Transformer 기반 음성 인식 및 생성 방법론

3조 - 나요셉, 김재민, 최원석, 조기흠

목차

CONTENTS

음성인식 활용

- 유튜브 자동 자막 생성
- iPhone Siri, AI 스피커
- 발음 교정 앱 ELSA speak

- 은행 업무를 도와주는 음성인식 기술들이 많아짐

→ 더 발전되면 본인인증까지 음성으로 가능하지 않을까?

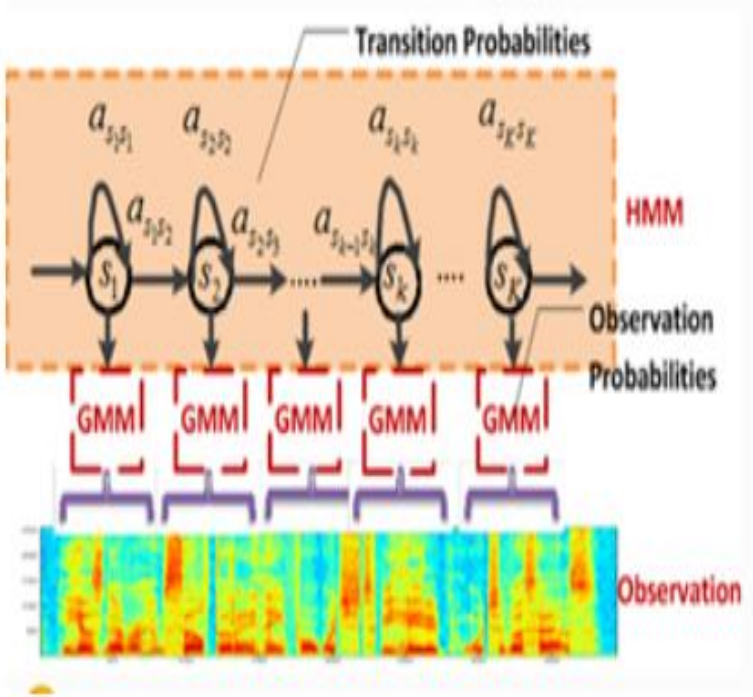
- 음성에서 감정을 파악해 분석하는 연구

  → 인간의 말에서 성격이나 가치관이 드러남

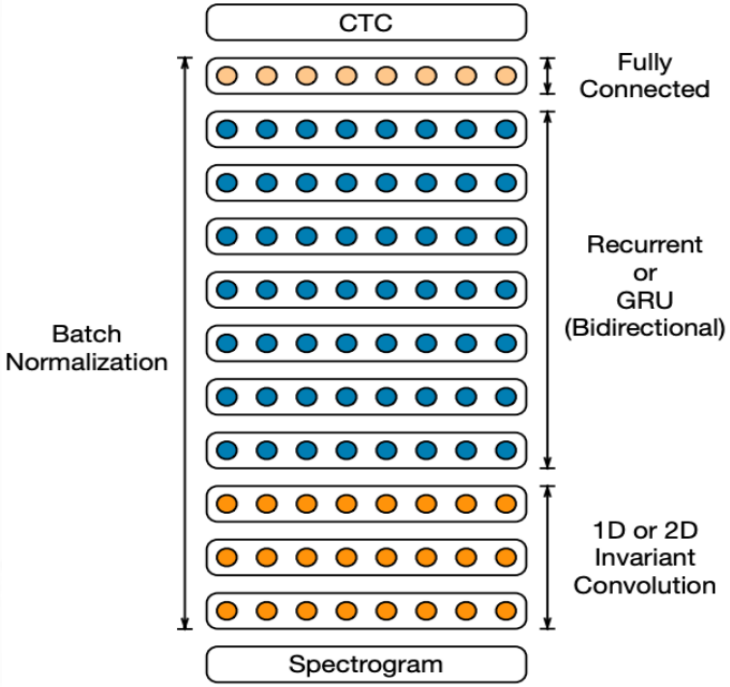
- 음성인식과 텍스트 확장 알고리즘을 이용한 인물 성향 분석에 관한 연구

기본적인 음성인식에 감정, 억양, 화자인식 같은 기술을 추가하면?

