

## Lasting emotions – An investigation of short- and long-term affective content remanence in speech

Serban MIHALACHE<sup>1,3</sup>, Dragos BURILEANU<sup>\*1</sup>, Eduard FRANTI<sup>2,3</sup>,  
Monica DASCALU<sup>1,3</sup>, and Costin-Andrei BRATAN<sup>1,3</sup>

<sup>1</sup>University *Politehnica* of Bucharest, Romania

<sup>2</sup>National Institute for Research and Development in Microtechnologies, Bucharest, Romania

<sup>3</sup>Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania

E-mail: serban.mihalache@upb.ro, dragos.burileanu@upb.ro\*,  
eduard.franti@imt.ro, monica.dascalu@upb.ro,  
costin.bratan@stud.etti.upb.ro

\* Corresponding author

**Abstract.** *Speech emotion recognition* (SER) is a promising ongoing research area with important applications for forensics and law enforcement operations, among others. Approaches have been previously proposed to integrate SER systems to assist in surveillance tasks, emergency services, police investigations, or other operations, especially in the attempt to anticipate and prevent potential criminal acts or even to counter terrorist activities. 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 and warrant further inquiry. In this work, we gain insight into these patterns and prove 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. In order to provide a reasonable partial proxy for the high-stakes conditions and triggers expected in real-life scenarios, we have developed a speech dataset comprising 270 recordings of 18 students behind on their university exams and about to attempt them for the second or third time; thus, the upcoming exams and the potential consequences of failing them represent the emotionally charged event. Human evaluators labeled the recordings in terms of the identified emotional classes (grouped into negative emotional classes and the neutral state) and of arousal-valence affect space values. Analyzing the annotations made by the evaluators, we prove that the subjects' affective response is significantly higher as the emotionally

charged event approaches, and emotional remanence can be observed even 15 minutes after the initial interaction, or even after 30 minutes when under the added influence of the event's imminence. We show that the arousal increases (higher intensity affective response) as the event draws closer, while the valence decreases (more negative affective response), again supporting the second hypothesis, and suggesting that such patterns would be relevant for the targeted applications. We propose and implement a SER system using *artificial neural networks* (ANNs) based on *multilayer perceptron* (MLP) models, obtaining good performance (up to 72.7% accuracy) when training in a speaker-independent manner, and yielding classification and regression results consistent with those given by human evaluation, supporting the possibility and usefulness of using *machine learning* (ML) systems to monitor affective responses in order to automatically detect the patterns associated with the behaviors relevant for forensic and law enforcement applications and to facilitate intervention and prevention.

**Key-words:** Speech emotion remanence; speech emotion recognition; machine learning; multilayer perceptrons; law enforcement.

## 1. Introduction

Emotions exert a powerful and ubiquitous influence throughout our lives. They change both the way we think and the way we perceive the world and other people. They manifest externally, visually and paralinguistically. Our facial expressions, our body language, our vocal characteristics, the specific words we use or avoid, all of these can reveal the nature of the emotions we are experiencing. Some psychologists consider that emotions have a stronger influence on human vocal expression than on facial expressions [1]. Several different emotions may trigger the same facial expression, but distinct speech features, while also noting that, when an emotion occurs, the vocal traits change almost instantly. It is also for these reasons that the audio modality remains of standalone interest in emotion recognition research, not just in combination with other modalities (video, text, biological information, etc.).

*Speech emotion recognition* (SER), the ability to automatically detect and infer the affective content hidden in human speech, particularly using *machine learning* (ML) and deep learning (DL) systems such as *artificial neural networks* (ANNs), is a promising ongoing research area [2–4], with important applications for human-machine interfaces, medical monitoring, clinical diagnostic assistance, and forensics and law enforcement operations, among others. Targeting the latter two fields, approaches have been proposed to integrate SER systems to assist in surveillance tasks, emergency services, police investigations, or other operations, especially in the attempt to anticipate and prevent potential criminal acts or even to counter terrorist activities [5–8].

Beyond the performance of SER systems themselves, which are not yet accurate enough for large-scale integration in such sensitive applications [6–8], 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 and warrant further inquiry.

The aim of this work is to gain insight into these patterns and our approach involves investigating the length of time for which the emotional response decays beyond a certain level, *i.e.*, the affective content remanence; the more intense the emotions, the stronger they are embedded within the vocal content, and for a longer period of time [1, 9]. However, most emotion recognition research focuses either on the very short term evolution of subjects' affective state (on the

order of a few minutes or less), or on very long timescales (on the order of months and years, with sparse observations) [2, 4, 10, 11], and in different contexts than the ones relevant for this article’s targeted applications. Additionally, while the concept of remanence has been investigated before [9], no quantitative study of the affective response evolution on the timescales we propose has been performed in the context of SER, to the best of our knowledge.

To this end, we first analyze the temporal evolution of the affective content of speech samples from a novel proprietary dataset on a short timescale (within 1 hour) and on a longer timescale (over 5 days). Then, in order to validate the applicability of automatic emotion recognition, we develop and train ANN-based systems and evaluate their performance for the same tasks.

The rest of the article is organized as follows. In section 2, we further discuss the theoretical background behind emotion recognition, and the challenges concerning SER. We clearly define the hypotheses of our study on speech emotion remanence in section 3, and also describe the experimental methodology employed, discuss the dataset developed for this work, and present the experimental setup and results for both human and automatic evaluation of the affective content of our dataset and its evolution in time, proving the validity of our assumptions regarding speech emotion remanence. Finally, we draw conclusions and consider our intended future work in section 4.

## 2. Theoretical background

The two main schools of thought in psychology provide two models of emotions: discrete categories [12, 13], where each affective state is viewed as a distinct, standalone class (*e.g.*, anger, fear, sadness, etc.), and is holistically distinguished from the others, leading to a classification task; and dimensional modeling [14–16], where a number of continuous valued psychological measures (*e.g.* arousal – a subjective evaluation of the intensity of the affective response; and valence – a subjective evaluation of the positive/negative character of the affective response) form an  $n$ -dimensional affect space (typically 2D), the position within being the target, leading to a regression task. Additionally, it is worth noting that the two paradigms are, in fact, related, since each emotion (in the sense of emotional class) is mappable to a subdomain within the affect space. Possible mapping techniques and models have been reported in previous literature [17, 18].

