IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Using Audio Events to Extend a Multi-modal Public Speaking Database with Reinterpreted Emotional Annotations

Esther Rituerto-González, Clara Luis-Mingueza, Carmen Peláez-Moreno

Signal Theory and Communications Department University Carlos III of Madrid, Spain
erituert@ing.uc3m.es, cluis@pa.uc3m.es, carmen@tsc.uc3m.es

Abstract

Emotions present in speech provide a lot of information about the emotional state of a speaker. Affective Computing is an emerging field that analyses these states and tries to improve human-computer interaction tasks.

In this paper we aim to present a preliminary study on the analysis of stress in speech and acoustic events that may possibly cause it. We merge four speech & audio technologies: speaker and emotion recognition and acoustic event detection and classification, and explore how they influence each other.

We perform initial experiments on BioSpeech, a multi-modal emotions database we have extended with acoustic events and discuss a novel labelling process targeted to improve the classification performance.

The current study is intended as a classification and detection baseline for the mono-modal speech tasks described, and presents a discussion on the future work and multi-modal architectures to be implemented in a cyberphysical system for gender-based violence automatic detection.

Index Terms: affective computing; speaker recognition; acoustic event detection; acoustic scene analysis

1. Introduction

Acoustic Scene Analysis and Interpretation aims to process and interpret the acoustic information in the environment usually captured by a multi-microphone acquisition system. In this paper, we are concerned with a specific type of acoustic scene: that in which Gender-based Violence (GV) appears. Indeed, GV is one of the biggest social problems in the world. Its eradication is essential for achieving gender equality, the fifth of the Sustainable Development Goals (SDG) adopted by all UN Member States in 2015, as part of the 2030 Agenda for Sustainable Development.

In particular, we aim at *detecting* this kind of scene by using Bindi [1, 2], a multimodal cyberphysical system. Bindi uses bio-signal processing technologies, wearable edge computing, and machine learning –among other disciplines. Unlike other technological solutions that mainly focus on providing the means for the victims to call for help (camouflaged panic buttons with geolocalization [3], voice-activated alert devices, etc.), Bindi employs intelligent bio-signal processing for affective computing to *autonomously detect a violent situation*. These bio-signals are collected by smart sensors, including a microphone, integrated in wearable edge devices.

In this preliminary paper, we will limit our scope to the processing of the auditory modality, leaving its effective integration with the physiological signals ([4, 5]) for further developments. It is important to remark that the complexity and energy restrictions typical of cyberphysical systems prevent the uninterrupted collection of this auditory information. In our setup, the microphone is woken up by an alarm from the physiological signals[1]. This alarm is configured in a conservative way with a high false alarm rate to prevent misses. Therefore, the auditory information is used to provide the *context* necessary to *disambiguate* the information carried by the physiological signals.

Thus the challenges we face in this paper are: first, the lack of appropriate datasets for this task that combines physiological and auditory signals and second, the interaction between speech and audio technologies in our setup. In particular, we describe a principled means to extend an existing multi-modal emotions database with violence related audio events and initial experiments to assess their effects on speaker and stress detection tasks on the one hand, and acoustic event detection and classification, on the other.

This paper is organized as follows: first, we outline the state of the art in Sec. 2, our reinterpretation and extension of the targeted dataset is in Sec. 3 followed by experiments and results in Sec. 4, and closing discussion and further work in Sec. 5.

2. Related work

Few studies investigate the relationship between acoustic events and the elicitation of fear [6], and there are no databases –to the knowledge of the authors– that include acoustic events and speech recorded consequentially. A large-scale dataset of manually annotated audio events is AudioSet [7].

As for databases containing real-life speech under panic or fear circumstances, literature is scarce, but at the moment there are a few in which stressed speech is either simulated or recorded under real conditions, such as SUSAS [8], UT-Scope [9], VOCE [10] and BioSpeech [11]. As films are one of the most effective ways to elicit emotions [12], our team UC3M4Safety is currently recording a database for fear recognition in the context of gender-based violence in the EMPATIA project [13].