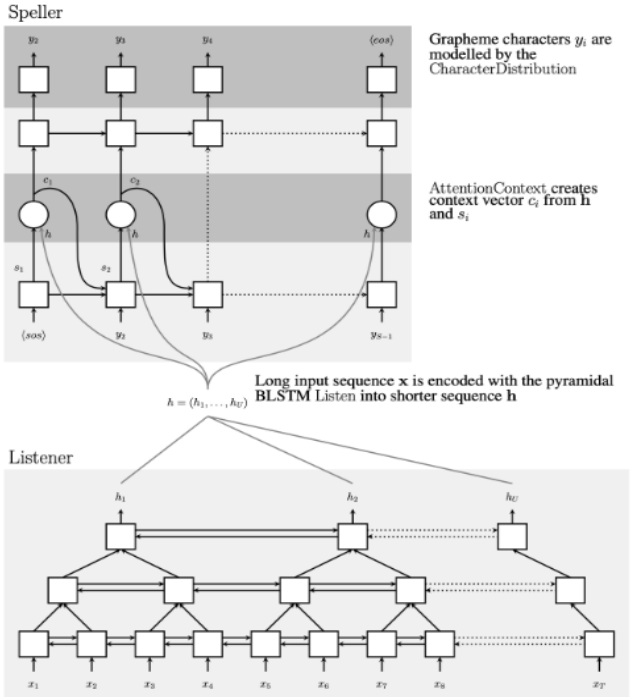
Acoustic Models



GMM-HMM
- 구성요소마다 다른 목적 함수를 독립적으로 학습하는 단점

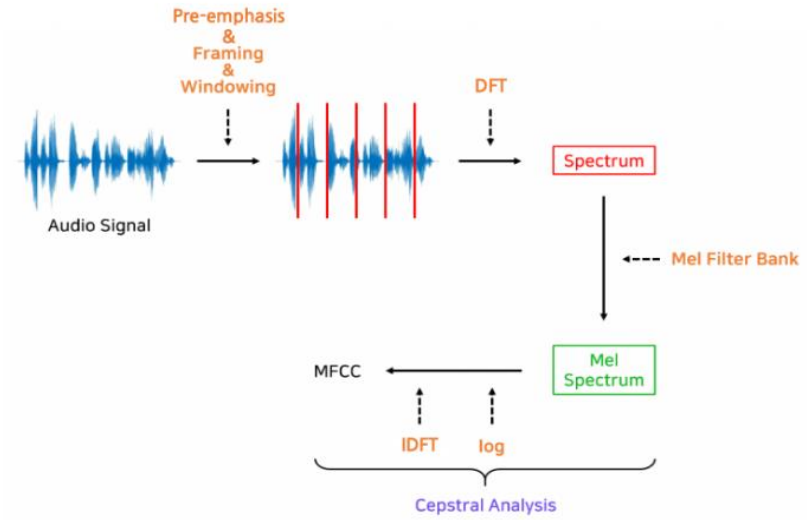


CTC
- 네트워크 출력을 조건부 독립으로 가정하는 문제

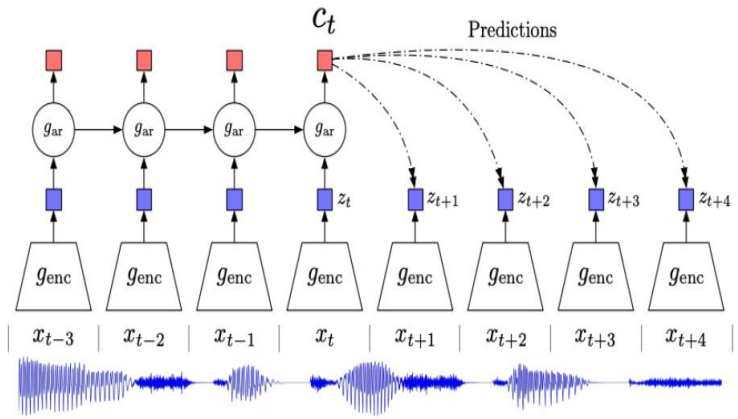


LAS
- 모델 출력이 음소 단위가 되도록 해 OOV해결

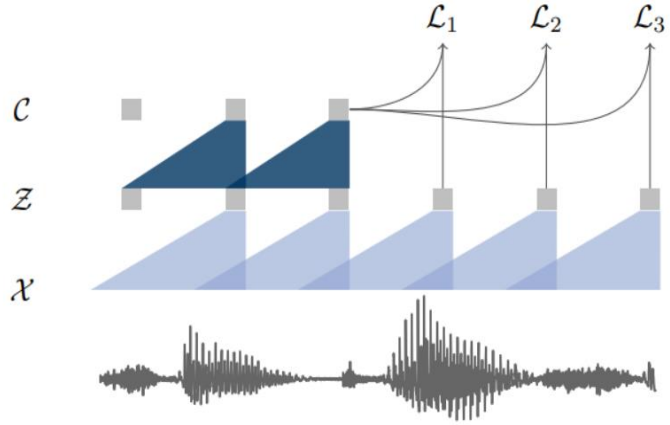
Feature Extraction



MFCC
- 낮은 신호 대 잡음비를 가지는 신호에서 성능 저하



CPC
- 고차원의 데이터를 간결한 latent embedding space로 압축해 사용



Wav2vec
- word2vec과 같은 loss를 최소화하는 방식으로 오디오 데이터의 representaion을 학습

