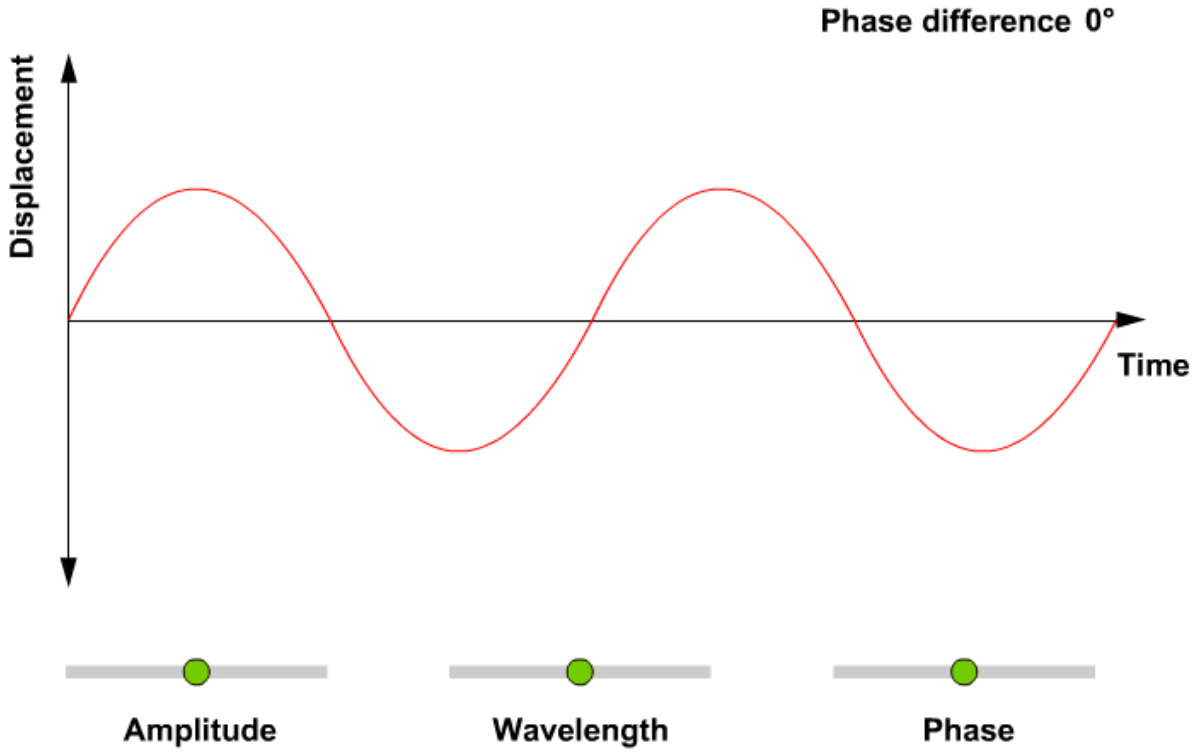

인공지능
Artificial Intelligence


08
오디오 음성 처리

1 음성 인식

오디오 처리(Audio Processing)

- 소리는 진동으로 인한 공기의 압축으로 생성
- 압축이 얼마나 됐느냐에 따라 진동하며, 공간이나 매질을 전파해 나가는 현상인 Wave(파동)으로 표현
- 파동에서 얻을 수 있는 정보
  - 위상(Phase; Degrees of displacement)
  - 진폭(Amplitude; Intensity)
  - 주파수(Frequency)



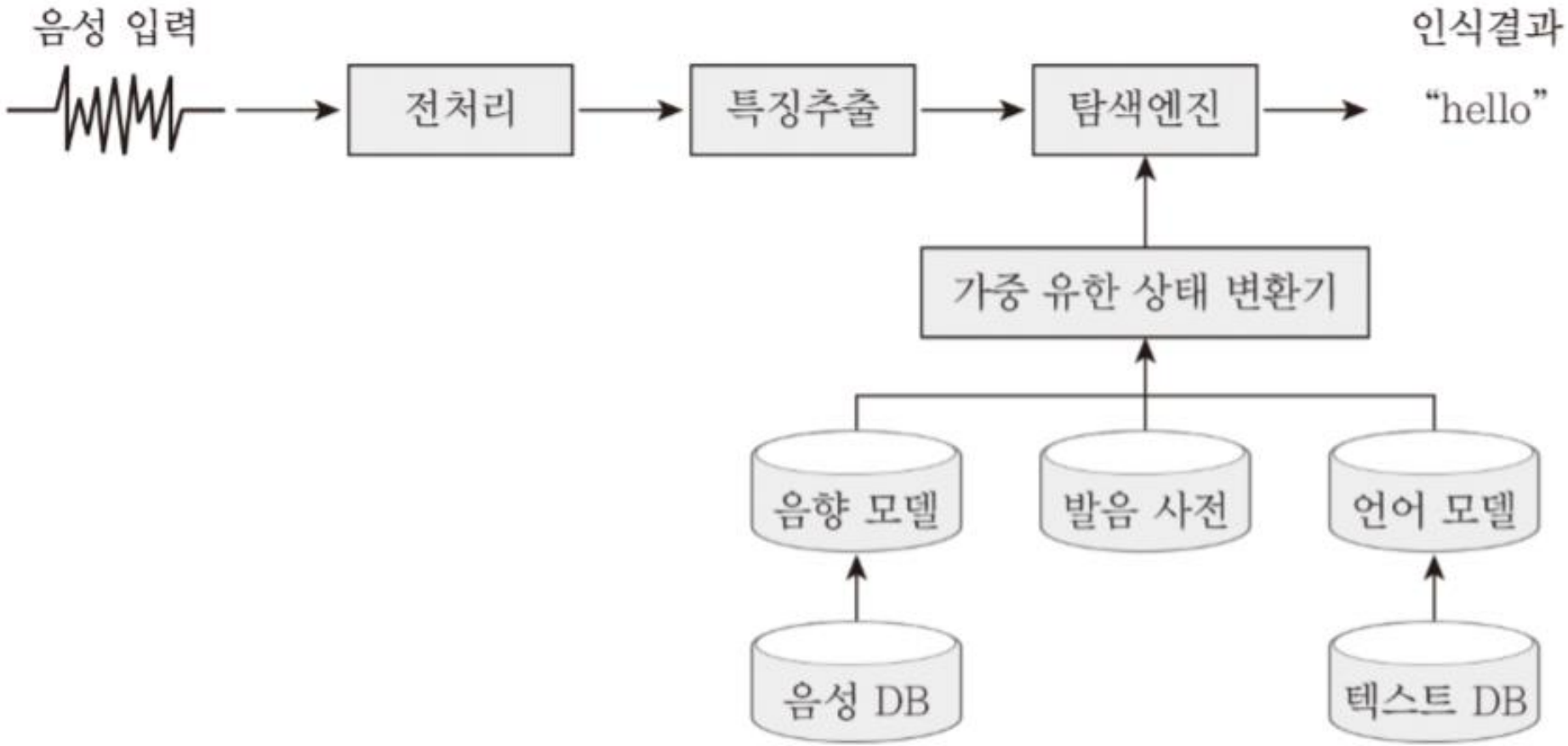
샘플링(Sampling)

- 음성을 처리하기 위해 아날로그 정보를 잘게 쪼개 이산적인 디지털 정보로 표현해야 함
- 이때 무한히 쪼개서 저장할 수는 없으므로, 기준을 세워 아날로그 정보를 쪼개 대표값을 사용, 이를 샘플링이라 함
- 주로 사용할 때 시간을 기준으로 아날로그 정보를 쪼개는 Time Domain 방식을 사용

- Sampling rate
  - sampling rate는 아날로그 정보를 얼마나 잘게 쪼갤지를 결정
  - 잘게 쪼갤수록 정보 손실이 줄어들지만, 데이터의 크기가 늘어남
- Sampling theorem
  - sampling rate가 최대 frequency보다 2배 커져야 함을 의미
  - 일반적으로 sampling은 인간의 청각 영역에 맞게 형성
    - Audio CD : 44.1 kHz(44100 sample/second)
    - Speech communication : 8 kHz(8000 sample/second)

