데이터 처리 방법들

- Spectrogram
  - STFT (Short time fourier transform)
- Cepstrum
- Mel Frequency Analysis
- Mel Frequency Cepstral Coefficients(MFCC)

Spectrogram

- 파형(Waveform)에서는 시간 축(time-domain)에 따른 진폭(Amplitude)의 변화를 보여준다.
- 스펙트럼(Spectrum)에서는 주파수 축(frequency-domain)에 따른 진폭(Amplitude) 의 변화를 보여준다.
- 음성 데이터에서 시간과 주파수, 진폭은 전부 의미있는 특징이기 때문에 이들을 전체적으로 보는 것은 중요하다.
- 스펙트로그램(Spectrogram)은 time-domain 과 frequency-domain의 변화에 따른 진폭의 차이를 색상 차이를 통해 시각화 시켜준다.
- Spectrogram 을 만드는 과정에서 **Fourier Transform** 을 사용한다.

Fourier Transform

- 푸리에 변환은 푸리에 적분을 시간 함수를 주파수 함수로 바꾸어서 생각했을 때의 다른 이름이다. 따라서 주파수 영역의 함수와 시간 영역의 함수를 잇는 수학적 연산 또는 공식 모두를 의미한다. 결국 푸리에 변환의 경우엔 복합파를 구성하는 정현파들을 주파수 함수로 표현한 것. 또한 역함수가 정의되어지는데 이를 푸리에 역변환이라고 하고 이를 통해 원래 함수로 다시 복원한다.

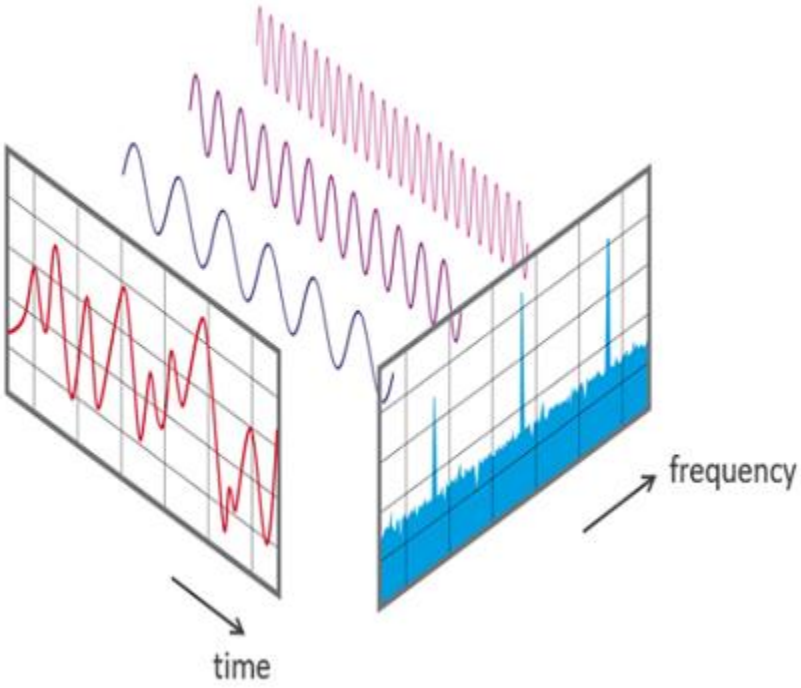
$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \qquad f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega$$

- 이산 푸리에 변환(Discrete Fourier Transform, DFT)는 길이가 N 인 이산시간 시계열 데이터에 대한 변환 공식이다. 기존의 푸리에 변환에서 이산시간시계열이 주기 N으로 계속 반복된다고 가정해서 진행된다.

- 고속 푸리에 변환(Fast Fourier Transform, FFT) 는 DFT의 속도 문제를 개선한 방식인데, 길이가 2^n 인 시계열 데이터에만 적용가능지만 DFT의 $O(n^2)$ 의 계산량에 비해 FFT 는 $O(n \log_2 n)$ 계산량으로 DFT 를 수행 할 수 있다.

https://ghebook.blogspot.com/2012/08/fourier-transform.html

Fourier Transform



FFT:

time-domain signal
→ Spectrum (frequency-domain signal)
(Digital Signal 이라서 주파수 bin 수 만큼의 size 를 갖는 벡터 생성)

https://youtu.be/EyVJtPg_Vr0

(링크를 누르시면 영상을 보실 수 있습니다.)

Short Time Fourier Transform (STFT)

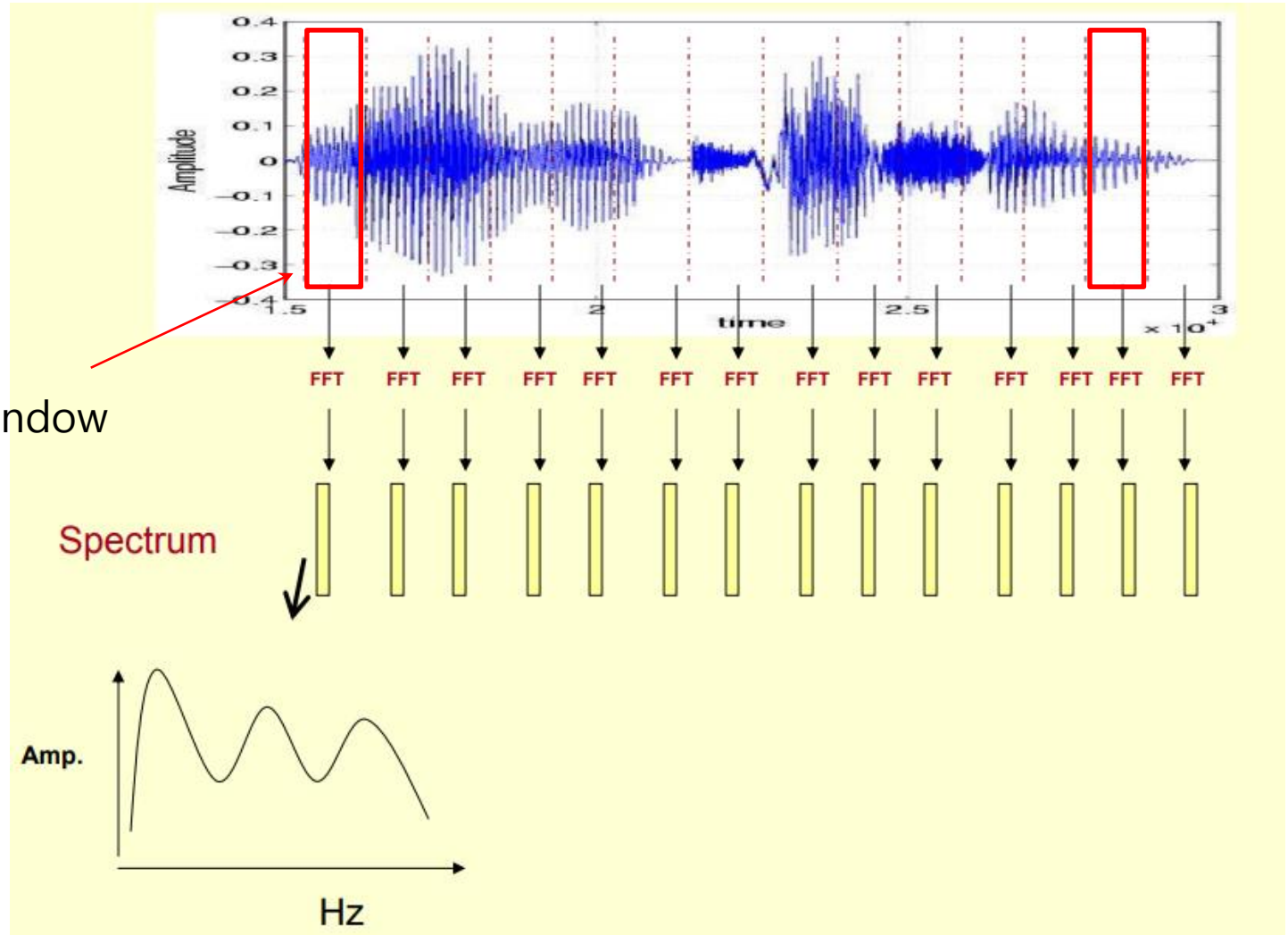
- 기존의 FFT의 문제점은 time-domain 의 함수를 frequency-domain 의 함수로 변환하는 과정에서 time-domain의 정보가 날라간다는 점이다.

- 주파수 정보와 시간 정보를 동시에 보기 위해 time-domain signal 을 일정한 크기(Window)로 잘라 각 구간마다 FFT를 한다.

- 그 결과 [주파수 bin 수 X 프레임 수(구간 수)] 크기의 복소수 (Complex-valued) 매트릭스 생성.

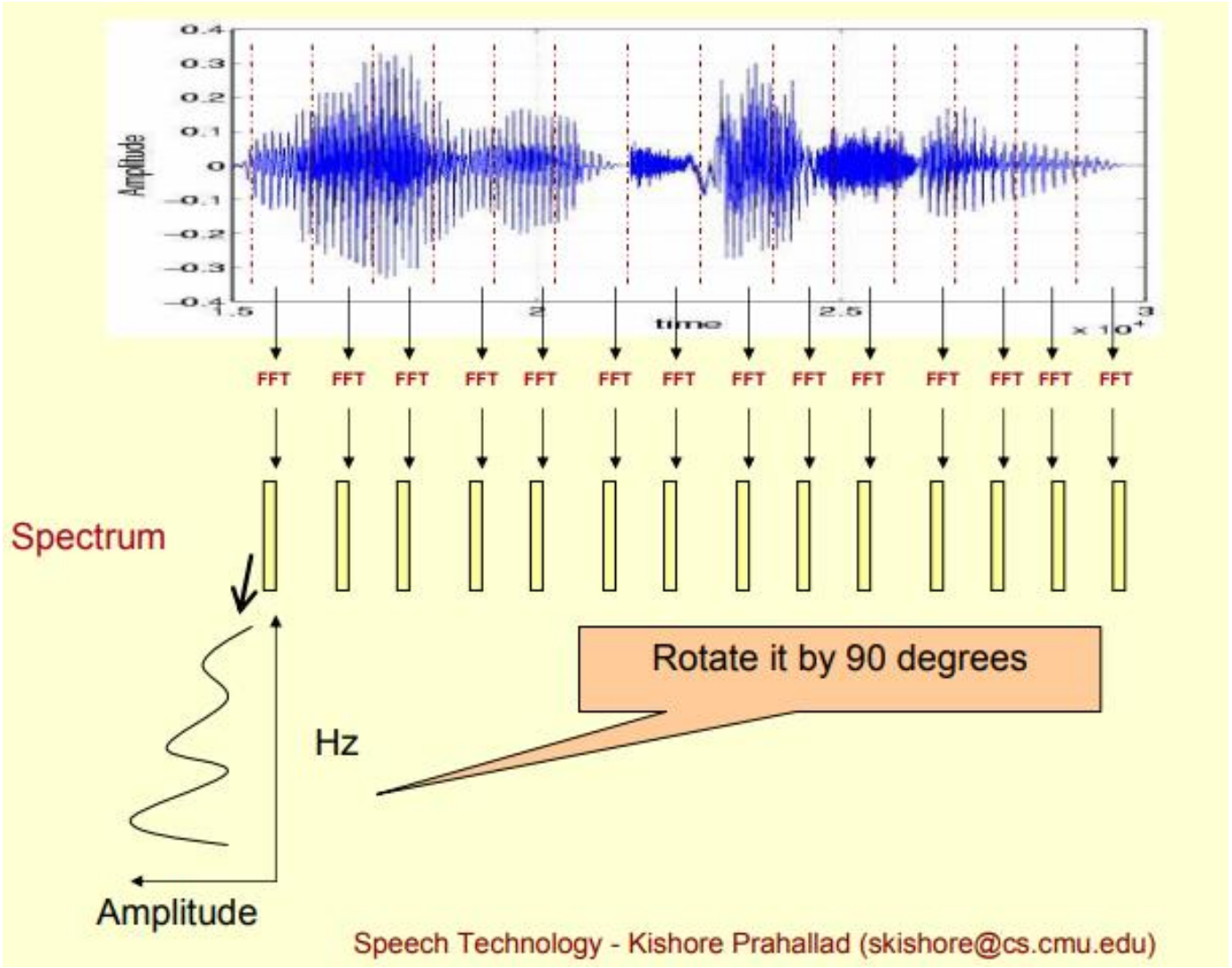
https://docs.google.com/presentation/d/1PlMQ1fEQrAlVQhP3mf3iMAy_JnLDtj7BPjlNGzffWTQ/edit#slide=id.g4f17be1e42_0_4

Short Time Fourier Transform (STFT)