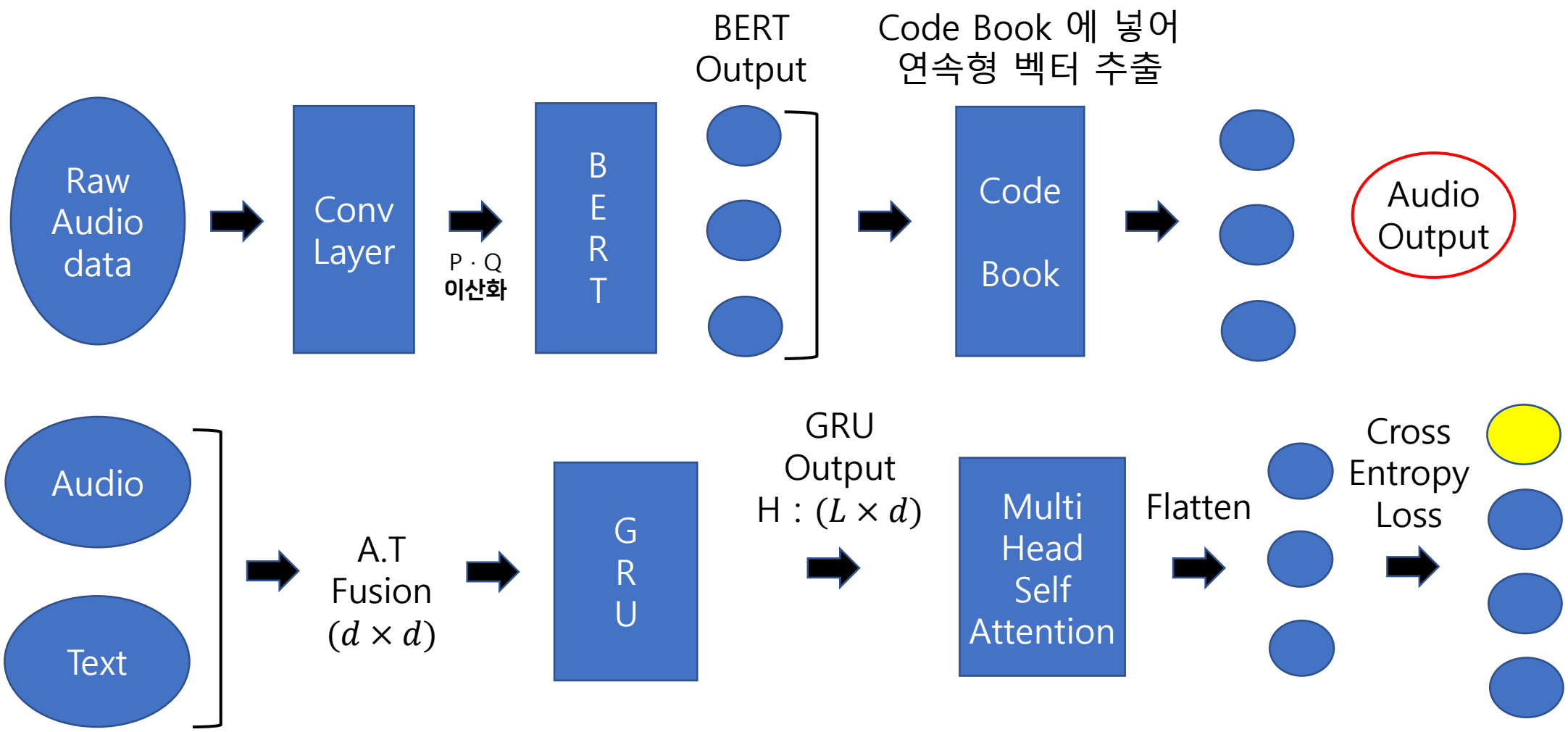
- AI Hub 감성 대화 말뭉치

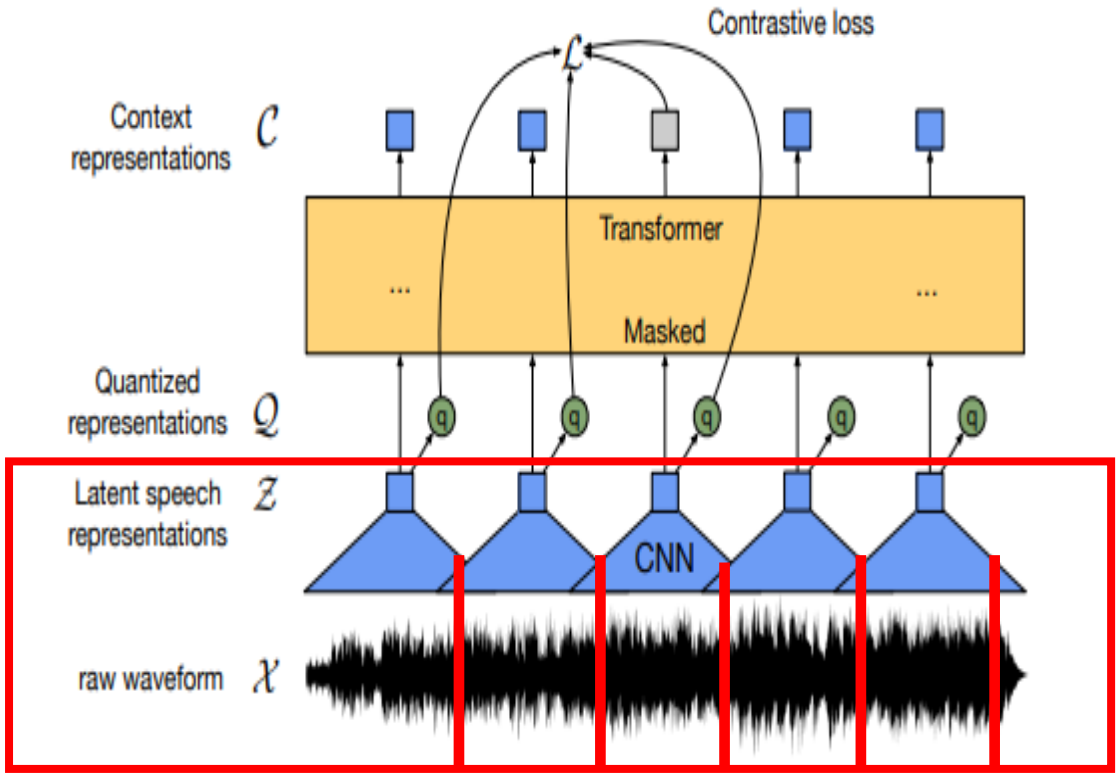
https://aihub.or.kr/aidata/7978

- 일반인 1,500명 대상으로 한 음성 10,000 문장 및 코퍼스 27만 문장

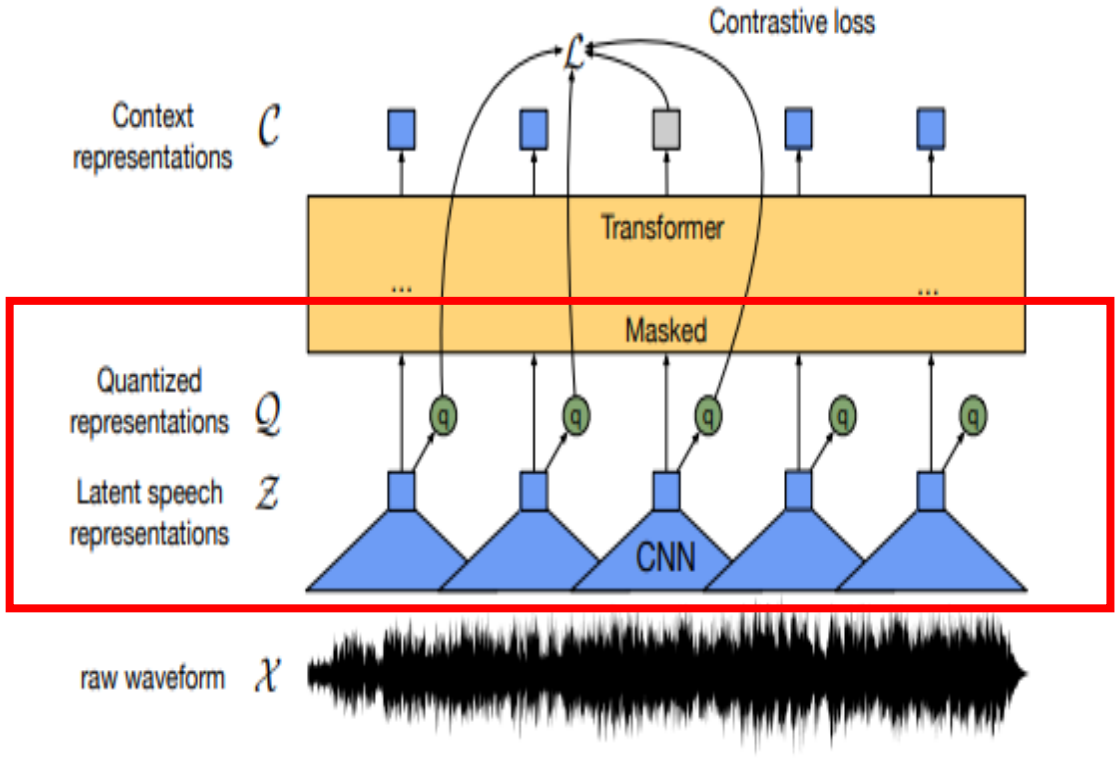
- 60가지의 감정 상태
- 연령, 성별의 다양성

- 음성 데이터의 공간적 정보를 담기 위해 Convolution layer 통과
- 해당 Input 을 Wav2vec 2.0 모델에 넣어 Self-supervised Learning
- 이후 나온 Vector 를 Audio Feature 로 간주하여 Text data 와 Fusion 진행
- GRU, Multi-head Attention Layer 에 통과시켜 Slow feature 를 포함한 Latent Vector 추출
- 해당 Vector 로 감정,화자에 대해 Classification 진행, Negative Sampling 으로 감정 및 화자 예측


Raw Audio data
Conv Layer
P · Q
이산화
BERT
BERT Output
Code Book 에 넣어 연속형 벡터 추출
Code Book
Audio Output
Audio
Text
A.T Fusion $(d \times d)$
GRU
GRU Output H : $(L \times d)$
Multi Head Self Attention
Flatten
Cross Entropy Loss



- 30ms음소 단위로 잘라서 Input으로 사용
- 음성 데이터를 5개의 Convolution Layer 을 통과 시킴
  Kernel size : (10,8,4,4,4) Strides : (5,4,2,2,2)
- 512개의 Channel 을 사용하며, ReLU 를 사용함
- Group Normalization 적용