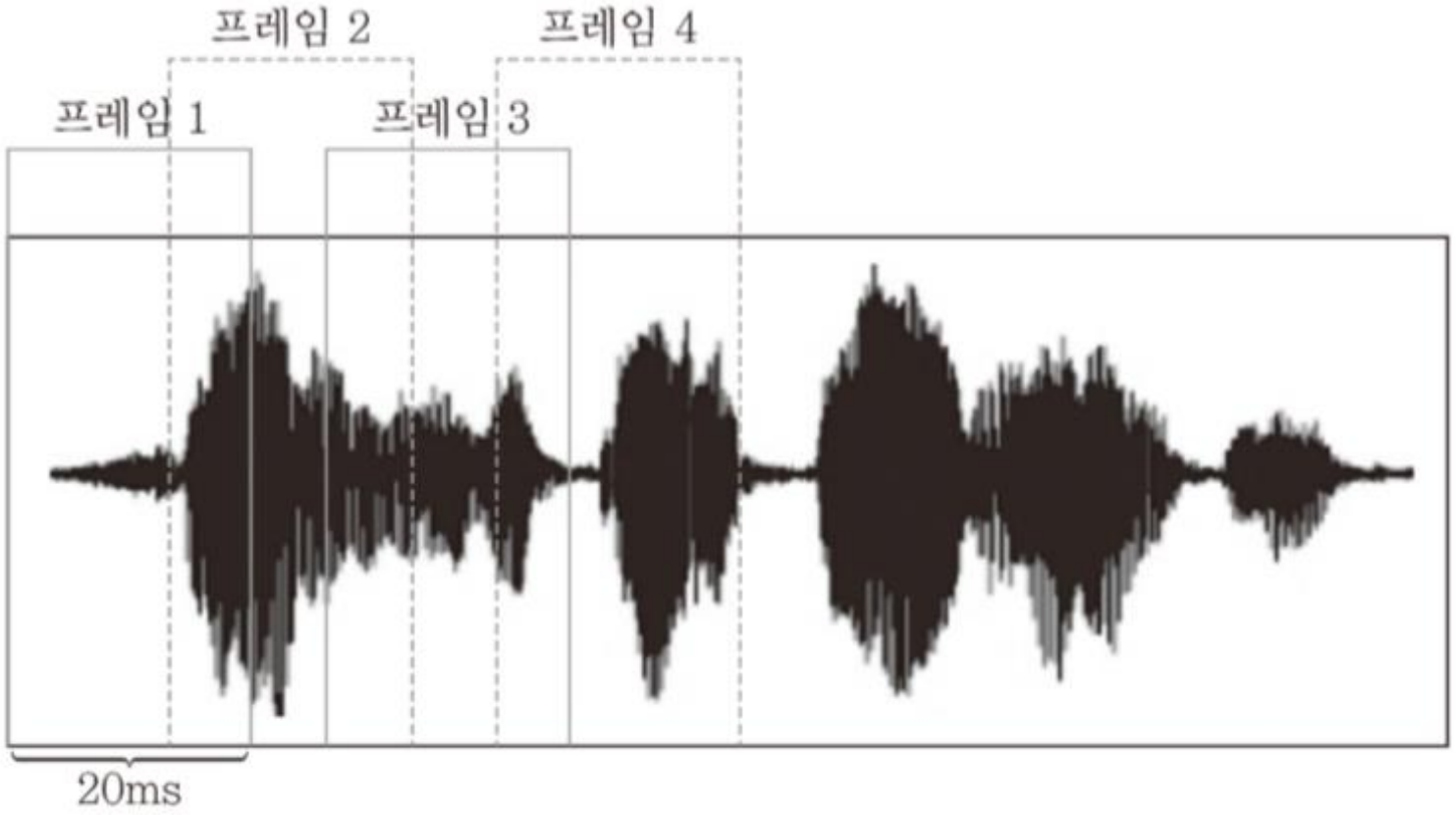
음성 인식

- 음성 신호를 텍스트로 자동으로 변환하는 것
- 텍스트를 생산하는 주요 기술



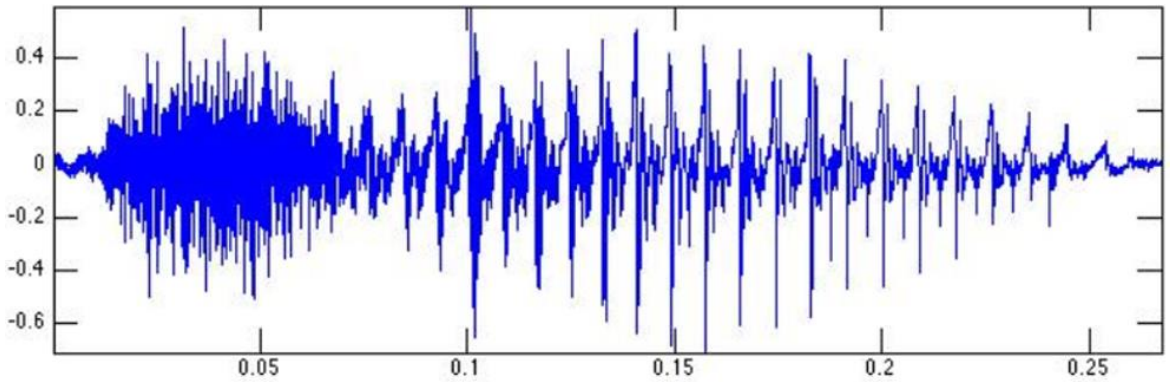
특징 추출

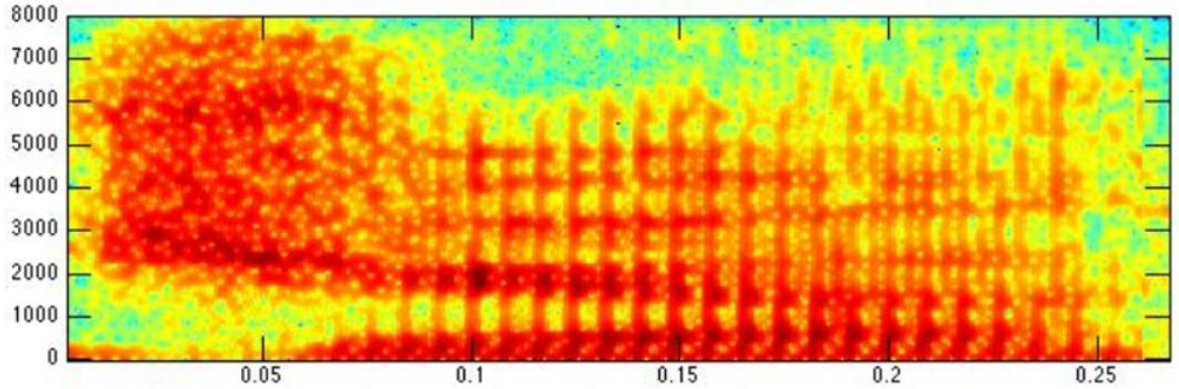
- 음성 신호를 프레임으로 구분
- 프레임 별로 특징 추출
  - 음성 프레임 → MFCC 데이터



스펙트로그램(spectrogram)

