A1003:
음성인식 및 기계학습

인공지능연구소, 복합지능연구실

박기영

강의내용: OVERVIEW

- 음성인식 이론
- 음성인식 실습
- 딥러닝 이론/실습 (Transformer)
- 딥러닝/리눅스 개발환경

강의일정

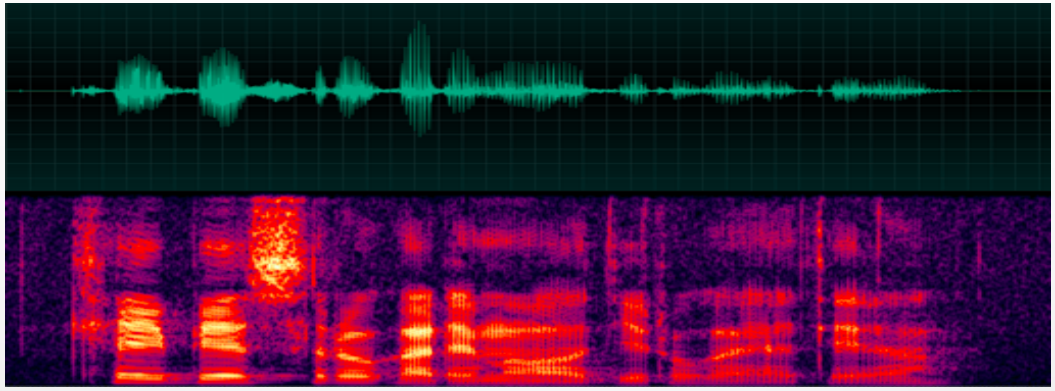
1일차	2일차	3일차
• 음성인식 개요 • (고전적) 음성인식 이론	• 딥러닝기반 (고전적) 음성인식 이론 • 종단형음성인식 개요	• 음성인식 이론
• 실습환경소개 • 인식률 측정하기	• ESPNet 소개 • 종단형 음성인식 recipe 살펴보기	• 음성인식 평가 실습 • 성능 측정
• (고전적) 음성인식 이론 • 특징추출	• 트랜스포머 소개	• 성능개선 방안 • 연구동향
• 훈련DB 소개 • 특징추출 실습	• 음성인식 훈련 실습	• 실습 마무리
Homework		

강의내용: 1일차

- What is Speech Recognition
- How to Evaluate Performance
- 실습
  - 실습환경 구성
  - WER 계산하기
- Feature Extraction
- 실습
  - Audio 들어보기, 멜스펙트럼 그려보기
- Q&A

WHAT IS SPEECH RECOGNITION

- ASR(Automatic Speech Recognition), STT: Speech-to-text



  - Isolated, Connected, Continuous, Keyword Spotting
  - Speaker Dependent/Independent
- Difference with Image/Video Classification
  - Sequence Generation Problem

HISTORY OF ASR

1950,60s

- Phonetic Recognizer
- 10 digit recognition
- DTW
- Idea of Continous ASR(CMU)

1970s

- IBM, Bell Lab, ...
- DARPA program
  - CMU Harpy:1,011 words vocab., FSN

1980s

- Connected words recognition (Fluently spoken)
- Template based → Statistical Methods
- HMM
- N-gram, Neural Nets.
- DARPA program
  - CMU SPHINX
  - BBN, SRI

1990s

- MCE, MMI
- DARPA programs
  - Natural Lanuage Recognition, ATIS, Broadcast news, Switchboard
- Robust ASR
- Applications

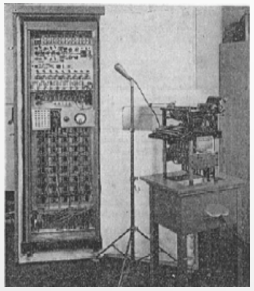
2000s

- Spontaneous speech
- Robust ASR
- Multimodal

50 Years of Progress in Speech and Speaker Recognition Research, ECTI Transactions On Computer And Information Technology, 2005

APPLICATIONS

1956,
RCA Labs



1975,
1997,
Nuance

DRAGON
NATURALLYSPEAKING

2012,
Google
Voice
Search

2011,
Apple
Siri

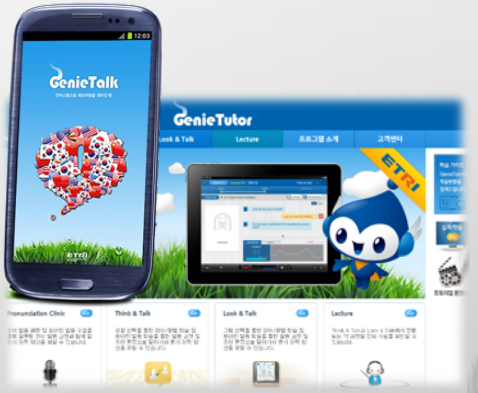
2014,
Amazon

2018,
Google
Duplex

1997,
삼성
애니콜

2008,
파인디지털

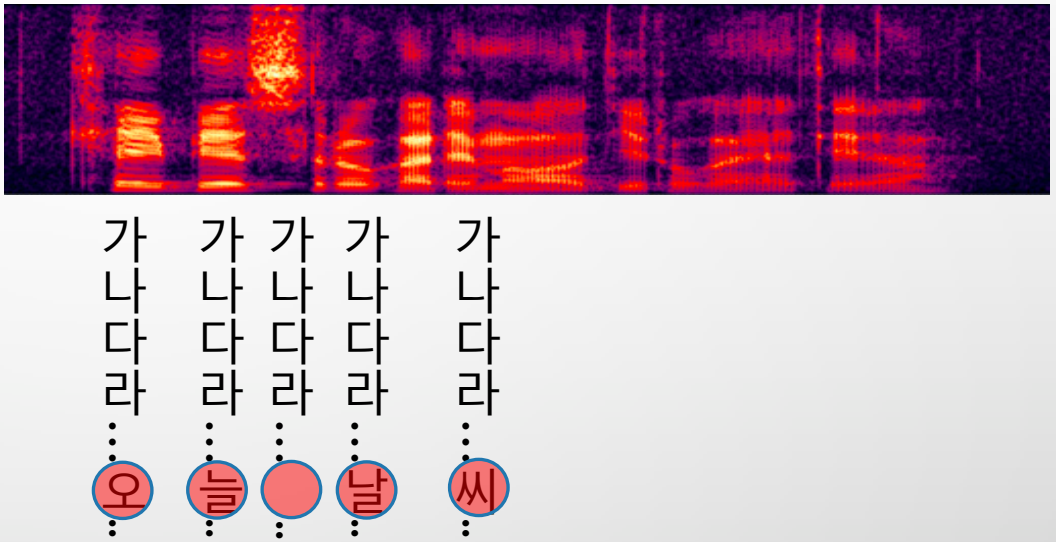
2012,
다음



2012~
ETRI

HOW IT WORKS

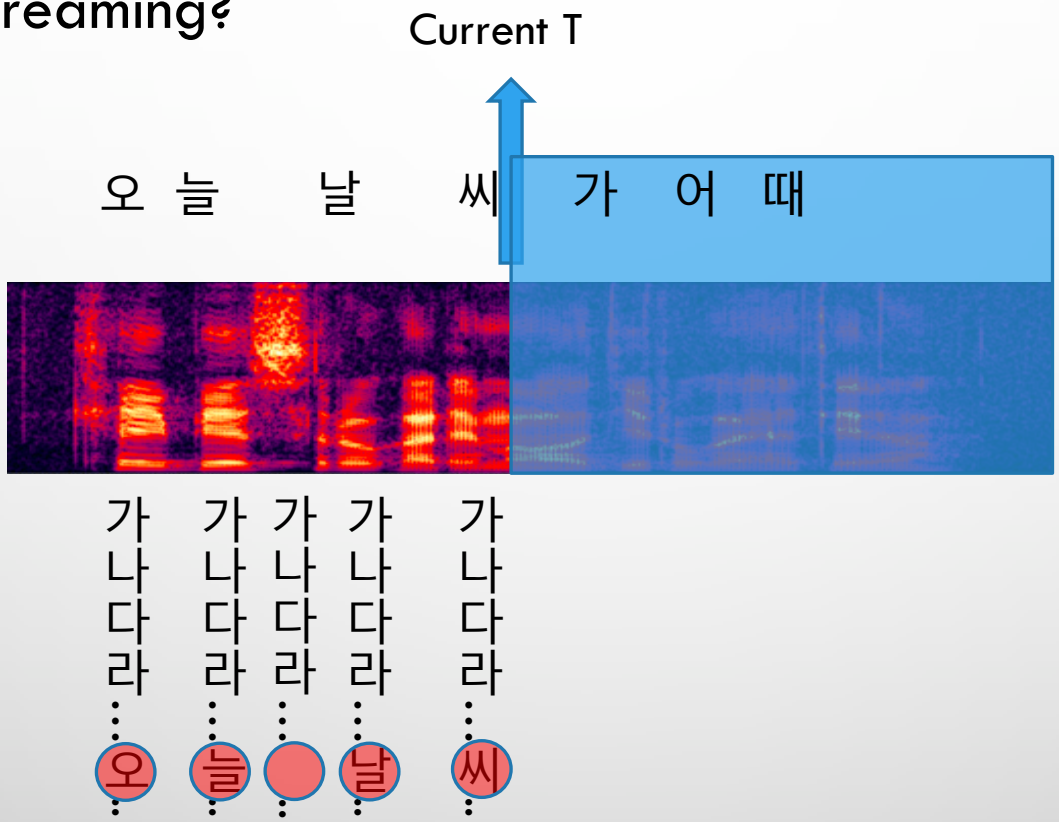
- W* = argmax P(W|X)
  - To Find Most Probable Word Sequence Given Input Signal/Feature



- Considerations
  - Boundary? Segmentation?
  - Output Units? Words, Characters, Phoneme, …
  - Classification Accuarcy? Unit Accuracy vs. Sentence Accuracy

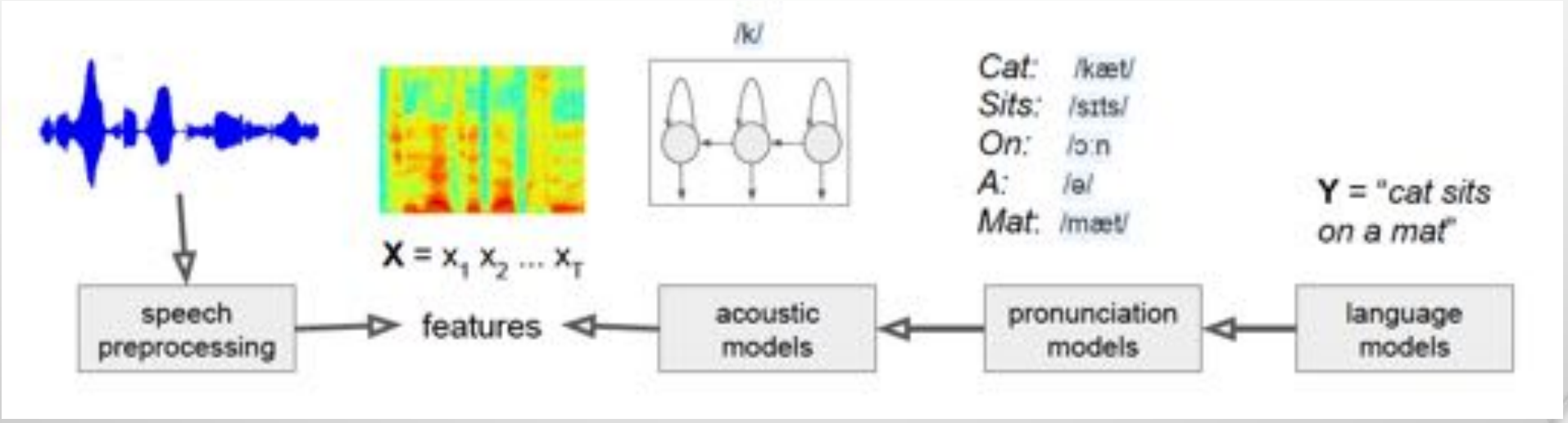
CONTEXT/LATENCY

Batch or Streaming?



