Speech Emotion Recognition Adapted to Multimodal Semantic Repositories[1]

Nikolaos Vryzas, Lazaros Vrysis, Rigas Kotsakis and Charalampos Dimoulas
Multidisciplinary Media & Mediated Communication Research Group (M3C)
Aristotle University of Thessaloniki
Thessaloniki, Greece
nvryzas@jour.auth.gr

***Abstract*—Speech emotion is an important paralinguistic element of speech communication, which undoubtedly involves high level of subjectivity, without concrete modeling of the implicated emotional states. Specifically, sentimental expression varies in great proportions among different spoken languages and persons. The current work is focused on the investigation of emotional states discrimination potentials, in an adaptive/ personalized approach, aiming at the creation of an effective multimodal speech emotion recognition service. In this context, an emotional speech ground truth database is formulated, containing semantically/ emotionally "loaded" utterances of a certain speaker in five basic sentiments. In the conducted experiments several classification algorithms are implemented and compared to the results of a generalized/ augmented multi-speaker emotional speech database. Furthermore, an audio-based application is designed for real time sentiment identification, while utilizing speech recording tools combined with camera and a Speech-to-Text modules. The audio, video and text files for every spoken utterance are labeled and stored via a user-friendly and functional GUI, for the subsequent augmentation of the personalized database.**

Keywords—speech emotion recognition; affective computing; crowdsourcing;

I. INTRODUCTION

Even before the wide spread of social media/ networking, the extraction of affective content information from broadcast audiovisual content for designing more personalized cognitive experiences, multimedia information retrieval systems and intelligent human machine interaction, has been a field of interest of scholars and researchers [1], [2]. Emotional expression is an important communication signal, containing paralinguistic information of speech, while simultaneously transmitting the carrier's sentimental state/ personality [3]. This is the reason that explains the exaggeration of emotional expression in human communication channels [3].

In social networking services and applications, the sentiment is represented/ encapsulated in various forms/ modalities (emoticons in text, emotional speech, etc.) [4], [5]. Consequently, the emotion recognition process is expected to play an important role as a semantic ontology, since analysis of affective information in social media and the Semantic Web may help into the improvement of communication possibilities and knowledge representation [4], [5].

In the scientific field of Speech Emotion Recognition (SER), much research work has been done in the past years. Specifically, several features have been proposed and evaluated, after being extracted from various ground truth speech corpora, while utilizing different machine learning classification schemes for emotional states discrimination [6], [7]. Recent deep learning approaches aim at the exploitation of automatically extracted representations of speech signals, skipping the feature generating procedure [8]. Furthermore, besides audio, additional modalities and the respective supported mechanisms can be combined towards the retrieval of affective speech information from audiovisual content [9].

In [10] a SER framework for theatrical automatic lighting augmentations is described, where different emotional classes are mapped/ linked to lighting colors. In [11], a novel data repository, named as Acted Emotional Speech Dynamic Database (AESDD), is formulated and proposed, that contains recorded utterances of professional actors under the supervision/ guidance of a theatrologist, addressed to 5 emotions; anger, disgust, fear, happiness, sadness. The dynamic property of AESDD is grounded on the subsequent addition/ inclusion of more utterances through theatrical crowdsourcing, featuring this way a continuous augmentation/ evolvement. In addition, the adoption of a collaborative model, where performance professionals and audio engineers/computer scientists can cooperate, will thereafter lead into generic semantic classification experiments based on the augmented versions of AESDD.

[1] cite as: N. Vryzas, L. Vrysis, R. Kotsakis and C. Dimoulas, "Speech Emotion Recognition Adapted to Multimodal Semantic Repositories," in *Proceedings of the 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP 18)*, Zaragoza, 2018.

II. A Framework Supporting Personalized SER In Broadcast And Social Media Communication

A. Aims and Overview

The development of the presented application is motivated by the necessity of an easy-to-use environment for model training in an adaptive/ personalized mode of SER, to be utilized in broadcasting and affective social communication.

