maum Academy
AI Professional Education

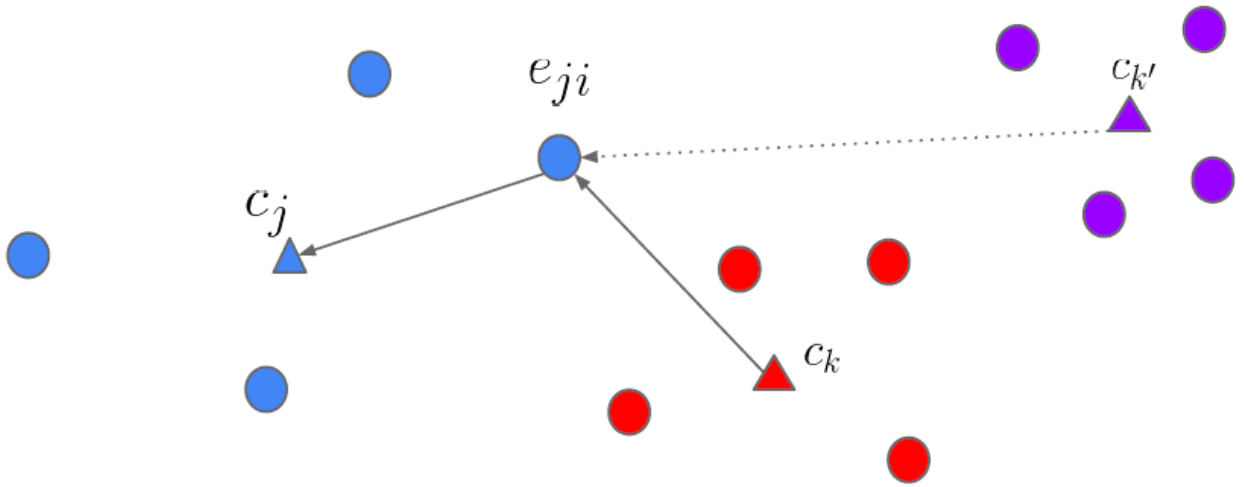
Level 2. How to customize AI engines

http://mindslab.ai

1. 화자인증
2. 화자분리
3. VoiceFilter
4. 음성 정제
5. 실습

화자인증

- 화자인증(Speaker Verification):
  - 2개의 음성이 같은 사람의 것인지 판별해주는 엔진
  - 임의의 길이의 음성을 고정 차원 화자벡터(Speaker Vector)로 만들어줌
  - 벡터 사이의 거리를 통해 유사도를 판별



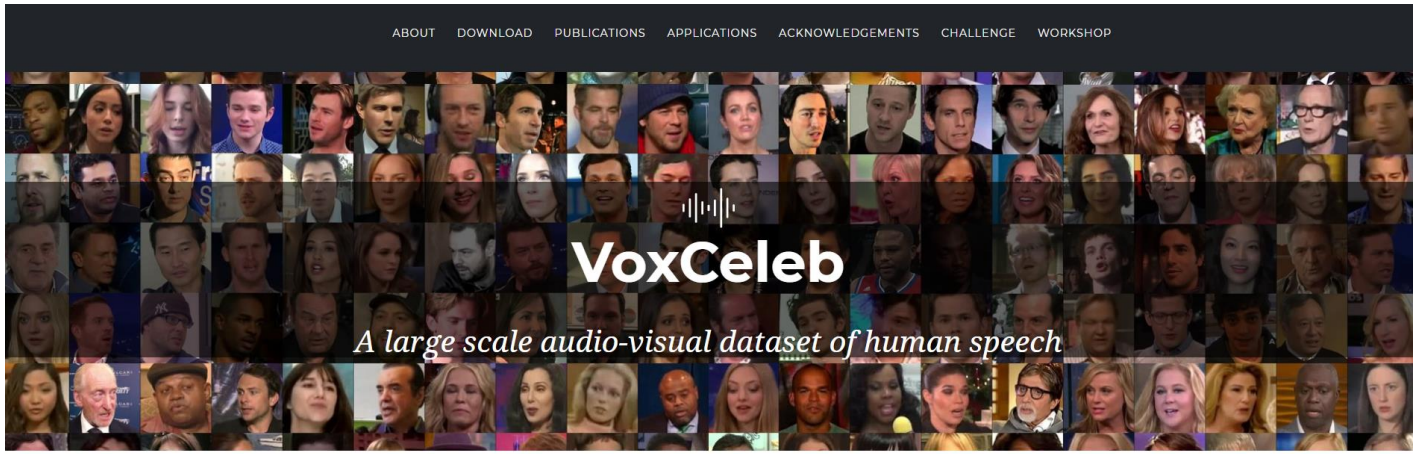
From “Generalized End-to-End Loss for Speaker Verification”, L. Wan et al.