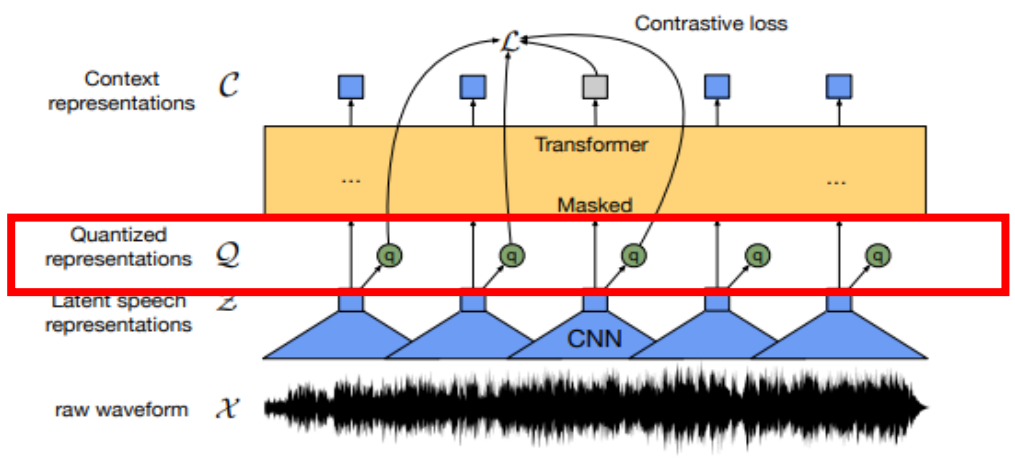


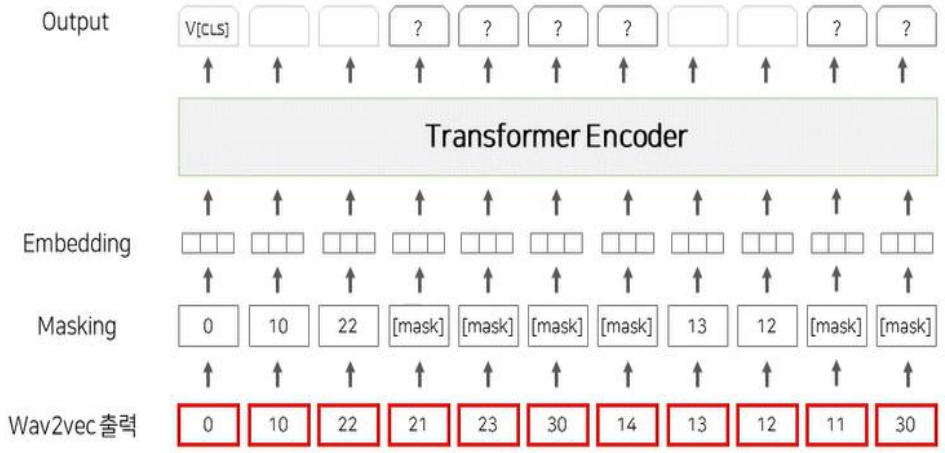
- Speech 의 Representation 을 표현한 Vector Z 를 양자화 시킴

  (Product Quantization)

- 이산화 된 벡터를 BERT 의 MLM 학습 방법 적용

- 앞 과정을 통해 Speech 안의 Slow feature 를 포함한 Vector 구함

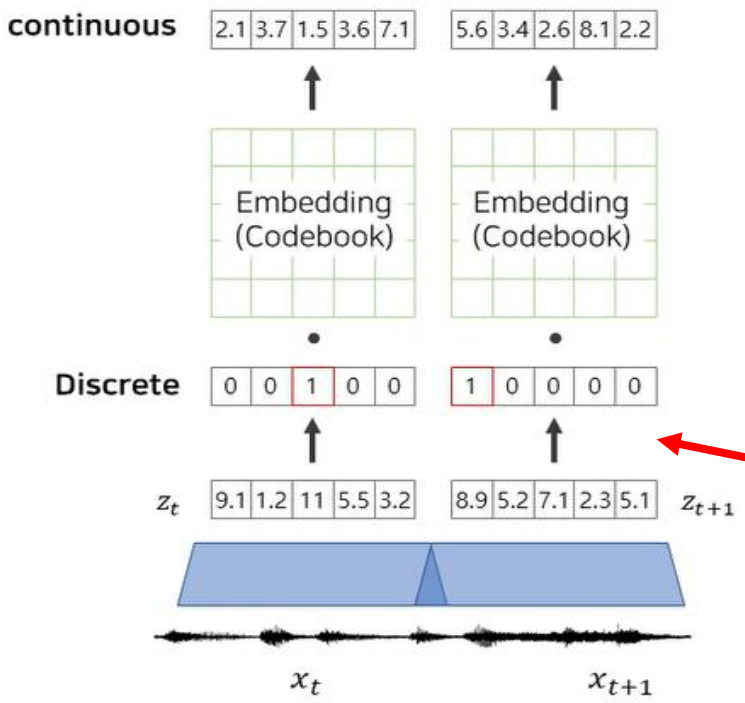




- Product Quantization: Self supervised Learning 을 하기 위해 벡터를 이산화 하여 제한된 표현으로 만들어 학습 진행
- Convolution 을 거친 Vector z 를 코드북 G를 통해 표현
- G 개의 R^($V \times d/G$) 벡터를 해당 e_1, ..., e_G 벡터에서 하나씩 선택
- 선택된 벡터를 단순 Concat 후 R^($V \times d$) 를 R^($V \times f$) 로 단순 선형 변환

$$p_{g,v} = \frac{\exp(l_{g,v} + n_v)/\tau}{\sum_{k=1}^{V} \exp(l_{g,k} + n_k)/\tau},$$

- 각 K 마다 해당 p가 최대가 되는 Codeword 를 선택, Argmax 는 학습이 불가능 하므로 위 식과 같은 Gumbel-Softmax 진행



- MLM 을 통한 학습으로 이산화 된 Token 을 맞추는 학습 진행
- Loss: CTC Loss 사용 $\mathcal{L} = \mathcal{L}_m + \alpha\mathcal{L}_d$ (α는 하이퍼파라미터)
- 학습이 진행된 Vector 를 학습된 Codebook 을 통해 다시 수치형 Vector 로 바꿈
- 해당 Vector 가 이후 모델의 Audio Vector 가 됨 (Slow Feature 가 학습 됨)

