Intelligent Systems with Applications 21 (2024) 200305

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

# Disruptive situation detection on public transport through speech emotion recognition

Eleonora Mancini *, Andrea Galassi, Federico Ruggeri, Paolo Torroni

*DISI, University of Bologna, Viale Risorgimento 2, Bologna, 40136, Italy*

ARTICLE INFO

*Keywords:*
Speech emotion recognition
Affective computing
Natural language processing
Machine learning
Data augmentation

ABSTRACT

Disruptive situations are emotionally-charged events diverging from ordinary behavior, like people fighting or screaming. Public transports are one type of social environment where disruptive situation may occur, and their timely detection may bring significant improvements to people's safety. Current approaches to disruptive situation detection, typically based on CCTVs, do not take the emotional dimension into account. Conversely, we propose to frame such a problem as a speech emotion recognition task.
To validate our hypotheses, we carry out an extensive experimental study focusing on the development of a model characterized by speaker/gender independence, robustness to noise, and robustness against multiple voices. We investigate a variety of audio features, classifiers, datasets, and data augmentation methods in an effort to define effective ways to address this under-investigated yet socially significant problem. Our experiments show that the proposed systems attain an F1 score of over 90% on the disruptive class, even when introducing noisy elements such as environmental noise or multiple overlapping voices. This robust performance is achieved with datasets characterized by speaker variability, gender diversity, and varying number of samples. Such promising results indicate that framing disruptive situation detection as a speech emotion recognition task could pave the way to the adoption of new types of intelligent systems with a positive impact on public safety.

## 1. Introduction

Recent developments in Artificial Intelligence (AI) led to the spread of innovative technologies specifically aimed at increasing safety. Application domains include domestic violence (Roa et al., 2018), health (Beltrán et al., 2019), autonomous vehicles (Wang et al., 2020) and public transports (Laffitte et al., 2016). These technologies brought quality-of-life improvements in private and public environments. In this context, we aim to devise a system for *disruptive situation detection* on public transport based on speech.[1] Public transport is essential, especially in urban areas where it is relied upon by millions of people daily. Unfortunately, it is not uncommon for disruptive situations, such as verbal altercations or physical assaults, to occur on public transport. These events create discomfort among passengers and, more significantly, pose a safety risk. Therefore, it is crucial to develop methods for detecting and preventing them. This problem, with a strong social impact, has not yet attracted the attention it deserves. Moreover, those who addressed it so far mostly framed it as the detection of scream and shouted speech (Laffitte et al., 2016). However, disruptive behavior is not necessarily associated with screaming or shouting. On the contrary, relevant scenarios may include passengers threatened quietly or paralyzed by fear.

The problem is especially challenging because the typical public transport environments are noisy and crowded with people of different genders and nationalities. Hence, issues to be addressed include multilingualism, robustness to environmental noise, presence of multiple overlapping voices, and actor- and gender-independence.[2] Finally, the lack of real-world data makes training of machine learning models even more challenging and hampers a complete and fair validation of any system. Given the many variables, we restrict the scope of our analysis

* Corresponding author.
*E-mail addresses:* e.mancini@unibo.it (E. Mancini), a.galassi@unibo.it (A. Galassi), federico.ruggeri6@unibo.it (F. Ruggeri), p.torroni@unibo.it (P. Torroni).

[1] This was a case study of the 5GMED EU project, aiming to leverage AI solutions for future railway systems, for example, in the detection of any tense situation that may happen inside the train, in combination with more traditional CCTV-based surveillance systems. More information on the project can be found at https://i2cat.net/projects/5gmed/. See Mancini (2021) for preliminary results.

[2] For the sake of consistency with the literature, we will use the term *gender* for the whole paper. However, we believe that the term *sex* would be more appropriate in this case.

https://doi.org/10.1016/j.iswa.2023.200305
Received 5 July 2023; Received in revised form 20 October 2023; Accepted 12 November 2023

Available online 17 November 2023

to robustness to noise, robustness against multiple voices, and actor-/gender-independence, leaving multilingualism to future investigation. Following a practice adopted in psychological science of categorizing emotions into *positive* and *negative* (An et al., 2017, Pekrun, 2006, Posner et al., 2005), in agreement with domain experts, we hypothesize that disruptive situations are characterized by the prevalence of a well-identified spectrum of emotions, such as *anger*, *sadness* and *fear*. Under this assumption, the detection of *disruptive situations* becomes akin to a Speech Emotion Recognition (SER) task.

Our starting point was an extensive analysis of the datasets for SER. It turned out that, with few exceptions (Oflazoglu & Yildirim, 2013, Sultana et al., 2021, Wang et al., 2014, Zhalehpour et al., 2017), almost all such datasets are in English. We thus decided to conduct our experiments in English. We selected four popular datasets, assigning samples to training, validation, and test splits so as to guarantee actor- and gender-independence. Furthermore, we designed a simple but effective data-augmentation strategy to deal with the limited number of training examples and the presence of environmental noise in the domain of interest. Our experimental analysis includes the evaluation of several audio features, classifiers, and datasets. Furthermore, we test our models for robustness against noise and in the presence of multiple overlapping voices. These aspects hold particular significance as the existing research has mainly focused on a specific kind of disruption that leads to screams (Laffitte et al., 2016). In contrast, real-world disruption scenarios are notably multifaceted. Therefore, investigations aimed at developing intelligent systems applicable within broader and less constrained contexts constitute a novel and valuable contribution. Additionally, we test the models in an ensemble setting. Our results suggest that the current technology seems adequate for the introduction of disruptive situation detection systems with the potential to benefit important social environments.

The code, datasets, and a working stand-alone prototype application developed for this study are publicly available.[3]

To summarize, our contributions are:

- a novel way to frame disruptive situation detection as an SER problem;
- a critical analysis of existing datasets for this task;
- an extensive empirical study designed to validate our hypotheses under relevant dimensions, such as speaker and gender variance, and presence of noise and multiple overlapping voices on four different datasets;
- a set of classification models and methods for data augmentation.

The structure of the paper is the following. We survey related work in Section 2. In Sections 3 and 4, we describe the datasets and the methodology. We discuss our experimental results in Section 5, and conclude in Section 6.

## 2. Related work

This section reviews related methods, architectures, and data.

**Disruptive Situation Detection.**

The problem of detecting disruptive situations on public transport from audio is scarcely addressed in the literature. One notable effort is presented by Laffitte et al. (2016), who address the problem as the detection of scream and shouted speech. However, we are the first ones to frame it as an SER task.

**Problem Formulation.**

An SER system is often depicted as comprising two core phases (Mustaqeem & Kwon, 2021). The first phase concerns the selection of robust, discriminative, and salient audio features, such as Mel-Frequency-Ceptrum coefficients (MFCCs), Chroma, Zero-Crossing-Rate (ZRC), Root-Mean-Square-Energy (RMSE) and Log-Mel Spectrogram. The second phase regards the definition of adequate classification methods. One of the trends in current literature employs classic methods like Support Vector Machines (SVMs) (Aouani & Ayed, 2020, Iqbal, 2021), Convolutional Neural Networks (CNNs), Long-Short Term Memory Networks (LSTMs) and Hidden Markov Models (HMMs) (An & Ruan, 2021, Chourasia et al., 2021, Fu et al., 2020, de Pinto et al., 2020, Venkataramanan & Rajamohan, 2019).