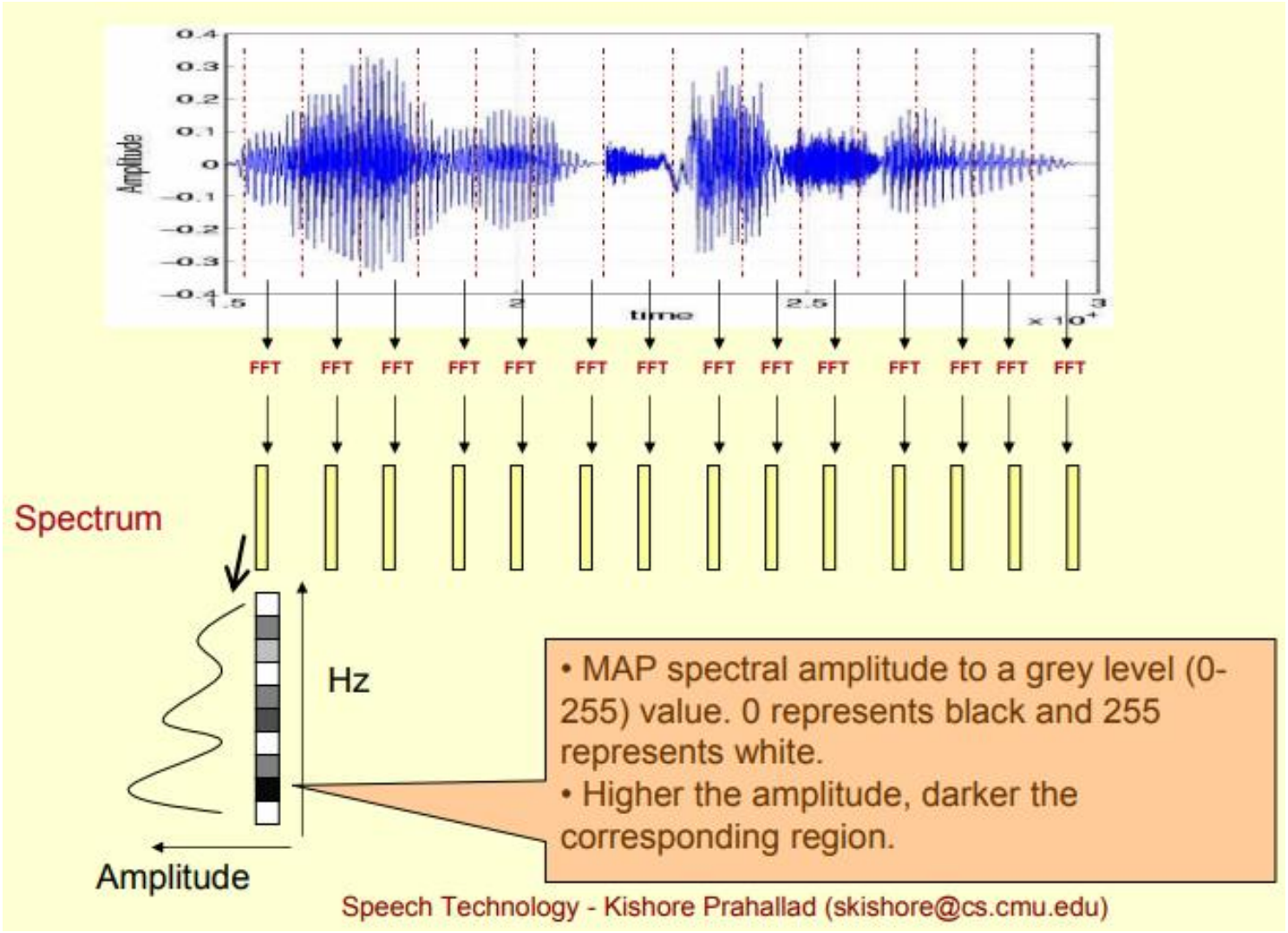
www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Short Time Fourier Transform (STFT)



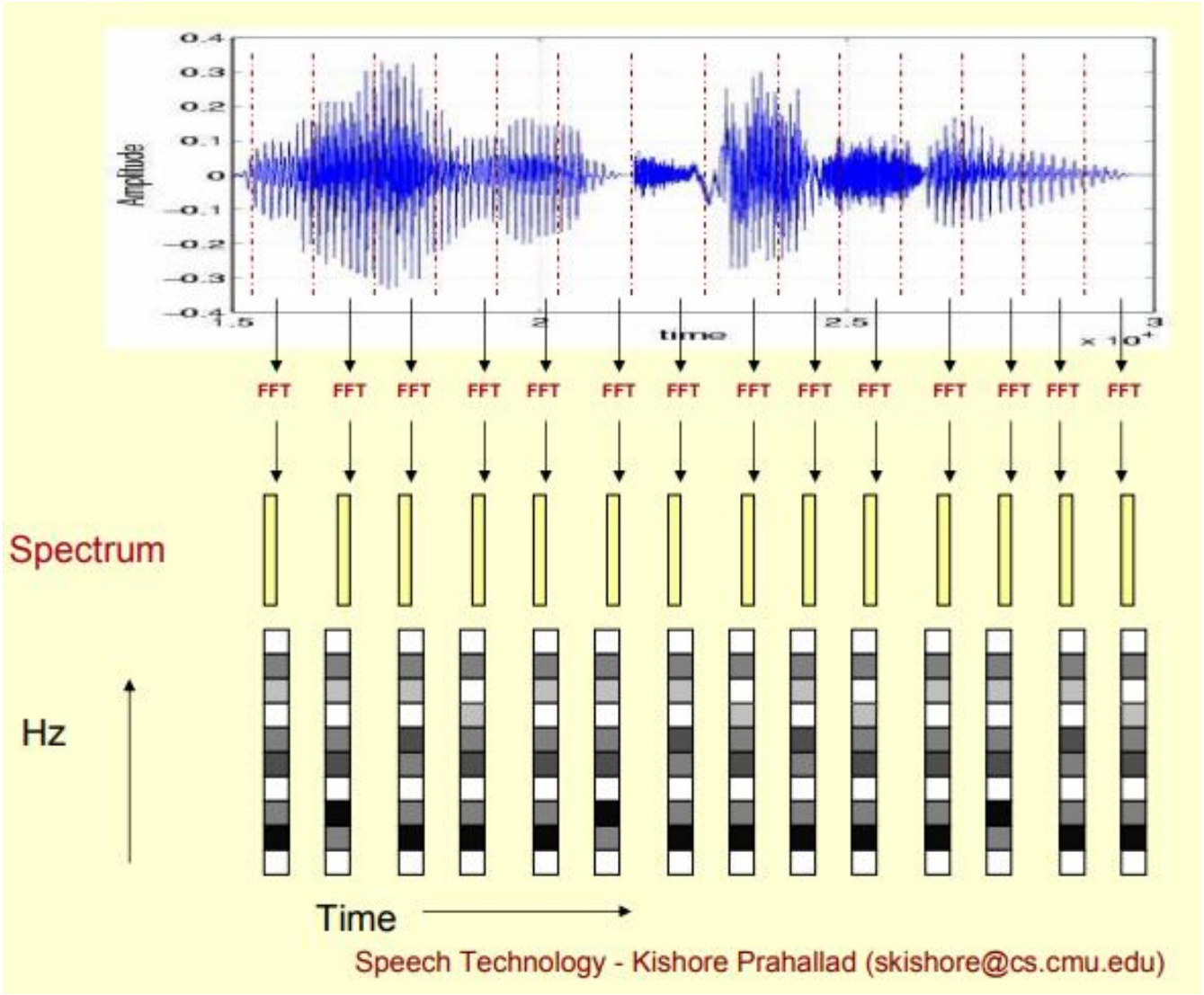
www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Short Time Fourier Transform (STFT)



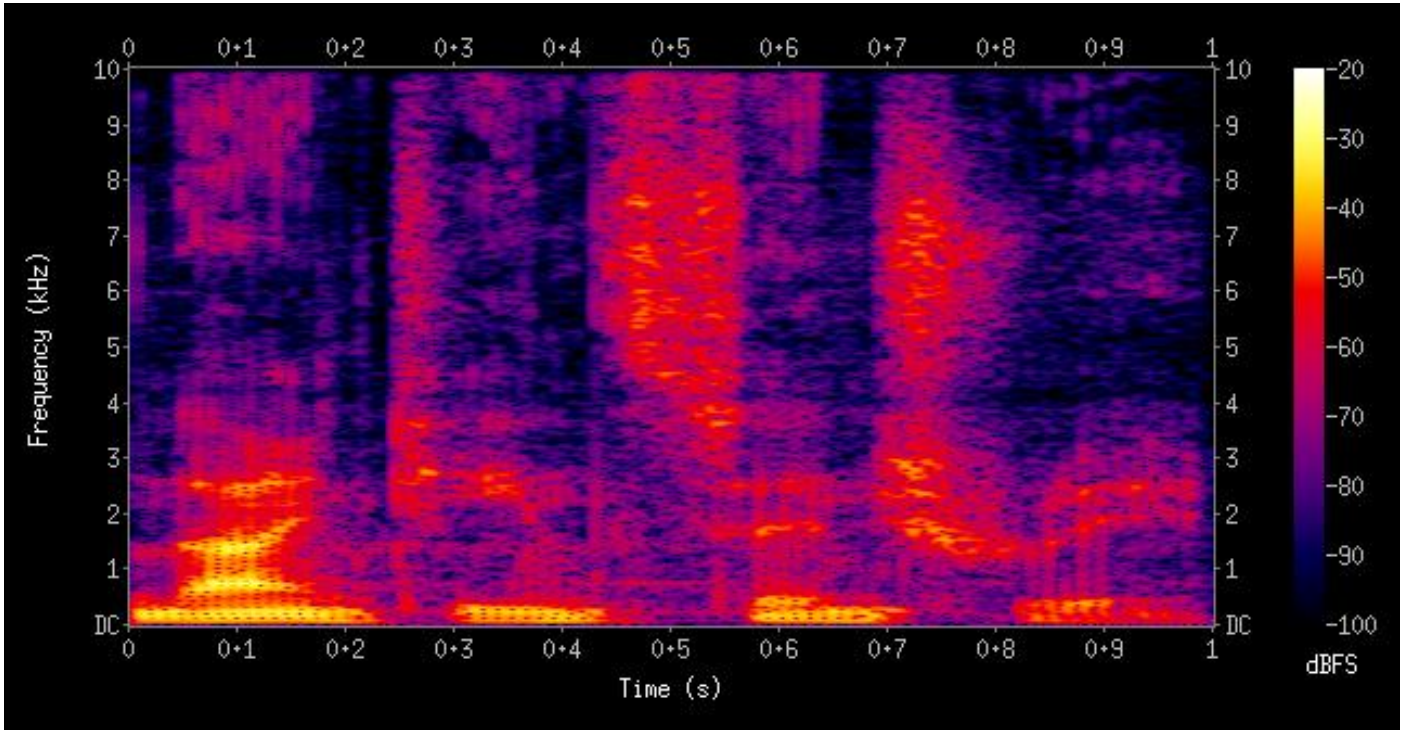
www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Short Time Fourier Transform (STFT)



www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Short Time Fourier Transform (STFT)



https://en.wikipedia.org/wiki/Spectrogram

Short Time Fourier Transform (STFT)

- 특징
  - Audio Signal의 time-domain 변화에 따른 frequency-domain의 정보를 시각적으로 표현 할 수 있다.
  - time-domain 변화에 따른 frequency-domain 의 정보가 중요한 musical sound 나 rhythm 분석에 활용될 수 있다.
  - 파형(Wave) signal로 재복원하기 쉽다.
  - 복소수로 값이 구성된다.

    $Y = a + bi$

Cepstrum

- 캡스트럼(Cepstrum)은 스펙트럼(Spectrum)에 log를 취하고 그 다음 푸리에 역변환을 취한 결과이다.
  - 'Ceps' 는 'Spec'을 뒤집은 단어이다. 뿐만 아니라 Spectrum에서 언급되는 용어들은 Cepstrum에서는 전부 뒤집어서 단어를 정의한다. Ex)'magn'itude -> 'ngam'itude...

$$\text{power cepstrum of signal} = \left|\mathcal{F}^{-1}\left\{\log\left(\left|\mathcal{F}\{f(t)\}\right|^2\right)\right\}\right|^2.$$

- Cepstrum 은 여러 개의 단순파가 합쳐진 복합파의 경우 그 구성을 알기가 힘든 경우에 좀 더 알기 쉽게 분석해준다.
  - 그래서 연산 과정에서 Log를 쓰는 이유가 곱의 형태를 덧셈으로 바꿔줄 수 있기 때문이다. 이 부분은 뒤에서 더 자세히...

[paper] Clinical Usefulness of Cepstral Analysis in Dysphonia Evaluation

Cepstrum

- 캡스트럼의 log가 과연 어떤 의미인지 이해하기 좋은 예시는 Speech signal 이다.