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

MLM 학습을 통해 나온 Vector C 와 이산화 한 Input q 와의 유사도를 구함.
이후 k 개의 q ~(negative sample) 을 구한 후 Q 와의 유사도가 최대가 되도록 학습 Loss 정의

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}$$

전체 Code Book 에서 특정 Entry에만 접근하는 것이 아니라 골고루 접근하게 하기 위해 Loss 추가

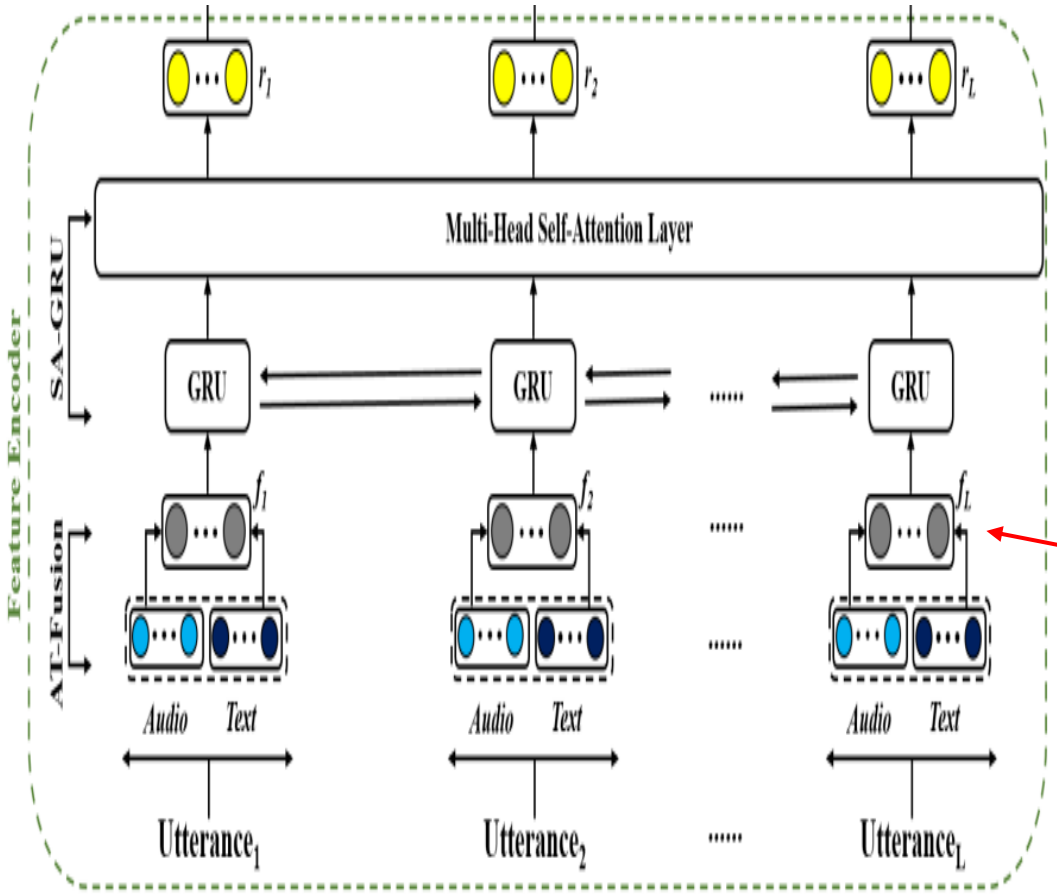


Figure 1: *Overall structure of the proposed framework.*

- Audio Feature와 해당 Text Feature를 결합

$u_i^{cat} = Concat(W_a a_i, W_t t_i)$ 학습 가능한 가중치 행렬을 부여하여 Concat

$\alpha_{fuse} = softmax(w_F^T tanh(W_F u_i^{cat}))$ u_i 에 Softmax 함수를 통과하여 Fusion 시 Audio, text 비율 학습

$f_i = u_i^{cat} \alpha_{fuse}^T$ f_i (dXd) Vector 를 추출
f_i를 GRU 에 통과시켜 $H \in R^{(L \times d)}$ [h_1,…,h_L] 생성

$$head_i = softmax((HW_i^Q)(HW_i^K)^T)((HW_i^V)$$

H 벡터에 Multi-Head-Self-Attention 진행 후 최종 $R^{'} \epsilon R^{(L \times d)}$ 생성

- 이후 각 j 번째 벡터 마다 M 개의 Labeling 된 감정 및 화자 분류의 Cross Entropy를 계산

- $L_y = \sum_{i=1}^{M}\sum_{j=1}^{L_i} -logP(e_j|r_j)$ 각 발화 데이터에서 Labeling 된 감정 or 화자를 예측

