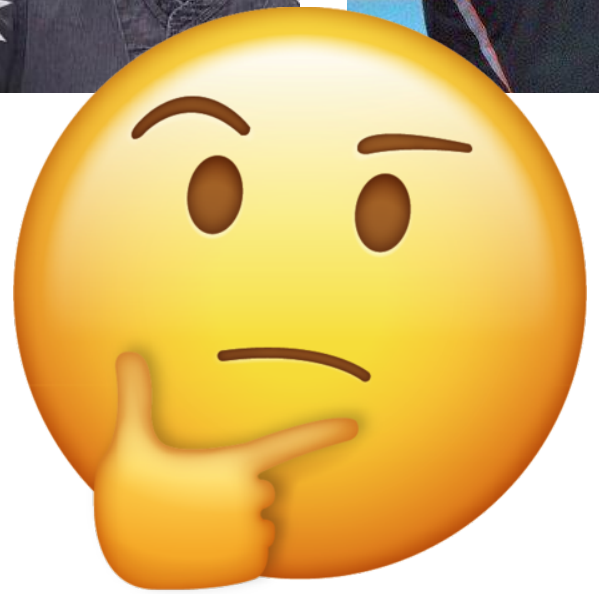
LangCon 2021

말로 하는 감정 인식

바벨피쉬 송치성

Somewhere over the Text

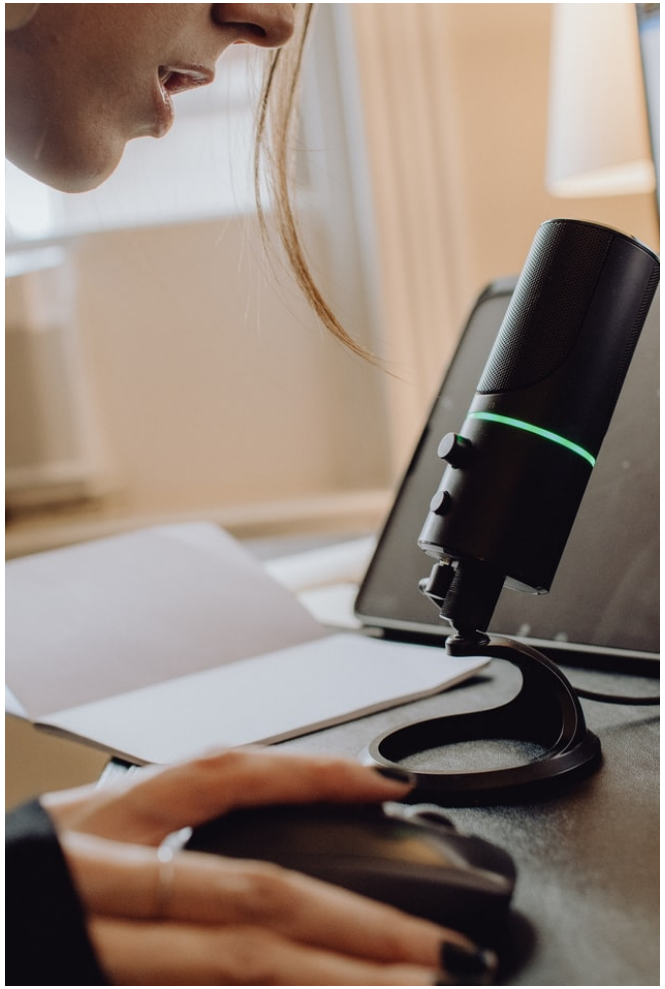
Language Conference



Somewhere over the Text

TEXT

AUDIO



SPEECH EMOTION RECOGNITION

1. Emotion?
2. Representations of Emotion
3. Speech Emotion Recognition