- 앞단에서 얘기했듯이 Speech signal 은 Voice tract(filter)과 Glottal signal(source)의 곱으로 표현할 수 있다. 아래는 Spectrum 식을 표현한 것이다.

$$S(f) = H(f) \times G(f)$$

- Magnitude Spectrum으로 표현하면 아래와 같다.

$$|S(f)| = |H(f)| \times |G(f)|$$

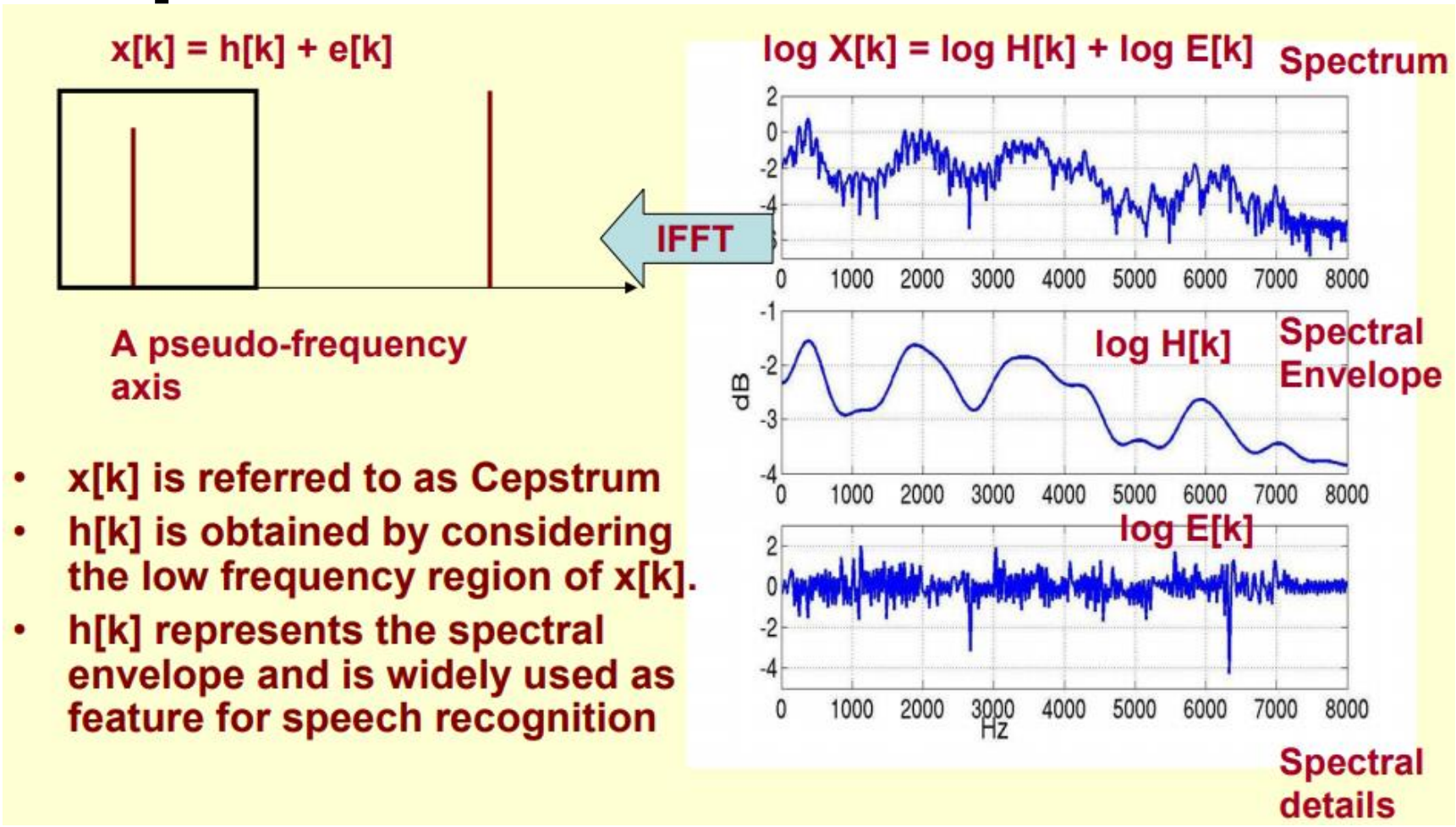
- 우리가 보는 Speech signal은 voice tract filter와 glottal source 가 곱해진 signal인데 이때 log는 이 둘의 곱을 합으로 다시 표현할 수 있게끔 해준다.

- $Log\ |S(f)| = log|H(f)| + log|G(f)|$ 고 할 수 있다. 분리한 뒤에는 더 관심이 있는 부분을 분석할 수 있게 된다.

https://slideplayer.com/slide/13590881/

http://www.cs.tut.fi/~sgn14006/PDF2015/S04-cepstrum.pdf

Cepstrum

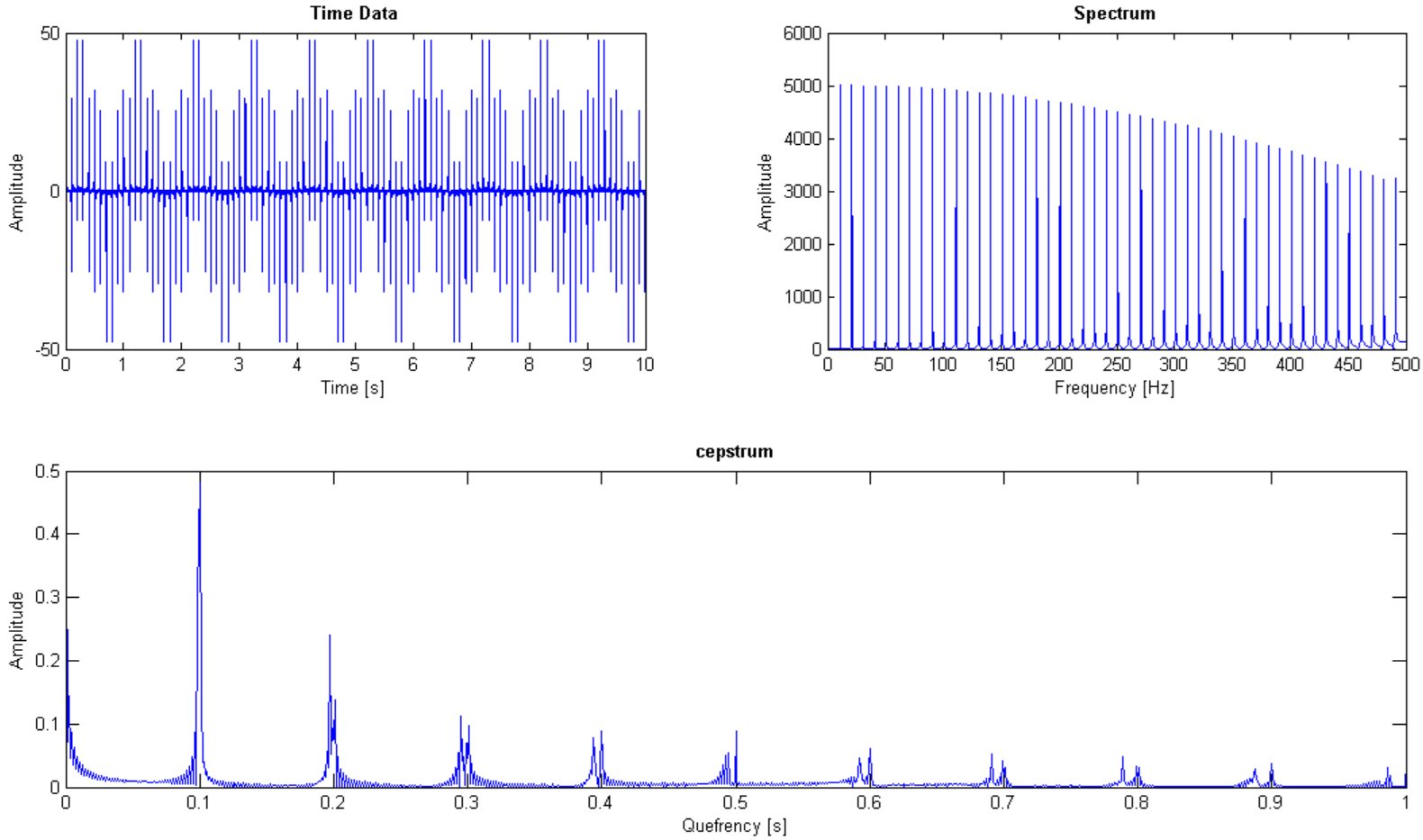


- x[k] is referred to as Cepstrum
- h[k] is obtained by considering the low frequency region of x[k].
- h[k] represents the spectral envelope and is widely used as feature for speech recognition

더 자세한 과정은 아래의 링크를 참조하시기 바랍니다.

www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

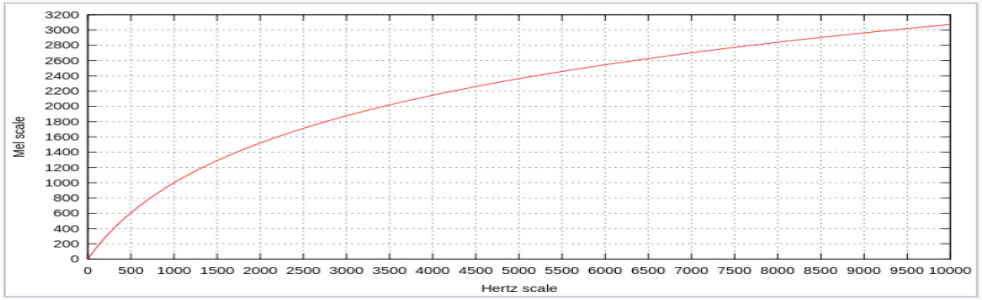
Cepstrum



https://www.researchgate.net/post/What_is_ceptrum_and_how_is_it_useful_in_signal_processing

Mel-Frequency Analysis

- Mel-Frequency analysis 란 인간의 지각 실험을 기반으로 한 분석법.
- 인간의 귀는 특정 주파수 성분에 집중하는 filter 역할을 하는 것으로 관찰되어진다. 이 때 이 필터는 저주파 영역의 정보는 더 많이 받아들이고 고주파 영역의 정보는 덜 받아들인다.
- 이러한 구조를 모사한 filter 가 Mel-Frequency filter 이다. 이 때 filter를 어느 간격으로 나눌지를 정하는 것이 mel-scale인데, mel-scale이란 청중들이 서로 거리가 같다고 판단한 Pitch(음정)의 지각적인 척도(perceptual scale)이다.



Plots of pitch mel scale versus Hertz scale

A popular formula to convert f hertz into m mels is:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

The corresponding inverse expressions are:

$$f = 700\left(10^{\frac{m}{2595}} - 1\right) = 700\left(e^{\frac{m}{1127}} - 1\right)$$

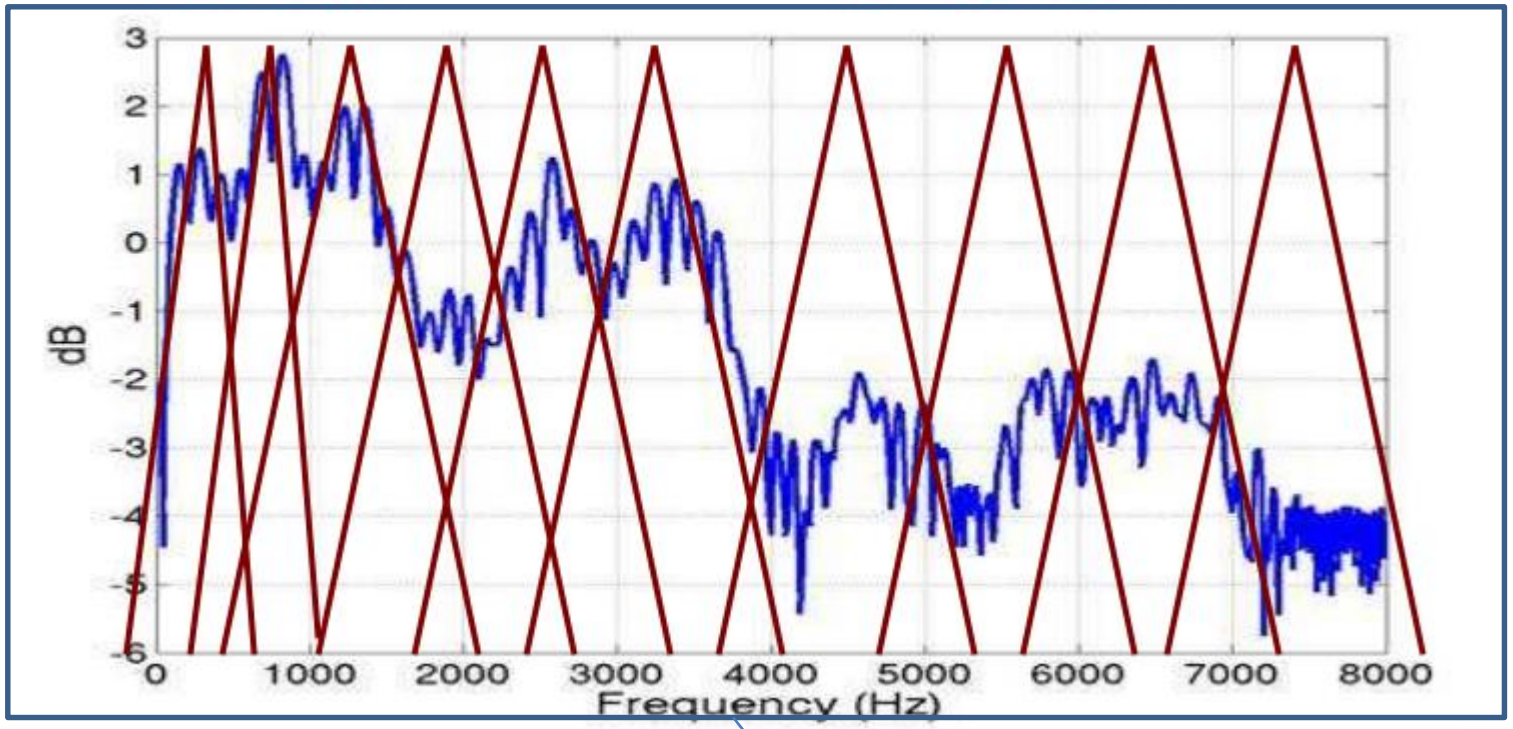
- STFT 에 Mel-Filter를 취한 Spectrogram을 Mel-Spectrogram 이라고 한다. Mel-Frequency 특징이 들어감.

