UNIVERSITY POLITEHNICA OF BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 1048 **from** 10-07-2023

PhD THESIS
SUMMARY

Matei-Șerban MIHALACHE

SPEECH SIGNAL ANALYSIS AND PROCESSING TECHNIQUES FOR AUTOMATIC RECOGNITION OF PARALINGUISTIC ELEMENTS, WITH APPLICATIONS IN FORENSIC SPEECH

TEHNICI DE ANALIZĂ ȘI PRELUCRARE A SEMNALULUI VOCAL PENTRU RECUNOAȘTEREA AUTOMATĂ A ELEMENTELOR PARALINGVISTICE, CU APLICAȚII ÎN EXPERTIZA CRIMINALISTICĂ A VORBIRII

THESIS COMMITTEE

Prof. Gheorghe BREZEANU University POLITEHNICA of Bucharest	President
Prof. Dragoș BURILEANU University POLITEHNICA of Bucharest	PhD Supervisor
Prof. Daniela TĂRNICERIU "Gh. Asachi" Technical University of Iași	Referee
Prof. Corneliu RUSU Technical University of Cluj-Napoca	Referee
Prof. Constantin PALEOLOGU University POLITEHNICA of Bucharest	Referee

BUCHAREST 2023

Contents

Chapter 1

Introduction

Within this work, existing machine learning (ML) and deep learning (DL) models are extended, and novel methods and techniques are proposed, developed, and validated within the context of speech signal analysis and processing for the automatic recognition of paralinguistic elements, with applications in forensic speech.

1.1. An overview of paralinguistic elements and recognition tasks

The concept of paralinguistics was first formulated by American linguist George Trager, and refers to the meta-information present in spoken communication, the nuances conveyed beyond the lexical / semantic content in regard to affective dimensions [Bac95, Bac99] or other psychological manifestations [Laz99, Vil12].

The most important fundamental paralinguistic elements and their associated recognition tasks for speech signals are defined as follows [Laz99, Mat09]:

- stress = a prolonged state of psychological and physiological arousal that negatively affects a subject's state of mind, mood, etc.
  $\Longrightarrow$ speech under stress detection (SSD);
- deception = behavior including actions such as withholding information, providing incomplete or false information (i.e., lying), etc. for the purpose of the subject's individual gain, usually at the expense of others
  $\Longrightarrow$ deceptive speech detection (DSD);
- emotions = higher-level, transient neurophysiological responses to stimuli, determining coordinated physical and mental responses for appraisal of the stimuli and preparation for the subject to take subsequent actions
  $\Longrightarrow$ speech emotion recognition (SER).

The main difficulties in this context arise from the subjective nature of the paralinguistic content evaluation, since the expressions are highly personal, and likely generated from a frame of reference to which that of the evaluator is hard to calibrate. For automatic recognition, however, the system must be trained not only using high-quality paralinguistic content annotations, but also high-quality paralinguistic content itself. In other words, the recorded spoken interactions should be natural, realistic, spontaneous, unguided, unrestricted, and as varied as possible.

This is, of course, seldom the case due to practical reasons. Developing a paralinguistic dataset is a considerably difficult task by itself, and ensuring that all the previously listed criteria are respected is often simply not an option, either due to data unavailability or to the prohibitively long time required to construct such a dataset. Thus, most often, actors are hired to record spoken interactions in which they mimic affective expressions and other paralinguistic elements to the best of their ability. However, this simulated nature, combined with the fact that the speech content is usually predetermined and rehearsed, leads to poor generalizability and reduced robustness for the trained automatic recognition systems, since the task-related quality of the available data has little in common with the instances later encountered in realistic scenarios, "in the wild".

1.2. Paralinguistic applications in forensic speech

One field to which the automatic recognition of paralinguistic content from speech lends itself directly is forensic speech analysis and, more generally, forensic and law enforcement operations. Due to the nature of these domains, the main focus should be on detecting negative emotions, manifestations of high stress levels, and especially engagement in deceptive behavior. For various possible applications, determining higher-level affective and behavioral patterns is also highly relevant, especially when concerned with complex and long-term actions undertaken by the subjects in question.

A list of a few key application examples is provided in the following:

- law enforcement active investigations in the form of conducting police interviews, questionings, or taking testimonies from persons of interest in criminal cases, suspects, witnesses, victims;
- criminal and terrorist activity anticipation and prevention through surveillance and recognition of suspicious long-term affective patterns;
- suspicious behavior monitoring at checkpoints, airports, tourist attractions, crowded areas, and other sensitive areas of interest;
- lie detection systems employing only on the audio modality; etc.

It must be emphasized that, in all of these examples and in any forensic or law enforcement applications, the position advocated for in this thesis is not complete automation of such tasks, removing the human element. One key ethical and security-related aspect is for such tasks to always have final decisions and actions taken solely by human agents, with the artificial intelligence (AI) system involvement taking only the form of non-definitive alarms, suggestions, and recommendations.

It is clear from the research conducted so far in the field that ML/DL models for SSD, DSD, and SER tasks that better performance is obtained when working with multimodal data (i.e., audio-video recordings, physiological data, etc.). Despite this, focusing on the research and development of such models using only speech data is of great interest for its advantages: the possibility to record speech data inconspicuously, reducing the subject's awareness and their chance to manipulate the data, or to approach situations where only speech data is available (e.g., telephone calls).

1.3. An interdisciplinary research area: scope and objectives

The scope of this doctorate can be understood as:

- Developing novel and high-performance ML/DL models and techniques for automatic speech under stress detection (SSD), deceptive speech detection (DSD), and speech emotion recognition (SER), focusing on negative and high-intensity affective and deceptive manifestations.
- Developing extensive and robust sets of speech signal features (i.e., key speech parameters and mathematical and/or physical measures) relevant for automatic recognition of paralinguistic elements from the speech content.
- Developing novel, high-quality, realistic datasets for paralinguistic element recognition, overcoming the limitations and disadvantages present for other available datasets, and allowing public access to the developed datasets.
- Determining long-term affective patterns and behaviors relevant for forensic and law enforcement applications, and providing affective models for their study in relation to suspicious behavior monitoring.

1.4. Thesis structure

The structure of this thesis includes eight chapters that comprise the following content:

Chapter 1 serves as an introduction into the general concepts and particular aspects of paralinguistic elements and the corresponding (automatic) recognition tasks, and how they are applicable to the field of forensic speech. The chapter also defines the scope and the objectives of the thesis and outlines its structure.

Chapter 2 is a comprehensive summary of the main theoretical knowledge required in the development of this work, concerning the fields of speech analysis and processing and machine learning and deep learning. The chapter presents an extensive set of algorithmically extracted (hand-crafted) features used successfully for the paralinguistic tasks that are part of the scope of this thesis, the ML/DL models employed in developing the systems, and the training and testing methodologies approached in order to ensure proper performance validation for the proposed systems.

Chapter 3 debuts with a more detailed discussion concerning the different paralinguistic elements that are the focus of this work (psychological stress, emotions,

and deception) and how they are related to each other. The chapter then covers the first task, detecting psychological stress from speech, first by describing the related work published in literature and the current state of the art, followed by presenting the proposed system architectures, and the experimental setup, including the datasets employed, the methodology, the results, and their discussion.

Chapter 4 is the first of two chapters related to the second paralinguistic task approached in this work, i.e., deceptive speech detection. The chapter begins with a description of the challenges and requirements for the development of high-quality and reliable datasets for paralinguistic tasks, especially for deception detection, as well as an overview of the available public datasets for the aforementioned task. The second half of the chapter describes in detail the novel Romanian Deva Criminal Investigation Audio Recordings (RODeCAR) dataset, developed as part of the doctorate, arguing for the improvements it provides over other similar corpora.

Chapter 5 expands upon the subject of deceptive speech detection. The chapter follows the same structure as Chapter 3, comprising an overview of other approaches publicly reported and the current state of the art, the proposed system architectures, and the experimental setup and results obtained, including the first published results for the RODeCAR dataset introduced in Chapter 4. The systems developed in this work for deceptive speech detection include a voice activity detection component, which is also described appropriately throughout the chapter in terms of providing a review of previously published literature concerning this subtask, as well as the subsystem architecture and experimental validation.

Chapter 6 presents the main body of work done on speech emotion recognition over the course of the doctorate, covering three types of systems: direct approaches, using algorithmically extracted features and classifiers based on neural networks; multidomain approaches that attempt to use dimensional modeling to establish a mapping between the continuous affect space and emotions as discrete categories, training the system to simultaneously solve both a classification and a regression problem; and transfer learning approaches, for which large, high performing neural network models designed for visual recognition of objects in images are repurposed and retrained in order to recognize distinctive patterns in speech spectrograms. The chapter follows the same structure as Chapter 3 and Chapter 5: related work, proposed system architectures, and experimental setup, results, and discussions.

Chapter 7 is a follow-up to the previous chapter, providing the theoretical background and experimental validation for the problem of speech emotion remanence, i.e., the twin hypotheses that, as an emotionally charged event approaches in time, subjects will exhibit stronger negative emotions that will be present (and detectable) in their speech for longer time intervals after being triggered.

Chapter 8 is reserved for conclusions, offering a summary of the developments followed and the best obtained results over the course of this work, as well as discussing the candidate's original contributions, list of articles and papers published as part of the doctorate, and the perspectives for further developments in the machine learning, deep learning, and speech processing fields, with particular focus on automatic recognition of paralinguistic elements for law enforcement and forensic applications.

Chapter 2

Theoretical background: speech signal analysis, machine learning

This chapter summarizes and offers insights into the theoretical knowledge required for the development of the work presented in this thesis.

2.1. An extensive hand-crafted feature set for speech processing

The audio input goes through the preprocessing steps (resampling, normalization, filtering), with the unframed and framed data subsequently being designated the *audio vector* and the *audio frame vector*. In this work, the framing scheme involved using Hamming windows of 25 ms duration with a 15 ms (60%) overlap.

The features extracted in the next stage form a set that extends the ComParE set [Sch14] by including other features proven to be relevant for paralinguistic tasks both in recent literature and in preliminary experiments performed in this work. The audio vector is segmented, and the segment-wise features (SWFs) are extracted. The other features are extracted frame-wise from the audio frame vector and consist of: (i) time-domain features (TDFs); (ii) frequency-domain features (FDFs); (iii) the first 13 Mel-frequency cepstral coefficients (MFCCs); and (iv) modulation-based features (MBFs), proposed in [Cha14]. Delta and delta-delta coefficients are computed for the MFCCs, the TDFs, and the FDFs, as well as for some of the MBFs. Together with the SWFs, these represent the time-step features (TSFs), on which the F(·) set of functionals (mean and standard deviation) is applied, resulting in the utterance-wise features (UWFs). Finally, the N(·) function, z-score normalization, is applied per speaker. The obtained normalized feature vector (FV) has a total size of 2,260 and is used in its

entirety or through various subsets as the input data for many of the systems developed in this work. The proposed preprocessing and feature extraction stages described previously are illustrated together in Figure 2.2.

The SWFs (pitch, HNR, local jitter, local shimmer) are extracted using the Python implementations of the Yet Another Algorithm for Pitch Tracking (YAAPT) algorithm and of Praat. The TDFs considered are the RMS energy, offering a measure of the intensity of each frame of the speech signal, and the zero-crossing rate (ZCR), serving as a measure of the high-frequency noise-like content of the signal, with higher values typically corresponding to unvoiced regions of speech. The FDFs considered are: the low-frequency energy (within the 250 – 650 Hz subband), the high-frequency energy (within the 1 – 4 kHz subband), and, for several Mel-spaced frequency subbands, the spectral centroids, the spectral spread, the spectral skewness, the spectral kurtosis, the spectral entropy, the spectral flux, the spectral slope, and the spectral roll-off points (computed for the 25%, 50%, 75%, and 90% thresholds). The MBFs are obtained by treating the speech signal as a series of amplitude and frequency micro-modulations (AM-FM) and computing features based on the estimated instantaneous amplitude and frequency at each time step by demodulating the signal. The MBFs capture non-linear, time-varying speech production phenomena, including the fine structure of speech formants. The MFCCs are some of the effective and often used features for speech analysis and processing applications. They can be interpreted as a compressed model of the vocal tract, i.e., offering a description of the quefrency response of the vocal tract in the source-filter model of speech. The delta coefficients, $\Delta(\cdot)$ of a feature F are measures of its local variation, from index to index (e.g., across frames). Thus, they can be interpreted as a description of its first-order (time) variation. By applying the Δ function to the delta coefficients, the delta-delta coefficients may be computed.

2.2. Machine learning and deep learning models employed

A machine learning (ML) system can be viewed as a self-adjusting system in which the operations required to relate the input and output data are not determined and structured by a human agent, but are the result of automatic convergence to a numerically optimal solution, i.e., for an ML system, developers use inputs and outputs to determine “rules”.

A K-means model (KMM), or the K-means clustering algorithm [Bis06], is one of the simplest, yet effective clustering algorithms, and can be interpreted as a particular case of the expectation-maximization (EM; or Baum-Welch) algorithm. The idea behind Gaussian mixture models (GMMs) is to model the underlying probability distribution of the data as a superposition (mixture) of normal (Gaussian) distributions [Bis06].