Stress is barely considered an emotion although it is intimately related to anxiety and nervousness, being closely connected to fear. Among the measurable physiological consequences of stress appear respiratory changes, increased heart rate, skin perspiration, and increased muscle tension of the vocal folds and vocal tract [14].

The most suitable database to our interests is BioSpeech (BioS-DB) [11], since it includes continuous-time annotations in the Arousal/Valence space for non-acted speech presumably stressed due to its public speaking setting, and incorporates physiological data (Blood Volume Pulse –BVP– and Skin Conductance –SC–) as in Bindi, which could be of great use for multi-modal models in the future. Note, however, that the purpose of this dataset collection is quite different from ours since their creators aimed at predicting bio-signals from speech.

[1] There are also short wake up periods to check the functionality and update certain parameters periodically.

10.21437/IberSPEECH.2021-13

In general, the main difficulty with emotionally labelled data relies on the proper labelling process. There is no universal agreement on how to categorize or measure emotions. The self-assessment measures annotations by a specific subject can differ from labels annotated by external evaluators observing the said subject. We will further discuss the impact of this issue on our reinterpretation of BioS-DB in Section 3.1.

On the other hand, the emotional state of the subject could influence negatively the performance of any speech technology and in particular, their identification. This constitutes a challenge we will refer as to *affective speaker recognition*. Tracking the user's voice separating it from the rest of the speakers opens an interesting possibility for situations where it would be desirable to identify all the speakers involved in the scene, e.g., in case of legal evidence required.

In this study we present a synthetic augmentation of BioS-DB with acoustic events that most likely could cause an stressful reaction loosely synchronized with the time instants where the labels denote an acute stress occurrence. We introduce a reinterpretation of the labelling of BioS-DB, more suitable for our classification task. Moreover, we introduce the problem of the relation of acoustic events and emotions, and use shallow deep learning models to establish a baseline for mono-modal emotion (ER) and speaker recognition (SR), and a pretrained deep learning model for acoustic events detection (AED) and classification (AEC) tasks.

3. Methodology

3.1. Relabelling of BioSpeech

BioS-DB [11] is a multi-modal public speaking database which includes continuous-time emotional annotations. It consists of 55 speakers reading two texts, one in German and one in English, while their physiological variables (BVP, SC) and speech are being recorded. This database responds to the idea that performance anxiety can happen when speaking aloud and can be reflected in the physiological variables and speech. Three annotators with previous training use a joystick to obtain continuous time labels for the emotional state of the speaker in a 2D space of which their axis represent *arousal* and *valence*.

The authors of BioS-DB used the evaluator weighted estimation (EWE) for computing the collective time-continuous ratings when creating a gold standard for the emotional labels from the three individual time-continuous annotations [15].

Though EWE is reliable when the number of annotators is large, in this case the possibility of disparity in the ratings is very high. The subjective evaluation of each scorer affects their ratings, besides the bias of the possible comparisons between consecutive speakers. These factors can induce to a lot of variability and discrepancies, and a weighted combination of the labels of each annotator may not be the optimal merging method. This was specially damaging for our purposes, that are different from those of the creators of the dataset.

Thus, we propose a re-labelling of BioS-DB Arousal and Valence values quantizing them into 4 categorical quadrants. This is crucial to define a classification task instead of using a regressor. These four quadrants are:

- High Arousal, High Valence: Excitement (Q1)
- High Arousal, Low Valence: Stress (Q2)
- Low Arousal, Low Valence: Sadness (Q3)
- Low Arousal, High Valence: Calmness (Q4)

We also believe that although BioS-DB counts with a very precise temporal resolution in the labelling, coarser time resolution for capturing the underlying emotions in speech is more suitable in classification tasks such as ours.