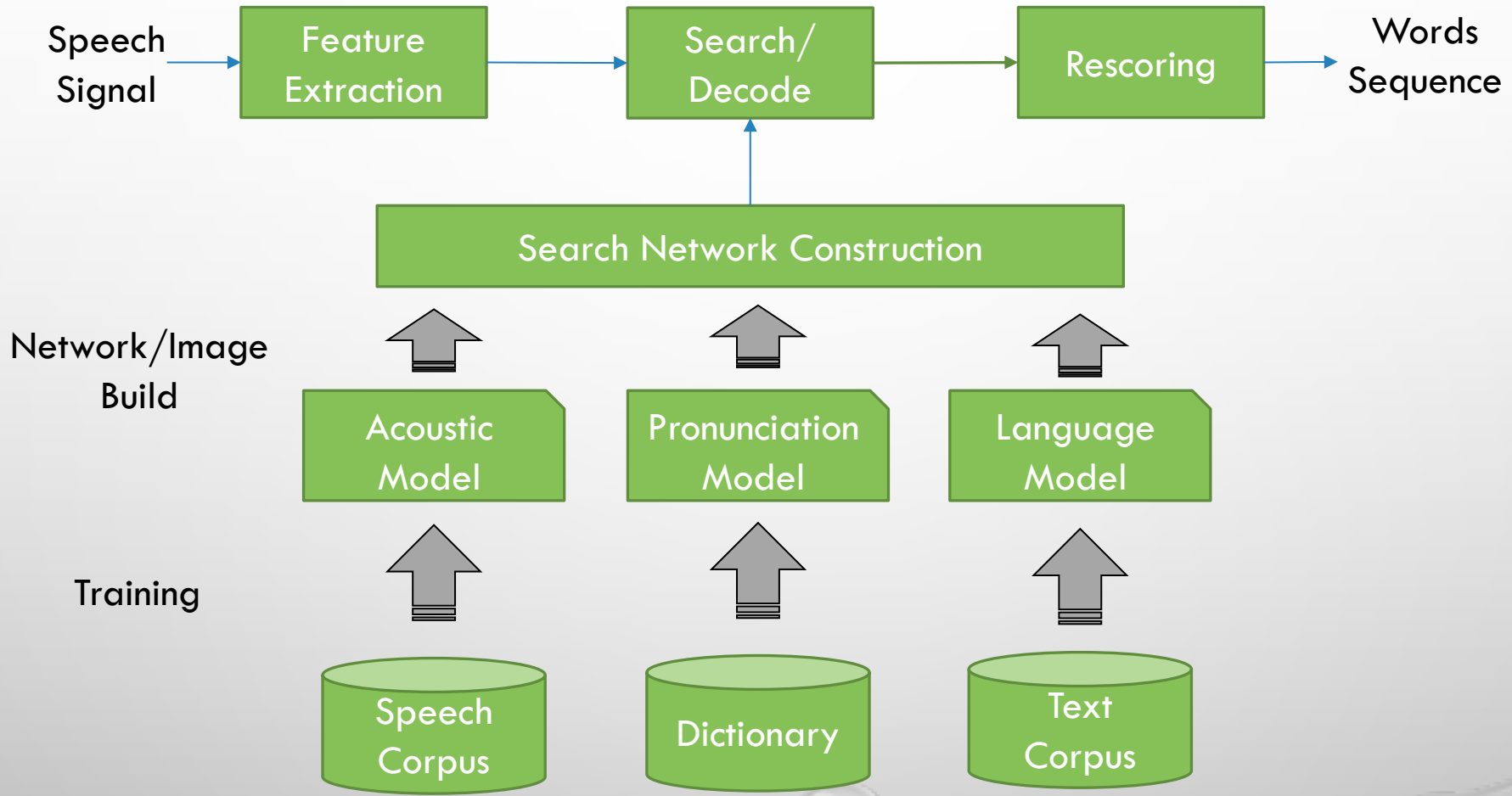
HOW IT REALLY WORKS

- W* = argmax log P(W|X)
- = argmax log P(X|Q)P(Q|W)P(W)
- To Find Most Probable Sequence Among Plausible Words Sequences



https://heartbeat.fritz.ai/the-3-deep-learning-frameworks-for-end-to-end-speech-recognition-that-power-your-devices-37b891ddc380

STRUCTURE OF TRADITIONAL ASR



EVALUATION METRIC

- Types of Error
  - Substitution
  - Deletion
  - Insertion
- Error Rate (cab be > 1)
  - (S + D + I)/N
- Accuracy (can be < 0)
  - 1-(Error Rate)
- WER/CER/SER:
  - Word/Character/Sentence Error Rate

REF : how is the weather today
REC/HYP: how was the better to day

In Words: WER = 100%, Acc=0%
- N= 5: how, is, the, weather, today
- S = 2
- D = 1
- I = 2

how is the weather today
how was the better to day

In Chars: CER = 25%, Acc=75%
- N= 20: h,o,w,i,s,t,h,e,w,e,a,t,h,e,r,t,o,d,a,y
- S = 3
- D = 1
- I = 1

how is the weather today
how was the better to day

In Sentence: SER = 100%, Acc=0%
- N= 1
- S = 1

QUIZ

- REF: 오늘 서울의 날씨가 어때
- REC: 음 오눌의 날씨 가 어때
- WER = ?

측정방법

- Edit distance 측정
  - https://en.wikipedia.org/wiki/Edit_distance
- 사용도구
  - HResults (HTK)
  - compute-wer (kaldi)
  - sclite (NIST, ESPnet)
- 예)
  - compute-wer ark:ref.txt ark:rec.txt

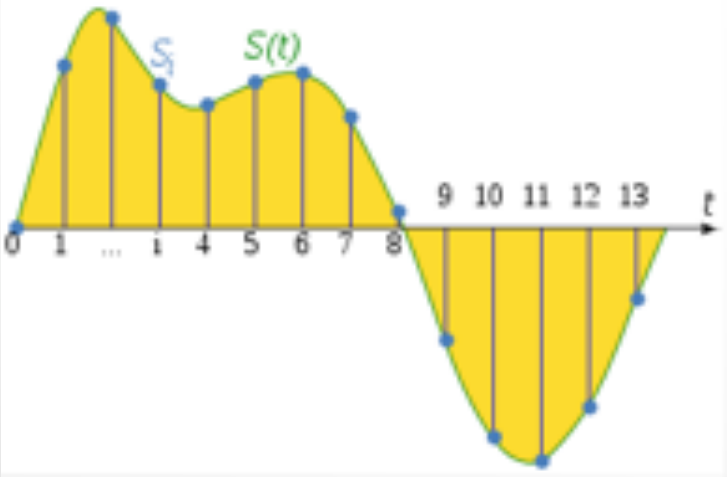
실습

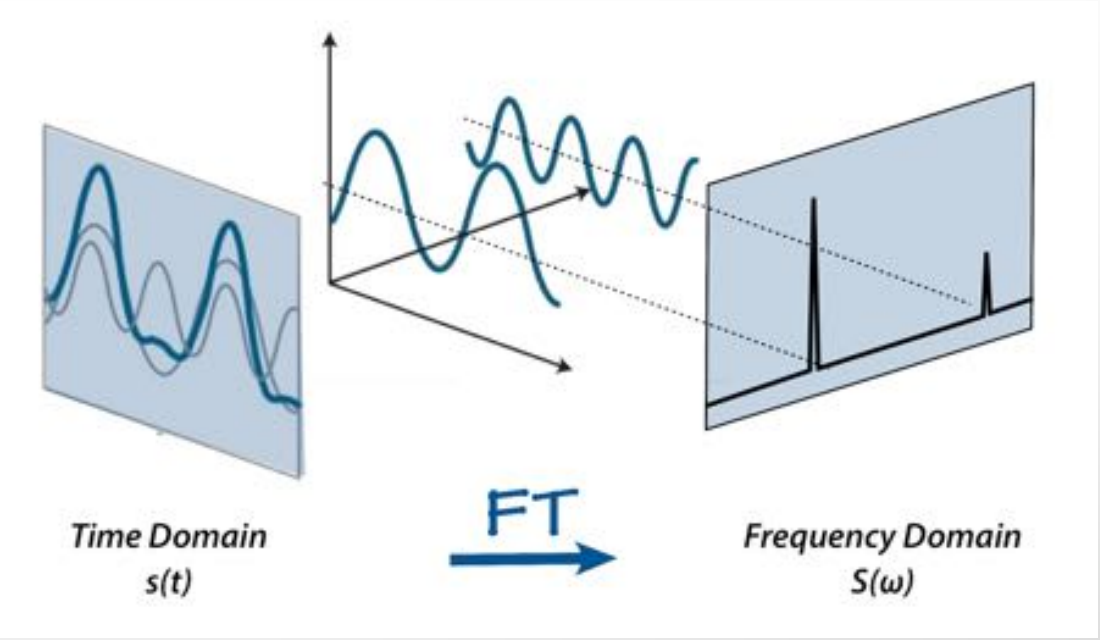
- 사용환경 설명/로그인/세션 생성
- VS Code/Jupyter/Python
- WER 측정

https://www.nvidia.com/ko-kr/data-center/dgx-a100/

FEATURE EXTRACTION

SAMPLING AND SPECTRUM



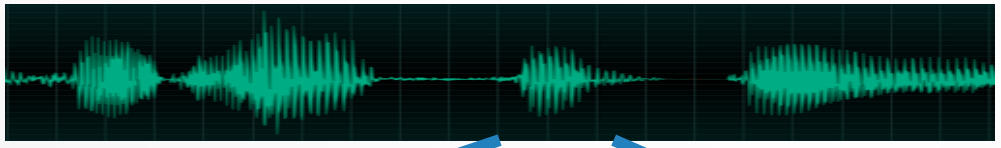


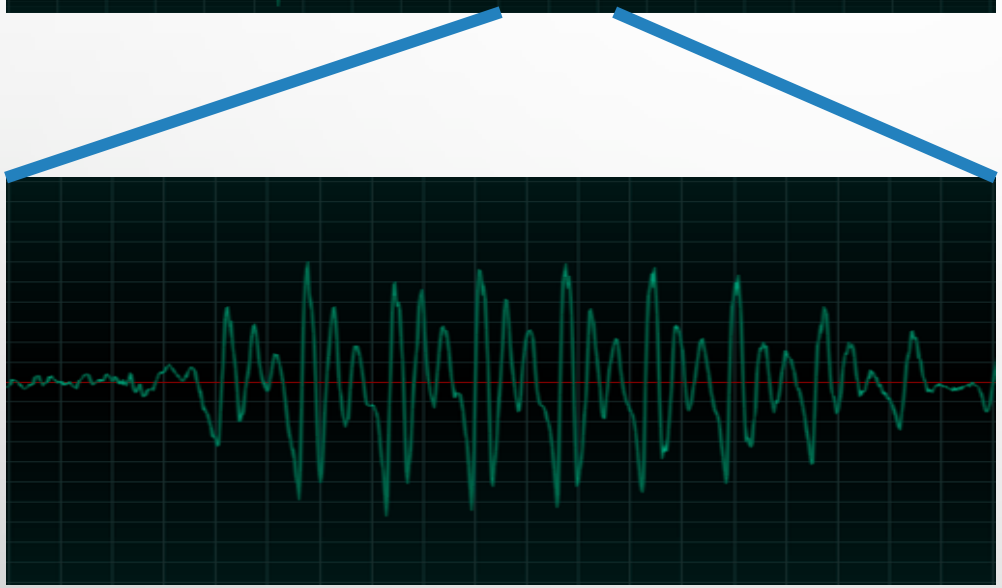
8kHz: Narrowband, 전화망
16kHz: Wideband
44.1/48kHz: High quality audio

https://en.wikipedia.org/wiki/Sampling_(signal_processing)
https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520

