

Nonverbal Vocalization Dataset

Why do we need the nonverbal vocalization dataset?

메라비언의 의사소통 이론에 따르면, 개인적인 의사소통은 7%의 말의 내용, 38%의 청각적 요소, 그리고 55%의 시각적 요소로 구성된다고 합니다. 그러나, 기존의 인간 음성 데이터 세트의 대부분은 인간의 언어에 집중되어 있고 수 년 동안 많은 연구가 수행되어 왔고, 비언어적 발성이 의사소통에 중요한 역할을 한다는 사실에도 불구하고, 매우 제한된 양의 인간의 비언어적 음성 데이터 세트만 이용할 수 있습니다. Nonverbal Vocalization Dataset은 인간의 비언어적 소리를 이해하려는 시도를 지원하기 위해 인간이 만든 최초의 비언어적 소리 데이터 세트입니다.

What's inside the nonverbal vocalization dataset?

Nonverbal Vocalization Dataset은 한국의 일반 대중이 제공한 1419명의 발화자가 녹음한 56.7시간의 짧은 클립으로 구성된 인간의 비언어적 발성 데이터 세트입니다. 데이터 세트에는 연령, 성별, 소음 수준 및 발언 품질과 같은 메타데이터가 포함됩니다. 포함된 인간의 비언어적 소리에는 '치아 부딪히는 소리', '이 가는 소리', '혀 차는 소리', '코 푸는 소리', '기침하는 소리', '하품하는 소리', '목청 가다듬는 소리', '한숨 소리', '입술을 오므렸다 터뜨리는 소리', '쩝쩝대는 소리', '훨떡이는 소리', '우는 소리', '웃는 소리', '재채기 소리', '끙끙 앓는 소리', '비명 소리'로, 총 16가지 소리가 포함되어 있습니다.

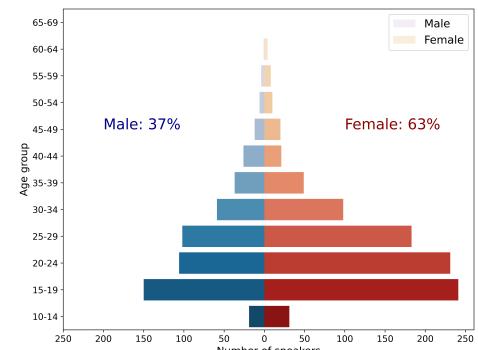
데이터셋의 모든 음성에는 연구 목적에 따라 데이터셋을 필터링할 수 있도록 노이즈 수준 및 품질 수준이 나누어져 있습니다. 소음 수준은 '소음 없음'과 '소음 있음'의 2단계로 정의되고, 품질 수준은 '높음'과 '낮음'의 2단계로 정의되며, 음성의 진정성 수준을 나타냅니다. 모든 메타데이터는 발화자가 입력한 뒤, 검수자의 검수를 거쳤습니다.

데이터셋에 기록된 모든 원본 음성은 44,100Hz, 16-bit PCM, 1-channel wav 파일이며 Android 운영 체제를 이용하는 스마트폰으로 녹음했습니다. 각 레코딩은 16,000Hz로 다시 샘플링되어 메타데이터와 함께 단일 H5 파일로 저장됐습니다. h5 파일을 다루는 것이 익숙하지 않다면 아래 참조해주세요.