For manual emotion classification, a number of human evaluators are tasked with listening to speech samples (either previously recorded or in real time) and selecting from a set of words representing emotional classes the one that best describes the perceived expressed emotion [19].

In the case of manual affect space regression, the evaluators instead indirectly assign a numerical value for each dimension (*e.g.*, arousal and valence), for each speech sample. Most often, the *self-assisted manikin* (SAM) technique [20] is employed, in which 5 drawings are presented for each dimension, representing the levels of arousal (ranging from very low to very high) and the type of valence (ranging from strongly negative to strongly positive), the evaluator picking the pair that is most applicable. Post-evaluation, each selected drawing is associated with the corresponding number from a set of 5 values (usually  $\{1, 2, 3, 4, 5\}$  or  $\{-1, -0.5, 0, 0.5, 1\}$ ).

However, a crucial aspect of the task consists of ensuring that the speech datasets used are appropriate for the considered application. As outlined in [19, 21], most of the publicly available corpora for paralinguistic tasks [10, 11] were developed in the context of simulated behavior, in a relaxed, low-stakes environment, constituting significant disadvantages, especially regarding

their departure from the characteristics of actual real-life situations. Consequently, a better approach would involve using more realistic databases, in which participants have not received any information regarding their expected behavior, have full autonomy over the content and form of their answers, and are placed in (and aware of) realistic scenarios. Moreover, previous research [22–25] 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 emotion expression *vs.* the culture, background, culture, age, etc. of the speaker, but there exists no conclusive evidence for or against this idea.

Concerning the targeted law enforcement and forensic applications considered in this article, the existing databases related to actual criminal or terrorist acts either comprise only metadata (*e.g.*, geographical location, historical information, ethnicity, age, gender, etc.) and no actual modalities (audio or video recordings, transcriptions of speech, etc.), or are restricted from being used in public research [26].

Concordantly, as part of this work, we recorded and annotated a novel affective dataset, to provide a reasonable partial proxy for the high-stakes conditions and triggers expected in real-life scenarios. This dataset is described in section 3.1. To investigate the automatic evaluation of the emotional content, we develop and train ANNs based on *Multilayer Perceptron* (MLP) models, since they are accessible in terms of computational cost and have consistently proven to be effective for paralinguistic tasks [27–29] concerning both classification and regression, *i.e.*, both the paradigm of emotions as classes and the dimensional modeling of emotions.

### 3. Study on speech emotion remanence

#### 3.1. Methodology and setup

The two main hypotheses of our study are the following:

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 affective state on the subject's part.
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.

The two are naturally also interrelated, since the closer in time to an emotionally charged event an interaction is, the higher the subject's initial affective response to the interaction will be, and the short-time affective response decay period will be increased. It should be noted that, due to the nature of the targeted law enforcement and forensic applications, the focus of this work is on monitoring negative emotions, or the affect space subdomain associated with it, *vs.* a neutral affective state; positive or mixed affective states are not within the scope.

To test both hypotheses, we constructed a dataset 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. Thus, the upcoming exams and the potential consequences of failing them represent the emotionally charged event. The students were instructed that they would be contacted by different people for this research and would have to respond to questions concerning their exam preparation, lookout, and other usual activities. All students accepted and signed a participation agreement by which they committed to answer the questions truthfully and interact openly with the interviewers.

The interactions were undertaken daily, starting from 5 days before the exam date. During each conversation, the students were initially asked about the difficulties encountered during their study and their lookout concerning their near future, thus triggering an affective response. Then, the discussion would be diverted towards ordinary and neutral topics for at least 30 minutes. The conversations were recorded at 44.1 kHz sampling rate and stored in PCM format, and particular utterances were extracted from the students' replies to the initial query, and after 15 and 30 minutes of neutral conversation, respectively. This process was repeated 5 times, including on the day before the exams. The duration of the extracted utterances ranges between 6 and 45 seconds, with an average of 15 seconds.

Afterwards, human evaluators were asked to listen to the recorded utterances and label them in terms of the identified dominant emotional class (from the 4-class set: *anger*, *fear*, *sadness*, *neutral*) and of *arousal valence* affect space values through the SAM technique described in section 2. Subsequently, the arousal and valence labels were associated with numerical values from the set 1, 2, 3, 4, 5 and averaged over the number of evaluations. Since this study is only concerned with the detection of negative emotions in general, and not discriminating between them, the class labels were merged into 2 groups: *negative* (grouping together *anger*, *fear*, and *sadness*) and *neutral*. Majority voting was used to decide the final class labeling of each utterance.

There were 18 students (4 female, 14 male) involved in the process, of ages between 19.7 years and 23.3 years, with a mean of 21.3 years. The number of evaluators participating was 5 (1 female, 4 male) to ensure no tied results can occur when applying majority voting. The total number of recorded utterances is, thus, 270, and the total duration of the dataset's speech content is 1 hour and 8 minutes. The dataset annotation was analyzed using linear regression to highlight the observed patterns and trends supporting the hypotheses, and the results are detailed in section 3.2.

In order to validate the applicability of automatic emotion recognition, we propose and develop the MLP-based ANNs illustrated in Fig. 1, taking as input an extensive set of descriptors obtained by applying high level statistical functions on extracted hand-crafted acoustic, prosodic, spectral, and cepstral features. The MLP model uses several hidden layers, with different numbers of nodes per layer, and an output layer whose size is equal to the number of class groups, *i.e.*, 2, in the case of classification, and equal to the number of affect space dimensions, *i.e.*, 2, in the case of regression. For classification, this configuration for the output layer was chosen empirically after observing that having a separate output probability of the sample belonging to each class allowed the optimization algorithm to better adjust the ANN weights than having a single output probability of the sample belonging to the positive class. 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 log2 law.

