A Dataset for Speech Emotion Recognition in Greek Theatrical Plays

Maria Moutti, Sofia Eleftheriou, Panagiotis Koromilas, Theodoros Giannakopoulos
National Center for Scientific Research "DEMOKRITOS", Institute of Informatics and Telecommunications
Multimedia Analysis Group - Computational Intelligence Lab, magcil.github.io

MagCIL
Multimodal Machine Learning

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 new 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 *RAVDESS* [14]
- Greek SER: Greek-based SER databases 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
- SER datasets for cultural content: databases of acted speech: CREMA-D [4], CaFE [9], IEMOCAP [2], EMOVO [5] and RAVDESS [14], database of elicited speech: MSP-IMPROV [3]. A study [8] examines *predicted* emotions of both the audience and the actors during a public performance. Cinematic films databases: EMOVIE [6], AVE [11].

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 μ and σ^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

Experiment	Arousal F1	Valence F1
Baseline	27%	26%
Prior-aware Baseline	31%	30%
SVM	53%	38%
SVM - Oversampling	55%	40%
SVM - Undersampling	54%	39%
CNN_iemocap	40%	37%
CNN_msp	36%	34%
CNN_merged	41%	34%

Table 1: GreThE evaluation results

- Session-independent validation: first baseline method, GreThE ID-based train/validation split. Improvement of 27.9% for arousal and 21.2% for valence compared to the prior-aware baseline
- Cross-domain validation: second baseline method, training on MSP-podcast [15] and IEMOCAP [2], tested on GreThE. Improvement of 20.6% for arousal and for 17.5% valence compared to the baseline.

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

[1] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008.

[3] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2017.

[4] Houwei Cao, David Cooper, Michael Keutmann, Ruben Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5:377–390, 10 2014.

[5] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. Emovo corpus: an italian emotional speech database. In *LREC*, 2014.

[6] Chenye Cui, Yi Ren, Jinglin Liu, Feiyang Chen, Rongjie Huang, Ming Lei, and Zhou Zhao. Emovie: A mandarin emotion speech dataset with a simple emotional text-to-speech model, 2021.

[7] T. Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12), 2015.

[8] Peter A Gloor, Keith April Araño, and Emanuele Guerrazzi. Measuring audience and actor emotions at a theater play through automatic emotion recognition from face, speech, and body sensors. In *Collaborative innovation networks conference of Digital Transformation of Collaboration*, pages 33–50. Springer, 2019.

[9] Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. A canadian french emotional speech dataset. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, page 399–402, New York, NY, USA, 2018. Association for Computing Machinery.

[10] Philip Jackson and SJUoSG Haq. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014.

[11] Sudarsana Reddy Kadiri, P. Gangamohan, Vinay Kumar Mittal, and Bayya Yegnanarayana. Naturalistic audio-visual emotion database. In *ICON*, 2014.

[12] Panagiotis Koromilas and Theodoros Giannakopoulos. Deep multimodal emotion recognition on human speech: A review. *Applied Sciences*, 11(17):7962, 2021.

[13] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjorn Schuller, and et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040, Mar 2021.

[14] Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35, 05 2018.

[15] Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. The msp-conversation corpus. *Interspeech 2020*, 2020.

[16] Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, Charalampos Dimoulas, and George Kalliris. Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society. Audio Engineering Society*, 66:457–467, 06 2018.