# Speech Emotion Recognition for Performance Interaction[1]

**NIKOLAOS VRYZAS,** *AES Student Member*, **RIGAS KOTSAKIS,** **AIKATERINI LIATSOU,**

(nvryzas@jour.auth.gr) (rkotsakis@gmail.com) (liatsou@thea.auth.gr)

**CHARALAMPOS DIMOULAS,** *AES Member*, **AND** **GEORGE KALLIRIS,**

(babis@eng.auth.gr) (gkal@jour.auth.gr)

*Aristotle University of Thessaloniki, Thessaloniki, Greece*

This paper investigates the applicability of machine-driven Speech Emotion Recognition (SER) towards the augmentation of theatrical performances and interactions (e.g. controlling stage color /light, stimulating active audience engagement, actors' interactive training, etc.). For the needs of the classification experiments, the Acted Emotional Speech Dynamic Database (AESDD) is developed, containing spoken utterances by 5 actors in 5 emotions. Several audio features and various classification techniques are implemented and evaluated, based on their performance with the AESDD, while also comparing to the well-known SAVEE database. The trained classifier is integrated in a novel application that performs live SER, fitting the needs of actors training, while simultaneously augmenting the AESDD repository.

[1] cite this paper as: Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C. A., & Kalliris, G. (2018). Speech Emotion Recognition for Performance Interaction. *Journal of the Audio Engineering Society*, *66*(6), 457-467.

## 0 INTRODUCTION

Speech Emotion Recognition (SER) has become a very popular scientific topic with extended multidisciplinary applicability. Many approaches for automatic SER have been proposed, yet the mechanisms of spoken emotion perception have not been fully revealed, neither the achieved performance has reached its full potentials [1]-[18]. The current paper investigates the usefulness of the emotional speech information towards automated control staging mechanisms in theatrical performances. In particular, a framework is introduced for detecting and recognizing spoken content emotional transitions, aiming at augmenting theatrical performance, production and management. In the simplest use scenario, this can be part of an automated lighting system, controlled by sentiment analysis modules. In addition, SER-driven documentation of theatrical plays can also be deployed, serving easier content archiving, searching and retrieval in batch mode, while promoting active audience engagement and actors' feedback /live training in a real-time scenario [1].

While senses are often considered independent from a cognitive point of view, cross-modal influence in perception has been reported [19]. Multimodal interaction is the most common strategy in everyday social activities [1], [5]-[6], [20]. Since multimodality may increase the efficiency in human-computer interaction, it can also address to more creative side of the human perception, such as multisensory excitement that has always been challenging in arts [21], [22]. Previous work on automatic sound-to-light transformations makes use of particular sound elements, such as audio pitch and intensity [23]. Temporal and spectral characteristics of musical pieces can be exploited for controlling the intensity, duration and color of the lights [11], [24]. On the other hand, the probabilistic aspect of interaction via SER matching supports the temporality /adaptivity of theatrical act, promoting a less solid character of the performance, thus becoming more transient. Every show provides a quite unique audio and visual experience [25], and moreover the spectators perceive much more than the theatrical play itself. In this context, a good proportion of contemporary performances are composed by a variety of physical and technological expressive means, aiming at developing a new aesthetic that differentiates from conventional theatrical tradition [26]. Multimodal emotion detection for the control of stage lighting in music shows have been proposed, using visual and motion, and music information retrieval cues [27].

The current work focuses on the use of SER modalities as a mean of augmenting the emotional extra-linguistic elements of theatrical speech (i.e. stage color, emotion tags, etc.) [1], [28]. Apart from aesthetic purposes, this augmentation aims at providing the possibility of multilevel perception of theatrical performance. The rest of the paper is organized as follows. Related work on speech emotion recognition is presented in the next section. The proposed framework is outlined in section 2, listing typical generalized and adaptive scenarios, with possible extension and augmentation facilities. An application that calculates speech emotion using pre-trained classifiers is presented,

providing a GUI for actors training, while also augmenting the existing databases. Emotional speech databases are reviewed next, stating the current limitations and the motivation for developing a new repository of combined datasets. The main SER module is described in detail in section 4, followed by the experimental results and discussion (section 5), which are presented along with conclusion and future work remarks.

## 1 RELATED WORK AND BASIC PRINCIPLES ON SPEECH EMOTION RECOGNITION

Speech has always been one of the most common modalities in human-machine interaction, which is quite expected, considering its role in physical and mediated communication. While speech recognition is a thoroughly researched scientific field, the vocabulary, syntax and grammar of spoken language cannot actually withhold the whole information of speech. This is because the deep understanding of the context and emotional state of the speaker can induce a different meaning. There is a wide area of speech (and emotion) recognition applications, including automatic translations, patient surveillance and assisted diagnosis, validation of the well-being of drivers and pilots, emotional state monitoring in various /violent situations, and others [1]-[7], [16].

SER process lies under the scientific fields of general audio detection and segmentation, including demanding semantic analysis tasks. Therefore, pattern classification and data mining strategies are usually involved, where feature-based Machine Learning (ML) algorithms are dominant. The more reliable approaches involve the training of a model, having as inputs multivariate and salient audio features, extracted from pre-labeled /annotated data. A predefined pattern classification taxonomy is also implicated, where both audio and audio-driven multimedia processing tasks are involved [5]-[6], [13]-[18], [29]-[37]. However, this task involves a certain degree of ambiguity, because human emotion is not always correctly /undoubtedly perceived and measured via strict quantitative metrics, while sometimes is not clearly defined. In this context, a lot of research has been conducted for defining applicable sentiment analysis models and classification schemes for recognizing the associated patterns of emotion. While a number of around 150 different emotional states has been identified for the English language, related psychological research discriminated two basic dimensions: activation and valence [1], [5]-[6], [28]-[29], [35]-[36], [38]. These components form a two-dimensional space (three in some cases, combined with emotional tension), where emotions are located /pinpointed based on their inclusion of the aforementioned properties.

In typical ML approaches, depending on the problem under study and the associated classification taxonomies, SER research is mainly a task of detecting and choosing the most salient acoustic features to feed the appropriate classifiers. The main goal is to achieve accurate /efficient classification scores that could be achieved even at unknown samples, thus offering good generalization. Among others, selection of time-frequency resolution and proper windowing

configuration is important, since it is related to the non-stationary signal behavior, as well as to the accurate estimation of the involved features. Also, time overlapping strategies, structural audio attributes and temporal feature integration techniques are engaged for capturing dependencies of successive audio frames, while serving the required resolutions in the multiple feature domains [30]-[37], [39]-[41].