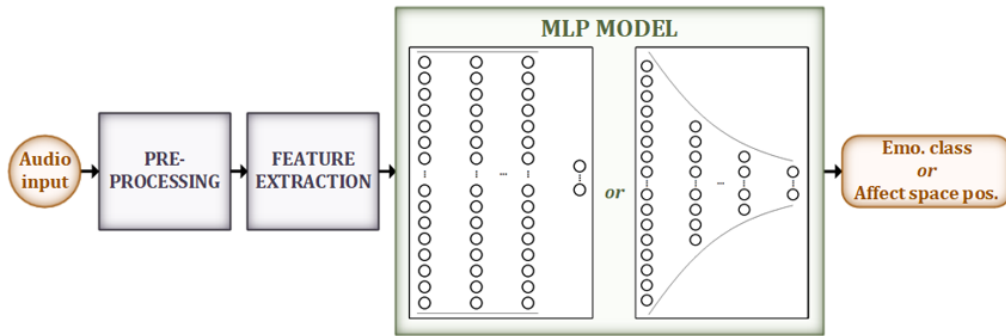
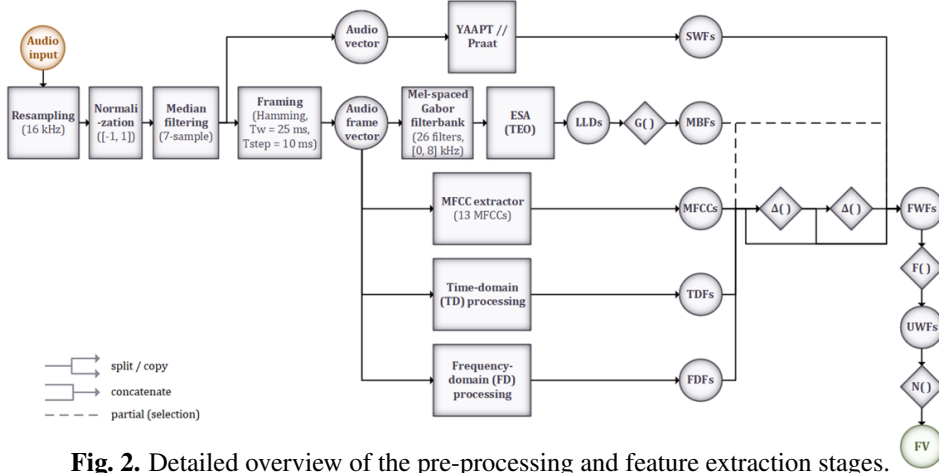


Fig. 1. Proposed ML system using MLP-based ANNs.

The pre-processing and feature extraction stages are detailed in Fig. 2. The former consists of resampling the audio input to 16 kHz, applying amplitude normalization and 7-sample median filtering, and framing the signal using Hamming windows of 25 ms duration, with a 15 ms overlap. The feature set used is an extension of the ComParE feature set [30], and also includes the *modulation-based features* (MBFs) proposed in [31] and utilized successfully in our previous work on speech emotion recognition [32] and speech stress detection [33]. The *Yet Another Algorithm for Pitch Tracking* (YAAPT) algorithm [34] and the Python implementation of Praat [35] are used to extract *segment-wise features* (SWFs), with the same stride employed for framing the input audio (10 ms): pitch, *harmonic-to-noise ratio* (HNR), local jitter and local shimmer. The other features are extracted frame-wise and consist of: i) the MBFs, obtained through applying a 26-filter Mel-spaced Gabor filterbank together with the Teager-Kaiser Energy Separation Algorithm (ESA-TEO) [31], followed by the  $G(\cdot)$  set of functionals (mean, standard deviation, weighted mean, weighted squared bandwidth) on the resulting *low-level descriptors* (LLDs); ii) the first 13 MFCCs; iii) *time-domain features* (TDFs); and iv) *frequency-domain features* (FDFs). The TDFs are: RMS energy and *zero-crossing rate* (ZCR). The FDFs are: loudness, low frequency band energy (250–650 Hz), high frequency band energy (1–4 kHz), spectral flux, spectral centroids, spectral spread, spectral skewness, spectral kurtosis, spectral entropy, spectral slope, spectral roll-off points (25%, 50%, 75%, 90%) and the log filterbank energies. Delta and double-delta coefficients are computed for the MFCCs, TDFs, and FDFs, as well as for some of the MBFs. Together with the SWFs, these represent the *frame wise features* (FWFs), on which the  $F(\cdot)$  set of functionals (mean and standard deviation) is applied, resulting in the *utterance-wise features* (UWFs).

The  $N(\cdot)$  function, z score normalization, is applied per speaker and is defined in (1), where  $x$  is a vector of the original values of each feature (for the entire dataset),  $m_x$  and  $s_x$  are the mean and standard deviation of  $x$ , and  $x_n$  is the normalized vector. The obtained normalized feature vector (FV) has a size of 2,258 and is subsequently fed to the ANNs. We also ran experiments using a reduced feature set comprising only the MFCC-related descriptors (the normalized mean and standard deviation functionals applied utterance wise to the first 13 MFCCs and their delta and double-delta coefficients), since this has sometimes proven to yield better results due to the more robust nature of these features and to the removal of correlations between other extracted features.



**Fig. 2.** Detailed overview of the pre-processing and feature extraction stages.

$$x_n = \frac{x - m_x}{s_x} \quad (1)$$

The MLP models use either the same number of nodes for each hidden layer (‘constant’ architecture) or a progressively smaller number, following a log2 law (‘log2dec’ architecture). The depth (number of hidden layers) was varied between 2 and 4, with an initial number of nodes (for the first hidden layer) of 256, 128, 64 or 32. Dropout was included after each hidden layer, with a rate between 20% and 50%, in order to reduce overfitting. Other hyperparameters chosen include: the *rectified linear unit* (ReLU) activation function for the hidden layers, and the softmax (for classification) or identity (for regression) function for the output layer; ‘Adam’ as the optimization algorithm [36]; and L1-norm regularization with a regularization parameter equal to  $10^{-4}$ . The batch size was set to 32. The experiments were run for up to 100 epochs, with early stopping and learning rate decay (between  $10^{-3}$  and  $10^{-6}$ , with a decay factor of 0.1).