SER CASE STUDY

1. SER Dataset
2. Model Architecture
3. Practice

SPEECH EMOTION RECOGNITION

LangCon 2021

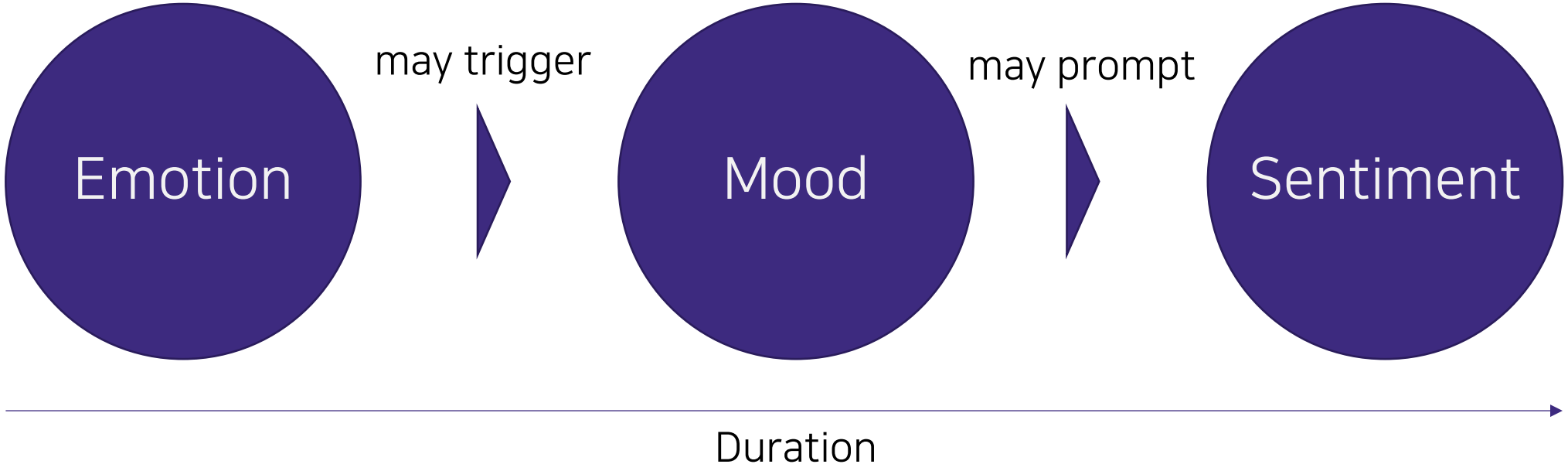
- Emotion :
  - 사건, 사물, 사람 등 어떤 대상에 의해 발생
  - 일반적으로 식별되는 얼굴 표정 등
    생리적, 정서적, 행동적, 인지적 요소를 포함
  - 지속시간이 비교적 짧음

EMOTION is different from...