For the SVM model, assuming the input data is linearly separable, i.e., there exists a hyperplane that perfectly separates the instances belonging to each of two classes, the hyperplane is chosen so that the margin (the minimum distance between the decision boundary and the points closest to the decision boundary, i.e., the *support vectors*) is maximized. The initial model was constructed to include a non-linear transformation

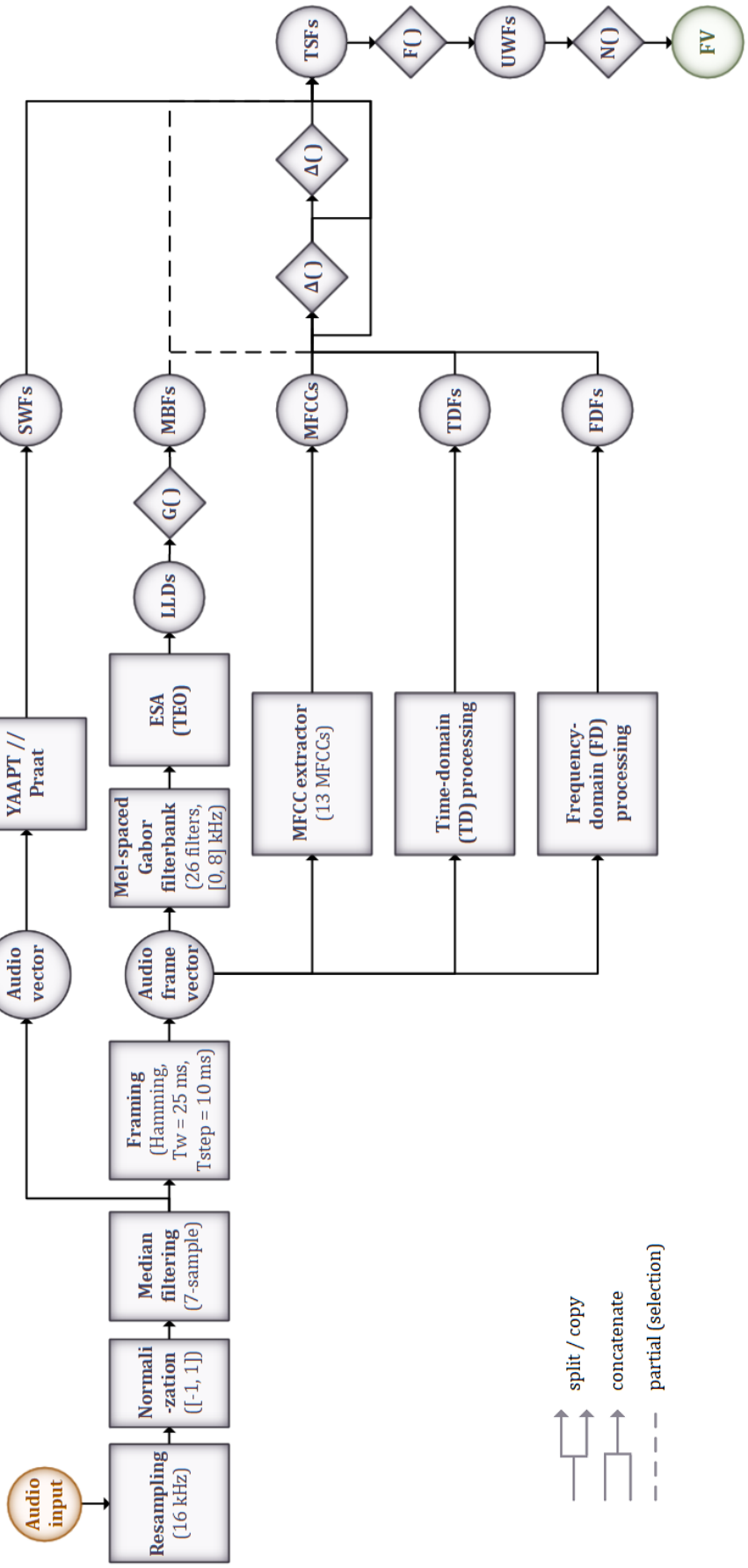

Figure 2.2 – Detailed block diagram of the preprocessing and feature extraction stages.

function, $\Phi(\cdot)$, applied to the feature vector space, because the input data is often not linearly separable in the original feature space, but might be in a higher-dimensional space, the mapping between the two being given by $\Phi(\cdot)$. Directly computing $\Phi(\cdot)$ is prohibitively expensive. Instead, a kernel function allows the indirect usage of $\Phi(\cdot)$ as a simple dot product in the original feature vector space, reducing the computational complexity considerably. For non-binary, K-class problems, multiple SVMs are used in ensemble via: the one-vs.-rest (OvR) strategy, which involves training K SVMs separately, one for each class, by first grouping together all instances that are not part of the current "positive" class into a "negative" class; or the one-vs.-one (OvO) strategy, by training $K \cdot (K-1)/2$ classifiers separately, one for each class combination.

The basic building block of a fully-connected neural network (FCNN) is the *neuron*, which models the equivalent biological cell found in the human nervous system. The artificial neuron takes as input variables arriving from previous neurons with associated weights, the total stimulus (the *activation*), representing the linear combination of the inputs, to which a non-linear *activation function* is applied. A single neuron would not offer relevant processing power, so several are organized together into *layers*. The main heuristic approach that has proven to be increasingly feasible and high-performing in the contemporary development of the field is to trade "breadth" for "depth": instead of attempting to extract information through a single transformation between the input data feature space and the output data, doing it in several stages, i.e., using (a large number of) hidden layers, with each one ideally obtaining a higher-level abstraction of the input data by stacked transformations.

Convolutional neural networks (CNNs) are the backbone of modern deep learning (DL) AI systems. They can be seen as a derivative of FCNNs based on a few key principles, including locality, invariance, and deep abstraction [Bis06, Goo16]. These principles translate into not having all the neurons in a layer connected to each of the neurons in the consecutive layer, and each layer creating a new representation of the input data, called a *feature map*. Ideally, every subsequent layer would extract feature maps that offer a higher and higher level of abstraction for the data.

2.3. Training and testing methodologies

The fundamental methodology involves splitting the data into several subsets [Has06]. Ideally, this involves three subsets: one for training, one for development validation (*dev* / *val*), and one for general evaluation (final testing; *eval* / *test*). The total size of the dataset should be as large as possible. But many available datasets are relatively small, and the alternative is to apply *cross-validation*, using repeated two-way training-validation splits. A number of advanced training techniques have been adopted for the systems developed in this work to improve the training process and to increase the models' performance: regularization, dropout, batch normalization, and others.

For regression problems, apart from the value of the loss function itself, the performance metrics are the correlation coefficient, ρ, defined in (2.91), where σ_y and $\sigma_{\hat{y}}$ are the standard deviations of the target values and of the model output values, and $K_{y\hat{y}}$

represents the covariance between the two vectors, and the concordance correlation coefficient, ρ_c, defined in (2.92), where μ_y, $\mu_{\hat{y}}$, σ_y^2, and $\sigma_{\hat{y}}^2$ are the means and variances of the target values and of the model output values. For classification tasks, considering N instances belonging to K classes, with N_k the number of instances in each class, let H_k represent the number of correct predictions made by the model for class k, given by (2.98a), where the function $\mathrm{h}_{(k)}(\cdot, \cdot)$ is defined in (2.98b); and F_k the number of incorrect membership predictions made for class k, given by (2.99a), where the function $\mathrm{f}_{(k)}(\cdot, \cdot)$ is defined in (2.99b). The precision (P), recall (R), unweighted and weighted accuracy (UA / WA) metrics are then defined according to (2.100) – (2.103).

$$\rho = \frac{K_{y\hat{y}}}{\sigma_y \cdot \sigma_{\hat{y}}} \tag{2.91}$$

$$\rho_c = \frac{2\rho \cdot \sigma_y \cdot \sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + \left(\mu_y - \mu_{\hat{y}}\right)^2} \tag{2.92}$$

$$H_k = \sum_{n=0}^{N-1} \mathrm{h}_{(k)}\left(y_n, \hat{y}_n\right) \tag{2.98a}$$

$$\mathrm{h}_{(k)}\left(y_n, \hat{y}_n\right) = \begin{cases} 1, & \hat{y}_n = y_n = k \\ 0, & \text{otherwise} \end{cases} \tag{2.98b}$$

$$F_k = \sum_{n=0}^{N-1} \mathrm{f}_{(k)}\left(y_n, \hat{y}_n\right) \tag{2.99a}$$

$$\mathrm{f}_{(k)}\left(y_n, \hat{y}_n\right) = \begin{cases} 1, & \hat{y}_n = k \quad \text{and} \quad y_n \neq k \\ 0, & \text{otherwise} \end{cases} \tag{2.99b}$$

$$\mathrm{P}_k = \frac{H_k}{H_k + F_k} \tag{2.100}$$

$$\mathrm{R}_k = \frac{H_k}{N_k} \tag{2.101}$$

$$\mathrm{WA} = \frac{1}{N} \sum_{k=0}^{K-1} H_k = \frac{1}{K} \sum_{k=0}^{K-1} \frac{K \cdot N_k}{N} \cdot \frac{H_k}{N_k} \tag{2.102}$$

$$\mathrm{UA} = \frac{1}{K} \sum_{k=0}^{K-1} \frac{H_k}{N_k} \tag{2.103}$$

2.4. Chapter conclusions

In this chapter, a summary of the main theoretical knowledge employed during the development of this work was presented, including an extensive set of algorithmically extracted features, ML/DL models, and fundamental and advanced training and testing methodologies, techniques, and metrics used to ensure proper system performance.

Chapter 3

Speech under stress detection

This chapter covers the task of detecting states of psychological stress from a subject's speech, also called *speech under stress detection* (SSD). Parts of the content herein were published as a conference paper by the candidate [Mih21b].

3.1. Background and related work

It is important to understand that there is a considerable conceptual and visceral overlap between the state of being subjected to psychological stress, externalizing affective states (emotions), and engaging in deceptive behavior. This leads to two fundamental principles that must be taken into account for paralinguistic tasks:

1) A complete separation between these concepts / states is neither possible, nor necessarily desirable, since many forensic speech applications (and beyond) will have end objectives focused on higher level behaviors or manifestations (e.g., suspicious behavior monitoring), in which all of these elements are relevant both as distinct subtasks and for holistic approaches.
2) Each concept / state implies relevance and cannot simply be subsumed within any of the others, since many individual edge cases encountered "in the wild" may very well fall just one of the states or under combinations of two of them, with only small manifestations of the third.

Among the features shown to offer promising results, spectral and cepstral features are prominent, such as spectrogram [He09] or wavelet [Zao14] decompositions, and the Mel-frequency cepstral coefficients (MFCCs) [Cas06, Li07], as well as other acoustic features, e.g., the fundamental frequency [Cas06], the jitter and the shimmer [Li07]. It was also shown that leveraging extended feature sets often improves system accuracy for the SSD task.

Several approaches have been reported using traditional machine learning (ML) models, including hidden Markov models (HMMs) [Cas06], Gaussian mixture models (GMMs) [He09, Zao14], support vector machines (SVMs) [Bes16], or hybrid HMM-GMM models [Li07]. More recently, deep learning (DL) solutions such as convolutional neural networks (CNNs) and hybrid convolutional-recurrent neural networks (CRNNs) have been reported as well [Avi19, Shi20].

3.2. Proposed system architectures

The basic proposed DL system consists of using a deep neural network (DNN) taking as input an extensive set of features obtained by applying high-level statistical functions on algorithmically extracted acoustic, spectral, and cepstral descriptors [Mih21b]. The DNN classifier is a feed-forward fully-connected neural network (FCNN) model, using between 2 and 4 hidden layers, with different numbers of nodes per layer, and an output layer whose size is equal to the number of classes taken into account for each experiment group (4-class, 3-class, or binary classification, i.e., 2-class). Two hidden layer node structures were taken into consideration: the 'constant' architecture, which consist of selecting the same number of nodes for each hidden layer; and the 'log2dec' architecture, which consists of selecting a progressively smaller number of nodes per active layer, following a decreasing $\log_2(\cdot)$ law.

A more advanced system using ensemble classifiers is also proposed. It is shown in Figure 3.3. To this end, a one-vs.-one (OvO) ensemble classification strategy was employed, inspired by the corresponding approach to multiclass SVM models. A total of $K\cdot(K-1)/2$ classifiers (where K is the number of classes) were trained independently for each pair of classes and their output was fed, together with its rounded values (the intermediate binary predictions of each OvO DNN classifier), to a similarly structured DNN that performs final classification.