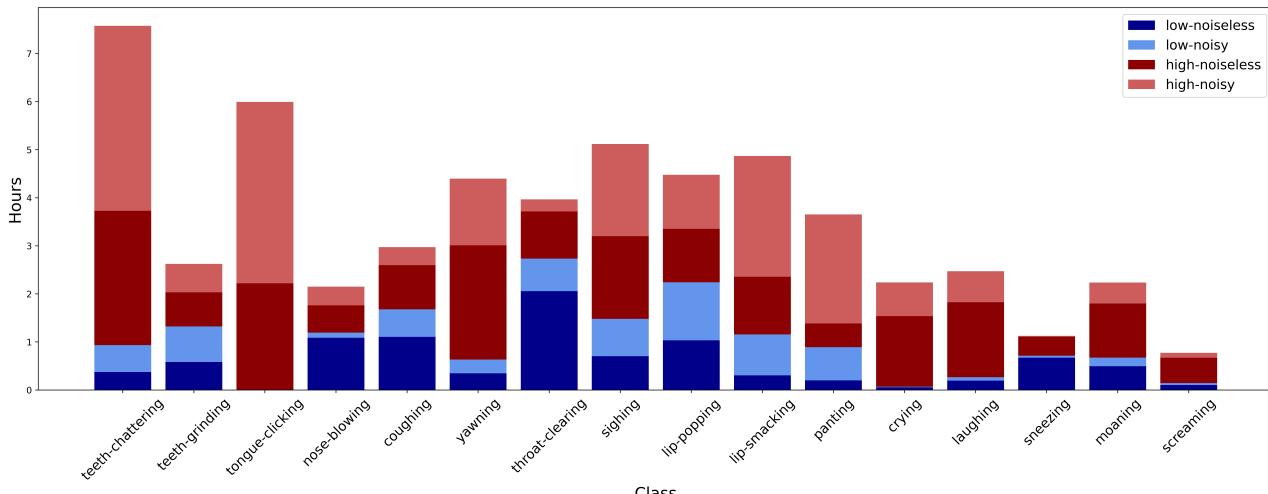
Information & Statistics

그림 1은 데이터 세트의 화자에 대한 인구통계 정보를 보여줍니다. 녹음에 참가한 화자들은 남성 37%, 여성 63%로 구성되어 있습니다. 그리고 크라우드소싱 방식 데이터 수집의 특성 때문에, 화자의 거의 70%가 15-24 세의 범위 내에 있습니다.

그림 2는 각 클래스의 총 길이(시간)를 보여줍니다. 각 막대 그래프의 다른 색상은 노이즈 및 품질 수준의 조합을 나타냅니다(이 문서의 뒷부분에서는 이 조합을 '필터'라고 하겠습니다). 예를 들어, 클래스 '치아 부딪히는 소리'의 전체 기록은 '낮음-소음없음' 부분 0.37시간, '낮음-소음있음' 부분 0.56시간, '높음-소음없음' 부분 2.79시간, '높음-소음있음' 부분 3.85시간으로 구성됩니다.



<그림 1. Age distribution by sex>



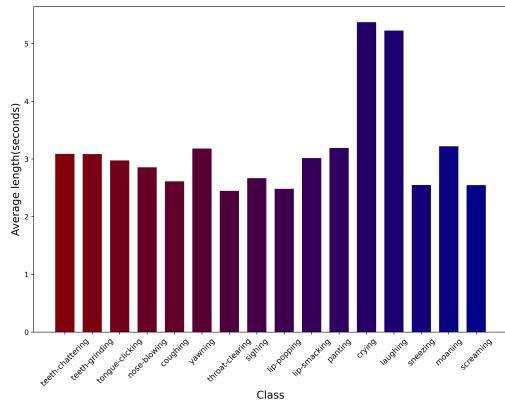
<그림 2. Total length(hours) by class and filter>

그림 3에서 볼 수 있듯, 14개의 클래스(‘울음 소리’와 ‘웃음 소리’ 제외)의 음성의 평균 길이(초)는 3초 정도, 나머지는 5초 정도 높여 있다. 그림 4에서는 클래스별 길이(초) 분포를 그림으로 표시했습니다. 이 그래프를 보면 위에서 언급한 14개 클래스의 길이 분포가 서로 유사하며, 나머지 클래스의 길이 분포도 유사하다는 것을 알 수 있습니다.

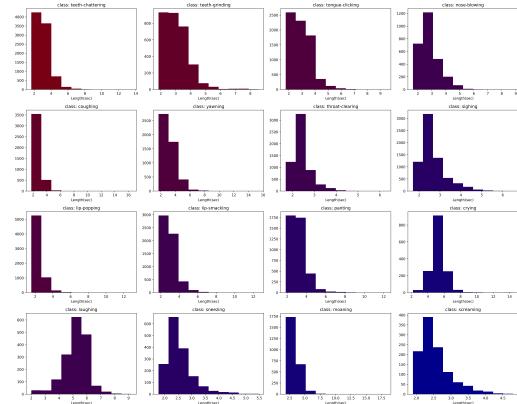
아래 그림 5는 각 클래스에 몇 명의 화자가 포함되어 있는지 보여줍니다. 앞서 언급했던 것처럼, 녹음의 평균 길이와 한 명의 화자가 제출한 녹음의 수는 클래스마다 크게 다르지 않기 때문에 그림 5의 분포는 그림 2의 분포와 매우 유사한 것을 볼 수 있습니다.