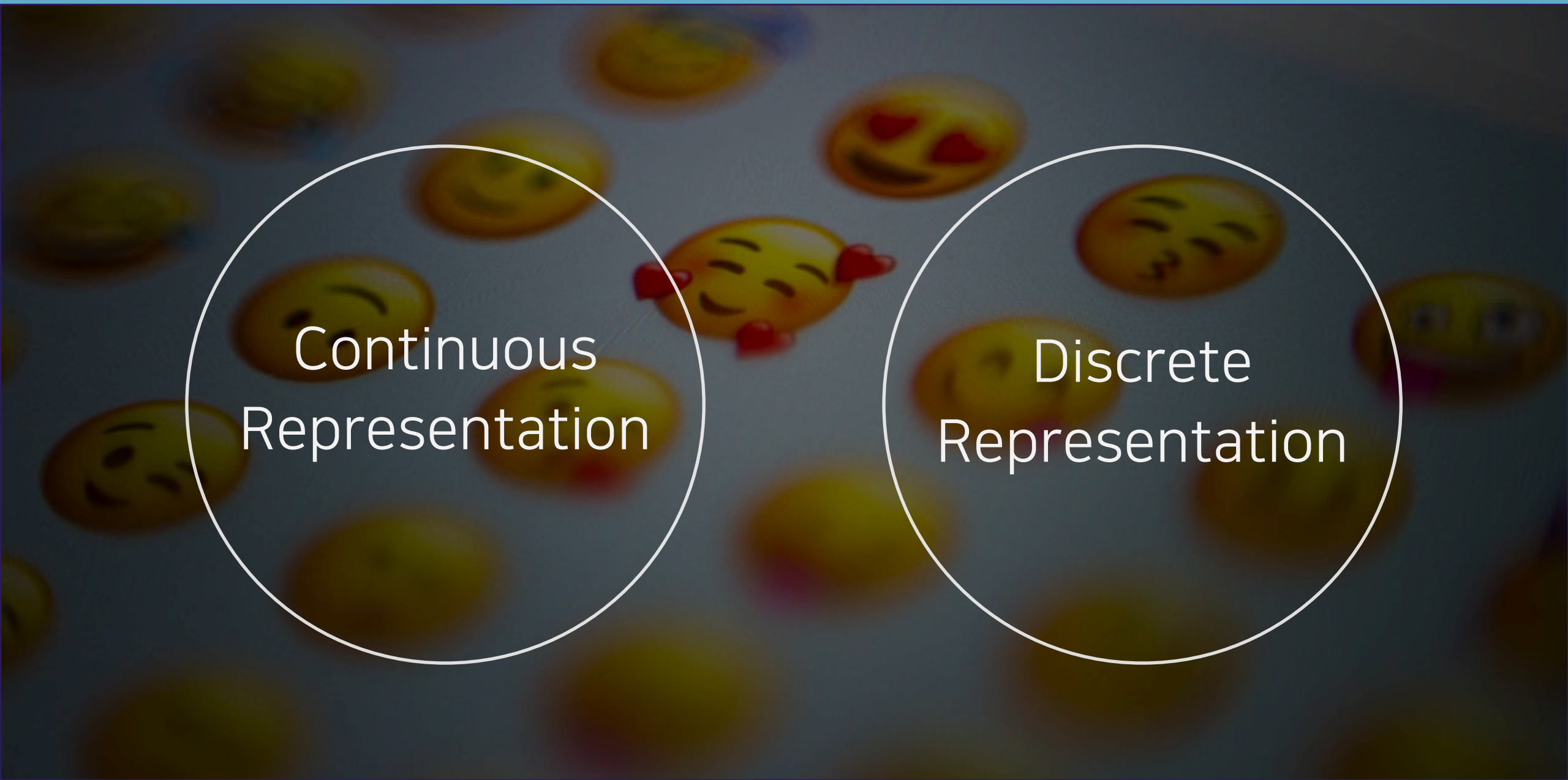
Language Conference

- **Mood :**
  - 원인이 전반적이며 명확하지 않음
  - Emotion보다 오래 지속
  - E.g. 너 때문에 슬프다 (Emotion) vs 우울하다 (Mood)

- **Sentiment :**
  - 어떤 대상에 대한 긍정/부정 반응
  - 판단의 근거는 직접적인 경험과 그 후의 일반화에서 나오지만, social learning을 통해서도 이루어짐



