DIMENSIONAL MODELS FOR CONTINUOUS-TO-DISCRETE AFFECT MAPPING IN SPEECH EMOTION RECOGNITION

Serban MIHALACHE[1], Dragos BURILEANU[2]

Speech Emotion Recognition (*SER*) *is an important research area, with two distinct approaches for modeling emotions: as discrete classes and as points within a continuous affect space. In this paper, we argue for the carefully considered fitting of dimensional models in order to allow for accurate mapping of emotion classes within the affect space, unifying the two approaches. To this end, we use machine learning algorithms* (*K-means clustering, Gaussian Mixture Models, and Support Vector Machines*), *fitted on the Interactive Emotional Dyadic Motion Capture database* (*IEMOCAP*), *which provides dual discrete and continuous annotation of emotional content. We also show how the reliability and generality of the results can be improved by initializing the dimensional model's class centroids using the Warriner-Kuperman-Brysbaert* (*WKB*) *corpus. The proposed approach can lead to an unweighted accuracy up to 74.3% ÷ 77.3%, which represents state-of-the-art results for the considered dataset.*

Keywords: dimensional models of affect, machine learning, speech emotion recognition

1. Introduction and Related Work

The science (and somewhat art) of Speech Emotion Recognition (SER) is a constantly growing research area, finding a foothold in a wide spectrum of applications, in fields such as human-machine interfaces, forensics, and medical science, to name a few [1], [2].

When designing a SER system, there are two ways to consider emotions: discrete classes and dimensional modeling. In the former case, each affective state is viewed as a distinct, standalone category (e.g., anger, fear, sadness, etc.) [3], leading to a classification problem. By contrast, the latter takes into consideration a number of continuous affective dimensions, giving rise to an abstract affect space (typically 2D) [4], [5], the position within being the target, leading to a regression problem. Most often, the affective dimensions used are *arousal* (a subjective evaluation of the level of the affective manifestation) and *valence* (a subjective evaluation of the positivity of the affective manifestation). It is worth mentioning that the terms *arousal* and *activation* are used interchangeably in SER

[1] Ph.D. student, Speech and Dialogue Research Laboratory (SpeeD), University POLITEHNICA of Bucharest, Romania, e-mail: serban.mihalache@upb.ro
[2] Professor, Speech and Dialogue Research Laboratory (SpeeD), University POLITEHNICA of Bucharest, Romania, e-mail: dragos.burileanu@upb.ro

literature, although *arousal* is defined as a measure of the physiological response in an affective state, whereas *activation* represents the subjective measure described previously [6].

Earlier research has been concerned separately with the two SER paradigms (discrete and continuous), i.e., trying to determine the prevalent emotion class or the position within the affect space. In this sense, promising results have been reported in literature using machine learning and deep learning models and techniques, including Support Vector Machines (SVMs) [7], Multilayer Perceptrons (MLPs) [8], [9], Recurrent Neural Networks (RNNs) with Long Short-term Memory (LSTM) cells [10]-[12], Convolutional Neural Networks (CNNs) [13], [14], hybrid models [15], or advanced Convolutional Recurrent Neural Networks (CRNNs) [16], [17], using either algorithmic or automatic ("true deep learning") feature extraction, with the trend favoring the latter type, especially for continuous emotion recognition [18]. In most cases, attention mechanisms are also employed, significantly boosting the systems' performance.

Using both the discrete and continuous paradigms in a form of joint learning has also been proposed and yielded good results [19]. However, to the best of our knowledge, this is the first time that multidomain strategies are proposed (unifying the discrete and continuous paradigms by directly tying them together, through mapping). Two examples of such envisioned strategies are illustrated in Fig. 1.