FRAME-WISE PROCESSING

1 sec





0.1 sec

Windowing

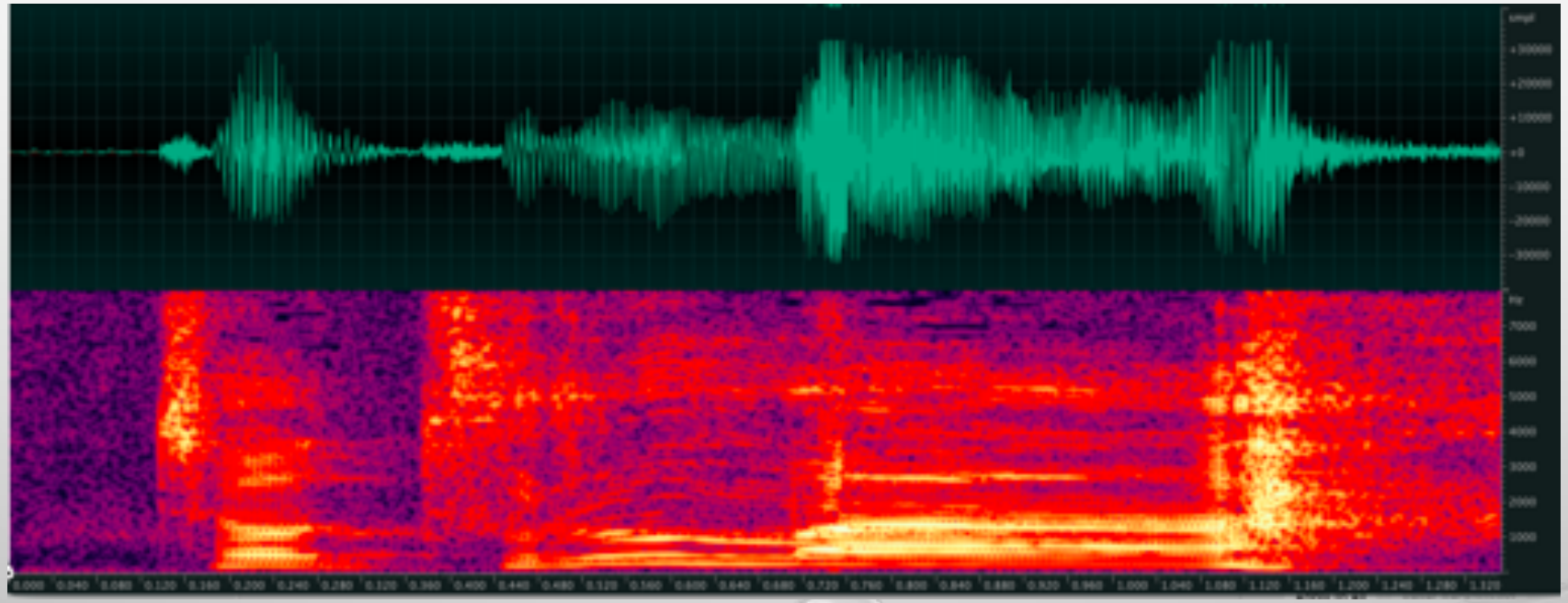
Window length,

FFT size,

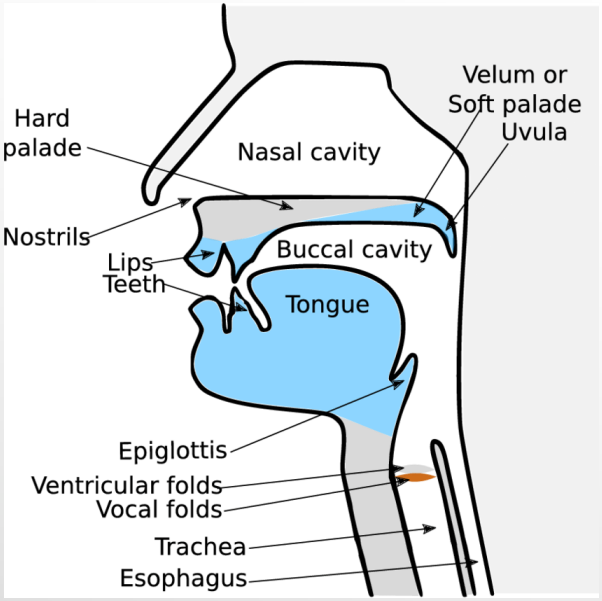
hop size

SPECTROGRAM

- Series Of Spectrum
- Matlab, Python, Adobe Audition, Audacity, …
- Frame Shift, Overlap, Window Length, Windowing, FFT points

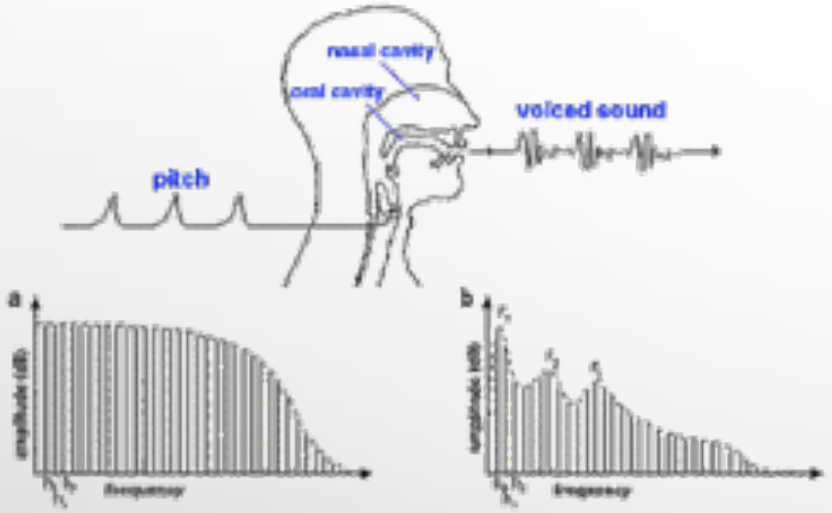


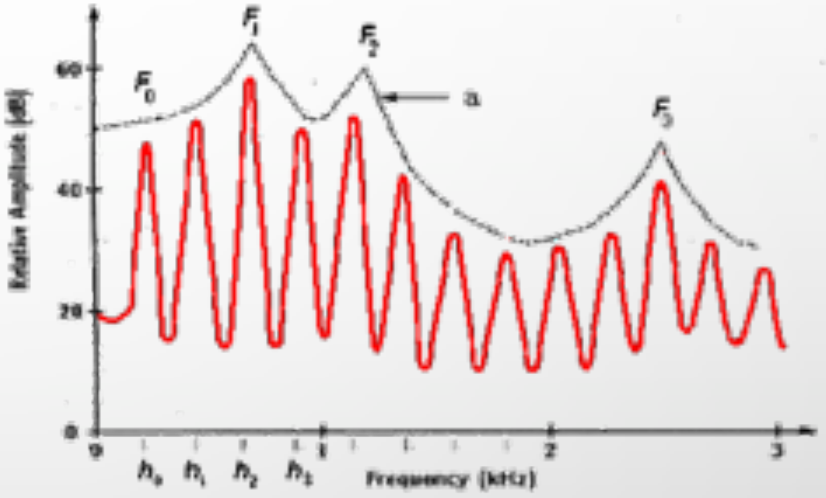
VOICE PRODUCTION



https://www.researchgate.net/publication/318814563_Analyzing_of_the_vocal_fold_dynamics_using_laryngeal_videos
https://www.youtube.com/watch?v=kfkFTw3sBXQ

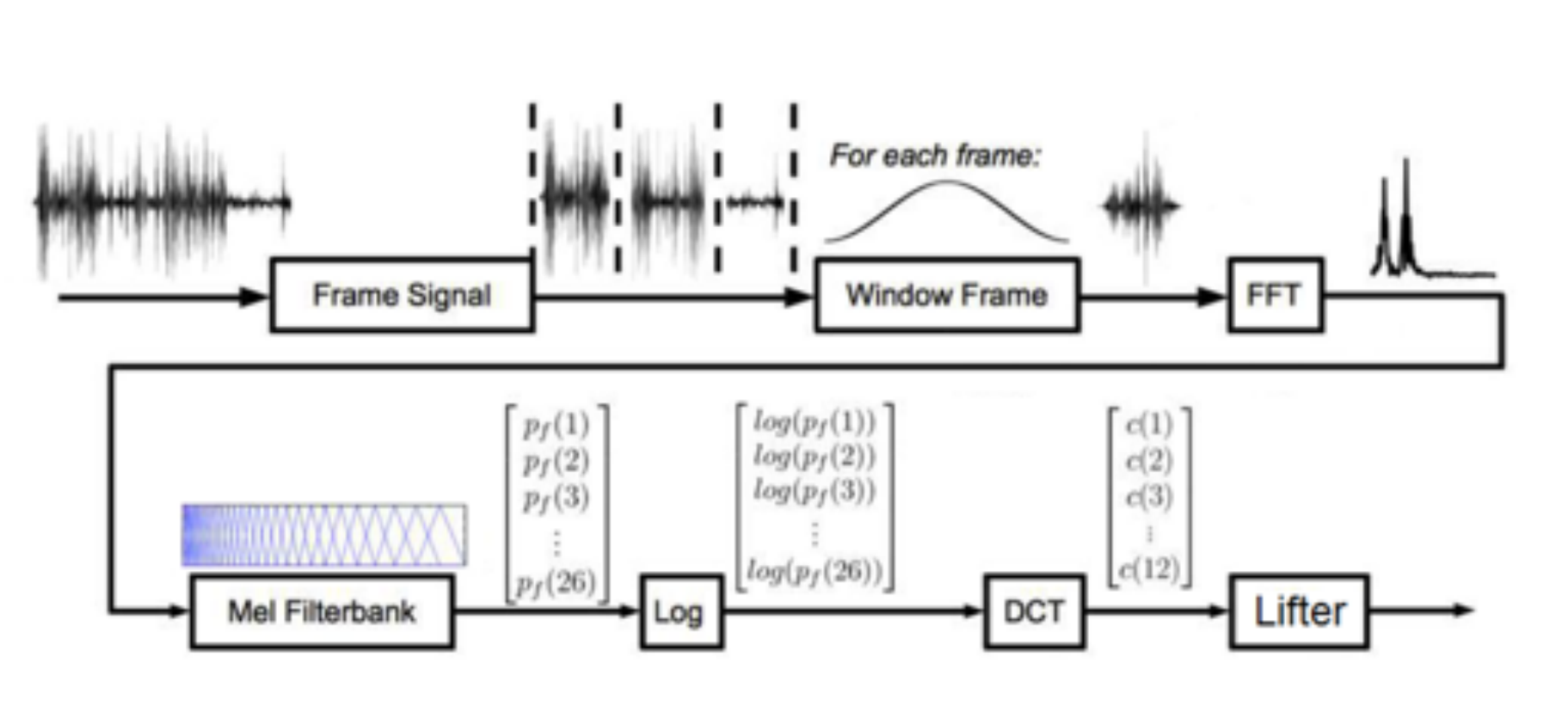
PITCH AND FORMANT





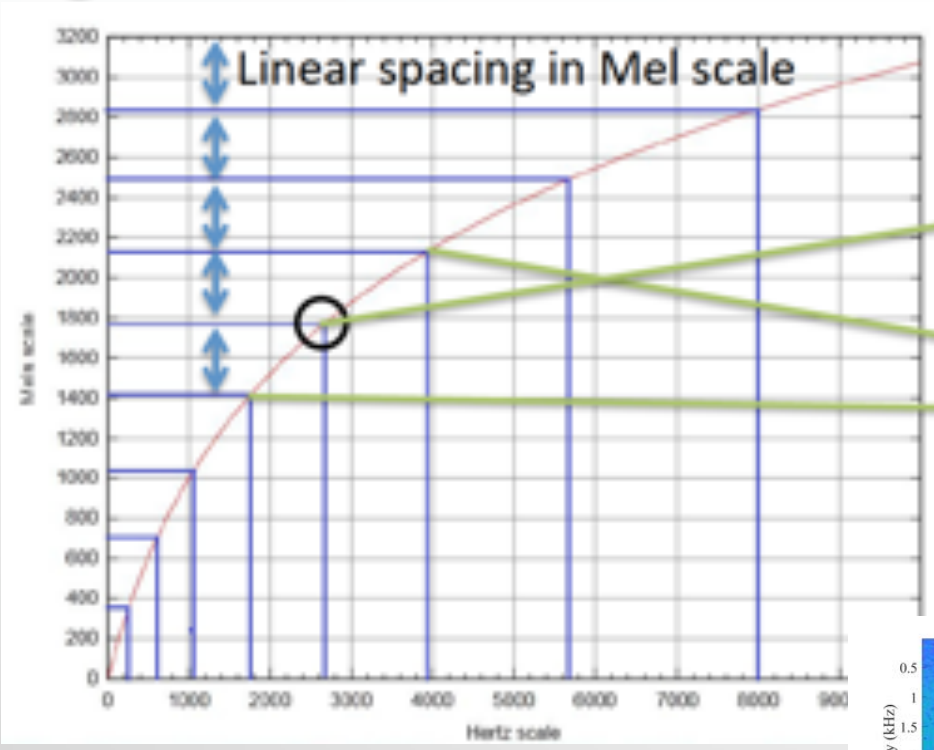
http://147.162.36.50/cochlea/cochleapages/theory/sndproc/sndcomm.htm