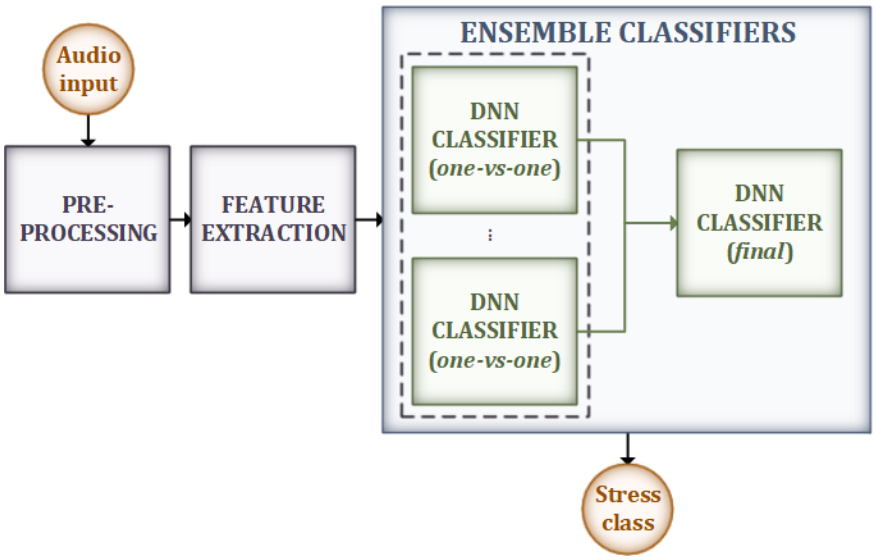

Figure 3.3 – Advanced architecture: ensemble classifiers, one-vs.-one (OvO) strategy.

3.3. Experimental setup and results

The Speech Under Simulated and Actual Stress (SUSAS) database [Han98] contains approximately 14,600 recordings in English, with an average duration of 0.6 s. The first half of the corpus, having 9 speakers (all male), comprises recordings under simulated stress conditions, resulting in 11 classes: 7 referring to the speaking style (Fast, Slow, Soft, Loud, Clear, Angry, Question), 3 to the environment in which the speakers were placed at the time of recording (Cond50 – recorded while solving a medium difficulty task; Cond70 – recorded while solving a higher difficulty task; Lombard – recorded while listening to high-intensity pink noise triggering the Lombard effect), and a neutral class. The second half, having 7 speakers (3 female, 4 male), contains recordings made under actual stress conditions, resulting in 5 classes: 2 referring to solving complex tasks (MeS – medium difficulty task; HiS – higher difficulty task), 2 referring to riding two different roller coasters (Freefall and Scream), and a neutral class.

In order to match other works presented in literature, as well as the main target SSD application, i.e., binary detection of psychological stress from speech, the following datasets were created by partitioning the SUSAS database:

- **Set A** – 4 classes under *actual* stress conditions: Scream (SCRM), HiS, MeS, and Neutral (NEU); total size: 3,567 recordings.
- **Set B** – 3 classes from set A: HiS, MeS, and NEU; total size: 3,179 recordings.
- **Set C** – 2 classes from set A: STRS (grouping together SCRM, HiS, and MeS) vs. NEU; same size as set A.
- **Set D** – 4 classes under *simulated* stress conditions: Angry (ANG), Lombard (LOM), Loud (LOU), and Neutral (NEU); total size: 2,518 recordings.
- **Set E** – 2 classes under *simulated* stress conditions: STRS (grouping together ANG, LOM, LOU, Cond50, and Cond70) vs. NEU; total size: 7,556 recordings.

The DNN classifier depth varied between 2 and 4, with an initial number of nodes (for the first hidden layer) of 256 or 128. Other hyperparameters chosen include: the rectified linear unit (ReLU) activation function for the hidden layers, and the softmax activation function for the output layer; and Adam as the optimization algorithm. The same configurations were tested for the OvO DNN classifiers, with the final DNN classifier having a fixed depth equal to 2 or 3, as well as 6 or 12 nodes per hidden layer, for the 3-class and 4-class experiments. For the *actual* stress experiments (datasets A, B, and C), a 10-fold cross-validation testing scheme was selected, with a 70% / 30% dataset split between training and validation, leaving out 2 out of 7 speakers (1 female, 1 male) for each validation subset. For the *simulated* stress experiments (datasets D and E), 10-fold cross-validation was also used, but with a 66% / 33% dataset split, leaving out 3 out of 9 speakers. For all unbalanced cases, class weighting was employed.

A comparison of the obtained results to other works presented in literature is made in Table 3.9, for all available cases, i.e., 4-class *actual* stress conditions (dataset A),

3-class *actual* stress conditions (dataset B), 4-class *simulated* stress conditions (dataset D), and 2-class *simulated* stress conditions (dataset E). The proposed system shows a significant performance (highlighted in green) increase for the 4-class *actual* stress (dataset A), 4-class *simulated* stress (dataset D), and 2-class *simulated* stress (dataset E) cases. It is noted that, in the 3-class *actual* stress (dataset B) case, the only reported results [He09] were for training KNN and GMM models (with the latter demonstrating better results) on frequency-scaled PSD spectrograms extracted from vowel samples. These were extracted from only 547 data instances vs. the much higher number of 3,179 instances used for the approach proposed in this work. This large discrepancy in the training and validation subset sizes can explain the higher accuracy obtained in [He09].

3.4. Chapter conclusions

In this chapter, deep learning systems were proposed, based on employing multiple feed-forward fully-connected deep neural networks (DNNs) connected together within an ensemble one-vs.-one (OvO) classification strategy configuration, and using as input an extensive set of algorithmically extracted acoustic, spectral, and cepstral features. The systems were tested on the SUSAS database, for 5 class grouping subsets (4-class, 3-class, and 2-class SSD tasks for speech under *actual* stress conditions; 4-class and 2-class tasks for speech under *simulated* stress conditions).

Significant performance improvements have been obtained over other relevant state-of-the-art results previously reported in literature, with an (unweighted / weighted) accuracy (UA / WA) of **68.8% / 65.5%** for the 4-class *actual* stress case, **59.2% / 62.4%** for the 3-class *actual* stress case, **66.7% / 81.4%** for the 2-class *actual* stress case, **75.5% / 75.5%** for the 4-class *simulated* stress case, and **76.1% / 78.4%** for the 4-class *actual* stress case.

Table 3.9 – Performance comparison between the best results achieved in this work and other relevant results published in literature.

Dataset	Method	Performance			
		Avg. P [%]	Avg. F1 [%]	Avg. R ≡ UA [%]	WA [%]
A	[Zao14] – GMM	–	–	64.0	–
	This work	69.7	68.9	**68.8**	65.5
B	[He09] – GMM	–	–	–	**73.8**
	This work	61.0	59.5	59.2	62.4
D	[Cas06] – HMM	–	–	72.9	–
	[Avi19] – CNN	–	–	–	71.0
	This work	75.6	75.3	**75.5**	**75.5**
E	[Avi19] – CNN	–	–	–	76.0
	This work	79.5	76.8	76.1	**78.4**

Chapter 4

RODeCAR: A novel dataset for deceptive speech detection

This chapter covers the end-to-end development of a novel, high-quality, objective dataset for *deceptive speech detection* (DSD): the Romanian Deva Criminal Investigation Audio Recordings (RODeCAR) dataset. Parts of the content herein were published as a conference paper by the candidate [Mih19b].

4.1. Background and related work

The results of a polygraph test provided by this test are not admissible as scientific evidence in a court of law. At best, they can influence the investigator to further pursue some lead or line of questioning. While a system for lie detection from speech would not provide legal support either, the results may prove to be of higher accuracy, since some speech characteristics are generally more difficult to alter voluntarily than the parameters tracked by conventional polygraph tests or newer psychological analysis methods, which can be manipulated into yielding false results [Ver09]. More so, even though a multimodal approach would lead to greater performance, an audio-only lie detector can be used inconspicuously in most relevant scenarios to reduce the subject's awareness of its presence, leading to lower success in manipulating the test results.

In order to develop machine learning systems capable of detecting untruthfulness in speech, large datasets with precise and methodical annotation are required for training. The nature of the audio content should be as authentic as possible. Simulated data may affect the system's ability to generalize from unrealistic to real case evaluation [Mor12]. For a sensitive task like lie detection, the concern is even greater, emphasizing the need for speech data from real-life situations.

The main issues revealed for the reviewed publicly available DSD datasets are the use of actors or pre-trained participants (i.e., simulated behavior); having a specific goal in a familiar or relaxed environment (i.e., simulated scenario); having low-stakes tasks (i.e., reduced incentive); and using self-assessment or subjective methods to annotate the data (i.e., subjective annotation) [Mih19b]. To overcome these common significant disadvantages, a different approach is proposed in this work:

1) Participants must not have prior guidance regarding their expected behavior and should have complete control over the content of their answers.
2) A specific scenario should not be created; instead, within reason, a free-form framework should be employed, so as to reduce the predictability of the participants' speech and/or that of the interviewer's line of questioning (if applicable).
3) Participants should be aware of relatively severe (high-stakes) consequences both for engaging in deceptive behavior or admitting incriminating truths.
4) Data annotation should be performed by an expert, after doing follow-up work in order to objectively and clearly determine the truthfulness of the participants' statements.
5) If uncertainty over the labeling is minimized, but not eliminated, a confidence score should be formulated and associated with each interaction.

4.2. The Romanian Deva Criminal Investigation Audio Recordings dataset

In order to address the first three requirements previously outlined, it can be argued that one of the best sources of material would be recordings of real law enforcement investigative activity, in which the participants are actual interviewed suspects, witnesses, etc., and the context is given by hearings and interviews conducted by trained professional law enforcement investigators. This way, the participants have very little prior knowledge regarding the interviewer's possible line of questioning, and there is great incentive both for honest parties to compellingly present a truthful account of what was asked about, as well as for suspects to convincingly hide any incriminating facts. Of course, the sensitive and often classified nature of such recordings is an immediate deterrent. However, the RODeCAR dataset was constructed using files covering 9 closed older criminal cases during which investigations of murder, sexual assault, and fraud were conducted. The fourth requirement on the list was satisfied by performing a meticulous manual review of the recordings and the associated case notes together with the prosecutor who originally investigated the cases, in order to determine the truthfulness of the content. Finally, the inevitable uncertainty concerning the finer details or arising from unavailable information regarding the investigations is addressed by associating a confidence level to each interaction (file), varying from 70% to 100%.

After an initial filtering of the content, to select only the actual interactions with the interviewees, all 20 involved speakers were manually identified (ID) and associated with an ID number, with a special value reserved for the prosecutor (not counted for the

final dataset content). Their gender was also taken into account, having 4 females and 16 males. The audio tracks were then extracted using the FFmpeg framework and saved in 16-bit PCM format, at 16 kHz sampling rate.

At this stage, the 26 processed audio recordings (in total, approximately 7.5 hours of material) are sorted into three distinct categories, depending on the content type and how the participants are involved:

- Questionings (Q): interrogations of participants by the prosecutor in a formal environment and following a strict procedure; this is the most stressful scenario for the participants.
- Interviews (I): interactions between the prosecutor and the participant in an informal environment (often more familiar to the questioned party); this is the least stressful scenario for the participants.
- Testimonies (M): uninterrupted, free-form recounting / confessions given by the participants, often as a follow-up to a previous interrogation.

For each file, semi-automatic segmentation was employed. A segment is defined as a portion of speech from a single speaker, either (i) separated by a pause of at least 200 ms from other portions of speech from the same speaker; or (ii) separated from other portions of speech from a different speaker, regardless of the onset delay duration.

The binary annotation (truthful, T / untruthful, UT) is made per speech segment, but in a global sense; e.g., a short segment containing factually accurate information, found within a longer speaker turn engaging in deceptive behavior, will also be labeled as untruthful. This is further supported by the argument that the state of mind the participants find themselves in when lying will be sustained by the long-term goal of deceiving the prosecutor, the cues still being present in the participants' speech.

The dataset, in its complete and public form, consists of 4 hours and 46 minutes of total content, acquired from 20 speakers (4 female, 16 male) during testimonies, interviews and questionings conducted by Romanian law enforcement agencies, in which all participants were persons of interest (guilty parties, suspects, witnesses, etc.). Out of the total content duration, 3 hours and 28 minutes represent the participants' speech segments; 2 hours and 6 minutes (60.5%) represent the truthful content, while 1 hour and 22 minutes (39.5%) represent the untruthful content. The RODeCAR dataset is available upon request and can be obtained by following the instructions provided here: https://speed.pub.ro/downloads/paralinguistic-datasets/.

4.3. Chapter conclusions

In this chapter, the Romanian Deva Criminal Investigation Audio Recordings (RODeCAR) dataset was introduced: a dataset of truthful and untruthful speech, constructed by analyzing, processing, and cross-examining archived original criminal investigation recordings. The most important advantage to be leveraged when using this dataset is the casework nature of the content, i.e., all the speakers were suspects or witnesses in real criminal investigations, and all interactions were spontaneous and were part of actual law enforcement activity.

Chapter 5

Deceptive speech detection

This chapter covers the task of detecting untruthful statements in a subject's speech, also called *deceptive speech detection* (DSD). Parts of the content herein were published by the candidate as a conference paper [Mih21a] and as a journal article [Mih22a].

5.1. Background and related work

For DSD, previous research was carried out using traditional algorithmically extracted speech features and descriptors, including the mean and standard deviation of the pitch [Sen22], the MFCCs and their delta and delta-delta coefficients [Fat21a, Men17], jitter, the harmonic-to-noise ratio (HNR) [Jai16], or other acoustic and prosodic features [Men17, Vel19] based on the ComParE feature set. As for the ML and DL models employed, these include support vector machines (SVMs) [Jai16, Men17, Mon16, Sen22], random forests (RFs) [Men17, Sen22, Vel19, Zha20], FCNNs [Kop19, Men17, Sen22, Vel19], logistic regression [Kop19, Men17], or ensemble methods with multiple classifiers and average/majority voting [Vel19].