While a lot of research has been conducted on explicit SER works [7]-[8], [12]-[17], more generic sentiment analysis of audio content has become also very popular (i.e. music and soundtrack classification and generation, emotional soundscape recognition, content evaluation and profiling, etc.). [3]-[6], [9]-[11], [41]-[43]. Sophisticated cross-modal methods utilize multiple decision-making systems, fusing the advantages of the involved modalities (i.e. text and natural language processing, speech /audio parameters, visual features and facial expressions in a multimodal fashion) [1], [16]-[18], [30], [35]-[37]. Deep Learning (DL) strategies form a recent trend in ML, which is widely used in Automatic Speech Recognition (ASR) applications. It is appreciated as an approach that achieves higher accuracy, consuming less human effort [42]. The input of a deep neural network for speech recognition can be typical acoustic features, or even raw sound information, such as spectrograms. In DL, the hidden layers of Neural Networks extract unsupervised features from the raw data that fit the training procedure, or higher-level representations of supervised extracted features [44]-[47]. Especially in SER applications, where the effectiveness of common features is unclear, DL feature extraction can improve performance [45]. However, big training datasets are a prerequisite for the training of accurate and generalized models [42].

## 2 THE PROPOSED FRAMEWORK

### 2.1 Framework for performance augmentations and the Collaborative Model for database dynamic evolution

A SER framework for controlling stage lighting has been proposed in [1], aiming at offering augmented drama experience. In this paper, the initial framework is further elaborated emphasizing mostly on the algorithmic aspects of SER, while offering additional performance automations (Fig.1a). The main goal of the current work is focused on the augmentation of the user interaction, based on the conduction of speech emotion analysis. As already implied, both live and offline /batch processing modes can be deployed depending on the targeted utilities and the available computational power. As Fig.1a depicts, a set of microphones is installed on stage in order to capture the acted speech of the performance. These audio signals are routed to a computing environment (i.e. a personal computer –PC or more sophisticated /embedded hardware, like DSPs), where the emotion analysis system is running. An audio feature-based SER module lays in the core of that system, comprising the feature extraction modality, which operates at sound frame level, and a pre-trained classifier, which uses a decision model to identify the dominant emotion in the current audio

frame. Since different emotions can be matched to different states of a “color wheel” of emotions, the estimated distinct colors can be thought as the classification outcomes, which can be then propagated to the interface controlling stage lighting. In this straightforward use-case scenario, the emotion-color matching process can be quite subjective and adaptive to the play, the settings and the artistic view. The contributors of each production have the freedom and capability to make their own correlation /selection, which serves the specific aesthetic and theatrical purposes of the production. Moreover, it has to be noted that besides the aesthetic selections for the theatrical productions, the hosting venue has to be taken into consideration because of specific requirements of gear setup, model implementation (stage position, lighting availability and number of spotlights, size of the venue etc.). The output (colored lighting selection) that is extracted by the classification model may serve either as a proposed exclusive lighting scheme or as an indication to supplement the original lighting design of the play, since it can be switched on and off during different parts. The output (e.g. colored lighting selection) that is extracted by the classification model may serve either as a proposed exclusive lighting scheme or as an indication to supplement the original lighting design of the play, since it can be switched on and off during different parts. While a more comprehensive review of the color mapping of emotions and the aesthetic concept of the basic framework can be found in [1], additional potential uses are explained in the next sections. Hence the extracted emotional outcomes can be further combined with other semantic tags to provide performance documentation and indexing, to trigger interesting content enhancements and interactions with the audience (including off-line user), to offer training utilities along with dataset augmentation functionalities.

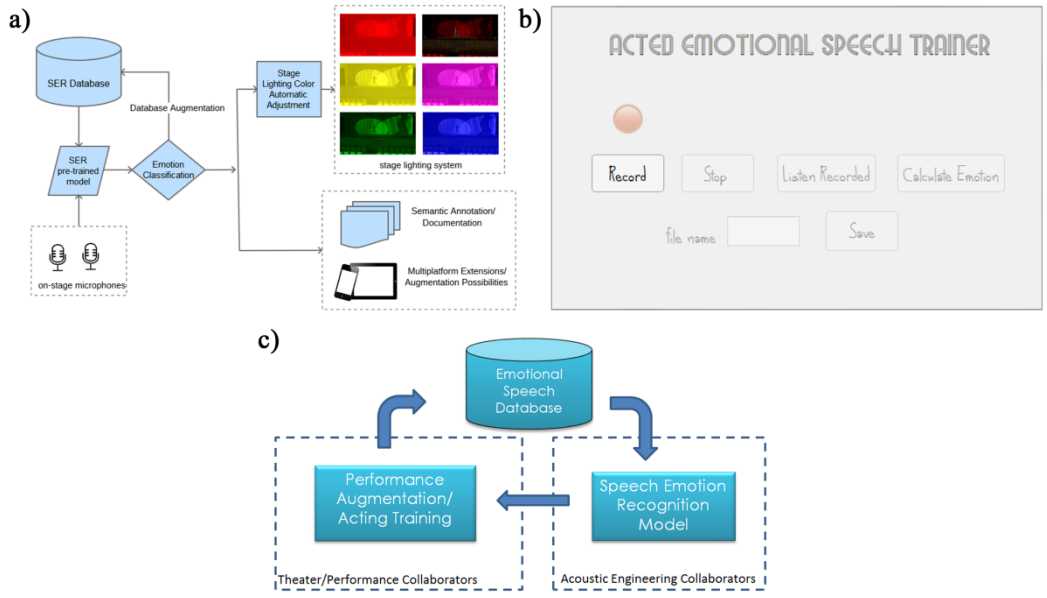


Fig. 1 a) The proposed framework for performance augmentation [1] b) GUI of the application for emotional acting training and database augmentation c) the Collaborative Model.

The microphone positions setup has a crucial role in the whole implementation and emotion identification procedure, because the sound source localization may lead to different classification results. Consequently, the actor's position and area of movement could also be determinative parameters for the speech controlled stage lighting process [1], [30]. Speech emotion recognition is considered a very difficult and sophisticated pattern recognition task, even when the discrimination process is developed and performed by humans [47]. In scenarios where increased level of control of the performance lighting is demanded, a generalized automatic classification model may not be suitable (because of the existence of multivariate patterns of voices, aesthetic choices, differentiated performance requirements, etc.). Aiming at eliminating ambiguity, the proposed system uses an automatic SER model that attempts to adaptively perform well, while taking into consideration the specific actors, the context and verbal content of the specific performance, etc. Even in cases where real-time processing (SER) is not feasible, detected sentiment transitions can help in the temporal annotation of the perceived emotions of the theatrical plays, thus offering many extensions and augmentation possibilities. As aforementioned, emotional tags can be used for proper documenting and managing the captured performance content, while both actors and audience can be involved in various interaction scenarios (e.g. live feedback /training, more active participation, non-linear storytelling, etc.).