FEATURE EXTRACTION: MFCC



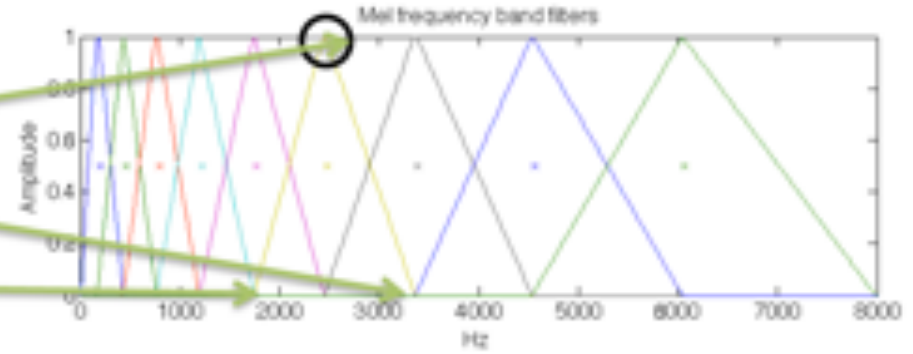
https://hyunlee103.tistory.com/46

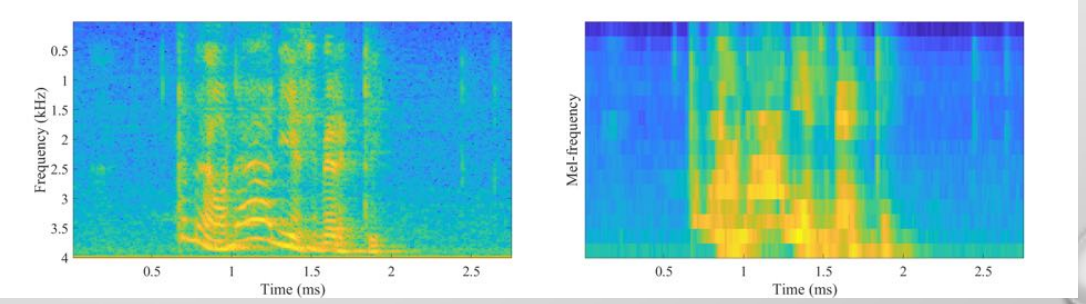
MEL FILTERBANK



https://hyunlee103.tistory.com/46

# of filters: 23 → 40 → 80+3





https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC

CMVN

- Cepstral Mean Variance Normalization
  - Zero-mean Unit Variance
- CMS: Cepstral Mean Substraction
  - Per Utterance
  - 채널/화자 효과를 제거하고 발성의 특성만 남김
- For Deep Learning
  - Global CMVN
  - For better convergence

HOMEWORK

- 음성인식 평가용 테스트데이터 수집
- 본인 (또는 근처 아무나) 목소리를 녹음
  - 16kHz, wav (uncompressed), mono
  - (일단 녹음하고 확인해봅시다)
- 인식률이 되도록 좋게 or 나쁘게 나오도록
  - But don't be too evil… noise, yell, whisper…
- 10문장 정도, 1문장당 10초 정도.
- wav/text pair (wav.scp, text)
- Due: 3일차 시작 전까지

강의내용: 2일차

- Classical ASR
- Introduction to End-to-End ASR
- 실습
  - ESPNet 소개
  - 종단형 음성인식 recipe 살펴보기
- Transformer
- 실습
  - 한국어 1,000시간 훈련 DB를 이용한 훈련 시작

CLASSICAL ASR

HMM-BASED ASR

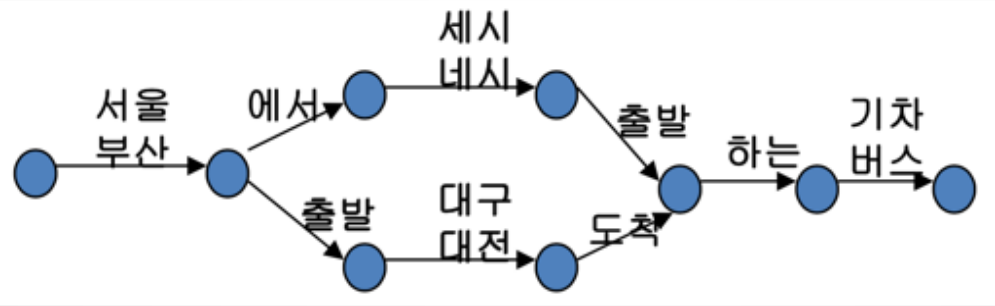
- How we call it?
  - Conventional
  - Traditional/Classical
  - Ancient
- Why?
  - 내부 동작을 이해하고 문제점 또는 성능 개선 방법을 찾기 위해서

ERA OF HIDDEN MARKOV MODEL

- Problem to Solve:
- $W^* = \text{argmax} \log P(W|X)$
- $= \text{argmax} \log P(X|Q)P(Q|W)P(W)$
- P(W): Language Model, P(Wt|Wt-1, Wt-2, …)
- P(Q|Wt): Prononciation Model
- P(X|Q): Acoustic Model

LANGUAGE MODEL

- 단어간의 연결 가능성을 이용하여 search space를 제한



- Deterministic Grammar
  - FSN (Finite State Network)
  - JSGF (Java Speech Grammar Format)

```
$time = 세시|네시;
$city = 서울|부산|대구|대전;
$trans = 기차|버스;
sent-start $city (에서 $time 출발 |
출발 $city 도착) 하는 $trans sent-end
```

