

MSP-PODCAST CORPUS

Speech Emotion Recognition in
Naturalistic Conditions Challenge

김동환

목차

Table of Contents

Introduction	03p
MSP-PODCAST의 개괄	
Related work	05p
기존 DB의 한계점	
부자연스러움, 편향, 크기 부족	
MSP-PODCAST	09p
새로운 DB 구축 방법	
ML+Crowdsourcing	
Analysis of DB	18p
MSP-PODCAST 분석	
속성, 카테고리 기반 분석	
Conclusion	21p
결론	

Abstract

[Status Pages](#)

Speech Emotional DB

keywords : Spontaneous, Balanced, Efficient, Scalable

Utilizing podcast recordings combined with machine learning (ML) and crowdsourcing to effectively ensure both efficiency and accuracy in building a balanced emotional database



Carlos Busso, Reza Lotfian, IEEE Transactions on Affective Computing

Intoduction

기존 *Speech Emotional DB*의 문제점

Table Page

항목	세부내용	Database
Size	딥러닝 모델 학습시킬 만큼의 데이터 확보 X, 화자 다양성 X	
Distribution	감정적으로 unbalance한 데이터 (데이터가 수집된 맥락에 따라서 편향이 발생) ex) 콜센터, 로봇과 상호작용하는 아이들 ...	FAU-AIBO VAM call centers
Spontaneous	초기 방법 : 배우들이 정해진 문장을 "연기" 최근 방법 : 두명 이상의 화자 사이의 대화를 시뮬레이션 -> 고비용	IEMO-CAP RECOLDA

Introduction

Background Factors

MSP-PODCAST



- 데이터 수집

무한한 녹음 파일이 존재하는 Podcast 선정
CC license 이용하여 저작권 문제 해결

- 데이터 전처리

팟캐스트를 단일 화자 세그먼트로 분할
침묵 구간, 배경 음악, 소음 구간, 중첩된 발화 제거

- 데이터 품질 (핵심)

ML로 전처리된 세그먼트 감정 분석 (효율성)
이후 크라우드소싱으로 지각적 평가 (정확성)

Related Work

[Analysis Page](#)

Lack of Naturalness

1

배우가 감정을 연기(read)

자연스러움 부족, 일상에서의 모호한
감정 표현이 아닌 "전형적" 행동

2

배우간 대화를 연기(simulate)

읽기 대신 대화형 상호작용으로 감정
유도, 보다 자연스럽지만 여전히 "연
기"라는 한계점

3

Corpus	Size	# Spkr	Type	Lang.
IEMOCAP [10]	12 h26 m	10	acted	English
MSP-IMPROV [19]	9 h35 m	12	acted	English
CREMA-D [2]	7,442 samples	91	acted	English
Chen Bimodal [20]	9,900 samples	100	acted	English
Emo-DB [6]	22 m	10	acted	German
GEMEP [21]	1,260 samples	10	acted	-

Related Work

[Analysis Page](#)

Unbalanced Emotional Content

1

일상 대화의 감정 분포는 불균형

일상 대화는 극단적인 감정이 표현되지 않음. 대다수는 "중립"적 감정

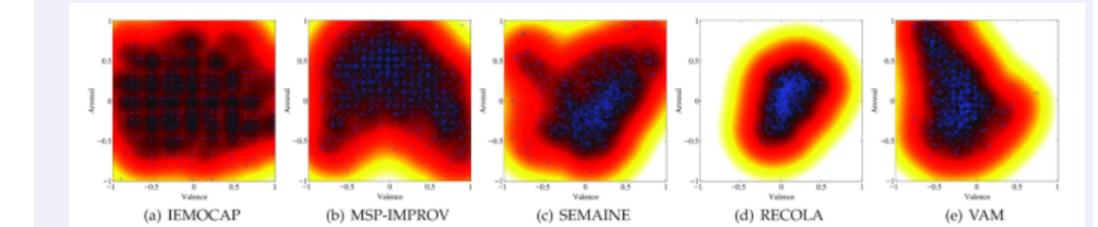
2

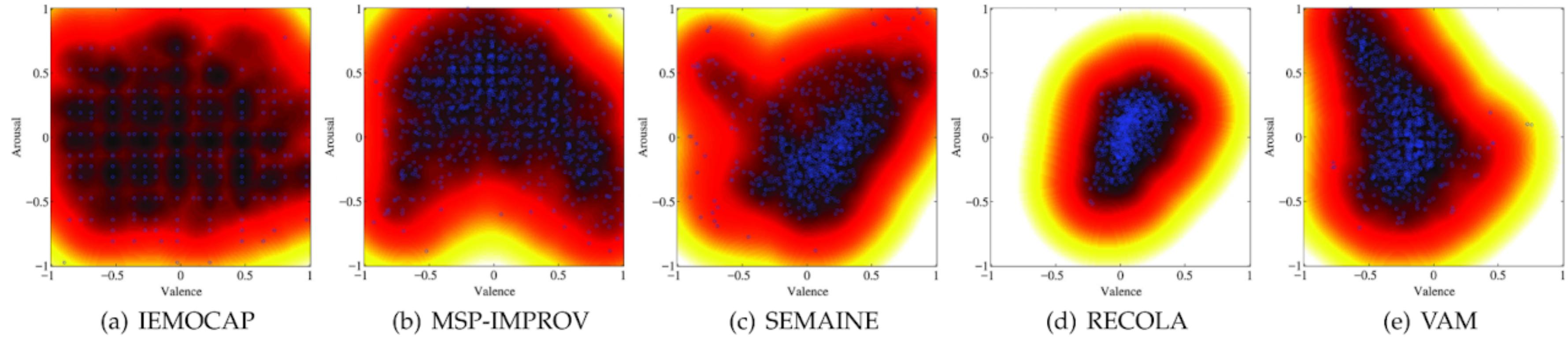
녹음 프로토콜은 감정을 편향시킴

ex) VAM(TV show : fatherhood, affairs, friendship)이면 negative valence로 편향

3

감정별 분포 (뒤에 다시)





1

IEMOCAP, MSP-IMPROV

배우가 연기한 감정을 포함하며, 특정한 연기 프로토콜에 따라 수집

2

SEMAINE, RECOLA, VAM

일상적인 데이터는 편향됨
자연스러움을 얻으려 하면 "균형"을
잃음

3

결론

균형 잡힌 데이터가 중요, 그러나
"자연스러움"도 함께

Related Work

[Analysis Page](#)