Current approaches to SER follow two problem formulations: *categorical* (Ekman, 1992, 1989), and *dimensional* (Posner et al., 2005). In the first, emotions define discrete classes for multi-class or multi-label classification. Generally, such discrete classes refer to Ekman's six basic emotions (Ekman, 1999): sadness, happiness, fear, anger, disgust, and surprise (Akçay & Oguz, 2020). Conversely, in the *dimensional* task formulation, emotions are defined as small numerical values over distinct emotion latent dimensions (e.g., valence and arousal) (Akçay & Oguz, 2020, Yang & Chen, 2012). The dimensional model accurately captures certain complex emotional states. However, it fails to discriminate between emotions such as *fear* and *anger* and makes it hard to characterize an emotion such as *surprise*, which may have a positive or negative valence depending on the circumstances (Akçay & Oguz, 2020). Since these emotions are key to the correct classification of disruptive situations, for the purposes of this study, we frame SER as a binary classification task, following a categorical approach.

**Architectures.**

Among other works that follow a categorical approach, one primary line of research is based on CNNs to extract spatial features. Notable examples are (de Pinto et al., 2020) and (Chourasia et al., 2021), who propose a 1D-CNN classifier with MFCCs audio features achieving good accuracy performance. Fu et al. (2020) address it with end-to-end training of an attention-based CNN-BLSTM model. Meng et al. (2019), as well as Zhao et al. (2019) integrate CNNs and LSTM networks to propose a novel architecture for emotion detection. Pandey et al. (2022) introduce a deep neural network that combines convolutional layers and LSTM. Meanwhile, Shahin et al. (2022) undertake a comprehensive evaluation of their CNN and LSTM-based model against conventional classifiers across various corpora, including RAVDESS and CREMA-D. They show their model's state-of-the-art performance across all datasets. Furthermore, Nagase et al. (2022) employ a deep neural network featuring convolutional and LSTM layers for emotion recognition. Notably, they propose the application of label smoothing as a technique to mitigate overfitting stemming from mislabeled information. Furthermore, An and Ruan (2021) use two parallel CNNs for spatial features and a transformer encoder network to extract temporal features. They also design a data augmentation method using Additive White Gaussian Noise (AWGN). In contrast, we seek the integration of domain-specific environmental noise. Andayani et al. (2022) recently propose to integrate LSTM and transformer architectures in order to capture long-term dependencies in speech signals, yielding superior accuracy. Mocanu and Tapu (2022) formulate a 2D CNN coupled with deep metric learning for emotion recognition. Their model demonstrates notable efficacy on the RAVDESS and CREMA-D datasets. These studies collectively contribute to advancing emotion detection through the fusion of convolutional and LSTM architectures and other techniques.

Another line of research primarily adopts SVM classifiers. Aouani and Ayed (2020) adopt several audio features and an autoencoder to extract more advanced features to be fed into the SVM classifier. Furthermore, Iqbal (2021) tackles SER using SVM and MFCCs on the TESS dataset, reaching satisfactory results.

In our study, we opt to start with simple CNNs and SVMs classifiers. This decision is driven by our desire to establish a robust proof of concept for our hypotheses, leaving the exploration of more advanced methods for future investigations. Furthermore, our commitment to lightweight models harmonizes with the practical integration of the deployed models into edge devices, necessitating the adoption of shallow architectures.

[3] https://github.com/helemanc/ambient-intelligence.

**Audio Features.**

Other approaches have been recently proposed for feature extraction. Patel et al. (2022) study the impact of an autoencoder architecture to extract high-level features for SER, achieving good performance on the RAVDESS and TESS datasets. Chattopadhyay et al. (2020) use Linear Predictive Coding (LPC) in conjunction with MFCCs. Additionally, they propose a novel application of Manta Ray optimization that achieves state-of-the-art performance on SAVEE and EMO-DB.

Venkataramanan and Rajamohan (2019) carry out an extensive comparison of various approaches for SER. In particular, they study different audio features, such as Log-Mel Spectrogram and MFCCs, in combination with several neural architectures like LSTMs, CNNs, and HMMs. They conclude that the choice of audio features has more impact on model performance than model complexity. Chen et al. (2023) incorporate a connection attention mechanism to effectively integrate frame-level manual features, utterance-level deep features. Liu et al. (2023) improve the imbalance of the sample distribution among emotional categories and increase feature diversity by employing balanced augmented sampling on triple-channel log-Mel spectrograms, implementing time and frequency-domain filters, and achieving remarkable SER performance on datasets like IEMOCAP and MSP-IMPROV. Moreover, Singh et al. (2023) contribute by introducing constant-Q transform-based modulation spectral features (CQT-MSF), offering emotion-specific representations that outperform conventional mel-scale spectrograms and modulation features, notably on datasets like Berlin EmoDB and RAVDESS. In our study, we opt for the use of MFCCs as our choice of audio representation, primarily to ascertain if this compact spectral feature representation can deliver good performance, reserving the exploration of more sophisticated audio representation techniques for future investigations.

**Multilingualism.**

Currently, there is a growing emphasis on the exploration of methods and techniques for dealing with multilingualism. To address the multilingual aspect of emotion recognition, Sultana et al. (2022) introduce a system based on CNN and LSTM networks. They perform transfer learning between the RAVDESS dataset and a second one in the Bangla language. Gerczuk et al. (2023) employ a transfer learning approach with a diverse multi-corpus database encompassing 26 freely available corpora. This corpus, called EmoSet, encompasses 84,181 multi-lingual audio recordings with a combined duration exceeding 65 hours. Their approach, leveraging various convolutional neural network architectures and spectrograms derived from original audio recordings, showcases promise in overcoming the challenges of multilingualism.

As emphasized in Section 1, our current analysis is centered on robustness against noise, multiple voices, and actor/gender independence, with multilingualism investigations reserved for future research. Therefore, we intend to explore the aforementioned techniques in the future to address challenges related to the presence of multilingualism in public transport environments.

**Data.**

Besides methods and architectures, another point that deserves attention is the data and how it has been used. Several works randomly split data among train, validation, and test splits. Random splitting may lead to biased situations in which an actor is shared among different splits. Such a phenomenon can alter experimental scenarios, yielding optimistic results that do not truly evaluate properties like actor independence. Moreover, as a side effect, an unbalanced gender distribution concerning actors can further bias an SER classifier. Previous work tackled this problem only partially by using a leave-one-speaker-out validation (Bitouk et al., 2010, Cao et al., 2015, Fu et al., 2020, Sato & Obuchi, 2007) or by dividing the data by gender (Zhu et al., 2017). Huang et al. (2016) suggest a feature normalization method for speaker-independent SER. However, in our opinion, the evidence for the generality of the latter approach is still limited. To guarantee actor independence and a rigorous evaluation, we split our data following Vogt and André (2006) and Venkataramanan and Rajamohan (2019) (see Section 4.1.3 for details).

**Table 1**
Emotions for each selected dataset concerning SER.

| | RAVDESS | TESS | SAVEE | CREMA-D |
|---|---|---|---|---|
| **fear** | ✓ | ✓ | ✓ | ✓ |
| **disgust** | ✓ | ✓ | ✓ | ✓ |
| **neutral** | ✓ | ✓ | ✓ | × |
| **calm** | ✓ | × | × | × |
| **happiness** | ✓ | ✓ | ✓ | ✓ |
| **sadness** | ✓ | ✓ | ✓ | ✓ |
| **surprise** | ✓ | ✓ | ✓ | × |
| **angry** | ✓ | ✓ | ✓ | ✓ |

## 3. Data

This section provides details on the data employed in our experiments.