In particular, the raw annotations in BioS-DB from each annotator were originally sampled at 2Hz and their range was [-1000, 1000]. Therefore for our purposes, we downsample the signals to 1Hz to obtain one label per second, which will be our baseline working frequency for future data fusion schemes. To compute a combined final label for each second, we chose the two annotators that had labelled closer in the 2D space[2], and based on the sign of the Arousal and Valence values, we convert these into a categorical label in each of the four quadrants. If the quadrant where the two labels considered lay coincides, it is chosen as the aggregated label value, otherwise, we assign an undetermined value, x.

Then, we analyze several cases for the undetermined labels: if x is due to a transition between quadrants (one annotator has crossed the boundary but the other has not yet), we randomly choose any of the two quadrants. Otherwise, we consider whether two annotators fall into the same quadrant even though they are not the closest in the 2D space. If so, the aggregated label is the corresponding to that quadrant. This process solves a great amount of undetermined labels. For the rest and those cases where we found several x in a row, we used a 5-second window and replaced the unknown labels with majority voting.

Our process takes into account the proximity of the labels of the raters, which provides confidence about the resulting label since the annotators interpret the 2D space in terms of the quadrants meaning. Transitions between quadrants are considered carefully since people do not leap from one emotional state to another suddenly. The smoothing window provides a smooth label signal by avoiding sharp changes between quadrants.

Finally, for our task of automatic detection of gender-based violence situations, the second quadrant $Q2$ where emotions related to stress, anxiety and fear rely, will be chosen as target. Thus, for the baseline experimentation we considered two types of labellings: quadrants and binary (considering $Q1$, $Q3$, $Q4$ as the negative label, and $Q2$ as the positive).

	Q1	Q2	Q3	Q4
Original BioS-DB	29.22	22.56	8.53	39.67
Reinterpreted BioS-DB	22.16	39.04	8.56	30.24

Table 1: *Percentage of labels in each quadrant*

3.2. BioSpeech+

As stated in previous sections, the ultimate goal of Bindi is to provide an autonomous and inconspicuous tool to detect Gender-based Violence. Regarding speech and audio, we aim at tracking and identifying the user's voice [1] and then use it to detect fear or panic. To improve the precision of the system, this is contextualized by the analysis of the acoustic scene (background sounds and noises) by using a Sound Event Detection and Classification (AED/C) system.

BioS-DB is being used as a proxy to our problem. However, for our specific purposes it is key to complement the spoken information with knowledge about the events present in the acoustic scene: in many cases, panic could cause a GV victim to remain in silence. That is why environmental sounds, that is, the characterization of the acoustic scene, may provide useful information for the detection system.

Therefore we introduce a preliminary procedure to extend the BioSpeech database, consisting of the original speech audio

[2]Preliminary analysis considered selecting labels in terms of 1) proximity or 2) quadrant concordance but experimental results proved that the first approach was more consistent and stable in time.

```
for each lang {'de' or 'en'} do
    for file in lang_foreground_path do
        compute file duration;
        define Scaper object {sample rate = 16 kHz, n_channels =
          1, set ref_db (loudness level)};
        reset previous event specifications;
        groupby: sequential Q2 labels (binary) from correspondent
          BioS-DB.csv file;
        for each Q2 group do
            define event_duration and start_time from Q2 labels;
            if binary_label == 1 then
                add background event fixing {event_duration,
                  start_time};
            end
        end
        add foreground event fixing {file};
        synthesize defined mix;
    end
end
```

Algorithm 1: Procedure for mixing BioSpeech and Audioset samples with Scaper

files synthetically enriched with environmental sounds. The process is an initial approach open for discussion.

We make use of AudioSet [7], a large-scale collection of human-labeled 10-second sound clips drawn from YouTube videos. Audioset provides 2,084,320 samples containing 527 weak annotations at clip level of sound events. We have selected a subset of 2108 samples from Audioset, belonging to 83 classes, to extend the original BioS-DB. To choose classes related to violent events, we have employed the audiovisual stimuli selected in the early stages of the UC3M4Safety dataset collection (in progress). The initial selection was made by experts in VG and later on validated by more than 1300 volunteers [13]. To identify the acoustic events present in the audiovisual stimuli we have employed a pre-trained sound event classification model: YAMNet [16].