Limited size, # of Speakers

1

데이터 셋 크기가 작음

DNN을 학습시키기 위해 충분한
9시간 이상의 데이터셋이 4개에
불과

2

화자 수가 적음

사람들간의 감정 표출 방법이 다르므로
다양한 화자가 포함되어야 하지만,
대부분 50명 미만

3

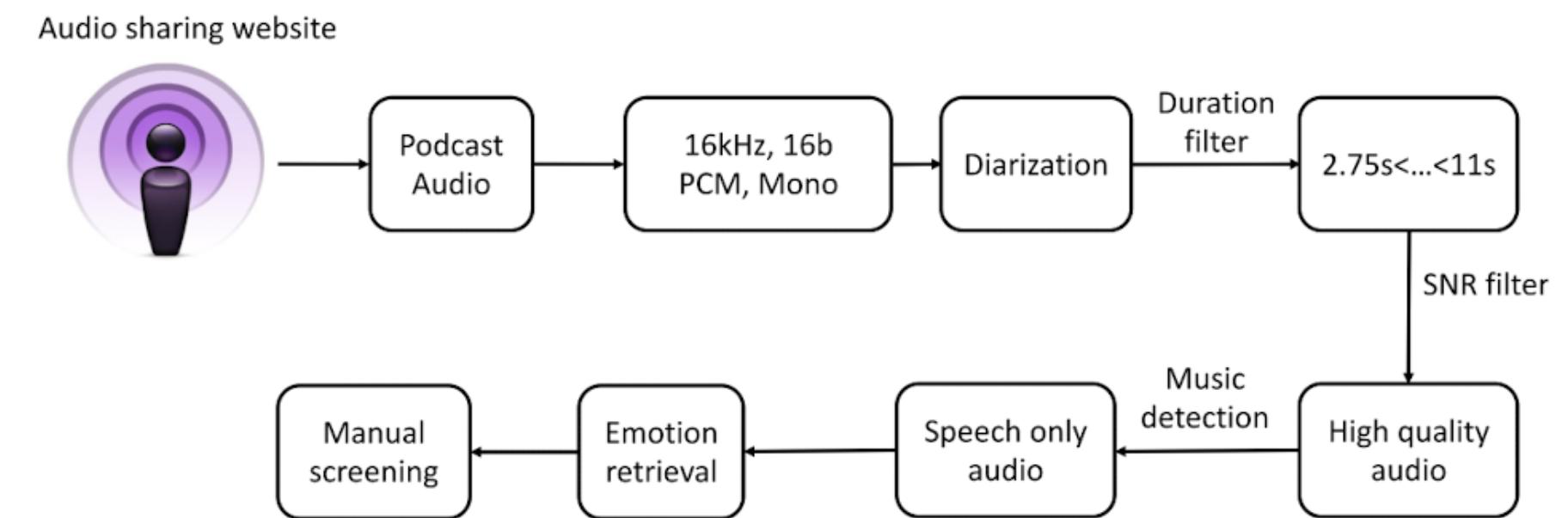
TABLE 1
Summary of Some of the Existing Emotion Corpora

Corpus	Size	# Spkr	Type	Lang.
IEMOCAP [10]	12 h26 m	10	acted	English
MSP-IMPROV [19]	9 h35 m	12	acted	English
CREMA-D [2]	7,442 samples	91	acted	English
Chen Bimodal [20]	9,900 samples	100	acted	English
Emo-DB [6]	22 m	10	acted	German
GEMEP [21]	1,260 samples	10	acted	-
VAM-Audio [15]	48 m	47	spont.	German
TUM AVIC [22]	10 h23 m	21	spont.	English
SEMAINE [13]	6 h21 m	20	spont.	English
FAU-AIBO [14]	9 h12 m	51	spont.	German
RECOLA [11]	2 h50 m	46	spont.	French

MSP-PODCAST

Status Pages

Scale, balanced DB



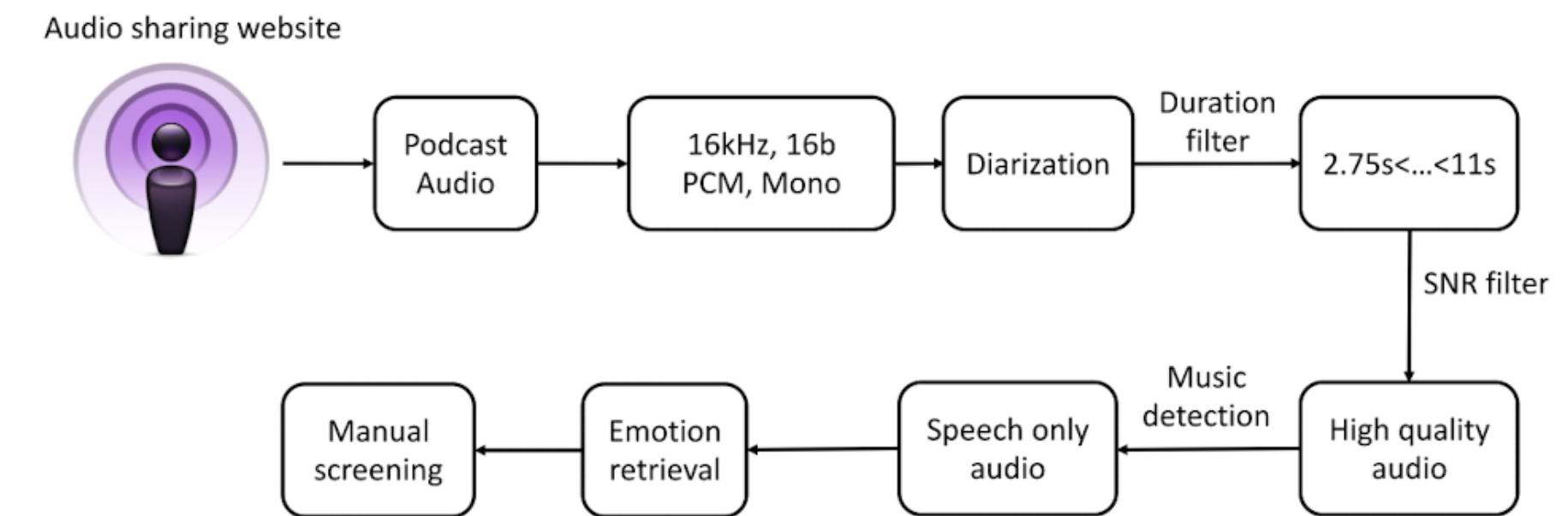
Block diagram of emotional audio speech collection

다양한 주제와 자연스러운 상호작용을 가진 팟캐스트를 활용하여,
감정적으로 균형 잡힌 대규모 음성 데이터베이스를 구축

MSP-PODCAST

Selection of Podcasts

팟캐스트 선택 기준



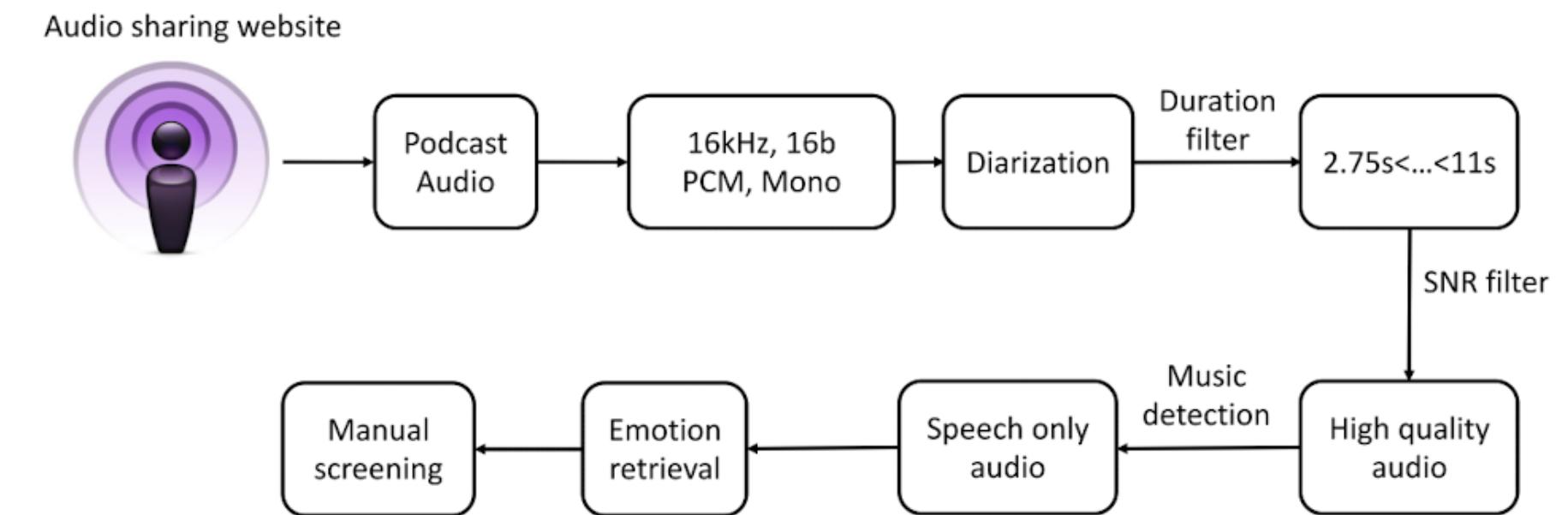
Block diagram of emotional audio speech collection

