A Dataset for Speech Emotion Recognition in Greek Theatrical Plays

Maria Moutti, Sofia Eleftheriou, Panagiotis Koromilas, Theodoros Giannakopoulos

Multimedia Analysis Group - Computational Intelligence Lab, magcil.github.io

1. Abstract

Machine learning methodologies can be adopted in cultural applications and propose new ways to distribute or even present the cultural content to the public. For instance, speech analytics can be adopted to automatically generate subtitles in theatrical plays, in order to (among other purposes) help people with hearing loss. Apart from a typical speech-to-text transcription with Automatic Speech Recognition (ASR), Speech Emotion Recognition (SER) can be used to automatically predict the underlying emotional content of speech dialogues in theatrical plays, and thus to provide a deeper understanding of how the actors utter their lines. However, real-world datasets from theatrical plays are not available in the literature. In this work we present GreThE, the Greek Theatrical Emotion Dataset, a newly publicly available data collection for speech emotion recognition in Greek theatrical plays. The dataset contains utterances from various actors and plays, along with respective valence and arousal annotations. Towards this end, multiple annotators have been asked to provide their input for each speech recording and inter-annotator agreement is taken into account in the final ground truth generation. In addition, we discuss the results of some indicative experiments that have been conducted with machine and deep learning frameworks, using the dataset, along with some widely used databases in the field of speech emotion recognition. (https://github.com/magcil/GreThE)

2. Related Work

- SER datasets: can be classified into 4 categories according to the recording procedure [12]: spontaneous, acted, elicited and annotated public speech. Commonly used datasets: IEMOCAP [2], Emo-DB [1], MSP-podcast [15], EMOVO [5], SAVEE [10] and RAIDNESS [14]
- Greek SER: Greek-based SER datasets are limited: AESDD [16], 5 actors and annotated with 5 emotional states (no neutral state). SEWA [13] is multi-lingual, 2000 minutes of data of 398 people coming from 6 cultures (including Greek), annotated among others in terms of continuously valued valence and arousal

3. The Dataset

- Data collection: at least 20 single speaker utterances x 23 Greek discrete theatrical plays = 500 recordings. Total duration 46 min, average duration 5.5 sec.
- Annotation process: four individuals annotated via the Label Studio tool. Arousal labels: (1)very weak (2)weak (3)neutral (4)strong (5)very strong. Valence labels: (1)very negative (2)negative (3)neutral (4)positive (5)very positive.
- Annotations aggregation:
  - Mean Thresholding (average annotation rating): Arousal: strong = [3.66, 5], neutral = (2.66, 3.66), weak = [0, 2.66], Valence: positive = [3.33, 5], neutral = (2.33, 3.33), negative = [0, 2.33]
  - Deviation Thresholding (mean absolute deviation (MAD)): \( \sigma < 1.3 \)
  - Inter-annotator (dis)agreement (mean of MAD): Arousal: 0.48, Valence: 0.49
  - Average disagreement for each annotator.

4. Baseline Classification Methods

- ML + hand-crafted audio features [7]:
  - utterances split into 50 msec frames and extract 34 spectral, time and cepstral features. Compute deltas. Get \( \mu \) and \( \sigma^2 \) of features (per 1 sec, then long-term-average). 168-D representation for the whole utterance.
  - SVM classifier with an RBF kernel
- DL based approach: mel-spectrograms and CNNs (deep_audio_features)

5. Experimental Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Arousal F1</th>
<th>Valence F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>27%</td>
<td>20%</td>
</tr>
<tr>
<td>Prior-aware Baseline</td>
<td>31%</td>
<td>30%</td>
</tr>
<tr>
<td>SVM - Oversampling</td>
<td>53%</td>
<td>30%</td>
</tr>
<tr>
<td>SVM - Undersampling</td>
<td>55%</td>
<td>40%</td>
</tr>
<tr>
<td>CNN_iemocap</td>
<td>40%</td>
<td>37%</td>
</tr>
<tr>
<td>CNN_msp</td>
<td>36%</td>
<td>34%</td>
</tr>
<tr>
<td>CNN_merged</td>
<td>41%</td>
<td>34%</td>
</tr>
</tbody>
</table>

Table 1: GreThE evaluation results

6. Conclusions & Future Work

- Conclusions:
  - (a) recognising emotion in theatrical data is challenging when training from scratch
  - (b) using state-of-the-art datasets from generic SER on cross-language theatrical data is not effective
- Future work: robust domain adaptation techniques using few-shot learning strategies.

7. References