5.2. Voice activity detection as a subtask

One of the first tasks that must be addressed within a typical speech processing pipeline is voice activity detection (VAD) [Mih21a], which is used in this work as a subtask for DSD for computing several prosodic features. In this work, *utterances* are defined as the content of intervals of an audio signal in which speech is present, separated from other such speech intervals by pauses of at least 200 ms in duration.

The deep neural network (DNN) models investigated are FCNNs, LSTM-based RNNs, and CNNs, together with three optimized postprocessing techniques: hysteresis

thresholding, minimum duration filtering, and bilateral extension. The proposed FCNN-based VAD subsystems use two hidden layers and are coupled with traditional algorithmically extracted features, i.e., the energy, the zero-crossing rate (ZCR), the HNR, the normalized autocorrelation coefficient, and the first 13 MFCCs, which are grouped into several feature subsets. The RNN-based subsystems employ an LSTM layer, followed by fully-connected layers for the actual classification, and take as inputs the same feature sets. The CNN-based subsystems implement three pairs of convolutional and max-pooling layers, followed by fully-connected layers for classification. The raw time-domain samples or the frequency-domain representation of the signal, given by the 127-point Discrete Fourier Transform (DFT) are provided to the CNN. All features are computed at the frame level. The model's output will represent the probability that the frames include speech content. A sliding window encompassing 3 consecutive frames is used to average these probabilities and the resulting value is compared to a threshold. If the value is above the threshold, the window will be considered *positive* (containing speech). Consecutive *positive* windows determine the utterance start and stop times. To boost performance, hysteresis is used to obtain two separate thresholds, the higher one being involved when switching from negative to positive predictions. Additionally, if a resulting utterance has a shorter duration than a reference value, Δt_{min}, it is discarded. For the remaining predicted utterances, a bilateral extension of their durations is implemented to compensate for the subsystem's tendency to underestimate the utterance length. The extension consists of lowering the utterance start time by a value Δt_{ext}, while its stop time is increased by the same value.

Final testing was conducted on the real ambient noise subset of the Corpus and Environment for Noisy Speech Recognition (CENSREC-1-C) [Kit07] corpus, with detailed results being shown in Table 5.3, compared to other literature. The VAD subsystem was subsequently adapted for the DSD datasets.

Table 5.3 – VAD utterance-level top test accuracy [%] vs. model type: CENSREC-1-C.

Model	**Ambient noise type**				
	Restaurant		**Highway**		**Average**
	High SNR	**Low SNR**	**High SNR**	**Low SNR**	
CENSREC-1-C baseline [Kit07]	74.20	56.50	39.40	41.40	52.88
[Esp11]	76.75	63.02	92.44	79.64	77.96
[Fuj10]	92.75	65.51	**100.00**	**100.00**	**89.57**
[Fuj14]	75.65	21.45	95.94	49.86	60.73
FCNN	88.11	57.46	56.65	54.34	64.14
RNN	74.25	39.10	65.80	51.50	57.66
CNN-DFT1	85.21	64.92	82.60	75.65	77.10
CNN-DFT2	97.10	59.13	95.36	89.56	85.29
CNN-DFT3	**99.13**	**68.69**	**97.97**	**90.72**	**89.13**

5.3. Proposed system architectures

The basic deep learning system proposed for the main task, i.e., deceptive speech detection (DSD), consists of a DNN that takes as input a modified version of the extensive 2,258-dimensional feature set described in Chapter 2. The DNN classifier is an FCNN model, using between 2 and 3 hidden layers with 64 or 128 nodes per layer, but with an output layer whose size is equal to the number of classes taken into account, i.e., 2 (*truthful* and *deceptive*), and which uses the softmax activation function, instead of a single neuron applying the sigmoid activation function.

Regarding the modification to the feature set, it refers to having additionally included two utterance-wise prosodic features (UPFs) [Mih22a]: the utterance duration and the leading pause duration, i.e., the time interval between the end of the previous utterance and the start of the current one, both proven to be relevant for the DSD task.

A second DSD system is proposed, leveraging the nature of automatic feature extraction offered by CNNs. The model takes as input the magnitude spectrogram of each preprocessed utterance, extracted using Hamming windows of 25 ms duration with a 15 ms overlap, scaled linearly into 257 frequency bins (corresponding to half the sampling frequency). Subsequently, three stages of 2D convolutional layers are applied with a small receptive field, with max-pooling applied after each one to reduce the data dimensionality. The output of the final pooling layer represents the set of automatically extracted feature maps that are then flattened into a 1D vector, and passed through a sequence of fully-connected hidden layers. Together with the output layer, these form the actual classifier stages, and adopt a configuration (number of layers, number of nodes per layer) according to the best structure determined previously.

Lastly, in order to boost performance further, a final hybrid CNN-MLP network is proposed (the MLP being represented by the final set of fully-connected layers after the concatenation layer, i.e., the classifier head), combining the automatically extracted features provided by the convolutional stages with the best subset of algorithmically extracted features determined using the basic DNN-based DSD system. These features, extracted at the utterance level as described beforehand, are provided as an additional input, and are concatenated with the output of the flattening layer before being fed to the classifier (fully-connected) stage of the hybrid network.

5.4. Experimental setup and results

The Real-Life Trial Data for Deception Detection (RLDD) [Per15] dataset comprises 121 audio-visual recordings in English of defendants and witnesses, obtained from trials conducted in the United States, with 61 recordings being labeled as *deceptive* and 60 as *truthful*. The content duration totals 56 min, with an average recording length of 28 s. Excluding prosecutors, lawyers, and other interviewers, the total number of speakers is 56 (22 female, 34 male). The second dataset, RODeCAR, was described in Chapter 4.

Other research previously published on DSD for the RLDD dataset uses a speaker-level [Sen22] or a recording-level approach [Fat21a, Jai16, Vel19]. The former

involves determining the overall *truthful* / *deceptive* attitude of each speaker (56 instances), while the latter classifies each entire audio recording (121 instances) as *truthful* / *deceptive*. In this work, the main focus is on a different and more challenging 'local lie' (utterance-level) approach, i.e., determining which particular utterances in each recording are *truthful* vs. *deceptive*. To this end, from each audio recording available in the RLDD and RODeCAR datasets, all utterances were extracted, totaling 931 (467 *truthful* and 464 *deceptive*) and 5,859 (3,136 *truthful* and 2,723 *deceptive*).

Unless otherwise specified, for all DSD experiments, 10-fold cross-validation with speaker separation was employed as the testing methodology, with an 80% / 20% training-validation split, ensuring the same ratio of *truthful* and *deceptive* samples were available in each subset, as well as having the same ratio of male to female speakers.

The basic FCNN-based models were evaluated for a depth (number of hidden layers) between 2 and 3, with 64 or 128 nodes per hidden layer. Other hyperparameters chosen included: the ReLU activation function for the hidden layers; Adam as the optimization algorithm; L1-norm regularization with the regularization parameter equal to 10^{-4}. Since the RODeCAR dataset is slightly imbalanced in terms of class distribution (53.5% of the utterances are *truthful*), class weighting was employed.

The basic systems were tested for the total feature set of size 2,260 described in Section 5.3, as well as for several feature subsets. Finally, using a novel proposed feature selection algorithm based on the two-sample Kolmogorov-Smirnov test (KS), an additional 5 subsets were obtained by selecting the most relevant 10, 20, 50, 100, and 200 features, in terms of the KS statistic, resulting in a total number of 36 feature subsets. For the other two proposed DSD systems, i.e., based on a CNN model or using a hybrid CNN-MLP network, the input spectrograms must all be of the same size, thus requiring zero-padding to the duration (in number of frames) of the longest utterance in each dataset. Additionally, a dropout of 0.4 was used before each pooling layer, and L2-norm regularization was implemented instead. The fully-connected stage consists of two layers, with 32 nodes per layer. All other applicable hyperparameters follow the same configuration as for the basic, FCNN-based DSD system described previously.

In order to provide a comparison between the proposed model and the other DSD systems reported in literature, the utterance-level results were postprocessed on a macro level in order to correspond to the alternative speaker-level and recording-level results, as applicable. For the RLDD dataset (the only one for which such comparisons can be made, since no other results have been published for the RODeCAR dataset by external parties up to the date of writing this thesis), the following steps are taken:

- for the speaker-level approach, all utterances that belong to each speaker are grouped together and a majority vote is taken over the utterance-level labels predicted by the CNN-MLP model;
- for the recording-level approach, a similar step is performed, but per recording instead of per speaker.

For the RLDD dataset, the speaker-level results and the recording-level results are given in Table 5.11 and Table 5.12, respectively, in terms of the (weighted) accuracy. For RODeCAR, the performance for adapting the system to the speaker-level approach has also been determined to be 83.5%. The recording-level approach is inapplicable due

Table 5.11 – Speaker-level approach test accuracy [%] comparison vs. other works published in literature: RLDD.

System	Accuracy [%]
[Sen22] – RF	71.2
[Sen22] – MLP	61.0
This work: CNN-MLP; input: 1,183×257 / 2,392×257 linear magnitude spectrogram + 10 / 340 feat.; 16 kHz s.r.	**85.6**

Table 5.12 – Recording-level approach test accuracy [%] comparison vs. other works published in literature: RLDD.

System	Accuracy [%]
[Fat21a] – SVM	81.5
[Vel19] – Ensemble (KNN + RF + MLP)	70.0
[Jai16] – SVM	34.2
This work: CNN-MLP; input: 1,183×257 / 2,392×257 linear magnitude spectrogram + 10 / 340 feat.; 16 kHz s.r.	**88.6**

to the nature of the RODeCAR dataset since its individual files have very long durations (from tens of minutes up to one hour).

5.5. Chapter conclusions

In this chapter, several voice activity detection (VAD) subsystems based on deep neural networks (DNNs) were proposed, implemented, and validated. The VAD subsystem is employed for the extraction of utterance-wise prosodic features.

For the main DSD task, it was shown that the utterance-level approach is better suited than other speaker-level or recording-level approaches for forensic applications. Several neural network-based DSD systems were proposed, implemented, and validated. The highest-performing architecture was a novel hybrid CNN-MLP-based network, leveraging a fusion of automatically extracted feature maps and subsets of hand-crafted features, selected based on a novel proposed feature selection algorithm. Within the most relevant utterance-level ('local lie') approach, the system reaches an accuracy of **63.7%** on the RLDD dataset, and of **62.4%** on the RODeCAR dataset.

For RLDD, the speaker-level performance was **85.6%**, representing a 20.22% increase vs. other comparable systems reported in literature; and the recording-level performance was **88.6%**, a corresponding 8.71% increase.

For the RODeCAR dataset, the recording-level approach is incompatible, but the speaker-level approach led to an accuracy of **83.5%**.

Chapter 6

Speech emotion recognition for suspicious behavior monitoring

This chapter covers the task of determining the affective content present in speech, also called *speech emotion recognition* (SER). Parts of the content herein were published as a conference paper [Mih19a] and as a journal article [Mih21c] by the candidate. Parts of the content herein were supported by the candidate's participation as a research assistant in project PN-III-P2-2.1-SOL-2016-02-0002, agreement 2SOL/2017, funded by the Romanian Government through UEFISCDI: *Intelligent Systems for Video and Audio Analysis – Technologies and Innovative Video Systems for Person Re-identification and Analysis of Dissimulated Behavior* (SPIA-VA) [Mih20].

6.1. Background and related work

When designing SER systems, there are the two main schools of thought in psychology that establish the conceptual modeling of emotions:

- discrete classes [Laz99], wherein each emotion (or, rather, each emotion class) is holistically distinguished from the others; and
- dimensional models [Wat99], where a number of continuous value psychological measures (e.g., arousal or valence) form a multidimensional affect space (typically 2D), each emotion being a sub-zone within it.

In this sense, promising results have been reported in literature using machine learning (ML) and deep learning (DL) models and techniques, including hidden Markov models [Sha23b], support vector machines (SVMs) [Jin15], multilayer perceptron (MLP) DNNs [Atm20, Lat20a, Rao17], recurrent neural networks (RNNs) with long short-term memory (LSTM) cells [Gha19, Liu20, Mir17], convolutional neural

networks (CNNs) [Tan21, Zha18a], hybrid models [Fah20], or advanced convolutional-recurrent neural networks (CRNNs) [Che18, Zha19], using either algorithmic or automatic ("true deep learning") feature extraction.

6.2. Dimensional models for continuous-to-discrete affect mapping

Since conventional dimensional models derived in the field of psychology only allow a vague "mapping" of the affect space, without exact numerical positioning or delineation of the corresponding emotion classes, more precise (yet compact and low-complexity) models would prove useful to be constructed using ML techniques. These models would allow determining the emotion class of an instance based on its continuous affect space position (e.g., its *arousal* and *valence* values). An alternative approach consists of adopting different multidomain strategies, in which the discrete and continuous paradigms are directly tied together through a-priori mapping [Mih21c].