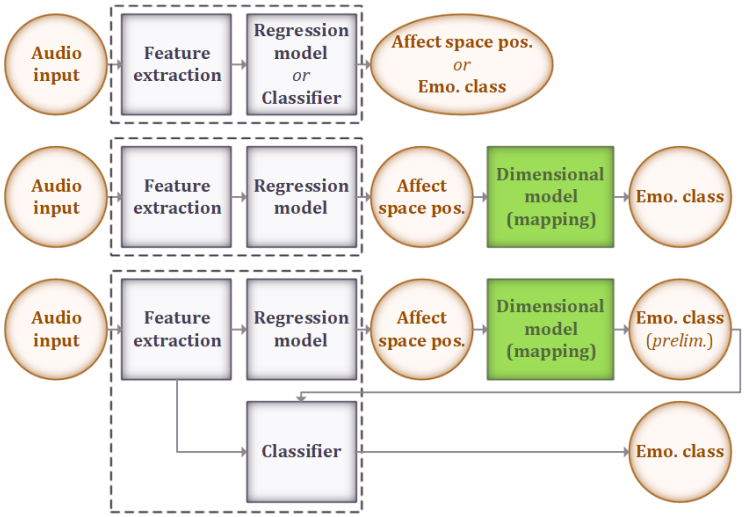

Fig. 1. Direct vs. multidomain SER. Top: standard systems. Middle: the dimensional model maps the estimated affect space position to the emotion class. Bottom: the dimensional model output is used as additional features for a classifier. Note: A single model may contain the enclosed blocks.

Instead of a standard system for only one task (top), the more nuanced nature of continuous emotion recognition serves to first determine the position within the affect space, and a pre-trained dimensional model then maps it to the corresponding emotion class (middle). Or, going further (bottom), the dimensional models' output could be used as an additional guiding feature for a second model, a classifier, trained together with the regression model through joint learning. In this paper, we present only the implementation and fitting of the dimensional model mapping, shown in green.

The main source of data for fitting the dimensional model would be a dataset with dual discrete and continuous annotation of emotional content (labels for emotion classes, and numerical values for the affective dimensions). The only such available corpus is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [20], having almost exclusively been used for emotion classification. In the rare case when the affective dimensions were taken into consideration, they were not used for true continuous emotion recognition, but were grouped into categories (low vs. medium vs. high values) [21].

Unfortunately, similar to the vast majority of available datasets for SER, the IEMOCAP database comprises simulated data, using actors that are more or less guided through specific scenarios with limited or no unpredictability [22]. Furthermore, only 6 evaluators annotated the data, which, considering the subjective and uncertain nature of the task, leads to lower confidence in the available data and suggests the fitted dimensional model would have poor generalization. Therefore, an additional source can be used, such as the Warriner-Kuperman-Brysbaert (WKB) corpus [23], which includes affective dimension annotations for a number of words, the relevant ones being those representing the emotion classes, such as "anger" (i.e., the "concept of anger"), etc. In our approach, these annotated values are used to initialize the class centroids (means). The reasoning is that, for the WKB corpus, there were hundreds of evaluators, greatly increasing confidence in its reliability and leading to better generalization.

In this paper, we used three machine learning algorithms – K-means clustering, Gaussian Mixture Models (GMMs), and SVMs – in order to create dimensional models for reliable continuous-to-discrete mapping from a 2D *arousal-valence* affect space to 4 emotion classes, as a key step towards a multidomain approach to SER. The main contributions are:

1) Proposing to combine the two emotion recognition paradigms (discrete and continuous) into multidomain strategies, with a dimensional model serving as the link.
2) Using machine learning algorithms and the only dually annotated database available in literature (IEMOCAP) for fitting the dimensional model.
3) Employing the WKB corpus for dimensional model initialization, improving confidence and generalization.

4) Demonstrating how the proposed approach can reach state-of-the-art SER classification performance for the IEMOCAP database.

2. Proposed Methodology

We tested three machine learning algorithms for developing the dimensional model for affect mapping. When the model was obtained using K-means clustering (we refer to this as the K-means model, KMM) or GMM fitting, then the class centroids (means) were initialized in one of two ways: i) using the values estimated from averaging over the IEMOCAP data (native initialization); ii) using the values estimated from the WKB data (WKB initialization). The first option allows for better data fitting, but, in the second case, model generalization is greater. SVMs lead to even better results thanks to their implicit higher dimensional transformation of the affect space but can only use IEMOCAP data.

The K-means clustering algorithm represents the simplest and fastest possible approach [24]. Its basic principle consists of grouping the data into a selected number of clusters so that each data point is assigned to the cluster whose centroid (mean) is closest (in the sense of minimum L2-norm) to the data point. This model implies linear decision boundaries between the clusters and can be seen as a particular case of a GMM with hard component assignment and with all mixture components sharing the same covariance matrix.