화자인증
Speaker Verification

화자인증 학습 데이터

- N명의 사람들이 발화한 음성(5초 이상) M개
  - N, M 은 다다익선

- 동일한 사람에 대해 녹음 환경, 날짜, 감정, 발화 언어 등이 다양해야 함
  - VoxCeleb2 데이터 사용
  - Language agnostic (영어, 한국어 등 동시 사용 가능)



VoxCeleb is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube

7,000 +	1 million +	2,000 +
speakers	utterances	hours

화자인증
Speaker Verification

화자인증 학습

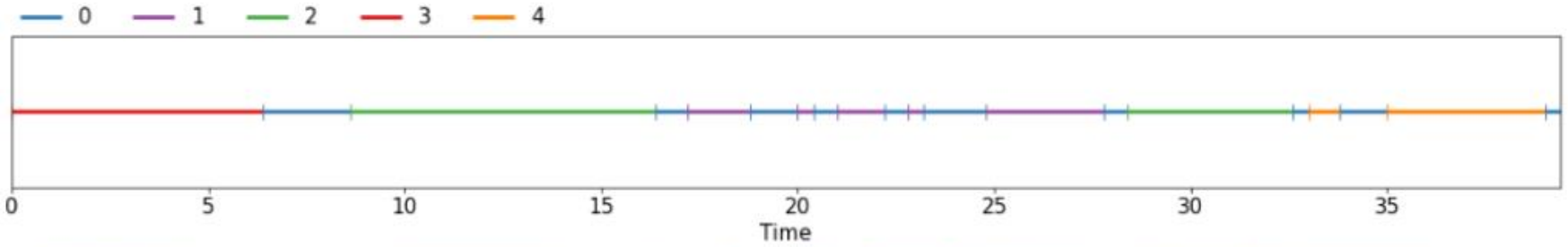
- VoxCeleb2로 한번 학습해두고 나면 추가 학습이 필요 없음
- 화자인증이 아닌 화자인식으로 사용하려 한다면 transfer learning 가능

maum Academy
AI Professional Education

화자분리
Diarization

화자분리 란

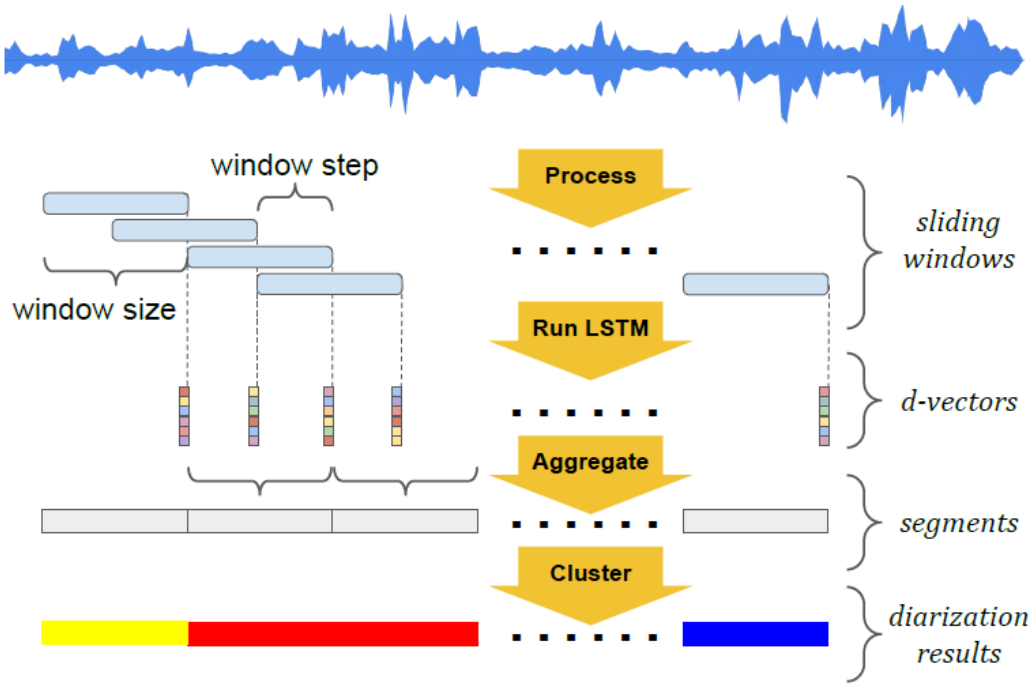
- "Who spoke when?": 대화에 등장하는 각각의 화자가 발화한 시각을 구하는 엔진
- 화자인증 엔진 기반이기 때문에 language agnostic