라이선스 제한이 적은 공개된 팟캐스트를 사용
음악 없는 대화가 포함된 자연스러운 데이터 선호
감정적 다양성을 확보하기 위해 다양한 주제의 대화 선택

MSP-PODCAST

Selecting Candidate Speaking Turns

대화자 선택



Block diagram of emotional audio speech collection

목표 : 화자가 한 명만 있는 세그먼트를 선택

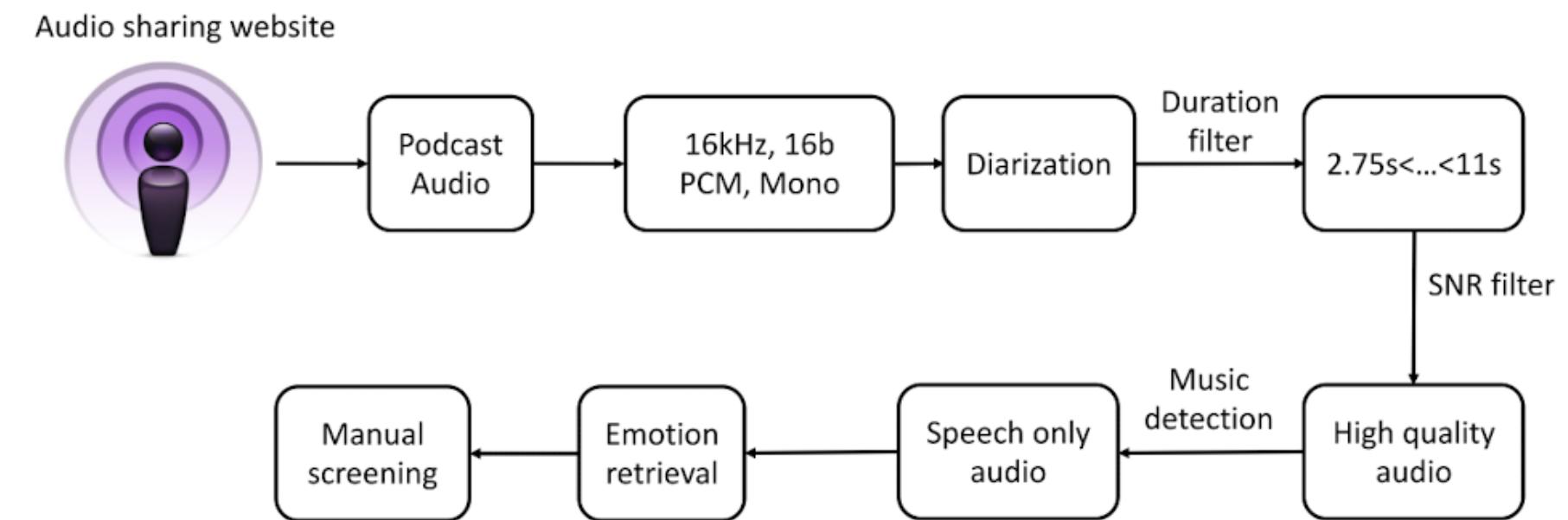
노이즈나 음악이 없고, 겹쳐진 발화 세그먼트를 제외

자동화된 파이프라인을 구현, 선정 기준을 충족하는 후보 세그먼트만 선택

MSP-PODCAST

Selection of Podcasts

데이터 전처리



Block diagram of emotional audio speech collection

다운로드한 팟캐스트를 모노 채널로 변환

샘플링 레이트 16kHz, 비트는 16-bit PCM 형식으로 설정

팟캐스트 녹음은 3분 ~ 190분, 발화 구간과 음악 구간이 포함

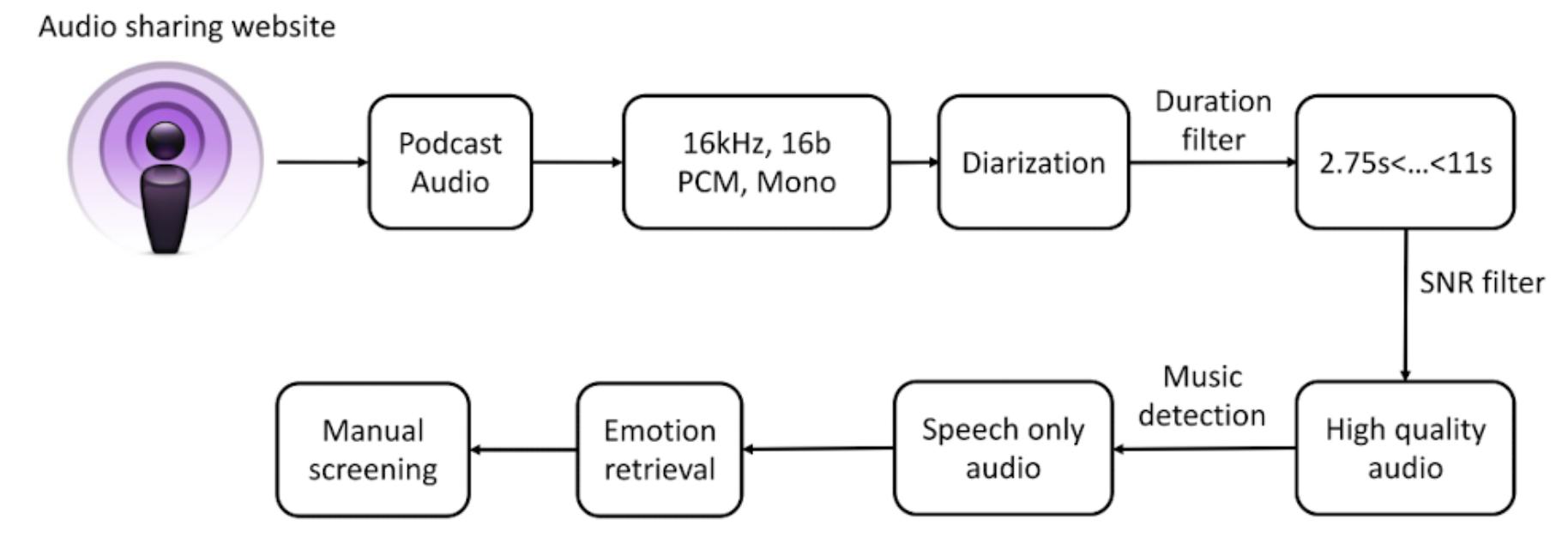
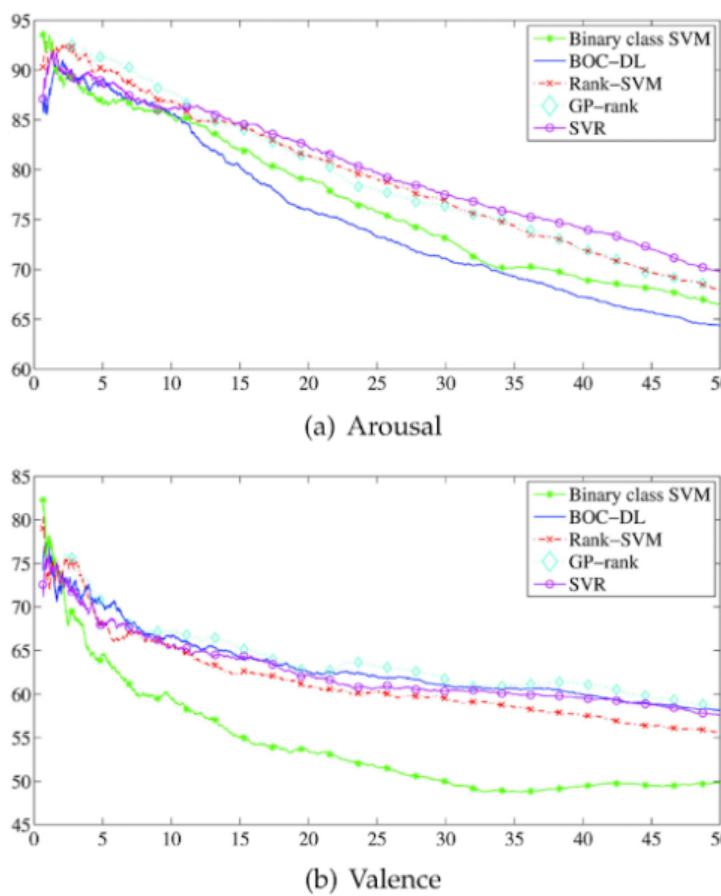
녹음에 한 명 또는 여러 명의 화자가 있을 수 있으므로,