Several speech datasets have been created in a wide variety of languages for developing emotional systems that work on audio (Swain et al., 2018). These datasets are generally divided into three categories: *acted* (or *simulated*), *invoked* (or *elicited*) and *spontaneous* (Akçay & Oguz, 2020).

Motivated by the lack of SER datasets in the public transport domains, we focus on acted speech datasets where actors are native English speakers. In particular, we select the following datasets: RAVDESS (Livingstone & Russo, 2018), TESS (Pichora-Fuller & Dupuis, 2020), SAVEE (Haq & Jackson, 2010) and *high-intensity* data part of CREMA-D (Cao et al., 2014).[4] We choose these datasets because they are standardized collections of emotions, allowing a straightforward comparison of classification results (Abbaschian et al., 2021). Moreover, unlike *invoked* or *spontaneous* datasets, *acted* datasets allow easier modeling and detection of emotions since (i) each audio file is characterized by a specific emotion[5]; (ii) the amount of different emotions found in these corpora is higher than in *invoked* and in *spontaneous*; (iii) recordings are not significantly altered by environmental noise. The absence of environmental noise also eliminates the need for a denoising pre-processing phase which might cause information loss. Nevertheless, we are aware that the selected speech datasets could not represent a realistic domain-specific scenario. For instance, as *simulated* datasets have synthesized emotions, models tend to overfit around emotions differently from day-to-day conversations (Abbaschian et al., 2021).

One of our primary objectives is to implement a classification model that is not biased towards one of the two genders. Therefore, we pay particular attention to gender distribution. We observe that RAVDESS and CREMA-D present a balanced number of female and male actors. In particular, RAVDESS has 12 female and 12 male actors. Similarly, CREMA-D has 48 male actors and 43 female actors. In contrast, TESS contains recordings from male actors only, whereas SAVEE has recordings from female actors only.

Not all datasets contain the same emotion categories (see Table 1). The average recording duration ranges from two to five seconds. Other aspects, such as the sample rate, the dB amplitude, and the RMSE are specific to each dataset. For instance, RAVDESS and CREMA-D have substantial differences in audio quality, since recordings in CREMA-D present more echoes and volume variations.

Concerning the number and variety of samples, RAVDESS is one of the datasets with the largest amount of recordings when considering

[4] For CREMA-D, we consider only the *high* intensity data since we deemed the quality of the other parts not sufficient for our purposes.

[5] The emotions do not vary dynamically during the course of speech and there is no concurrence of different emotions.

**Table 2**
Distribution of emotions, genders, and actors for each dataset for SER. We differentiate the number of samples in training, validation, and test splits with the / symbol.

| | RAVDESS (R) | TESS (T) | SAVEE (S) | CREMA (C) |
|---|---|---|---|---|
| disruptive D | 640 / 64 / 64 | 1000 / — / 800 | 120 / 60 / 60 | 256 / 48 / 48 |
| disgust | 160 / 16 / 16 | 200 / — / 200 | 30 / 15 / 15 | 64 / 12 / 12 |
| sadness | 160 / 16 / 16 | 200 / — / 200 | 30 / 15 / 15 | 64 / 12 / 12 |
| angry | 160 / 16 / 16 | 400 / — / 200 | 30 / 15 / 15 | 64 / 12 / 12 |
| fear | 160 / 16 / 16 | 200 / — / 200 | 30 / 15 / 15 | 64 / 12 / 12 |
| non-disruptive ND | 560 / 56 / 56 | 400 / — / 600 | 120 / 60 / 60 | 64 / 12 / 12 |
| happy | 160 / 16 / 16 | — / — / 200 | 30 / 15 / 15 | 64 / 12 / 12 |
| neutral | 240 / 24 / 24 | 200 / — / 200 | 60 / 30 / 30 | — / — / — |
| surprise | 160 / 16 / 16 | 200 / — / 200 | 30 / 15 / 15 | — / — / — |
| Total samples | 1200 / 120 / 120 | 1400 / — / 1400 | 240 / 120 / 120 | 320 / 60 / 60 |
| Female actresses | 10 / 1 / 1 | 1 / — / 1 | — / — / — | 32 / 6 / 6 |
| Male actors | 10 / 1 / 1 | — / — / — | 2 / 2 / 2 | 32 / 6 / 6 |
| Total actors | 20 / 2 / 2 | 1 / — / 1 | 2 / 2 / 2 | 64 / 12 / 12 |

multiple actors (see Table 2). Additionally, RAVDESS contains an equal and adequate number of actors for the definition of gender- and actor-independent classification models. Lastly, recordings in RAVDESS are not particularly subject to environmental noise that might downgrade audio quality. For these reasons, we consider RAVDESS as the reference dataset of our experimental setting.

## 4. Method

This section provides an in-depth exploration of our research methodology, encompassing pre-processing, classifiers, and the experimental framework. Within *Pre-Processing* (Section 4.1), we delve into data loading and equalization (Section 4.1.1), labels aggregation (Section 4.1.2), data splitting (Section 4.1.3), data augmentation (Section 4.1.4), cut and pad techniques (Section 4.1.5), as well as feature extraction (Section 4.1.6). In *Classifiers* (Section 4.2) and *Experimental Setting* (Section 4.3), we describe in detail architectures, experiments and model calibration.

### *4.1. Preprocessing*

#### *4.1.1. Data loading and equalization*

We load and resample all audio files through `librosa` (McFee et al., 2015) using a sample rate of 16 kHz, which is the minimum sample rate of the original sources.

#### *4.1.2. Labels aggregation*

We frame the problem as the binary classification of *disruptive* (D) and *non-disruptive* (ND) emotions. We consider disruptive situations in the public transport domain and make assumptions about which emotions (among those available in the dataset) characterize them. Therefore, we split the emotions as follows:

- *Disruptive Emotions* (D): *anger*, *sadness*, *fear*, *disgust*;
- *Non-Disruptive Emotions* (ND): *neutral*, *happiness*, *surprise*, *calm*.

As can be seen in Table 2, CREMA-D turns out to be unbalanced after aggregating labels. This imbalance towards the D-class is caused by the fact that CREMA-D contains only *happiness* among all the emotions belonging to the ND class.

#### *4.1.3. Data splitting*

Existing work on SER (Padi et al., 2020, Patel et al., 2022, de Pinto et al., 2020) defines train, validation, and test sets via random splitting. Nonetheless, such an approach leads to a biased model evaluation (Venkataramanan & Rajamohan, 2019) since the same actor can appear in multiple splits. To better create and evaluate a truly *actor-* and *gender-independent* model, we define dataset splits according to the following criteria: (i) splits should have a balanced amount of male and female actors; (ii) each actor can only appear in one split.[6] The final distribution of labels, actors, and genders across the three splits is reported in Table 2. As can be noticed in Table 2, TESS has two splits since it contains two actors only.