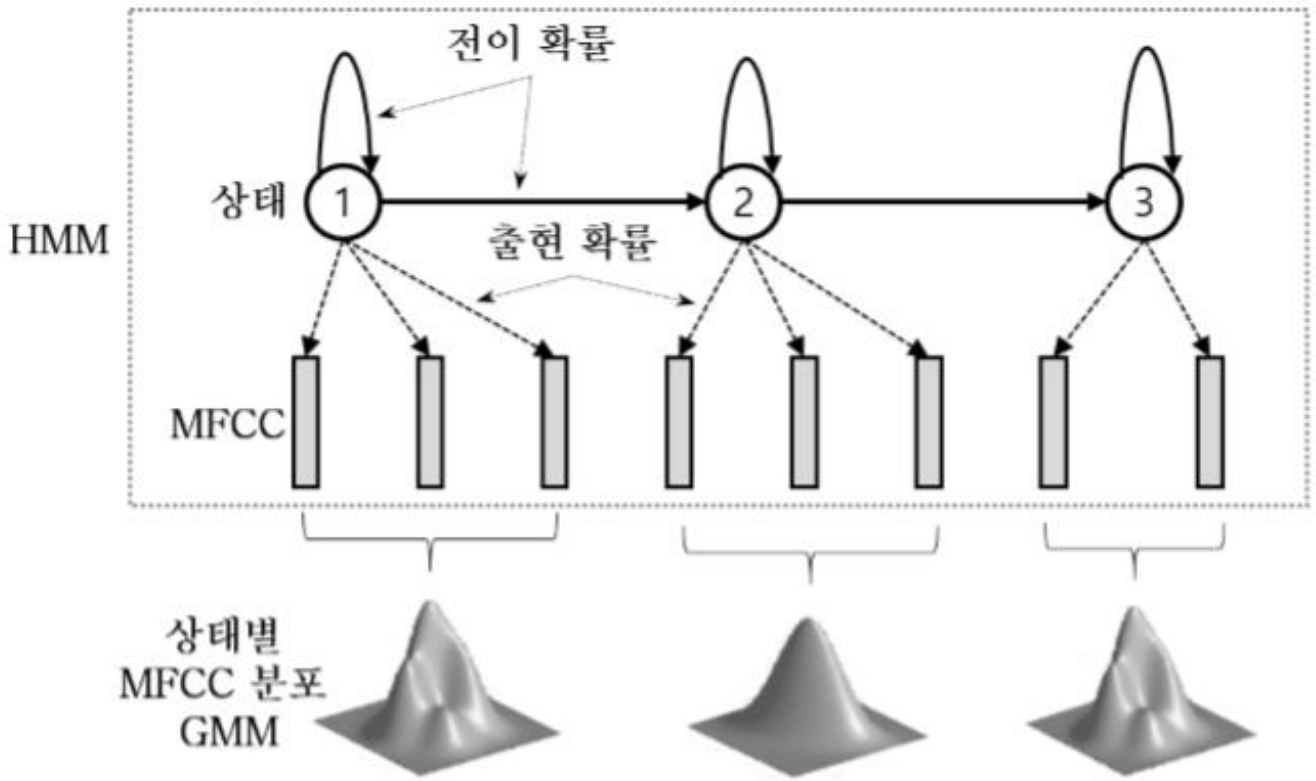
- 음성 프레임들의 MFCC 데이터의 시각적 표현
- 딥러닝 모델을 음성 인식 등에 적용할 때 사용





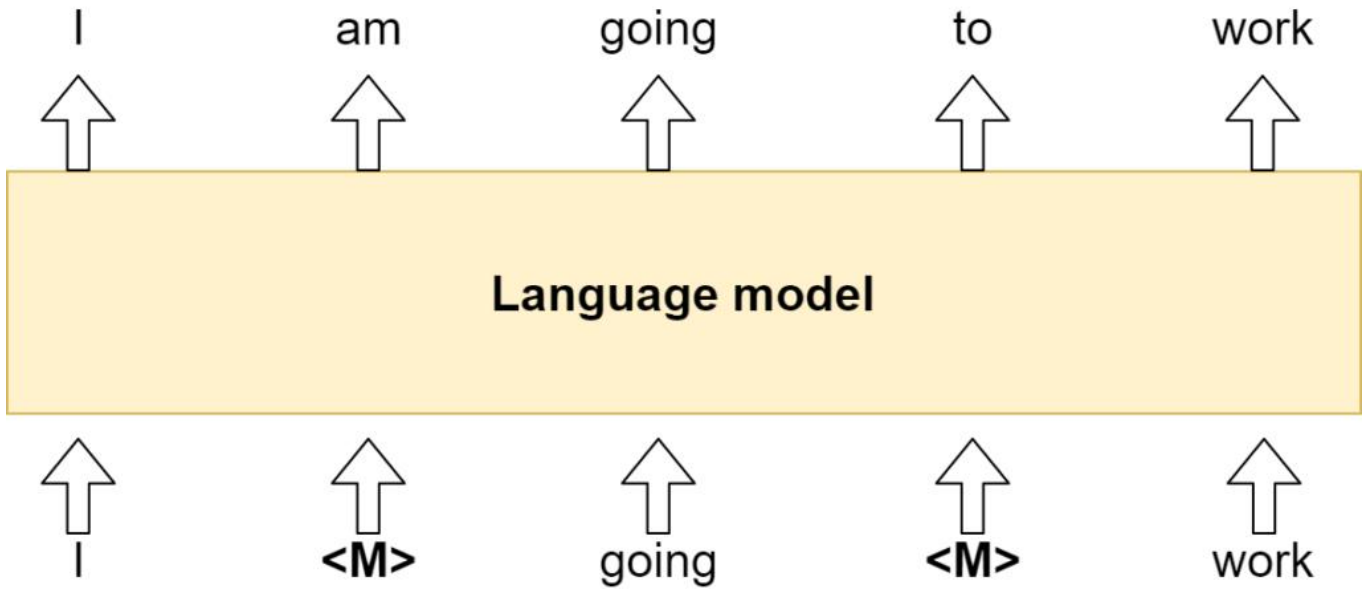
음향 모델(acoustic model)

- 대량의 학습용 음성 데이터베이스를 이용하여 음소(phoneme, 音素)별로 특성 정보를 구성해 놓은 것



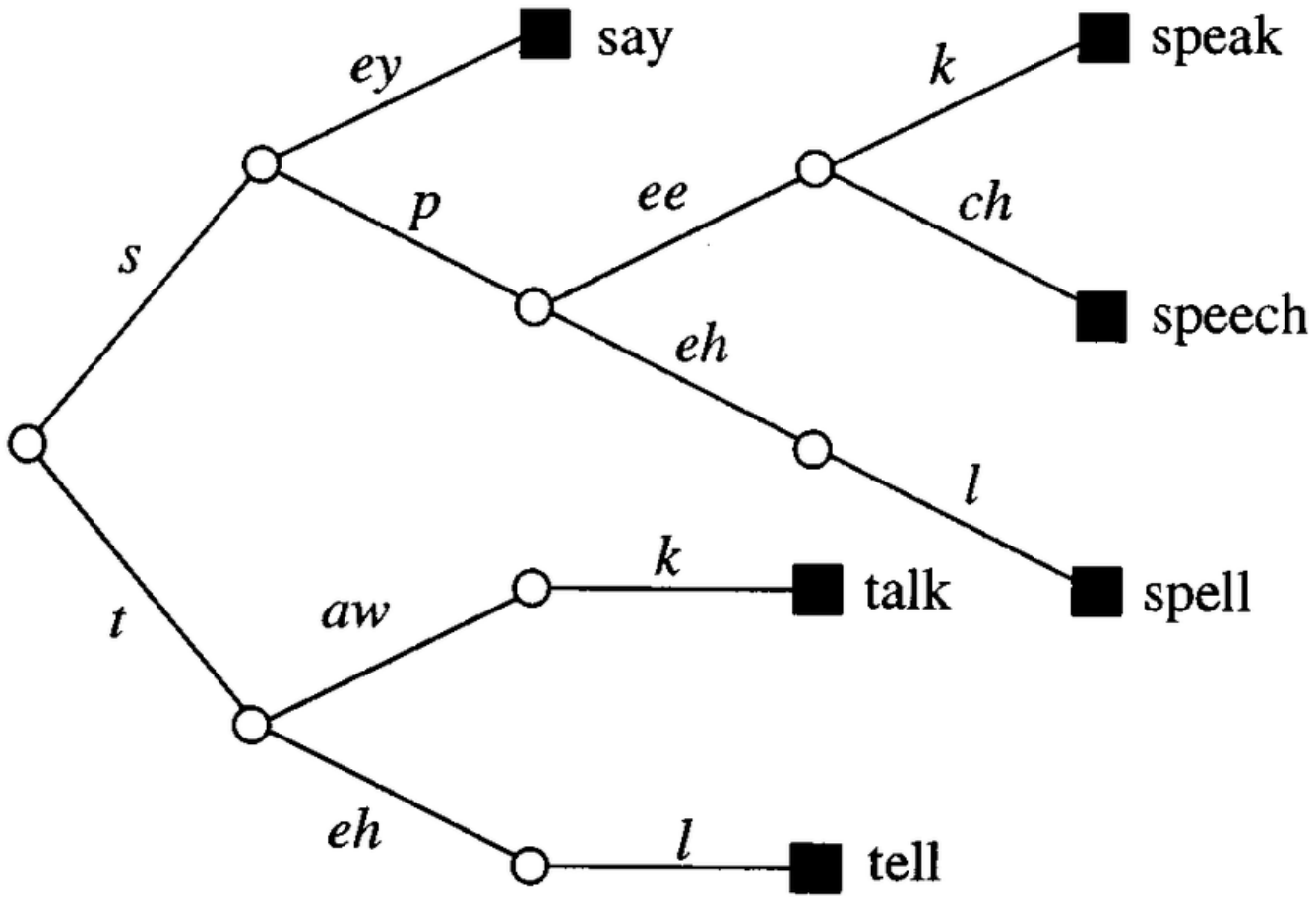
언어 모델(language model)

