



Analysis of Trustworthiness Recognition models from an aural and emotional perspective

Cristina Luna-Jiménez¹, Ricardo Kleinlein¹, Syaheerah Lebai Lutfi², Juan M. Montero¹, and Fernando Fernández-Martínez¹

¹ Grupo de Tecnología del Habla y Aprendizaje Automático (T.H.A.U. Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, 28040, Madrid, Spain; ²School of Computer Sciences, Universiti Sains Malaysia, 11800, Minden, Pulau Pinang, Malaysia

crisrina.lunaj@upm.es, ricardo.kleinlein@upm.es, syaheerah@usm.my,
juanmanuel.montero@upm.es and fernando.fernandezm@upm.es

Abstract

Trustworthiness and deception recognition attracts the research community attention due to their relevant role in social negotiations and other relevant areas.

Despite the increasing interest in the field, there are still many questions about how to perform automatic deception detection or which features explain better how people perceive trustworthiness.

Previous studies have demonstrated that emotions and sentiments correlate with deception. However, not many articles employed deep-learning models pre-trained on emotion recognition tasks to predict trustworthiness. For this reason, this paper will compare traditional statistical functional feature sets proposed for performing emotion recognition, such as eGeMAPs, with features extracted from deep-learning models, like AlexNet, CNN-14 or xlsr-Wav2Vec2.0 pre-trained on emotion recognition tasks. After obtaining each set of features, we will train a Support Vector Machine (SVM) model on deception detection.

These experiments provide a baseline to understand how methodologies exploited in emotion recognition tasks could be applied to speech trustworthiness recognition. Utilizing the eGeMAPs feature set on deception detection achieved an accuracy of 65.98% at turn level, and employing transfer-learning on the embeddings extracted from a pre-trained xlsr-Wav2Vec2.0 let improve this rate until a 68.11%, surpassing the baseline on audio modality from previous works by an 8.5%.

Index Terms: trustworthiness recognition, audio, deception detection human-computer interaction, emotion, deep-learning, functionals

1. Introduction and Related Works

Automatic deception detection and trustworthiness recognition could have a beneficial impact on our societies. These systems could be used for noticing when a person is lying during a negotiation, or in conversational agents to create more engaging interactions between users and devices by detecting loss of trust or changes in emotional states of the user caused by the machine decisions [1]. For this reason, this field is gaining attention from the research community in recent years.

Previous studies have demonstrated that emotions play a crucial role in trustworthiness perception [2, 3, 4]. For example, in the investigation of J. J. Lee et al. [4], they reported that non-verbal features like smiling correlated positively with higher trustworthiness perception, encouraging a collaborative

aptitude in an exchange money game. In the same way, the research of J. R. Dunn et al. [5] reported that people experiencing happiness usually gave higher ratings on trustworthiness surveys compared to those feeling sadness or anger. In [6], they also explored the influence of anger emotions on trust behaviors.

Regarding emotion recognition, before the apparition of deep-learning models, the most common trend was to study spectral, voice quality, and prosodic features considered paralinguistics attributes that could encapsulate emotional information from speech signals. Some tools, such as OpenSmile [7] and Praat [8], appeared to simplify the feature generation process, extracting these sets of characteristics automatically. Among the attribute collections, eGeMAPs [9], and those derived from Interspeech Emotion and paralinguistic challenges [10, 11, 12] were the most popular, employed in several emotion recognition tasks as baselines [13, 14]. However, with the development of Convolutional Neural Networks (CNNs) and deep-learning models able to process and handle aural signals, the research community focused on exploiting pre-trained models to solve emotion or sentiment recognition by applying transfer learning [15, 16].

Following this line of study, this article aims to understand trustworthiness from computer science and emotional perspectives. With this aim, we explored how typical features and models employed in emotion recognition problems perform when they are used for solving deception detection tasks.