#### *4.1.4. Data augmentation*

Additive noise interference is a significant obstacle to the practical use of SER systems (Tiwari et al., 2020, Zhang et al., 2018). To overcome these issues, there exist three main approaches: at the signal level (e.g., using a denoising module or a voice activity detector), at the feature level (e.g., enhancement through Wiener filtering), at the model level (e.g., training a model on noise-corrupted data). We rely on the latter and propose a data augmentation strategy to simultaneously deal with the limited number of training samples and the presence of typical environmental noises associated with the public transport domain. Since the introduced datasets for SER are defined in a *noiseless* environment, we create new synthetic training data by injecting domain-specific noise. In particular, noise-corrupted data is created based on three samples[7]: (i) a recording of the inside of a train; (ii) a recording of a freight train with squeaky wheels passing by; (iii) a recording of a small crowd of children playing. We create a noise-corrupted version of a recording by overlaying a randomly chosen noise sample on a randomly chosen temporal position. The noise's volume is set to 2 dB less than each sample's volume to keep the recorded voice audible. We apply this technique to the training set of each dataset and merge the new samples with the original ones, resulting in six augmented training sets with twice the number of training samples.

Another form of data augmentation is the combination of audio samples to simulate the presence of multiple voices and evaluate our system in such conditions. To simulate this scenario, we select combinations of two, five, and ten audio files from each test set of each dataset, overlaying them for each of the two categories being analyzed (disruptive and non-disruptive). We overlay the samples belonging to each combination. The resulting distribution of samples is reported in Table 7. We make sure that each sample is contained in only one combination.

#### *4.1.5. Cut and pad*

We trim and pad recordings to make them uniform. We fix the audio duration of recordings to 5 seconds. Shorter recordings are padded us-

[6] It has been recently observed that gender imbalance in data could lead to decreased ASR performance on the least represented gender category (Garnerin et al., 2021).

[7] The audio files were retrieved from the Youtube and Soundible platforms.

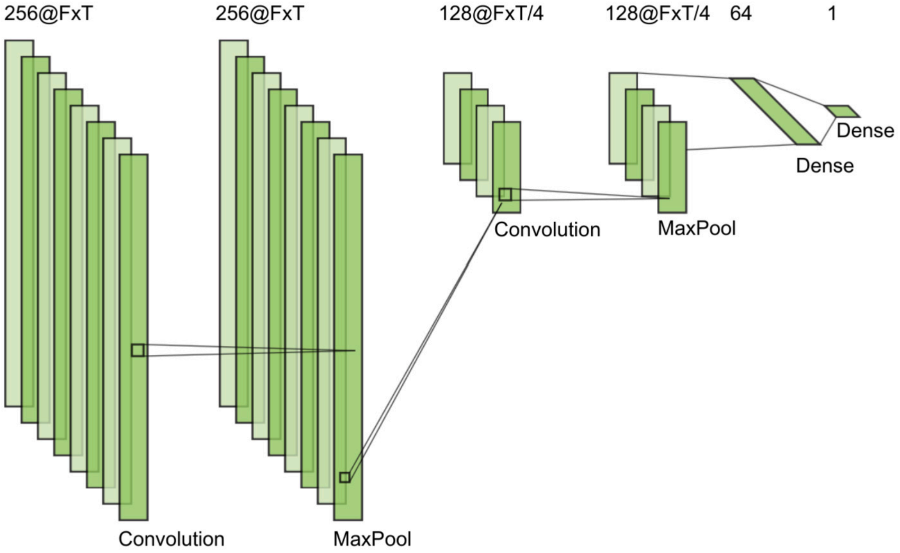

**Fig. 1.** CNN architecture. The labels in the figure follow the format of *n_filters@FxT*, where *n_filters* represents the number of filters applied, $F$ denotes the number of features, and $T$ is the number of timesteps.

ing the audio file's median value, while longer ones are truncated. The sampling rate for recordings is set to 16 kHz. Thus, each recording is represented by 80000 samples.

#### 4.1.6. Features

We extract Mel-Frequency-Cepstral Coefficients (MFCCs) (Guðnason & Brookes, 2008) and Root-Mean-Square-Energy (RMSE) (Er, 2020) for each frame to represent recordings. The number of frames is obtained by dividing the number of samples by the hop length. We set the hop length to 512,[8] resulting in 157 $t$ time-steps per recording. MFCCs are then represented as a $t \times c$ matrix, where $c$ is the number of coefficients. Conversely, RMSE is a $t$-dimensional vector.

Regarding MFCCs, we experiment with the first 13 and 26 $c$ coefficients. In both cases, we remove the first MFCC component since it carries little information (Fahmy, 2010), thus obtaining 12 and 25 MFCCs. RMSE is concatenated to MFCCs when considered as an additional feature. We normalize and standardize the MFCCs and RMSE of each input example at the dataset level to facilitate learning.[9]

Lastly, some machine learning models (e.g., SVM) require mono-dimensional inputs. In this case, we reduce MFCCs by averaging over the $t$ dimension.

### 4.2. Classifiers

Following previous work on SER, we employ SVM[10](Cortes & Vapnik, 1995) and CNN (Goodfellow et al., 2016) as classifiers. These are light-weight classifiers with square or linear computational complexity (Bottou et al., 2007, Vaswani et al., 2017), routinely used in edge devices. Our CNN consists of three blocks. The first two blocks comprise of a 1-D convolutional layer with 5×5 kernel and ReLU activation function, followed by 1-D pooling with $4 \times 4$ kernel and dropout layers. The dropout rate is set to 0.6 for the first block and 0.5 for the second one. The final block comprises two dense layers for classification. In Figure 1, we present an overview of our CNN structure. We train each CNN model with binary cross-entropy loss function and Adam (Kingma & Ba, 2015) optimizer.

### 4.3. Experimental setting

We devise four experiments to assess the effectiveness of audio features, machine learning models, and data augmentation techniques for SER:

*a)* **Disruptive Situation Detection.** We train and evaluate SVM and CNN models on selected datasets for SER.
*b)* **Gender Independence.** We assess the robustness and potential gender bias of the best-performing models developed for *Disruptive Situation Detection*. In this experiment, the focus is on evaluating the models' performance when tested on datasets containing actors of the same gender.
*c)* **Noise Robustness.** We evaluate the noise robustness of employed classifiers on a noise-corrupted RAVDESS dataset version.
*d)* **Multiple Voices.** We evaluate the effectiveness of our approach for detecting emotional changes in speech with multiple overlapping voices. To construct this setting, we overlay groups of audio files with identical labels (either disruptive or non-disruptive).
*e)* **Ensemble.** We evaluate the learning stability of employed classifiers by considering an ensemble of the best configurations.

#### 4.3.1. Model calibration

We calibrate each model hyper-parameter set via a randomized search. We evaluate models via a three-fold cross-validation routine. For the SVM model, we calibrate the kernel ($RBF$, $Linear$) and the C regularization (0.1, 1.0, 10, 100). Regarding the CNN model, we calibrate the weight initializer (*uniform*, *LeCun*, *Glorot*, *He normal*, *He uniform*), the batch size $(4, 8, 16)$ and the optimizer learning rate $(10^{-3}, 10^{-4}, 5 \cdot 10^{-5})$. All the remaining hyper-parameters are set to their default values. We regularize models via early stopping regularization during training. We set early-stopping patience to 45. Models are early stopped on the validation loss, except TESS, since no validation set is defined. In this case, models are early stopped on the training loss.

#### 4.3.2. Disruptive situation detection

Our primary goal is to analyze how the choice of data, audio features, and classification models might affect an SER system. In addition to RAVDESS, TESS, SAVEE, and CREMA-D, we create new datasets by combining these datasets and generating noise-corrupted examples (Section 4.1.4). A summary of the described approach is shown in Table 3. Overall, we consider 4 feature sets, 2 classifiers, and 8 datasets,

[8] We relied on `librosa` (McFee et al., 2015) default's hop length.
[9] All data were processed using the mean and standard deviation values computed on the respective training set.
[10] We relied on the `libsvm` library (Chang & Lin, 2011).