- 대규모 텍스트 데이터로부터 이전에 나타난 단어열 정보로부터 현재 단어가 나타날 확률을 계산하는 모델
- 음성인식 단계에서 탐색 엔진이 가장 적절한 단어열을 찾을 수 있도록, 단어열의 확률을 계산할 때 사용



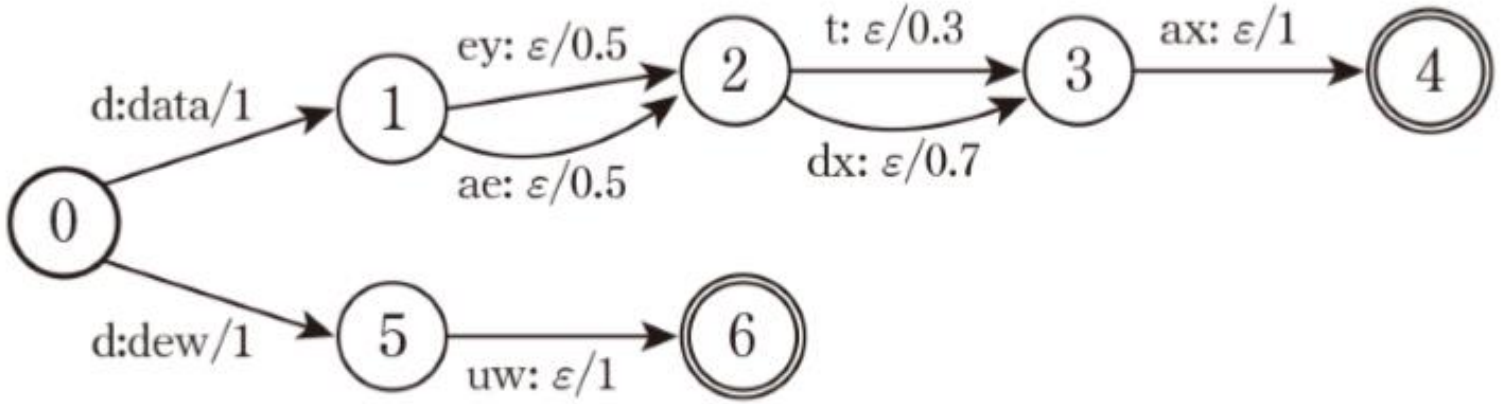
발음 사전(pronunciation lexicon)

- 단어별로 단어를 구성하는 글자인 문자소(grapheme)와 음소로 대응시켜 기록한 사전



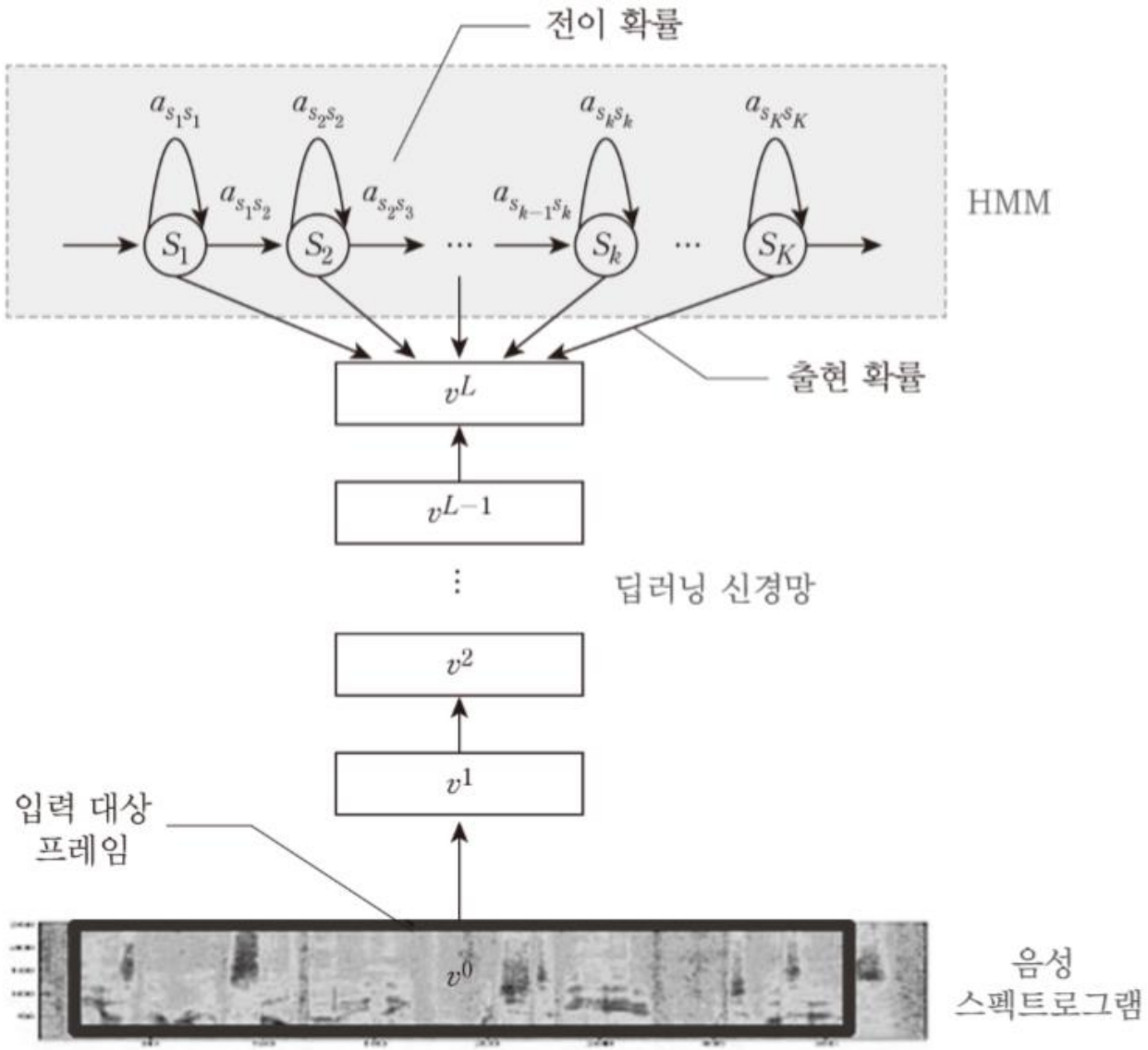
탐색 엔진

- 주어진 MFCC 벡터 배열이 나타내는 단어 또는 문장을 효과적으로 탐색할 수 있도록, 음향 모델, 언어 모델, 발음사전을 결합한 자료구조 사용
- **가중 유한상태 변환기**(weighted finite state transducer) 구조로 표현