www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Mel-Frequency Filters

More no. of filters in low freq. region

Lesser no. of filters in high freq. region



an array of band-pass filter : Filter bank

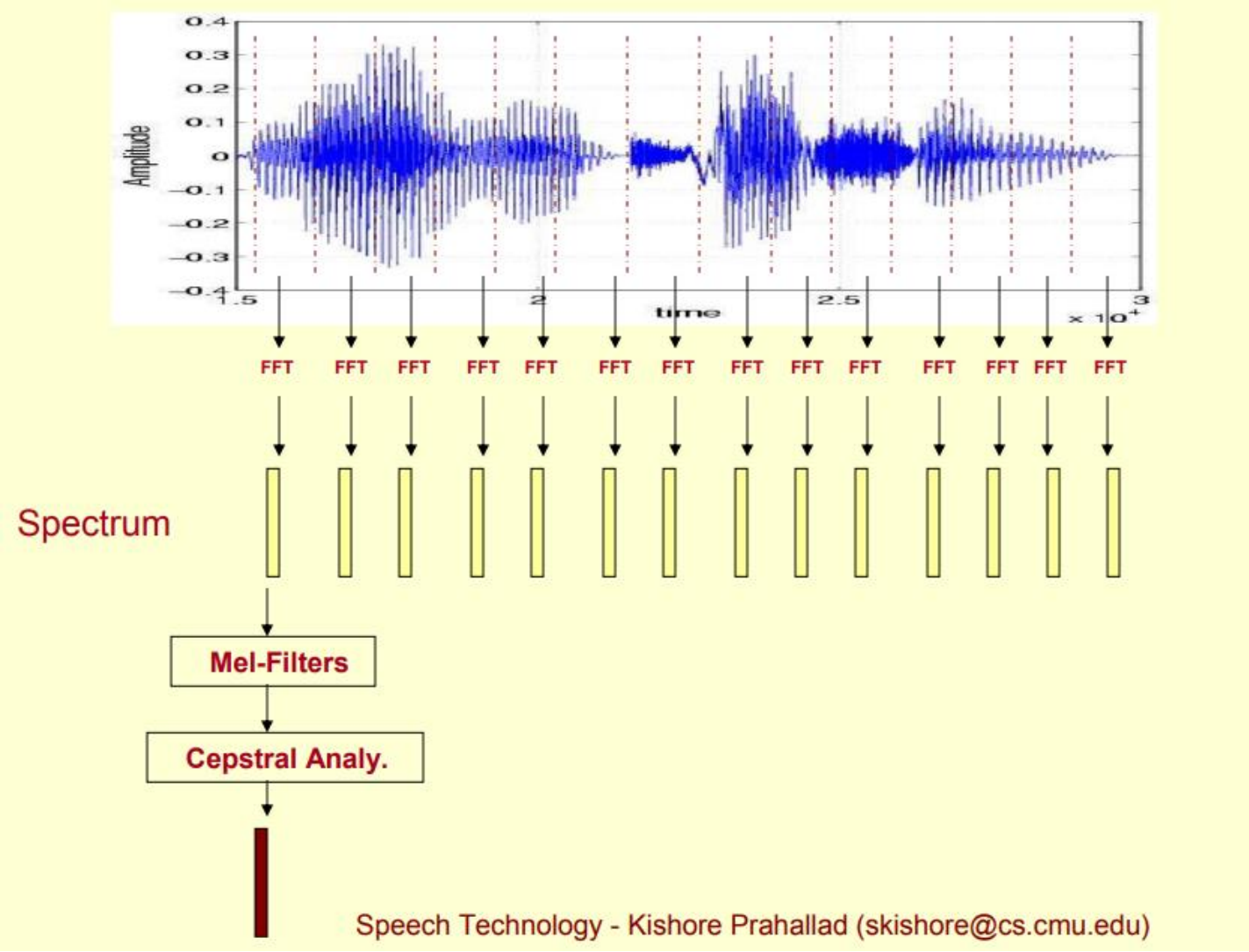
www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Mel-Frequency Cepstral Coefficients (MFCC)

- MFCC 는 집합적으로 MFC(Mel-Frequency Cepstrum) 를 구성하는 계수(Coefficients)이다. 이 계수가 음성 데이터에서 추출된 특징값 (feature value) 이고 vector로 이루어진다. Feature extraction으로 널리 사용된다.

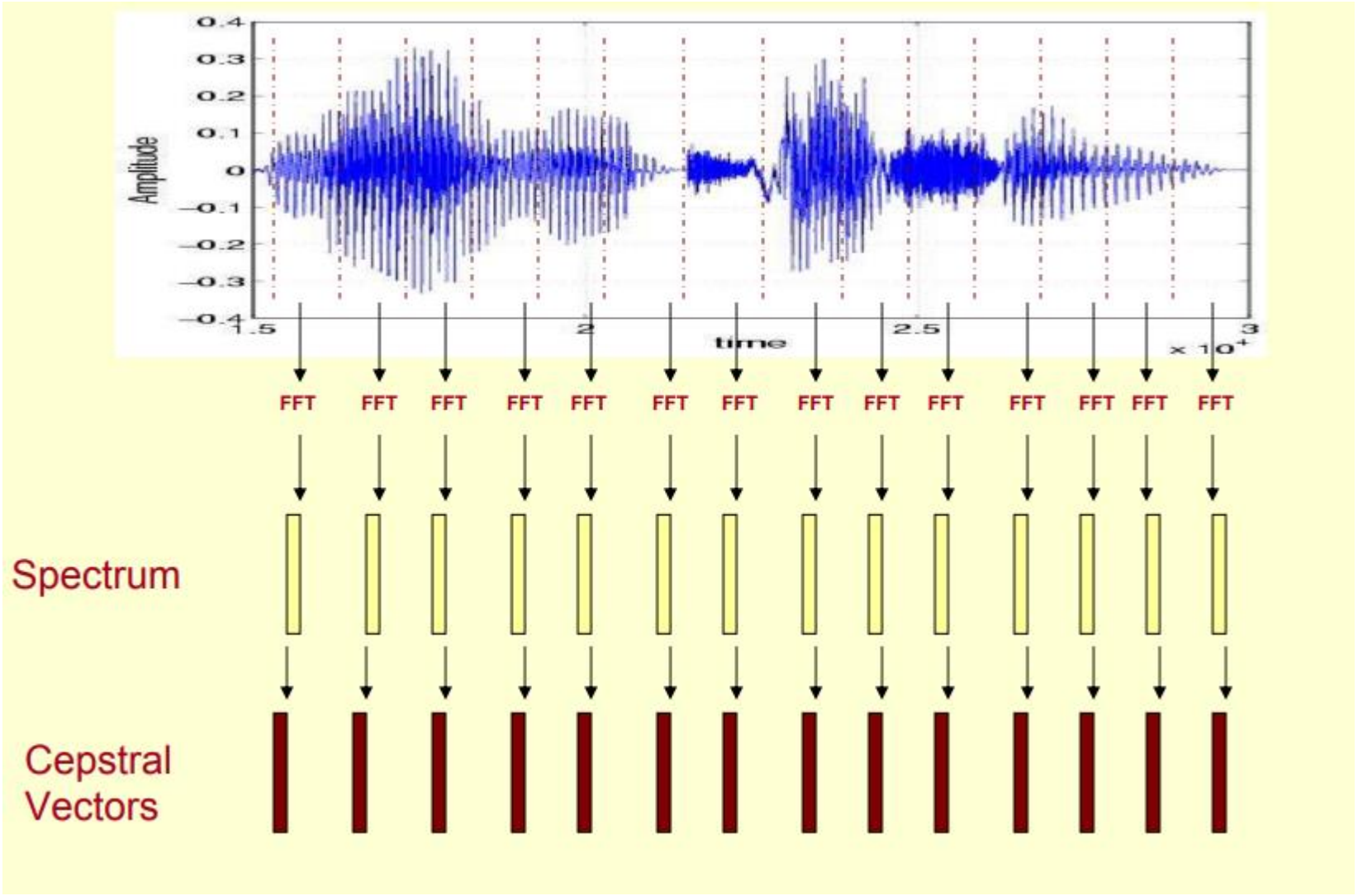
- 아래와 같은 순서로 계산되어진다.
  - Signal을 짧은 간격으로 쪼갠다 → Windowing
  - Frequency-domain으로 변환 → FFT
  - 각 Spectrum에 mel-filter bank를 적용 → Mel-Spectrum
  - Mel-Spectrum 에 log 적용
  - Mel-log-Spectrum list 전체에 DCT(Discrete Cosine transform) 적용
  - 얻어진 Coefficients 에서 앞에서부터 N개만 남기고 버린다.
    - 나머지의 계수가 음성 인식 등에 크게 기여하지 못하기 때문이다.

Mel-Frequency Cepstral Coefficients (MFCC)



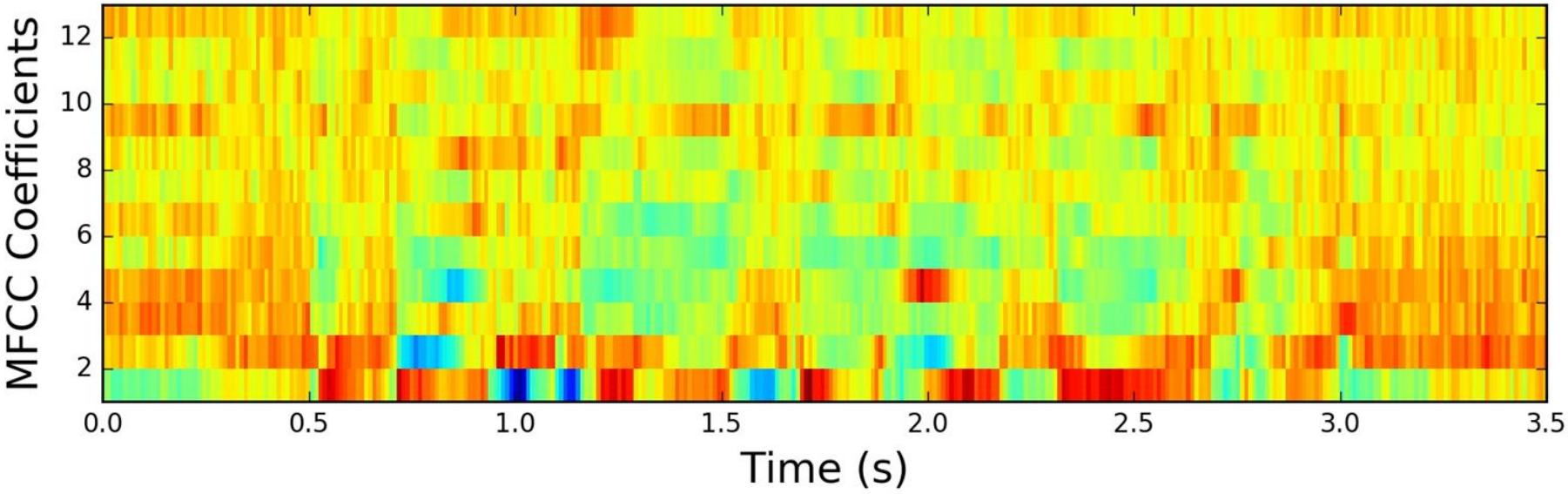
www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Mel-Frequency Cepstral Coefficients (MFCC)



www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Mel-Frequency Cepstral Coefficients (MFCC)



https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

When do I use these data?

- 시각적 형태로 작업할 경우 Spectrogram으로 접근해본다.
  - 음악, 리듬 관련은 STFT
  - 사람의 perception에 관련된 경우 또는 Speech 는 Mel-Spectrogram
  - 전반적인 '소리'에 관련된거면 실험적으로 결정해본다.

- 특징값 추출인 경우 (EX. 음성인식)
  - MFCC

- 접근법을 대략적으로 나눈 것이지 절대적인 것은 아니니 다각도로 접근해 볼 것!

Speech Separation

- Background Interference (방해 요소)
  - Nonspeech noise (비음성 잡음)
  - interfering speech (방해되는 음성 or target이 아닌 음성)
  - room reverberation (잔향)
    - 목표가 아닌것들은 결국 다 noise

- Cocktail Party Problem
  - 여러 화자의 speech가 섞인 음원에서 해당 speech를 분리하는 작업.

- Speech Enhancement(Denoising)
  - 음성(speech)와 비음성(nonspeech)을 분리하는 작업.