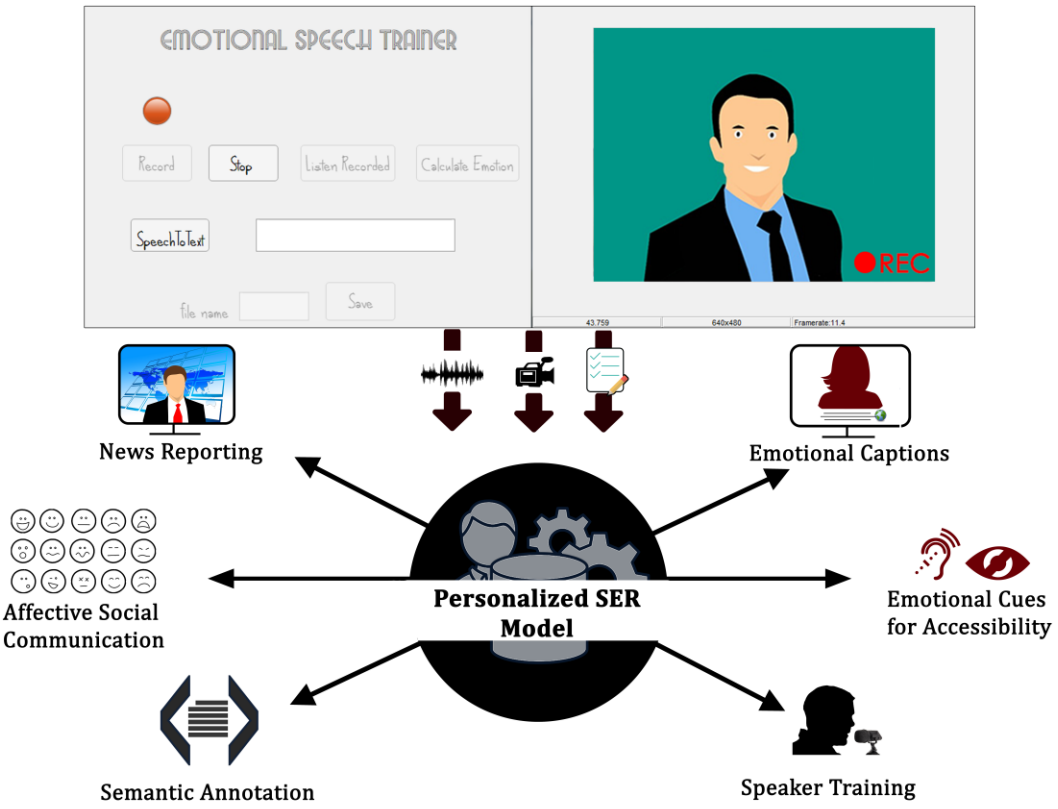

Fig. 1 The GUI of the application for personalized SER and multimodal database creation along with the proposed framework for integrating personalized SER in mediated communication, broadcasting and social media communication

While there is vast previous work in automatic SER, we considered that it is very suitable for the new media needs to create adaptive models, designed for single-person analysis. Despite the fact that there is a degree of universality in the verbal expression of emotions that favors the use of more generalized schemes, it has also to be considered that every person has a distinctive expression of emotional speech. This empirical observation triggered intuitively the research hypothesis that personalized SER would be way more efficient for adaptive real-world scenarios (streaming channels, web broadcasters etc.). Moreover, there is a fruitful field of applications of personalized semantic modeling in modern communications and web services.

News reporting and television talk shows are probably the most traditional channels, in which a single person is the main transmitter of information. In this common scenario, the show is focused on a main headliner, who can be intelligently modelled. However, television stations no longer hold the monopoly of broadcasting, since a plethora of users maintain broadcasting channels in social media for information, entertainment etc. purposes.

Affective computing is evolving into an important aspect of Human Computer Interaction and social communication. Semantic annotation concerning the paralinguistic information of broadcast and shared comment can enhance content management and description. While language can play crucial role in expressing and perceiving emotion, sentimental cues can augment end-to-end translation, where Automatic Speech Recognition cannot transmit/ convey all the information. The emotion extraction process from speech can be supported with multiple modalities; audio, visual and textual context. Automatic SER can promote accessibility and give people, with vision or hearing disabilities, estimations of expressed emotions for every modality. Finally, successful broadcasting usually requires training through feedback mechanisms, elements that can be offered by the presented application that can serve as an educational environment for carrying out emotions verbally.

