Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1040–1046
Marseille, 20-25 June 2022

A Dataset for Speech Emotion Recognition in Greek Theatrical Plays

Maria Moutti[1], **Sofia Eleftheriou**[2], **Panagiotis Koromilas**[2], **Theodoros Giannakopoulos**[2 3]
[1]University of the Peloponnese, [2]National Center for Scientific Research - Demokritos,
[3]Behavioral Signal Technologies Inc.
{mar.moutti}@gmail.com, {seleftheriou, pakoromilas, tyianak}@iit.demokritos.gr

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.

Keywords: GreThE dataset, speech emotion recognition, Greek theatrical plays, valence, arousal

1. Introduction

The task of recognizing the underlying emotion from speech, irrespective of its semantic content, is rather important in various applications. However, it is hard to notate even by human beings, mostly due to the subjectiveness of the emotional content. The ability to automatically conduct it, is a demanding task and still an ongoing subject of research. Several open-source databases exist in the field of Speech Emotion Recognition (SER) which may contain audio-only or multimodal information and are usually annotated on two categories: categorical attributes (distinct classes of emotions) (Ortony and Turner, 1990; Ekman, 1992) and dimensional attributes (continuous values of valence, arousal and intensity) (Wundt and Judd, 2011; Russell, 1980). The emerge of the fields of Machine and Deep Learning in the past decade, gave the chance to researchers to apply research outputs on real-world problems. However, despite the fact that SER is a task that has gained great attention in the literature (Koromilas and Giannakopoulos, 2021), the industrial applications of the proposed works are either centered around web content (online video or podcast analysis) or have been applied on actual conversations to enrich understanding (eg. empathetic dialogue (Ma et al., 2020)).

At the same time, cultural events are a base factor for every human civilization and, as such, they can also benefit from modern Machine Learning (ML) applications. Automatic review mining, automatic summary generation of movies, content-based movie recommendation systems, music and movie retrieval, production of subtitles/transcriptions in guided tours, movies or theatrical performances are just some examples. These applications of ML on cultural content change the way the content is generated and distributed. Moreover, it provides solutions to increase *inclusion* of particular groups of the population: e.g. automatic generation of subtitles, enriched with paralinguistic attributes such as emotional arousal could help people with hearing loss to actually understand and ”feel” a theatrical play.

The widely used SER approaches have not properly been used to address challenges that are set from cultural content data, with the best example being the theatrical plays.

This is mostly due to the fact that methods that are trained on SER datasets cannot be properly applied on theatrical content. This claim is based on the following facts:

1. actors that perform in theatrical plays are more expressive and thus the emotional levels are aroused. That is, the arousal classes are shifted towards more energetic emotions (e.g. the weak class is expressed in a more ”aroused” - energetic way). As far as the valence classes are concerned, they also differ from the respective valence classes in other SER datasets. For example, depending on play type (e.g. dramas) the neutral class itself can also include negative or positive emotions, compared to SER datasets that try to capture a real-world (non-theatrical) context.

2. interaction with the audience make the actors express their actions in different ways so as to be better perceived by the attendants. That is, actors

use emotional states that are not common in real-life conversations and thus are not included in the general SER datasets

3. the recording setups and conditions of theatrical plays differ from that of the datasets found in the literature, as the former may include complex microphone systems and fine post-editing procedures.

In this work, we propose the Greek Theatrical Emotion (GreThE) dataset with the aim of filling the existing gap in the literature. GreThE is a collection of speech utterances from 23 Greek theatrical plays annotated with regards to the respective levels of emotional valence and arousal. We also provide a baseline evaluation for the presented dataset and we examine whether the domain knowledge of general SER can be used to achieve robust performance on GreThE.

The paper is organized as follows: in section 2 we report the related works on SER datasets, section 3 describes the GreThE dataset, section 4 contains the used classification methods, section 5 reports the experimental results, section 6 comment on the availability of the dataset and section 7 concludes the paper.

2. Related Work

2.1. Speech emotion recognition datasets

Remarkable effort has been given in the literature, to create emotion-based datasets that accurately represent the basic human emotions and reactions in speech signals. The existing datasets can be classified into four categories according to the recording procedure that is followed through the data collection process (Koromilas and Giannakopoulos, 2021). Specifically, these methods include one of the following: *(i) spontaneous* speech: the participants are unaware of the recording while their speech and reactions are recorded with hidden mechanisms in a real environment (Cao et al., 2015); *(ii) acted* speech: the emotional condition of the speakers is acted; *(iii) elicited* speech: where the speaker is placed in a situation which evokes a specific emotional state (Basu et al., 2017); and *(iv) annotated public* speech: data from public sources, such as YouTube, are annotated to associate them with a range of emotional states.