The main source of data for fitting the dimensional model would be a dataset with dual discrete and continuous annotation of emotional content (labels for emotion classes, and numerical values for the affective dimensions). The only such available corpus is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [Bus08], having almost exclusively been used for emotion classification. An additional source can be used, such as the Warriner-Kuperman-Brysbaert (WKB) corpus [War13], which includes affective dimension annotations for a number of words, the relevant ones being those representing the emotion classes, such as "anger" (i.e., the "concept of anger"), etc. In the proposed approach, these annotated values are leveraged to initialize the class centroids (means). The reasoning is that including the WKB corpus greatly increases reliability and leads to better generalization.

Three ML algorithms were tested for developing the dimensional model for affect mapping: K-means clustering (referred to as the K-means model, KMM), GMM fitting and SVM fitting. For the KMM and GMM approaches, the class centroids (means) were initialized in one of two ways: (i) using the values estimated from averaging over the IEMOCAP data (native initialization); (ii) using the values estimated from the WKB data (WKB initialization). The first option allows for better data fitting, but, for the second, model generalization is greater. SVMs lead to even better results thanks to the higher-dimensional transformation of the affect space, but can only use IEMOCAP data. In all experiments, 5-fold cross-validation was employed, reserving one speaker session for testing (i.e., 20% of the data). The results are given using both the unweighted accuracy (UA) and the weighted accuracy (WA) as metrics.

In Table 6.2, the proposed dimensional model mapping approach is compared to other works using standard classification systems for discrete emotions. As can be seen, dimensional models can lead to best performance, as long as reliable dimensional affective dimension data exists; in other words, if the overall affect space coordinates of speech segments can be correctly predicted by a regression model.

Table 6.2 – Maximum performance comparison to other works using standard classification systems for discrete emotions.

Context	Best results
[Che18]	UA = 64.7%
[Fah20]	UA = 66.0%, WA = 70.5%
[Jin15]	WA = 68.6%
[Lat20a]	UA = 61.0%
[Liu20]	UA = 65.0%, WA = 66.1%
[Mir17]	UA = 58.8%, WA = 63.5%
[Zha18a]	UA = 63.9%, WA = 70.4%
[Zha19]	UA = 67.0%, WA = 68.1%
Dimensional model (DM) mapping	**UA = 74.3%, WA = 72.5%**

6.3. Proposed system architectures

In this section, multiple complete system architectures to be employed for SER are presented, including types based on multidomain strategies leveraging the dimensional model mapping concept detailed in Section 6.2.

The proposed system architectures for the SER task were iteratively developed, and present increasing levels of complexity, falling within six categories (*approaches*):

1) single DNN models, for classification or regression;
2) ensemble classification strategies comprising multiple DNN models;
3) multidomain systems, uniting the two emotion recognition paradigms;
4) single DNN classification models adapted through transfer learning (TL);
5) heterogeneous fusion classification through TL-DNN models;
6) homogeneous ensemble classification using TL-DNN models.

The single DNN model approach takes as input the extensive 2,258-dimensional set of acoustic, spectral, and cepstral features. The DNN classifier is a feed-forward fully-connected neural network (FCNN) model, using between 1 and 4 hidden layers, with different numbers of nodes per layer, and with an output layer of size equal to either the number of classes applicable or having two neurons, corresponding to the 2D affect space dimensions. The second, more advanced type of system revisits the ensemble classification approach presented in Chapter 3, extending it to the two main strategies adopted by SVM models for multiclass problems: one-vs.-one (OvO) and one-vs.-rest (OvR). For the multidomain systems, seven proposed forms are defined:

- type 1: the native form of joint learning;
- type 2: the first active layers are common for the two tasks, then splitting into an output layer for one of the tasks and a second active section used only for the other task – (A) regression is modeled only by the first section, classification is modeled by both; (B) classification is modeled only by the first section, regression is modeled by both;

- type 3: the first active section is focused on one of the tasks, with the second active section being trained for the other task on the same initial input data together with the outputs of the first active section – (A) regression is modeled by the first section, classification is modeled by the second section; (B) classification is modeled by the first section, regression is modeled by the second section;
- type 4: these sequential approaches leverage pretrained dimensional model (DM) mappings as described in Section 6.2 to establish the link between the continuous affect space and the position within (determined by regression) and the emotion class – (A) direct application of the DM is used to determine the emotion class based on the output of the DNN regression model; (B) a second DNN is trained on the initial input data together with the preliminary classification provided by the DM.

Transfer learning (TL) is a deep learning (DL) technique that leverages the data space transformations corresponding to a task for which DNNs were designed by adapting them (through retraining) for a different, but related task. In the context of SER, this is achieved by adopting very deep, high-performing image recognition models and representing data instances in a form compatible with images, e.g., spectrograms. The modern top-performing image recognition DNNs (hereafter denoted as TL-DNNs) are: Xception, VGG16 and VGG19, ResNet50, ResNet50V2, InceptionV3, InceptionResNetV2, NASNetMobile and NASNetLarge, and EfficientNetB0 through EfficientNetB7, trained on the ImageNet dataset.

In this work, the first proposed TL-based approach consists of retraining the top layers of each TL-DNNs in order to develop single DNN classification models. Going further, a form of ensemble information representation through fusion is proposed in the form of a heterogeneous TL-DNN system: the core of each of the TL-DNN models is used to extract deep feature map representations of the input data. All representations are then flattened and concatenated into a single FV that is subsequently fed to a DNN classifier. Finally, the homogeneous TL-based approach, unlike the heterogeneous system, this architecture leverages the ensemble classification strategies (OvO and OvR), but with TL-DNN models. The homogeneity property refers to the fact that, for each class combination, the same TL-DNN model is used within a single system.

The three TL-DNN-based approaches for SER systems are denoted as the fourth, fifth, and sixth overall approach for SER. For the TL-DNN-based approaches, the input given to the networks must be in the form of spectrograms. These were extracted as linear or log magnitude spectrograms using Hamming windows of 25 ms duration (15 ms overlap), with linear or Mel scaling, and 3 color (RGB) channels.

6.4. Experimental setup and results

The Berlin Database of Emotional Speech (EMODB) [Bur05] is a German language dataset comprising 535 short utterances recorded by 10 actors (5 female, 5 male)

specially chosen by an expert jury with spoken language naturalness and recognizability as the main criteria. The utterances have an average duration of 2.5 s and a maximum duration of 8 s. The 7 emotion classes considered are: Anger (ANG), Disgust (DIS), Fear (FEA), Sadness (SAD), Boredom (BOR), Happiness (HAP) and Neutral (NEU).

Since the focus of this work is on forensic and law enforcement applications, particularly suspicious behavior monitoring, negative emotion classes are more relevant and important to detect (individually, for applications requiring more detail and nuance), as well as negative affective manifestations in general (considered together as a single group, without subdivisions, as an overall monitored class). Apart from the full set of 7 classes, three additional subsets were considered. The 4 considered subsets are:

- **EMODB-7**: all 7 classes: ANG, DIS, FEA, SAD, BOR, HAP, and NEU;
- **EMODB-5N**: 5 classes: ANG, DIS, FEA, SAD, and NEU;
- **EMODB-4**: 4 classes: ANG, SAD, HAP, and NEU;
- **EMODB-2N**: 2 classes: Negative (NEG; grouping together ANG, DIS, FEA, and SAD) vs. NEU.

The Crowd-sourced Emotional Multimodal Actors Dataset (CREMAD) [Cao14] is an English language dataset comprising 7,442 recordings of facial and vocal affective content manifested in sentences spoken by 91 directed actors (43 female, 48 male). The encompassed 6 emotion classes were: Anger (ANG), Disgust (DIS), Fear (FEA), Sadness (SAD), Happiness (HAP), and Neutral (NEU). The average duration of the recordings is 2.5 s. The 4 subsets utilized in this work are:

- **CREMAD-6**: all 6 classes: ANG, DIS, FEA, SAD, HAP, and NEU;
- **CREMAD-5N**: 5 classes: ANG, DIS, FEA, SAD, and NEU;
- **CREMAD-4**: 4 classes: ANG, SAD, HAP, and NEU;
- **CREMAD-2N**: 2 classes: Negative (NEG; grouping together ANG, DIS, FEA, and SAD) vs. NEU.

The IEMOCAP dataset [Bus08] includes 10 actors (5 female, 5 male) working in pairs to solve scripted and improvised English speaking tasks, with a total number of 10,039 audio-visual recordings. A total of 10 discrete emotion classes (Anger, Fear, Disgust, Sadness, Happiness, Frustration, Excitement, Surprise, Neutral, and Other) are available, but with many strongly underrepresented. This results in having to group only a smaller subset of them into 4 new classes, i.e., Neutral (NEU); Sadness (SAD); Anger + Frustration (ANG); and Happiness + Excitement (HAP). For the continuous dimensions, *arousal* and *valence* were chosen. The 2 considered subsets are:

- **IEMOCAP-4**: 4 classes: ANG, HAP, SAD, and NEU;
- **IEMOCAP-2N**: 2 classes: Negative (NEG; grouping together ANG and SAD) vs. NEU.

For *Approach 1* (single DNN models), the number of hidden layers was chosen between 1 and 4, with the number of neurons for the first hidden layer being chosen from the set {8, 16, 32, 64, 128, 256, 512, 1024}, and varying the dropout rate between 0.1 and 0.5. Other hyperparameters chosen included: the rectified linear unit (ReLU) activation function for the hidden layers, and either the softmax (for classification) or the linear output activation function (for regression). The same configurations were

Table 6.13a – Performance comparison between the best results for SER classification achieved in this work and other relevant results published in literature.

Data subset	System	Perf.	
		UA [%]	WA [%]
EMODB-7	[Ker19] – SVM + recursive feature elimination	–	**86.2**
	[Che18] – CRNN	–	82.8
	[Lot17] – SNN + LSM + Gammatone filterbank	–	82.4
	[Bis13] – SVM + gender recognition	–	81.5
	[Cha14] – GMM	79.8	–
	[Yil21] – SVM + feature selection	78.6	79.1
	[Kan21] – GA + clustering	77.5	–
	[Cha14] – SVM	77.0	–
	This work: *Approach 2* – ensemble classification (OvR) with multiple DNNs (FCNNs).	**82.6**	**82.9**
EMODB-4	[Vas15] – GMM + SVM	–	84.3
	This work: *Approach 2* – ensemble classification (OvR) with multiple DNNs (FCNNs).	**88.9**	**89.1**
EMODB-5N	[He15] – MLP + GA-based modified backpropagation	–	80.4
	This work: *Approach 2* – ensemble classification (OvR) with multiple DNNs (FCNNs).	**91.2**	**91.4**
EMODB-2N	[Cas08] – SVM	–	95.8
	[Vas15] – GMM + SVM	–	94.9
	This work: *Approach 1* – single DNN (FCNN) classifier.	**95.1**	**98.3**
CREMAD-6	[Gha20] – SVM	–	**57.2**
	[Gha19] – LSTM	–	**57.0**
	[Bea18] – LSTM	–	41.5
	This work: *Approach 6* – homogeneous ensemble classification (OvO) with multiple TL-DNN models (EfficientNetB1).	**51.8**	**54.6**
CREMAD-4	**This work**: *Approach 6* – homogeneous ensemble classification (OvO) with multiple TL-DNN models (EfficientNetB1).	**65.8**	**70.3**
CREMAD-5N	**This work**: *Approach 6* – homogeneous ensemble classification (OvO) with multiple TL-DNN models (EfficientNetB0).	**54.7**	**58.7**
CREMAD-2N	**This work**: *Approach 1* – single DNN (FCNN) classifier.	**72.8**	**72.6**
IEMOCAP-4	[Yi22] – DNN + adversarial data augmentation	**63.7**	**63.2**
	[Lat20a] – MLP + GAN-based synthetic data	**61.0**	–
	[Yi22] – SVM + adversarial data augmentation	**60.0**	**64.7**
	[Yil21] – SVM + feature selection	**59.4**	59.5
	[Mir17] – MLP + LSTM + attention	58.7	**63.5**
	[Pan20] – LSTM	48.7	57.1
	[Rao17] – MLP + i-vectors	–	48.8
	This work: *Approach 2* – ensemble classification (OvO) with multiple DNNs (FCNNs).	**58.7**	**61.6**
IEMOCAP-2N	[Rah12] – SVM + feature adaptation	–	69.8
	This work: *Approach 1* – single DNN (FCNN) classifier.	**69.0**	**71.2**

Table 6.13b – Performance comparison between the best results for SER regression achieved in this work and other relevant results published in literature.
The results are separated for each affective dimension: A = arousal, V = valence.

Dataset	System	Performance					
		MSE (loss)		ρ		ρ_c	
		A	V	A	V	A	V
IEMOCAP	[Atm20] – MLP	–	–	–	–	0.611	0.301
	[Zha18b] – DNN	–	–	–	–	0.392	**0.715**
	This work: *Approach 1* – single DNN (FCNN).	**0.073**	**0.180**	**0.677**	**0.408**	**0.621**	**0.343**

tested afterwards for the OvO and OvR DNN classifiers (*Approach 2*), with the final DNN classifier having a fixed depth of either 1, 2, or 3 layers. The configurations were also employed for the several types of multidomain systems (*Approach 3*), as well as for the fully-connected classification heads in the TL-DNN experiments.