### 2.2 A SER application for acting training and database augmentation

The creation of a big emotional speech database, suitable for the training of a speaker- or content-independent generalized model, can be a painful and costly process.

The current framework attempts to break the barrier between the engineering and theater world, proposing a collaborative model of mutual benefit. Performance collaborators (theatrical teams or drama schools) can produce new, high-quality data, while exploring the potentials of the proposed augmentation.

In this context, audio data can be collected during two long-term processes:

- rehearsing projects that adopt the augmentation framework,
- acting training

Staging a show requires time-consuming and painful preparations and experimentations, especially when this show plans to implement speech emotion recognition for performance augmentations. Nonetheless, actors' training to express different emotions is an every-day practice in drama schools and theatrical projects. In several common theatrical exercises, the expressed emotion does not have to correspond to the affective context of the utterance, thus separating lingual and extra-lingual information of speech. Hence, the availability of an easy-to-use platform, for capturing and labeling acted utterances of emotional speech, facilitates the gradual growth of the constructed AESDD repository. Moreover, since both rehearsals and acting education processes are supervised by directors and field experts, the quality

of the produced utterances is guaranteed. As the database grows, engineering collaborators can train more accurate and generalized SER models to provide for future performance augmentations, as exhibited in Fig.1c.

To assist this cause, an application was created, and a screenshot of the GUI is shown in Fig.1b. The environment is user friendly and provides the actor with the tools to record a theatrical utterance and listen to it. A pre-trained SER classifier is integrated into the application code that calculates the emotion of the spoken utterance. The actor can repeat the process until he is satisfied with the outcome, before saving the results into an appropriated file. This application serves four major purposes:

1. It can be used as an educational tool for acting training (i.e. on how to map /project specific emotional expressions), which can serve additional performances besides drama (e.g. news-casting, lectures - education, presentations on working meetings, project, conference, etc.).
2. It can be useful in the context of a specific theatrical project. The director can utilize it to guide the actors or to build a dedicated database for the training of a project-dependent classifier (e.g. for later exploitation in emotional transcription /annotation and indexing purposes).
3. It can be used for the augmentation of the initialized emotional speech database, purposing to address the SER problem more robustly.
4. It will allow monitoring of actor-adapted salient audio features, thus leading to personalized, more adaptive and accurate SER classifiers, while providing useful insights for the applicability of the various audio features in different SER scenarios. For this reason, an initial superset of audio features is also required.

Using a pre-trained classifier based on the already known samples allows the evolution of the application in future versions. While more data is gathered by theatrical collaborators using the system, the audio engineering collaborators are capable of training new discrimination models on the augmented database. The classification models are interchangeable, thus leading to evolved versions of the application with the same functionalities and more robust/generalized SER performances. In project-dependent SER, adaptive data concerning a given project can be collected, referring to the specific actors and utterances that are present. The audio engineering collaborators can build classifiers based on the dedicated dataset as input training samples, to design automations /augmentations that are suitable in the specific context of the performance. In any case, the recorded files will be used for the formulation of an extended database, serving the more efficient exploration of SER and the creation of more robust and generalized models. Figure 2 depicts the diagram with the data-flow of this joint “acting training” / “database augmentation” functionality.

A fully operational prototype has been implemented in Matlab 2017a, taking advantage of the offered App Designer and the associated Graphical User Interfacing (GUI) components (Fig 1b), thus allowing some initial proof-of-concept

experimentation with representative users. Based on previous experience [31], the outmost target is the elaboration of this utility into a Software as a Service cloud module, which could support the database augmentation process through Web or mobile crowdsourcing procedures.

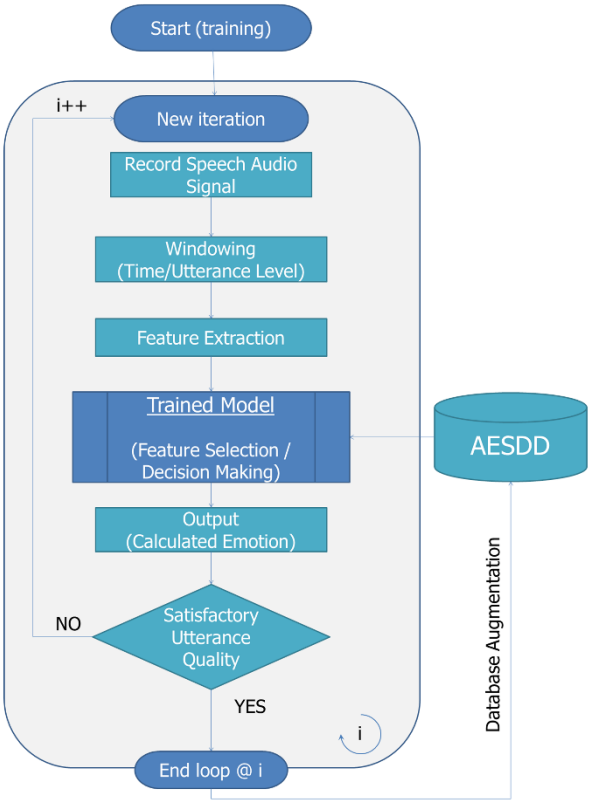


Fig. 2 Data-flow of the joint "acting training"/"database augmentation" functionality

## 3 EMOTIONAL SPEECH DATABASE

### 3.1 Databases short review and limitations

There are several databases proposed for SER purposes that contain utterances of acted or natural speech, with some of them being publicly available. Most databases differ in certain aspects concerning the number and selection of emotion classes, besides the spoken language. Many scientific works on SER utilize some of these known databases in order to evaluate, rank and compare different features and discrimination techniques [17]. It has to be noted that the initial creation of such emotionally labeled speech dataset is a very difficult and sophisticated task. It involves the collection of spontaneous or acted emotional speech that needs to be labeled and segmented, either in predefined or unknown emotion categories. A possible approach is using resampling techniques for altering the samples of existing databases, enlarging their population [4]. Taking into consideration that there are limited collections of the aforementioned data, the formulation of an original (from scratch) database is attempted in the presented work. This initiative could serve applicability in various emotion recognition problems and, of course, it poses significant amount

of originality and innovation. In this context, many factors have to be carefully addressed, such as the selection of an experienced acting team, the kind of speech /utterances that have to be pronounced, professional guidance and supervision that has to be supported, etc. This venture implicates provisions for augmentation of the database, featuring long-term collaborative properties.