diarization tool을 사용해 짧은 세그먼트로 분할

MSP-PODCAST

Emotion retrieval

EMOETI 검색



Block diagram of emotional audio speech collection

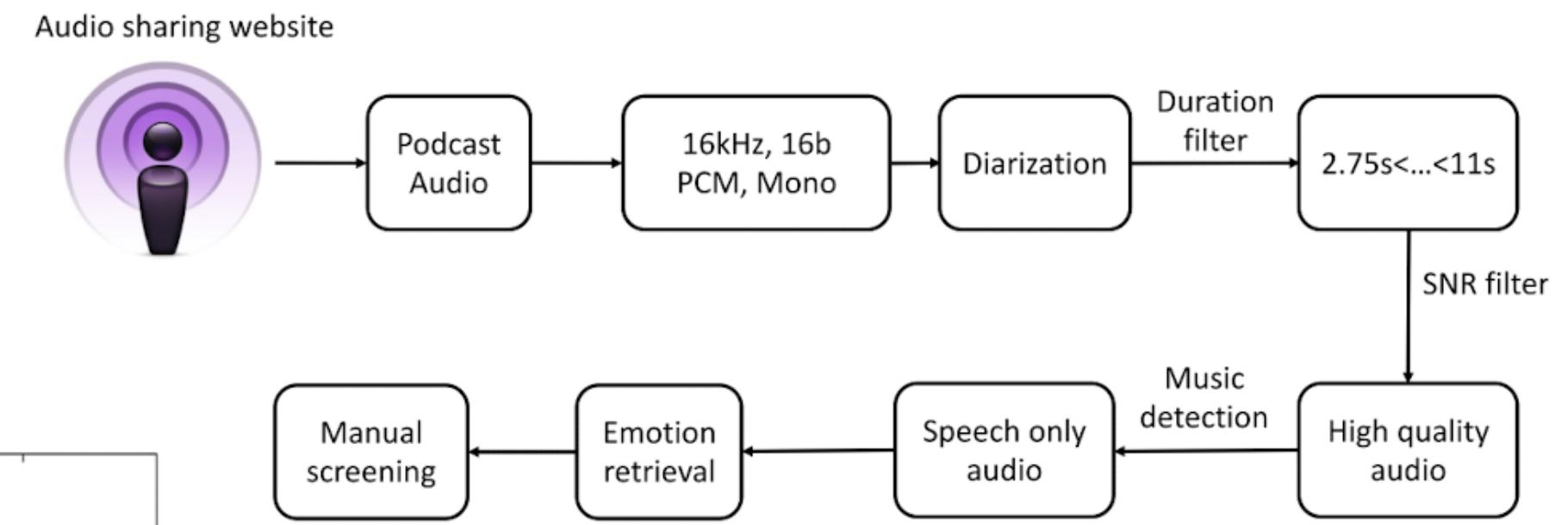
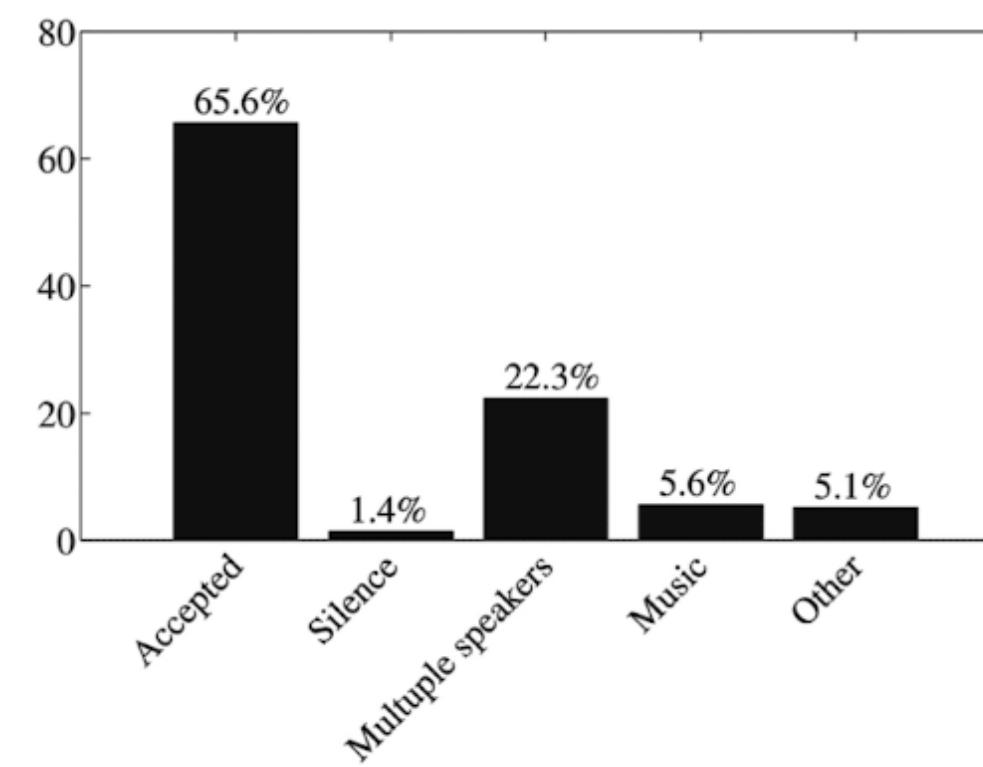
ML 이용하여 감정을 잘 나타내는 세그먼트만 추출
BOC-DL, GP-rank, SVR 알고리즘 사용

MSP-PODCAST

Emotion retrieval

E/I/O/I/E/I 검토

수작업으로 데이터 최종 확인



Block diagram of emotional audio speech collection

Perceptual Evaluations

Using Crowdsourcing

감정 속성 기반 평가

세그먼트를 7점 척도로 평가:

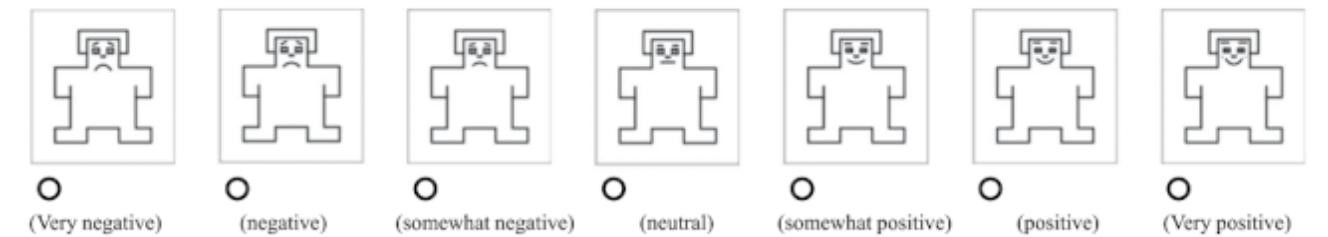
Valence(유쾌함): 매우 부정적(-) ~ 매우 긍정적(+).