화자분리
Diarization

화자분리 작동 원리: 클러스터링

- 대화 음성을 0.8초씩 잘라, 각각에 대해 화자벡터 추출
- 화자벡터의 분포를 비지도 클러스터링 (unsupervised clustering)
  - 이 과정은 **학습이 필요 없음**



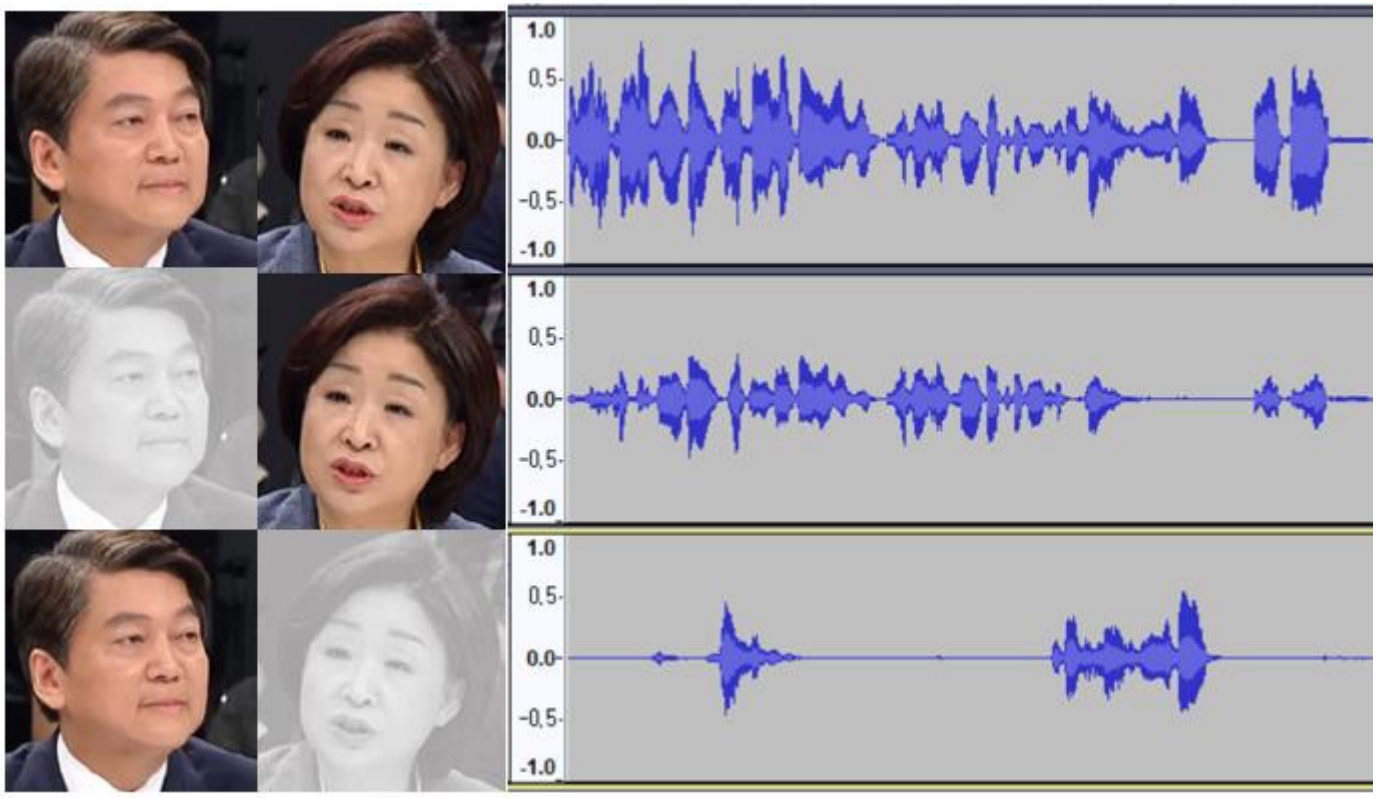
From “Speaker Diarization with LSTM”, Q. Wang et al.