For all experiments, 10-fold cross-validation was employed as the testing methodology, with an 80% / 20% training-validation split, ensuring as best as possible that each emotion class and each gender was proportionally represented in each training and/or validation subset. Speaker separation was ensured for all experiments.

Performance comparisons between the proposed systems and others reported in literature are made in Table 6.13a and Table 6.13b.

6.5. Chapter conclusions

In this chapter, a detailed introduction into the SER task and its challenges was given, establishing the two fundamental emotion modeling paradigms. SER systems based on deep neural networks (DNNs) spanning six levels of complexity were proposed, developed, and tested: single DNNs, multiple DNNs connected together following ensemble classification strategies (one-vs.-one, OvO, and one-vs.-rest, OvR), and systems leveraging transfer learning (TL) for the top modern image recognition deep learning models, either as standalone TL-DNN models or as heterogeneous or homogeneous ensemble classifiers. The systems were tested on the most relevant SER datasets available: EMODB, CREMAD, and IEMOCAP, for the standard full set of classes, as well as for additional negative emotion subsets relevant for suspicious behavior monitoring and other applications that fall within the scope of this work.

The proposed systems achieved state-of-the-art results (up to **83%** accuracy) for the EMODB all-class subset, while the performance on the corresponding CREMAD and IEMOCAP subsets was lesser (up to **55%** accuracy for CREMAD and **62%** accuracy for IEMOCAP), but still comparable to other published research. Additionally, for all negative-emotion-only subsets, the proposed solutions offered top performance.

Chapter 7

Speech emotion remanence

This chapter covers the study of *speech emotion remanence* in the context of forensic speech, and how speech emotion recognition (SER) can be applied to the topic. Parts of the content herein were published as a journal article by the candidate [Mih22b].

7.1. Background and related work

Beyond the performance of SER systems themselves, one of the challenges presented by these tasks consists of discerning patterns in the temporal evolution of the affective content that would indicate suspicious behavior. To this end, the temporal evolution of the affective content of speech samples was analyzed, using a novel proprietary dataset, on a short (within 1 hour) and on a longer timescale (over 5 days) [Mih22b].

Previous research [Liu21, Su21, Zha21a, Zha22] has shown that ML and DL models still perform relatively poorly cross-corpus, i.e., when evaluating the model on different corpora than the ones it was trained on, even when using advanced and costly techniques for input data adaptation. This reduced generalization power may be caused, at least in part, by differences in emotional expression vs. the culture, background, age, etc. of the speaker, but there exists no conclusive evidence for or against this idea.

7.2. Study on speech emotion remanence

The two main hypotheses of this study were the following [Mih22b]:

1) If a human interaction is emotionally triggering for the subject, then their affective response will not decay instantly after the interaction ends, but over a longer time period, and subsequent emotionally neutral interactions will still be accompanied by an aroused negative affective state.

2) In the context of the existence of a forthcoming emotionally charged event for the subject (and of which they are aware), as the event approaches, the subject will experience higher intensity emotions and will exhibit a correspondingly increased affective response.

To test both hypotheses, a dataset was constructed using recordings of recurrent spoken interactions with a number of students who were behind on their university exams, and were studying in order to attempt them for the second or third time. There were 18 students (4 female, 14 male) involved in the process, of ages between 19.7 years and 23.3 years. The total number of recorded utterances was 270, and the total duration of the dataset's speech content is 1 hour and 8 minutes.

In order to validate the applicability of automatic speech emotion recognition within this context, systems based on fully-connected neural networks (FCNNs) were developed, following the same approach illustrated in Chapter 6. Two hidden layer node structures were taken into consideration: the 'constant' architecture, consisting of the same number of nodes for each hidden layer; and the 'log2dec' architecture, consisting of a progressively smaller number of nodes per layer, following a $\log_2(\cdot)$ law. The depth varied between 2 and 4, with an initial number of nodes of 256, 128, 64 or 32. Dropout was included after each hidden layer, with a rate between 20% and 50%. Other hyperparameters chosen include: the rectified linear unit (ReLU) activation function for the hidden layers, and the softmax (for classification) or identity (for regression) activation function for the output layer.

The data is considered vs. each timestamp in Figure 7.2, with the labels referring to the initial affective response, and the affective response after 15 and 30 minutes of neutral conversation, respectively. For the regression problem, the values for arousal and valence vs. each day are represented in Figure 7.3 for each timestamp: for each speaker individually (thin lines) and the mean over all speakers (thick lines).

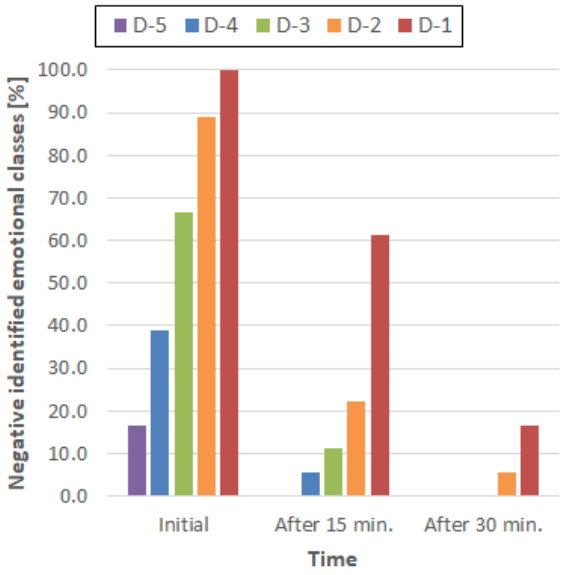

Figure 7.2 – Percentage ratio of speakers identified as expressing *negative* emotions.

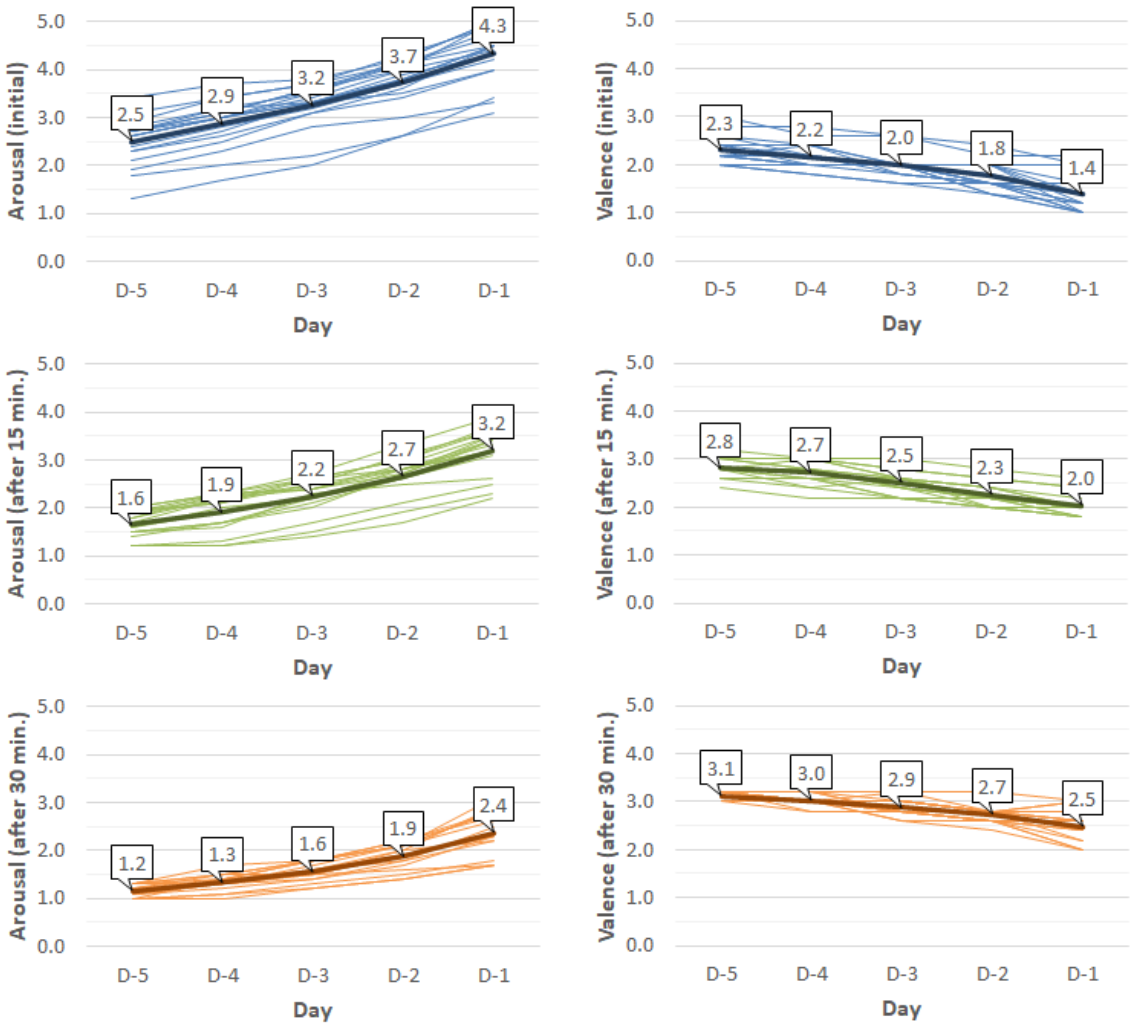

Figure 7.3 – Arousal and valence evolution vs. day, for each timestamp. Thin lines represent individual speaker evolutions, while the thick labeled data lines represent the average values for all speakers.

7.3. Chapter conclusions

In this chapter, insight was gained into speech emotion remanence by investigating short (under 1 hour) and long (5 day) timescales, different than the ones used in other speech emotion recognition research, more relevant for the targeted applications.

It was proven that: (1) if a human interaction is emotionally triggering for the subject, then their affective response will not decay instantly, but over a longer time period, and subsequent emotionally neutral interactions will still be accompanied by an aroused negative affective state; and (2) if an emotionally charged event is forthcoming for the subject, as the event draws closer, the subject will experience higher intensity emotions and will exhibit a correspondingly increased affective response.

Chapter 8

Conclusions

8.1. Developments and obtained results

In Chapter 1, the scope and objectives were defined, covering the development of AI systems for automatic recognition of paralinguistic elements using only speech data, with a focus on manifestations of negative emotions, high psychological stress levels, and engagement in deceptive behavior, the main application area being forensic speech.

Chapter 2 provided a summary of the main theoretical concepts employed during the development of this work in terms of the algorithmically extracted speech features used for analysis and processing, the models adopted in developing the systems, and the training and testing methodologies, techniques, and performance metrics.

In Chapter 3, systems were proposed and developed for speech under stress detection (SSD), employing ensembles of feed-forward fully-connected DNNs. Performance increases on the SUSAS dataset have been obtained over previously reported state-of-the-art results, with an (unweighted / weighted) accuracy of **68.8% / 65.5%** for the 4-class *actual* stress case, **59.2% / 62.4%** for the 3-class *actual* stress case, **66.7% / 81.4%** for the 2-class *actual* stress case, **75.5% / 75.5%** for the 4-class *simulated* stress case, and **76.1% / 78.4%** for the 4-class *actual* stress case.

Chapter 4 focused on introducing the Romanian Deva Criminal Investigation Audio Recordings (RODeCAR) database: a dataset of truthful and untruthful speech, constructed by the candidate by analyzing, processing, and cross-examining archived original criminal investigation recordings. The dataset consists of **3 h 28 min** of speech segments acquired from 20 speakers (4 female, 16 male): 2 h 6 min truthful (60.5% of the dataset), 1 h 22 min untruthful (39.5% of the dataset), objectively annotated.

In Chapter 5, the main task is developing systems for DSD. However, for the proposed DSD processing pipeline, a voice activity detection subsystem is required, which is first developed and discussed. After showing that the utterance-level approach

is better suited for forensic and law enforcement applications than other speaker-level or recording-level approaches, four neural network-based DSD systems were proposed, implemented, and validated. The best-performing architecture was a novel hybrid network. Within the utterance-level approach, the system reaches an accuracy of **63.7%** on the RLDD dataset, and of **62.4%** on the RODeCAR dataset. The proposed system was also tested to determine its speaker-level and recording-level performance. For RLDD, the speaker-level performance was **85.6%**, representing a 20.22% increase vs. other comparable systems reported in literature; and the recording-level performance was **88.6%**, a corresponding 8.71% increase. The recording-level approach was not compatible with RODeCAR; the speaker-level approach led to an accuracy of **83.5%**.

Chapter 6 provided a detailed introduction into the speech emotion recognition (SER) task and its challenges. DNN-based systems spanning six levels of complexity were proposed, developed, and tested, including single DNNs, multiple DNNs connected together following ensemble classification strategies (one-vs.-one, OvO, and one-vs.-rest, OvR), as well as systems leveraging transfer learning (TL) for the top modern image recognition deep learning models, either as standalone TL-DNN models or as heterogeneous or homogeneous ensemble classifiers. The systems were tested on the most relevant SER datasets: EMODB, CREMAD, and IEMOCAP. The proposed systems achieved state-of-the-art results (up to **83%** accuracy) for the EMODB all-class subset, while the performance on the corresponding CREMAD and IEMOCAP subsets was lesser (up to **55%** accuracy for CREMAD and **62%** accuracy for IEMOCAP), but still comparable to other published research. Additionally, for all subsets comprising only negative affective content, the proposed solutions offered the top performance.

