Intermediary Fuzzification in Speech Emotion Recognition

1st Gustavo Assunção
Institute of Systems and Robotics
University of Coimbra
Coimbra, Portugal
gustavo.assuncao@isr.uc.pt

2nd Paulo Menezes
Institute of Systems and Robotics
University of Coimbra
Coimbra, Portugal
PauloMenezes@isr.uc.pt

***Abstract*—Affective systems are getting increasingly more attention from researchers and high-tech companies in order to enable the acknowledgment or adaptation to a user's mood. Emotion classification is typically a hard problem due to the number of subtle cues which are present in human facial and body expressions, or in voiced utterances. Another critical factor is that typically used models tend to map emotions into all-or-nothing regions with artificially sharp divisions among them, a view which is rather unsupported in the field of psychology and human behavioral analysis. In this paper we propose the inclusion of an intermediary fuzzy layer in a VGGVox-based NN, whose aim is to deal with the inherently foggy transitions between emotional states. This neuro-fuzzy model was trained and evaluated against four emotional speech databases and has shown improvements in the classification performance over a non-fuzzy counterpart. Observed performances were also on-par or above those of other current state-of-the-art techniques.**

***Index Terms*—Fuzzy Neural Networks, Fuzzy Clustering, speech emotion recognition, VGG**

I. Introduction

Emotion is a largely impactful aspect of human life, influencing many if not all interactions and decision making processes. As according to Cowie [1] this may be due to our evolutionary track, which has led emotion to develop as an means of behavioral adaptation to distinct scenarios. Hence, researching emotional development and its potential emulation in machines may be a worthwhile endeavor in order to enhance their response fluidity and adaptability. Quite a large amount of systems already part of everyday use could benefit from this, such as the home assistant *Google Home*.

A clear path to endowing machines with emotional understanding first requires some degree of human awareness and a necessary recognition of the emotional states displayed by the machine's users. A job that can be achieved by analyzing information gathered from different modalities, being vision and audio the currently most prominent ones. Evidently, emotion has yet to be fully understood and many distinct designs have been proposed which attempt to model emotionality, such as Plutchik's emotional wheel [2] or the more widely used pleasure-arousal-dominance (PAD) model [3]. A common version of the former is shown in Figure I, where the entire emotional space suffered discretization to only 6 emotional states plus neutrality. Despite the glaring differences among these conventional models, they all suggest the existence of inter-dependency between archetypal emotional states, something that is supported by different studies such as [4]. Interestingly enough, emotion recognition machine learning approaches tend to completely disregard these interrelations when classifying an instance of data. This in turn is maybe due to supervised learning datasets only including instances with single labels, instead of each instance being assigned a vector of membership degrees with respect to each emotional state.

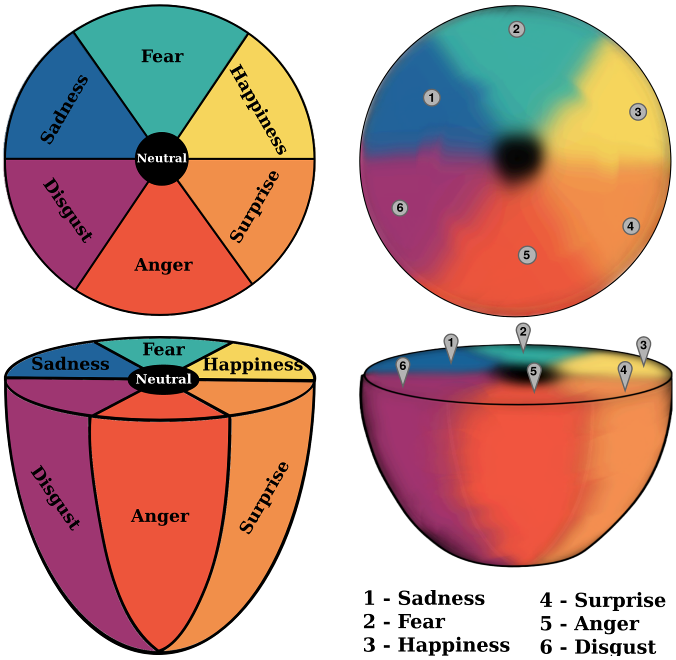

Fig. 1. Distribution of 6 emotional states plus neutrality, following the emotional wheel model. The color palette is arbitrary. Conventional discretization is shown on the left for comparison with the hypothesized fuzziness on the right.