First, we will extract eGeMAPs features, a set of hand-crafted attributes, establishing a baseline. Later, we will apply a transfer-learning strategy, extracting features from deep-learning models and training a final model with them. These transfer-knowledge analyses will be subdivided into two experiments: (1) extracting features from models trained on generic tasks not related to emotion recognition, and (2) extracting features from models pre-trained on an emotion recognition problem using RAVDESS dataset [17].

The structure of the paper is as follows: Section 1 provides a review of existing publications in the trustworthiness field. Section 2 summarises the followed methodology, including the description of the datasets, the evaluation strategy, and the experimentation pipeline. Next, in Section 3 we analyze the results obtained by comparing hand-crafted features against deep-learning models. Here, we will also comment on the differences between applying fine-tuning compared to feature extraction in this scenario, and the effects of removing background noise from the recordings on the deception detection scores. Finally, in Sec-

tion 4 and Section 5, we provide concise conclusions regarding our work and results, as well as its main limitations.

2. Methodology

2.1. Datasets

Two main datasets were employed in this study: Bag Of Lies [18] and RAVDESS [17]. Bag Of Lies is annotated according to honest and deceptive descriptions, whereas RAVDESS contains labels of emotions. In this section, we will detail the characteristics of both datasets and explain how they were used in our experiments.

2.1.1. Bag Of Lies

The main corpus used for this project is Bag-of-lies (BOL)[18], an open-source dataset for the research community, annotated in terms of objective trustworthiness [19]. This dataset includes 325 recordings from 35 subjects (10 female and 25 male) describing an image in a true or deceptive manner. Overall, 162 samples contained lies, and 163 had trustful descriptions. Each recording is composed of a video, the image used as stimuli to describe, gaze features extracted automatically using the Gaze-Point Open Gaze API, and EEG features generated by the Emotiv EPOC+EEG headset device. Of the 35 users available, we only employed 34 during the evaluation because user 12 had some corrupted fields. Hence, a total of 315 recordings with an average duration of 13.38 seconds (Min: 3.81 seconds - Max: 42.19 seconds), 157 non-honest, and 158 honest.

2.1.2. RAVDESS

RAVDESS is the dataset chosen for training the deep-learning models in an emotional recognition task to study whether the trained weights could generate valuable attributes for solving a deception recognition task or not.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [17] contains 7356 recordings with acted-emotional content. These files are divided into three modalities (full AV, video-only, and audio-only) and two vocal channels (speech and song). Each file contains a single actor representing an emotion that could be one of the eight following categories: calm, neutral, happy, sad, angry, fearful, surprised, and disgusted. These expressions are produced at two levels of emotional intensities (regular and strong) except for the neutral emotion that only contains regular intensity.

For our experiments, we only used the full AV material and the speech channel. This selection reduced the number of files to 1440 videos that had a maximum and minimum duration of 5.31 and 2.99 s, respectively. Among its advantages, it had a balanced number of files per emotion and gender, which avoided problems derived from training algorithms with non-balanced data.

2.2. Evaluation

Initially, the BOL corpus only contained 315 recordings for learning whether a person is deceiving or not. As this amount is usually not enough for training deep models, we replicated the set-up proposed in [20]. First, the audios were down-sampled to 16kHz and the first seconds of the videos were cut to remove a spurious voice that indicated the beginning of the experiment. Later, each video was divided into windows of 3 seconds, with an overlap of 2.5 seconds.

After this pre-processing, the total number of segments of

3 seconds was 5.249, of which 2.711 belonged to people lying and 2.538 with honest descriptions.

For evaluating our proposals, we adopted the subject-wise 3-fold cross-validation method presented in the reference study of A. Gallardo-Antolín et. al [20]. In this work, we replicated their baseline speech-based experiment using SVMs to compare the reached performance by our extracted features in the same model, reporting the average accuracy of the cross-validation strategy on each experiment.

In the experimentation phase, models were trained with the 5.249 segments derived from the audios, calculating the turn level accuracy by applying a majority voting procedure on the predictions of all the chunks that composed the original recording.