### 3.2 Initial experiments with the SAVEE database

The data collection for the initial experiments of speech emotion classification derives from the Surrey Audio-Visual Expressed Emotion (SAVEE) database [48]. Specifically, this database consists of 480 phrase recordings of 4 British English male actors in 7 different emotion patterns, namely anger, fear, surprise, happiness, sadness, disgust and neutral. The 6 basic emotions that were initially considered are *anger*, *fear*, *happiness*, *sadness*, *disgust* and *surprise* [48]. Since surprise could refer to /indicate ambiguous overlapping sentimental states (e.g. Surprise could be positive or negative oriented) the most clear /undisputable sentiments were implicated in the classification experiments, resulting in the selection of the first five emotions. These emotions correspond also to the categories of the classification problem. This approach is common in related sentiment analysis tasks, where bimodal and hierarchical classification schemes are adopted to serve specific emotional classification needs [5]-[6]. The applicability of the formed scheme was empirically validated, following the guidelines and references of such previous /generic works [5]-[6], [35]-[36].

### 3.3 Acted Emotional Speech Dynamic Database (AESDD)

Having the SAVEE database as reference, the need for a big database that overcomes the limitations of existing emotional speech databases, led to the creation of the original Acted Emotional Speech Dynamic Database (AESDD).

As the name implies, the main characteristics of this database are:

- It consists of acted speech utterances
- It is an ever-growing database

The database of acted emotional speech is initialized with the first audio data collection and the experiments on its first form are presented in this paper. This initialization process sets the foundation for the creation of subsequent augmented versions of the database, thus favoring future experiments for the development of generalized and robust SER models.

#### 3.3.1 Creating the Acted Emotional Speech Dynamic Database (AESDD)

For the creation of the initial form of the AESDD, five (5) professional actors were employed, aging from 25 to 30 years old, while the group consisted of 2 male and 3 female actors (in order to succeed both genders existence, contrary to SAVEE database). Furthermore, while the SAVEE database contains utterances in English, the AESDD database utterances are in Greek.

The recordings took place in the sound studio of the Laboratory of Electronic Media (Aristotle University of Thessaloniki, Greece), offering an appropriate acoustic environment, thus ensuring high quality recordings.

Since the current work is focused on theatrical productions, the spoken /recorded phrases had to derive from theatrical scripts. Specifically, 19 utterances were chosen from different theatrical plays for the database formulation, based on the criterion of the ambiguity of their emotional context. The actors expressed these 19 sentences in Greek language in 5 different emotional contexts, namely *happiness*, *sadness*, *anger*, *fear* and *disgust*, because of their undisputed form, as stated above. Moreover, for every emotion, one extra /improvised utterance was recorded, while more than one recording were used for some utterances, resulting in around 500 utterances of emotional speech (5 actors **x** 5 emotions **x** 20 utterances). As all actors recorded the same utterances for all 5 emotions, it is ensured that the training process is not user-dependent or the verbal content of the utterances. A scientific expert in dramatology was present in order to supervise the recordings, to guide the actors and to make the proper adjustments/ corrections when needed, ensuring the quality and suitability of the acted speech.

During the preprocessing, all the utterances were properly normalized (at -3dB peak) and given appropriate filenames (related info is provided in the readme file of the link below). The database was uploaded to the cloud along with proper documentation, explaining important creation and organizing details, and is publicly available at https://goo.gl/ezbmGA

#### 3.3.2 Dynamic evolution of AESDD

The current initial state of the AESDD database is used for the training of the classifier in the first version of the application presented in 2.2. The size of the database (5 actors and 500 utterances in total) was chosen to fit the scenario of a typical project-dependent database, crafted to suit the needs of an implementation for a single performance. In this context, the SER performance for such a scenario is tested. While the application is in use by the theater collaborators, new utterances will be added and uploaded online publically, making the AESDD and ever-growing database through time.

## 4 SPEECH EMOTION RECOGNITION MODULE

### 4.1 Feature extraction and Evaluation

For the training process that has been conducted for both databases (SAVEE, AESDD), audio features are extracted from the labeled samples, forming the initial ground truth basis of the experiments. The initial feature vector consists of popular audio parameters that have already been tested in previous research works [1], [6], [30]-[37], resulting in efficient performances during the experimentation step. Specifically, as depicted in Table 1, the set feature vector

includes time domain characteristics (number of peaks -Npeaks, RMS, Low Energy, Event Density, Tempo and Pulse Clarity, Zero Cross Rate, etc.), spectral domain parameters (Rolloff and Brightness measures for different frequency thresholds, Irregularity, Inharmonicity, Mode, Centroid, Spread, Skewness, Kurtosis, Flatness, Entropy), as well as thirteen cepstral (Mel Frequency Cepstral Coefficients -MFCCs).

Table 1 Feature Vector.

| # | Domain | Features |
|---|---|---|
| 1-7 | **Time** | Npeaks, RMS, Low Energy,<br>Event Density, Tempo, Pulse Clarity, Zero Cross Rate |
| 8-21 | **Frequency** | Rolloff($F_R$), $F_R$= 30, 50, 70, 90<br>Bright(B), B= 0.5, 1, 1.5, 2, 3, 4, 8 kHz<br>Irregularity, Inharmonicity, Mode |
| 22-27 | **Spectral Statistics** | Centroid, Spread, Skewness, Kurtosis, Flatness, Entropy |
| 28-40 | **Cepstral** | MFCC(i), i=1-13 |
| 41-43 | **Emotion** | Activity, Valence, Tension |

In addition, features from Music Information Retrieval (MIR) literature that address the nature of dimensional emotion analysis were used, namely Activity, Valence and Tension. These are regression models integrating many other low- and high-level musical features, but in the current work we consider their outputs as features [49]-[52]. The elusive nature of the emotion recognition problem [52] makes the intuitive process of choosing the most relevant features a difficult process, since SER is subjective and unclear, even for human listeners. As already explained, an initial superset of audio features is required for being able to monitor actor-, play- or class-/emotional state-adapted salient feature ranking. In addition, even in cases that sole MIR parameters are not very useful in non-music audio semantics, the experience has shown that their combination with other features is quite advantageous even for General Audio Detection and Classification Tasks (GADC, i.e. radio broadcasted shows indexing, noise – silence detection, speech – music discrimination, speaker identification or even audio driven emotion recognition, etc.) [5]-[9], [30]-[37]. This is the reason why we extracted an initial extended audio feature vector to explore the efficiency of audio properties in the subsequent experiments.

Before proceeding to the training process, the chosen parameters are subjected to evaluation through different attribute-selection algorithms, ordering to select the most salient features for the given problem, while avoiding increased cross-correlation between them. In this context, the InfoGain Attribute Evaluation algorithm was applied in the formulated

ground truth data, aiming at evaluating features based on entropy metrics. Fig.3 presents ranking outcomes for both databases (AESDD and SAVEE). The depicted results verify the applicability of the initially selected parameters, where slight variations are observed in the order of appearance of some features (RMS, Valence, MFCCs, Activity, etc.).