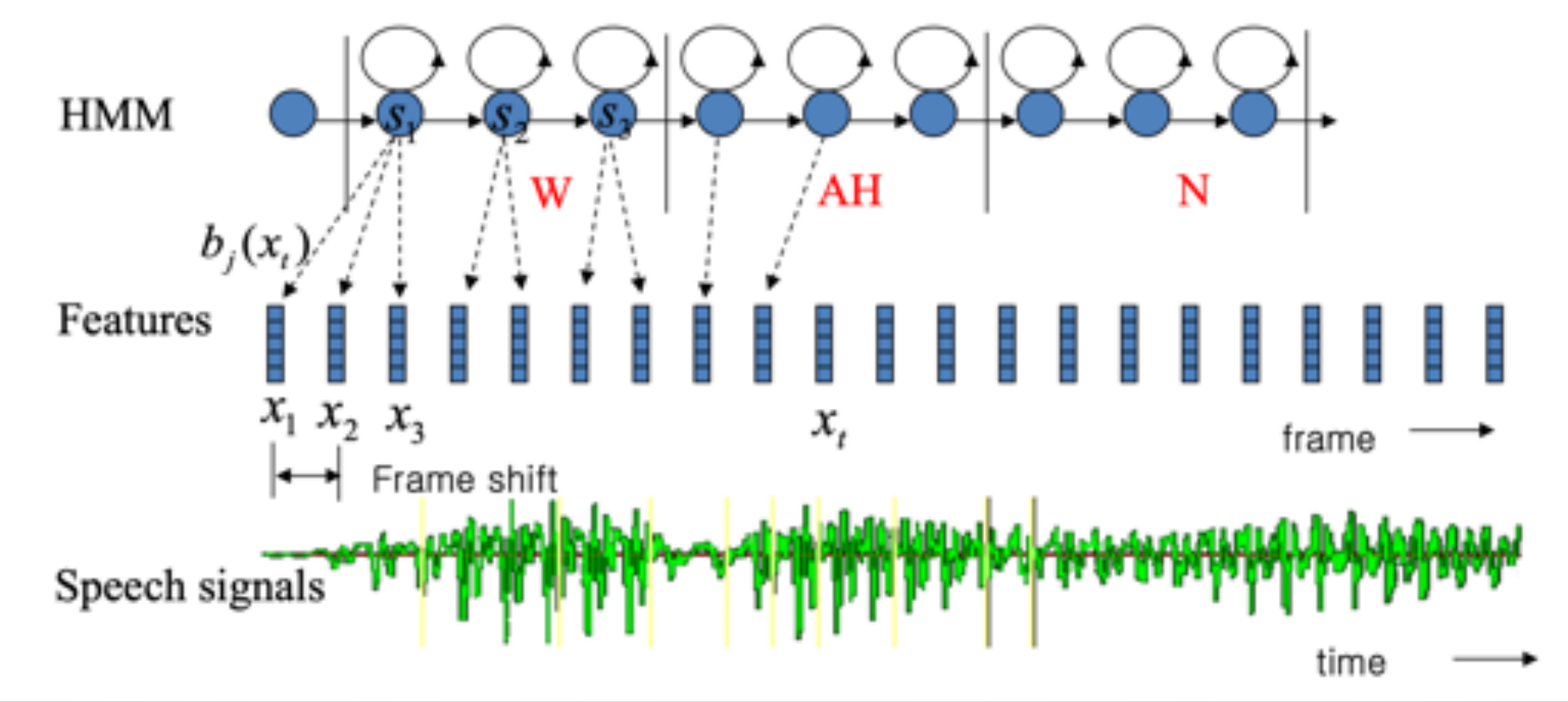
- Stochastic Grammar
  - N-gram

P(에서|서울)=0.2 P(세시|에서)=0.5
P(출발|세시)=1.0 P(하는|출발)=0.5
P(출발|서울)=0.5 P(도착|대구)=0.9
…

PRONONCIATION MODEL

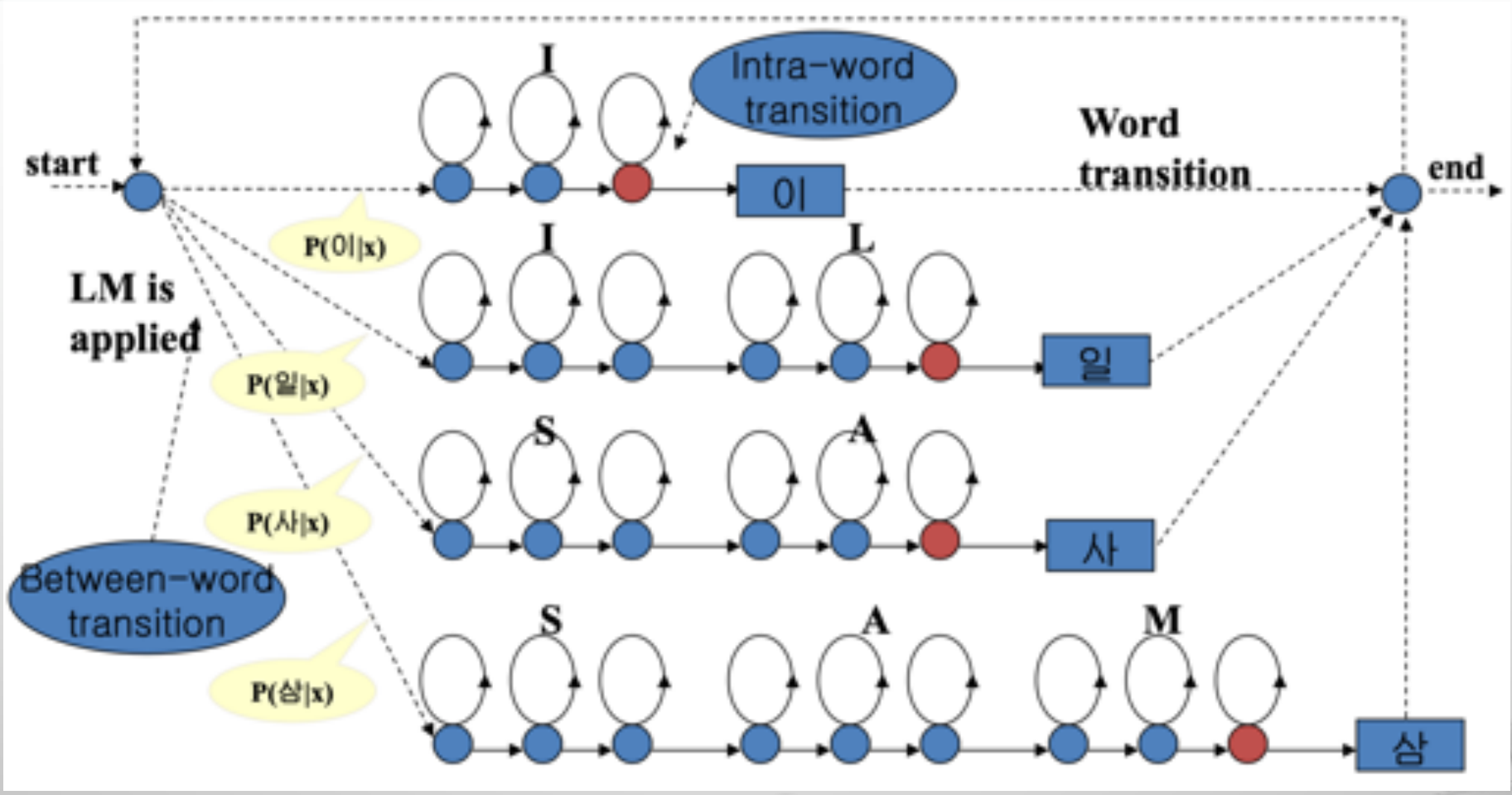
- How a word is pronunciated
- Very Langauge-Dependent and Requires Expert Knowledge
  - 대한민국: /d E h a xn m i xn g u xg/
  - 2NE1, 야탑역, 맨유
- Phoneset
  - 한국어: ETRI 46 phoneset
  - 영어: CMUDict(48), TIMIT(61) → CMU 39 phoneset
- Rule-based, Statistical Approach, Neural Approach

ACOUSTIC MODEL



http://speech.cbnu.ac.kr/srhome/technology/index.html

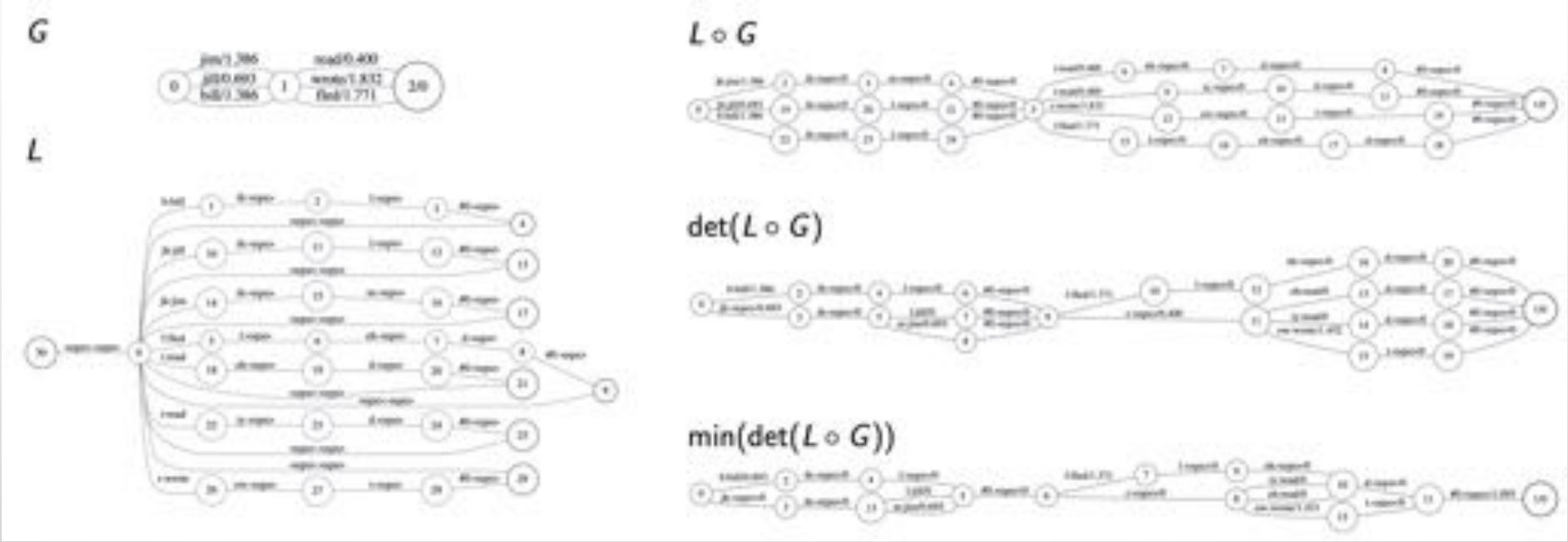
SEARCH NETWORK



http://speech.cbnu.ac.kr/srhome/technology/index.html

SEARCH NETWORK: wFST

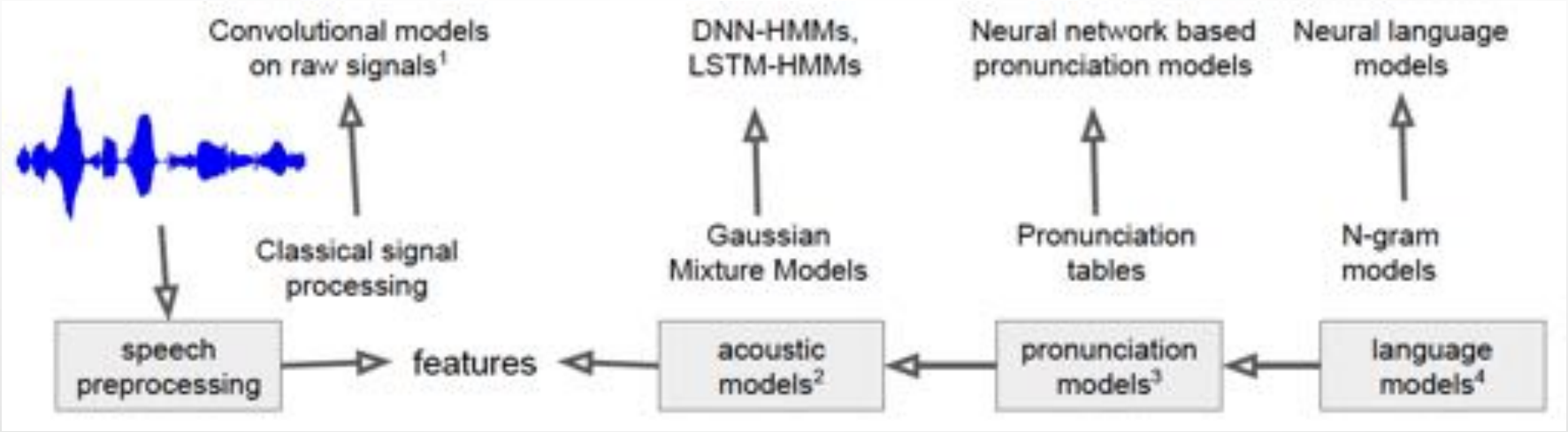
- Weighted Finite Statue Transducer



https://medium.com/@jonathan_hui/speech-recognition-weighted-finite-state-transducers-wfst-a4ece08a89b7

DEEP LEARNING FOR ASR

DEEP LEARNING FOR ASR



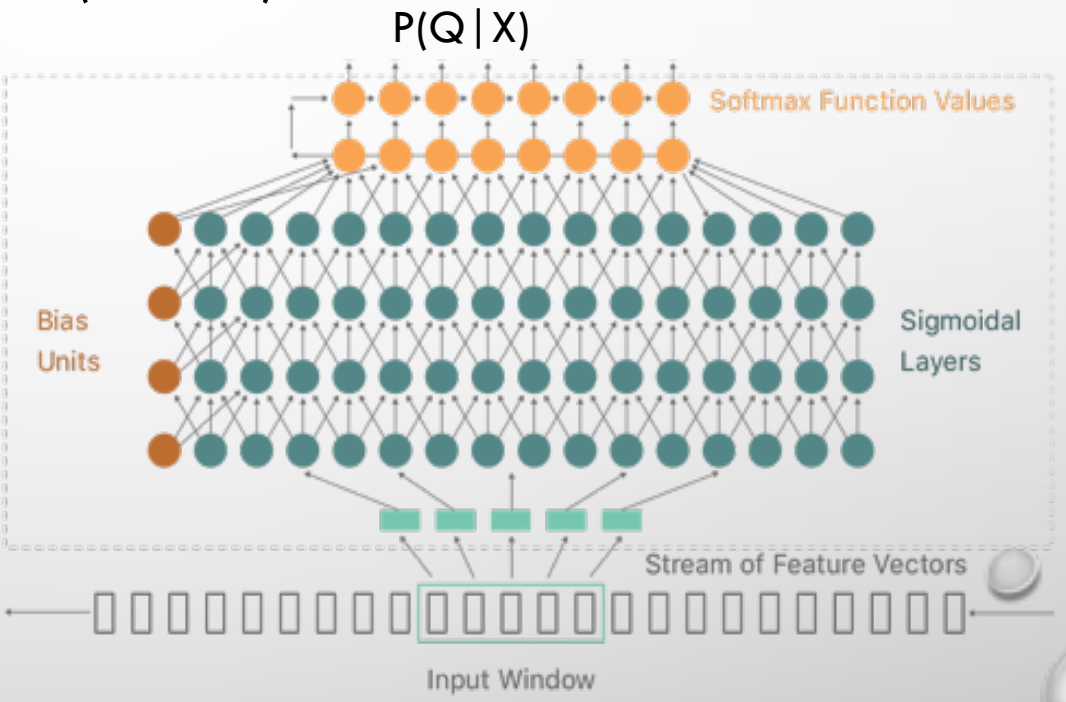
https://heartbeat.fritz.ai/the-3-deep-learning-frameworks-for-end-to-end-speech-recognition-that-power-your-devices-37b891ddc380

DNN-HMM (1)

- Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, 2012, IEEE Trans. on Audio, Speech and Language Processing, Microsoft
- Large Vocabulary Continuous Speech Recognition With Context-dependent DBN-HMMS, 2011, ICASSP, Microsoft
- Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, 2012, Hinton et al.