For the training of the emotion recognition models, we used a subject-wise train-test division of the RAVDESS dataset in which users 3,6,7,13 and 18 formed the test set and the remaining 19, the training set. We tested three different architectures: AlexNet [21], CNN-14 [22], and xlsr-Wav2Vec2.0 [23]. In each case, we chose the checkpoint that achieved the highest accuracy in the test set of RAVDESS for being used in the transfer-learning.

Following the hypothesis that models trained on emotion recognition tasks should generate suitable features for solving deception detection. We extracted the d-vectors of the pre-trained models on emotion recognition for the BOL recordings.

2.3. Feature Extraction

The studied features can be classified into hand-crafted and D-vectors. The hand-crafted features are statistics calculated on the raw waveform or from spectral components of the audio. On the other hand, D-vectors are obtained from pre-trained deep-learning models that in this study are: AlexNet, CNN-14 and xlsr-Wav2Vec2.0.

2.3.1. Hand-crafted Features

The first set of hand-crafted features was proposed by A. Gallardo-Antolín et. al [20]. It consists of 192 statistical functionals calculated as a result of computing the average, standard deviation, kurtosis, skewness, max, and min over the output of the 32 filters applied to the mel spectrograms of each audio-chunk. This set is adopted as the baseline.

The second set of hand-crafted features is eGeMAPS [9]. One of the main advantages of this set is that it combines attributes of different natures: frequency-based, energy/amplitude, spectral, temporal, and quality-based parameters. Overall, it contains 42 low-level descriptors, from which statistics such as the mean or the standard deviation are obtained, resulting in 88 functional. Due to its diversity, this set was used on emotion recognition tasks as baseline [13, 24].

2.3.2. D-vectors

Apart from the classic feature sets used in emotion recognition, d-vectors are extracted from several deep architectures to solve speech trustworthiness recognition. Specifically, the evaluated models are: AlexNet [21], CNN-14 from PANNs library[22] and xlsr-Wav2Vec2.0 [23].

AlexNet embeddings were obtained from its 4.096-dimensional last-fully connected layer after the ReLU activation function. Though the default version of AlexNet available in Pytorch [25] was trained on the ImageNet dataset for object recognition, several studies have also used this convolutional

Strategy	# Features	Type of Norm.	Segment Level		Turn Level	
			Val ACC \pm CI	Test ACC \pm CI	Val. ACC \pm CI	Test ACC \pm CI
Baseline [20]	192	z-score	49.27 \pm 1.35	62.60 \pm 1.31	53.54 \pm 5.51	59.61 \pm 5.42
eGeMAPS	88	z-score	53.97 \pm 1.34	64.35 \pm 1.30	53.35 \pm 5.51	65.98 \pm 5.23
FE-AlexNet -ImageNet pre-trained-	4096	z-score	47.83 \pm 1.35	64.66 \pm 1.29	51.44 \pm 5.52	64.29 \pm 5.29
FE-AlexNet -RAVDESS pre-trained-	4096	z-score	49.31 \pm 1.35	60.62 \pm 1.32	50.17 \pm 5.52	58.56 \pm 5.44
FE-PANNs (CNN-14) -AudioSet pre-trained-	2048	No Norm.	48.15 \pm 1.35	60.18 \pm 1.32	50.33 \pm 5.52	58.65 \pm 5.44
FE-PANNs (CNN-14) -RAVDESS pre-trained-	2048	No Norm.	46.78 \pm 1.35	57.66 \pm 1.34	48.98 \pm 5.52	59.50 \pm 5.42
FE-xlsr-Wav2Vec2.0 -CommonVoice pre-trained-	512	Scale Min-Max	50.48 \pm 1.35	65.76 \pm 1.28	54.14 \pm 5.50	67.72 \pm 5.16
FE-xlsr-Wav2Vec2.0 -RAVDESS pre-trained-	512	No. Norm	49.36 \pm 1.35	66.53 \pm 1.27	47.30 \pm 5.51	68.11 \pm 5.14

Table 1: Comparison of SVM performances trained on hand-crafted features or features extracted from deep-learning models trained on non-emotional datasets (ImageNet, AudioSet) and on an emotional dataset (RAVDESS). *ACC = Accuracy, *CI= Confidence Interval, *FE=Feature Extraction

network for solving aural-related problems [26]. Thus, we followed a similar procedure and converted the speech signals into images of the spectrogram of the recordings to feed them into the network and receive the embeddings.