Arousal(각성): 매우 침착(-) ~ 매우 활발(+).

Dominance(지배력): 매우 약함(-) ~ 매우 강함(+).

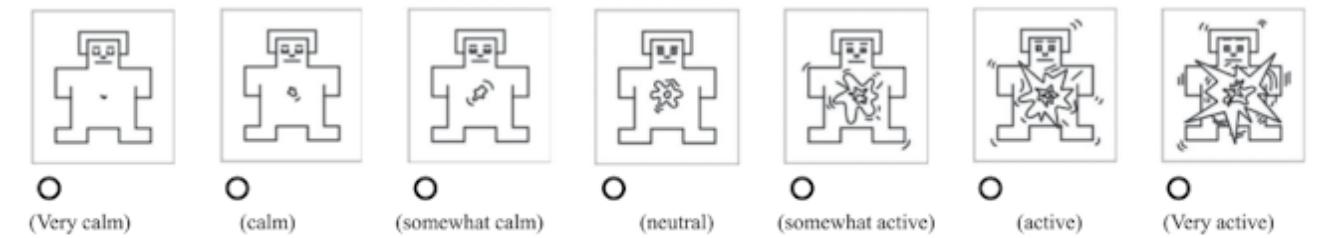
시각적 지표로 SAM(Self-Assessment Manikins)를 활용

Please rate the negative vs. positive aspect of the video
Click on the image that best fits the video.



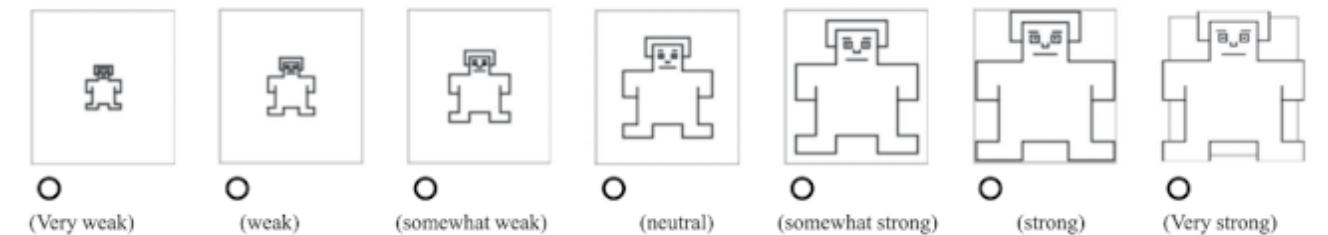
(a) Valence

Please rate the calm vs. excited aspect of the video
Click on the image that best fits the video.



(b) Arousal

Please rate the weak vs. strong aspect of the video
Click on the image that best fits the video.



(c) Dominance

Is any of these emotions the primary emotion in the audio? If not, select Other and specify the emotion.

Angry Sad Happy Surprise Fear Disgust Contempt Neutral Other

(d) Primary emotion

Please pick all the emotional classes that you perceived in the audio (Include the primary emotions selected in previous question)

Angry Sad Happy Amused Neutral
 Frustrated Depressed Surprise Concerned
 Disgust Disappointed Excited Confused
 Annoyed Fear Contempt Other

(e) Secondary emotion

Perceptual Evaluations

Using Crowdsourcing

감정 범주 기반 평가

평가자가 세그먼트에서 느껴지는 주 감정(primary emotion)을 선택

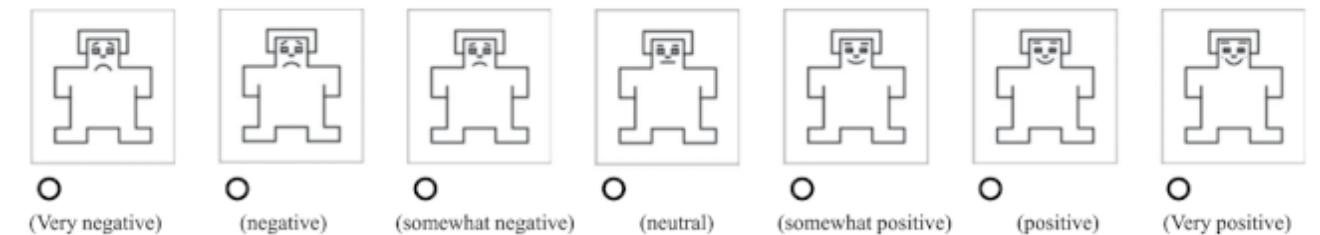
선택지: 분노, 슬픔, 행복, 놀람, 공포, 혐오, 경멸, 중립 상태

추가로 "기타(other)"를 선택 가능 -> 보조 감정이 더 다양함

보조 감정(secondary emotion)도 주석으로 추가 가능(복수 가능)

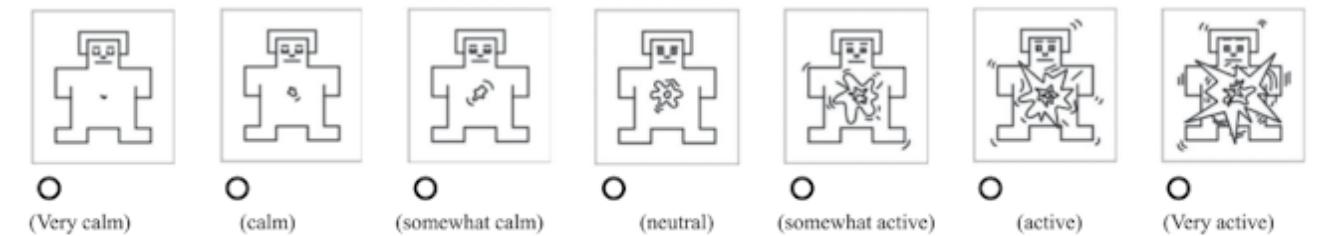
예: 슬픔 + 좌절.

Please rate the negative vs. positive aspect of the video
Click on the image that best fits the video.



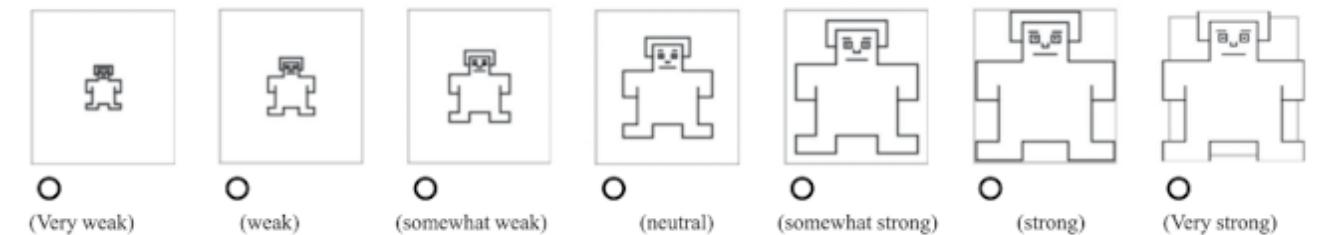
(a) Valence

Please rate the calm vs. excited aspect of the video
Click on the image that best fits the video.