Some of the most commonly used datasets in that field are: *IEMOCAP* (Busso et al., 2008), a multimodal database which includes recordings from 10 actors annotated in categorical and dimensional attributes, *Emo-DB* (Burkhardt et al., 2005), an emotional speech database containing recordings of 10 speakers that simulates 7 emotional states, *MSP-podcast* (Martinez-Lucas et al., 2020) that contains speech segments from podcast recordings which are annotated with emotional labels using attribute-based descriptors and categorical labels, *EMOVO* (Costantini et al., 2014) an Italian emotional speech database created by 6 actors that simulates 7 emotional state, *SAVEE* (Jackson and ul haq, 2011) database which describes the emotion in 6 distinct categories and *RAVDESS* (Livingstone and Russo, 2018) that provides speeches of 24 actors and songs in audio and video format and includes 7 emotional expressions among with two levels of emotional intensity.

2.2. Greek emotion recognition

The respective work on Greek-based speech emotion recognition databases is limited. In particular, one of the first approaches has been introduced in the *AESDD* dataset (Vryzas et al., 2018), which is a publicly available SER database that contains utterances of acted emotional speech in the Greek language created by 5 actors and annotated with five emotional states (without containing the neutral state). Furthermore, *SEWA* (Kossaifi et al., 2021) is a multi-lingual database for audio-visual emotion and sentiment research in the wild containing more than 2000 minutes of data of 398 people coming from 6 cultures (including Greek), annotated among others in terms of continuously valued valence and arousal.

2.3. Emotion datasets for cultural content

As discussed in 2.1, the participation of actors in the recording of emotional databases in order to perform acted emotional speech has been a popular approach in the study of emotions. There is a range of databases that are based on actors, such as CREMA-D (Cao et al., 2014), CaFE (Gournay et al., 2018), IEMOCAP (Busso et al., 2008), EMOVO (Costantini et al., 2014) and RAVDESS (Livingstone and Russo, 2018).

The MSP-IMPROV corpus (Busso et al., 2017) is an example of the elicited speech (category *(iii)*) datasets. It proposes an alternative, approach according to which the authors define hypothetical scenarios for each sentence that are carefully designed to elicit a particular emotion. Two actors improvise these emotion-specific situations, leading them to utter contextualized, non-read renditions of sentences that have fixed lexical content and convey different emotions. In this way, they manage to produce more natural behaviors. However, neither the recording conditions or the emotional reactions can be considered to be close to these of an actual theatrical play which is more expressive and differs from real-life reactions.

Regarding the task of recognizing emotions in actual theatrical plays, to our knowledge, the only study in the literature is presented in (Gloor et al., 2019), where the authors developed a system to measure both audience and actor satisfaction during a public performance. They used smartwatches to gather physiological signals from the actors, as well as video cameras to capture facial expressions from the audience and finally speech signals from the actors to be used in SER. Then, predictions of emotions were extracted on the three channels of information using pretrained models from existing external datasets and they presented results of correlation metrics between the emotions predicted from the individual channels. Therefore, the particular work does not present a new speech emotion recognition dataset rather than it examines relationships between *predicted* emotions of both the audience and the actors from different channels of information.

Cinematic films is a type of content with limited representation in the literature of emotion recognition. Specifically, the EMOVIE (Cui et al., 2021) dataset that includes 9,724 samples from seven movies with audio files annotated in the emotion polarity and the AVE (Kadiri et al., 2014) dataset which is based on an Indonesian Movie study (Muljono et al., 2019) are two of the few examples.

Our proposed dataset, GreThE, aims to fulfill the gap in the literature of language resources, by proposing a publicly available non-english speech emotion recognition dataset for real-world theatrical recording conditions.