Additionally, for classification, since the dataset is imbalanced in terms of class group distribution (*negative vs. neutral*), class weighting was employed, boosting the contribution of the deceptive samples in computing the loss function by their representation ratio (*i.e.*, the ratio of the two class group sizes).

Two types of experiments were performed: speaker-dependent and speaker-independent. In the first case, we trained the ANNs using the data available from all speakers from the first 2 days, and evaluated their performance on the data from the subsequent 3 days. In the second case, we divided the dataset, reserving 50% for training and 50% for validation and ensuring adequate speaker separation between the two subsets (2 female and 7 male speakers for each). For the speaker-independent experiments, we employed 10-fold cross-validation, distributing the speakers over the folds as uniformly as possible.

The performance metrics used for classification were the *unweighted accuracy* (UA) and the *weighted accuracy* (WA). These are defined in (2) and (3), where  $K$  is the number of classes, *i.e.*, 2,  $N_k$  is the size of class  $k$ ,  $H_k$  is the number of correct predictions made for class  $k$ , and  $N$  is the size of the entire subset. For regression, we used the *mean squared error* (MSE), defined in (4), where  $N$  is the same,  $y_{A,i}$  and  $y_{V,i}$  are the real values of arousal and valence for sample  $i$ , and  $\hat{y}_{A,i}$  and  $\hat{y}_{V,i}$  are their predicted values.

$$UA = \frac{1}{K} \sum_{k=1}^K \frac{H_k}{N_k} \quad (2)$$

$$WA = \frac{1}{N} \sum_{k=1}^K H_k = \frac{1}{K} \sum_{k=1}^K \frac{K \cdot N_k}{N} \cdot \frac{H_k}{N_k} \quad (3)$$

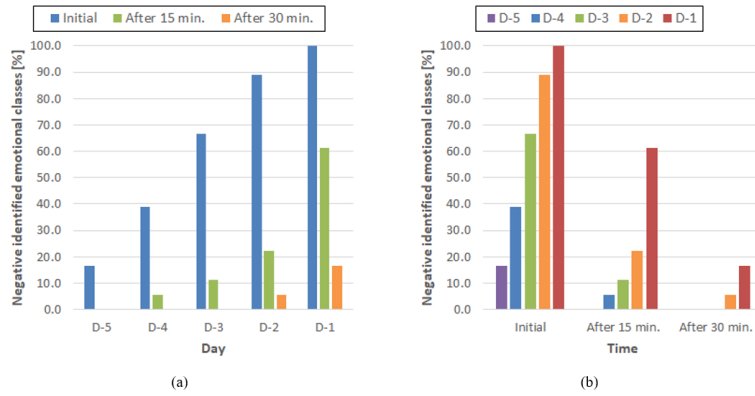
$$MSE = \frac{1}{2N} \left[ \sum_{i=1}^N (y_{A,i} - \hat{y}_{A,i})^2 + \sum_{i=1}^N (y_{V,i} - \hat{y}_{V,i})^2 \right] \quad (4)$$

The proposed ANNs were developed in Keras, a neural networks framework for Python. All experiments were performed on a Linux machine running Ubuntu 18.04, with a 16-core Intel Xeon E5 1680 CPU at 3 GHz, 192 GB of RAM at 2133 MHz, and an 8 GB Nvidia Quadro M4000 GPU, and the results are given and discussed in section 3.3.

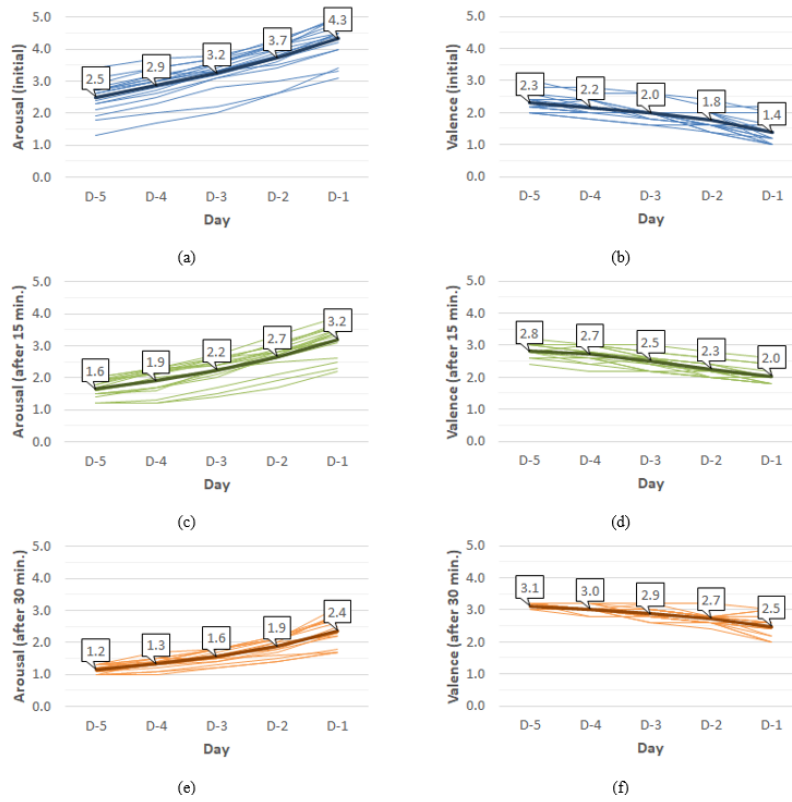
### 3.2. Experimental results – human evaluation

The ratio between the number of speakers identified as expressing a *negative* emotion and the total number of speakers (*i.e.*, 18) is represented *vs.* each day in Fig. 3a, where D-5 is the first day of spoken interaction (5 days before the exam) and so on up to D-1 (the day before the exam), and *vs.* each timestamp in Fig. 3b, with the labels referring to the initial affective response, and after 15 and 30 minutes, respectively, of neutral conversation. As expected and hypothesized, the affective response is significantly higher as the emotionally charged event draws closer, which can be inferred from the increasing ratio of *negative* identified emotional classes in Fig. 3b, especially for the initial timestamp. Additionally, the short-time remanence of the speech affective content can be observed in Fig. 3a, since *negative* emotional classes have been identified even 15 minutes after the initial interaction, or even after 30 minutes if under the added influence of the exam’s imminence.