In fuzzy logic [5], where an hypothesis truthiness is evaluated quantitatively or qualitatively rather than in a strictly binary fashion, the barriers dividing data are blurred and consid-

This work has been partially supported by OE - National funds of FCT/MCTES under project UIDB/00048/2020.

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

ering mutual relationships among distinct instances becomes increasingly achievable. More specifically, fuzzy classification allows for bordering classes to overlap with one another, meaning the same data object may comprise several distinct classes with corresponding degrees of membership. Given these facts, fuzzy approaches to emotion recognition seem nearly perfect candidates for the study of the mentioned dependencies among emotional states. These inter-dependencies are visually exemplified in Figure I, where each emotional state is considered a point in space represented by a combination of memberships to archetypal categories rather than the categories themselves. This way emotion is not limited and data may be classified based on how it correlates with each category separately and altogether. This can become even more useful when combined with state-of-the-art techniques such as neural networks, a synergy which has been somewhat explored for facial-centered emotion recognition [6]–[10], but not so much in the speech-centered counterpart. As such, and given the success of the facial-centered techniques, similar ones should be explored as well using the audio modality.

In this study we propose a novel approach to the speech emotion recognition task, which fuses fuzzy logic with machine learning by means of incorporating a fuzzy method in-between the layers of a neural network. With this we intended to explore how inter-connected the emotional states are with respect to their acoustic features extracted from human voice, and whether or not these relationships can be employed to accurately classify emotion. Applications of such a system are varied and mostly real-time, from interrogation purposes in forensic science [11] to integration in virtual interactive assistants [12], [13], endowing them with the ability to formulate appropriate responses to the emotional state of a user.

This paper is organized in the following manner. First, an overview of recent and related work is provided in Section II, in order for the reader to have a better understanding of the topic's state-of-the-art. Our proposed approach is described in Section III, succeeded by Section IV where the performed experiments are explained. The results obtained from these experiments are shown in Section V, as well as their corresponding discussion and critique. Lastly, a conclusion is provided in Section VI.

II. Related work

This overview is divided in the following two sub-entries. First, recent approaches to fuzzy logic-neural network combinations are analyzed so as to contextualize the proposed method. Subsequently, we go over some state-of-the-art speech emotion recognition techniques and what novelties allowed them to be successful.

A. Fuzzy Logic and Neural Networks

As previously mentioned, incorporating fuzziness into neural networks may be advantageous to several areas of research. This is largely due to fuzzy logic's ability to deal with instance relationships and reduce uncertainty in raw data. This type of approaches are commonly referred to as neuro-fuzzy systems. To give some examples of this kind of methods, Korshunova [14], [15] *et al.* integrated a self-organized fuzzy c-means clustering layer into a common CNN architecture, in order to exploit object similarities for improved classification. In [16], Sharma *et al.* employed a novel fuzzy pooling layer for dimensionality reduction to dominant features alone, effectively replacing conventional pooling layers in a CNN architecture so as to counter their inability to only preserve useful information. In [17] Greeshma *et al.* incorporated a fuzzy rule layer into a CNN so as to efficiently reduce contamination in reconstruction of images at a higher resolution, as a result of rule-driven feature map learning between inputs and targets. An interesting approach by Rajurkar *et al.* [18] presented an alternative to conventional ANNs where each neuron was designed as a standalone Takagi Sugeno fuzzy inference system [19]. This allowed for simultaneous modeling of diverse fuzzy structures, leading to increased robustness against ambiguity and vagueness.

Generally speaking, neuro-fuzzy methods outperform their non-fuzzy counterparts with considerable improvement. Such advancements should also be attempted for the speech emotion recognition topic, following the mentioned approaches.