**Table 3**
Summary of features, classifiers, and training sets used for disruptive situation detection.

| Features | Classifiers | Training Sets | |
|---|---|---|---|
| 12 MFCCs | | RAVDESS | Augmented RAVDESS |
| 12 MFCCs + RMSE | CNN | TESS | Augmented TESS |
| 25 MFCCs | SVM | SAVEE | Augmented SAVEE |
| 25 MFCCs + RMSE | | CREMA-D | Augmented CREMA-D |

totaling 64 different combinations. Since this experiment is based on noiseless test sets, its results will represent an overestimation of the performance that could be achieved in the field.

### 4.3.3. Gender independence

The goal of this experiment is to investigate possible gender biases in the classification models. To do so, we evaluated the performance of the best-performing models for *Disruptive Situation Detection* with gender-specific data subsets. If such models performed significantly worse on either the female-only or male-only test sets compared to their performance on the original, mixed-gender test set, that would indicate a possible gender-related bias. To achieve this, we curate two distinct test sets: the *Female-Only Test Set*, consisting exclusively of scenarios featuring female actors, and the *Male-Only Test Set*, containing scenarios exclusively involving male actors. Given the limited availability of datasets encompassing both genders, namely RAVDESS and CREMA-D, we conduct this experiment exclusively on these two datasets.

### 4.3.4. Noise robustness

In the public transport domain, developing SER models that are robust to noise is crucial. To have a more objective evaluation of our use case, we experiment with a noisy test set. In particular, we create a noise-corrupted RAVDESS test set following the procedure described in Section 4.1.4, and use it to evaluate models that were trained on the original RAVDESS dataset.

### 4.3.5. Multiple voices

To test the effectiveness of our approach in detecting emotional changes in multiple voice conditions, we conduct an experiment exploiting the models that obtained the best performance in Section 4.3.2.

### 4.3.6. Ensemble

To improve performance and to evaluate learning stability, we consider a supervised ensemble setting (Akçay & Oguz, 2020, Dong et al., 2020). In particular, we take the three best models of the *Disruptive Situation Detection* experiment and aggregate their predictions via a voting schema. As voting schema, we consider *majority voting* and *averaged probability*. The former chooses the class that the majority of the models have predicted. For the latter, we pick the maximum class probability averaged over the models. Additionally, we limit ensemble voting to confident models only. We experiment by excluding models that output probability scores in the $[0.4, 0.6]$ and $[0.3, 0.7]$ ranges. Applying average probability voting schema to CNNs is straightforward as they output a probability score via the sigmoid activation function. In contrast, we rely on Platt scaling (Platt, 1999) to compute probability scores for SVMs.

## 5. Results and discussion

This section explains the performance metrics (Section 5.1) employed to evaluate our experiments and presents our results, with a specific focus on *Disruptive Situation Detection* (Section 5.2), *Gender Independence* (Section 5.3), *Noise Robustness* (Section 5.4), *Multiple Voices* (Section 5.5), and *Ensemble* (Section 5.6) scenarios.

### 5.1. Performance metrics

The result of the experiments can be expressed as the number of positive instances that are correctly classified (True Positives TP), those that are misclassified as negatives (False Negatives FN), the correctly classified negatives (True Negatives TN), and those that are misclassified as positives (False Positive FP). Accuracy is defined as the percentage of correctly classified instances (Equation (1)). The F1 score on the positive class is instead defined as in Equation (2).

$$A = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$F1 = \frac{2TP}{2TP + FN + FP} \tag{2}$$

We compute the F1 score on the disruptive class to evaluate the performance of individual models. A model with a high F1 score on

**Table 4**
Best three performing CNN and SVM models for each dataset. We report the F1-score for the disruptive class D and the Accuracy (A). Models are sorted on the F1-score in descending order. The *Improvement* columns denote performance improvement with respect to the majority baseline.

| Dataset | Classifier | MFCCs | RMSE | Noise | A | A Improvement | F1 (D) | F1 (D) Improvement |
|---|---|---|---|---|---|---|---|---|
| RAVDESS | CNN | 25 | | ✓ | 0.90 | +0.37 | **0.91** | +0.21 |
| | CNN | 25 | | | 0.89 | +0.36 | 0.90 | +0.20 |
| | CNN | 25 | ✓ | | 0.89 | +0.36 | 0.90 | +0.20 |
| | SVM | 25 | | | 0.73 | +0.20 | 0.75 | +0.05 |
| | SVM | 25 | | ✓ | 0.68 | +0.15 | 0.68 | -0.02 |
| | SVM | 25 | ✓ | ✓ | 0.70 | +0.17 | 0.68 | -0.02 |
| TESS | CNN | 25 | | | 0.68 | +0.11 | **0.77** | +0.04 |
| | CNN | 25 | | ✓ | 0.67 | +0.10 | 0.76 | +0.03 |
| | CNN | 25 | ✓ | | 0.66 | +0.09 | 0.73 | +0.00 |
| | SVM | 12 | | ✓ | 0.57 | +0.00 | 0.73 | +0.00 |
| | SVM | 25 | ✓ | ✓ | 0.57 | +0.00 | 0.73 | +0.00 |
| | SVM | 12 | | | 0.57 | +0.00 | 0.72 | -0.01 |
| SAVEE | CNN | 25 | ✓ | ✓ | 0.52 | +0.02 | **0.67** | +0.00 |
| | SVM | 25 | ✓ | | 0.57 | +0.07 | 0.62 | -0.05 |
| | CNN | 12 | | ✓ | 0.60 | +0.10 | 0.50 | -0.17 |
| | CNN | 25 | | | 0.52 | +0.2 | 0.43 | -0.24 |
| | SVM | 12 | ✓ | | 0.50 | +0.00 | 0.40 | -0.27 |
| | SVM | 12 | | ✓ | 0.55 | +0.05 | 0.22 | -0.45 |
| CREMA-D | CNN | 12 | ✓ | ✓ | 0.87 | +0.07 | **0.92** | +0.03 |
| | CNN | 12 | | ✓ | 0.87 | +0.07 | 0.92 | +0.03 |
| | CNN | 25 | | | 0.85 | +0.05 | 0.91 | +0.02 |
| | SVM | 25 | | ✓ | 0.80 | +0.00 | 0.89 | +0.00 |
| | SVM | 25 | ✓ | ✓ | 0.80 | +0.00 | 0.89 | +0.00 |
| | SVM | 12 | ✓ | ✓ | 0.80 | +0.00 | 0.89 | +0.00 |

**Table 5**
Performance comparison of best three CNN and SVM Models on Disruptive Class (D) F1 Scores, differentiated by gender-specific test sets (Males - *M* and Females - *F*).

| Dataset | Classifier | MFCCs | RMSE | Noise | F1 | F1 (D) | F1 (D) |
|---|---|---|---|---|---|---|---|
| | | | | | (D) | M | F |
| RAVDESS | CNN | 25 | | ✓ | **0.91** | 0.94 | 0.89 |
| | CNN | 25 | | | 0.90 | 0.90 | 0.89 |
| | CNN | 25 | ✓ | | 0.90 | 0.90 | 0.89 |
| | SVM | 25 | | | 0.75 | 0.70 | 0.78 |
| | SVM | 25 | | ✓ | 0.68 | 0.55 | 0.79 |
| | SVM | 25 | ✓ | ✓ | 0.68 | 0.56 | 0.77 |
| CREMA-D | CNN | 12 | ✓ | ✓ | **0.92** | 0.92 | 0.92 |
| | CNN | 12 | | ✓ | 0.92 | 0.94 | 0.89 |
| | CNN | 25 | | | 0.91 | 0.92 | 0.91 |
| | SVM | 25 | | ✓ | 0.89 | 0.89 | 0.89 |
| | SVM | 25 | ✓ | ✓ | 0.89 | 0.89 | 0.89 |
| | SVM | 12 | ✓ | ✓ | 0.89 | 0.89 | 0.89 |

the disruptive class can identify instances of disruptive behavior while minimizing false positives, which reduces the need for costly human intervention. For the sake of completeness, we also report accuracy.