CNN-14 last-layer embeddings have a dimension of 2.048. This model is more complex than AlexNet concerning the number of trainable parameters. However, the main difference is that this network was trained in a sound recognition task on the AudioSet corpus, with raw audios. Hence, the only pre-processing performed on the audios was downsampling them to 16 kHz before passing them to the network.

The last d-vectors of the experiments were extracted from the xlsr-Wav2Vec2.0 transformer of HuggingFace library [27] ('jonatasgrosman/wav2vec2-large-xlsr-53-english'). We obtained 512-dimensional vectors from the output of the convolutional feature encoder and calculated the average of these embeddings along its temporal dimension. This pooling step is required because the framework internally processes the audios by dividing them into windows of 25 ms with an overlap of 15 ms and a stride of 20ms. Hence, the outputs of each recording should be combined with an agglomerative strategy, that in our case is an average across the temporal dimension. As it happened with CNN-14, this model was pre-trained on a speech processing task, specifically, for performing Automatic Speech Recognition, using the CommonVoice dataset [28].

For all these models, we compared the performance between using embeddings from their default described pre-trained tasks, against the performance achieved by the models after training them on an emotion recognition task on the RAVDESS dataset.

2.4. SVM Model

After extracting the features and before training the final model, we compared three types of normalizations in each experiment: (1) not using normalization, (2) scale features in a min-max range, and (3) z-score normalization. In Section 3, only the normalization that achieved the best scores is reported in each

case.

After the normalization step, attributes or D-vectors were passed to an SVM model [29]. All the SVMs were implemented in sklearn [30] with the default parameters of the library except for the regularization parameter (C) that we set to 1. We selected this model because of its simplicity, relatively high performance in many tasks, and because it was used in [20] as the baseline too.

3. Results and Discussion

Table 1 summarizes the performance of the different SVMs models trained on the hand-crafted features (Baseline and eGeMAPS) and with the extracted features from the three deep-learning models tested (AlexNet, CNN-14, and xlsr-Wav2Vec2.0). In each case, the normalization with the top performance appears in the table. Also, the table reveals the differences between using embeddings extracted from a default task against the models trained on emotion recognition with the RAVDESS dataset. Results exhibit an improvement concerning the accuracy achieved in the baseline work [20].

Focusing on the performance of eGeMAPS features, it reaches an accuracy of 64.35% at the segment level and 65.98% at the turn level. Despite the reduced number of attributes of the set (88), their metrics surpass the baseline and the achievements of AlexNet and CNN-14.

Analyzing the metrics accomplished by the deep-learning models, we can see that the average embeddings extracted from xlsr-Wav2Vec2.0 reached the maximum accuracy at the turn level (68.11%), over the accuracy of eGeMAPS, although not in a statistically significant way.

Finally, comparing the results attained applying transfer-learning on default model tasks (image recognition, sound classification, or ASR) against emotion recognition, experiments suggest a slight improvement in the test set at turn level when employing pre-trained models on emotion recognition tasks, except for AlexNet architecture. These differences between networks could be related to their performance on the emotion

recognition task using RAVDESS (68.30% with AlexNet, 81% with CNN-14, and 84.67% with xlsr-Wav2Vec2.0). These outcomes suggest that not only the similarity between tasks may influence transfer-learning strategies, but also the accuracy of the model used as a warm-start. Nevertheless, more experiments should be accomplished with other datasets and analyze different stop points on the emotion recognizers to conclude that applying transfer-learning on pre-trained emotion recognition tasks implies an improvement in a deception detection task.