B. Speech Emotion Recognition

As should be expected, it is necessary to consider a tremendously large set of features when attempting to accurately classify an utterance with an emotional state. Given this need, it is not surprising how deep neural networks (DNN) and other machine learning related approaches have been demonstrating better performances than classical techniques [20]. To provide some examples, Gideon *et al.* [21] used a progressive neural network (ProgNet) approach to emotion recognition in order to also take advantage of relevant non-emotional information (e.g. speaker gender). Lim *et al.* [22] combined convolutional neural nets (CNNs) with long-short term memory recurrent neural nets (LSTM-RNNs) to synthesize sequential dynamics in speech and explore temporal associations for emotion recognition. In a more related work [23], Zhang *et al.* diversified training and exploited the relations between possible emotional schemes by making hidden layers of a multi-task deep neural network available to all considered schemes. Another approach [24] was presented by Zhang *et al.*, where a neural network with bounded and weighted fuzzy membership functions (BSWFMs) on the hidden layer was used to map extracted speech features from emotional utterances to a valence-arousal 2D model. Overall it is quite perceptible how the current state-of-the-art in terms of speech emotion recognition is dominated by machine learning techniques and more specifically by artificial neural networks, whose performance far surpasses that of conventional audio analysis methods.

As can be perceived, neuro-fuzzy approaches to speech emotion recognition are still somewhat rudimentary when considering the state-of-the-art of neuro-fuzzy systems. Furthermore, much of what has been proposed for other areas of research could also be applied in classification of emotional speech. Hence, we propose the following approach which is outlined in the next section.

III. Methodology

This sections provides a detailed overview of the proposed method. For that, each component is gone over individually in order to allow for a better understanding of the pipeline as a whole. As stated, and following the studies presented in [15], the main idea of this work was to take advantage of fuzzification to assess the relationships between different emotional states and whether or not these would be suitable for their recognition. Following the state-of-the-art trends, we employed machine learning in our approach allied with spectral representations of raw audio, rather than extracted artificial features.

A. Spectral Representations

In order to analyze audio, some representation of it must first be obtained usually in form of artificial features extracted through various manipulation and sampling techniques, such as the well known Mel-frequency cepstral coefficients [38]. Though these have been extensively employed before in the evaluation of emotional audio, they require a considerable amount of preprocessing leading to additional overhead and manual tuning which negates their alleged success. As such, our method focused on using audio in its raw spectral form, composed of energy variations at different frequencies over time. These images termed spectrograms are quite suitable for audio analysis as they retain all information regarding general features such as tone or pitch, within which emotional information is believed to be embedded. Following this premise, a sliding Hamming window of width 25ms was applied to each considered audio clip, with a step of 10ms, in order to generate the corresponding narrowband spectrogram. In addition, the respective means and variances were also normalized at every frequency bin of the spectrum. No additional procedures besides the former were carried out on the raw data.

B. Convolutional NNs & VGGVox

Artificial neural networks or ANNs have long been a focus of research as they are modelled after the biological connections which make up the brain. These machine learning algorithms are essentially layered pipelines of operations which perform operations over some input data in order to obtain a desired output. A particular type of ANNs are convolutional neural networks (CNNs) [25], which are characterized by having convolutional layers which strive to emulate the capabilities of the visual cortex, the brain sector responsible for processing visual information taken in by the eye retina. This is achieved through series of filters whose weights accentuate different picture areas deemed important, allowing for a better response to even faint stimuli and making them ideal for image processing. In light of how spectrograms are visual representations of audio, they may be analyzed as images using CNNs.

The VGGVox model first presented in [26] is a VGG-M CNN architecture [27] specifically designed for the analysis of audio in its spectral form, with the end goal of closed set speaker recognition. Its architecture has been trained extensively with over 2000 hours of audio, and considering its high performance it is quite able of extracting highly robust audio features from spectrograms of any form of speech. This has been verified before specifically for classification of emotional speech in [28]. To this end, only its final classification layers must be replaced for emotion recognition whilst the rest of the already trained convolutional and pooling layers may be retained.

C. Fuzzy Layer

In order to study the associations between different emotional states and reduce the separators believed to exist between them in discrete classification, fuzzification of data becomes a necessity. Hence, a new layer was developed which employed the method of fuzzy c-means (FCM) [29], [30] for fuzzification of embeddings during propagation of data through the network. This algorithm was chosen not only given its success with object recognition in [15], which we intended to extrapolate for emotion recognition, but also due to its implementation simplicity which deems it ideal for baseline establishment and subsequent comparison with more complex techniques such as possibilistic-c-means or random forest.