YAMNet is a Convolutional Neural Network (CNN) pre-trained on 521 classes of AudioSet, ready to perform inference over audio files to classify occurring sound-events. At the preprocessing stage, the audio signal is normalized, and converted to a 16 kHz mono. Then a log-mel spectrogram of 64 bins is computed to extract a time-frequency representation of the audio signal as an image. These features are fed in patches of 0.96s to the network. The inference stage provides the final score averaged over all the input frames, time-dependent output scores of each class for every 960 ms of raw audio data. It also allows extracting 1024-dimensional embeddings corresponding to the activations of the top convolutional layer.

Regarding the synthetic mixing, the process is based on the data-augmentation pipeline followed in Task 4 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge [17]. Scaper [18] allows us to define probability distributions for the occurrence and duration of the sound events. Thus, the system generates as many synthetic mixes as desired from audio previously classified as foreground or background. In our particular case, foreground events are the original BioS-DB samples and background events are the samples of the Audioset subset. The number of generated mixes has been set to 110: we generate one mix per BioS-DB file, considering recordings captured by the lavalier microphone, i.e. 55 German and 55 English-speaking *audio1* recordings.

The details about the mixing procedure, taking into account the new binarized labels explained in Section 3.1, is presented in pseudocode format in Algorithm 1.

The rationale for this is to provide a non-deterministic relationship between stressing sounds and the appearance of stress in the speaker. In addition to managing probability

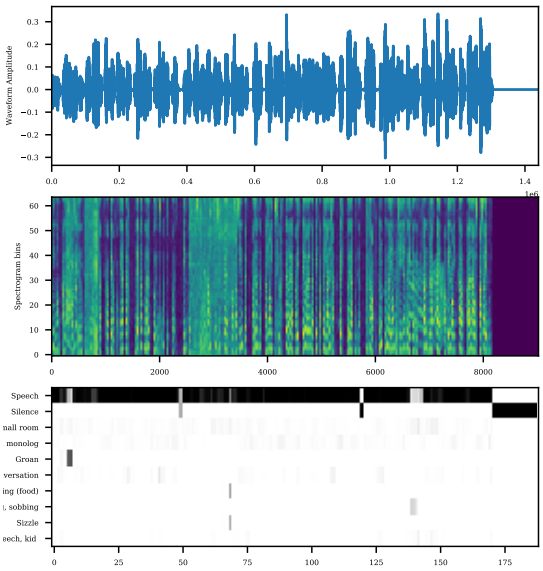


Figure 1: *YAMNet output of a sample of BioSpeech+ with the temporal representation (top), spectrogram (middle, bands spanning 125 to 7500 Hz) and top events found (bottom)*

distributions and timing of the events, Scaper allows pitch shifting and time stretching operations over foreground samples, that could be used for further augmenting the dataset.

4. Preliminary Experiments and Results

To find out if we could use AED/C of the background events to assist the Speaker (SR) and Emotion Recognition (ER) tasks, we have combined them at different SNR[3] (-5, 5 and 15 dB).