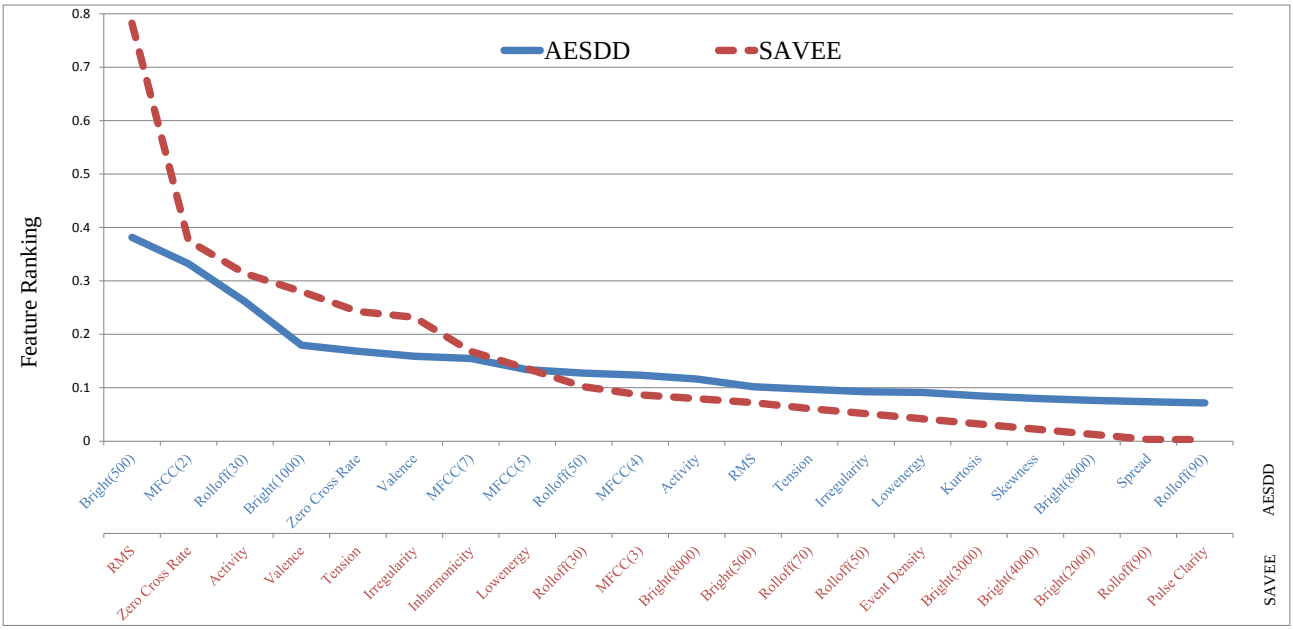


Fig. 3 Feature ranking for both databases (AESDD and SAVEE) using the WEKA Infogain Evaluation alogorithm

### 4.2 Training and classification process

The audio features that were extracted for each audio frame of the databases were utilized for training different classification algorithms. Several ML models were tested and compared in order to determine the most efficient configurations, using the popular Weka software, an open-source data mining environment [53]. In this context, Multi-Layer Perceptron (MLP) neural networks, Linear and Logarithmic Regressions (Log.Reg.) schemes, Decision Trees (J48, forests, LMT), Bayesian models, Support Vector Machines (SVM with polynomial kernel) and Nearest Neighbor approaches were implemented, taking into consideration the comparative advantages of complexity, computational cost, effectiveness, etc. of each method.

For the training purposes, the k-fold validation process was employed; the input data were divided in k-subsets, while the (k-1) subsets were utilized for training the classifier and the remaining one for testing and estimating the discrimination accuracy after k iterations. Several k-fold sessions were experimented (with various number of folds), concluding in k=8 folds as the more effective, for both AESDD and SAVEE sets. The pattern recognition rate was used as the classification performance metric (P), derived from the respective confusion matrices as the ratio of the number of

the correctly classified instances to the total number of input samples. Partial accuracy scores (Pc) were also estimated for evaluating the partial class (c) efficiencies [33]-[34].

## 5 RESULTS AND DISCUSSION

Following a hierarchical approach, the input samples were initially divided in two coarse categories (positive P+ and negative P- emotions) [6], [35]. Thus, the ground truth of the AESDD database involved 400 audio samples of negative emotions (*anger*, *fear*, *disgust*, *sadness*), with a single positive emotion (*happiness*), represented by 100 samples. This imbalance was somehow treated through oversampling (OS) of the deficient positive class [53].

Table 2 demonstrates the performance rates before and after OS, with the latter being used with MLP (one hidden layer consisting of 21 neurons with sigmoid trigger function and a linear output layer) and Log.Reg., which provided the highest classification rates. These configurations were validated based on theoretical assumptions, previous experience [31]-[37] and fine tuning through trial and error testing. Since high performance rates were noted in the first hierarchical classification step, only the happiness class was efficiently identified (98.23% for ANS and 80.56% for Log.Reg.), while the negative emotions are yet to be discriminated, after removing the misclassified audio samples, in order to avoid the error propagation. Table 3 and Fig.4 present the classification performances (P, Pa, Pd, Pf, Ps) for the negative emotions (*anger*, *disgust*, *fear* and *sadness*) of 5 different ML methods, leading to about 80% accuracy for MLP and LMT. It is easy to observe the increased partial discrimination rates (Pa, Ps) for the emotions of *anger* and *sadness* (above 80%).

Table 2 Classification Performances for Positive vs Negative Emotions before and after oversampling

| | P | P+ | P- |
|---|---|---|---|
| Log.Reg. | 83.93 | 50.51 | 92.1 |
| MLP | 81.75 | 49.49 | 89.63 |
| Log.Reg. (OS) | 79.03 | 80.56 | 77.53 |
| MLP (OS) | 93.38 | 98.23 | 88.64 |

Table 3 Pattern recognition rates of negative emotions, using various ML methods

| | P | Pa | Pd | Pf | Ps |
|---|---|---|---|---|---|
| Log.Reg. | 74.6 | 73.27 | 63.73 | 61 | 79.41 |
| MLP | 78.57 | 81.19 | 67.65 | 64 | 81.37 |
| Tree (LMT) | 79.56 | 83.17 | 65.69 | 64 | 85.29 |
| J48 | 59.72 | 61.39 | 44.12 | 33 | 60.78 |
| SVM | 73.83 | 86 | 66 | 59 | 88 |