- **주제에 대한 용어의 차이일뿐 접근법은 크게 다르지 않다.**

How to measure noise level

- Signal-to-Noise-Ratio(SNR)
  - 원하는 신호의 크기가 잡음의 크기보다 얼마나 큰지 나타내는 비율. 단위는 dB(데시벨)이다.
  - 음성(Speech)을 기준으로 얘기하면 Speech 이외의 소리는 다 Noise.

$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} = \left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right)^2$$

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) = P_{\mathrm{signal,dB}} - P_{\mathrm{noise,dB}}$$

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}\left[\left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right)^2\right] = 20\log_{10}\left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right) = (A_{\mathrm{signal,dB}} - A_{\mathrm{noise,dB}})$$

  - 여기서 P_{signal} 와 P_{noise} 은 신호의 전력(power)과 노이즈의 전력에 해당한다. 즉, SNR은 신호의 'power'와 노이즈의 'power'의 비율. 이때 'P' 는 average power 이다.

  - 신호와 잡음이 동일한 임피던스에서 측정되는 경우엔 SNR은 진폭(Amplitude) 비율의 제곱을 계산하여 얻을 수 있다. 이것이 각각 A_{signal} 와 A_{noise} 이고 A 는 root mean square(RMS) 로 구한 Amplitude이다.

Traditional Approach

- Spectrum subtraction
  - 잡음이 있는 신호 spectrum에서 추정한 noise spectrum을 뺀 후 다시 복원하여 clean 소리 신호를 만들어 내는 방법.

- Supervised Speech Separation Based on Deep Learning: An Overview

- Ideal Binary Mask(IBM)
  - 시간 주파수(time-frequency) 영역에서 SNR 값이 임계값(threshold) 보다 낮은 영역의 신호엔 '0' 의 마스크를 할당하여 제거하고 그렇지 않는 영역은 '1' 의 마스크를 적용하여 신호를 살린다. 간단히 말해 잡음이 많은 부분은 제거하지만 그렇지 않은 부분은 건들지 않음으로써 음성 왜곡을 최소화 시킨다. 입력 음성에 대한 해당 IBM을 분류하는 확률모델은 감독 학습으로 이루어진다.

- A Post-processing for Binary Mask Estimation Toward Improving Speech Intelligibility in Noise

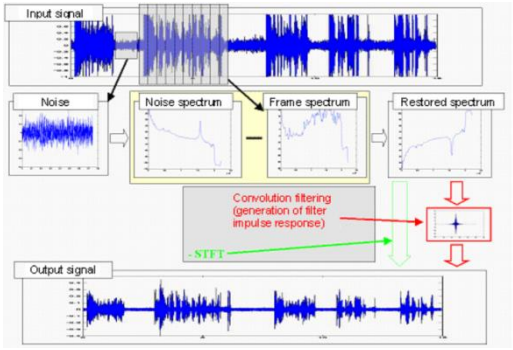
- Beamforming
  - Multi-microphone(array-based) 일 때 사용하는 방식이다. delay-and-sum-beamforming 이 대표적인데,대상 방향의 다중 마이크 신호를 위상(Phase)에 추가 하고 위상 차이를 사용해서 다른 방향의 신호를 감쇠한다. 단, 잔향이 있는 환경에선 빔포밍 방식이 방해 받을 수 있다.

- Supervised Speech Separation Based on Deep Learning: An Overview

- 이외에도 여러 방법들이 많이 존재한다.

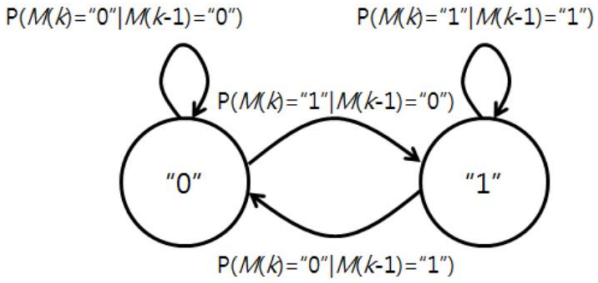
Traditional Approach

- Spectrum subtraction

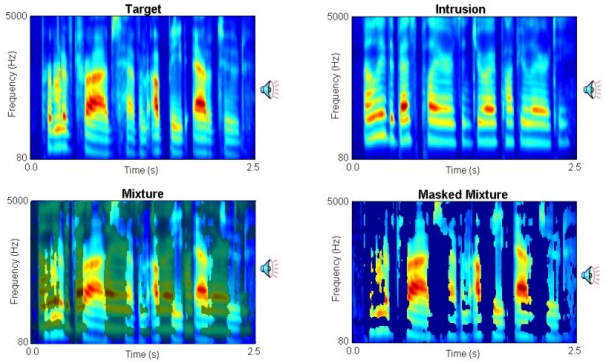


참조 링크

- Ideal Binary Mask(IBM)

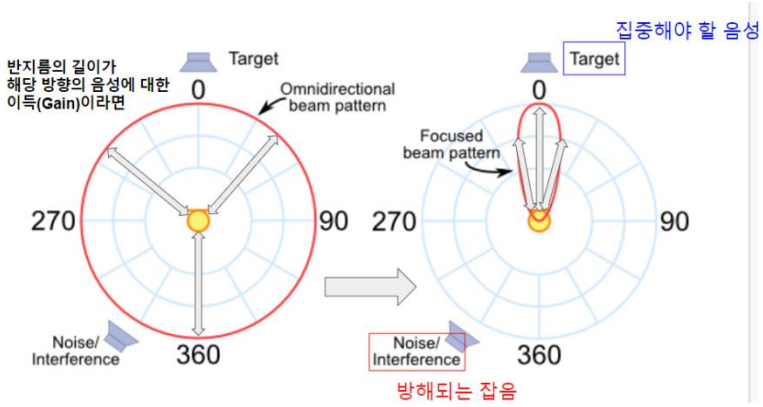


참조 링크



참조 링크

- Beamforming



참조 링크

Deep Neural Network Approach - Separation

- MULTI-SCALE MULTI-BAND DENSENETS FOR AUDIO SOURCE SEPARATION - Naoya Takahashi, Yuki Mitsufuji [17.06]

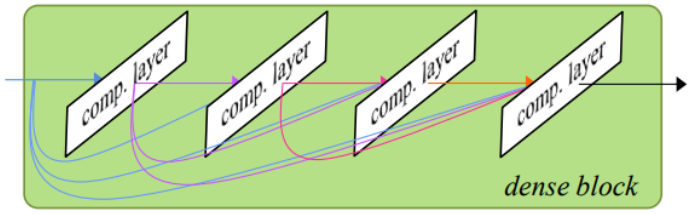


Figure 1: dense block architecture. The input of a composite layer is the concatenation of outputs of all preceding layers.

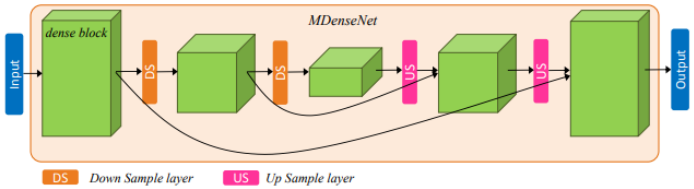


Figure 2: MDenseNet architecture. Multi-scale dense blocks are connected though down- or up-sampling layer or through block skip connections. The figure shows the case $s = 3$.

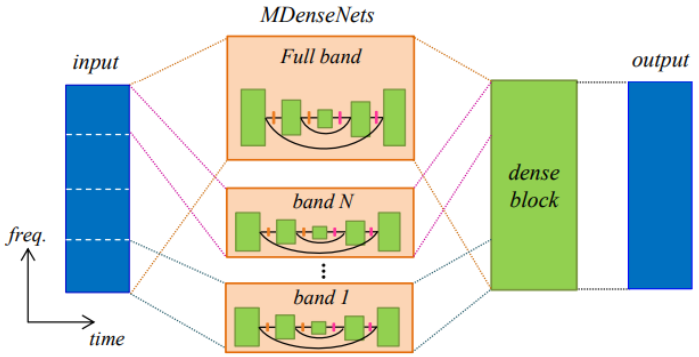


Figure 3: MMDenseNet architecture. Outputs of MDenseNets dedicated for each frequency band including full band are concatenated and the final dense block integrate features from these bands to create final output.

End-to-End 방식으로 CNN을 이용해서 target 음원(source)를 구해내는 방식이다.
음악 신호에서 4가지의 음원을 분리하는 실험을 시도했다.

STFT를 사용해 Spectrogram 으로 input data를 넣어준다.

Deep Neural Network Approach - Enhancement

- SEGAN: Speech Enhancement Generative Adversarial Network
  - Santiago Pascual , Antonio Bonafonte , Joan Serra [17.03]

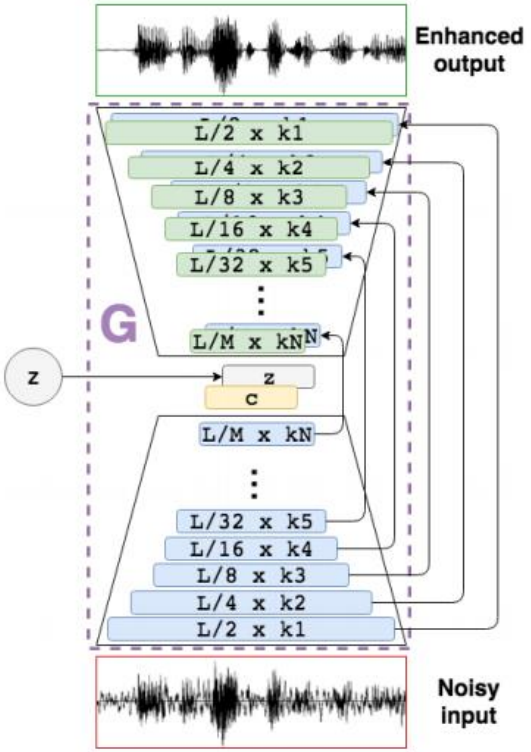


Figure 2: *Encoder-decoder architecture for speech enhancement (G network). The arrows between encoder and decoder blocks denote skip connections.*

End-to-End 방식으로 Convolutional Network 구조의 GAN을 이용해서 Noisy 한 Speech를 입력으로 넣어 Enhanced Speech를 생성해낸다.

Waveform(time-domain signal) 으로 input data를 넣어주었다.

Deep Neural Network Approach - Enhancement

- MULTI-RESOLUTION FULLY CONVOLUTIONAL NEURAL NETWORKS FOR MONAURAL AUDIO SOURCE SEPARATION - Emad M. Grais, Hagen Wierstorf, Dominic Ward, and Mark D. Plumbley [17.10]

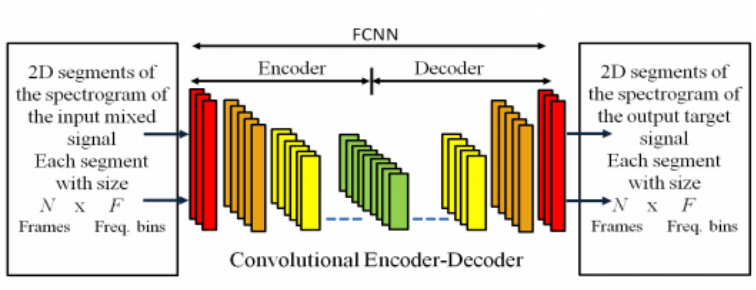


Fig. 1: The overview of the structure of a FCNN that separates one target source from the mixed signal. Each layer consists of a single set of filters with the same size followed by a rectified linear unit (ReLU) as activation function. The set of filters in the input and output layers have large filter sizes and small number of filters. The number of filters increases and the size decreases when getting further from the input and output layers [21]. There is symmetric in the filter sizes and numbers of filters between the encoder and decoder sides.

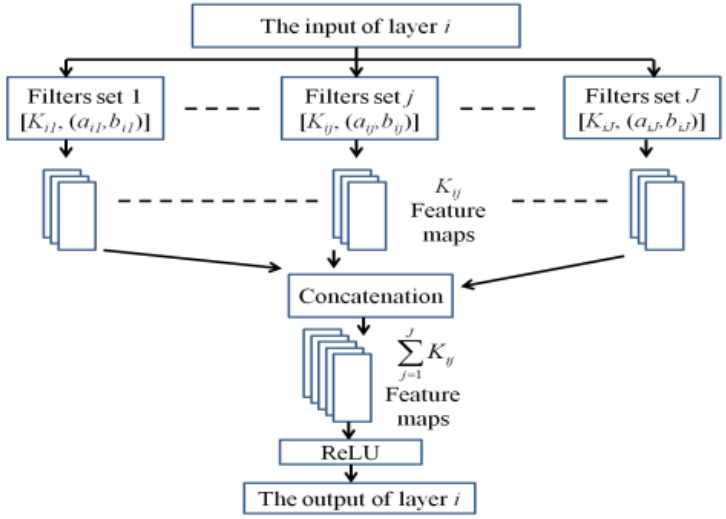


Fig. 2: The overview of the proposed structure of each layer of the MR-FCNN. K_{ij} denotes the number of filters with size $a_{ij} \times b_{ij}$ in set j in layer i. a_{ij} is the dimension in the time direction of the filters and b_{ij} is the dimension in the frequency direction of the filters in set j and layer i. The filters in different sets have different sizes and the filters within a set have the same size. Each set j in layer i generates K_{ij} feature maps. The number of feature maps that each layer i generates equal to the sum of the number of feature maps that all the sets in layer i generate ($\sum_{j=1}^{J} K_{ij}$). ReLU denotes a rectified linear unit (ReLU) as an activation function.