VoiceFilter

음성분리

보이스필터

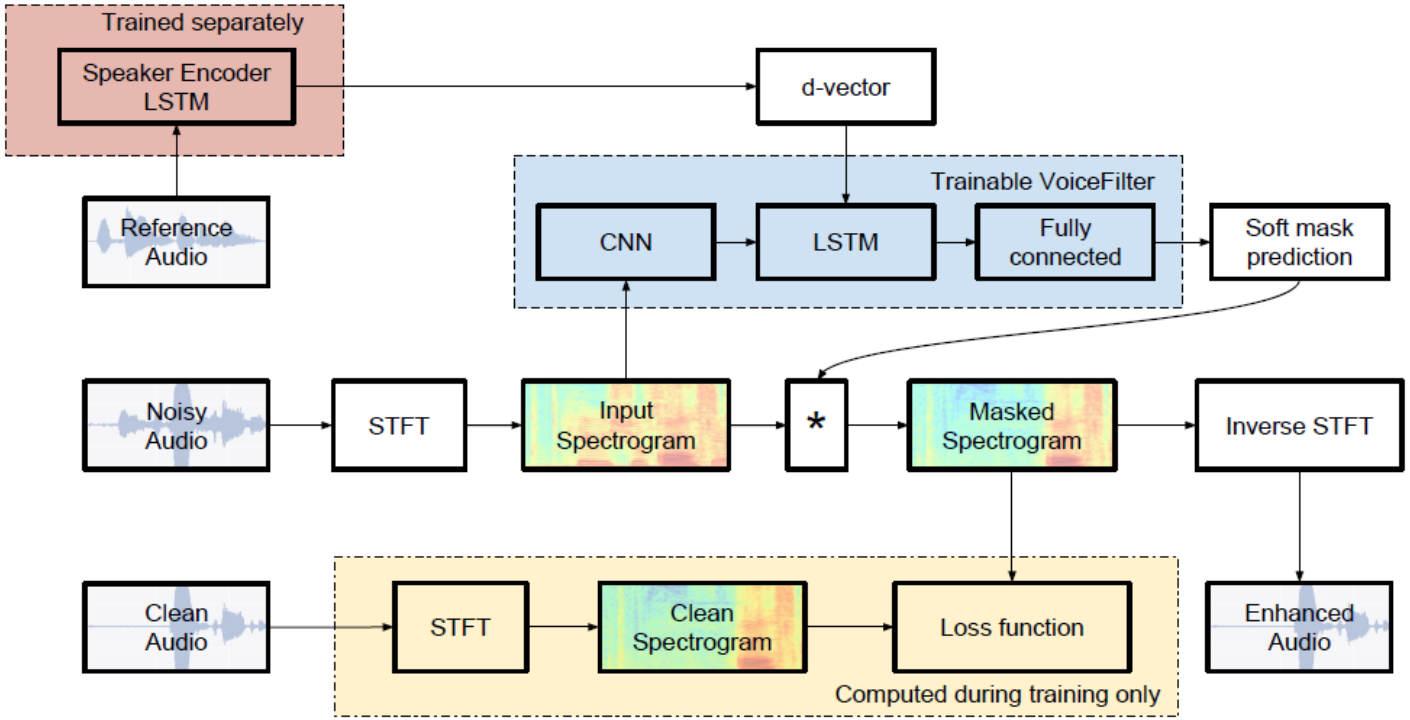
- 음성 내 겹침 발화로부터 원하는 화자의 음성을 추출/제거
- 원하는 화자에 대한 STT 성능을 높일 수 있음