The fuzzy layer uses the embeddings of training data it receives from the preceding layer to perform clustering in a way such that each data instance may belong to more than one cluster, thus becoming fuzzified. In addition, this clustering process does not equate to final classification, given how the number of clusters may be greater than that of classes. Considering how each embedding x_j is made up of L features $\mathbf{F} = (f_1, ..., f_L)$ and given a pre-set number of clusters K, the algorithm starts with $C = \{c_1, ..., c_K\}$ cluster centroids initialized according to a uniform distribution, which are then iteratively updated to better match the data. These are computed through a weighted average of all points:

$$c_k = \frac{\sum_{j=1}^{N} \omega_{jk}^{m} \cdot x_j}{\sum_{j=1}^{N} \omega_{jk}^{m}} \tag{1}$$

Where $m > 1$ is the fuzziness intensity of a cluster, which asserts how much it may overlap with others, and w_{jk} is the membership degree of embedding x_j in cluster k obtained through:

$$\omega_{jk} = \frac{1}{\sum_{i=1}^{K} (\frac{||x_j - c_k||}{||x_j - c_i||})^{\frac{2}{m-1}}} \tag{2}$$

Using these expressions, minimization of the weighted sum of all possible instance-centroid pair squared norms is performed until a maximum number of iterations is reached or centroid updates become negligible, to find optimized locations. Post-training, this fuzzy layer takes as input an

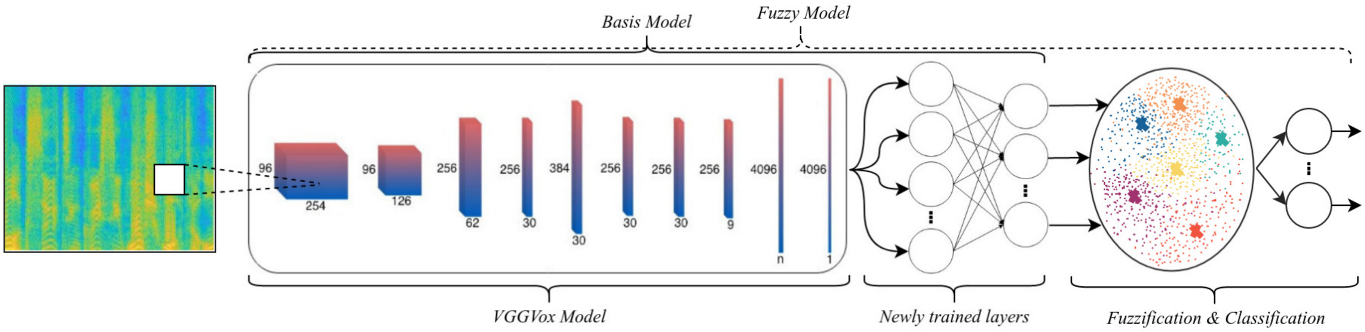

Fig. 2. Structure overview of the proposed model incorporating a fuzzy layer.

embedding of features **F** and provides a vector of membership degrees of ω respective to each cluster and its centroid.

D. System Overview

The integration of the fuzzy layer into the employed CNN architecture is quite straightforward. In order to take full advantage of *VGGVox*'s pre-trained weights, its intermediate layers are kept and two fully-connected layers are added for feature dimensionality reduction. Subsequently, the fuzzy layer receives these reduced size vectors and can successfully perform their fuzzification. The resulting membership degrees vectors are then progressed through another fully-connected layer for final classification. This pipeline is depicted in Figure 2 for easier understanding.

The approach we used differs from that presented in [15] for object classification given that we first execute dimensionality reduction of the embeddings obtained by VGGVox through pre-fuzzification fully-connected layers. The motivation for this comes from the work [31], which extensively demonstrated the inability of the fuzzy c-means algorithm to deal with data at dimensions greater than a few dozen. Hence it would be senseless to employ this fuzzy clustering algorithm right after the encoding layers of a CNN as these generally output embeddings composed of hundreds or even thousands of features.

IV. Experiments

In order to evaluate the performance of the proposed technique, two models were run in tandem with the same hyperparameters and inputs. One was the system described in the previous section, including the fuzzification layer for assessment of inter-emotion relationships and its suitability for speech emotion recognition. The other model, used as a comparison term to its former, maintains the exact same architecture with the exception of the last two layers (fuzzy and classification) which were removed. The two models were then trained separately but always with the same inputs.