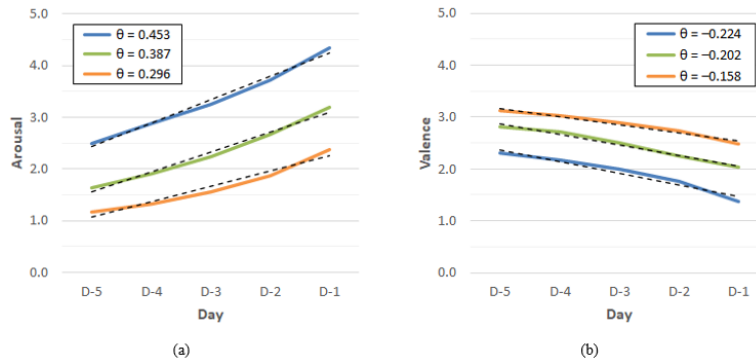
In the case of positioning within the affect space, the values for arousal and valence *vs.* each day are represented in Fig. 4 for each timestamp, for each speaker individually (thin lines) and on average (thick lines). On average, and for most speakers, the arousal increases (higher intensity affective response) as the emotionally charged event draws closer, while the valence decreases (more negative affective response), again supporting the second hypothesis.



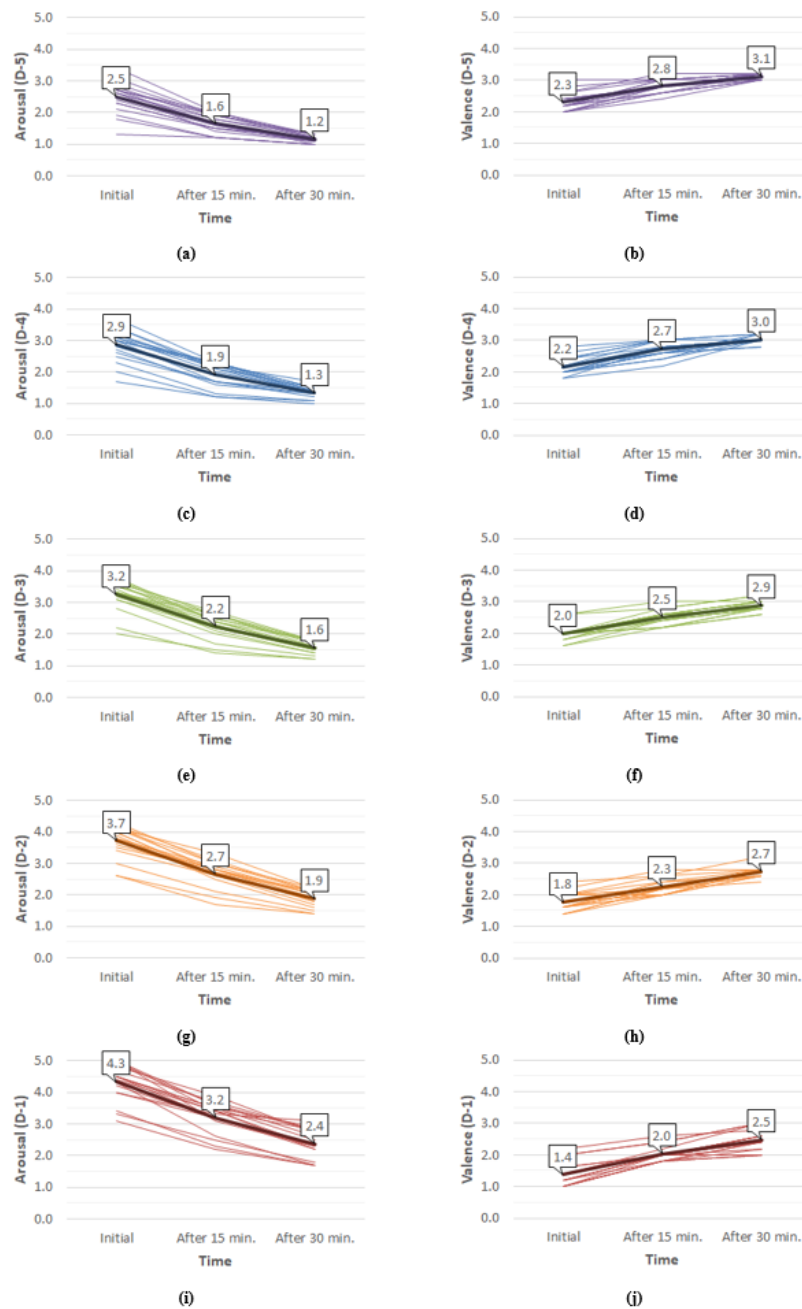
**Fig. 3.** Percentage ratio of speakers identified as expressing negative emotions.



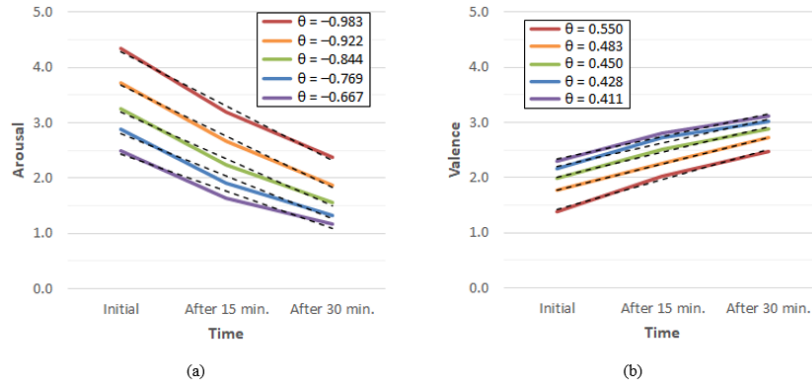
**Fig. 4.** Arousal and valence evolution vs. day, for each timestamp. Thin lines represent individual speaker evolutions, while the data labeled thick lines represent the average values for all speakers.