VoiceFilter
음성분리

보이스필터

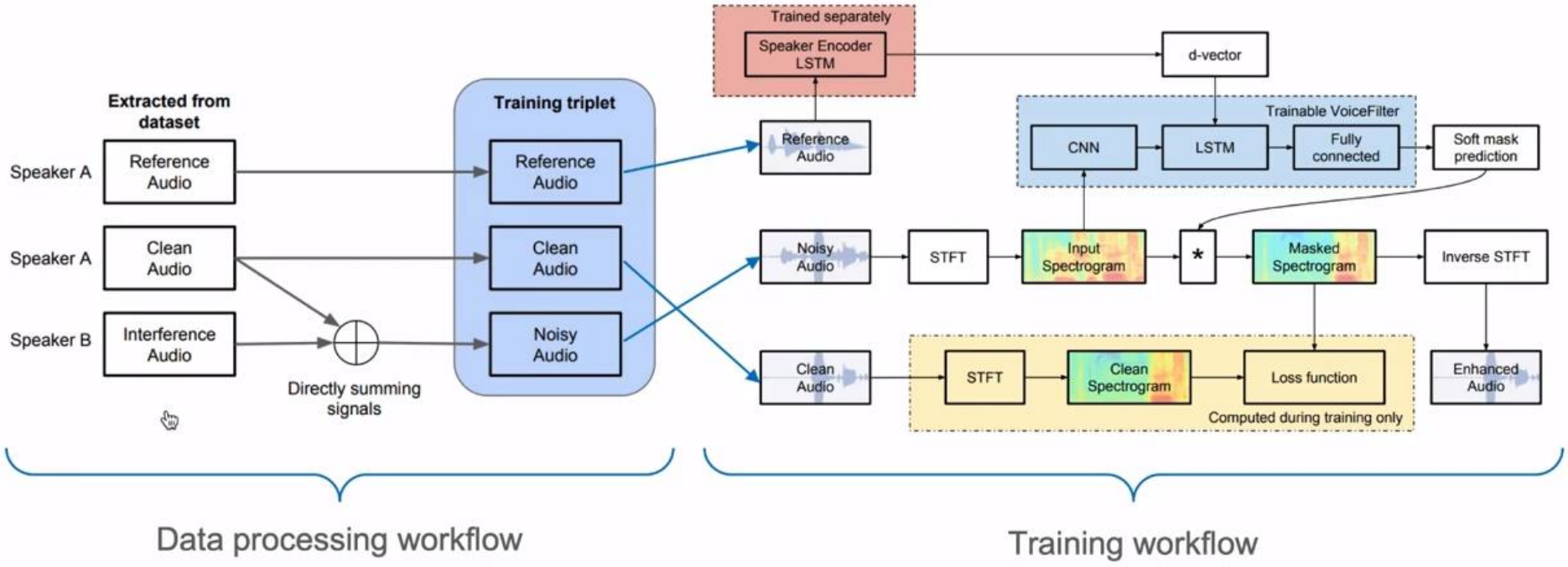
- A의 음성과 B의 음성을 합친 뒤, 그로부터 A의 음성만 뽑아내도록 학습
  - 단, A의 음성에 대한 사전 지식을 제공 (A의 화자벡터)



From “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking”, Q. Wang et al.