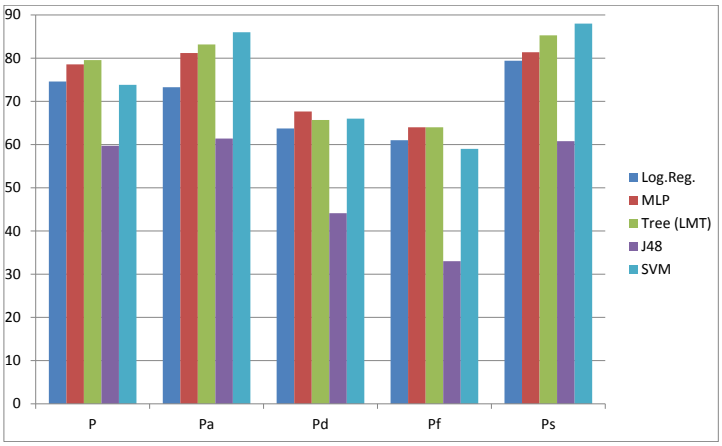


Fig. 4 Classification performances (P, Pa, Pd, Pf, Ps) for the negative emotions (anger, disgust, fear and sadness) via 5 machine learning methods (Logistic Regression, Multi Layer Perceptron, LMT-Tree, J48,SVM).

Taking into consideration the aforementioned results, the AESDD database proved to serve as an efficient data source for SER classification problems. This is important considering that the corresponding SAVEE scores reach about 70% performance through ML models, while even the subjective classification (conducted by humans) has been found to be between 66.5±2.5%, thus reflecting the increased ambiguity [47]. Related empirical observations were indicatively conducted for the newly formed dataset (AESDD), validating the (wanted) decreased vagueness /overlapping between classes, thus making it appropriate for initiating self-learning SER training sessions. Fig.5 depicts this comparative

advantage of AESDD database in all classes while utilizing the Log.Reg classification method, which received the best SAVEE scores (with the only exception of the *disgust* class).

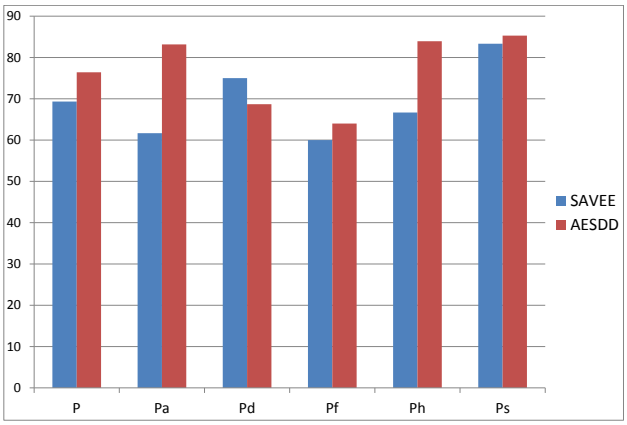


Fig.5 Comparative performance of logistic regression (Log. Reg.) for all 5 classes in both databases (SAVEE, AESDD).

## 6. FUTURE WORK

The results of this first experimentation are quite promising, especially concerning the construction of the new AESDD dataset. Equally important and encouraging is considered the gained experience, while collaborating with the theatrical acting team. Future plans include the implementation of the described collaborative model in practice. This involves:

- The support of staging shows that make use of the presented framework.
- The expansion of the AESDD repository.
- Testing mixed-language speech emotion classifiers.

Staging shows that implement the current framework of automated drama augmentation (i.e. assisted lighting, live emotional tagging and interaction, etc.) will allow the multivariate evaluation of the perceived experience, for both theater professionals and spectators. Meanwhile, the extension of the AESDD will be able to support better performing and more generalized classification models in the future. In this context, significant enlargement of the samples population will allow the deployment of Deep Learning strategies that seem to be very promising.

## 7. ACKNOWLEDGEMENTS

Part of this research has been financially supported by General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI), which supports N. Vryzas' PhD research. (Scholarship Code: 1900).

## REFERENCES


[1] N. Vryzas, A. Liatsou, R. Kotsakis, C. Dimoulas, and G. Kalliris, "Augmenting Drama: A Speech Emotion–Controlled Stage Lighting Framework," *Proceedings of AM '17* (August 2017) https://doi.org/10.1145/3123514.3123557

[2] B.Kostek, *Perception-based data processing in acoustics: applications to music information retrieval and psychophysiology of hearing* , vol. 3(Springer Science & Business Media, 2005).

[3] J.Fan, M. Thorogood and P. Pasquier, "Automatic Soundscape affect recognition using a dimensional approach," *J. Audio Eng. Soc*., vol.64,no.9, pp. 646-653 (2016), https://doi.org/10.17743/jaes.2016.0044.

[4] D.Williams, "Toward emotionally-congruent dynamic soundtrack generation," *J. Audio Eng. Soc*., vol.64, no.9, pp. 654-663 ( 2016), https://doi.org/10.17743/jaes.2016.0038.

[5] G. Kalliris, M. Matsiola, C. Dimoulas and A. Veglis, "Emotional Aspects and Quality of Experience for Multifactor Evaluation of Audiovisual Content", *International Journal of Monitoring and Surveillance Technologies Research*, vol.2,no.4, pp. 40-61 (2014 Oct./Dec.), https://doi.org/10.4018/IJMSTR.2014100103.

[6] R.Kotsakis, C. Dimoulas, G.Kalliris and A.Veglis, "Emotional Prediction and Content Profile Estimation in Evaluating Audiovisual Mediated Communication", *International Journal of Monitoring and Surveillance Technologies Research*,vol.2,no.4, pp.62-80, (2014 Oct./Dec.), https://doi.org/10.4018/IJMSTR.2014100104.

[7] J. Higueras-Soler, R. Gil-Pita, E. Alexandre and M. Rosa-Zurera, "Violence Prediction through Emotional Speech," presented at the *128th Convention of the Audio Engineering Society* (2010 May), convention paper 8003.

[8] H.K. Ha, N.K. Kim, W.K. Seong and H.K. Kim, "Noise-Robust Speech Emotion Recognition Using Denoising Autoencoder," presented at the *140th Audio Engineering Society Convention* (2016 May), convention e-Brief 260.

[9] F.Noroozi, D. Kaminska, T. Sapinski and G. Anbarjafari, "Supervised Vocal-Based Emotion Recognition Using Multiclass Support Vector Machine, Random Forests and Adaboost," *J. Audio Eng. Soc*. (Abstracts), vol. 65, pp.562-572 (2017 Jul.), https://doi.org/10.17743/jaes.2017.0022.