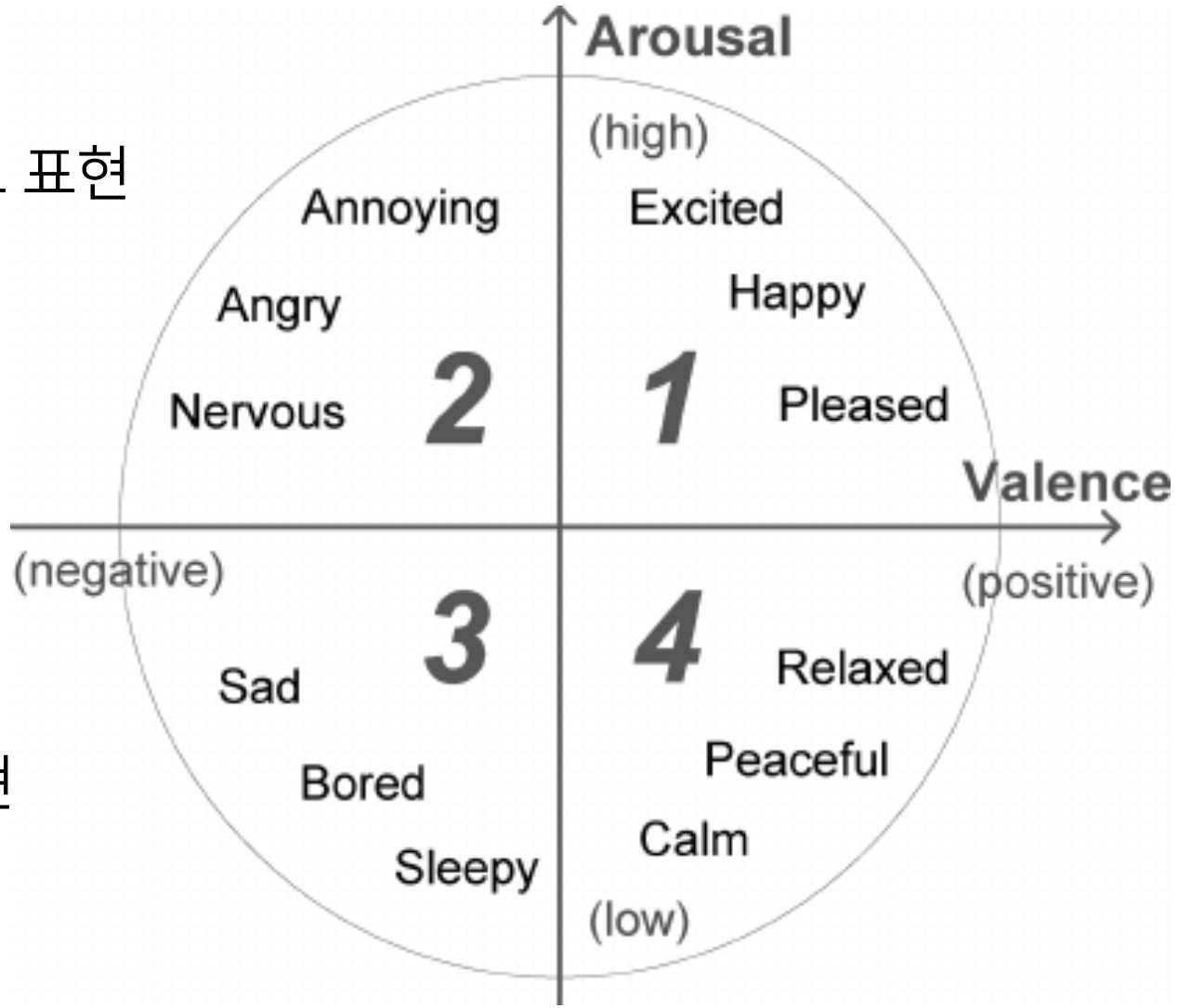
Representations of Emotion



Representations of Emotion

Continuous

- 감정을 Arousal, Valence 두가지 차원으로 표현
  - Arousal : 감정의 강도가 큰지/작은지 (각성/이완)
  - Valence : 긍정인지/부정인지 (유쾌/불쾌)



Discrete

- 감정을 이산적이고 특정한 카테고리로 표현
  - E.g. angry, sad, happy…

Speech Emotion Recognition (SER)

- 발화의 음성신호로부터 화자의 감정 상태를 인식
- 음성인식 결과(텍스트)의 정보를 활용하기도함

SER의 활용

- 스마트 스피커, 음성비서 시스템에서 음성언어 이해(SLU) 성능을 개선
- 이러닝, 게임, 인터랙티브 무비 등에서 사용자의 상태 파악
- 항공/자동차 운전자 상태 파악하여 사고 예방