### 5.2. *Disruptive situation detection*

Table 4 shows the three best-performing configurations for each classifier. We observe that CNNs yield the best performance in all four datasets. In particular, CNNs strongly outperform SVMs on the RAVDESS dataset, where they strongly outperform the baseline. On TESS and CREMA-D, their performance is comparable to the SVMs and to the baseline, denoting the challenging setup of these datasets. As for SAVEE, while CNNs achieve comparable outcomes to the baseline, SVMs experience a significant drop in performance. A qualitative analysis confirmed that files within the disruptive class are not easily discernible from those in the non-disruptive category.

Regarding audio features, results do not show a clear winning configuration over all datasets, but rather, specific combinations are preferred for each setting.

Similarly, our results show that data augmentation with noise-corrupted data has almost no impact on model performance.

### 5.3. *Gender independence*

Table 5 shows that our experiment yields remarkably high F1 scores for both the *female-only* and *male-only* subsets in most cases. For some configurations, we observe differences between the two test sets, with CNNs performing slightly better on the male test set and the SVM trained on RAVDESS performing better on the female test set. However, our best-performing models yield consistent results across gender-specific data, indicating robustness and gender-independence.

### 5.4. *Noise robustness*

We compare the SVMs and CNNs that scored best on RAVDESS by testing them on our new synthetic test set and present the results in Table 6. As expected, the CNN model trained with additional noise-corrupted recordings is the best performing one, losing at most 4 F1-score percentage points. In contrast, CNNs trained without data augmentation lose up to 17 F1-score percentage points. Conversely, noise affects SVMs less, losing only a few percentage points. It is important to remark that these noisy test samples have been created through the same procedure that was applied to create noisy training samples. Thus, we have no guarantee that a model trained on such noise-corrupted data will be robust to other types of noise.

**Table 6**
Results of the *Noise Robustness* experiments on the best-performing CNNs and SVMs models that are trained on RAVDESS. The reference test set for these experiments is the noise-corrupted RAVDESS test set. The two Δ columns report the deterioration with respect to the experiments on the original RAVDESS test set in terms of Accuracy (A) and F1-score for the disruptive class. In bold, the best-performing model.

| Classifier | #MFCC | RMSE | Noise | A | Δ-A | F1 | Δ-F1 |
|---|---|---|---|---|---|---|---|
| CNN | 25 | | | 0.74 | -0.15 | 0.73 | -0.17 |
| CNN | 25 | | ✓ | **0.86** | -0.04 | **0.87** | -0.04 |
| CNN | 25 | ✓ | | 0.76 | -0.13 | 0.76 | -0.14 |
| SVM | 25 | | | 0.65 | -0.08 | 0.70 | -0.05 |
| SVM | 25 | | ✓ | 0.68 | -0.00 | 0.67 | -0.01 |
| SVM | 25 | ✓ | ✓ | 0.67 | -0.03 | 0.66 | -0.02 |

**Table 7**
Results of the *Multiple Voices* experiments on the best-performing CNNs and SVMs models that are trained on RAVDESS, TESS, SAVEE, and CREMA-D. The table includes information such as the number of overlayed audio files (N.OAF) and the number of samples (D/ND). Additionally, the table provides a breakdown of the F1 score based on different categories (D and ND) and the F1 score on the disruptive class obtained with the best model on the original test set composed of single files (SF).

| Dataset | N. | N. Samples | F1 | F1 | F1 |
|---|---|---|---|---|---|
| | OAF | (D/ND) | (SF) | (D) | (ND) |
| RAVDESS | 2 | 32/28 | 0.91 | 0.89 | 0.84 |
| | 5 | 12/11 | 0.91 | 0.83 | 0.71 |
| | 10 | 6/5 | 0.91 | 0.92 | 0.88 |
| TESS | 2 | 400/300 | 0.77 | 0.75 | 0.29 |
| | 5 | 160/120 | 0.77 | 0.73 | 0.05 |
| | 10 | 80/60 | 0.77 | 0.73 | 0.0 |
| SAVEE | 2 | 30/30 | 0.67 | 0.38 | 0.67 |
| | 5 | 12/12 | 0.67 | 0.25 | 0.63 |
| | 10 | 6/6 | 0.67 | 0.50 | 0.75 |
| CREMA-D | 2 | 24/6 | 0.92 | 0.90 | 0.44 |
| | 5 | 9/2 | 0.92 | 0.95 | 0.67 |
| | 10 | 4/1 | 0.92 | 0.89 | 0.0 |

### 5.5. *Multiple voices*

Table 7 shows that in all experiments, except for SAVEE, the performance is comparable to that achieved with single voices. This demonstrates the robustness of our approach against multiple overlapping voices.

### 5.6. *Ensemble*

As can be observed in Table 8 the *averaged probability* is the best-performing voting strategy, obtaining slightly better scores than the best single model. We also notice that excluding models based on probability confidence ranges leads to lower model performance. This seems to indicate that even limited-confidence predictions provide a useful contribution to the ensemble. On average each network is excluded in 21% of the prediction, and never in more than 31% of the cases, indicating that models are often fairly confident in their predictions.

## 6. Conclusions

We present a novel approach for the detection of disruptive situations in the domain of public transport, framing the problem as an SER task. We analyze popular datasets for SER, highlight the limits of previous approaches for training unbiased classifiers, and re-define data

**Table 8**
Ensemble models performance on the RAVDESS dataset. In each prediction, models that output a probability within the exclusion range are not considered. Column *Average Exclusions* reports the average number of predictions from which the models are excluded.

| Strategy | Exclusion range | Average Exclusions(%) | Accuracy | F1 |
|---|---|---|---|---|
| Majority Voting | None | 0 | 0.90 | 0.91 |
| Majority Voting | [0.4-0.6] | 21 | 0.89 | 0.90 |
| Majority Voting | [0.3-0.7] | 21 | 0.89 | 0.90 |
| Averaged Probability | None | 0 | **0.92** | **0.92** |
| Averaged Probability | [0.4-0.6] | 21 | 0.88 | 0.89 |
| Averaged Probability | [0.3-0.7] | 21 | 0.88 | 0.89 |

splitting as a solution. To overcome the scarcity of real data for the domain of interest, we introduce a data-augmentation process to obtain realistic domain-specific recordings. Our experiments confirm that data augmentation is particularly beneficial in some cases. Furthermore, we explore the use of ensemble to obtain better performances and evaluate the models on noisy data. We find that our SVM models are more robust against noise than CNNs and that ensemble settings have limited impact on the system's performance. Importantly, the performances of our best models on male-only/female-only datasets are comparable, suggesting model fairness.