3. Dataset

3.1. Audio data collection

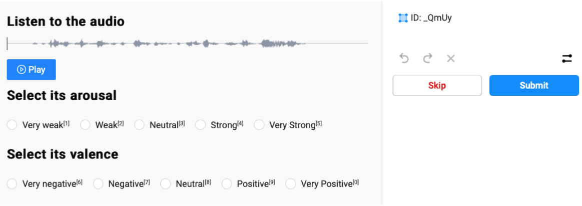
We have selected to adopt the `Audacity` open source audio editing tool to manually segment at least 20 single speaker utterances, randomly selected from each theatrical play. This process led to 95 single-speaker utterances (90 unique speakers) from 23 Greek discrete theatrical plays, resulted to a total of 500 recordings/speeches. The initial recordings have been gathered from various online resources, covering a wide range of recording conditions and respective equipment used. The total duration of speech is 46 minutes and their average duration is 5.5 seconds (the shortest utterance is 2.1 seconds and the longest utterance is 10.9 seconds).

3.2. Utterance annotation process

Each speech utterance that has been collected from the various theatrical plays, as described in Section 3.1, has been annotated with respect to its emotional content. Towards this end, we have selected to use the dimensional emotional representation of Valence and Arousal. Our goal was to adopt the standard in SER 3-class approach (Metallinou et al., 2012), according to which the classes for Valence are: negative, neutral and positive and the classes for Arousal are weak, neutral and strong.

However, in the initial annotation process we asked the individual annotators to provide their feedback in a 5-valued scale for both tasks. Then we used aggregates on these 5-scale estimates to map them to the three distinct final ground truth classes as described in Section 3.3. So in the individual annotation procedure, the following 5 labels were used: (1) very weak (2) weak (3) neutral (4) strong (5) very strong, and (1) very negative (2) negative (3) neutral (4) positive (5) very positive, for the arousal and valence task respectively.

The annotation process has been carried out by four individuals. We have selected to adopt the `Label Studio` open source data labeling tool, that provides a web-enabled dynamic graphical interface for annotating multimodal content. [1]



Listen to the audio

Select its arousal

○ Very weak[1] ○ Weak[2] ○ Neutral[3] ○ Strong[4] ○ Very Strong[5]

Select its valence

○ Very negative[6] ○ Negative[7] ○ Neutral[8] ○ Positive[9] ○ Very Positive[0]

ID: _QmUy

Figure 1: Screen shot from the LabelStudio project used for annotating the utterances.

Each individual annotated the whole dataset of 500 utterances and the final ground truth has been generated by an annotation aggregation procedure described in Section 3.3.

3.3. Annotations aggregation

After the previously described data annotation process, we applied an aggregation step to extract the final 3-class ground-truth label for each data point (utterance) from the initial individual 5-scale annotations. Furthermore, for each data point we calculated the agreement between the individual annotators so as to quantify the level of complexity of the adopted classification tasks (valence and arousal) for the particular dataset (Eleftheriou et al., 2021).

Final ground truths were obtained by first calculating an average annotation rating (in the 1-5 scale for both classification tasks, as described above) and comparing it against some predefined thresholds, so as to conclude a final 3-class classification for the respective sample. In particular, two thresholds are required to map the aggregated annotations to a 3-class taxonomy. An obvious selection would be to uniformly select the thresholds in the 1-5 range: in that case the thresholds would be $T1 = 2.33$ and $T2 = 3.66$, i.e. any utterance with an average annotation in the $[T1, T2]$ range would be finally assigned to class ”neutral”.

However, we have made some slight modifications from that uniform selection of thresholds, to satisfy the aforementioned (Section 1) shift of the theatrical utterances to more aroused emotional states. For the valence task, the $T1$ was defined to be equal to 2.33 and $T2$ equal to 3.33. Concerning the task of arousal, we used slightly different thresholds: the lower threshold was defined to 2.66 and the higher threshold to 3.66. This minor differentiation between the thresholds for the two tasks stems from the fact that the theatrical data are characterized by a high dominance of negative and strong emotions. So, for instance, the fact that ”weak” arousal would be rather rare, setting the lower

[1] `https://labelstud.io`

threshold $T1 = 2.66$ instead of the "default" uniform value (2.33) makes it possible to decide in favor of the weak class even in borderline cases such as the following: suppose the 4 annotators have given the ratings "weak" (value 2), "weak" (value 2), "neutral" (value 3), "neutral" (value 3), the average value of these ratings is $\frac{2+2+3+3}{4} = 2.5$. In that case, having the uniform $T1$ would make that sample be finally classified as "neutral", while the threshold $T1 = 2.66$ finally classifies that sample as "weak".