Speech Emotion Recognition

Features for SER

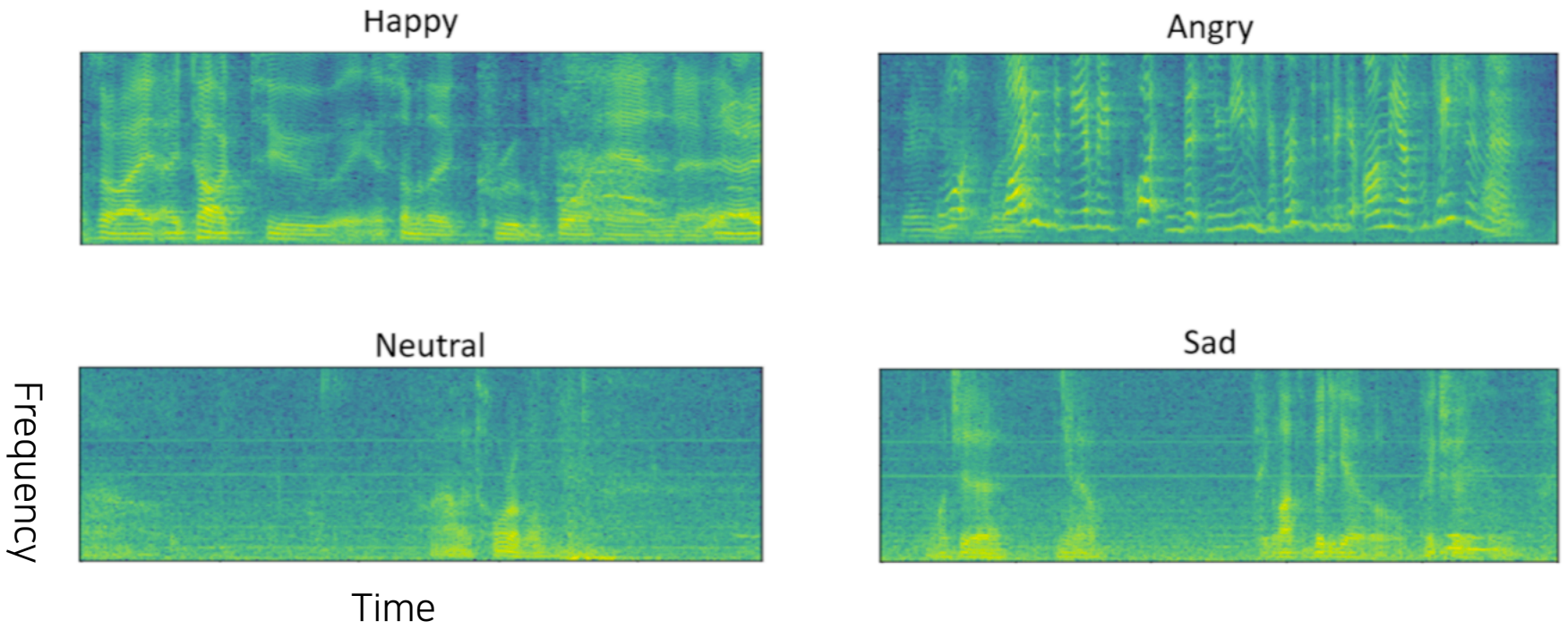
- Lexical Feature :
  - Content
- Acoustic Feature :
  - Spectral Feature, Prosody, Pitch, Voice quality…

TABLE 4.2. Voice and Emotion

	Fear	Anger	Sadness	Happiness	Disgust
Speech rate	Much faster	Slightly faster	Slightly slower	Faster or slower	Very much slower
Pitch average	Very much higher	Very much higher	Slightly lower	Much higher	Very much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Normal	Higher	Lower	Higher	Lower
Voice quality	Irregular voicing	Breathy chest tone	Resonant	Breathy blaring	Grumbled chest tone
Pitch changes	Normal	Abrupt on stressed syllables	Downward inflections	Smooth upward inflections	Wide downward terminal inflections
Articulation	Precise	Tense	Slurring	Normal	Normal

Features for SER

- Acoustic Feature :
  - Spectral Feature, Prosody, Pitch, Voice quality…



Difficulties

- 딥러닝하기에 충분한 양의 데이터 부재 (e.g. ImageNet)
- 연기자가 연기한(simulated) 데이터셋이 많음 -> 실제와 다름
- 문화 및 언어에 종속적
- 주석의 불확실성 (주관성)
- 실제 환경과 다름 (크로스톡)