(b) Arousal

Please rate the weak vs. strong aspect of the video
Click on the image that best fits the video.



(c) Dominance

Is any of these emotions the primary emotion in the audio? If not, select Other and specify the emotion.

Angry Sad Happy Surprise Fear Disgust Contempt Neutral Other

(d) Primary emotion

Please pick all the emotional classes that you perceived in the audio (Include the primary emotions selected in previous question)

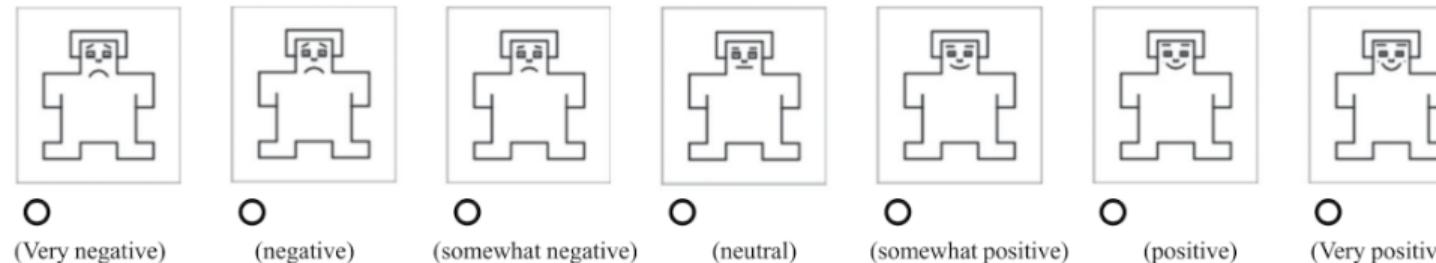
<input type="checkbox"/> Angry	<input type="checkbox"/> Sad	<input type="checkbox"/> Happy	<input type="checkbox"/> Amused	<input type="checkbox"/> Neutral
<input type="checkbox"/> Frustrated	<input type="checkbox"/> Depressed	<input type="checkbox"/> Surprise	<input type="checkbox"/> Concerned	
<input type="checkbox"/> Disgust	<input type="checkbox"/> Disappointed	<input type="checkbox"/> Excited	<input type="checkbox"/> Confused	
<input type="checkbox"/> Annoyed	<input type="checkbox"/> Fear	<input type="checkbox"/> Contempt	<input type="checkbox"/> Other	<input type="text"/>

(e) Secondary emotion

Perceptual Evaluations

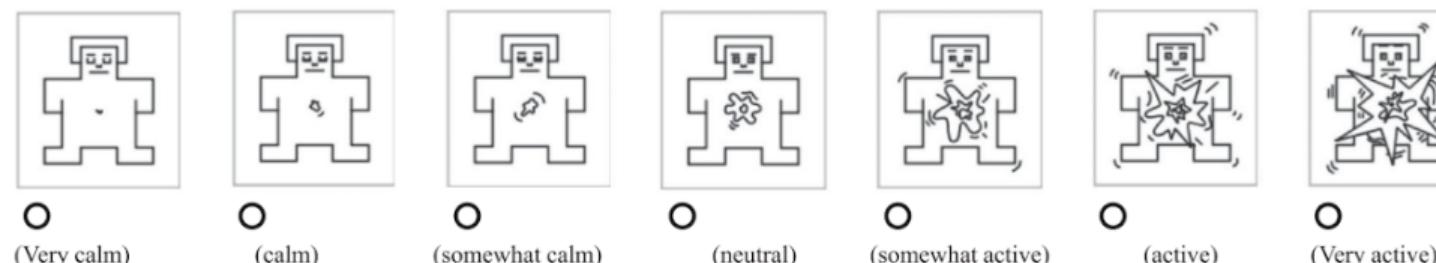
Using Crowdsourcing

Please rate the negative vs. positive aspect of the video
Click on the image that best fits the video.



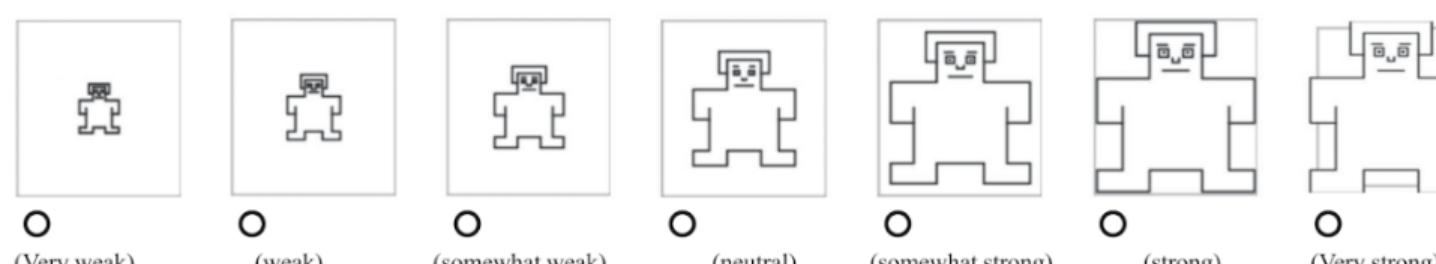
(a) Valence

Please rate the calm vs. excited aspect of the video
Click on the image that best fits the video.



(b) Arousal

Please rate the weak vs. strong aspect of the video
Click on the image that best fits the video.



(c) Dominance

Is any of these emotions the primary emotion in the audio? If not, select **Other** and specify the emotion.