[10] I. Mohino-Herranz, H.A. Sánchez-Hevia, R. Gil-Pita and M. Rosa-Zurera, "Creation of New Virtual Patterns for Emotion Recognition through PSOLA," presented at the *136th Convention of the Audio Engineering Society* (2014 Apr.), convention paper 9037.

[11] T. Eerola, "Modeling Emotions in Music: Advances in Conceptual, Contextual and Validity Issues," presented at the *53rd International Audio Engineering Society Conference*, p. S1-1. (2014 Apr.).

[12] A.B. Ingale and D.S. Chaudhari, "Speech emotion recognition", *International Journal of Soft Computing and Engineering* (IJSCE) 2, 1.pp. 235-238 (2012).

[13] T. Vogt, E. André and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization," *Affect and emotion in human-computer interaction*, pp. 75-91 (Springer Berlin Heidelberg, 2008), https://doi.org/10.1007/978-3-540-85099-1_7.

[14] L.Chen, X. Mao, Y. Xue and L.L. Cheng, "Speech emotion recognition: Features and classification models," *Digital signal processing*, vol.22, no.6, pp. 1154-1160 (2012), https://doi.org/10.1016/j.dsp.2012.05.007.

[15] S. Wu, T.H. Falk and W.Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol.53, no.5, pp.768-785 (2011), https://doi.org/10.1016/j.specom.2010.08.013.

[16] M. El Ayadi, M.S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition* vol.44,no. 3, pp. 572-587 (2011), https://doi.org/10.1016/j.patcog.2010.09.020.

[17] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol.48, no.9, pp.1162-1181 (2006), https://doi.org/10.1016/j.specom.2006.04.003.

[18] S. Emerich, E. Lupu, and A. Apatean, "Emotions recognition by speech and facial expressions analysis," presented at the *17th European Signal Processing Conference (EUSIPCO)*, pp.1617-1621 (2009).

[19] M. Turk, "Multimodal interaction: A review," *Pattern Recognition Letters*, vol.36, pp.89-195 (2014), https://doi.org/10.1016/j.patrec.2013.07.003.

[20] L. Mondada and L, "Challenges of multimodality: Language and the body in social interaction," *Journal of Sociolinguistics*, vol.20, no.3, pp.336-366 (2016), https://doi.org/10.1111/josl.1_12177.

[21] C. Sinke, J. Neufeld, D. Wiswede, H.M. Emrich, S. Bleich, T.F. Münte and G.R. Szycik, "N1 enhancement in synesthesia during visual and audio–visual perception in semantic cross-modal conflict situations: an ERP study," Developing Synaesthesia, vol. 45 (2014), https://doi.org/10.3389/fnhum.2014.00021.

[22] A.J. Cohen, "From auditory focus, to auditory influence, to multimodal experience in psycho-musicological research," *Psychomusicology: Music, Mind, and Brain*, vol.26, no.2, pp.99-100 (2016 Jun.), http://dx.doi.org/10.1037/pmu0000149

[23] N. Osmanovic, "Low-Latency Conversion of Audible Guitar Tones into Visible Light Colors," presented at the *127th Convention of the Audio Engineering Society* (2009 Oct.), convention paper 7859.

[24] R. Dahyot, G.Kearney and C.Kelly, "Visual Enhancement Using Multiple Audio Streams in Live Music Performance," presented at the *31st International Conference of the Audio Engineering Society: New Directions in High Resolution Audio*, p.12 (2007 Jun.).

[25] E. Fischer-Lichte, *The transformative power of performance: a new aesthetics*, (Routledge, 2008).

[26] W. Puchner, *Theatrical Science in 21st Century*, (Kichli, 2014).

[27] E.O. Bonde, E.K. Hansen, G. Triantafyllidis, "Auditory and Visual based Intelligent Lighting Design for Music Concerts," Eai Endrosed Trasactions on Creative Technologies, (2017), http://dx.doi.org/10.4108/eai.10-4-2018.154452

[28] N. Campbell, "Databases of emotional speech", presented at the *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (2000).

[29] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J.G.Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no.1,pp.32-80 (2001), https://doi.org/10.1109/79.911197

[30] N. Vryzas, R. Kotsakis, C.A. Dimoulas and G. Kalliris, "Investigating Multimodal Audiovisual Event Detection and Localization," *Proceedings of the Audio Mostly 2016*, pp. 97-104 (ACM, 2016), https://doi.org/10.1145/2986416.2986426.

[31] L. Vrysis, N. Tsipas, C.A. Dimoulas and G.Papanikolaou, "Crowdsourcing Audio Semantics by Means of Hybrid Bimodal Segmentation with Hierarchical Classification", *J. Audio Eng. Soc*. (Abstracts), vol. 64, issue 12, p. 1042 (2016 Dec.), https://doi.org/10.17743/jaes.2016.0051.

[32] L. Vrysis, N. Tsipas, C. Dimoulas and G. Papanikolaou, "Extending Temporal Feature Integration for Semantic Audio Analysis," presented at the *142th Convention of the Audio Engineering Society* (May 2017), convention paper 9808.

[33] R. Kotsakis, G. Kalliris and C. Dimoulas, "Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification," *Speech Communication*, vol.54,no. 6,, pp.743-762 (July 2012), https://doi.org/10.1016/j.specom.2012.01.004.

[34] R. Kotsakis, G. Kalliris and C. Dimoulas, "Investigation of salient audio-features for pattern-based semantic content analysis of radio productions," presented at the 132nd Convention of the Acoustic Engineering Society (April 2012), convention paper 8663.

[35] R. Kotsakis, C. Dimoulas, G. Kalliri and A. Veglis, "Emotional descriptors and Quality of Experience (QoE) metrics in evaluating Mediated Learning," *Proceedings of the IEEE 5th International Conference on Information, Intelligence, Systems and Applications* (IISA 2014), pp.232-237 (2014 Jul.), https://doi.org/10.1109/IISA.2014.6878744.

[36] G. Kalliris, M. Matsiola, C. Dimoulas and A. Veglis, "Emotional aspects in Quality of Experience and Learning (QoE & QoL) of Audiovisual Content in Mediated Learning," *Proceedings of the IEEE 5th International*

*Conference on Information, Intelligence, Systems and Applications (IISA 2014)*, pp.198-203 (2014 Jul.), https://doi.org/10.1109/IISA.2014.6878743

[37] N. Tsipas, L. Vrysis, C. Dimoulas, G. Papanikolaou, "Efficient audio-driven multimedia indexing through similarity-based speech/ music discrimination", *Multimedia Tools and Applications*, pp. 1-19 (2017), https://doi.org/10.1007/s11042-016-4315-0.

[38] R. Plutchik, *The psychology and biology of emotion*, (HarperCollins College Publishers, 1994).