SER CASE STUDY

LangCon 2021

AI Hub 감정 분류를 위한 대화 음성 데이터셋

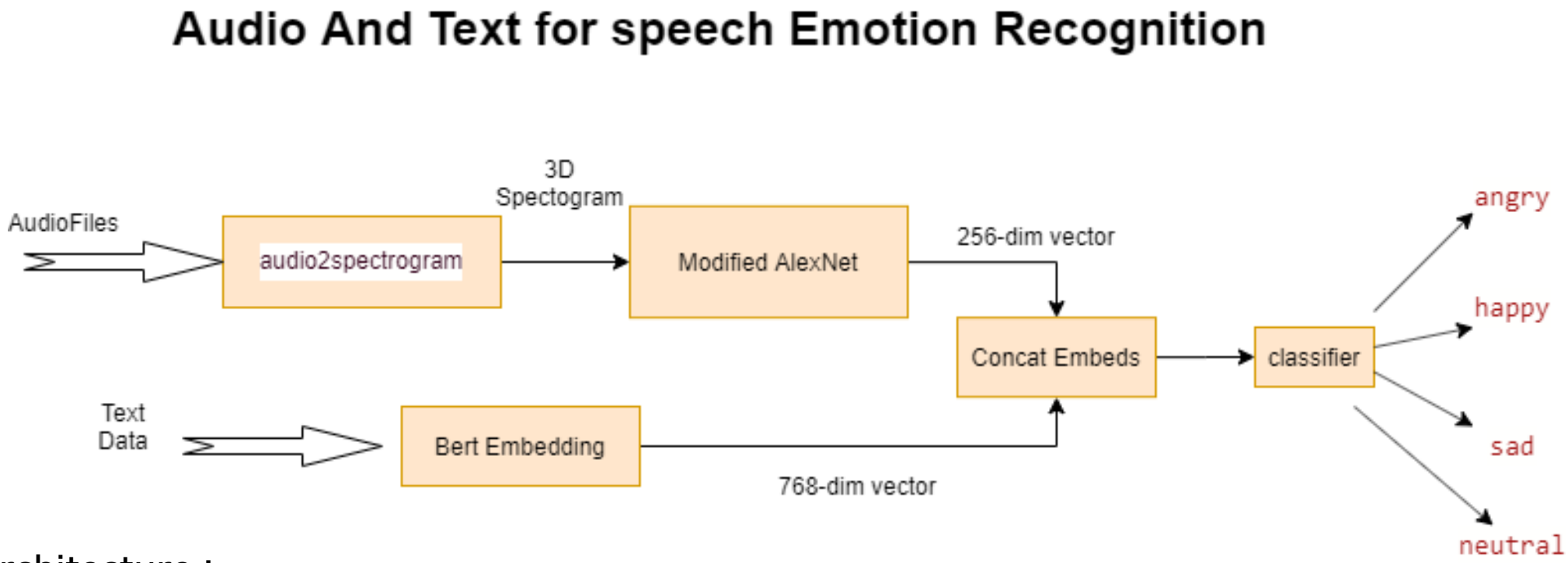
- 영어 감정 인식 멀티모달 데이터셋은 IEMOCAP, CMU-MOSEI이 있지만, 적절한 한국어 데이터셋은 없음
- 구축 내용 :
  - 감성대화 어플리케이션을 이용한 수집
  - 일정 기간동안 사용자들이 어플리케이션과 자연스럽게 대화하고,
    수집된 데이터를 정제 작업을 거쳐 선별
  - 7가지 감정에 대해 5명이 각각 레이블링

    - **감정 레이블 : happiness, angry, disgust, fear, neutral, sadness, surprise**
- 데이터 :
  - 규모 : 약 44K
  - 구조 : 음성 포맷 : 16bit, 48kHz wav 파일
  - csv 파일 메타정보 : 대화의 상황, 음성인식 결과, 감정 레이블, 화자의 성별/나이 정보
- 링크 : https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-002