Meanwhile, the annotated databases of audio, video, and text content that can be generated from speech can be further utilized for the augmentation of the aforementioned AESDD database [11]. In this way, along with the formulated personalized databases, a

vast generalized multimodal repository of emotional speech utterances will be developed, aiming at addressing and further investigating the research field of multimodal SER in future work.

B. Functionalies of the Application

The application, as far as it is developed, supports the three aforementioned modalities in terms of data gathering/ storing, but in the grounds of SER processing is (for now) audio-driven. This choice is not only justified on the importance of speech in human communication, but also on the research/ experiments that have been conducted in previous work [10], [11] .

The main functionalities/ operations/ capabilities of the application can be described as following:

- The user can record an emotionally charged utterance
- The user can playback the recorded sample
- The user is able to watch a video preview of the selected webcam capturing while recording
- The user can use an editable field of a third-party speech-to-text transformation of the recorded utterance. If the transformation is not accurate, the user is able to correct it.
- A post processing module of audio signal performs automatic editing of the recorded content
- A feature extraction generator computes the selected feature vector from the processed audio signal
- The application can determine/ predict the emotion of the spoken utterance via a pre-trained classifier based only on audio information.
- The user can save the utterance in three separate file representations/ forms; a wav file containing audio data, a video file with the visual/facial information and a text file with the linguistic content.

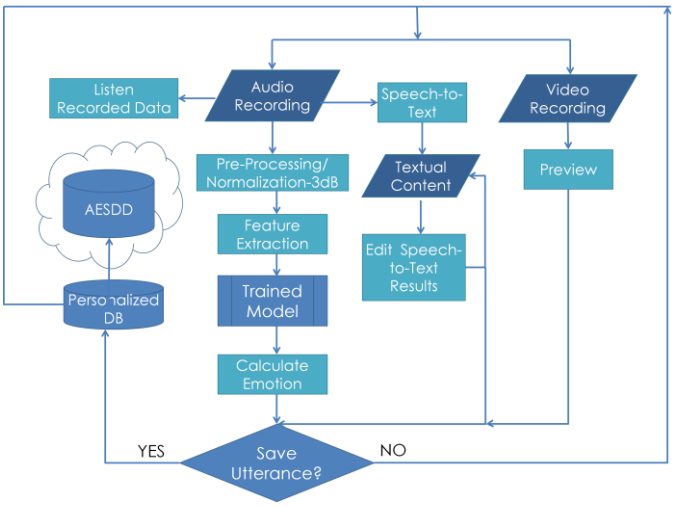

Fig. 2 The flowchart of the application functionalities

While the user is logged onto the environment, in order to produce/ gather multimodal data, in the same time the personalized database continually grows, in terms of the collected emotional speech. In this context, new models can be trained on the augmented future versions of the dataset, and simultaneously the new classifiers can be imported/ integrated in the SER module of the application, without affecting the rest of the functionalities. In addition, the extraction of the complementary visual and textual information serves the development of big multimodal datasets for future robust multimodal decision-making schemes. In its contemporary form, the prototype application was developed in the Matlab 2017a App Designer environment and it is "programmingly packaged" into a single standalone service, in order to be thereafter installed as a Matlab add-on app in end users.

III. Database Creation and Feature Extraction

In order to elaborate experimentally on the intuitive assumption that personalized SER will give more accurate results, a novel database was created. This dataset consisted of 20 utterances in Greek, for each of the five emotions; Happiness, Sadness, Anger, Fear and Disgust, resulting in a total number of 100 audio samples. All the recordings derived from the same non-professional

actor, a member of the research team, so that the conducted experiments to simulate the scenario of emotional analysis of a single, non-actor user. It has to be noted that the choice of sentiments along with the recorded utterances, were in accordance with previous experiments that took place while utilizing the generalized database AESDD (with multiple actors). The choice of a member of the team as a speaker for the formulation of the database serves the purposes of testing and presenting the application for evaluation. Empirical observations show similar results for other speakers as well. Concerning the audio technical characteristics, all the recordings were coded in PCM wav format, with a sampling frequency of 44100Hz and 16-bit dynamic range.