End-to-End 방식으로 Convolutional Network 구조.
Convolutional Autoencoder를 채용했다.

Noisy 한 Speech를 입력으로 넣어 Enhanced Speech를 생성해 낸다.

STFT를 사용해 Spectrogram 으로 input data를 만들었다.

Deep Neural Network Approach - Enhancement

- PHASE-AWARE SPEECH ENHANCEMENT WITH DEEP COMPLEX U-NET
  -[NAVER] Hyeong-Seok Choi, Janghyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, Kyogu Lee [18.12]

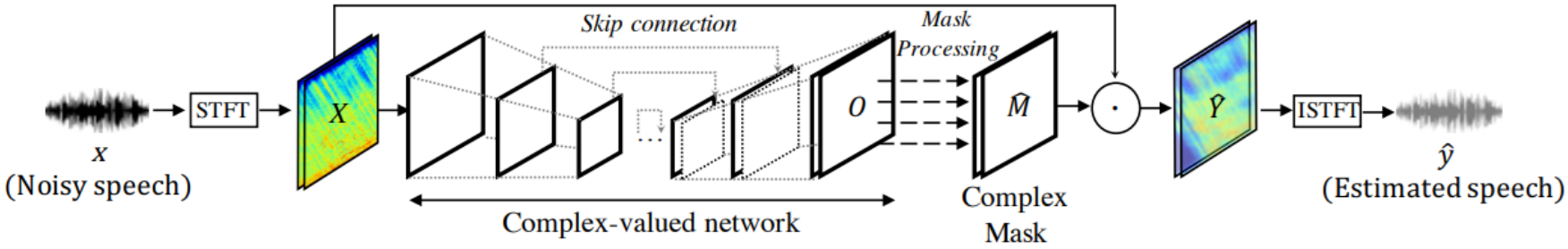


Figure 2: Illustration of speech enhancement framework with DCUnet.

End-to-End 방식으로 Convolutional Network 구조로 U-net을 채용했다.

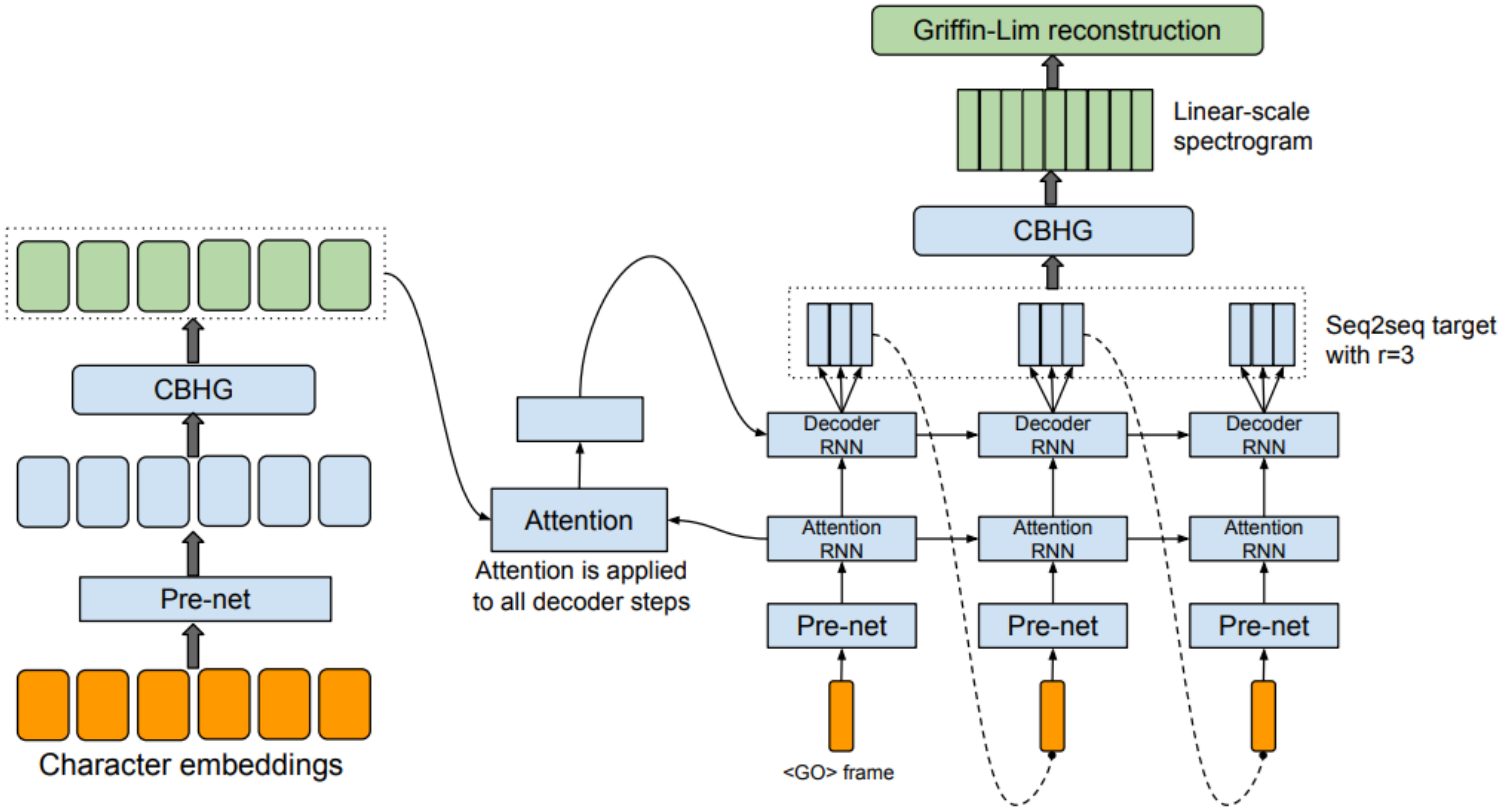
Spectrogram 특성상 복소수(Complex-value) 값으로 구성되어있다. 따라서 복소수 연산을 고려해 Complex-valued convolutional network 로 cnn을 구성해주었다.
(참고 코드: https://gist.github.com/DevKiHyun/c7c07aebb4a5fa0971b0f945b75f0537)

Mask를 학습하는 방식으로 진행되며 input signal에 대해 추정된 mask를 가지고서 noisy input을 Masking 해준다.

STFT를 사용해 Spectrogram으로 input data를 넣어주었다.

Deep Neural Network Approach - Synthesis

- TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS
  - 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



Deep Neural Network Approach – Abnormal Detection

- [Fast Adaptive RNN Encoder–Decoder for Anomaly Detection in SMD Assembly Machine]
  - YeongHyeon Park, Il Dong Yun

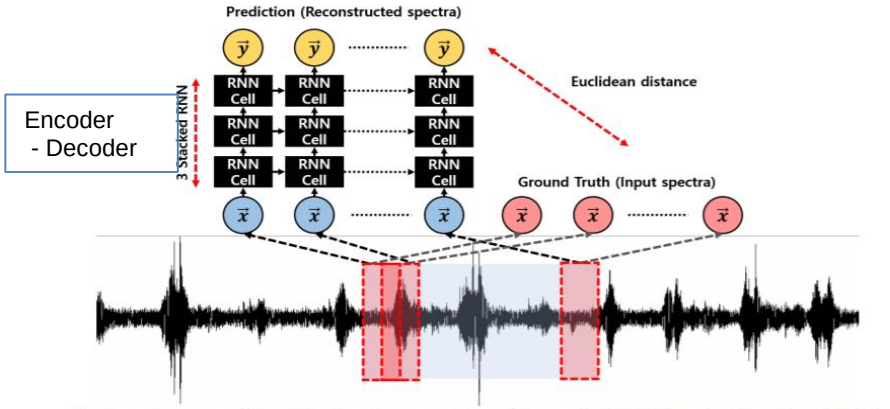
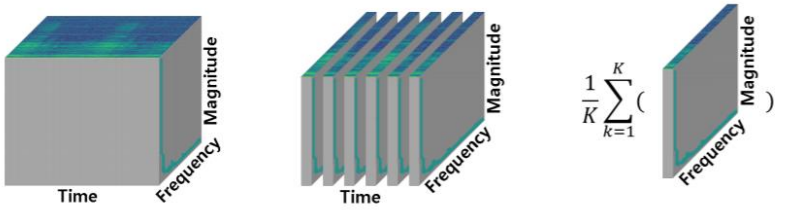


Figure 3. Structure of Fast Adaptive Recurrent Neural Network (RNN) Encoder–Decoder (FARED). The input is constructed by sequential spectrum (red box). Each spectrum has 50% of overlapping in the time domain. The output is a sequential reconstructed spectrum from input sequences. The Euclidean distance between prediction and ground truth is used for training and anomaly detection.



(a) the original spectra (b) sliced spectra (c) the averaging with spectra

Figure 1. The process for obtaining the average spectrum of spectra. (a) original spectra obtained from Short-Time Fourier Transform (STFT) or Mel-Frequency Cepstral Coefficients (MFCC) feature extraction; (b) sliced by time axis spectra according to the number of window used in STFT or MFCC feature extraction; (c) average of sliced spectra.

End-to-End 방식의 Recurrent Neural Network 구조.

정상적인 기계 소리만 Encoder-Decoder 에 넣고 output이 input과 동일하게 나오도록 학습시키면, 실전에서 정상적이지 않는 기계소리가 들어가게 될 때 제대로 복원이 되지 않는 점을 이용해서 비정상 소리를 검출하게 된다. 이때의 오차는 MSE를 통해서 계산한다.

Autoencoder 로 하면 비정상 소리도 정상으로 복원하기 때문에 Encoder-Decoder는 AE와 다르게 dimension reduction을 하지 않는다.

모델이 가볍고 빨라 산업 현장에 바로 응용할 수 있다는 점이 장점.

입력 데이터를 Spectrogram으로 변환 한 후 특정 간격(시간)을 기준으로 쪼개 평균을 취한 하나의 Spectrum 으로 만들어준다.

학습 단계에선 해당 입력 데이터를 input 과 label 둘 다 동일하게 넣어준다.

감사합니다

책 읽어주는 딥러닝

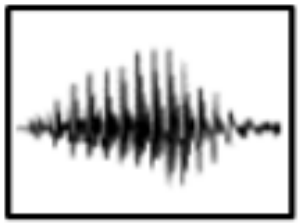
음성합성 기술

TTS(Text-to-Speech)

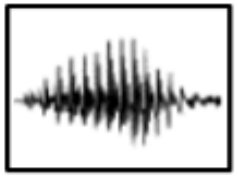
음성 합성

텍스트 → 음성 합성 → 음성

“오늘도 즐거운 하루 되세요”



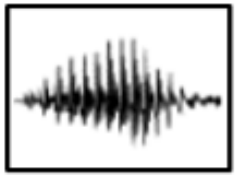
음성



음성 분류

긍정/부정

음성



음성 인식

텍스트

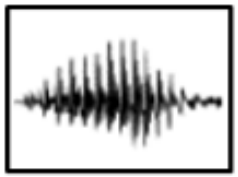
"오늘도 즐거운 하루 되세요"

텍스트

"오늘도 즐거운 하루 되세요"

음성 합성

음성



어디에 사용될 수 있을까요??