- Angry Sad Happy Surprise Fear Disgust Contempt Neutral Other

(d) Primary emotion

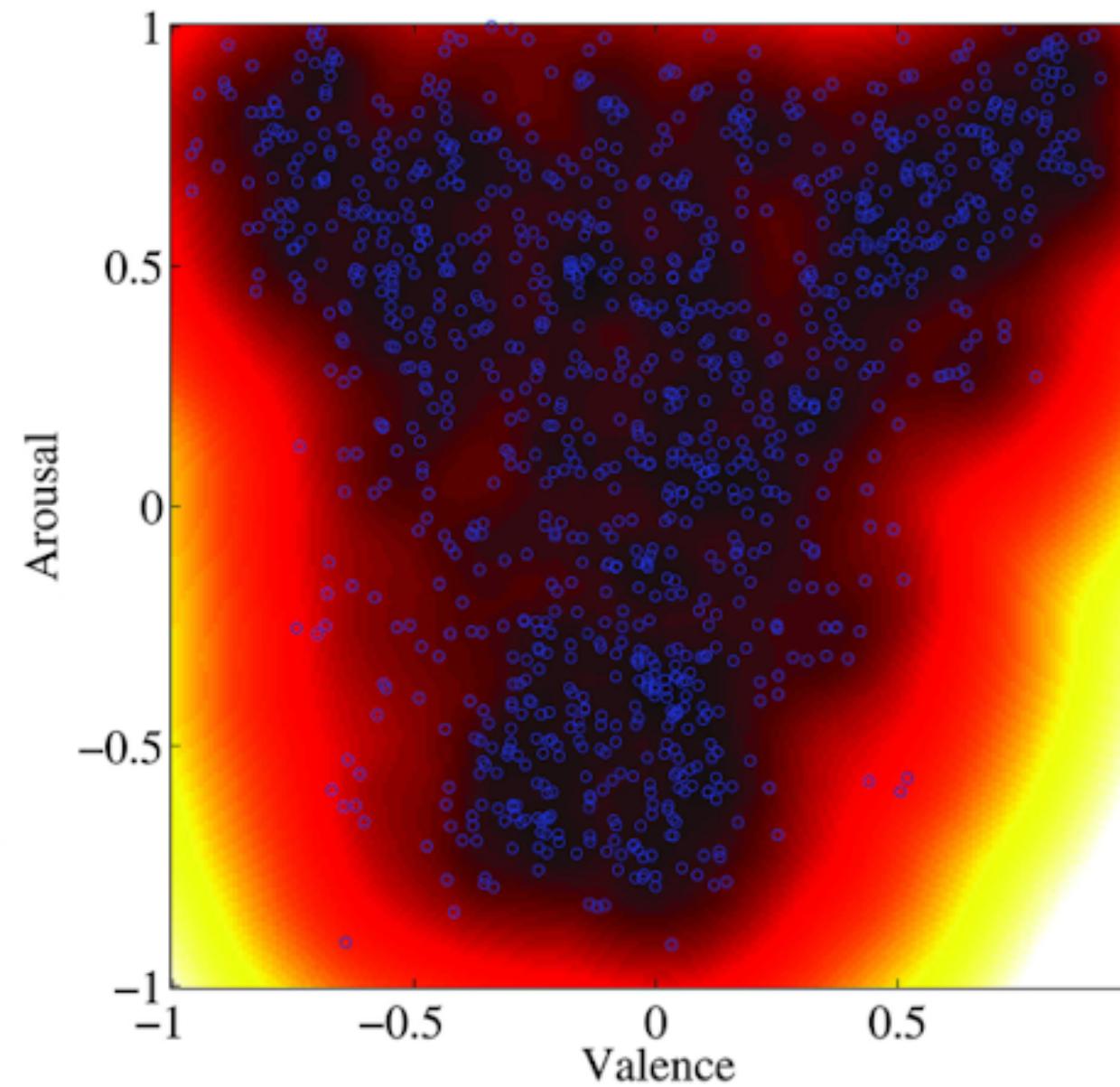
Please pick all the emotional classes that you perceived in the audio (Include the primary emotions selected in previous question)

- | | | | | |
|-------------------------------------|---------------------------------------|-----------------------------------|------------------------------------|----------------------------------|
| <input type="checkbox"/> Angry | <input type="checkbox"/> Sad | <input type="checkbox"/> Happy | <input type="checkbox"/> Amused | <input type="checkbox"/> Neutral |
| <input type="checkbox"/> Frustrated | <input type="checkbox"/> Depressed | <input type="checkbox"/> Surprise | <input type="checkbox"/> Concerned | |
| <input type="checkbox"/> Disgust | <input type="checkbox"/> Disappointed | <input type="checkbox"/> Excited | <input type="checkbox"/> Confused | |
| <input type="checkbox"/> Annoyed | <input type="checkbox"/> Fear | <input type="checkbox"/> Contempt | <input type="checkbox"/> Other | <input type="text"/> |

(e) Secondary emotion

Analysis of Emotional Content

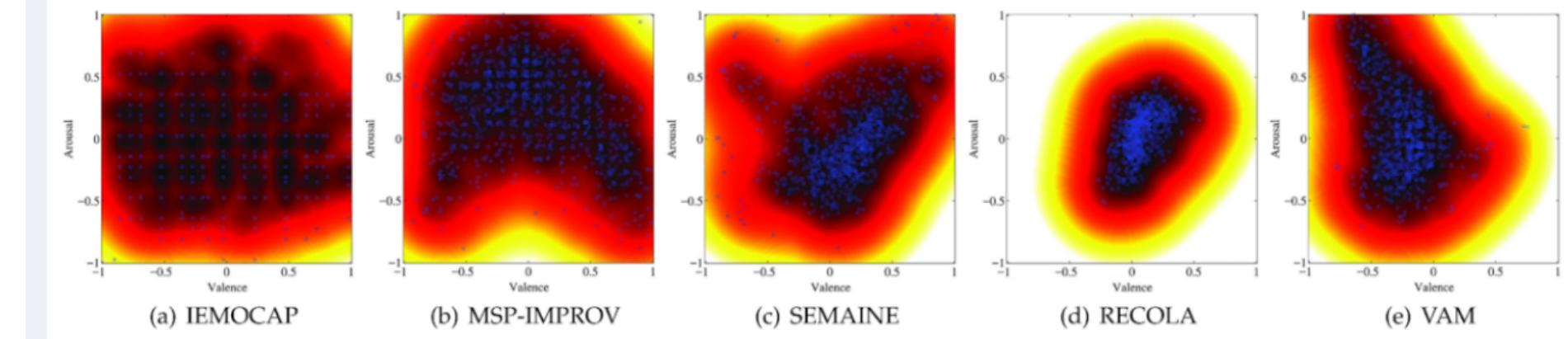
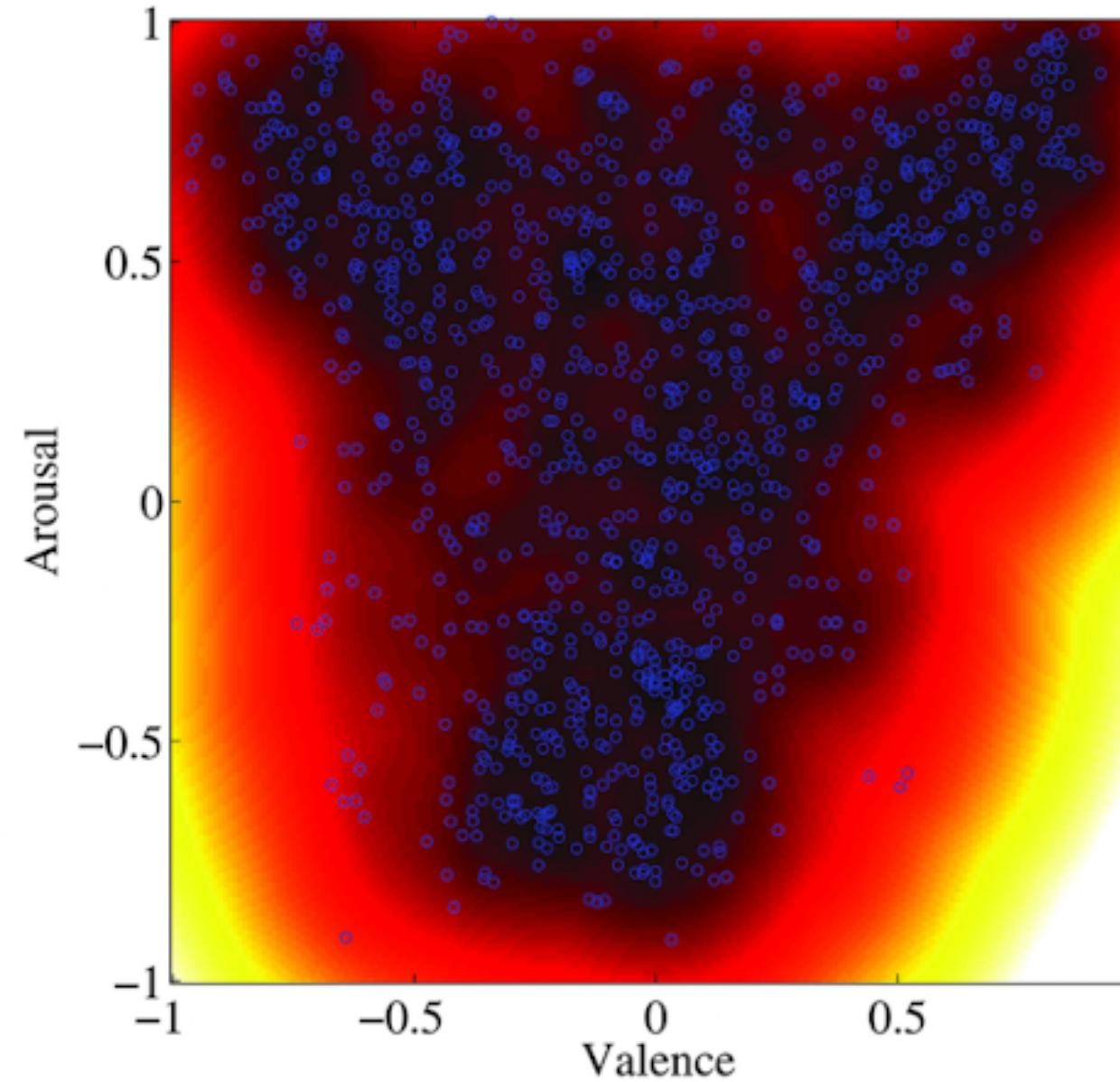
MSP-PODCAST