딥러닝 기반 음성 인식

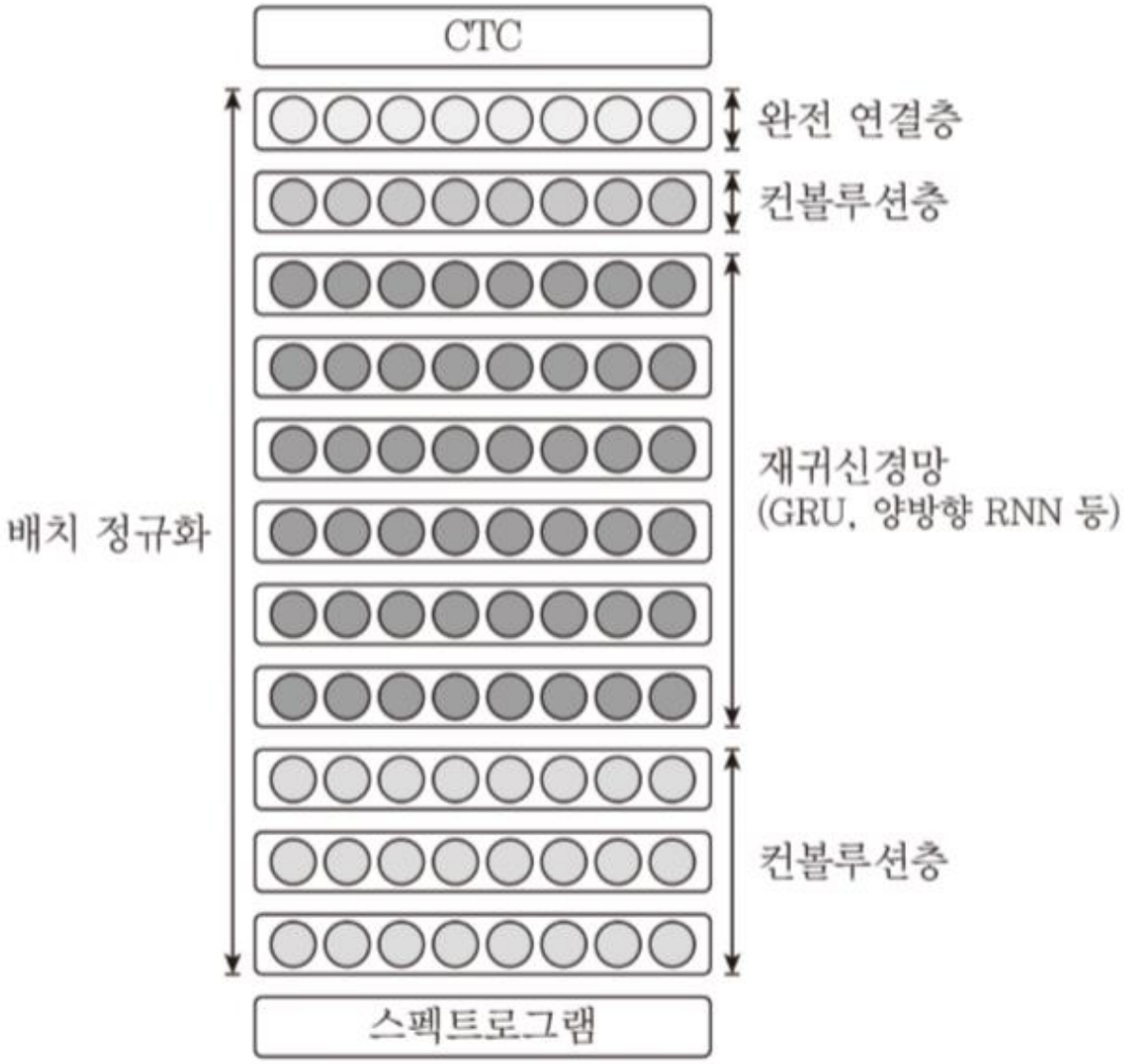
- 음성인식 과정의 특정 단계에서 딥러닝 모델 사용



What happens if I feed my own voice into a **neural network**?