DNN-HMM (2)

- $P(X|Q) = P(Q|X)P(X)/P(Q)$
- Output Units: Senone, States of HMM(5k~20k)
- FC-DNN
- CNN
- RNN: GRU, LSTM, Bi-LSTM
- TDNN
- Longer Context Helps



TRAINING OF DNN-HMM

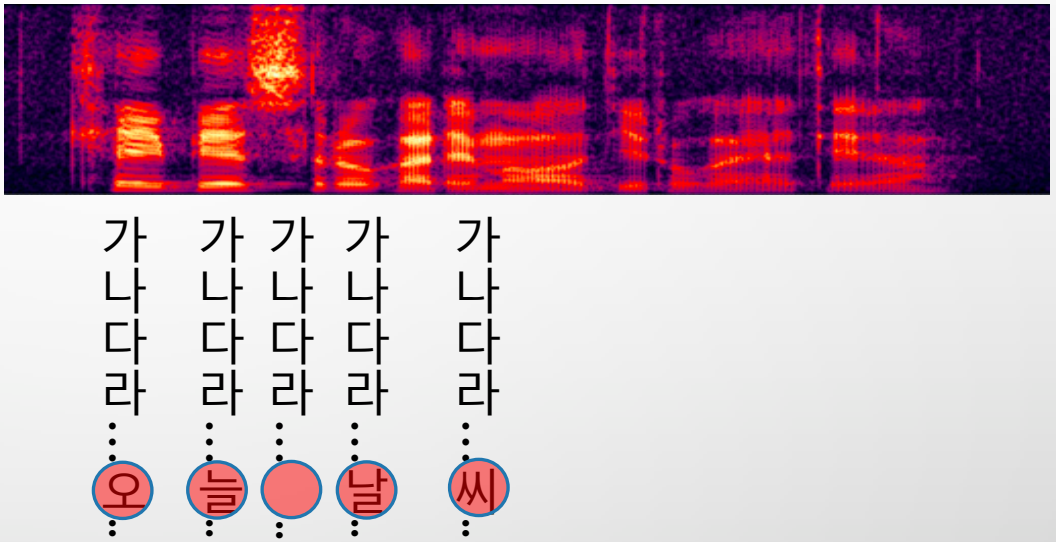
- Requires Frame-wise Label
- Forced-Alignment using Seed Model (Usually GMM-HMM Model)
  - Speech/Text Pair → g2p → State level alignment

- Kaldi Toolkit (2009~, JHU)
  - https://github.com/kaldi-asr/kaldi
- HTK Toolkit (1989~, Cambridge)
  - http://htk.eng.cam.ac.uk/
  - https://github.com/open-speech/HTK

END-TO-END ASR

HOW IT WORKS: REVISITED

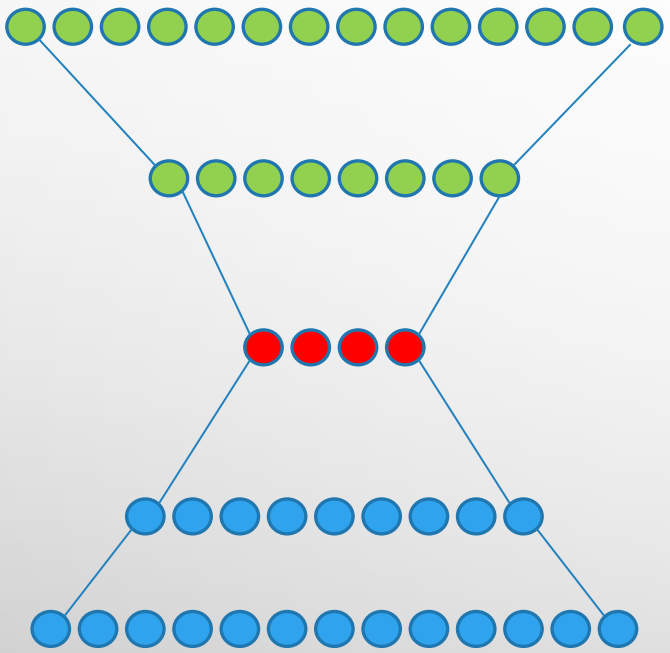
- W* = argmax P(W|X)
  - To Find Most Probable Word Sequence Given Input Signal/Feature



- Considerations
  - Boundary? Segmentation?
  - Output Units? Words, Characters, Phoneme, …
  - Classification Accuarcy? Unit Accuracy vs. Sentence Accuracy

ENCODER-DECODER

- Auto Encoder



Y: Output

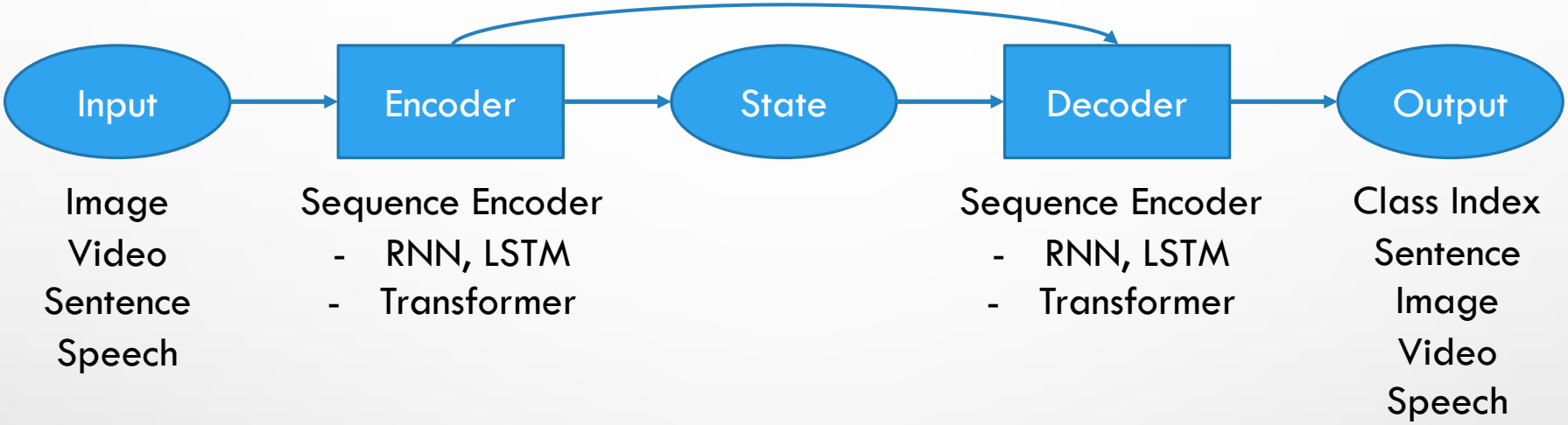
Decoder

Z: Latent Variable, Hidden Representation
Context

Encoder

X: Input

ENCODER-DECODER FOR SEQUENCE

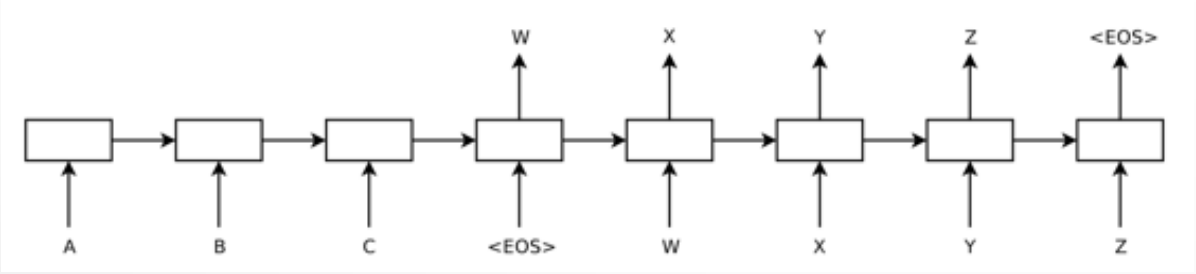


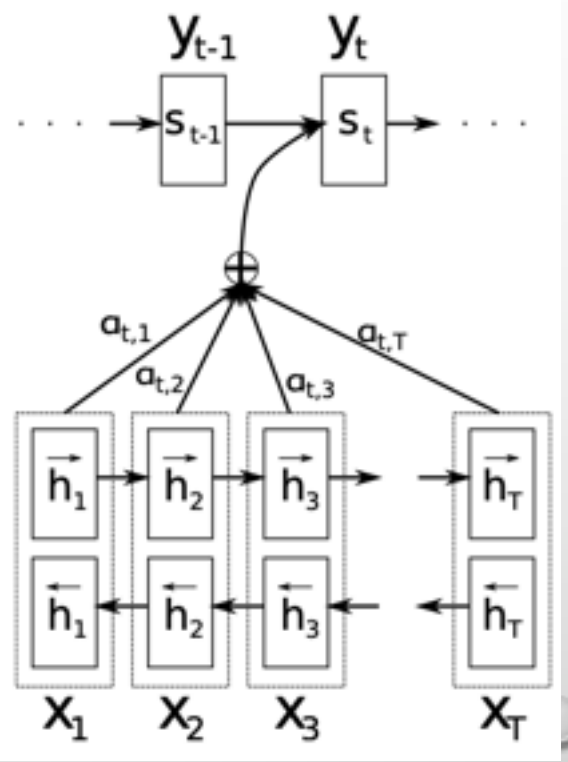
- Translation
- Image/Video Captioning
- Q&A, Document Summarization
- Speech
  - Recognition, Synthesis, Translation, Dialog System(Google Duplex, 2018)

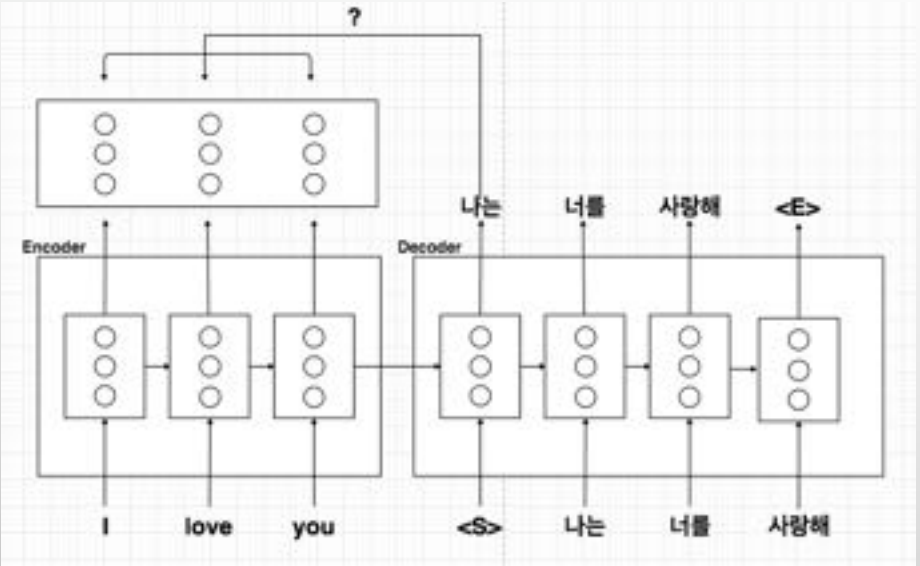
ERA OF SEQUENCE-TO-SEQUENCE

- Natural Language Processing
- Sequence to Sequence Leanring with Neural Networks, NeurIPS, 2015
- Neural Machine Translation By Jointly Learning To Align And Translate, ICLR, 2016
- Attention Is All You Need, NuerIPS, 2017
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ACL, 2019