As the top test accuracy at the turn level was obtained with the xlsr-Wav2Vec2.0 model, we decided to fine-tune it on the deception detection task with the 5.249 chunks of the BOL dataset to try to adapt the model more to the problem. We adjusted the network by introducing a global average pooling on top of the output of the transformer module, as described in [15]. Contrary to expected, the achieved accuracy was lower than applying the feature-extraction strategy, 62.41% on the test set at the segment level, and 64.09% at the turn level. One of the possible explanations for these outcomes could be a reduced number of samples available or the low variability of the data which was not enough for fine-tuning a transformer-type model.

Strategy	Segment Level	Turn Level
	Test ACC \pm CI	Test ACC \pm CI
Baseline [20]	60.92 \pm 1.32	58.38 \pm 5.44
eGeMAPS	59.23 \pm 1.33	61.87 \pm 5.36
FE-xlsr-Wav2Vec2.0 -RAVDESS pre-trained-	58.98 \pm 1.33	63.81 \pm 5.31

Table 2: Comparison of top models after removing noise from audios using NVIDIA Audio Effects SDK. *ACC = Accuracy, *CI= Confidence Interval, *FE=Feature Extraction

Since the audios contained high background noise, we considered using a denoising tool. We passed the audios through the NVIDIA Audio Effects SDK to remove spurious frequencies. After that, we repeated the top experiments using these cleaned audios. Nonetheless, as we can see in Table 2, results are below the obtained with ‘noisy-default’ audios. These scores point out that despite improving the comprehensibility of the recordings, the removed frequencies could include relevant information for recognizing trustworthiness that disappears when these frequencies are removed.

4. Limitations

One of the main limitations of the study is the small number of samples available for training a deep-learning model or fine-tuning it. For this reason, metrics obtained when fine-tuning xlsr-Wav2Vec2.0 models are lower than applying feature extraction. However, this dataset was chosen because it is one of the biggest available in the field and, because it was used in previous studies. Another constraint is the emotion recognition dataset. Maybe using another dataset with not-acted emotions could be more related to the automatic deception detection task with ‘in the wild’ data.

5. Conclusions

This study aimed to explore how traditional features and deep-learning models that were applied successfully in emotion recognition tasks could be used in automatic deception detection problems, due to the correlation between emotions and trustworthiness reported in previous studies [2, 3, 4].

Our results reveal that improvements can be achieved when

applying eGeMAPs feature sets or extracting features from xlsr-Wav2Vec2.0 transformer, surpassing the baseline based on statistical features proposed in [20] by 3.93 points and 8.50 points, on the test set at turn level respectively. These results encourage us to think that previous methods employed in emotion recognition could be also applied for automatizing the recognition of other psychological states, such as deception detection.

Regarding the experiments applying transfer-learning of emotion recognition models for solving deception detection, the evidence achieved is not conclusive because improvements between strategies are not significant. Several reasons could explain this behavior, although the most likely is that acted emotions annotated in RAVDESS do not correlate with deception recognition as natural-expressed emotions might do.

In future works, we would explore feature selection techniques to reduce the number of features used during the training of the SVMs since complexity and the high dimensions of the vectors could affect the final performance of the models. Also, we would wish to explore ‘in-the-wild’ emotion-recognition datasets to understand if acted emotion could not be as related to deception detection as genuine expressions that appear in a real scenario. Finally, we will invest some efforts in analyzing different performances obtained on deception detection by scanning several stop points of the networks for the emotion recognition task to understand how much the accuracy of the source task (emotion recognition) could affect the target task (deception detection).

6. Acknowledgements

The work leading to these results was supported by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C21 and PID2020-118112RB-C22, funded by MCIN/AEI/10.13039/501100011033), and AMIC-PoC (PDC2021-120846-C42, funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU/PRTR”).

The authors thank Universiti Sains Malaysia for the partial funding of this work from the grant no. 304/PKOMP/6315137.

We also thanks the ‘Ayudas del Programa Propio UPM 2022’ for partial funding the mobility for performing a stay to complete this study.