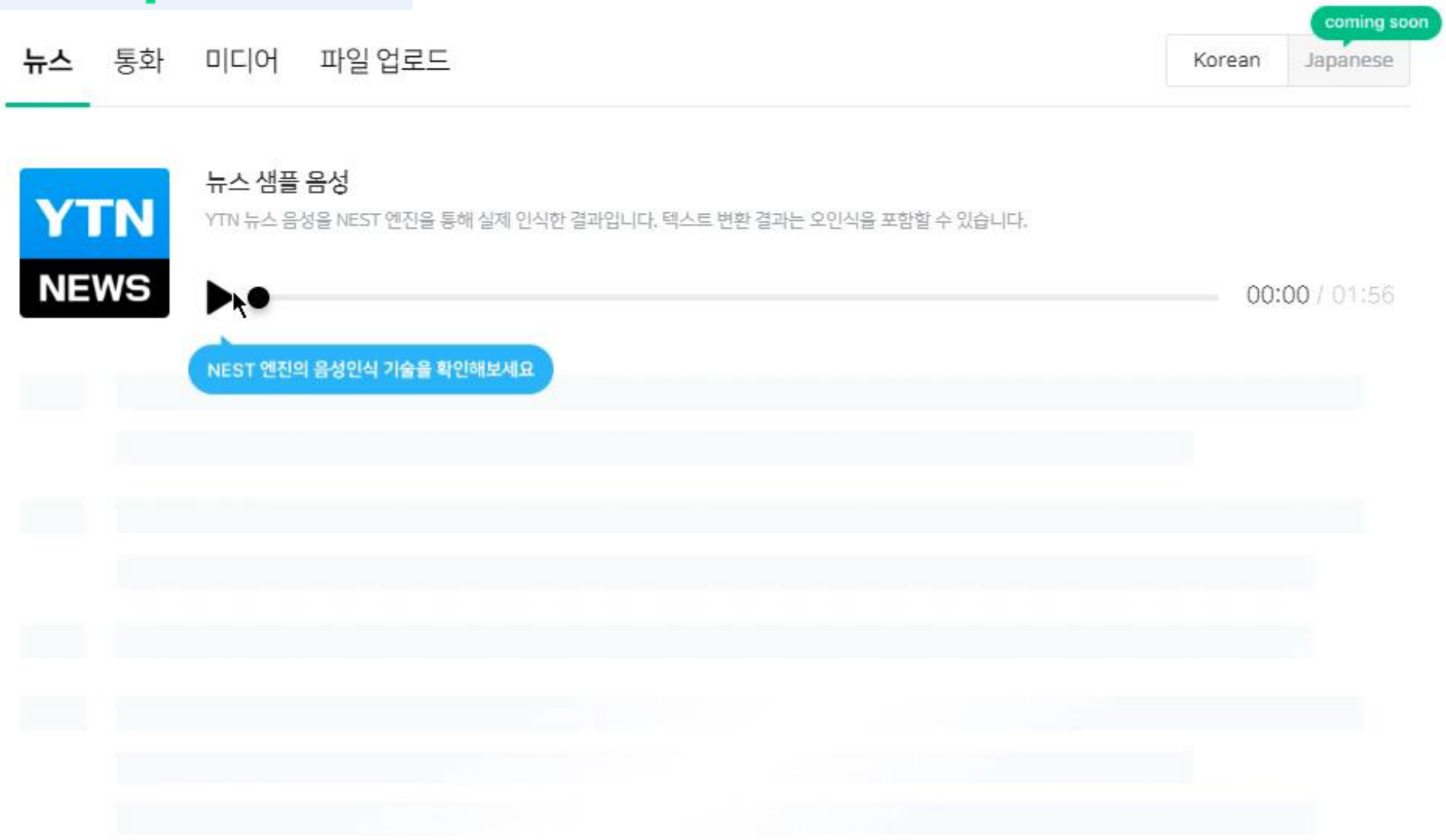
딥러닝 기반 음성 인식

- 딥러닝을 이용한 종단간 학습으로 음성인식

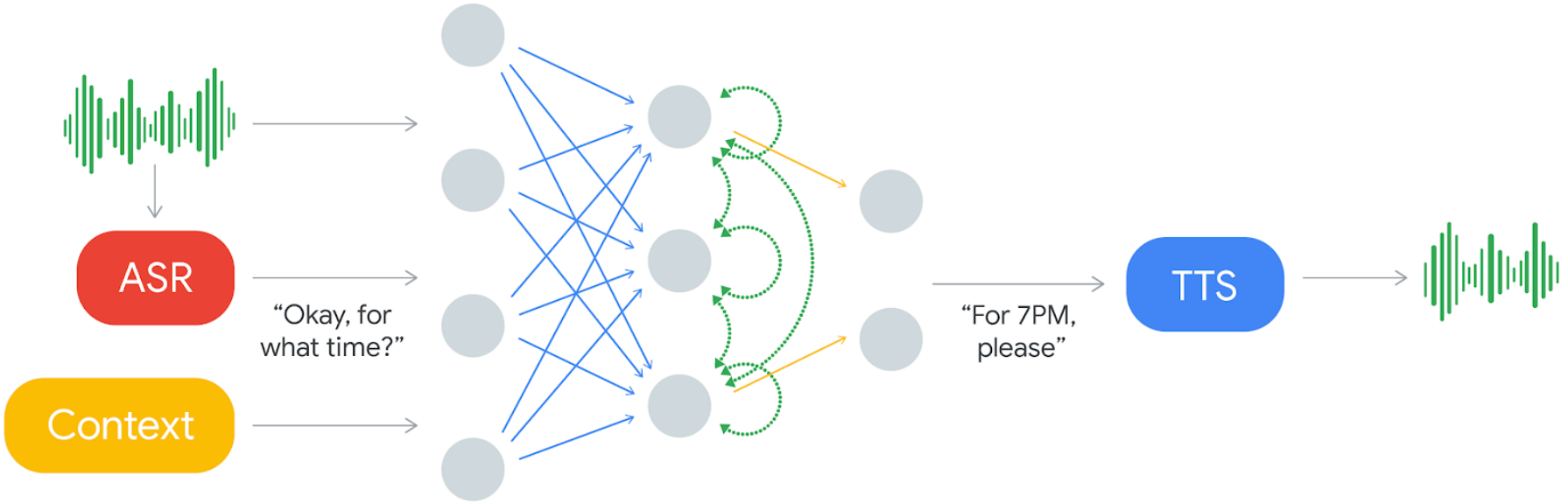


CLOVA Speech

https://clova.ai/speech



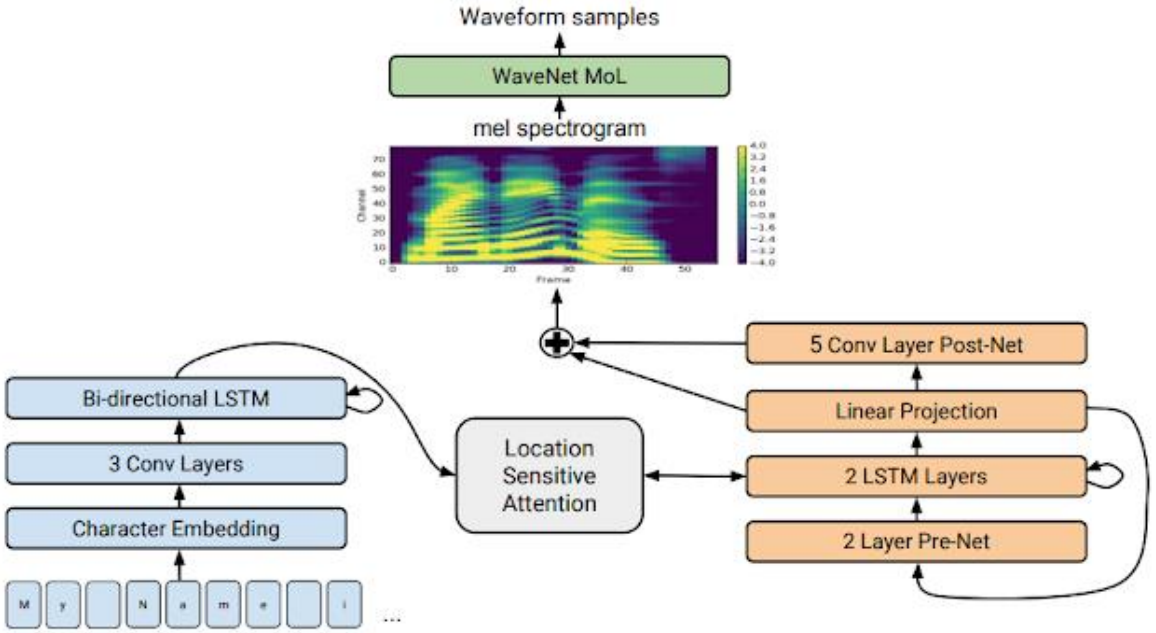
Google Duplex

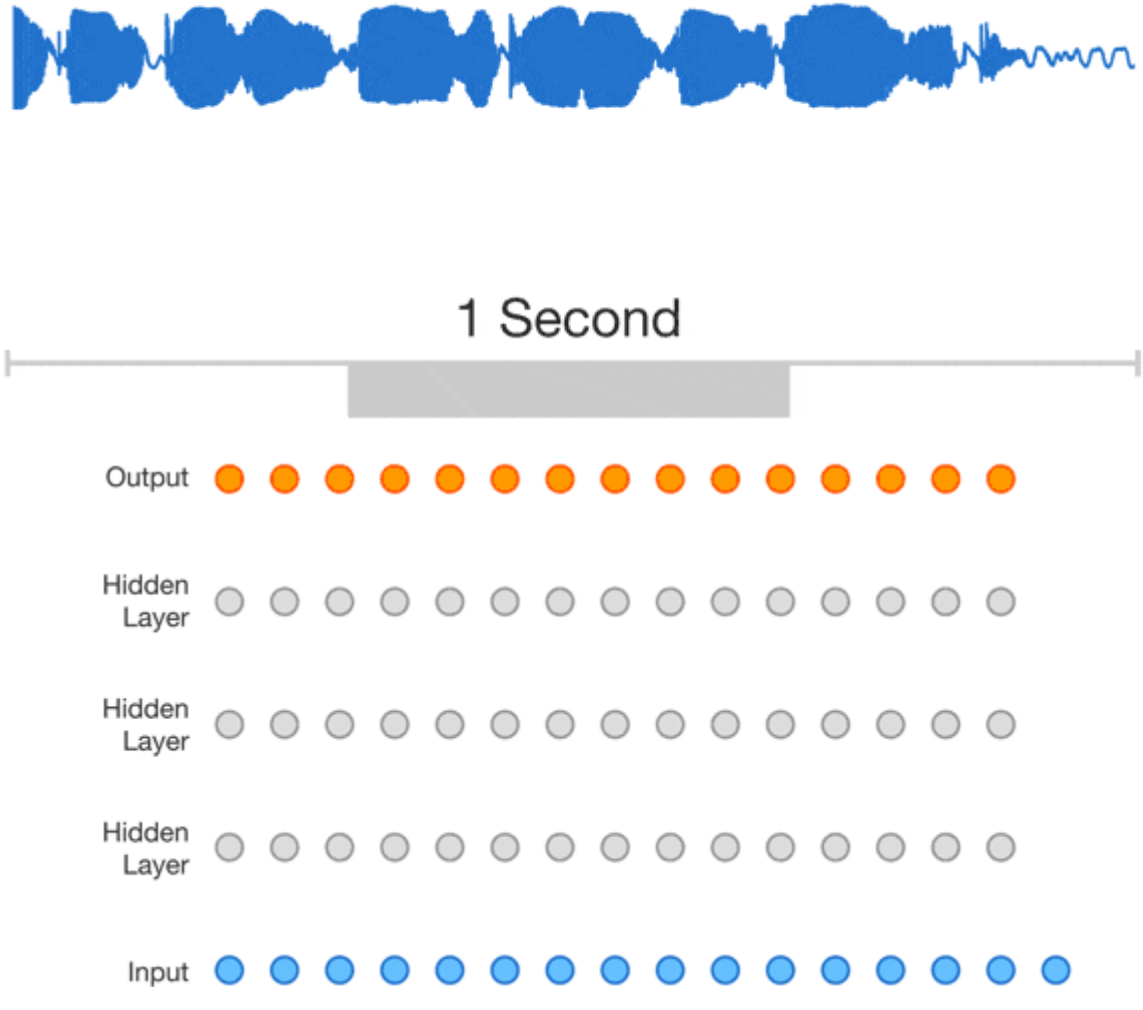


Duplex scheduling a hair salon appointment:

Duplex calling a restaurant:

Tacotron 2: Generating Human-like Speech from Text

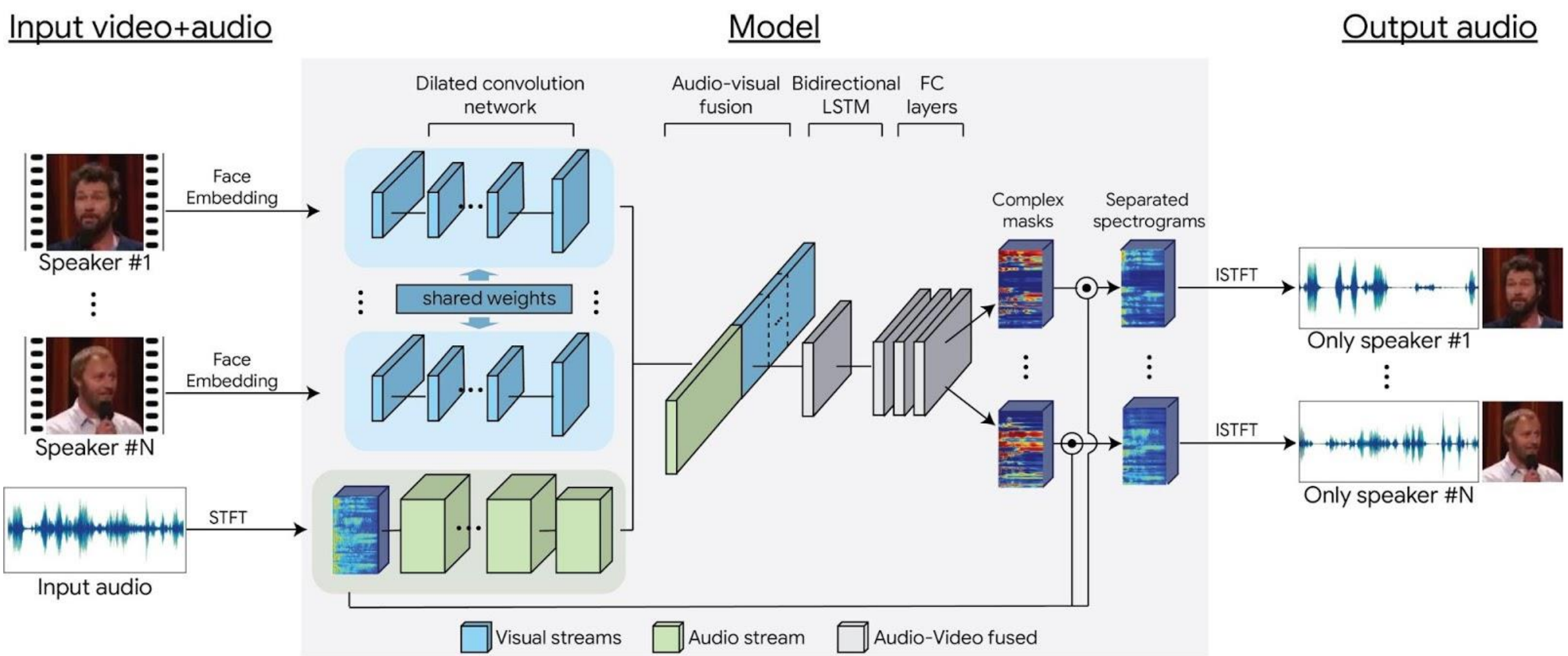


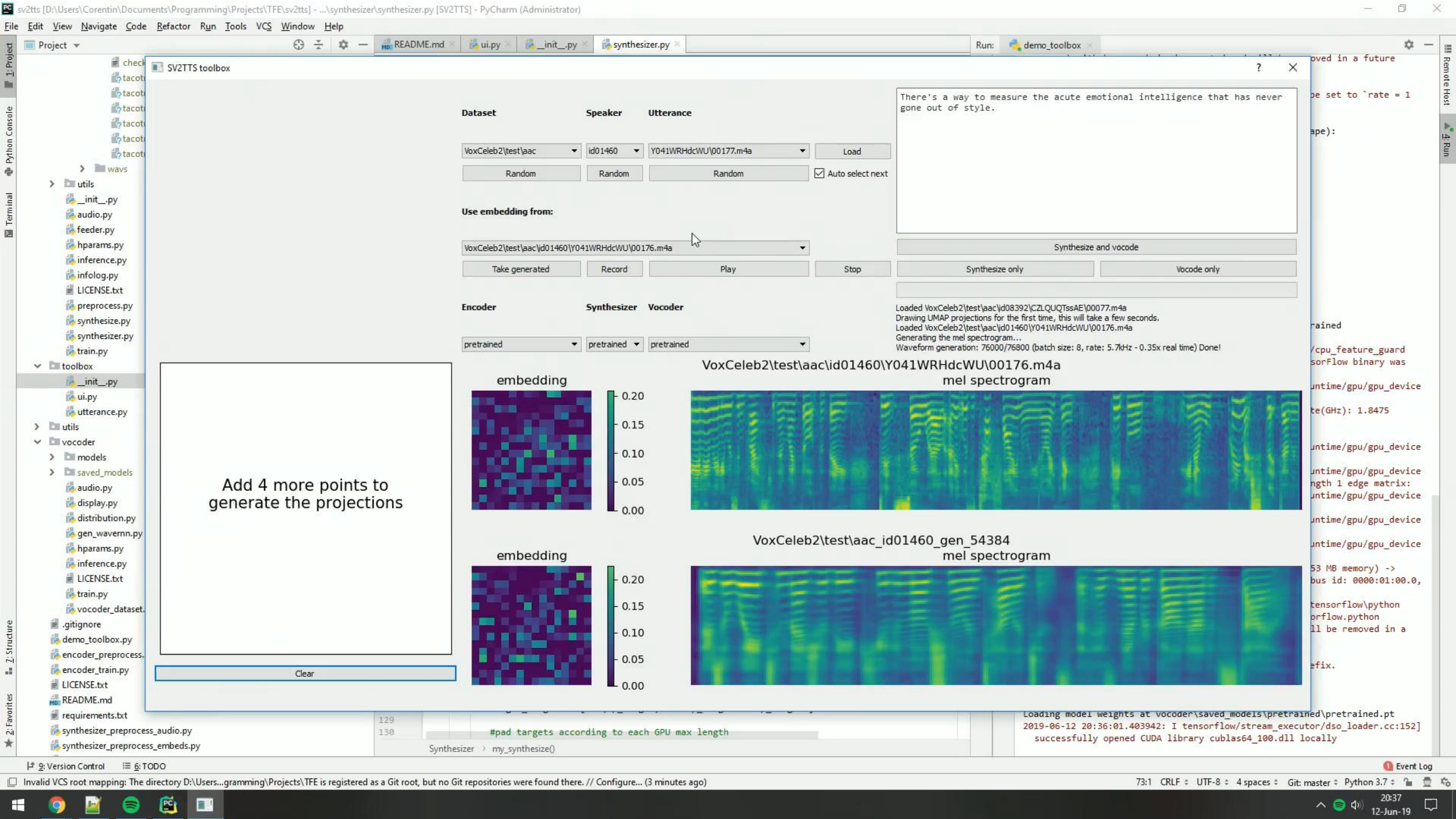


Input video (two people speaking together)

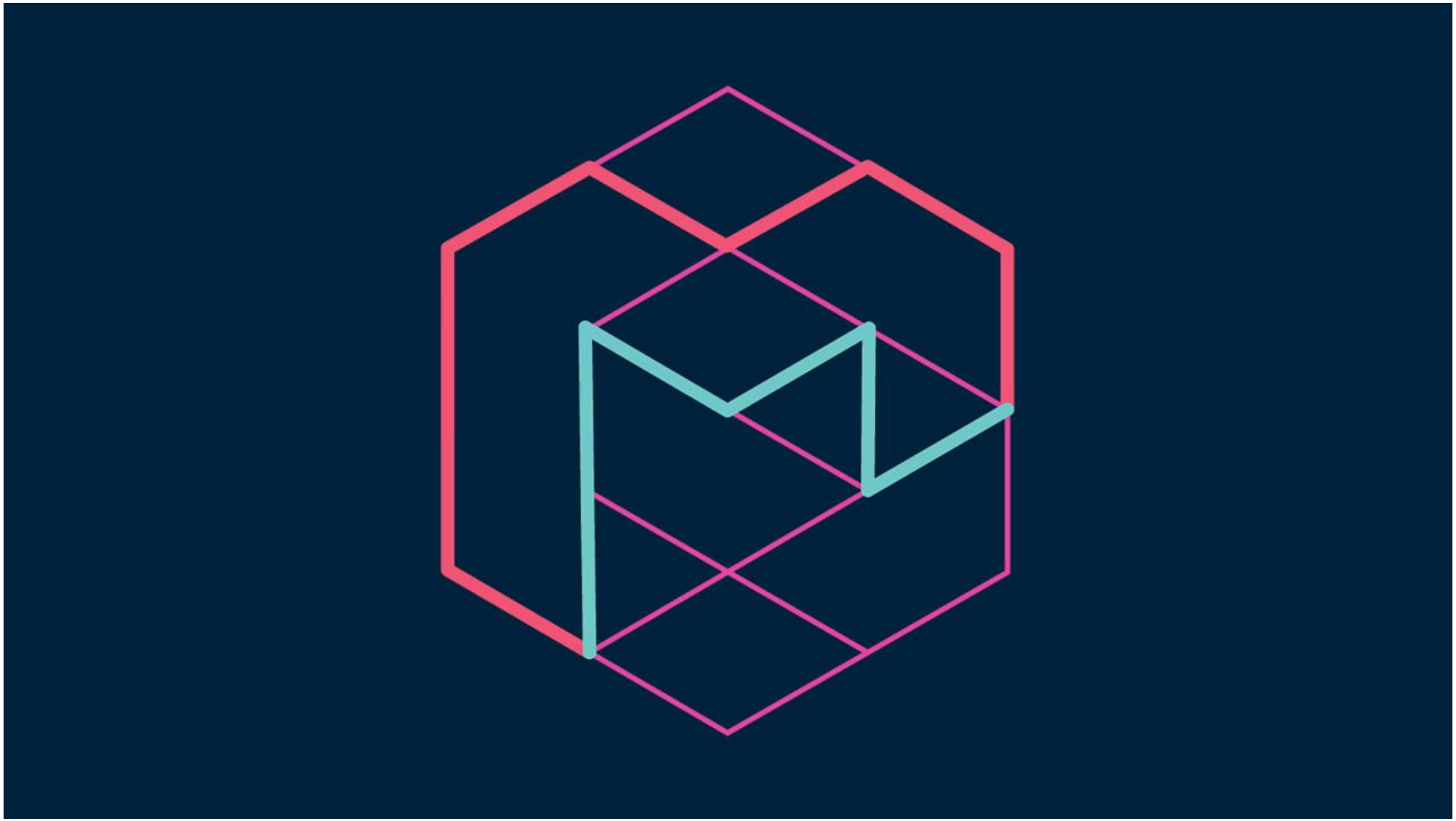
Video source: Team Coco, https://www.youtube.com/watch?v=UT7h4nRcWjU