A. Emotional Speech Data

As is typical with emotion recognition, real data acquisition is uncommon and considerably difficult to obtain due to ethic related issues and the spontaneity inherent to real emotional expression. Thus research and testing is often carried out using acted databases, where a set of emotional states is represented either visually or vocally by professional or volunteer actors. Notable datasets such as these include EMODB [32] in German, EMOVO [34] in Italian, SAVEE [33] in British English and ELRA-S0329 [40] in Spanish, which have been employed and exhaustively tested on before for validation of the current state-of-the-art techniques. As such, these were chosen for application of our fuzzy model not only to assert its validity but also to compare its performance with other methods from recent research.

TABLE I
Fuzzy layer performance variation with number of clusters on EMODB, using 30 epochs and 10-fold corss-validation.

#Clusters	10	25	50
Acc (Std)	63.75% (4.00%)	64.12% (5.32%)	65.42% (8.25%)
#Clusters	100	150	200
Acc (Std)	67.10% (4.82%)	68.93% (5.35%)	63.33% (6.68%)

B. Model Evaluation

In the case of the non-fuzzy model, the training and testing phases followed typical procedure. The model was built using the Keras framework [39], the *VGGVox* weights loaded and the added fully-connected layers trained using the mentioned datasets, one at a time. As for the fuzzy version of the model, training following this same scheme until the Fuzzy Layer. Here, the described fuzzy c-means clustering algorithm was applied to the accumulated embeddings of the training data and the resulting membership vectors used to train the final fully-connected layer for classification. The number of clusters was set to 150, a considerably greater than the number of classes, as we hypothesized extracting membership degrees from a large set of intermediate emotional states, rather than only the 6 archetypal ones, could help with classification. To this end, the number of clusters was steadily raised until a performance drop was observed, as shown in Table I. Both models were trained with 30, 50 and 100 epochs to evaluate the fuzzification effect and all training and testing was performed using 10-fold cross-validation in order to more accurately assess the

TABLE II
Comparison of accuracy results between the non-fuzzy and fuzzy models trained with 30, 50 or 100 epochs and the same hyperparameters, using 10-fold cross-validation.
Mean Accuracy (Standard Deviation)

Epochs	**Non-Fuzzy**				**Fuzzy**			
	EMODB	*SAVEE*	*EMOVO*	*S0329*	*EMODB*	*SAVEE*	*EMOVO*	*S0329*
30	67.05% (5.61%)	58.54% (8.08%)	51.02% (8.02%)	88.38% (3.47%)	68.93% (5.35%)	61.25% (5.88%)	50.85% (7.19%)	90.61% (3.03%)
50	71.94% (5.30%)	63.54% (4.61%)	55.42% (4.55%)	88.97% (2.97%)	74.17% (5.71%)	66.04% (6.27%)	55.93% (3.26%)	90.31% (2.11%)
100	76.62% (8.41%)	68.54% (2.98%)	62.20% (7.42%)	90.91% (3.19%)	78.48% (7.86%)	71.04% (4.76%)	64.07% (7.72%)	91.51% (2.68%)

generalization of our technique. Results are presented in Table II for each database considered and for both models.

As can be observed from Table II, results were quite successful with an increased number of epochs even though accuracies varied from database to database.

V. Discussion

When analyzing the obtained results, the performance raise from the basic to the fuzzy-layer model is quite clear as it happens in nearly every case. Even despite the performance variation between different datasets, which may stem from the differences in amount of clips, recording specifications and/or cultural background, the fuzzy upgraded VGGVox model always performed better than its non-fuzzy counterpart when the number of epochs was increased. This serves to show how archetypal emotional states are not absolute, in fact sharing common traits and frequently being correlated with one another to form more of an emotional spectrum rather than a discrete set of categories. Furthermore, it can be concluded that these emotion inter-relationships are present in sound and are suitable for emotional state classification or at the very least increased robustness, considering the success of the performed experiments.

TABLE III
Performance comparison with current state-of-the-art techniques.