**Fig. 5.** Arousal and valence evolution vs. day, for each timestamp. Blue lines represent initial values, while the values after 15 and 30 minutes are illustrated in green and orange, respectively. Linear regression trendlines are illustrated with black dashed lines, with  $\theta$  representing the trendline slope.



**Fig. 6.** Arousal and valence evolution vs. timestamp, for each day. Thin lines represent individual speaker evolutions, while the data labeled thick lines represent the average values for all speakers.



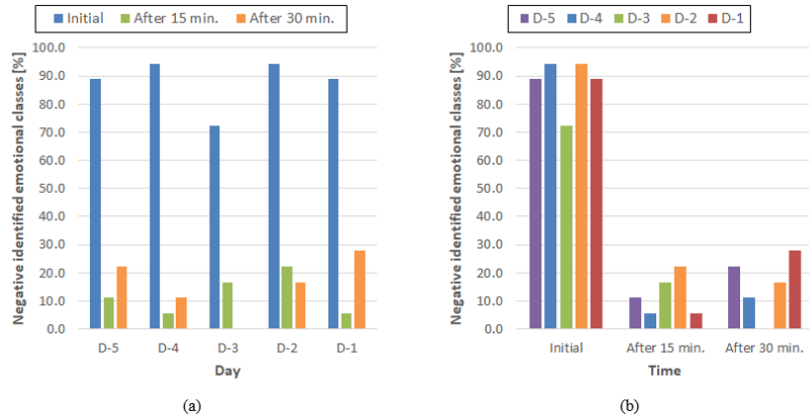
**Fig. 7.** Arousal and valence evolution vs. timestamp, for each day. The values for days D-5 up to D-1 are shown in purple, blue, green, orange, and red, respectively. Linear regression trendlines are illustrated with black dashed lines, with  $\theta$  representing the trendline slope.

The average curves are also illustrated again in Fig. 5, together with their evolution modeled through linear regression (black dashed lines), where  $\theta$  represents the trendline slope. A 14.5%/day increase in initial arousal and an 11.7%/day decrease in initial valence can be observed, both reaching levels associated with negative emotions [17, 18] on day D-2, suggesting that such patterns would be relevant for the targeted applications and warrant further inquiries if detected when monitoring a subject.

For the short-term evolution, the arousal and valence curves vs. each timestamp are illustrated in Fig. 6 for each day, for each speaker individually (thin lines) and on average (thick lines), and the average curves are redrawn in Fig. 7, together with the linear regression trendlines (black dashed lines). For all speakers, the arousal and valence do not decrease, respectively increase, immediately towards *neutral* values following the initial affective response, consistent with the first hypothesis. Moreover, as the emotionally charged event draws closer, the rate of arousal decrease and valence increase diminishes (e.g., the average arousal decrease on day D-5 is 36% after 15 min. and 52% after 30 min., but becomes 25% and 44%, respectively, on day D-1) suggesting an increased emotional remanence factor.

### 3.3. Experimental results – automatic evaluation

For both classification and regression, for each set of experiments (speaker-dependent and speaker-independent), the total number of tested ANN configurations was 192 (as described in section 3.1). In the speaker-dependent cases, all of the models demonstrated poor performance. The likely explanation is that, by training the models only on data from the first few days, farther away from the emotionally charged event, when the subjects' affective responses are lower, the ANNs are exposed only to a very small number of samples expressing *negative* emotions and only learn from arousal and valence values closer to the *neutral* state, thus being unable to properly learn to identify *negative* emotions and extrapolate towards more extreme values of arousal and valence. An example of these poor classification results is shown in Fig. 8. The system wrongly identifies more *negative* emotional classes (or none at all) after 30 minutes from the initial interaction than after 15 minutes, and the number of *negative* identified emotional classes does not increase monotonically vs. each day, for none of the three timestamps.



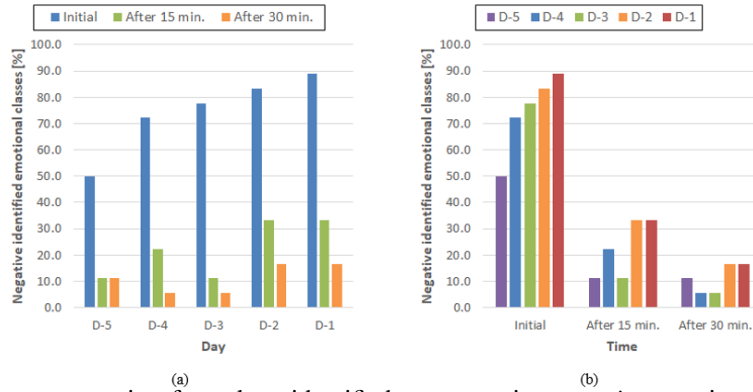
**Fig. 8.** Percentage ratio of speakers identified as expressing *negative* emotions. Evaluation example of a system trained in the speaker-dependent manner, demonstrating poor performance.

However, the speaker-independent approach yielded good results. The best performing system for classification used the reduced, MFCC-based feature set (78 features), the ‘constant’ architecture with a depth of 2 and 256 initial nodes (*i.e.*, 256 nodes in each of the 2 hidden layers), and a dropout rate of 40%. The resulting UA was 67.8%, and the WA was 72.7%. For regression, the full set of features was used (2,258 features), the model architecture was also ‘constant’, but with a depth of 4 and 256 initial nodes, and the best dropout rate was 20%. The resulting MSE was 0.726 (0.995 for arousal and 0.458 for valence). These configurations are summarized in Table 1, together with the corresponding performance metrics.