1. 대화 챗봇
   - 인공지능 스피커, 빅스비, Siri

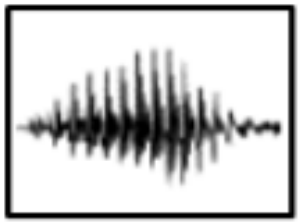
2. 오디오 북
   - 엄마 목소리로 아이에게 책을 읽어줌
   - 시각장애인이 읽고 싶은 책을 오디오로 만들어줌

3. 음성 안내 시스템
   - 박물관, 공공 시설

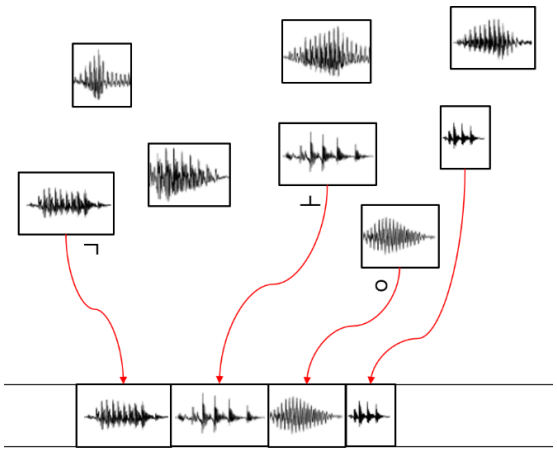
[책 읽어 주는 유인나]

과거의 음성 합성 기술들

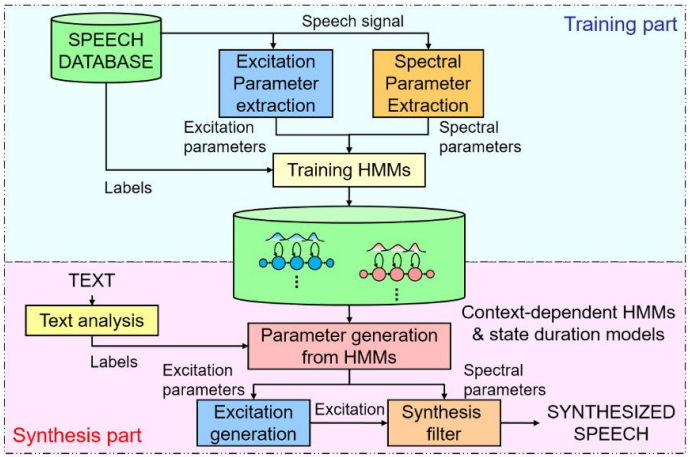
텍스트 → 음성 합성 → 음성



"오늘도 즐거운 하루 되세요"

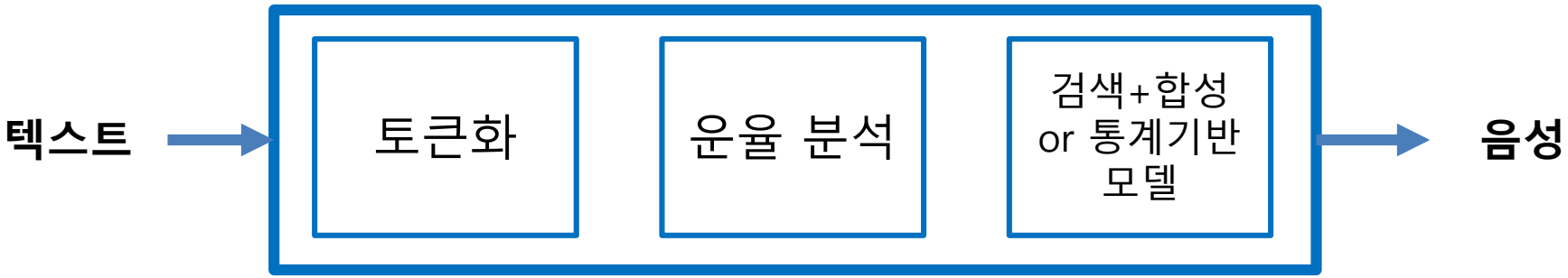


1. 연결 합성

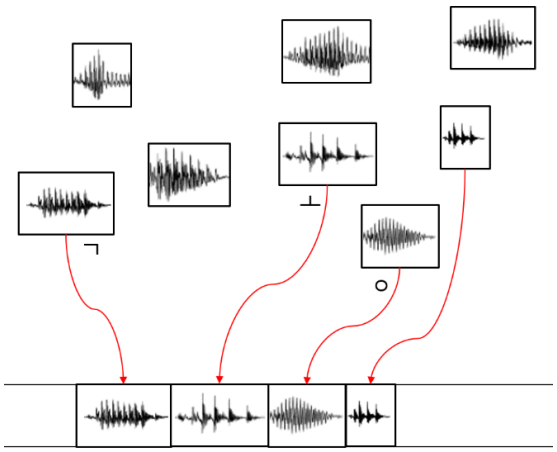


2. 통계 기반 합성

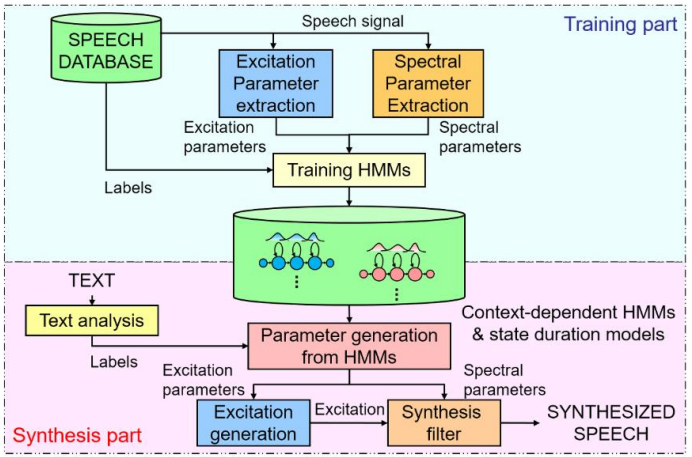
과거의 음성 합성 기술들



"오늘도 즐거운 하루 되세요"

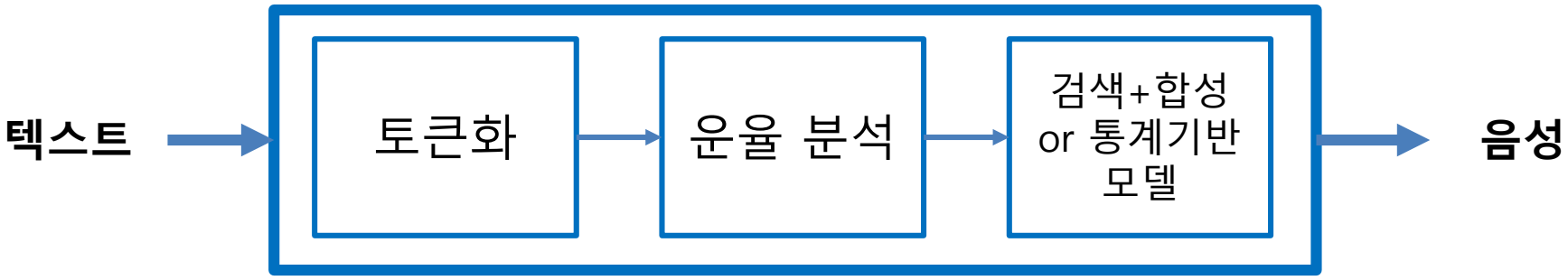


1. 연결 합성



2. 통계 기반 합성

딥러닝 기반 음성 합성 기술

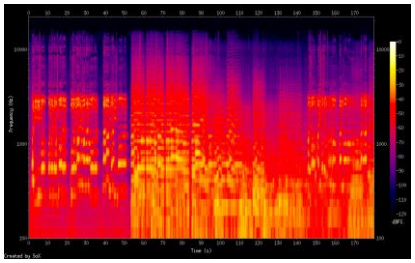
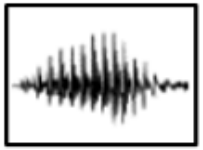


과거 음성 합성 기술

딥러닝 기반 음성 합성 기술

딥러닝 기반 음성 합성 기술

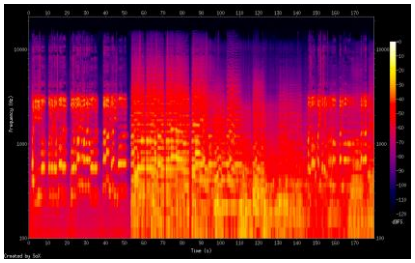
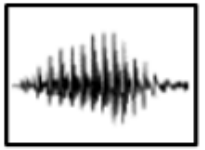
"오늘도 즐거운 하루 되세요"

텍스트 → Encoder Decoder →  → Vocoder →  음성

스펙트로그램 생성
(음성의 주파수 표현)

딥러닝 기반 음성 합성 기술

"오늘도 즐거운 하루 되세요"

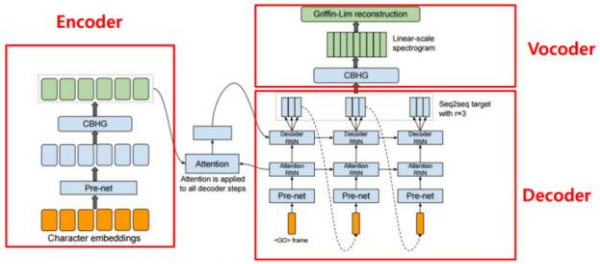
텍스트 → Encoder Decoder →  → Vocoder →  음성

스펙트로그램 생성
(음성의 주파수 표현)

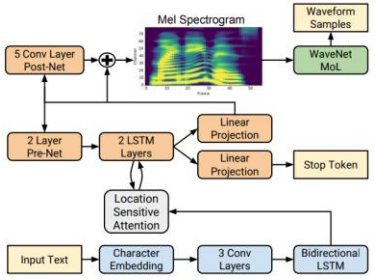
Encoder-Decoder 기술
1) Tacotron 1
2) Tacotron 2
3) DCTTS

모두 구현 후 성능비교
Tacotron2 선택

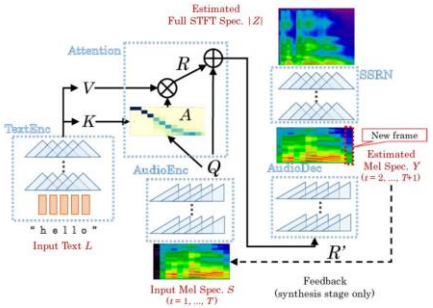
Encoder-Decoder 기술



Tacotron 1
https://arxiv.org/pdf/1703.10135.pdf



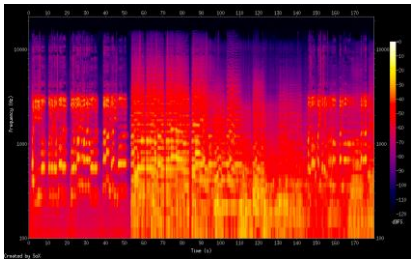
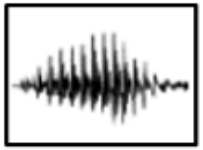
Tacotron 2
https://arxiv.org/pdf/1712.05884.pdf



DCTTS
https://arxiv.org/abs/1710.08969

딥러닝 기반 음성 합성 기술

"오늘도 즐거운 하루 되세요"

텍스트 → Encoder Decoder →  → Vocoder →  음성

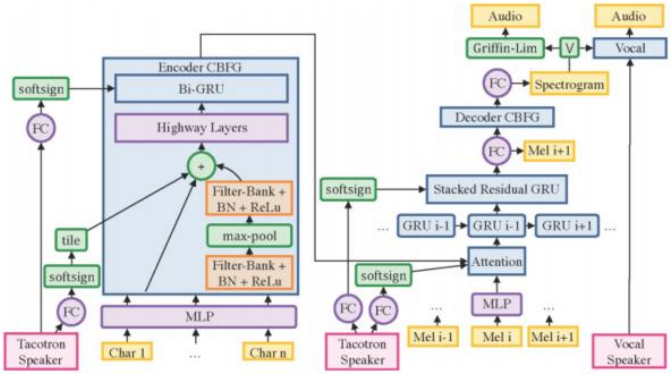
Encoder-Decoder 기술
1) Tacotron 1
2) Tacotron 2
3) DCTTS

스펙트로그램 생성
(음성의 주파수 표현)

Vocoder 기술
1) Griffin-Lim
2) Wavenet
3) Waveglow

Multi-Speaker 기술로 확장

+

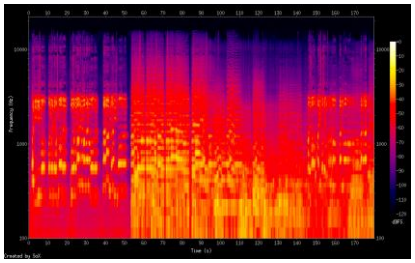
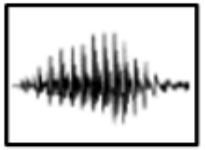


Deep Voice 2

https://arxiv.org/abs/1705.08947

딥러닝 기반 음성 합성 기술

"오늘도 즐거운 하루 되세요"

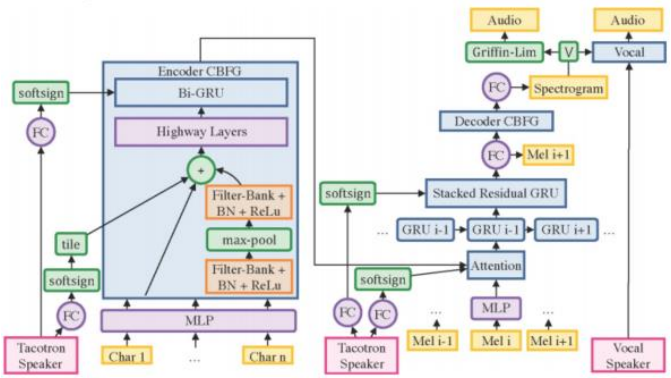
텍스트 → Encoder Decoder →  → Vocoder →  음성

Encoder-Decoder 기술
1) Tacotron 1
2) Tacotron 2
3) DCTTS

스펙트로그램 생성
(음성의 주파수 표현)

Vocoder 기술
1) Griffin-Lim
2) Wavenet
3) Waveglow

Multi-Speaker 기술로 확장



Deep Voice 2

https://arxiv.org/abs/1705.08947

한글에 적용

- 한글 text를 (초성/중성/초성)으로 나누어진 sequence로 만들어야 한다.
  - ✓ jamo package를 이용하면 된다.
  - ✓ '존경하는' → ['ㅈ', 'ㅗ', 'ㄴ', 'ㄱ', 'ㅕ', 'ㅇ', 'ㅎ', 'ㅏ', 'ㄴ', 'ㅡ', 'ㄴ']