	Kerkeni [35]	Latif [36]	Sidorov [37]	**Proposed Method**
EMODB	69.6%	72.4%	74.6%	**78.5%**
SAVEE	-	56.8%	63.8%	**71.0%**
EMOVO	-	76.2%	-	64.1%
S0329	90.1%	-	-	**91.5%**

In addition to the fuzzy model results having surpassed those of the non-fuzzy model, the former's performance was also on par with that of current state-of-the-art techniques. This can be observed in Table III, where our proposed method either reached higher accuracies than other recent models for the same databases, or nearly matched their results. Moreover, the model takes only a fraction of a second to evaluate a new data instance, making it highly suitable for real-time applications.

VI. Conclusion

In this paper, we proposed a method which integrated fuzzification in a neural network in order to perform speech emotion recognition. The method was validated against four standard and widely employed emotional speech databases. From the obtained results it was empirically determined that archetypal emotional states are indeed correlated and showed how these correlations can be successfully employed in the classification and robustness increase of speech emotion recognition machine learning methods. This backs our initial hypothesis of a fuzzy rather than discrete emotional distribution model and supports the notion that fuzzy emotion recognition should be given more attention than its absolute counterpart. In addition, our technique showed great performance when compared with other state-of-the-art methods, either matching or surpassing results of previous research and showing this mentioned importance of fuzzification for speech emotion recognition.

In the future, we intend to apply other more robust methods of fuzzification and fuzzy clustering as layers of NNs. This is intended to resolve the curse of dimensionality issue recurrent to fuzzy c-means and allow for the integration of fuzzy layers between its encoding counterparts instead of just on the classification section of the architecture. Additionally, we will evaluate the possibility of pooling together fuzzy representations from different modalities (e.g. audio and video) as a fuzzy multi-modal emotion recognition approach.

Acknowledgements

The authors thank the respective curators for providing access to their emotional speech databases and allowing for their use in our work. Appreciation is also extended to designer Rita Nogueira for her help with the presented images.

References

[1] 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, 18(11):32–80, 2001.

[2] R. Plutchik. Emotion: A psychoevolutionary synthesis. New York: Harper and Row, 1980.

[3] A. Mehrabian. Basic dimensions for a general psychological theory, pages 39–53. Oelgeschlager, Gunn and Hain, 1980.

[4] R.J. Larsen and E. Diener, "Promises and problems with the circumplex model of emotion", Review of Personality and Social Psychology, vol. 13, pp. 25–59, 1992.

[5] L. A. Zadeh, “Fuzzy sets”, Inf. Control, vol. 8, no. 3, pp. 338-353, jun 1965.

[6] Y. Guo and H. Gao, ”Emotion Recognition System in Images Based On Fuzzy Neural Network and HMM,” 2006 5th IEEE International Conference on Cognitive Informatics, Beijing, 2006, pp. 73-78.

[7] W. Shi and M. Jiang, ”Fuzzy Wavelet Network with Feature Fusion and LM Algorithm for Facial Emotion Recognition,” 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), Chongqing, China, 2018, pp. 582-586.

[8] E. Lotfi, A. Khosravi and S. Nahavandi, ”Facial emotion recognition using emotional neural network and hybrid of fuzzy c-means and genetic algorithm,” 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, 2017, pp. 1-6.

[9] Dae-Jin Kim, Zeungnam Bien and Kwang-Hyun Park, ”Fuzzy neural networks (FNN)-based approach for personalized facial expression recognition with novel feature selection method,” The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ ’03., St Louis, MO, USA, 2003, pp. 908-913 vol.2.

[10] X. Pan, ”Research on the Emotion Recognition Based on the Fuzzy Neural Network in the Intelligence Education System”, 2011 Second International Conference on Digital Manufacturing and Automation, Zhangjiajie, Hunan, 2011, pp. 1030-1033.

[11] G. Assunção, F. Perdigão, and P. Menezes. ”Premature overspecialization in emotion recognition systems”. In Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics, Jun 2019.

[12] B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K. R. Choo and M. Jamshidi, ”Toward Artificial Emotional Intelligence for Cooperative Social Human–Machine Interaction”, in IEEE Transactions on Computational Social Systems, vol. 7, no. 1, pp. 234-246, Feb. 2020.

[13] Z. Liu et al., ”A multimodal emotional communication based humans-robots interaction system,” 2016 35th Chinese Control Conference (CCC), Chengdu, 2016, pp. 6363-6368.