또한, 그림 6에는 단일 화자가 녹음한 클래스 수를 나타내는 히스토그램이 나와 있습니다. 예를 들어, 432명의 화자가 하나의 클래스만 녹음했고(그림 6의 맨 왼쪽 막대 그래프), 16명의 화자는 모든 클래스(16 클래스)를 녹음했습니다.(맨 오른쪽 막대)

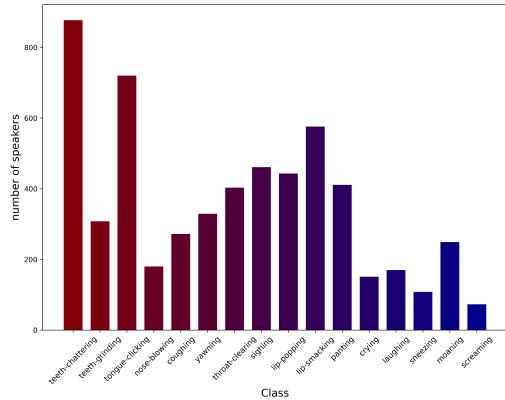
마지막으로, 음성 파일의 98.76%는 실내에서 기록되었고 1.24%는 실외에서 기록되었습니다.



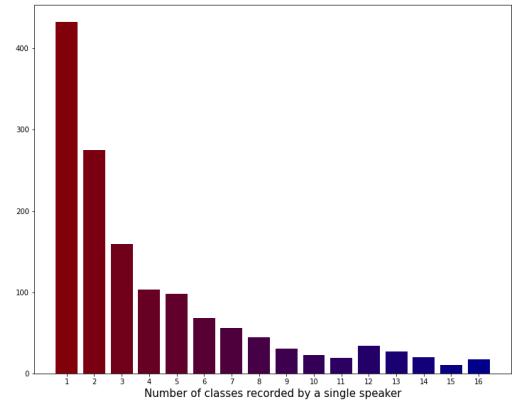
<그림 3. Average length of recordings by class>



<그림 4. Length of recordings distribution by class>



<그림 5. Average number of speakers by class>



<그림 6. Number of classes recorded by speaker>

Filename convention

원본 녹음 파일의 파일명은 다음과 같은 규칙을 따릅니다.

{speaker ID}_{class}_{index}_{sex}_{age}_{location}_{quality}_{noise}.wav
Example: 3CHZ_9_8_0_29_0_0_1.wav

Speaker ID 각 화자의 고유 4자리 영숫자 코드이며, **Class**는 비언어적 사운드의 유형을 나타내는 디지털 코드이며, **Index**는 특정 클래스에서 동일한 스피커의 다른 녹음 파일을 식별하는 데 사용되는 숫자이며, **Location**은 녹화가 발생한 위치를 나타내는 디지털 코드이며, **Quality**은 녹음 파일의 진정성 수준을 나타냅니다. **Noise**은 녹음 시의 소음 수준을 나타냅니다.

How to decode?

Class

- 0: 'teeth-chattering'('치아 부딛히는 소리'),
- 1: 'teeth-grinding'('이 가는 소리'),
- 2: 'tongue-clicking'('혀 차는 소리'),
- 3: 'nose-blowing'('코 푸는 소리'),
- 4: 'coughing'('기침하는 소리'),
- 5: 'yawning'('하품하는 소리'),
- 6: 'throat-clearing'('목청 가다듬는 소리'),
- 7: 'sighing'('한숨 소리'),
- 8: 'lip-popping'('입술을 오므렸다 터뜨리는 소리'),
- 9: 'lip-smacking'('쩝쩝대는 소리'),
- 10: 'panting'('헐떡이는 소리'),
- 11: 'crying'('우는 소리'),
- 12: 'laughing'('웃는 소리'),
- 13: 'sneezing'('재채기 소리'),
- 14: 'moaning'('끙끙 앓는 소리'),
- 15: screaming('비명 소리')

Sex

- {0: '여성', 1: '남성'}

Location

- {0: '실내', 1: '실외'}

Quality

- {0: '높음', 1: '낮음'}

Noise

- {0: '소음없음', 1: '소음있음'}

Metadata

metadata.json

License

Contact & Purchase