To sum up, the allocation of the final labels followed the below thresholding rules:

$$arousal_i = \begin{cases} \text{weak}, & E[A_i] \leq 2.66. \\ \text{neutral}, & 2.66 < E[A_i] < 3.66. \\ \text{strong}, & E[A_i] \geq 3.66 \end{cases}$$

$$valence_i = \begin{cases} \text{negative}, & E[V_i] \leq 2.33. \\ \text{neutral}, & 2.33 < E[V_i] < 3.33. \\ \text{positive}, & E[V_i] \geq 3.33 \end{cases}$$

where $A_{i,\alpha}$ is the arousal annotated value for a sample i of the annotator α (in the 1-5 range) and $E[A_i]$ is the mean annotation of sample i (similarly for valence, let $V_{i,\alpha}$ be the annotation value for valence in the 1-5 range)

In addition to the mean values of valence and arousal annotations that has been used to extract the final ground truths, through the thresholding steps described above, a deviation threshold is also considered, in order to filter out controversial annotations. To this end, the mean absolute deviation (MAD) of the annotations of each sample i is calculated for each annotation X for both tasks:

$$MAD_i = E(|X_{i,\alpha} - E(X_{i,\alpha})|)$$

This is obviously a metric of inter-annotator agreement for the given sample, therefore samples with high MAD values should be excluded from the dataset. In this work, we have adopted the aforementioned threshold to be equal to 1.3. This rule is used as a safety net step to exclude possible examples with extreme inter-annotator disagreements. As an indicative example, consider that two of the annotators have given an "1" ("very weak") label for the arousal task, and two of the annotators "4" ("strong"). This combination yields to a MAD equal to 1.5, therefore the respective example would be excluded.

Inter-annotator agreement: Apart from using MAD to filter out possible highly questionable utterances, we also used it to demonstrate the overall inter-annotator (dis)agreement for each of the classification tasks, by computing its mean value over all samples in the task. High values of average disagreement would have indicated that annotators had different points of view on the ratings.

On top of the total disagreement for all data, the average disagreement for each annotator was also calculated, to evaluate each human annotator compared to the overall ground truth. To this end, we first calculated the disagreement of the annotation of annotator a for the i-th sample. Let $A_{i,a}$ be the value of that annotation (in the 1-5 range), and M be the total number of annotators. Then the deviation of $A_{i,a}$ from the average value of all annotators for the same sample is:

$$u_{i,a} = \left| A_{i,a} - \frac{\sum_{j=1}^{M} A_{i,j}}{M} \right|$$

The results of the filtering procedure and the calculated agreement metrics for the arousal and valence task are listed in Table 1. It has to be noted that the average disagreement of both tasks is below 0.5 (0.48 and 0.49 respectively), which is half the size of the neutral class.

4. Baseline Classification Methods

4.1. Traditional machine-learning-based approach

As a baseline audio classification technique, we have selected to use a set of handcrafted audio features from the time, spectral and cepstral domains, such as Zero Crossing Rate, Spectral Centroid and Mel Frequency Cepstral Coefficients (MFCCs), along with Support Vector Machines as classifiers. In particular, each speech utterance is first split into a sequence of non-overlapping 50 msec short-term windows (frames), and for each frame a set of 68 audio features is computed. At this stage, each speech utterance is represented by a sequence of short-term feature vectors (short-term representation).

Then, the mean and standard deviation of these features are extracted in a long-term segment of 3 seconds, using 1 second step. According to that, each utterance is represented by a sequence of (68 x 2 = 136) feature statistics. Finally, we apply a long-term averaging step, which results in a single-vector 168-D representation for the whole utterance (long-term representation). Note that both short-term (matrix) and long-term (vector) representations are provided in the repository of the dataset.

As a baseline classification method we have experimented with training and evaluating an SVM classifier with an RBF kernel, using the long-term vector representation described above. All respective experiments and feature extraction procedures have been carried out with the pyAudioAnalysis library (Giannakopoulos, 2015).