[39] C. Dimoulas and G. Kalliris, "Investigation of wavelet approaches for joint temporal, spectral and cepstral features in audio semantics," presented at the *134th Convention of the Audio Engineering Society* (2013 May), convention paper 8858.

[40] N. Tsipas, L.Vrysis, C.A. Dimoulas and G. Papanikolaou, "Content-Based Music Structure Analysis Using Vector Quantization," presented at the *138th Convention of the Audio Engineering Society* (2015, May), convention paper 9269.

[41] M.Barthert, S. Essid, T. Fillon, J. Prado and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software", *ISMIR 2010*, pp. 441-446 (2010).

[42] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models." *Proc. CMMR* (2012): 492-507.

[43] M. Barthet, G. Fazekas, and M. Sandler, "Music emotion recognition: From content-to context-based models," In *International Symposium on Computer Music Modeling and Retrieval*, pp. 228-252. Springer, Berlin, Heidelberg, 2012, https://doi.org/10.1007/978-3-642-41248-6_13.

[44] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," presented at the *31st International Conference on Machine Learning (ICML-14)*, p. 1764-1772 (2014).

[45] Y.Kim, H. Lee and E.M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition", presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3687-3691 (2013), https://doi.org/10.1109/ICASSP.2013.6638346.

[46] K.Han, D. Yu and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," presented at the *Fifteenth Annual Conference of the International Speech Communication Association* (2014).

[47] Choi, K., Fazekas, G., Sandler, M. and Cho, K., "Convolutional recurrent neural networks for music classification," In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (pp. 2392-2396), 2017. https://doi.org/10.1109/ICASSP.2017.7952585.

[48] P. Jackson & S. Haq, *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*, University of Surrey (2014).

[49] O. Lartillot and P. Toiviainen, “Mir in matlab (ii): A toolbox for musical feature extraction from audio,” *Proceedings of the 8th International Conference on Music Information Retrieval* (Vienna, Austria, 2007 Sep.).

[50] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol.10, no.5, pp. 293-302 (2002 Jul.). https://doi.org/10.1109/TSA.2002.800560.

[51] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, “Multi-feature modeling of pulse clarity: Design, validation, and optimization,” presented at the *International Conference on Music Information Retrieval*, (Philadelphia, 2008).

[52] T. Eerola, O. Lartillot and P. Toiviainen, “Prediction of Multidimensional Emotional Ratings in Music From Audio Using Multivariate Regression Models,” presented at the *International Conference on Music Information Retrieval*, (Kobe, 2009).

[53] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, “The WEKA Data Mining Software: An Update,” *ACM SIGKDD Explorations Newsletter 11*, vol.1, pp.10-18 (June 2009) https://doi.org/10.1145/1656274.1656278.

## THE AUTHORS

**Nikolaos Vryzas** was born in Thessaloniki in 1990. He studied Electrical & Computer Engineering in the Aristotle University of Thessaloniki. After graduating, he received his master degrees on Information and Communication Audio Video Technologies for Education & Production from the Interdepartmental/Interdisciplinary Postgraduate Program on Advanced Computer and Communication Systems of the Aristotle University. His master thesis was on spatio-temporal audio recognition and segmentation. He is currently a PhD candidate in School of Journalism & Mass Media Communication, dealing with topics concerning semantically enhanced/ intelligent interaction tools for Journalism applications. Since August 2017 his research is supported by the Hellenic Foundation for Research and Innovation (HFRI).

**Dr. Rigas Kotsakis** received his diploma in Electrical and Computer Engineering at the Polytechnic School of Aristotle University of Thessaloniki. He continued his academic studies in MSc in Management and Business Administration at the University of Macedonia, MSc in Advanced Computer and Communication Systems at the Polytechnic School of Aristotle University of Thessaloniki and PhD degree at the School of Journalism and Mass Communications. He currently is a Tenured Senior Teaching Fellow in the School of Journalism and Mass Communications. His research interests include pattern recognition and semantic analysis techniques in multimedia content, management of audiovisual content and web-based multimedia applications.

**Aikaterini Liatsou** was born in Athens in 1991. She is a senior student in the School of Drama of the Faculty of Fine Arts of the Aristotle University of Thessaloniki, with expertise in dramatology and performance theory. She is a charter member and main dramatologist of the theatrical group "Ouk Nouk". She has participated in many theatrical production as stage manager, director assistant and dramatologist (National Theater of Northern Greece, Black Box, Municipal and Regional Theater of Kavala etc.).

**Charalampos Dimoulas** was born in Munich, Germany on August 14, 1974. He received his diploma and PhD from the School of Electrical and Computer Engineering, Faculty of Engineering, Aristotle University of Thessaloniki (AUTh) in 1997 and 2006, respectively. In 2008, he received scholarship on post-doctoral research at the Laboratory of Electronic Media of the School of Journalism and Mass Communications of AUTh. Both his doctoral dissertation and his post-doc research deal with advanced audio-visual processing and content management techniques for intelligent analysis of prolonged multi-channel recordings. He was elected Lecturer (November 2009) and Assistant Professor (June 2014) of Electronic Media in the School of Journalism and Mass Communications, AUTh, where he is currently serving. His current scientific interests include media technologies, signal processing, machine learning, multimodal systems, multimedia semantics, audiovisual content description and management automation. Dr. Dimoulas is member of IEEE, EURASIP and AES.

Dr. George Kalliris, http://kalliris.blogspot.com/, is professor of audio and audiovisual media technologies at the School of Journalism and Mass Communication of the Aristotle University of Thessaloniki (AUTh) Greece. He holds a 5-year MSc equivalent degree from the School of Electrical Engineering – specialty in Telecommunications, Faculty of Engineering, AUTh, with a scholarship from the Cyprus Government for all years of study 1984 – 1989, a doctorate degree from the same school, Lab of Electroacoustics & TV Systems 1990 – 1995. During and after completing his doctoral studies he worked in several research, development and innovation projects as well as a part-time higher education teacher. In 1998 he was elected Lecturer of Electronic Media Technology at AUTh. In his current position he

is the Deputy Head of the School, Former Master Program Director and the Head of Electronic Media Lab. He has also taught and/or teaching in four master's degree programs, to the Film Studies School and as a visiting professor of Frederick University Cyprus. His current research interests and publications include audiovisual technologies for the new media, radio and television studio design, digital audio-video processing–production–broadcasting-webcasting, multimedia content, restoration, management and retrieval. He is a member of the Audio Engineering Society (AES), member of the Association for Computing Machinery (ACM), board member of the Steering Committee of AUDIOMOSTLY yearly international conference and board member of the Hellenic Institute of Acoustics.