- GP-Rank가 추출한 감정 데이터가 얼마나 Arousal과 Valence의 스펙트럼을 균형 있게 포함하는지를 확인
- Arousal(흥분도)**
상위와 하위 수준 데이터를 효과적으로 분리해냈음을 보여줌.
즉, 흥분도가 높은 데이터와 낮은 데이터를 균형 있게 잘 나눔.
- Valence(긍정/부정)**
긍정적이거나 부정적인 데이터를 구분하는 데는 한계가 있음.
대부분의 데이터가 중립 영역(0 근처)에 몰려 있는 것을 알 수 있음.

Analysis of Emotional Content

MSP-PODCAST



● Arousal(흥분도)

상위와 하위 수준 데이터를 효과적으로 분리해냈음을 보여줌.
즉, 흥분도가 높은 데이터와 낮은 데이터를 균형 있게 잘 나눔.

● Valence(긍정/부정)

긍정적이거나 부정적인 데이터를 구분하는 데는 한계가 있음.
대부분의 데이터가 중립 영역(0 근처)에 몰려 있는 것을 알 수 있음.

Analysis of Primary Emotions

MSP-PODCAST

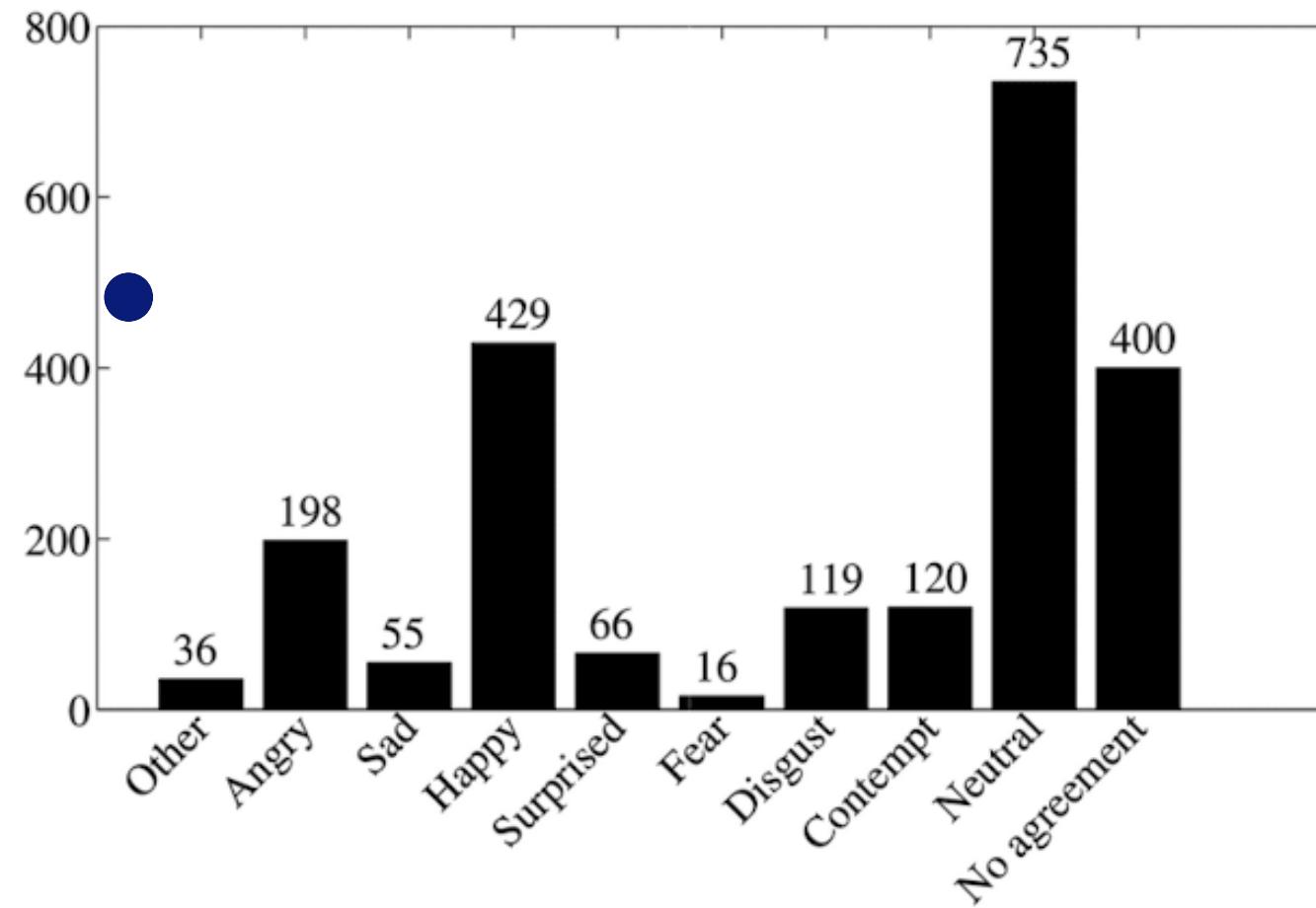


Fig. 10. Distribution of primary discrete emotions of retrieved samples by all three methods. Labels are assigned based on the majority vote consensus. Label *No Agreement* indicates that majority agreement does not exist for those samples.

- 주요 감정 분포

세 가지 방법으로 추출된 샘플에서 Primary Discrete Emotions의 분포. 샘플의 라벨은 다수결 합의(majority vote consensus)를 기반으로 할당

- Majority Vote Consensus

각 샘플은 여러 평가자(worker)가 감정 라벨을 할당한 후, 다수결로 결정된 라벨이 최종 라벨로 할당. 예를 들어, 5명의 평가자 중 3명이 "Anger"를 선택하면 해당 샘플의 라벨은 "Anger"로 결정

- No Agreement

평가자들 간의 의견이 일치하지 않아 다수결 합의를 도출할 수 없는 경우 해당 샘플은 "No Agreement" 라벨로 지정. 이는 특정 샘플의 감정이 명확히 구분되지 않음을 의미

Conclusion

Report Conclusion

자연스러운 감정 DB

팟캐스트와 같은 자연스러운 대화 데이터를 활용

ML과 Crowdsourcing을 결합하여 효율성과 정확성 확보

평가된 문장들은 Arousal(각성도)와 Valence(감정적 가치) 축을 대부분 커버, 기존 감정 데이터베이스보다 더 균형 잡힌 감정 분포를 보임