The overall performance results show that simple classifiers like SVMs and CNNs are capable of achieving satisfying performance for SER on multiple datasets when relying on informative audio features like MFCCs and RMSE. We shall however remark that while the SVM may be more suitable for small samples, its effectiveness may diminish in real-world situations with large-scale training data, as the complexity of the model and computational demands increase. This is a limitation that should be considered in future studies using larger datasets.

Although our employed architectures are less sophisticated than current state-of-the-art models, we posit that, for the purposes of our application, simplicity is an advantage. The reason is that in our envisaged real-world deployments, such systems ought to seamlessly integrate into edge systems where computing resources may be limited.

These results show that the formulation of disruptive situation detection as an SER task is a promising research direction that deserves further investigation and an investment in datasets for validation.

With respect to previous literature (Laffitte et al., 2016), which restricted disruptive situation detection to a scream recognition task, we address disruptive situation detection in a broader sense, paving the way to the development of intelligent systems applicable in more versatile real-life contexts. As far as limitations, our study does not address multi-cultural, multi-lingual scenarios, and has not been tested in real-world deployments. However, these shortcomings are mainly due to the lack of reference data.

As future work, we plan to train and evaluate models on different dataset combinations following a transfer learning formulation, similarly to what has been proposed by Gerczuk et al. (2023). For this purpose, a re-annotation process might be considered to overcome annotation inconsistencies between datasets (Abbaschian et al., 2021). We also intend to investigate advanced techniques for feature extraction and more sophisticated model architectures, such as attention-based models (Galassi et al., 2021). Furthermore, we intend to address the problem of multilingualism by experimenting on datasets such as BAUM-1 (Zhalehpour et al., 2017) or SUBESCO (Sultana et al., 2021). Additionally, we plan to explore adaptive speech features in an attempt to increase our SER performance, as suggested by Wu et al. (2018). Finally, we believe that the creation of a real-world dataset is needed in order to enable progress in the field of public transport safety and security. This would enable testing advanced solutions in a real-world setting, as proposed by Wu et al. (2015).

## Funding

This work has been partially funded by the H2020 5GMED project under Grant agreement #951947, by the H2020 StairwAI project under Grant #101017142, by the H2020 HumanE AI Network under contract #952026, and by the European Commission NextGeneration EU programme, PNRR-M4C2-Investimento 1.3, PE00000013-"FAIR" - Spoke 8.

## CRediT authorship contribution statement

**Eleonora Mancini:** Data curation, Formal analysis, Investigation, Resources, Software, Validation. **Andrea Galassi:** Supervision. **Federico Ruggeri:** Supervision. **Paolo Torroni:** Funding acquisition, Project administration, Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The paper only uses publicly available data.

## References

Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, *21*, 1249. https://doi.org/10.3390/s21041249.

Akçay, M. B., & Oguz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56–76. https://doi.org/10.1016/j.specom.2019.12.001.

An, S., Ji, L. J., Marks, M., & Zhang, Z. (2017). Two sides of emotion: Exploring positivity and negativity in six basic emotions across cultures. *Frontiers in Psychology*, *8*, 1–14. https://doi.org/10.3389/fpsyg.2017.00610.

An, X. D., & Ruan, Z. (2021). Speech emotion recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. *Journal of Physics. Conference Series*, *1861*, Article 012064. https://doi.org/10.1088/1742-6596/1861/1/012064.

Andayani, F., Theng, L. B., Tsun, M. T. K., & Chua, C. (2022). Hybrid lstm-transformer model for emotion recognition from speech audio files. *IEEE Access*, *10*, 36018–36027. https://doi.org/10.1109/ACCESS.2022.3163856.

Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. In *KES* (pp. 251–260). Elsevier. Online.

Beltrán, J., Navarro, R. F., Chávez, E., Favela, J., Soto-Mendoza, V., & Ibarra, C. (2019). Recognition of audible disruptive behavior from people with dementia. *Personal and Ubiquitous Computing*, *23*, 145–157. https://doi.org/10.1007/s00779-018-01188-8.

Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. *Speech Communication*, *52*, 613–625. https://doi.org/10.1016/j.specom.2010.02.010.

Bottou, L., Chapelle, O., DeCoste, D., & Weston, J. (2007). *Support vector machine solvers* (pp. 1–27).

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, *5*, 377–390. https://doi.org/10.1109/TAFFC.2014.2336244.

Cao, H., Verma, R., & Nenkova, A. (2015). Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Computer Speech & Language*, *29*, 186–202. https://doi.org/10.1016/j.csl.2014.01.003.

Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27:27. https://doi.org/10.1145/1961189.1961199.

Chattopadhyay, S., Dey, A., & Basak, H. (2020). Optimizing speech emotion recognition using manta-ray based feature selection. arXiv preprint arXiv:2009.08909 [abs]. https://arxiv.org/abs/2009.08909. arXiv:2009.08909.

Chen, Z., Li, J., Liu, H., Wang, X., Wang, H., & Zheng, Q. (2023). Learning multi-scale features for speech emotion recognition with connection attention mechanism. *Expert Systems with Applications*, *214*, Article 118943. https://doi.org/10.1016/j.eswa.2022.118943.

Chourasia, M., Haral, S., Bhatkar, S., & Kulkarni, S. (2021). Emotion recognition from speech signal using deep learning. In *ICICI* (pp. 471–481). Singapore: Springer.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241–258. https://doi.org/10.1007/s11704-019-8208-z.

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*, 550–553. https://doi.org/10.1037/0033-295x.99.3.550.

Ekman, P. (1989). *The argument and evidence about universals in facial expressions of emotion*In *Wiley handbooks of psychophysiology* (pp. 143–164). Oxford, England: John Wiley & Sons.

Ekman, P. (1999). Basic emotions. In T. Dalgleish, & M. J. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60).

Er, M. B. (2020). A novel approach for classification of speech emotions based on deep and acoustic features. *IEEE Access*, *8*, 221640–221653. https://doi.org/10.1109/ACCESS.2020.3043201.

Fahmy, M. M. (2010). Palmprint recognition based on mel frequency cepstral coefficients feature extraction. *Ain Shams Engineering Journal*, *1*, 39–47. https://doi.org/10.1016/J.ASEJ.2010.09.005.

Fu, C., Liu, C., Ishi, C. T., & Ishiguro, H. (2020). An end-to-end multitask learning model to improve speech emotion recognition. In *EUSIPCO* (pp. 1–5). Amsterdam, Netherlands: IEEE.

Galassi, A., Lippi, M., & Torroni, P. (2021). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*, 4291–4308. https://doi.org/10.1109/TNNLS.2020.3019893.

Garnerin, M., Rossato, S., & Besacier, L. (2021). Investigating the impact of gender representation in asr training data: A case study on librispeech. In *3rd workshop on gender bias in natural language processing, association for computational linguistics* (pp. 86–92).

Gerczuk, M., Amiriparian, S., Ottl, S., & Schuller, B. W. (2023). Emonet: A transfer learning framework for multi-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, *14*, 1472–1487. https://doi.org/10.1109/TAFFC.2021.3135152.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org.

Guðnason, J., & Brookes, M. (2008). Voice source cepstrum coefficients for speaker identification. In *ICASSP* (pp. 4821–4824). Las Vegas, NV, USA: IEEE.