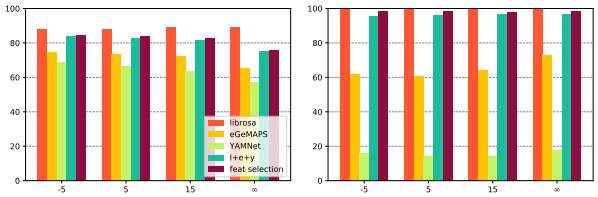
We extracted features from well-known libraries used for SR, ER and AED/C, respectively: librosa [19], eGeMAPS [20] from the openSMILE toolkit [21] and YAMNet embeddings [16]. The size of our working window is one second. This is a compromise between computational complexity and speed and a requirement in Bindi. Thus, from librosa we extracted 19 features with a window size of 20ms and a 10ms overlap and then their mean and standard deviations every second resulting in 38 features per second. Using openSMILE we extracted the eGeMAPS feature set with 88 features. For extracting features suitable for audio events we used the 1024-dimensional embeddings corresponding to the activations of the top convolutional layer of YAMNet. A feature selection method where the correlation of the concatenation of the three feature sets was used to remove the features with a correlation higher than 95%. This resulted in a reduction of the 68% of the features. Examining the correlation matrices we confirmed that most YAMNet features were highly correlated with each other. All features were standardized by using z-score normalization.

With the chosen window size, BioS-DB contains approximately 5000 samples. This is a small size for the use of deep neural networks, so a simple Multi-Layer Perceptron (MLP) implemented with scikit-learn [22] and two shallow architectures implemented with Keras [23] were tested, working towards maintaining a low computational complexity. The first of them consists of two hidden fully-connected layers. The second is a combination of a convolutional 1D layer, a bidirectional GRU layer and a fully-connected layer. This model responds to the idea that it is important to extract

[3]For the SNR measure we consider the foreground speech from BioS-DB as the 'signal´ and the violent audio events as 'noise´.

Model	LIBROSA	*p*	EGEMAPS	*p*	YAMNET	*p*	L+E+Y	*p*	FEAT SEL	*p*
EMOTIONS RECOGNITION BINARY										
MLP	89.1±0.9	*12k*	65.4±1.8	*27k*	57.2±1.4	*307k*	75.3±1.7	*345k*	75.8±1.3	*111k*
K2D	82.4±1.0	*3k*	54.2±0.8	*5k*	32.7±9.0	*52k*	66.3±1.4	*58k*	65.1±1.2	*19k*
KCGD	80.9±1.8	*9k*	54.3±2.7	*12k*	30.4±5.6	*72k*	66.7±1.3	*80k*	67.2±1.3	*30k*
EMOTIONS RECOGNITION 4-Q										
MLP	90.0±0.9	*12k*	45.5±1.1	*27k*	35.8±1.7	*307k*	59.5±1.0	*346k*	60.4±1.6	*112k*
K2D	73.2±1.0	*3k*	47.7±2.0	*6k*	37.6±1.0	*52k*	56.8±1.0	*59k*	57.8±1.2	*19k*
KCGD	73.2±0.9	*9k*	47.9±1.0	*12k*	37.6±0.9	*72k*	58.7±1.2	*80k*	56.9±1.7	*30k*
SPEAKER RECOGNITION										
MLP	100±0	*28k*	72.7±0.6	*43k*	17.8±1.4	*324k*	96.4±1.0	*361k*	98.35±0.3	*128k*
K2D	99.9±0.1	*4k*	64.3±2.0	*7k*	15.21±1.4	*53k*	95.9±0.8	*60k*	96.6±0.7	*20k*
KCGD	100±0	*10k*	50.9±0.7	*13k*	12.6±1.9	*73k*	90.8±1.3	*81k*	95.7±0.9	*31k*

Table 2: *F1-score results for clean BioS-DB. MLP refers to the Multi-Layer Perceptron, K2D refers to the 2-dense layers model in Keras and KCGD refers to the Keras model composed of a Convolutional 1D, Bidirectional GRU and Dense layers. Mean and standard deviation results are shown for a 5-fold validation.*



(a) Emotion Recognition *(b) Speaker Recognition*

Figure 2: *F1-score results with Multi-Layer Perceptron*

information from the temporal context distribution of the features data. The models were compiled using Adam with a learning rate of 0.001, categorical cross-entropy as the loss function and f1-score as the metric to evaluate performance due to the imbalance of the dataset. For all experiments we used a 5-fold cross-validation.