4.2. Deep-learning-based approach

Using audio spectrograms (or mel-spectrograms) as image inputs to Convolutional Neural Networks (CNNs) is a widely adopted approach in the literature of emotion recognition (Koromilas and Giannakopoulos, 2021). In this work we incorporate

	GreThE					
	Arousal			Valence		
	Strong	Neutral	Weak	Positive	Neutral	Negative
Mean Thresholding	$\mu >= 3.66$	$2.66 < \mu < 3.66$	$\mu <= 2.66$	$\mu >= 3.33$	$2.33 < \mu < 3.33$	$\mu <= 2.33$
Number of samples after Mean Thresholding	227	180	93	91	218	191
Deviation Thresholding	$\sigma < 1.3$	$\sigma < 1.3$	$\sigma < 1.3$	$\sigma < 1.3$	$\sigma < 1.3$	$\sigma < 1.3$
Number of samples after Deviation Thresholding	227	180	93	91	218	191
Average Disagreement	**0.48**			**0.49**		

Table 1: Definition of Arousal and Valence Dataset

this methodology as a baseline from the field of Deep Learning. While this method is expected to underperform when applied to small datasets, it is perfectly suited for testing whether knowledge from large amount of emotion data can be applied to our proposed dataset. For that reason we use the deep_audio_features library to extract mel-spectrograms and train a CNN for our two speech emotion recognition tasks.

5. Experimental Results

5.1. Experimental Setup

5.1.1. Session-independent validation

The traditional feature extraction and SVM approach described in Section 4.1, has been evaluated using a repeated random shuffling train/validation split. The ”session” ID used to split the data was based on the ID of the theatrical play. In this way, we guarantee that the evaluation results we report in the Results Section are not assuming dependence on the theatrical plays and are therefore realistic. Note, that this ID is provided with the dataset, as well. Moreover, we have used the aforementioned validation strategy to experiment against different values of the C parameter. Finally, since the datasets is imbalanced in both classification tasks (valence and arousal), we have adopted a data balancing step using either a basic random subsampling (of the dominant classes) or SMOTE oversampling (synthetic minority over-sampling technique, (Chawla et al., 2002)).

5.1.2. Cross-domain validation

Intending to examine whether the domain knowledge of Speech Emotion Recognition can be transferred on our dataset, we train a CNN in the way described in 4.2, on two widely used datasets, the MSP-podcast (Martinez-Lucas et al., 2020) and IEMOCAP (Busso et al., 2008). Following (Metallinou et al., 2012) we define two three-class problems on these datasets, namely arousal (or valence) with classes *(i)* weak (or negative) for values in the range [1, 2] (or [1,3]) for the IEMOCAP (or MSP-podcast) dataset, *(ii)* neutral (or neutral) for values in the range (2, 4) (or (3,5]) for the IEMOCAP (or MSP-podcast), and *(iii)* strong (or positive) for values in the range [4, 5] (or (5,7]) for IEMOCAP (or MSP-podcast).

Three CNNs are trained in total, one for each of the aforementioned datasets and one for their merged combination. The resulted models, namely CNN_iemocap, CNN_msp and CNN_merged, are subsequently used for testing in the GreThE dataset. That is, none of the GreThE instances are used for the CNNs' training, and thus the used models are completely unaware of the recording conditions and acted speech of a theatrical play.

Our CNN architecture consists of 4 convolutional and 3 linear layers, including batch normalization (Ioffe and Szegedy, 2015) and the LeakyRelu activation function. We trained the model using the Cross-Entropy loss as loss function, while Adam was chosen to be the optimizer with initial learning rate of 0.002 and a reduce-on-plateau learning rate scheduler.

5.1.3. Baseline models

In order to compare our methods with the threshold of a random classifier, one randomized classifier is defined for each of the two approaches. Specifically, since the traditional machine-learning-based approach (section 4.1) is trained on the GreThE dataset and thus is aware of the sample distribution, it can only be compared to a prior-aware randomized (prior-aware baseline) classifier, ie. a classifier that randomly predicts class a based on the prior probability of class a in the dataset distribution. On the other hand, the cross-domain trained models, as defined in section 5.1.2, are not familiar with GreThE and thus their predictions can only be compared with a completely randomized (baseline) classifier.

5.2. Results

In table 2 we report the f1 metrics that emerged from our evaluation.

As can be clearly inferred from the provided results, *(i)* the SVM session-independent method achieves

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 2: GreThE evaluation results

27.9% relative improvement for arousal and 21.2% for valence, and *(ii)* the CNN cross-domain method achieves 20.6 % for arousal and for 17.5% valence.

The small, in absolute values, improvement when using cross-domain validation indicates that *(i)* general knowledge from the task of emotion recognition can be transferred to the new domain, but *(ii)* the problem of theatrical speech emotion recognition differs from that of speech emotion recognition. The second point is easily verified by the fact that the actors expressiveness in theatrical plays usually result in increased arousal, shifting (in terms of energy and frequency) the weak and neutral classes towards the strong one.