Lastly, Chapter 7 served as a follow-up on SER, offering a deep dive into speech emotion remanence, investigating short (under 1 h) and long (5 day) timescales, which are more relevant for the applications within the scope of this work. It was proven that: (1) if a human interaction is emotionally triggering for the subject, then their affective response will not decay instantly, but over a longer time period, and subsequent emotionally neutral interactions will still be accompanied by an aroused negative affective state (emotional remanence); and (2) if an emotionally charged event is forthcoming for the subject, as the event draws closer, the subject will experience higher intensity emotions and will exhibit a correspondingly increased affective response.

8.2. Original contributions

General and global contributions

- Classification strategies using ensembles of neural networks in OvO and OvR configurations were developed and leveraged successfully for SSD, DSD, and SER. The description of the strategies was given in Chapter 3, and results obtained employing them were detailed in Chapter 3 and Chapter 5, and published in [Mih21b, Mih22a].

- A set of algorithmically extracted features for automatic paralinguistic element recognition tasks was developed, based on extending the most used reference feature set available in literature with several additional features. In all tasks (SSD, DSD, SER), the proposed feature set was employed successfully. Described in Chapter 2, the results obtained with neural networks trained on it were discussed in Chapter 3, Chapter 5, and Chapter 6, and published in [Mih21b, Mih22a].

- A feature selection algorithm for binary classification problems based on the two-sample Kolmogorov-Smirnov test was proposed and applied for DSD. The description of the algorithm was given in Chapter 5, and the results obtained incorporating it into the DSD systems were published in [Mih22a].

- A voice activity detection (VAD) subsystem was developed and employed to extract prosodic features used for DSD. The description of the system was given in Chapter 5, and the results were published in [Mih21a].

Speech under stress detection (SSD) contributions

- Novel approaches were employed for SSD in the context of forensic speech applications, using class groupings and analyses specific to the scope of this work. The results obtained demonstrated improved performance over most of the state-of-the-art literature previously published in the field. The results were presented in Chapter 3 and published in [Mih21b].

Deceptive speech detection (DSD) contributions

- The Romanian Deva Criminal Investigation Audio Recordings (RODeCAR) dataset was developed end-to-end for DSD tasks. Its complete description was given in Chapter 4 and published in [Mih19b]. This novel dataset is:
  i) one of the very few publicly available datasets that provides reliable, realistic, objectively annotated data for DSD by leveraging recordings with non-simulated behavior in realistic high-stakes scenarios;
  ii) the only such dataset available for the Romanian language;
  iii) a consistent database for paralinguistic applications, particularly DSD, comprising approximately 3.5 hours of spoken content.

- A novel approach, more challenging, more detailed, and better suited for forensic and law enforcement applications, was employed for DSD by training the proposed systems to discern between truthful and deceptive speech at the utterance-level (i.e., short-term) instead of the recording-level (i.e., long-term) or the speaker-level (i.e., overall profiling):

i) the reasoning was explained in Chapter 5 and published in [Mih22a], representing the first published results using the utterance-level proposed approach;
ii) the results obtained at the recording-level and speaker-level were also given in Chapter 5, and represent significant performance increases over the state-of-the-art literature previously published.

- Hybrid deep neural network architectures were developed for DSD, combining the automatic feature extraction function of convolutional neural networks with the relevance of well-chosen algorithmically extracted (i.e., hand-crafted) features. The detailed description of the hybrid architectures was illustrated in Chapter 5, and the results obtained with this approach were published in [Mih21a, Mih22a].

Speech emotion recognition (SER) contributions

- A theoretical and experimental investigation of speech emotion remanence was conducted to validate two important hypotheses for law enforcement and forensic applications: (1) affective responses to emotionally charged events decay over long time periods, with subsequent neutral interactions being accompanied by aroused negative affective states; and (2) imminent emotionally charged events determine higher intensity emotions and manifestations of correspondingly increased affective responses. It was presented in Chapter 7 and published in [Mih22b]. The applied study:
  i) is one of the few studies conducted on the topic, and the only study performed at the timescales chosen (within 1 hour and over 5 days), which were justified as the most relevant for the targeted applications;
  ii) included experimental validation of using SER systems to monitor the emotional manifestations and their temporal evolution relevant for the envisaged applications.

- Improved dimensional model mappings were developed for SER, refining the link between the discrete emotional class paradigm and the continuous affect space modeling paradigm for emotion analysis and recognition. They were discussed in Chapter 6 and published in [Mih21c]. The models:
  i) were developed by correlating the data in one of the few available dually-annotated SER datasets, also one of the most often cited and used in recent research in the field;
  ii) were improved through multimodality, by refining the initial versions obtained based on the audio data with relevant emotional class textual data from a large corpus.

- Novel approaches were employed for SER in the context of forensic speech applications, with a focus on negative emotions as specific to the scope of

this work. Part of the obtained results demonstrated improved performance over most of the state-of-the-art literature previously published in the field. The results were presented in Chapter 6 and partially published in [Mih19a].

- Systems based on transfer learning were developed for SER using each of the best-performing modern image recognition deep neural networks (VGG16, VGG19, Inception, Xception, and several versions of ResNet50, NASNet, and EfficientNet), both standalone and via ensemble classification strategies. The methodology and results were presented in Chapter 6.

Research acknowledgment

Parts of this work were supported by the candidate's participation between 2017 and 2020 as a research assistant in project PN-III-P2-2.1-SOL-2016-02-0002, agreement 2SOL/2017, funded by the Romanian Government through UEFISCDI: *Intelligent Systems for Video and Audio Analysis – Technologies and Innovative Video Systems for Person Re-identification and Analysis of Dissimulated Behavior* (SPIA-VA) [Mih20].

8.3. List of original publications

Within the scope of this doctorate, 3 journal articles (one ranked Q1, one ranked Q2) and 4 conference papers were published by the candidate as the first author.

Journal articles

1) **Ș. Mihalache** and D. Burileanu, "Dimensional models for continuous-to-discrete affect mapping in speech emotion recognition," in *University Politehnica of Bucharest Scientific Bulletin, Series C – Electrical Engineering and Computer Science*, vol. 83, iss. 4, Politehnica Press, Bucharest, pp. 137-148, Dec. 2021. ISSN: 2286-3540. [Mih21c]
ISI WOS: 000741473700013 (Q4, IF: 0.3)

2) **Ș. Mihalache** and D. Burileanu, "Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection," in *Sensors*, vol. 22, iss. 3, art. no. 1228, MDPI, Basel, Switzerland, pp. 1-21, Feb. 2022. ISSN: 1424-8220. DOI: 10.3390/s22031228. [Mih22a]
ISI WOS: 000759983300001 (Q1, IF: 3.576 – Feb. 2022)

3) **Ș. Mihalache**, D. Burileanu, E. Franți, M. Dascălu, and C.A. Brătan, "Lasting emotions – An investigation of short- and long-term affective content remanence in speech," in *Romanian Journal of Information Science and Technology*, vol. 25, iss. 1, Publishing House of the Romanian Academy, Bucharest, pp. 20-35, Mar. 2022. ISSN: 1453-8245. [Mih22b]
ISI WOS: 000775912300002 (Q2, IF: 3.5)

Conference papers

4) **Ș. Mihalache**, D. Burileanu, G. Pop, and C. Burileanu, "Modulation-based speech emotion recognition with reconstruction error feature expansion," in *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timișoara, Romania, pp. 1-6, Oct. 2019, IEEE NY. ISBN: 978-1-7281-0983-1. DOI: 10.1109/SPED.2019.8906537. [Mih19a]
ISI WOS: 000571718700004

5) **Ș. Mihalache**, G. Pop, and D. Burileanu, "Introducing the RODeCAR database for deceptive speech detection," in *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timișoara, Romania, pp. 1-6, Oct. 2019, IEEE NY. ISBN: 978-1-7281-0983-1. DOI: 10.1109/SPED.2019.8906542. [Mih19b]
ISI WOS: 000571718700006

6) **Ș. Mihalache**, I.A. Ivanov, and D. Burileanu, "Deep neural networks for voice activity detection," in *Proc. International Conference on Telecommunications and Signal Processing* (*TSP*), Brno, Czech Republic, pp. 191-194, Jul. 2021, IEEE NY. ISBN: 978-1-6654-2933-7. DOI: 10.1109/TSP52935.2021.9522670. [Mih21a]
ISI WOS: 000701604600041

7) **Ș. Mihalache**, D. Burileanu, and C. Burileanu, "Detecting psychological stress from speech using deep neural networks and ensemble classifiers," in *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Bucharest, Romania, pp. 74-79, Oct. 2021, IEEE NY. ISBN: 978-1-6654-2786-9. DOI: 10.1109/SpeD53181.2021.9587430. [Mih21b]
ISI WOS: 000786794700014

8.4. Perspectives for further developments

Regarding the candidate's future research and work in the fields of machine learning, deep learning, speech analysis and processing, automatic recognition of paralinguistic elements, and forensic speech applications, there are several promising avenues available and of continued great interest and relevance.

For each of the automatic paralinguistic element recognition tasks within the scope of this work, i.e., SSD, DSD, and SER, further improvements could be attained by leveraging different conventional models such as stacked autoencoders (SAEs), or more extreme forms such as extreme learning machines (ELMs). Additionally, beyond investigating alternative models, an additional attractive idea is to adapt techniques specific to other types of deep neural networks, e.g., the attention mechanisms used in recurrent neural networks (RNNs) or in transformers. Particularly for the DSD task, a multimodal approach involving both audio and textual data will likely prove to lead to better discerning *vocal lie detection* systems. For SER, further research is required into developing language-independent automatic recognition systems.

References

[Atm20] B.T. Atmaja and M. Akagi, “Deep multilayer perceptrons for dimensional speech emotion recognition,” in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Auckland, New Zealand, pp. 325-331, Dec. 2020.

[Avi19] A.R. Avila et al., “Speech-based stress classification based on modulation spectral features and convolutional neural networks,” in *Proc. European Signal Processing Conference* (*EUSIPCO*), A Coruna, Spain, pp. 1-5, Sep. 2019.

[Bac95] J.A. Bachorowski and M.J. Owren, “Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context,” in *Psychological Science*, vol. 6, iss. 4, pp. 219-224, Jul. 1995.

[Bac99] J.A. Bachorowski, “Vocal expression and perception of emotion,” in *Current Directions in Psychological Science*, vol. 8, iss. 2, pp. 53-57, Apr. 1999.

[Bea18] R. Beard et al., “Multi-modal sequence fusion via recursive attention for emotion recognition,” in *Proc. Conference on Computational Natural Language Learning* (*CoNLL*), Brussels, Belgium, pp. 251-259, Oct. 2018.

[Bes16] S. Besbes and Z. Lachiri, “Multi-class SVM for stressed speech recognition,” in *Proc. International Conference on Advanced Technologies for Signal and Image Processing* (*ATSIP*), Monastir, Tunisia, pp. 782-787, Mar. 2016.

[Bis06] C. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., Springer, New York, New York, United States of America, 2006.

[Bis13] I. Bisio et al., “Gender-driven emotion recognition through speech signals for ambient intelligence applications,” in *IEEE Transactions on Emerging Topics in Computing*, vol. 1, no. 2, pp. 244-257, Dec. 2013.

[Bur05] F. Burkhardt et al., “A database of German emotional speech,” in *Proc. INTERSPEECH*, Lisbon, Portugal, pp. 1517-1520, Sep. 2005.

[Bus08] C. Busso et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” in *Language Resources & Evaluation*, vol. 42, no. 4, art. no. 335, Nov. 2008.

[Cao14] H. Cao et al., “CREMA-D: crowd-sourced emotional multimodal actors dataset,” in *IEEE Transactions on Affective Computing*, vol. 5, iss. 4, pp. 377-390, Dec. 2014.

[Cas06] S. Casale, A. Russo, and S. Serrano, “Classification of speech under stress using features selected by genetic algorithms,” in *Proc. European Signal Processing Conference* (*EUSIPCO*), Florence, Italy, pp. 1-5, Sep. 2006.

[Cas08] S. Casale et al., “Speech emotion classification using machine learning algorithms,” in *Proc. IEEE International Conference on Semantic Computing*, Santa Monica, California, United States of America, pp. 158-165, Aug. 2008.

[Cha14] T. Chaspari, D. Dimitriadis, and P. Maragos, “Emotion classification of speech using modulation features,” in *Proc. European Signal Processing Conference* (*EUSIPCO*), Lisbon, Portugal, pp. 1552-1556, Sep. 2014.

[Che18] M. Chen et al., “3-D convolutional recurrent neural networks with attention model for speech emotion recognition,” in *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440-1444, Oct. 2018.

[Esp11] M. Espi, “Using spectral fluctuation of speech in multi-feature HMM-based voice activity detection,” in *Proc. INTERSPEECH*, Florence, Italy, pp. 2613-2616, Aug. 2011.