The results for BioS-DB without the audio events are shown in Table 2, where p represents the number of parameters of each model. For the three tasks under consideration, MLP with librosa achieves the best performance. It is worth noting that librosa features achieve the maximum score for the SR task.

The differences in performance between features can be due to multiple reasons: their nature –librosa and eGeMAPS features are manually extracted whereas YAMNet's are automatically extracted from a pretrained sound-event detection network–, their number –38, 88 and 1024 respectively–, and their specific potential to represent emotions or speaker information. Further examination is certainly needed.

Figs. 2 (a) and (b) provide the results for ER and SR respectively for different SNRs. Specifically, Fig. 2 (a) shows the results for binary ER for the model that performed the best (MLP). All the feature sets –except maybe librosa, which remains stable– show a trend to improve the f1-score as the SNR value gets lower, that is, when the acoustic events overlay the speech. This demonstrates that extending our database with stressful events comes in handy for the recognition of stress in speech. All the feature sets, in a greater or lesser extent, are able to capture information about the acoustic events which are considered stress triggers.

As for Fig. 2 (b) we can observe an almost perfect performance for librosa features, and a considerable decrease in efficiency for YAMNet embeddings. However, the performance decreases with lower SNRs contrary with what we observed in Figure 2 (a). This means that acoustic events do not facilitate the SR task. Besides, YAMNet embeddings do not seem to capture relevant information about the acoustic cues of the speech that could help distinguish between speakers.

A for the AED/C task, we illustrate an example in Fig. 1 of the performance of YAMNet classifying a 90 s mixed audio. The dataset has been pre-processed to match YAMNet's requirements ($f_s = 16KHz$, mono, amplitude normalized to $[-1, 1]$) and then fed into the model. The only free parameter is `patch_hop`, which was set to $0.48s$. The audio corresponds to a women reading a text in German. When the first Q2 annotations occur, a background event of a man groaning can be heard. Some yelling and a snoring occur right after but the network only captures it by decreasing the confidence on the speech class. The next background event is a wind noise that is misslabeled as 'Frying (food)'. Lastly, a man whining sound is classified as crying/sobbing with low confidence.

5. Dicussion and Future Work

We draw from the premise that detecting violent situations involves taking into account speech and acoustic contexts since they could be correlated. However, there are no non-acted datasets that allow to elicit this relationship. In Section 3.1 we reinterpreted BioS-DB labels. The samples labelled $Q2$ were interpreted as those related to fear, anxiety or stress, but we should note that without the *dominance* dimension, emotions such as anger or rage could lay in that quadrant too. Using those labels we have extended BioS-DB with stressful sound events, as described in Section 3.2.

In this preliminary study we focused on speaker, stress, and acoustic events in the background. Both the feature sets and algorithms were used with the aim of keeping low the computing load and taking into account the number of samples of the database used. Stressful acoustic events with a non-deterministic correlation to stressed speech utterances proved to be beneficial to some extent for the classifications of binary emotional utterances. On the contrary, they were not helpful (eGeMAPS or YAMNet) or irrelevant (librosa) in the recognition of the speaker.

This research leaves many open questions and future lines of work. Since Scaper allows us to define probability distributions for the appearance and duration of the sound events, the procedure defined in Section 3.2, ready to perform the addition of background events when `binary_label` is $Q2$, could be extended by proceeding in a similar way with non-stressful events whenever `binary_label` is not $Q2$, making the resulting mix sound more realistic.

Also background sounds in the mixing process can be adapted to any kind of problem, resulting into new combinations of the BioS-DB and other datasets. As the main goal of Bindi is to detect and prevent Gender-based Violence, these background events could correspond to audio clips of movie scenes representing a GV scenario, selected with expert knowledge and guidance. This way it could be possible to count with a synthetic dataset of Gender-based Violence situations or other different kinds of situations.