An Audio-Visual Speech Separation Model




SV2TTS toolbox
Dataset
Speaker
Utterance
VoxCeleb2\test\aac
id01460
Y041WRHdcWU\00177.m4a
Load
Random
Random
Random
Auto select next
There's a way to measure the acute emotional intelligence that has never gone out of style.
Use embedding from:
VoxCeleb2\test\aac\id01460\Y041WRHdcWU\00176.m4a
Synthesize and vocode
Take generated
Record
Play
Stop
Synthesize only
Vocode only
Encoder
Synthesizer
Vocoder
pretrained
pretrained
pretrained
Loaded VoxCeleb2\test\aac\id08392\CZLQUQTssAE\00077.m4a
Drawing UMAP projections for the first time, this will take a few seconds.
Loaded VoxCeleb2\test\aac\id01460\Y041WRHdcWU\00176.m4a
Generating the mel spectrogram...
Waveform generation: 76000/76800 (batch size: 8, rate: 5.7kHz - 0.35x real time) Done!
Add 4 more points to generate the projections
Clear
embedding
0.20
0.15
0.10
0.05
0.00
VoxCeleb2\test\aac\id01460\Y041WRHdcWU\00176.m4a
mel spectrogram
embedding
0.20
0.15
0.10
0.05
0.00
VoxCeleb2\test\aac_id01460_gen_54384
mel spectrogram

RNN Music Generation



Daddy's Car

- 소니 CSL 리서치 랩의 과학자들은 인공 지능으로 구성된 최초의 노래 인 "Daddy's Car"와 "Mister Shadow"를 제작
- 거대한 데이터베이스에서 음악 스타일을 학습하는 시스템 인 FlowMachines를 개발
- 스타일 전송, 최적화 및 상호 작용 기술의 독특한 조합을 활용하여 FlowMachines는 다양한 스타일의 새로운 노래를 제작
- "Daddy's Car"는 비틀즈 스타일로 구성
- 프랑스 작곡가 인 Benoît Carre는 노래를 준비하고 제작하여 가사를 작업
- 두 곡은 인공 지능으로 구성된 앨범의 발췌 부분으로 2017 년에 발매

Supplementary audio samples to the paper:

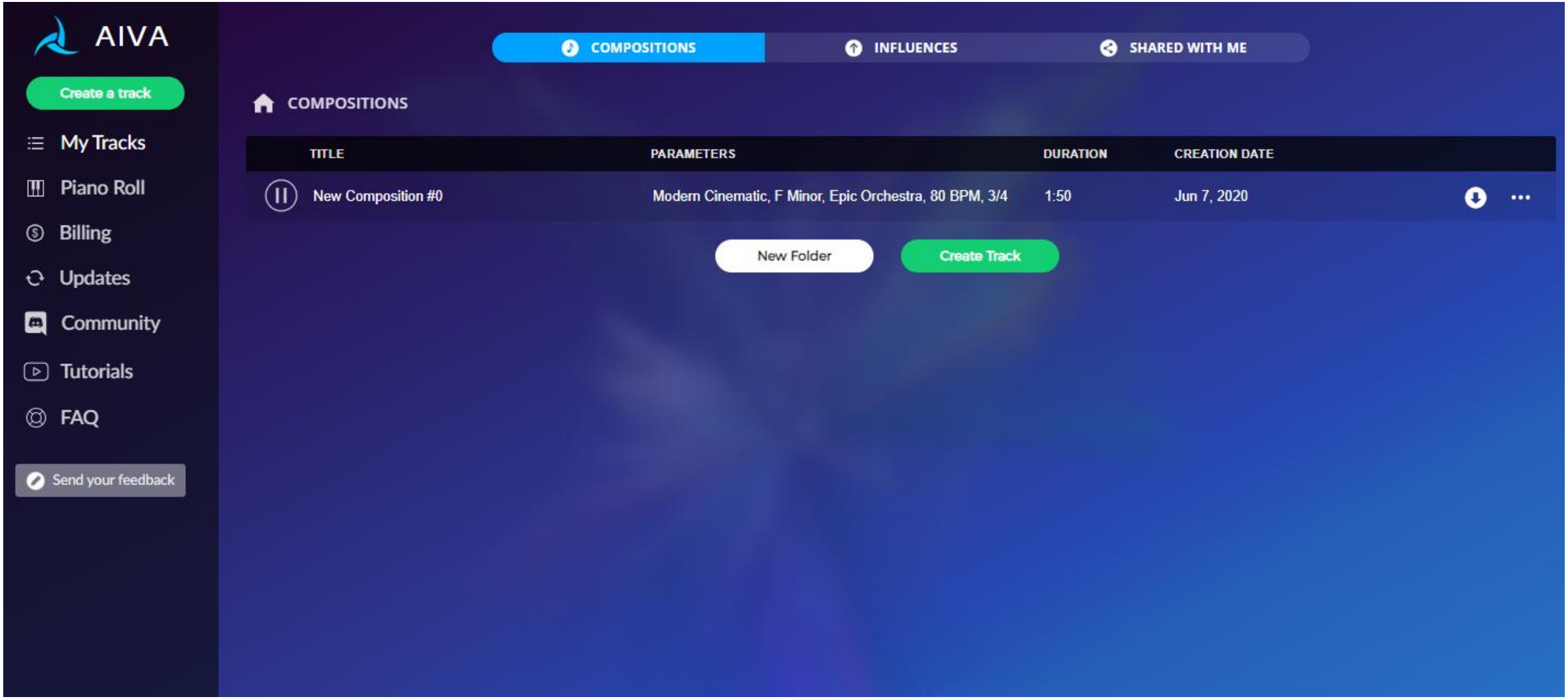
A Universal Music Translation Network

Noam Mor, Lior Wolf, Adam Polyak, Yaniv Taigman
Facebook AI Research


Drums
Input Clip
2 Electric Piano
Electric Piano
Temperature
1.0
Output clip to Clip Slot 2
Generate
MAGENTA
2 Electric Piano
Bass
Drums
Bass 2
Master
Electric Piano
Bass
1/1 [Electric Pi
Bass 2
Drop Files and Devices Here
126 bpm
89 bpm
85 bpm
4
5
MIDI From
All Ins
All Channels
Monitor
In Auto Off
MIDI To
No Output
Audio From
Ext. In
Audio To
Master
Cue Out
ii 1/2
Master Out
ii 1/2
-11.5
-81.4
-11.0
-8.87
Solo
Clip
Sample
Bass
4-Audio 0002 [2019
44.1 kHz 24 Bit 1 Ch
Edit Save Rev.
HiQ Fade RAM
Warp
Seg. BPM
83.35
Transpose
0 st
Detune
0 ct
0.00 dB
Preserve
Trans
Start
End
Loop
Position
Length
Signature
Groove
None
Commit
1.2
1.3
1.4
2
2.2
2.3
2.4
1/16

AIVA – AI 음악 작곡

https://beta.aiva.ai/



I am AI

Composed by AIVA for NVIDIA

00:00

Aiva Technologies

www.aiva.ai


suanlab
이수안 컴퓨터 연구소
suan computer laboratory