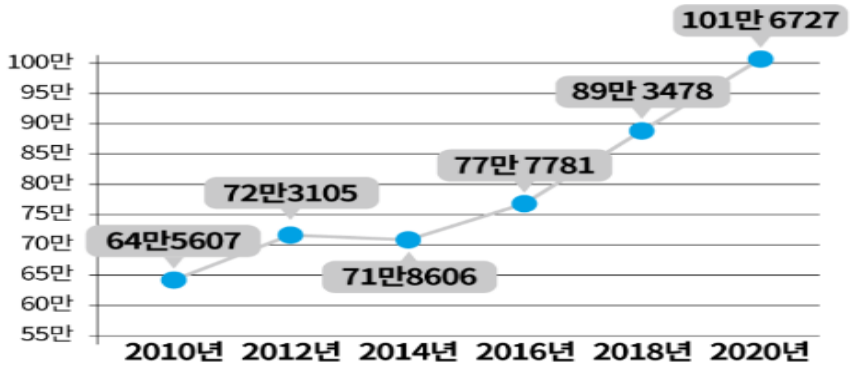
- 임베딩 Dimension 은 앞선 Code book 의 Dimension f 와 동일
- L2 정규화와 p= 0.2 의 Drop out 진행 , Num_head = 4
- Audio Feature 에 단순 Convolution Layer 를 통과한 피처를 적용 시 감정 인식 분야에서 82.68% 의 State of the art 모델 달성
- Slow Feature 를 추가한 Audio Feature 를 적용하면 성능 향상 예상

데이터 라벨링 세팅에 따른 성능 변화

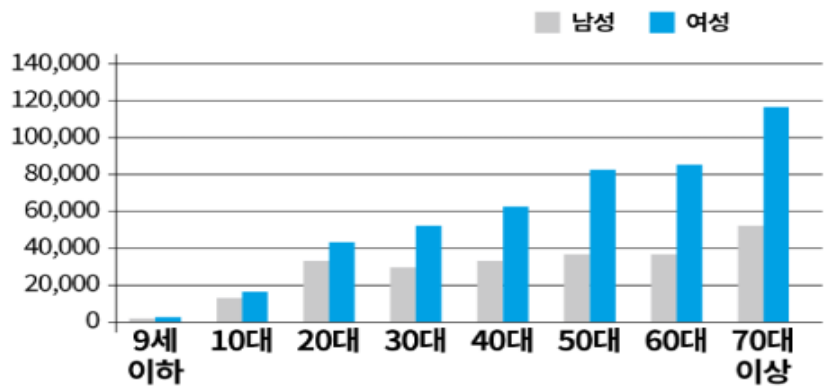
	TS_1234	TS_123	TS_134	TS_234	TS_23
Cmp	81.06	80.82	79.85	78.89	77.60
Our	**81.14**	**82.68**	**82.27**	**82.43**	**81.39**
Δ	+0.08	+1.86	+2.42	+3.54	+3.79

Approaches	WA (%)
Abdelwahab et al. (2018) [19]	56.68
Li et al. (2019) [18]	58.62
Rozgić et al. (2012) [29]	67.40
Jin et al. (2015) [30]	69.20
Poria et al. (2017) [11]	74.31
Li et al. (2018) [31]	74.80
Hazarika et al. (2018) [17]	77.62
Li et al. (2019) [32]	79.20
Proposed method	**82.68**

[연도별 '우울증' 진료인원 현황] (단위 : 명)



[연령별 '우울증' 진료인원 현황]



1. 우울증 예측 및 예방 (feat. 감정분석)

- AI 음성인식 기술을 통해 사용자의 음성데이터 감정분석을 통해 우울증 사전 예측

- 핸드폰이나 IoT등 데이터를 수집할 수 있는 기기에서 화자의 내재된 감정을 예측하여 우울증 예방 수칙을 행동할 수 있음

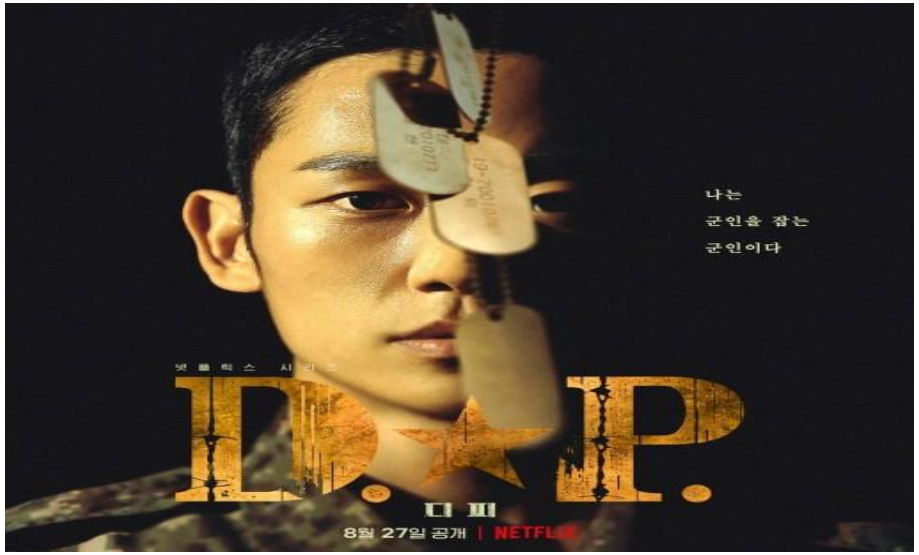
음성인식을 통한 우울증 케어 사례

CLOVA CareCall

- 네이버 클로바(Care Call) AI가 독거 어르신들의 안부를 확인하고 말벗이 되는 서비스
- 대화에서 감정을 파악 이에 맞춰 Care Call 서비스 제공
- 시범 사업 대상의 95%가 지속적으로 서비스를 이용하고 싶다고 응답