6. Acknowledgements

This work has been partially supported by the Dept. of Research and Innovation of Madrid Regional Authority, in the EMPATIA-CM research project (reference Y2018/TCS-5046). We thank NVIDIA for the donation of the TITAN Xp. The authors also thank the Spanish Ministry of Science, Innovation and Universities (MCIU) for the FPU grant FPU19/00448 and the rest of the members of UC3M4Safety for their support.

7. References

[1] E. Rituerto-González, A. Gallardo-Antolín, and C. Peláez-Moreno, "Speaker Recognition under Stress Conditions," in *Proc. IberSPEECH 2018*, 2018, pp. 15–19. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2018-4

[2] E. Rituerto-González, A. Mínguez Sánchez, A. Gallardo-Antolín, and C. Peláez-Moreno, "Data augmentation for speaker identification under stress conditions to combat gender-based violence," *Applied Sciences*, vol. 9, p. 2298, 06 2019.

[3] M. de Igualdad. (2020) Telematic control devices of measures and withdrawal penalties. [Online]. Available: https://violenciagenero.igualdad.gob.es/informacionUtil/recursos/disposi-tivosControlTelematico

[4] J. A. Miranda-Calero, R. Marino, J. M. Lanza-Gutierrez, T. Riesgo, M. Garcia-Valderas, and C. Lopez-Ongil, "Embedded emotion recognition within cyber-physical systems using physiological signals," in *Conf. on Design of Circuits and Integrated sys. (DCIS)*, 2018.

[5] E. Rituerto-González, J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, C. Peláez-Moreno, and C. López-Ongil, "A hybrid data fusion architecture for bindi: A wearable solution to combat gender-based violence," in *Multimedia Coms., Services and Security*. Springer Intl. Publishing, 2020, pp. 223–237.

[6] T. Garner and M. Grimshaw, "A climate of fear: Considerations for designing a virtual acoustic ecology of fear," in *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, ser. AM '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 31–38. [Online]. Available: https://doi.org/10.1145/2095667.2095672

[7] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, *"Audio set: An ontology and human-labeled dataset for audio events"*.

[8] J. Hansen and S. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," vol. 4, 01 1997.

[9] A. Ikeno, V. Varadarajan, S. Patil, and J. H. L. Hansen, "Ut-scope: Speech under lombard effect and cognitive stress," in *2007 IEEE Aerospace Conference*, 2007, pp. 1–7.

[10] A. Aguiar, M. Kaiseler, M. Cunha, J. Silva, M. H., and P. Almeida, "Voce corpus: Ecologically collected speech annotated with physiological and psychological stress assessments." 05 2014.

[11] A. Baird, S. Amiriparian, M. Berschneider, M. Schmitt, and B. Schuller, "Predicting biological signals from speech: Introducing a novel multimodal dataset and results," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–5.

[12] Y. Deng, M. Yang, and R. Zhou, "A new standardized emotional film database for asian culture," *Frontiers in Psychology*, vol. 8, p. 1941, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2017.01941

[13] M. Blanco-Ruiz, C. Sainz-De-Baranda, L. Gutiérrez-Martín, E. Romero-Perales, and C. López-Ongil, "Emotion elicitation under audiovisual stimuli reception: Should artificial intelligence consider the gender perspective?" *International Journal of Environmental Research and Public Health*, vol. 17, pp. 1–22, 11 2020.

[14] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 07 2019.

[15] S. Hantke, E. Marchi, and B. Schuller, "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2156–2161. [Online]. Available: https://www.aclweb.org/anthology/L16-1342

[16] M. Plakal and D. Ellis, "Yamnet," https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, Accessed: 2020-12-30.

[17] N. Turpault, R. Serizel, P. Shah, J. Salamon, and A. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis." [Online]. Available: https://hal.inria.fr/hal-02160855v2

[18] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 344–348.

[19] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

[20] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, 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, pp. 1–1, 01 2015.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile – the munich versatile and fast open-source audio feature extractor," 01 2010, pp. 1459–1462.

[22] 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.

[23] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015, Accessed: 2020-12-30.