AI Hub 감정 분류를 위한 대화 음성 데이터셋

wav_id	발화문	상황	1번 감정	1번 감정세기	2번 감정	2번 감정세기	3번 감정	3번 감정세기	4번 감정	4번감정세기	5번 감정	5번 감정세기
5ee1e7939aa8ea0eec53fac7	나 어제 헤어졌어	sad	Sadness	1	Sadness	1	Sadness	1	Sadness	1	Sadness	1
5ee1e7a379bf120ed2b81ba2	어쩌다 보니까 그렇게 됐네	sad	Sadness	1	Sadness	1	Neutral	0	Sadness	1	Sadness	1
5ed9793b9aa8ea0eec53f7f1	유기견 다큐멘터리를 봤는데 무책임한 사람들 때문에 너무 화가 나	disgust	Sadness	2	Sadness	2	Angry	1	Angry	1	Angry	1
5ed979582880d70f286128c3	버려지는 유기견들의 생활을 다른 다큐멘터리 없어	disgust	Sadness	1	Sadness	2	Sadness	1	Angry	1	Sadness	2
5ed9797b1dcf350eeded50b8	작년보다 5배는 더 늘어나는 것 같대 최근 들어 더 늘어나고 있고	disgust	Fear	1	Sadness	1	Sadness	1	Angry	1	Sadness	2
5ed9799ac90a530ee56b5f6b	정부보조금이랑 사람들의 기부금으로 운영되고 있어	disgust	Neutral	0	Neutral	0	Sadness	1	Sadness	1	Sadness	2
5ed979dc7e21a10eee253e90	맞아 벌어지는 대부분의 아이들이 다 큰 강아지 들였어 아직도 상처	disgust	Sadness	2	Sadness	2	Sadness	2	Sadness	2	Sadness	2
5ed97a22c90a530ee56b5f6c	어제저녁에 진짜 무서웠어	fear	Fear	2	Fear	2	Fear	1	Fear	1	Fear	2
5ed97a381dcf350eeded50bc	친구랑 약속 있어 나갔는데 밥 먹고 이따가 지진이 발생하는 거야 얼	fear	Fear	2	Fear	2	Fear	2	Fear	1	Fear	1
5ed97a5cc90a530ee56b5f6d	큰 지진은 지나갔는데 여진이 조금씩 있는 거 같기도 해	fear	Fear	2	Fear	2	Fear	1	Fear	1	Fear	1
5ed97a779aa8ea0eec53f7f5	다른 테이블에서 밥 먹던 사람들이 도망치다가 넘어지고 그 와중에	fear	Sadness	1	Fear	2	Sadness	1	Fear	1	Fear	1
5ed97a977e21a10eee253e92	그렇지 않아도 오늘 친구 데리고 병원 가기로 했어	fear	Neutral	0	Sadness	1	Sadness	1	Sadness	1	Fear	2
5ed97ac29aa8ea0eec53f7f8	룸메이트와 너무 자주 싸우게 돼	anger	Disgust	2	Angry	2	Angry	1	Angry	1	Fear	1

Model Architecture



- Architecture :
  - 임베딩을 얻기위해 Audio와 Text로 각각 모델 학습
  - 임베딩은 Concat되어서 Classification Layer에 연결
  - 파인튜닝할 때에는 Classification Layer만 학습
- Code : https://github.com/aris-ai/Audio-and-text-based-emotion-recognition
- KoBERT 사용 : https://github.com/monologg/KoBERT-Transformers 😍

- Code : https://github.com/daydrill/speech-emotion-recognition

References

- Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models." Sensors 21.4 (2021): 1249.
- Brave, Scott, and Cliff Nass. "Emotion in human-computer interaction." The human-computer interaction handbook. CRC Press, 2007. 103-118.
- Gangamohan, P., Sudarsana Reddy Kadiri, and B. Yegnanarayana. "Analysis of emotional speech—A review." Toward Robotic Socially Believable Behaving Systems-Volume I (2016): 205-238.
- Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung. "Multimodal speech emotion recognition using audio and text." 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018.
- https://github.com/aris-ai/Audio-and-text-based-emotion-recognition

THANK YOU - !

daydrilling@gmail.com | 송치성