VoiceFilter
음성분리

Data Processing for VoiceFilter



https://www.youtube.com/watch?v=gnRX2lzepz0&t=542s

VoiceFilter

음성분리

음성분리 학습

- 기본적으로는 타겟 화자에 대한 **추가 학습 필요 없음**

- LibriSpeech 와 같이 깨끗한 음성 데이터 사용
  - 잡음이 최소 수준이어야 함
  - 화자 수가 충분해야 함 (LibriSpeech 의 경우 2000명 이상)
  - language agnostic 함

- 각각의 폴더에는 각 화자의 음성 여러 개
  - 음성마다 화자벡터를 뽑아 ".gvlad" 포맷으로 저장

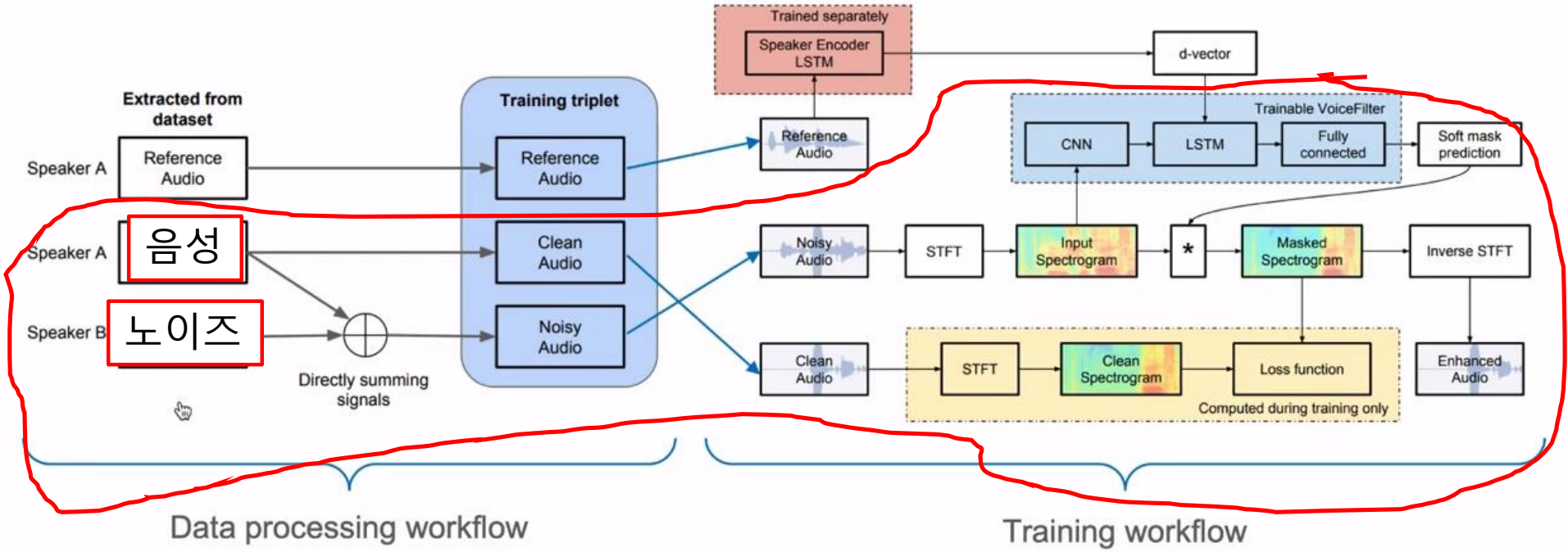
- 주의: 실행할 때에도 같은 화자벡터 추출기를 사용해야 함

음성정제

CNNoise

음성정제 작동 원리

- VoiceFilter 와 동일한 구조, 데이터만 변경
  - 사람 음성: LibriSpeech 등
  - 노이즈/음악: MUSAN (음악의 경우 가사 있는 음악 제외)
  - 사람 음성을 바꿔 domain 학습 가능하지만 일단은 **별도 학습 없이 사용 가능**



관련 자료


Smart City
AI Contact Center
AI English Education
Smart Factory
AI Robot

maum Academy
AI Professional Education

- https://arxiv.org/abs/1710.10467 (화자인증)
- https://arxiv.org/abs/1710.10468 (화자인증을 이용한 화자분리)
- https://arxiv.org/abs/1810.04826 (화자인증을 이용한 음성분리 – VoiceFilter)
- https://arxiv.org/abs/1902.10107 (화자인증 고도화)