[14] K. P. Korshunova, ”A Convolutional Fuzzy Neural Network for Image Classification,” 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, 2018, pp. 1-4.

[15] V. V. Borisov and K. P. Korshunova, ”Multiclass classification based on the convolutional fuzzy neural networks”, In Artificial Intelligence, pages 226–233, Cham, 2019. Springer International Publishing.

[16] T. Sharma, V. Singh, S. Sudhakaran and N. K. Verma, ”Fuzzy based Pooling in Convolutional Neural Network for Image Classification”, 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 2019, pp. 1-6.

[17] M. S. Greeshma and V. R. Bindu, ”Single image super resolution using fuzzy deep convolutional networks”, 2017 International Conference on Technological Advancements in Power and Energy (TAP Energy), Kollam, 2017, pp. 1-6.

[18] S. Rajurkar and N. K. Verma, ”Developing deep fuzzy network with Takagi Sugeno fuzzy inference system,” 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, 2017, pp. 1-6.

[19] Takagi, Tomohiro, and Michio Sugeno, “Fuzzy identification of systemsand its applications to modeling and control,”IEEE Trans. Syst., Man,Cyber., vol. SMC-15, no. 1, pp. 116-132, 1985.

[20] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition”, Neural Networks, vol.92, pp. 60–68, 8 2017.

[21] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis and E. M. Provost, ”Progressive Neural Networks for Transfer Learning in Emotion Recognition”, Interspeech, 1098-1102, 2017.

[22] W. Lim, D. Jang, and T. Lee, ”Speech emotion recognition using convolutional and recurrent neural networks”, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016.

[23] Y. Zhang, Y. Liu, F. Weninger and B. Schuller, ”Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 4990-4994.

[24] Z. Zhang and J. S. Lim, ”Emotion Recognition Algorithm Based on Neural Fuzzy Network and the Cloud Technology,” 2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA), Krakow, 2015, pp. 576-579.

[25] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”, Biological Cybernetics, vol. 36, no. 4, pp. 93–202, 1980.

[26] A. Nagrani, J. S. Chung and A. Zisserman, “VoxCeleb: A Large-ScaleSpeaker Identification Dataset”, INTERSPEECH, 2017.

[27] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets”. In Proceedings of the British Machine Vision Conference, 2014.

[28] G. Assunção, F. Perdigão, and P. Menezes. ”Premature overspecialization in emotion recognition systems”. In Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics, Jun 2019.

[29] J. C. Dunn (1973): ”A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters”, Journal of Cybernetics 3: 32-57.

[30] J. C. Bezdek (1981): ”Pattern Recognition with Fuzzy Objective Function Algoritms”, Plenum Press, New York.

[31] R. Winkler, F. Klawonn, R. Kruse, ”Fuzzy C-Means in High Dimensional Spaces”, International Journal of Fuzzy System Applications, 1, Jan. 2011, 1-16.

[32] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. ”A database of german emotional speech”. INTERSPEECH, 2005.

[33] S. Haq, P.J.B. Jackson, and J.D. Edge. ”Audio-Visual Feature Selection and Reduction for Emotion Classification”. In Proc. Int’l Conf. on Auditory-Visual Speech Processing, pages 185-190, 2008.

[34] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco. ”Emovo corpus: an italian emotional speech database”. LREC (2014).

[35] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub and C. Cleder. ”Automatic Speech Emotion Recognition Using Machine Learning”, Social Media and Machine Learning, 2019.

[36] S. Latif, R. Rana, S. Younis, J. Qadir, J. Epps. ”Transfer Learning for Improving Speech Emotion Classification Accuracy”, arXiv:1801.06353, 2018.

[37] M. Sidorov, S. Ultes, and A. Schmitt, ”Emotions are a personal thing: Towards speaker-adaptive emotion recognition”, ICASSP, pp. 4803–4807, 05 2014.

[38] P. Mermelstein, ”Distance measures for speech recognition, psychological and instrumental,” in Pattern Recognition and Artificial Intelligence, 1976, C. H. Chen Ed., pp374–388, Academic, New York.

[39] F. Chollet and others. Keras. https://keras.io, 2015.

[40] E. L. R. Association, “Emotional speech synthesis database elra-s0329”, https://catalogue.elra.info/en-us/repository/browse/ELRA-S0329, 2011.