2. 영상 컨텐츠 AI성우

- 한국 컨텐츠가 세계적으로 주목을 받게 되면서 외국인 시청 비율 증가
- 현재는 컨텐츠에 자막을 통해 시청하므로 성우의 음성과 감정을 생생하게 전달 받기 어려움
- 휴멜로 회사는 배우의 음성 특징들을 잘 잡아, 외국어로 번역하여 성우를 제공하는 서비스 제공



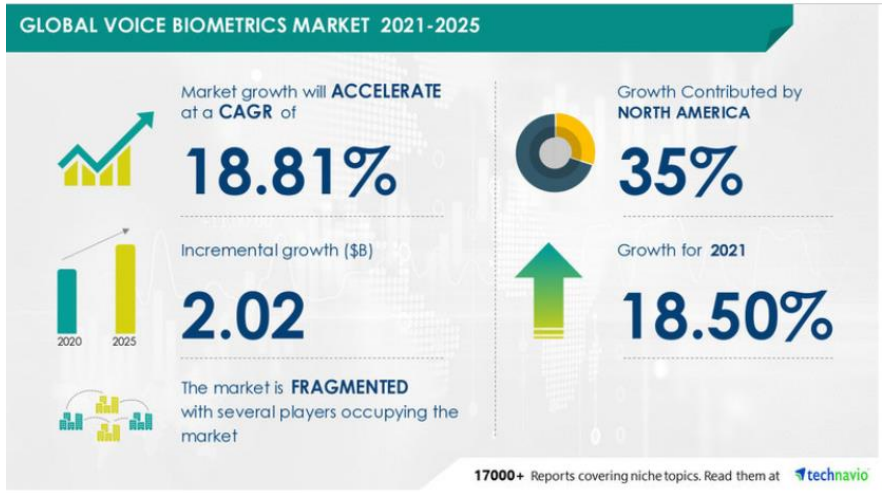
AI 성우 사례

- 인공지능이 음성 데이터를 학습 한 후 감정을 입혀 목소리를 추출
- 일반 목소리 + 잠재 벡터(감정이 담긴 벡터)를 학습한 후 감정을 담은 노래나 호소력 있는 연기 톤을 학습
- 현재 9가지 감정(화남, 차분함, 실망, 흥분, 공포, 행복, 중립, 슬픔, 졸림) 숫자를 개발했으며 한국 컨텐츠 더빙을 영어 및 중국어까지 번역하여 제공할 계획

3. 본인 인증(Voice Biometric)

- 세계적으로 음성을 통한 보안 인식 시장 규모는 18.1% 씩 증가, 209억 달러의 시장 규모가 될 예정
- 대표적인 기업 Pindrop은 전화 통화를 통해 본인 식별 및 사기 감지를 하고 전화 통화에 대한 위험 점수를 제공하는 형식의 서비스를 제공
- 최근에는 Google Cloud 와 ISV 파트너십을 맺어 음성 보안 및 사기 방지 솔루션을 클라우드 플랫폼에 내에서 구축

Pindrop®



AI 화자 인식 사례

- 미리 녹음된 사용자의 목소리와 발성을 비교하여 신원확인을 하는 '화자 인식' 기술
- 보컬 패스워드 기술처럼 고정된 암호문을 얘기하는 것과 달리 자유롭게 상담원과 전화 하는 과정에서 녹음된 목소리의 특징(억양, 말 빠르기) 활용
- 실시간으로 본인 여부를 탐지하기 때문에 사칭 전화 감지 가능
- 플레이백 채널 디텍션(Playback Channel Detection) 기술을 활용 해 녹음본으로 사칭하려는 음성 시도도 방지 가능

참고 문헌

CONTENTS

참고 사이트

https://drive.google.com/file/u/1/d/1u03zPLpTBG0cIlA5YXISElfH6G-xmwt_/view?usp=sharing (이병헌 목소리-영중일어 오디오 샘플)

https://clova.ai/aicontactcenter (네이버 클로바 Care Call)

https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002684238

(음성 인식과 텍스트 확장 알고리즘을 이용한 인물 성향 분석에 관한 연구 -영화 내 주연 배우의 성향 중심으로)

https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO202128837807055&dbt=NART

(음성감정인식 성능 향상을 위한 트랜스포머 기반 전이학습 및 다중작업학습)

https://www.mk.co.kr/news/culture/view/2022/01/36606/ (한국 컨텐츠 관련 기사)

https://www.chosun.com/economy/smb-venture/2021/12/09/RZ2CQQQ4DFFLNKGV32NJD737BE/ (휴멜로 관련 기사)

https://it.donga.com/23864/(뉘앙스 커뮤니케이션즈 관련 기사)

http://www.aitimes.kr/news/articleView.html?idxno=24171(음성인식을 활용한 핀테크 분야 기사)

참고 논문

https://arxiv.org/pdf/1508.01211.pdf

https://www.cs.toronto.edu/~graves/icml_2006.pdf

https://arxiv.org/abs/1807.03748

https://arxiv.org/abs/1904.05862

https://www.koreascience.or.kr/article/JAKO202128837810056.pdf

https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002761154

Thank you

3조 - 나요셉, 김재민, 최원석, 조기흠