From every utterance, several audio properties from multiple domains (time, spectral, cepstral) were extracted, while utilizing the Matlab MIR toolbox [12]. The selected feature vector was formulated based on previous experimentation and evaluation for SER [6], [7], [10], [11], and more generally broadcast audio content management research works [13], [14]. Specifically, the selected audio properties are categorized and presented in Table I.

TABLE I. FEATURE VECTOR

Domain	***Features***
Time	Npeaks, RMS, Low Energy, Event Density, Tempo, Pulse Clarity, Zero Cross Rate
Frequency	Rolloff(FR), FR= 30, 50, 70, 90 Bright(B), B= 0.5, 1, 1.5, 2, 3, 4, 8 kHz
Spectral Statistics	Centroid, Spread, Skewness, Kurtosis, Flatness, Entropy
Cepstral	MFCC(i), i=1-13

IV. EXPERIMENTATION AND RESULTS

The conducted machine learning experiments were dependent on the classification scheme of the five emotional states of the labeled instances, which represent the five output classes of the speech emotion recognition problem. Several classifiers were tested while using the WEKA environment, as well as the Matlab's Classification Learner Toolbox [15].

The classification models were trained and validated by the input feature values. The tested algorithms included Multi-Class Logistic Regression (Log.Reg.), Multi-Layer Perceptrons (MLP), Support Vector Machines (SVM) with polynomial kernel, Logistic Model Trees (LMT) and a Subspace Ensemble Learning classifier (Ensemble) with 30 Discriminant Learners. The 10-fold cross-validation technique was employed for training the algorithmic models, as well as to compute the respective performances by the extracted confusion matrices (in terms of the ratios of the correctly classified instances to the total number of input samples or the respective class size). The classification accuracy results (**P**) are depicted in Fig.3 for the five emotion classes (**a**nger, **d**isgust, **f**ear, **h**appiness, **s**adness) along with the overall performance of the implemented training algorithm.

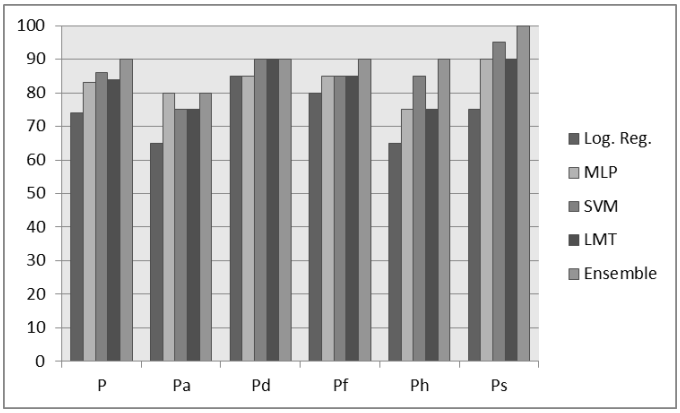

Fig. 3 Classification results for the personalized database for different classifiers and classes

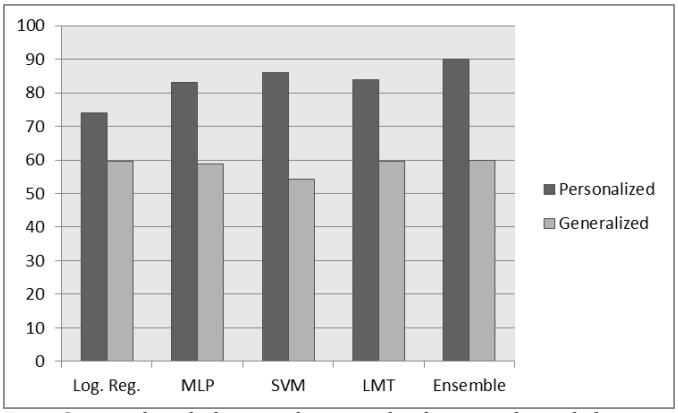

Fig. 4 Compared results between the personalized emotional speech dataset and the augmented AESDD generalized dataset