Furthermore, Ricardo Kleinlein’s research was supported by the Spanish Ministry of Education (FPI grant PRE2018-083225).

7. References

- [1] E. J. de Visser, R. Pak, and T. H. Shaw, “From ‘automation’ to ‘autonomy’: the importance of trust repair in human-machine interaction,” *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, 2018, pMID: 29578376. [Online]. Available: <https://doi.org/10.1080/00140139.2018.1457725>
- [2] D. F. Galinsky, E. Erol, K. Atanasova, M. Bohus, A. Krause-Utz, and S. Lis, “Do I trust you when you smile? effects of sex and emotional expression on facial trustworthiness appraisal,” *PLoS One*, vol. 15, no. 12, p. e0243230, Dec. 2020.
- [3] F. Caulfield, L. Ewing, S. Bank, and G. Rhodes, “Judging trustworthiness from faces: Emotion cues modulate trustworthiness judgments in young children,” *Br. J. Psychol.*, vol. 107, no. 3, pp. 503–518, Aug. 2016.
- [4] J. J. Lee, B. Knox, J. Baumann, C. Breazeal, and D. DeSteno, “Computationally modeling interpersonal trust,” *Frontiers in Psychology*, vol. 4, 2013. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fpsyg.2013.00893>

- [5] J. R. Dunn and M. E. Schweitzer, "Feeling and believing: The influence of emotion on trust," *Journal of Personality and Social Psychology*, pp. 736–748, 2005.
- [6] K. Zhang, T. Goetz, F. Chen, and A. Sverdluk, "Angry women are more trusting: The differential effects of perceived social distance on trust behavior," *Front. Psychol.*, vol. 12, p. 591312, Jul. 2021.
- [7] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: <https://doi.org/10.1145/1873951.1874246>
- [8] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glott International*, vol. 5, no. 9/10, pp. 341–345, 2001.
- [9] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GEMAPs) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016, open access.
- [10] B. W. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *INTER_SPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*. ISCA, 2009, pp. 312–315. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2009/i09_0312.html
- [11] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *INTER_SPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, T. Kobayashi, K. Hirose, and S. Nakamura, Eds. ISCA, 2010, pp. 2794–2797. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2010/i10_2794.html
- [12] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The interspeech 2014 computational paralinguistics challenge: cognitive & physical load," in *INTER_SPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, H. Li, H. M. Meng, B. Ma, E. S. Chng, and L. Xie, Eds. ISCA, 2014, pp. 427–431. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2014/i14_0427.html
- [13] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 478–484. [Online]. Available: <https://doi.org/10.1145/3123266.3123371>
- [14] F. Haider, S. Pollak, P. Albert, and S. Luz, "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods," *Computer Speech & Language*, vol. 65, p. 101119, 2021.
- [15] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset," *Applied Sciences*, vol. 12, no. 1, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/1/327>
- [16] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment analysis using image-based deep spectrum features," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017, pp. 26–29.
- [17] S. R. Livingstone and F. 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*, vol. 13, no. 5, p. e0196391, May 2018.
- [18] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-lies: A multimodal dataset for deception detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 83–90.
- [19] G. Lucas, G. Stratou, S. Lieblisch, and J. Gratch, "Trust me: Multimodal signals of trustworthiness," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 5–12. [Online]. Available: <https://doi.org/10.1145/2993148.2993178>
- [20] A. Gallardo-Antolín and J. M. Montero, "Detecting deception from gaze and speech using a multimodal attention lstm-based framework," *Applied Sciences*, vol. 11, no. 14, 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/14/6393>
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, May 2017. [Online]. Available: <https://doi.org/10.1145/3065386>
- [22] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [23] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.
- [24] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.
- [25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. NY, United States: Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [26] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Interspeech 2017*. ISCA, August 2017, pp. 3512–3516.
- [27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, October 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [28] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: <https://aclanthology.org/2020.lrec-1.520>
- [29] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 2000.
- [30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.