[Fah20] S. Fahad et al., “DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features,” in *Circuits, Systems, and Signal Processing*, vol. 40, iss. 1, pp. 466-489, Jul. 2020.

[Fat21a] E.P. Fathima Bareeda, B.S. Shajee Mohan, and K.V. Ahammed Muneer, “Lie detection using speech processing techniques,” in *Journal of Physics: Conference Series*, vol. 1921, pp. 12-28, Mar. 2021.

[Fuj10] M. Fujimoto, S. Watanabe, and T. Nakatani, “Voice activity detection using frame-wise model re-estimation method based on Gaussian pruning with weight normalization,” in *Proc. INTERSPEECH*, Makuhari, Chiba, Japan, pp. 3102-3105, Sep. 2010.

[Fuj14] H. Fujimura, “Simultaneous gender classification and voice activity detection using deep neural networks,” in *Proc. INTERSPEECH*, Singapore, pp. 1139-1143, Sep. 2014.

[Gha19] E. Ghaleb, M. Popa, and S. Asteriadis, “Multimodal and temporal perception of audio-visual cues for emotion recognition,” in *Proc. International Conference on Affective Computing and Intelligent Interaction* (*ACII*), Cambridge, United Kingdom, pp. 552-558, Sep. 2019.

[Gha20] E. Ghaleb, M. Popa, and S. Asteriadis, “Metric learning-based multimodal audio-visual emotion recognition,” in *IEEE MultiMedia*, vol. 27, iss. 1, pp. 37-48, Mar. 2020.

[Goo16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, Massachusetts, United States of America, 2016.

[Han98] J.H.L. Hansen et al., *Getting started with the SUSAS: Speech Under Simulated and Actual Stress database*, Technical report RSPL-98-10, Robust Speech Processing Laboratory, Duke University, Durham, United States of America, Apr. 1998.

[Has06] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, New York, New York, United States of America, 2006.

[He09] L. He et al., "Stress detection using speech spectrograms and sigma-pi neuron units," in *Proc. International Conference on Natural Computation*, Tianjian, China, pp. 260-264, Aug. 2009.

[He15] L. He, Y. Bo, and G. Zhao, "Speech-oriented negative emotion recognition," in *Proc. Chinese Control Conference* (*CCC*), Hangzhou, China, pp. 3553-3558, Jul. 2015.

[Jai16] M. Jaiswal, S. Tabibu, and R. Bajpai, "The truth and nothing but the truth: multimodal analysis for deception detection," in *Proc. IEEE International Conference on Data Mining Workshops* (*ICDMW*), Barcelona, Spain, pp. 938-943, Dec. 2016.

[Jin15] Q. Jin et al., "Speech emotion recognition with acoustic and lexical features," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Brisbane, Queensland, Australia, pp. 4749-4753, Apr. 2015.

[Kan21] S. Kanwal and S. Asghar, "Speech emotion recognition using clustering-based GA-optimized feature set," in *IEEE Access*, vol. 9, pp. 125830-125842, Sep. 2021.

[Ker19] L. Kerkeni et al., "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," in *Speech Communication*, vol. 114, pp. 22-35, Nov. 2019.

[Kit07] N. Kitaoka, K. Yamamoto, and T. Kusamizu, "Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance," in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding* (*ASRU*), Kyoto, Japan, pp. 607-612, Dec. 2007.

[Kop19] D. Kopev et al., "Detecting deception in political debates using acoustic and textual features," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), Singapore, pp. 652-659, Dec. 2019.

[Lat20a] S. Latif et al., "Augmenting generative adversarial networks for speech emotion recognition," in *Proc. INTERSPEECH*, Shanghai, China, pp. 521-525, Oct. 2020.

[Laz99] R.S. Lazarus, *Stress and Emotion: A new synthesis*, 1st ed., Springer, New York, New York, United States of America, 1999.

[Li07] X. Li et al., "Stress and emotion classification using jitter and shimmer features," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Honolulu, Hawaii, United States of America, pp. 1081-1084, Apr. 2007.

[Liu20] S. Liu et al., "Hierarchical component-attention based speaker turn embedding for emotion recognition," in *Proc. International Joint Conference on Neural Networks* (*IJCNN*), Glasgow, United Kingdom, pp. 1-7, Jul. 2020.

[Liu21] N. Liu et al., “Transfer subspace learning for unsupervised cross-corpus speech emotion recognition,” in *IEEE Access*, vol. 9, pp. 95925-95937, Jul. 2021.

[Lot17] R. Lotfidereshgi and P. Gournay, “Biologically inspired speech emotion recognition,” in *Proc. IEEE International Conference on Acoustics Speech, and Signal Processing* (*ICASSP*), New Orleans, Louisiana, United States of America, pp. 5135-5139, Mar. 2017.

[Mat09] D. Matsumoto, *The Cambridge Dictionary of Psychology*, 1st ed., Cambridge University Press, Cambridge, United Kingdom, 2009.

[Men17] G. Mendels et al., “Hybrid acoustic-lexical deep learning approach for deception detection,” in *Proc. INTERSPEECH*, Stockholm, Sweden, pp. 1472-1476, Aug. 2017.

[Mih19a] S. Mihalache, D. Burileanu, G. Pop, and C. Burileanu, “Modulation-based speech emotion recognition with reconstruction error feature expansion,” in *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timisoara, Romania, pp. 1-6, Oct. 2019.

[Mih19b] S. Mihalache, G. Pop, and D. Burileanu, “Introducing the RODeCAR database for deceptive speech detection,” in *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Timisoara, Romania, pp. 1-6, Oct. 2019.

[Mih20] S. Mihalache and D. Burileanu, *Speech emotion recognition for dissimulated behavior monitoring in surveillance applications*, Final report, Project 2SOL/2017 – PN-III-P2-2.1-SOL-2016-02-0002, *Intelligent Systems for Video and Audio Analysis – Technologies and Innovative Video Systems for Person Re-identification and Analysis of Dissimulated Behavior* (*SPIA-VA*), Apr. 2020.

[Mih21a] S. Mihalache, I.A. Ivanov, and D. Burileanu, “Deep neural networks for voice activity detection,” in *Proc. International Conference on Telecommunications and Signal Processing* (*TSP*), Brno, Czech Republic, pp. 191-194, Jul. 2021.

[Mih21b] S. Mihalache, D. Burileanu, and C. Burileanu, “Detecting psychological stress from speech using deep neural networks and ensemble classifiers,” in *Proc. International Conference on Speech Technology and Human-Computer Dialogue* (*SpeD*), Bucharest, Romania, pp. 74-79, Oct. 2021.

[Mih21c] S. Mihalache and D. Burileanu, “Dimensional models for continuous-to-discrete affect mapping in speech emotion recognition,” in *University Politehnica of Bucharest Scientific Bulletin, Series C*, vol. 83, iss. 4, pp. 137-148, Dec. 2021.

[Mih22a] S. Mihalache and D. Burileanu, “Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection,” in *Sensors*, vol. 22, iss. 3, art. no. 1228, pp. 1-21, Feb. 2022.

[Mih22b] S. Mihalache, D. Burileanu, E. Franți, M. Dascălu, and C.A. Brătan, “Lasting emotions – An investigation of short- and long-term affective content remanence in speech,” in *Romanian Journal of Information Science and Technology*, vol. 25, iss. 1, pp. 20-35, Mar. 2022.

[Mir17] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), New Orleans, Louisiana, United States of America, pp. 2227-2231, Mar. 2017.

[Mor12] G. S. Morrison, P. Rose, and C. Zhang, “Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice,” in *Australian Journal of Forensic Sciences*, vol. 44, no. 2, pp. 155-167, Jun. 2012.

[Pan20] Z. Pan et al., “Multi-modal attention for speech emotion recognition,” in *Proc. INTERSPEECH*, Shanghai, China, pp. 364-368, Oct. 2020.

[Per15] V. Perez-Rosas et al., “Deception detection using real-life trial data,” in *Proc. ACM on International Conference on Multimodal Interaction*, New York, New York, United States of America, pp. 59-66, Nov. 2015.

[Rah12] T. Rahman and C. Busso, “A personalized emotion recognition system using an unsupervised feature adaptation scheme,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Kyoto, Japan, pp. 5117-5120, Mar. 2012.

[Rao17] W. Rao et al., “Investigation of fixed-dimensional speech representations for real-time speech emotion recognition system,” in *Proc. International Conference on Orange Technologies* (*ICOT*), Singapore, pp. 197-200, Dec. 2017.

[Sch14] B. Schuller et al., “The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load,” in *Proc. INTERSPEECH*, Singapore, pp. 427-431, Sep. 2014.

[Sen22] U.M. Sen et al., “Multimodal deception detection using real-life trial data,” in *IEEE Transactions on Affective Computing*, vol. 13, iss. 1, pp. 306-319, Mar. 2022.

[Sha23b] D. Sharma et al., “Speech emotion recognition system using SVD algorithm with HMM model,” in *Proc. International Conference for Advancement in Technology* (*ICONAT*), Goa, India, pp. 1-5, Jan. 2023.

[Shi20] H.K. Shin et al., “Speaker-invariant psychological stress detection using attention-based network,” in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Auckland, New Zealand, pp. 308-313, Dec. 2020.

[Su21] B.H. Su and C.C. Lee, “A conditional cycle emotion GAN for cross corpus speech emotion recognition,” in *Proc. IEEE Spoken Language Technology Workshop* (*SLT*), Shenzhen, China, pp. 351-357, Jan. 2021.

[Tan21] D. Tang et al., “Adieu recurrence? End-to-end speech emotion recognition using a context stacking dilated convolutional network,” in *Proc. European Signal Processing Conference* (*EUSIPCO*), Amsterdam, Netherlands, pp. 1-5, Jan. 2021.

[Vas15] J.C. Vasquez-Correa et al., “Emotion recognition from speech under environmental noise conditions using wavelet decomposition,” in *Proc. International Carnahan Conference on Security Technology* (*ICCST*), Taipei, Taiwan, pp. 247-252, Sep. 2015.

[Vel19] A. Velichko et al., “Applying ensemble learning techniques and neural networks to deceptive and truthful information detection task in the flow of speech,” in I. Kotenko et al. (Eds.) *Intelligent Distributed Computing XIII. Studies in Computational Intelligence*, vol. 868, Springer, Cham, Switzerland, pp. 477-482, Oct. 2019.

[Ver09] B. Verschuere, V. Prati, and J. De Houwer, “Cheating the lie detector,” in *Journal of Psychological Science*, vol. 20, iss. 4, pp. 410-413, Apr. 2009.

[Vil12] G. Villar, J. Arciuli, and D. Mallard, “Use of ‘um’ in the deceptive speech of a convicted murderer,” in *Applied Psychoacoustics*, vol. 33, iss. 1, pp. 83-95, Jan. 2012.

[War13] A.B. Warriner, V. Kuperman, and M. Brysbaert, “Norms of valence, arousal, and dominance for 13,915 English lemmas,” in *Behavior Research Methods*, vol. 45, no. 4, pp. 1191-1207, Dec. 2013.

[Wat99] D. Watson et al., “The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence,” in *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 820-838, May 1999.

[Yi22] L. Yi and M.W. Mak, “Improving speech emotion recognition with adversarial data augmentation network,” in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 1, pp. 172-184, Jan. 2022.

[Yil21] S. Yildirim, Y. Kaya, and F. Kilic, “A modified feature selection method based on metaheuristic algorithms for speech emotion recognition,” in *Applied Acoustics*, vol. 173, art. no. 107721, Feb. 2021.

[Zao14] L. Zao, D. Cavalcante, and R. Coelho, “Time-frequency feature and AMS-GMM mask for acoustic emotion classification,” in *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 620-624, May 2014.

[Zha18a] Y. Zhang et al., “Attention based fully convolutional network for speech emotion recognition,” in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Honolulu, Hawaii, United States of America, pp. 1771-1775, Nov. 2018.

[Zha18b] H. Zhao, N. Ye, and R. Wang, “Transferring age and gender attributes for dimensional emotion prediction from big speech data using hierarchical deep learning,” in *Proc. IEEE International Conference on Big Data Security on Cloud* (*BigDataSecurity*), *IEEE International Conference on High Performance and Smart Computing* (*HSPC*), *and IEEE International Conference on Intelligent Data and Security* (*IDS*), Omaha, Nebraska, United States of America, pp. 20-24, May 2018.

[Zha19] Z. Zhao et al., “Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition,” in *IEEE Access*, vol. 7, pp. 97515-97525, Jul. 2019.

[Zha20] J. Zhang, S.I. Levitan, and J. Hirschberg, “Multimodal deception detection using automatically extracted acoustic, visual, and lexical features,” in *Proc. INTERSPEECH*, Shanghai, China, pp. 359-363, Oct. 2020.

[Zha21a] J. Zhang et al., “Cross-corpus speech emotion recognition using joint distribution adaptive regression,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Toronto, Ontario, Canada, pp. 3790-3794, Jun. 2021.

[Zha22] W. Zhang et al., “Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression,” in *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, iss. 2, pp. 588-598, Jun. 2022.