VOINOSIS
Voice Health Program Developer

2025.01.08 (WED)

MSP-PODCAST CORPUS

Speech Emotion Recognition in
Naturalistic Conditions Challenge

Q&A

김동환

목차

Table of Contents

PODCAST에
장르 라벨링이 있는가?

제거한 데이터의 기준이
어떻게 되는가?

ML로 어떻게 감정이
잘 드러나는 세그먼트를 검색했는가?

PODCAST에 장르 라벨링이 있는가?

Q&A Pages

아니요, 장르 라벨링은 따로 기재되어 있지 않습니다.

다만, 해당 논문에서는 다양성을 확보하기 위해
science, technology, politics, economics, business, arts,
culture, medicine, lifestyle and sport 주제를 수집했다고 밝혔습니다.

또한, 자연스러움을 위해 "non-acted" recodings을 수집하였고
이를 위하여 conversations, interviews, talk shows, news, discussion,
education, storytelling and debates 키워드를 통해 검색하였습니다.

Carlos Busso, Reza Lotfian, IEEE Transactions on Affective Computing

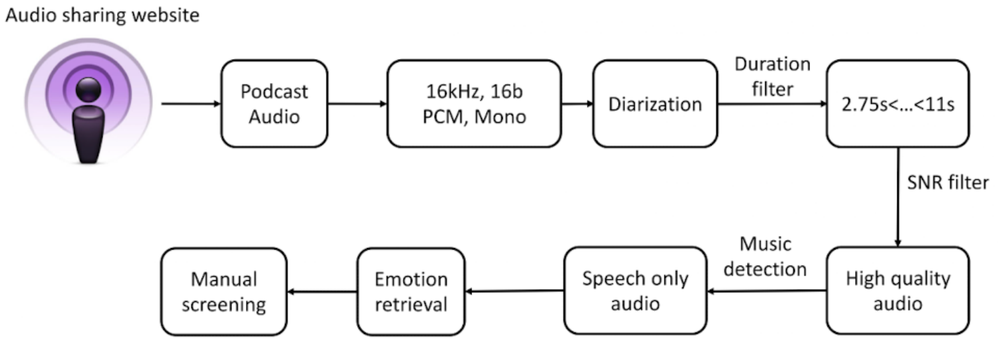
제거한 데이터의 기준이 어떻게 되는가?

Q&A Pages

아래와 같은 데이터를
1차적으로 기계적(수작업X)으로 제거하였습니다.

1. Diarization 후 2.75s~11s 이외의 길이 데이터
2. 침묵구간으로만 구성된 데이터
3. SNR (잡음비율) 20dB 미만인 데이터
4. 고주파 정보가 부족한 데이터 (감정분석에 주요한 특징은 고주파 대역에 포함)
5. 음악, 배경음악이 포함된 데이터
6. ML(classification, preference learning, regression)로
   "감정이 잘 드러나지 않는다고 판단"된 데이터 (뒤에 자세히)

Framework of Collecting MSP-PODCAST



제거한 데이터의 기준이 어떻게 되는가?

Q&A Pages

*이후 정제된 데이터를 **수작업**으로 검토하여 기준에 부합하지 않는 데이터들을 제거하였습니다.*

기준 : 단일 화자, 음성만(배경 음악 X) 존재, 2.75s~11s 길이

수작업으로 검토한 데이터의 결과는 오른쪽 표와 같습니다.

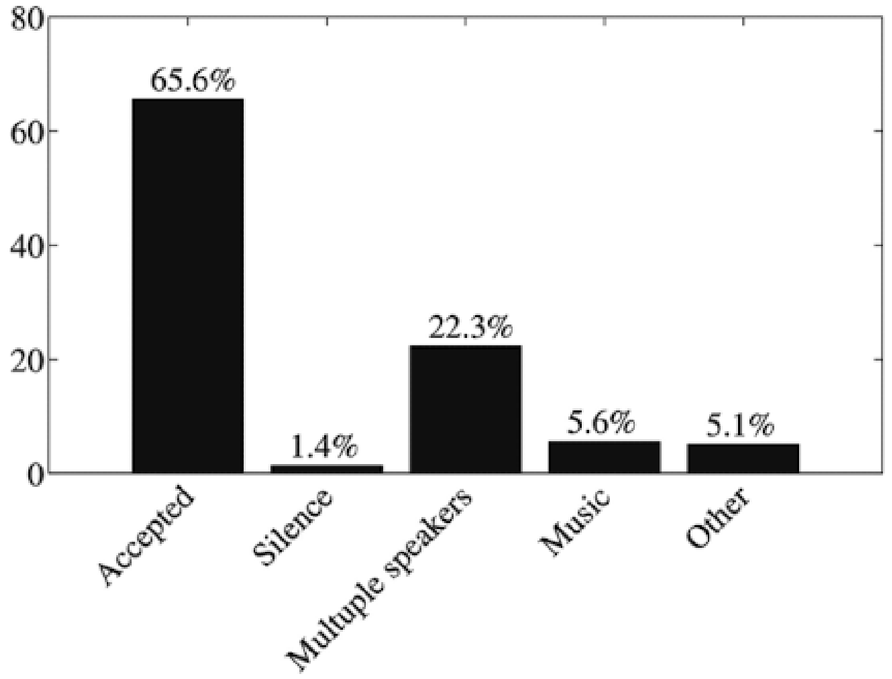


Fig. 3. Histogram of speech turns evaluated manually.