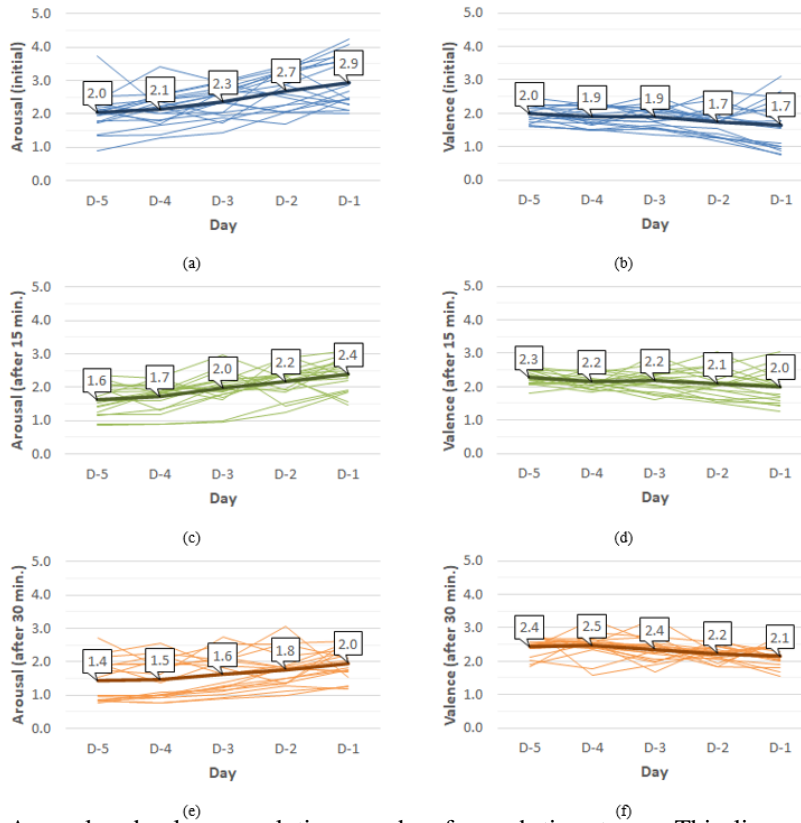
**Table 1.** Summary of best performing system configurations for speaker-independent classification and regression

Task	System configuration			Performance metrics		
	Feature set	MLP node structure	Dropout	UA	WA	MSE
Classification	Reduced (78 MFCC-based features)	[256, 256] (‘constant’ arch., depth = 2)	40%	67.8%	72.7%	-
Regression	Full (2,258 features)	[256, 256, 256, 256] (‘constant’ arch., depth = 4)	20%	-	-	0.726 (A: 0.995, V: 0.458)

The classification results provided by the speaker-independent system are represented in Fig. 9, and the arousal-valence values *vs.* each day are illustrated in Fig. 10. The same trends revealed through human evaluation can also be observed here, especially regarding the initial affective response. The number of *negative* identified emotional classes shows a long-term increase as the emotionally charged event draws closer, while it decreases in the short-term, but not immediately, still indicating the presence (remanence) of *negative* affective content even after 30 minutes from the initial interaction. Concurrently, for most speakers, the arousal increases (higher intensity affective response) each day, while the valence decreases (more negative affective response), consistent with the human evaluation results obtained in section 3.2. However, for some of the speakers, the system has difficulty in correctly determining the arousal-valence values, reflected in the presence of the non-monotonic evolution of their arousal and valence curves.



**Fig. 9.** Percentage ratio of speakers identified as expressing *negative* emotions. The system trained in the speaker-independent manner demonstrates good performance, consistent with human evaluation.



**Fig. 10.** Arousal and valence evolution vs. day, for each timestamp. Thin lines represent individual speaker evolutions, while the data labeled thick lines represent the average values for all speakers. The system trained in the speaker-independent manner demonstrates good performance, mostly consistent with human evaluation.

## 4. Conclusions

We have discussed the relevance and importance of monitoring the affective content of a subject's speech for sensitive applications such as forensics and law enforcement operations (surveillance, emergency services, etc.). We have investigated speech emotion remanence on short (under 1 hour) and long (5 days) timescales, different than the ones used in most speech emotion recognition research (usually very short, under a few minutes; or very long and sparse, over months and years) and more relevant for the targeted applications. We hypothesized 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.

We have developed a speech dataset comprising 270 recordings acquired over 5 days through conversations with 18 students who were behind on their university exams, and were studying in order to attempt them for the second or third time. Thus, the upcoming exams and the potential consequences of failing them represent the emotionally charged event. Human evaluators were asked to listen to the recordings and label them in terms of the identified emotional classes (grouped into *negative* emotional classes and the *neutral* state) and of arousal valence affect space values.

Analyzing the annotations made by the evaluators, we proved that the subjects' affective response was significantly higher as the emotionally charged event approached, and short time remanence was observed even 15 minutes after the initial interaction, or even after 30 minutes when under the added influence of the event's imminence. We show that the arousal increases (higher intensity affective response) as the event draws closer, while the valence decreases (more negative affective response), again supporting the second hypothesis, and suggesting that such patterns would be relevant for the targeted applications.

We proposed and implemented a speech emotion recognition system using artificial neural networks based on multilayer perceptron models, obtaining good performance (up to **72.7%** accuracy) when training in a speaker-independent manner, and yielding classification and regression results consistent with those given by human evaluation, supporting the possibility and usefulness of using machine learning systems to monitor affective responses in order to automatically detect the patterns associated with the behaviors relevant for forensic and law enforcement applications, facilitating intervention and prevention.

## References

- [1] EKMAN P., *Emotions revealed: recognizing faces and feelings to improve communication and emotional life*, 2nd ed., Henry Holt and Company: New York, NY, 2007, pp. 38–82.
- [2] SCHULLER B., *Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends*, Communications of the ACM **61**(5), pp. 90–99, 2018.
- [3] SCHULLER D., SCHULLER B., *The age of artificial emotional intelligence*, Computer **51**(9), pp. 38–46, 2018.
- [4] GUNES H., SCHULLER B., *Categorical and dimensional affect analysis in continuous input: Current trends and future directions*, Journal of Image and Vision Computing **31**(2), pp. 120–136, 2013.