SEQUENCE TO SEQUENCE WITH ATTENTION





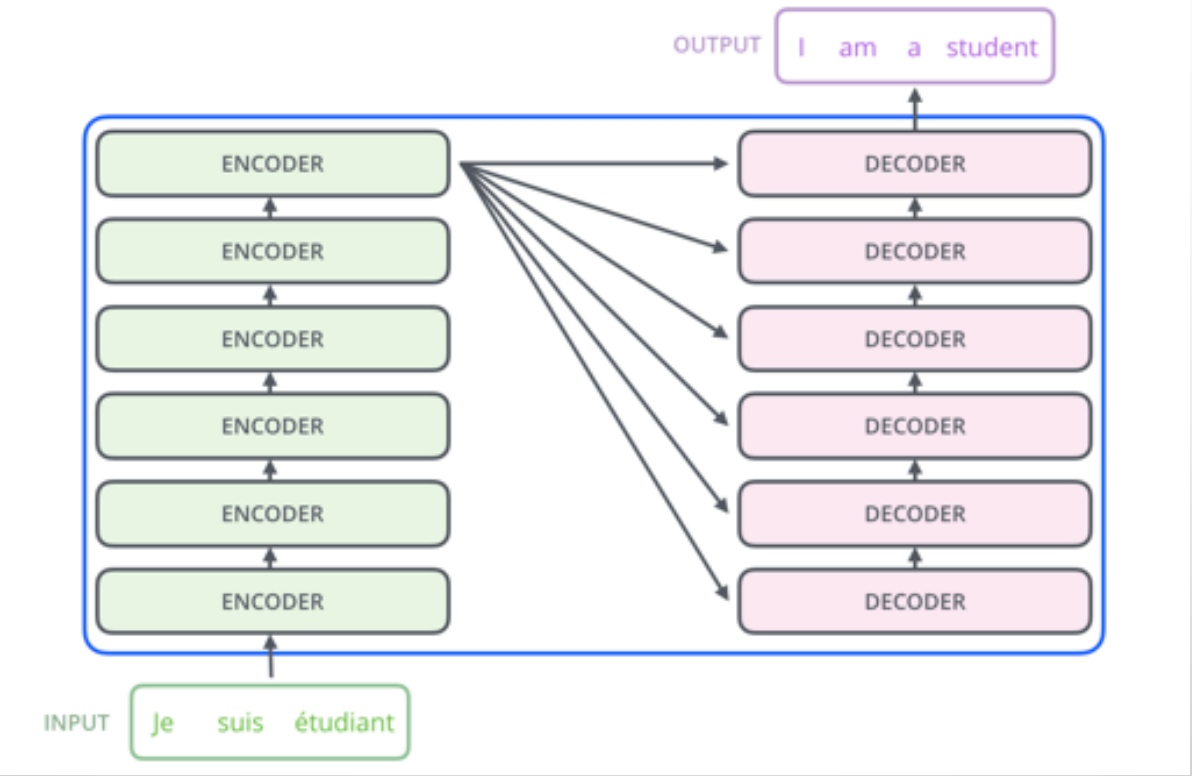


https://medium.com/platfarm어텐션-메커니즘과-transfomer-self-attention-842498fd3225

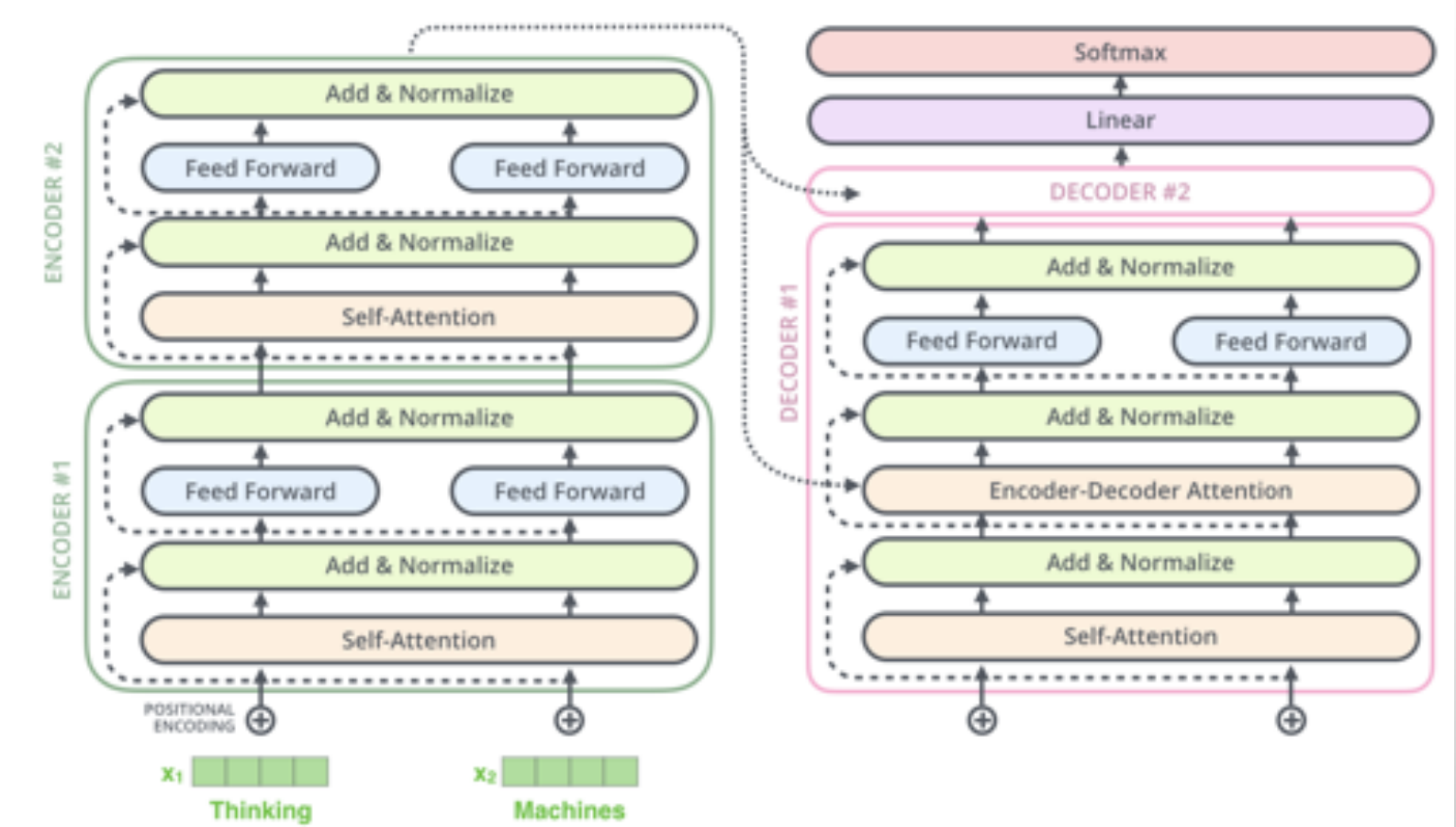
TRANSFORMER

TRANSFORMER: OVERALL STRUCTURE

- https://jalammar.github.io/illustrated-transformer/

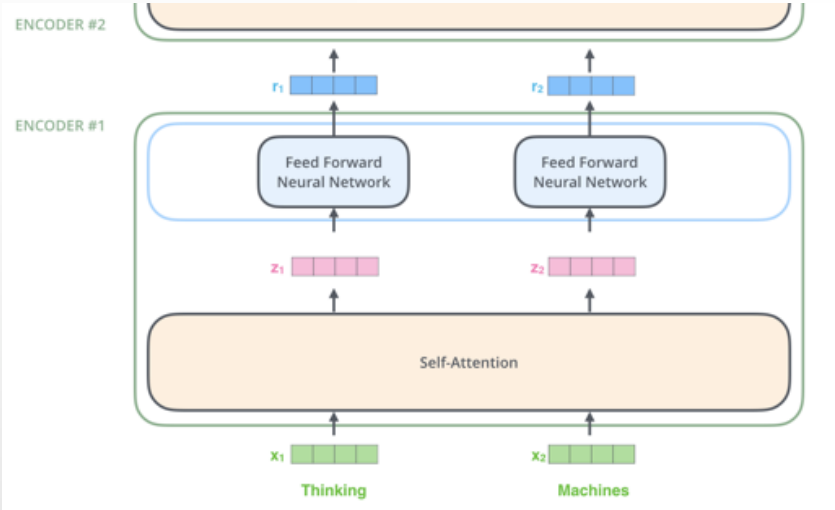


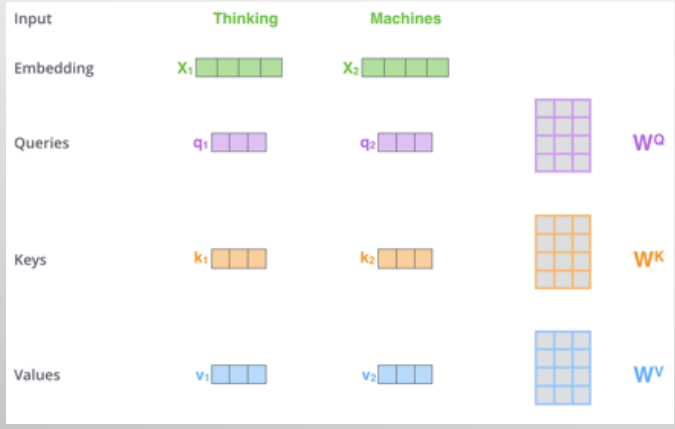
TRANSFOMER: DETAILED STRUCTURE

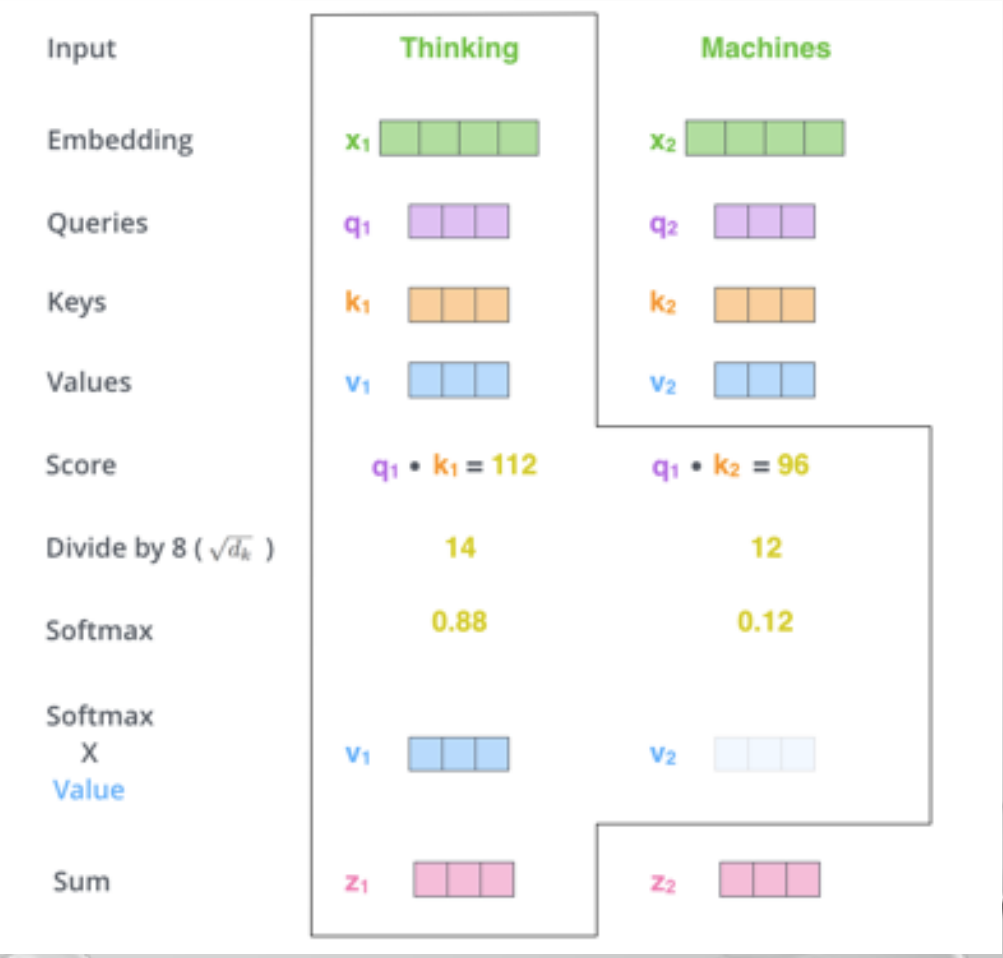


https://jalammar.github.io/illustrated-transformer/

SELF ATTENTION

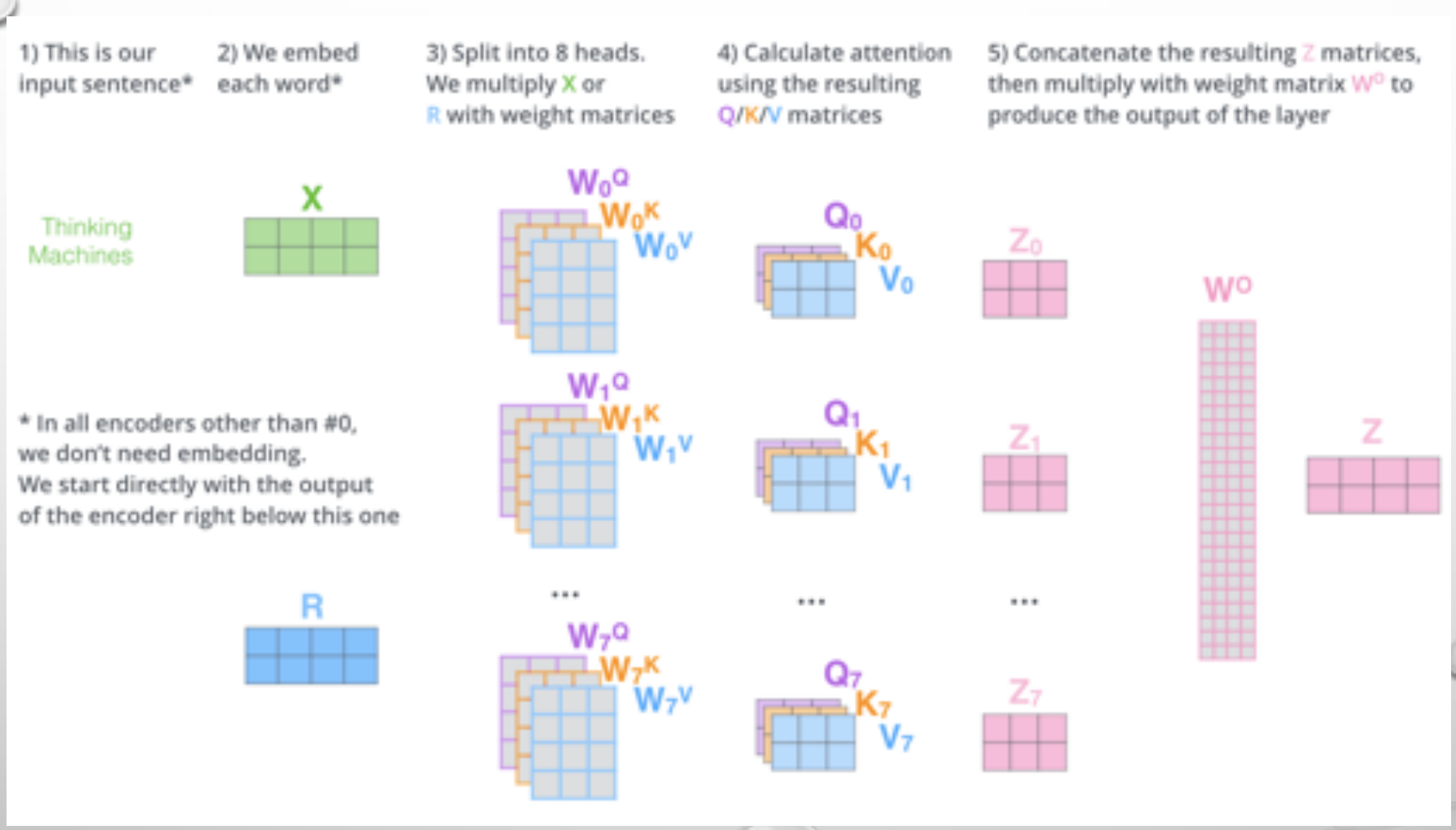






https://jalammar.github.io/illustrated-transformer/

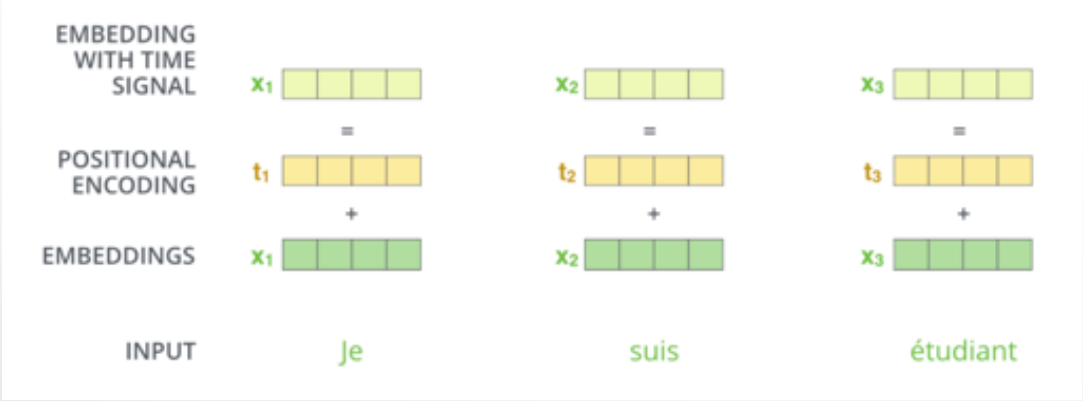
MULTIHEAD ATTENTION



https://jalammar.github.io/illustrated-transformer/

POSITIONAL ENCODING

- No Position Depenent Computation in Transfomer



- Absolute/Relative Position Encoding
  - Sinusoidal Positional Encoding

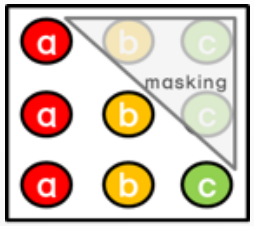
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

```
0 :  0 0 0 0
1 :  0 0 0 1
2 :  0 0 1 0
3 :  0 0 1 1
4 :  0 1 0 0
5 :  0 1 0 1
6 :  0 1 1 0
7 :  0 1 1 1
```

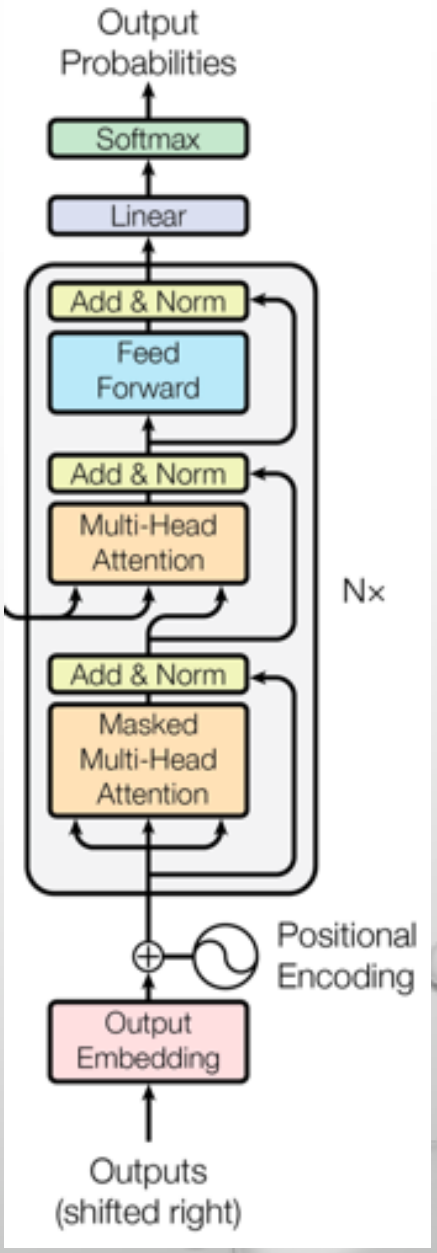
https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

DECODER

- Masked Multi-Head Self Attention

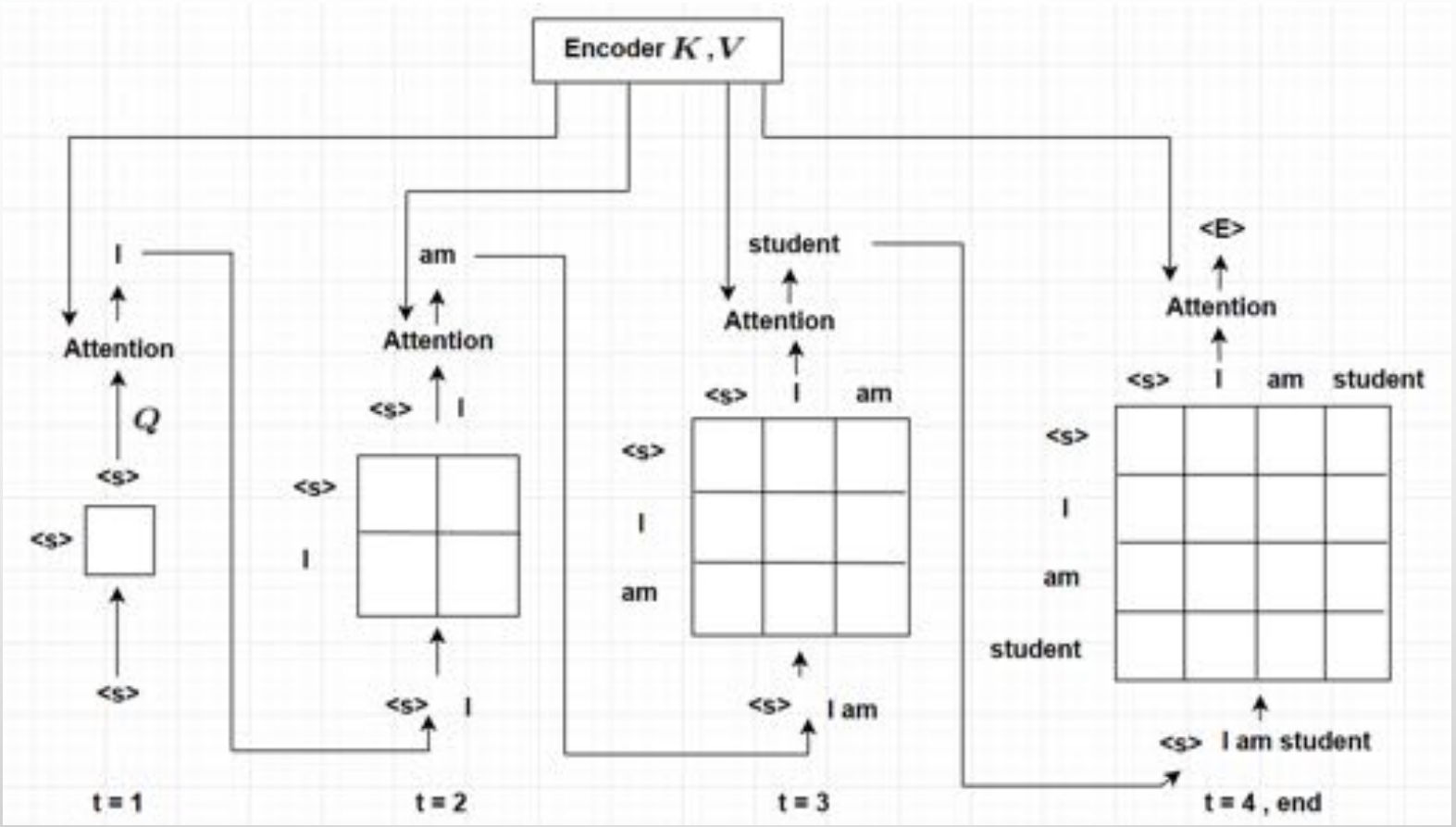


- Encoder-Decoder Attention
  - K, V from Encoder Last Layer
  - Q from Self Attention
- Beam Search



Attention Is All You Need, NuerIPS, 2017

DECODER IN ACTION