Haq, S., & Jackson, P. J. (2010). *Multimodal emotion recognition. IGI global*. UK: University of Surrey.

Huang, C., Song, B., & Zhao, L. (2016). Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering. *International Journal of Speech Technology*, *19*, 805–816. https://doi.org/10.1007/s10772-016-9371-3.

Iqbal, M. Z. (2021). Mfcc and machine learning based speech emotion recognition over tess and iemocap datasets. *Foundation University Journal of Engineering and Applied Sciences*, *2*, 25–30. https://doi.org/10.33897/FUJEAS.V1I2.321.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR (poster)* (pp. 1–15). San Diego, CA, USA: ICLR.

Laffitte, P., Sodoyer, D., Tatkeu, C., & Girin, L. (2016). Deep neural networks for automatic detection of screams and shouted speech in subway trains. In *ICASSP* (pp. 6460–6464). Shanghai, China: IEEE.

Liu, Z. T., Han, M. T., Wu, B. H., & Rehman, A. (2023). Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning. *Applied Acoustics*, *202*, Article 109178. https://doi.org/10.1016/j.apacoust.2022.109178.

Livingstone, S. R., & 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*, Article e0196391. https://doi.org/10.1371/journal.pone.0196391.

Mancini, E. (2021). Disruptive situations detection on public transports through speech emotion recognition. Master's thesis, Italy: University of Bologna. http://amslaurea.unibo.it/24721/.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. In *SCIPY* (pp. 18–25). Austin, Texas: SCIPY. https://librosa.org. (Accessed 18 May 2023).

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, *7*, 125868–125881. https://doi.org/10.1109/ACCESS.2019.2938007.

Mocanu, B., & Tapu, R. (2022). Emotion recognition from raw speech signals using 2d CNN with deep metric learning. In *IEEE international conference on consumer electronics* (pp. 1–5). IEEE.

Mustaqeem, & Kwon, S. (2021). 1d-cnn: Speech emotion recognition system using a stacked network with dilated cnn features. *Computers, Materials & Continua*, *67*, 4039–4059. https://doi.org/10.32604/cmc.2021.015070.

Nagase, R., Fukumori, T., & Yamashita, Y. (2022). Speech emotion recognition using label smoothing based on neutral and anger characteristics. In *4th IEEE global conference on life sciences and technologies* (pp. 626–627). IEEE.

Oflazoglu, C., & Yildirim, S. (2013). Recognizing emotion from Turkish speech using acoustic features. *EURASIP Journal on Audio, Speech, and Music Processing*, *2013*, 26. https://doi.org/10.1186/1687-4722-2013-26.

Padi, S., Manocha, D., & Sriram, R. D. (2020). Multi-window data augmentation approach for speech emotion recognition. pp. 1–5, arXiv preprint arXiv:2010.09895 [abs]. https://arxiv.org/abs/2010.09895.

Pandey, S. K., Shekhawat, H. S., & Prasanna, S. R. M. (2022). Attention gated tensor neural network architectures for speech emotion recognition. *Biomedical Signal Processing and Control*, *71*, Article 103173. https://doi.org/10.1016/j.bspc.2021.103173.

Patel, N., Patel, S., & Mankad, S. H. (2022). Impact of autoencoder based compact representation on emotion detection from audio. *Journal of Ambient Intelligence and Humanized Computing*, *13*, 867–885. https://doi.org/10.1007/s12652-021-02979-3.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*, 315–341. https://doi.org/10.1007/s10648-006-9029-9.

Pichora-Fuller, M. K., & Dupuis, K. (2020). Toronto emotional speech set (tess). https://doi.org/10.5683/SP2/E8H2MF.

de Pinto, M. G., Polignano, M., Lops, P., & Semeraro, G. (2020). Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In *EAIS* (pp. 1–5). Bari, Italy: IEEE.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margins classifiers* (pp. 61–74). Redmond, WA: MIT Press.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*, 715–734. https://doi.org/10.1017/S0954579405050340.

Roa, J., Jacob, G., Gallino, L., & Hung, P. C. K. (2018). Towards smart citizen security based on speech recognition. In *CACIDI* (pp. 1–6). Buenos Aires, Argentina: IEEE.

Sato, N., & Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *Journal of Natural Language Processing*, *14*, 83–96. https://doi.org/10.5715/jnlp.14.4_83.

Shahin, I., Hindawi, N. A. A., Nassif, A. B., Alhudhaif, A., & Polat, K. (2022). Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Systems with Applications*, *188*, Article 116080. https://doi.org/10.1016/j.eswa.2021.116080.

Singh, P., Sahidullah, M., & Saha, G. (2023). Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Communication*, *146*, 53–69. https://doi.org/10.1016/j.specom.2022.11.005.

Sultana, S., Iqbal, M. Z., Selim, M. R., Rashid, M. M., & Rahman, M. S. (2022). Bangla speech emotion recognition and cross-lingual study using deep CNN and BLSTM networks. *IEEE Access*, *10*, 564–578. https://doi.org/10.1109/ACCESS.2021.3136251.

Sultana, S., Rahman, M. S., Selim, M. R., & Iqbal, M. Z. (2021). SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLoS ONE*, *16*. https://doi.org/10.1371/journal.pone.0250173.

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology*, *21*, 93–120. https://doi.org/10.1007/s10772-018-9491-z.

Tiwari, U., Soni, M. H., Chakraborty, R., Panda, A., & Kopparapu, S. K. (2020). Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions. In *ICASSP* (pp. 7194–7198). Barcelona, Spain: IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017* (pp. 5998–6008). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion recognition from speech. pp. 1–14, arXiv preprint arXiv:1912.10458 [abs]. http://arxiv.org/abs/1912.10458.

Vogt, T., & André, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *LREC, European Language Resources Association (ELRA)* (pp. 1123–1126). http://www.lrec-conf.org/proceedings/lrec2006/pdf/392_pdf.pdf.

Wang, K., Zhang, Q., & Liao, S. (2014). A database of elderly emotional speech. In *Proc. int. symp. signal process. biomed. eng. informat.* (pp. 549–553).

Wang, S., Cao, J., Sun, K., & Li, Q. (2020). SIEVE: Secure in-vehicle automatic speech recognition systems. In *RAID* (pp. 365–379). Berkeley, California: USENIX Association. https://www.usenix.org/conference/raid2020/presentation/wang-shu.

Wu, C., Huang, C., & Chen, H. (2015). Automatic recognition of emotions and actions in bi-modal video analysis. In C. Hsu, F. Xia, X. Liu, & S. Wang (Eds.), *Internet of vehicles - safe and intelligent mobility - second international conference, proceedings* (pp. 427–438). Springer.

Wu, C., Huang, C., & Chen, H. (2018). Text-independent speech emotion recognition using frequency adaptive features. *Multimedia Tools and Applications*, *77*, 24353–24363. https://doi.org/10.1007/s11042-018-5742-x.

Yang, Y., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, *3*, 40:1–40:30. https://doi.org/10.1145/2168752.2168754.

Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2017). BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, *8*, 300–313. https://doi.org/10.1109/TAFFC.2016.2553038.

Zhang, Z., Geiger, J. T., Pohjalainen, J., Mousa, A. E., Jin, W., & Schuller, B. W. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology*, *9*, 49:1–49:28. https://doi.org/10.1145/3178115.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1d & 2d CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035.

Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. *Sensors*, *17*, 1694. https://doi.org/10.3390/s17071694.