6. Dataset Availability

The dataset is public available at `https://github.com/magcil/GreThE`. Each utterance is represented by either (a) a sequence of short-term feature vectors or (b) a spectrogram. Ground-truth is provided in a simple tabular CSV format for both classification tasks (valence and arousal). Finally, the repository also contains Python scripts that demonstrate (a) the aforementioned SVM experimental setup and (b) the adopted feature extraction methods (so that reproducability can also be made possible and combined with other sources of data).

7. Conclusion

In this paper we presented the Greek Theatrical Emotion dataset *GreThE*, a new publicly available data collection for speech emotion recognition in Greek theatrical plays. The dataset contains 500 utterances that have been annotated in terms of their emotional content (valence and arousal). Multiple-annotator data have been used to assure annotation quality. To our knowledge, this is the first dataset with real-world speech data from theatrical plays annotated in terms of the underlying emotion.

Also, we have presented classification performance results for a baseline machine learning classification approach cross-validated on *GreThE*, along with results using *GreThE* as a test dataset for cross-domain deep emotional models, trained on popular datasets of the English language. Results have proven that (a) the task of recognising emotion - and mostly valence - is rather challenging in theatrical data when training from scratch (b) using state-of-the-art datasets from generic SER on cross-language theatrical data is not effective. This indicates that future works in the field of recognizing emotions in theatrical data should probably consider robust domain adaptation techniques using few-shot learning strategies.

8. Acknowledgements

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-01359)

9. Bibliographical References

Basu, S., Chakraborty, J., Bag, A., and Aftabuddin, M. (2017). A review on emotion recognition using speech. In *2017 International conference on inventive communication and computational technologies (ICICCT)*, pages 109–114. IEEE.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of german emotional speech. volume 5, pages 1517–1520, 09.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S., and Narayanan, S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. (2017). Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Cao, H., Cooper, D., Keutmann, M., Gur, R., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5:377–390, 10.

Cao, H., Verma, R., and Nenkova, A. (2015). Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Computer speech & language*, 29(1):186–202.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M. (2014). Emovo corpus: an italian emotional speech database. In *LREC*.

Cui, C., Ren, Y., Liu, J., Chen, F., Huang, R., Lei, M., and Zhao, Z. (2021). Emovie: A mandarin emotion speech dataset with a simple emotional text-to-speech model.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Eleftheriou, S., Koromilas, P., and Giannakopoulos, T. (2021). Automatic assessment of speaking skills using aural and textual information. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 166–175, Trento, Italy, 12–13 November. Association for Computational Linguistics.

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

Gloor, P. A., Araño, K. A., and Guerrazzi, E. (2019). Measuring audience and actor emotions at a theater play through automatic emotion recognition from face, speech, and body sensors.

Gournay, P., Lahaie, O., and Lefebvre, R. (2018). A canadian french emotional speech dataset. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, page 399–402, New York, NY, USA. Association for Computing Machinery.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org.

Jackson, P. and ul haq, S. (2011). Surrey audio-visual expressed emotion (savee) database, 04.

Kadiri, S. R., Gangamohan, P., Mittal, V. K., and Yegnanarayana, B. (2014). Naturalistic audio-visual emotion database. In *ICON*.

Koromilas, P. and Giannakopoulos, T. (2021). Deep multimodal emotion recognition on human speech: A review. *Applied Sciences*, 11(17):7962.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B., and et al. (2021). 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.

Livingstone, S. R. and Russo, F. A. (2018). 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.

Ma, Y., Nguyen, K. L., Xing, F. Z., and Cambria, E. (2020). A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Martinez-Lucas, L., Abdelwahab, M., and Busso, C. (2020). The msp-conversation corpus. *Interspeech 2020*.

Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., and Narayanan, S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198.

Muljono, Prasetya, M. R., Harjoko, A., and Supriyanto, C. (2019). Speech emotion recognition of indonesian movie audio tracks based on mfcc and svm. In *2019 International Conference on contemporary Computing and Informatics (IC3I)*, pages 22–25.

Ortony, A. and Turner, T. (1990). What's basic about basic emotions? *Psychological review*, 97:315–31, 08.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12.

Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C., and Kalliris, G. (2018). Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society. Audio Engineering Society*, 66:457–467, 06.

Wundt, W. and Judd, C. (2011). *Outlines of Psychology*. W. Engelmann.