'_': 0, '~': 1, 'ㄱ': 2, 'ㄲ': 3, 'ㄴ': 4, 'ㄷ': 5, 'ㄸ': 6, 'ㄹ': 7, 'ㅁ': 8, 'ㅂ': 9, 'ㅃ': 10,
'ㅅ': 11, 'ㅆ': 12, 'ㅇ': 13, 'ㅈ': 14, 'ㅉ': 15, 'ㅊ': 16, 'ㅋ': 17, 'ㅌ': 18, 'ㅍ': 19, 'ㅎ': 20,
'ㅏ': 21, 'ㅐ': 22, 'ㅑ': 23, 'ㅒ': 24, 'ㅓ': 25, 'ㅔ': 26, 'ㅕ': 27, 'ㅖ': 28, 'ㅗ': 29, 'ㅘ': 30,
'ㅙ': 31, 'ㅚ': 32, 'ㅛ': 33, 'ㅜ': 34, 'ㅝ': 35, 'ㅞ': 36, 'ㅟ': 37, 'ㅠ': 38, 'ㅡ': 39, 'ㅢ': 40,
'ㅣ': 41, 'ㄳ': 42, 'ㄵ': 43, 'ㄶ': 44, 'ㄺ': 45, 'ㄻ': 46, 'ㄼ': 47, 'ㄽ': 48, 'ㄾ': 49, 'ㄿ': 50,
'ㅀ': 51, 'ㅄ': 52, 'ㅋ': 53, 'ㅌ': 54, 'ㅍ': 55, 'ㅎ': 56, 'ㅁ': 57, 'ㅂ': 58, 'ㅆ': 59, 'ㅅ': 60,
'ㅆ': 61, 'ㅇ': 62, 'ㅈ': 63, 'ㅊ': 64, 'ㅋ': 65, 'ㅌ': 66, 'ㅍ': 67, 'ㅎ': 68, '!': 69, '"': 70,
'(': 71, ')': 72, ',': 73, '-': 74, '.': 75, ':': 76, ';': 77, '?': 78, ' ': 79

80개 token

딥러닝 기반 음성 합성 기술

Tacotron2 + Waveglow + Multi-speaker + 한글 적용 기술 확보

Tacotron2 : 양질의 스펙트로그램 생성

Waveglow : 스펙트로그램에서 고음질의 음성 생성

Multi-speaker : 여러 명의 목소리를 하나의 모델로 생성

한글 적용 : 기존 모델들은 영어 생성에 특화된 모델

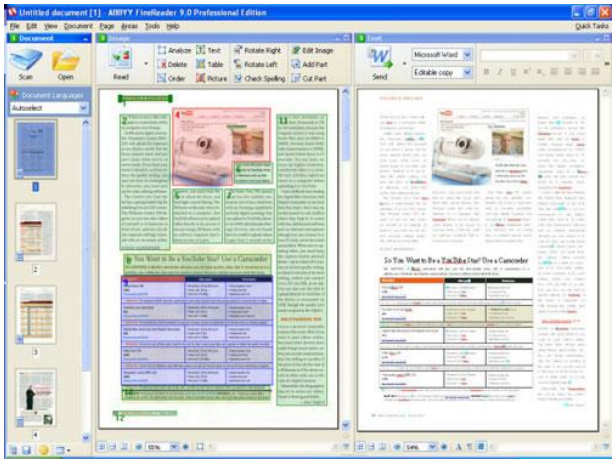
한글 음성 생성을 위해 한글 jamo 단위로 모델 구현

감사합니다

OCR

- **OCR (Optical Character Recognition): 광학문자 인식**
  - 인간이 종이 위에 써 놓은 글씨를 인지하여 텍스트 데이터로 치환
  - 필기체보단 정자로 또박또박 잘 쓴 글씨가 인식률이 더 높음
  - 프린터로 인쇄했던 문서라면 잘 인식

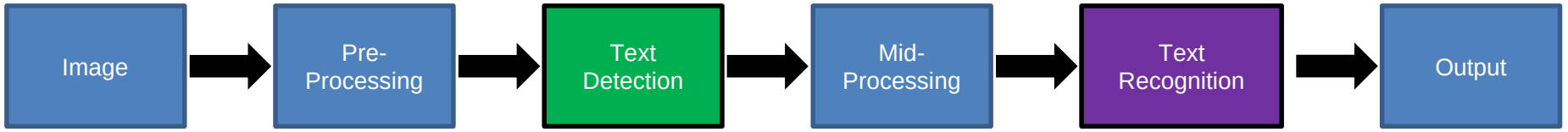
Clova OCR

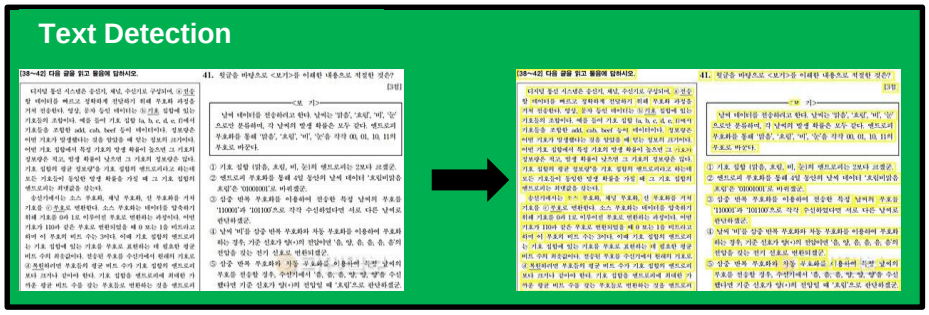


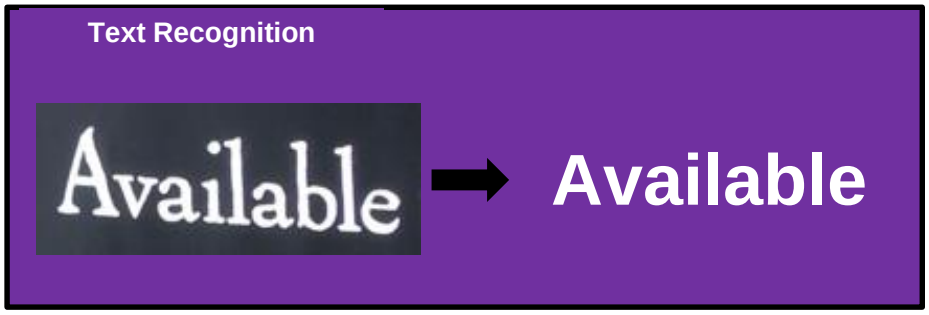
ABBYY

OCR

width_ths = 0.5 width_ths = 0.7







Dataset Images

Ground Truth (text files)

```
11, 15, 42, 28, "How"
41, 16, 61, 28, "to"
8, 33, 36, 46, "Find"
41, 32, 64, 46, "the"
11, 50, 61, 65, "Perfect"
16, 69, 56, 82, "HDTV"
```

```
22 249 113 286 "The"
142 249 287 286 "Photo"
326 245 620 297 "Specialists"
```

```
0, 1, 177, 32, "\"HIGHER"
11, 35, 182, 63, "SAVINGS"
12, 67, 183, 103, "RATES\""
50, 114, 56, 120, "A"
60, 114, 91, 120, "reward"
96, 114, 108, 120, "for"
112, 116, 126, 120, "our"
130, 114, 164, 120, "current"
...
```

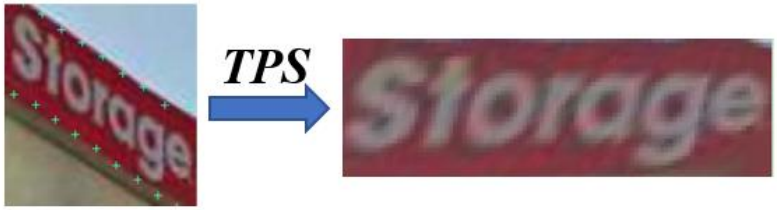
Ground Truth Visualisation

OCR

Irregular text recognition methods

Multi-directional feature-based method[1]

Segmentation-based method[2]



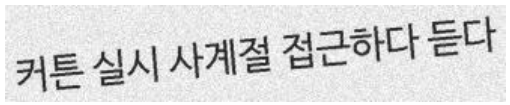
Rectification-based method[3]

2D-attention based method[4]

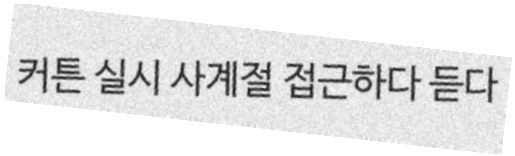
OCR

Image → Pre-Processing → Text Detection → Mid-Processing → Text Recognition → Output

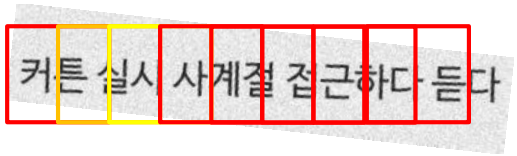
Text Recognition

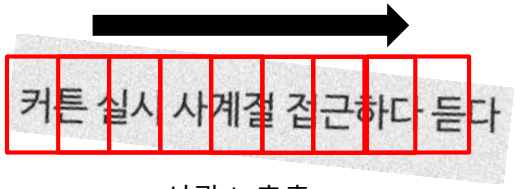


변환
TPS



특징추출
Resnet/VGG
Efficientnet



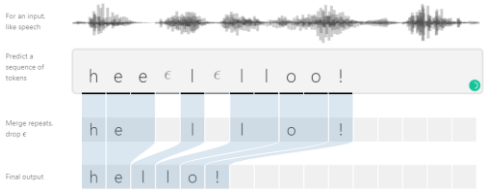


시퀀스 추출
BiLSTM

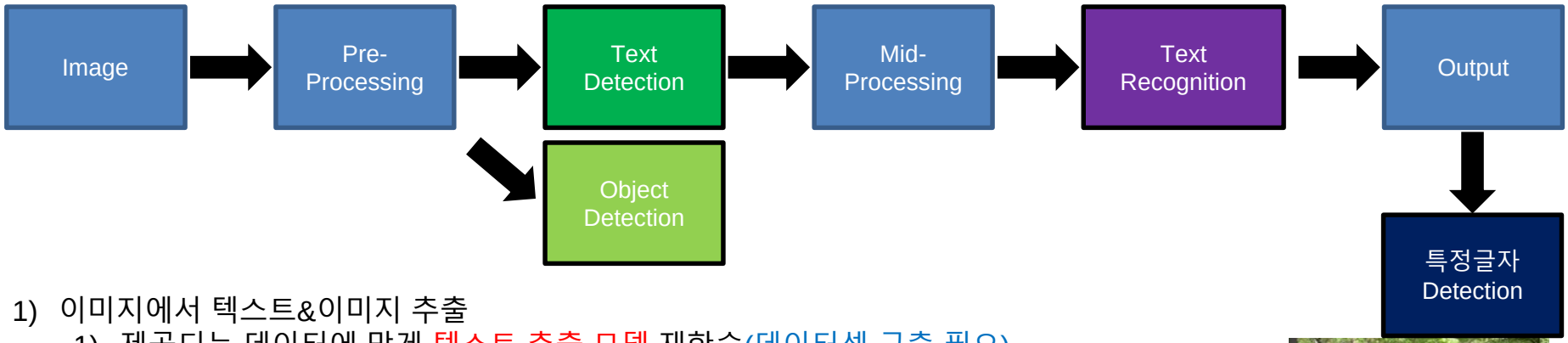
Prediction
CTC/Attn

a-----v--a-i-l-a-bb-l-ee---

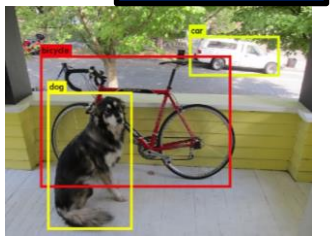
available



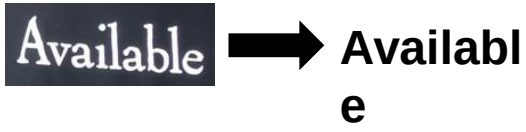
Connectionist Temporal Classification 예시



1) 이미지에서 텍스트&이미지 추출
   1) 제공되는 데이터에 맞게 텍스트 추출 모델 재학습(데이터셋 구축 필요)
   2) 이미지 추출 모델 구축 및 학습(데이터셋 구축 필요)
      1) 물체검출 최고성능 알고리즘 구현 및 학습



2) OCR 구현(한국어)
   1) 제공되는 데이터의 텍스트 글꼴 예측
   2) 해당 글꼴 학습데이터 생성
   3) 실제 데이터 학습데이터 구축
   4) 학습 12가지 케이스, best 선정
   5) 영어 : 26자, 한글 11,172자(가~힣)



- 한국어/영어 Text recognition 모델 학습을 위한 학습데이터 생성기술

3) 특정글자 블러처리
   1) 특정 글자 위치 찾기
   2) 숫자는 쉽다(10개), 한글은 11,172 가지 CASE

감사합니다