While observing Figure 3, the Ensemble Learning classifier scored the highest performance rates for the given dataset. Thereafter, the personalized database that was created for the current work was integrated into the AESDD database, serving in the same time the augmentation process. As mentioned before [11], the AESDD contains around 500 utterances from 5 different actors, consequently the addition resulted in the a total number of 600 recorded phrases of emotional speech. The same experimental procedure of classification and evaluation was followed for the augmented generalized dataset, and the classification accuracy results are presented in Fig.4, in contradiction to the personalized dataset results.

V. Conclusions And Future Work

The experimental results and subjective evaluation of the application inside the team proved quite promising. Personalized SER resulted to much more robust results than previous generalized models, so a new whole field for future work has opened. Some of the next steps of the research project will include:

- Subjective evaluation of the developed application by a selected audience of media information and broadcasting practitioners, students and researchers.
- Creation of an application with only open-source dependencies, favoring more effective broad distribution
- Integration of multimodal decision models with the combination of the three modalities (audio/video/text),that are supported by the application.
- Generalization of the experiments for multiple personalized databases.

Acknowledgment

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] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV." IEEE Signal Processing Magazine, vol 23.2, 2006, pp. 90-100.

[2] B. Schuller, M. Wöllmer, F. Eyben, and G. Rigoll, "Retrieval of paralinguistic information in broadcasts. Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance, and Authoring", pp. 273-287, 2012.

[3] J.L. Tracy, D. Randles, and C.M. Steckler, "The nonverbal communication of emotions," Current opinion in behavioral sciences, vol. 3, pp. 25-30, 2015.

[4] M. Grassi, E. Cambria, A. Hussain, and F. Piazza, "Sentic web: A new paradigm for managing social media affective information," Cognitive Computation, vol. 3.3, pp. 480-489, 2011.

[5] P. Weerasinghe, A. Marasinghe, R. Ranaweera, S. Amarakeerthi,and M. Cohen, "Emotion Expression for Affective Social Communication," International Journal of Affective Engineering, vol.13.4, pp. 261-268, 2014.

[6] C.N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artificial Intelligence Review, vol.43.2, pp. 155-177, 2015.

[7] C. Busso, M. Bulut, S. Narayanan, J. Gratch, and S. Marsella, "Toward effective automatic recognition systems of emotion in speech. Social emotions in nature and artifact: emotions in human and human-computer interaction," J. Gratch and S. Marsella, Eds, pp. 110-127,2013.

[8] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200-5204, March 2016.

[9] S.E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, V., Konda, et al., "Emonets: Multimodal deep learning approaches for emotion recognition in video," Journal on Multimodal User Interfaces, vol. 10.2, pp. 99-111, 2016.

[10] 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.

[11] N. Vryzas, A. Liatsou, R. Kotsakis, C. Dimoulas, and G. Kalliris, "Speech Emotion Recognition for Performance Interaction", Journal of the Audio Engineering Society, vol.66.6, pp. 457-467, 2018.

[12] O. Lartillot and P. Toiviainen, "Mir in matlab (ii): A toolbox for musical feature extraction from audio," in Proceedings of the 8th International Conference on Music Information Retrieval, September 2007.

[13] R. Kotsakis, G. Kalliris and C. Dimoulas, "Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification," Speech Communication, vol.54.6, pp.743-762, July 2012.

[14] R. Kotsakis, G. Kalliris and C. Dimoulas, "Investigation of salient audio-features for pattern-based semantic content analysis of radio productions," in 132nd Convention of the Acoustic Engineering Society, April 2012.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, "The WEKA Data Mining Software: An Update," in ACM SIGKDD Explorations Newsletter 11, vol.1, pp.10-18, June 2009.