- [5] GOMEZ I., *Artificial intelligence & machine learning in public safety*, EENA project report, 2019. [Online]. Available: <https://eena.org/knowledge-hub/documents/artificial-intelligence-machine-learning-in-public-safety/>. [Accessed: 14 Feb. 2022].
- [6] KRETZ D.R., GRANDERSON C.W., *An interdisciplinary approach to studying and improving terrorism analysis*, Proceedings of the 2013 IEEE International Conference on Intelligence and Security Informatics (ISI), Seattle, WA, USA, pp. 157–159, 2013.
- [7] COLANGELO F., BATTISTI F., CARLI M., NERI A., CALABRO F., *Enhancing audio surveillance with hierarchical recurrent neural networks*, Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, pp. 1–6, 2017.
- [8] HOU J., LI X., ZHU R., ZHU C., WEI Z., ZHANG C., *A neural relation extraction model for distant supervision in counter-terrorism scenario*, IEEE Access **8**, pp. 225088–225096, 2020.
- [9] VERDUYN P., LAVRIJSEN S., *Which emotions last longest and why: The role of event importance and rumination*, Motivation and Emotion **39**, pp. 119–127, 2015.
- [10] AKCAY M.B., OGUZ K., *Speech emotion recognition: emotional models, databases, features, pre-processing methods, supporting modalities, and classifiers*, Speech Communication **116**, pp. 56–76, 2020.
- [11] SWAIN M., ROUTRAY A., KABISATPATHY P., *Databases, features, and classifiers for speech emotion recognition: a review*, International Journal of Speech Technology **21**, pp. 93–120, 2018.
- [12] EKMAN P., *An argument for basic emotions*, Cognition and Emotion **6**(3), pp. 169–200, 1992.
- [13] GENDRON M., BARRETT L.F., *Reconstructing the past: a century of ideas about emotion in psychology*, Emotion Review **1**(4), pp. 316–339, 2009.
- [14] RUSSELL J.A., *A circumplex model of affect*, Journal of Personality and Social Psychology **39**(6), pp. 1161–1178, 1980.
- [15] WATSON D., WIESE D., VAIDYA J., TELLEGEN A., *The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence*, Journal of Personality and Social Psychology **76**(5), pp. 820–838, 1999.
- [16] RUBIN D.C., TALARICO J.M., *A comparison of dimensional models of emotion: evidence from emotions, prototypical events, autobiographical memories, and words*, Memory **17**(8), pp. 802–808, 2009.
- [17] MIHALACHE S., BURILEANU D., *Dimensional models for continuous-to-discrete affect mapping in speech emotion recognition*, University Politehnica of Bucharest Scientific Bulletin, Series C **83**(4), pp. 137–148, 2021.
- [18] TRNKA M., DARJAA S., RITOMSKY M., SABO R., RUSKO M., SCHAPER M., STELKENS-KOBSCHE T., *Mapping discrete emotions in the dimensional space: an acoustic approach*, Electronics **10**(23), 2950, 2021.
- [19] MORRISON G.S., ROSE P., ZHANG C., *Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice*, Australian Journal of Forensic Sciences **44**(2), pp. 155–167, 2012.
- [20] BRADLEY M.M., LANG P.J., *Measuring emotion: The self-assessment manikin and the semantic differential*, Journal of Behavior Therapy and Experimental Psychiatry **25**(1), pp. 49–59, 1994.
- [21] MIHALACHE S., POP G., BURILEANU D., *Introducing the RODECAR database for deceptive speech detection*, Proceedings of the 10th International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, pp. 1–6, 2019.
- [22] SU B.-H., LEE C.-C., *A conditional cycle emotion GAN for cross corpus speech emotion recognition*, Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, pp. 351–357, 2021.

- [23] ZHANG J., JIANG L., ZONG Y., ZHENG W., ZHAO L., *Cross-corpus speech emotion recognition using joint distribution adaptive regression*, Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pp. 3790–3794, 2021.
- [24] LIU N., ZHANG B., LIU B., SHI J., YANG L., LI Z., ZHU J., *Transfer subspace learning for unsupervised cross-corpus speech emotion recognition*, IEEE Access **9**, pp. 95925–95937, 2021.
- [25] ZHANG W., SONG P., CHEN D., SHENG C., ZHANG W., *Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression*, accepted for publication in IEEE Transactions on Cognitive and Developmental Systems, 2021. [Early Access, doi:10.1109/TCDS.2021.3055524].
- [26] BOWIE N.G., *40 terrorism databases and data sets: a new inventory*, Perspectives on Terrorism **15**(2), pp. 147–161, 2021.
- [27] SCHULLER D., SCHULLER B., *A review on five recent and near-future developments in computational processing of emotion in the human voice*, Emotion Review **13**(1), pp. 44–50, 2021.
- [28] LATIF S., RANA R., KHALIFA S., JURDAK R., QADIR J., SCHULLER B., *Survey of deep representation learning for speech emotion recognition*, accepted for publication in IEEE Transactions on Affective Computing, 2021. [Early Access, doi:10.1109/TAFFC.2021.3114365].
- [29] ROY T., MARWALA T., CHAKRAVERTY S., *A survey of classification techniques in speech emotion recognition*, in: S. Chakraverty (Ed.): *Mathematical methods in interdisciplinary sciences*, John Wiley & Sons: Hoboken, NJ, 2020, pp. 33–48.
- [30] SCHULLER B., STEIDL S., BATLINER A., EPPS J., EYBEN F., RINGEVAL F., MARCHI E., ZHANG Y., *The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load*, Proceedings of INTERSPEECH 2014, Singapore, pp. 427–431, 2014.
- [31] CHASPARI T., DIMITRIADIS D., MARAGOS P., *Emotion classification of speech using modulation features*, Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, pp. 1552–1556, 2014.
- [32] MIHALACHE S., BURILEANU D., BURILEANU C., *Detecting psychological stress from speech using deep neural networks and ensemble classifiers*, Proceedings of the 11th International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, pp. 74–79, 2021.
- [33] MIHALACHE S., BURILEANU D., POP G., BURILEANU C., *Modulation-based speech emotion recognition with reconstruction error feature expansion*, Proceedings of the 10th International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, pp. 1–6, 2019.
- [34] KASI K., ZAHORIAN S.A., *Yet another algorithm for pitch tracking*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, USA, pp. I-361-I-364, 2002.
- [35] JADOUL Y., THOMPSON B., DE BOER B., *Introducing Parselmouth: A Python interface to Praat*, Journal of Phonetics **71**, pp. 1–15, 2018.
- [36] KINGMA D.P., BA, J.L., *Adam: A method for stochastic optimization*, Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, pp. 1–15, 2015.