ML로 어떻게 감정이 잘 드러나는 세그먼트를 검색했는가

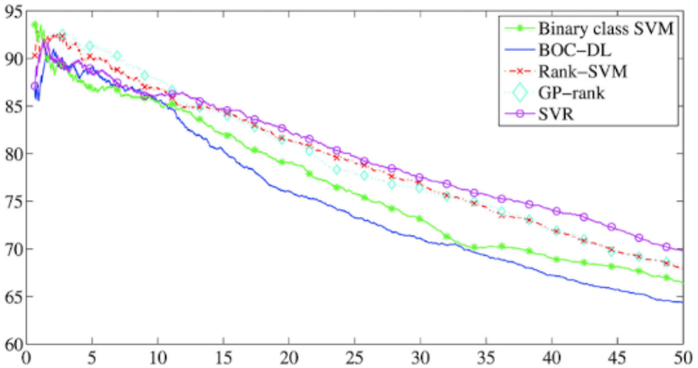
ML(classification, preference learning, regression)로 "감정이 잘 드러난다고 판단"된 데이터를 선별

MSP-PODCAST는 attribute-based balancing이 목표이므로
다른 데이터셋(IEMOCAP,MSP-IMPROV) 으로 훈련시킨 ML로
high and low arousal, high and low valence인 데이터를 선별하였습니다.

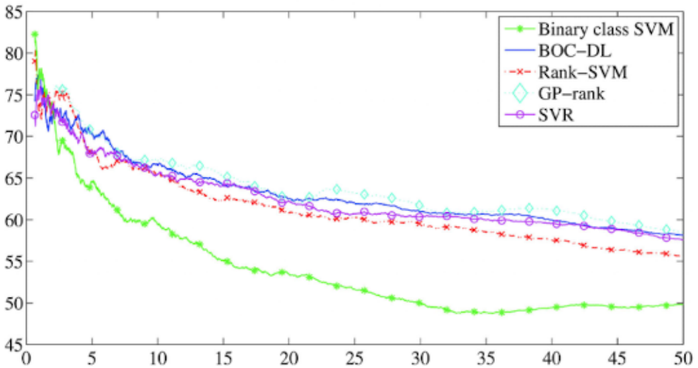
PAD emotion state는 Arousal, Valence, Dominance로 감정을 분류하지만,
dominance는 arousal과 높은 상관성을 보여 dominance는 무시하였습니다.

ML로 어떻게 감정이 잘 드러나는 세그먼트를 검색했는가

ML의 training, evaluation



(a) Arousal



(b) Valence

총 5가지 ML 방법을 사용하여
기존 데이터셋(IEMOCAP + MSP-IMPROV)에서 학습 후 평가를 진행하였습니다.
참고) 해당 데이터셋들은 모두 labeling이 수작업으로 이루어진 데이터셋입니다.

평가 지표로는 P@K metric를 사용했으며, 상위 k%의 정확도를 의미합니다.
예를 들어 P@10 == 87%는
상위 10% +하위 10% arousal 로 분류된 데이터중 87%가 올바르게 분류됨을 뜻합니다.

저자들은 최종적으로 좋은 성능을 내는 BOC-DL, GP-rank, SVR을 선택하여 세그먼트를 검색하였습니다. (high/row - arousal/valence)

ML로 어떻게 감정이 잘 드러나는 세그먼트를 검색했는가

ML의 training, evaluation

총 5가지 ML 방법을 사용하여
기존 데이터셋(IEMOCAP + MSP-IMPROV)에서 학습 후 평가를 진행하였습니다.

평가 지표로는 P@K metric를 사용했으며, 상위 k%의 정확도를 의미합니다.
예를 들어 P@10 == 87%는
상위 10% +하위 10% arousal 로 분류된 데이터중 87%가 올바르게 분류됨을 뜻합니다.

TABLE 2
Number of Sentences Retrieved Under Different Settings

Set	# Turns	Retrieval approach
High Arousal	200	GP-rank
Low Arousal	200	GP-rank
High Valence	200	GP-rank
Low Valence	200	GP-rank
High Arousal	200	SVM regression
Low Arousal	200	SVM regression
High Valence	200	SVM regression
Low Valence	200	SVM regression
High Arousal	200	BOC-DL
Low Arousal	200	BOC-DL
High Valence	200	BOC-DL
Low Valence	200	BOC-DL
Random	100	Random
Total	2,317	

Due to overlapped sets, the total number of distinct sentences is 2,317.

저자들은 최종적으로 좋은 성능을 내는 BOC-DL, GP-rank, SVR을 선택하여 세그먼트를 검색하였습니다. (high/row - arousal/valence)