https://medium.com/platfarm/어텐션-메커니즘과-transfomer-self-attention-842498fd3225

강의내용: 3일차

- End-to-End ASR In Practice
- ASR 성능 개선 방안
- 실습
  - 훈련된 모델의 평가(공통평가셋)
  - 개인별 평가셋을 이용한 평가
- Recent trends in ASR
- 실습
  - Wrapup and Backup
  - Q&A

END-TO-END FOR ASR

- ESPNet: End-to-end Speech Processing Toolkit
  - ASR, TTS, Speech Translation
  - https://github.com/espnet/espnet
- CTC Hybrid*: Connectionist Temporal Classification
  - Multi-task training with CTC Criteiron
  - Increase Stability while Training

* Hybrid CTC/Attention Architecture for End-to-End Speech Recognition, IEEE Journal of Selected Topics in Signal Processing, 2018

END-TO-END ASR IN PRACTICE

- Output Units
  - 영어: Alphabet, BPE(Byte Pair Encoding), Word
  - 한국어: Char(음절~2500), BPE(~5000), 형태소분석기
- Relative Performance
  - WER/CER
  - For eg. 25% (GMM-HMM) → 15% (DNN-HMM) → 10% (LSTM-HMM)
  - *7*% Transformer
- Limitation
  - Process Whole Sentence → Streaming ASR

성능개선방안

- 데이터!
  - 실환경 데이터수집: 적응훈련/연결학습
  - 음향모델/언어모델?
- 데이터!!
  - 데이터 증강
  - SpecAug, Speed/Volume perturbation, Noise addition, Simulated data
- 모델 파라미터
  - Number of epoch
  - Number of paramerters: layers, dimension etc
  - Gradient scale: batchsize, learning rate etc
  - Robustness: dropout rate,

ONGOING RESEARCHES

- Semi/Un-Supervised Training
  - Training without labeled data
  - wav2vec, …
- Data augmentation
  - Generative models
- Transfer learning
  - Domain transfer
- Domain adaptation
- Streaming Transformer

DISCUSSION
and
Q&A