Defined as a linear combination of Gaussian distributions, GMMs can be interpreted as a generalization of K-means clustering to account for clusters with non-identical distribution and involving soft component assignment (i.e., the assignment function for data points and clusters is no longer binary, but rather represents the probability of the data points to be part of each cluster). By allowing each mixture component to have a separate and non-diagonal covariance matrix, the resulting clusters end up stretched and rotated in the affect space so as to better fit the data. Additionally, the soft component assignment and the shape of the probability density functions allow the fitting to lead to more accurate, non-linear decision boundaries between the clusters.

The third method, which improved the emotional mapping, was represented by SVMs, since they transform an original input space (the affect space) into a higher-dimensional one, where the data may be linearly separable. When the transformation is nonlinear, by means of a nonlinear kernel, such as a radial basis function (RBF) or a polynomial kernel, then the resulting decision boundaries in the original space is also nonlinear, leading to better separation. Since this is a multiclass problem, we adopted the one versus one (OvO) classification strategy, which implied training a different SVM for each pair of classes, resulting in $N\cdot(N-1)/2$ classifiers, where N represents the number of classes.

One extra approach was to not do any fitting at all, but simply use the WKB data to define the class centroids and directly perform classification based on the minimum distance (L2 norm) from a data point to the centroids (we refer to this as the WKB model). This would offer the best generalization due to the WKB corpus' advantages, but also implies using only annotations provided within an abstract conceptualization of emotions (strictly the "concept of anger", etc., without a visceral side to the data as there exists in speech recordings), whose nature would obviously be insufficient for SER tasks.

3. Experimental Setup and Results

3.1. The IEMOCAP database

The database [20] contains 5 sessions of audio-visual recordings, with 10 actors (5 female, 5 male) working in pairs to solve speaking tasks (scripted and improvised), totaling 10,039 recordings. The corresponding audio files are stored in uncompressed 16-bit PCM format, sampled at 16 kHz.

A total number of 10 discrete emotion classes (anger, fear, disgust, sadness, happiness, frustration, excitement, surprise, neutral and other) is available, with many of them strongly underrepresented, however, forcing us to use a smaller subset, grouped into 4 new classes (neutral; sadness; anger + frustration, and happiness + excitement, grouping the last two pairs together due to their closeness), similar to [9], [10], [15], [17]. For each sample, there are 3 evaluators. As expected, in the vast majority of cases, not all agreed on the annotation, given the uncertainty involved in the task [25], [26]. In order to obtain the final annotation (the ground truth), we propose 3 voting techniques: unanimity (requiring 3/3 consensus; resulting in 2,200 samples); majority (requiring 2/3 consensus; 7,577 samples); and ranked majority (starting with 2/3 consensus and doing a second pass over the remaining samples and labeling them according to the most trustworthy available evaluator per case, i.e., the evaluator who was most often in agreement with the initial ground truth; 9,641 samples).

For the continuous dimensions, we used *arousal* (*activation*) and *valence*. Each sample was rated by 2-3 evaluators, on a scale from 1 to 5 (low to high *activation*, negative to positive *valence*), which we then averaged and normalized to the $[-1, 1]$ range. We also noticed that the annotation files contained an error: the values were reversed (1 was high *activation* / positive *valence*, and 5 was low *activation* / negative *valence*, instead of the other way around for both dimensions), requiring an additional sign inversion.

Additionally, due to the unrealistically low granularity of this annotation data, and to allow for better model fitting and regularization, white Gaussian noise was added to the dataset [12], [27]. The resulting data distribution within the affect space is illustrated in Fig. 2. While the classes have clearly separated

centroids, a large degree of overlap between data points for each pair of classes can be observed.

3.2. The WKB corpus

The Warriner-Kuperman-Brysbaert [23] corpus consists of a large number of lemmas (words in canonical form), with the corresponding affective dimensions being annotated by 1,827 participants through Amazon's crowdsourcing service, with all words having at least 15 ratings and 87% of words having 24 ratings on average. These ratings were given on a scale from 1 to 9 (low to high *arousal*, negative to positive *valence*).

As discussed, we only use the words defining the 4 classes (neutral, sadness, anger + frustration, happiness + excitement) and, to reduce uncertainty, extended the list to those part of the corresponding word families (e.g., anger, angry; frustrated, frustrating, frustration; etc.), but not semantic fields (e.g., fury; annoyed; etc.). For each class group, we average the affective dimension values and normalize the resulting means to the same [−1, 1] range.

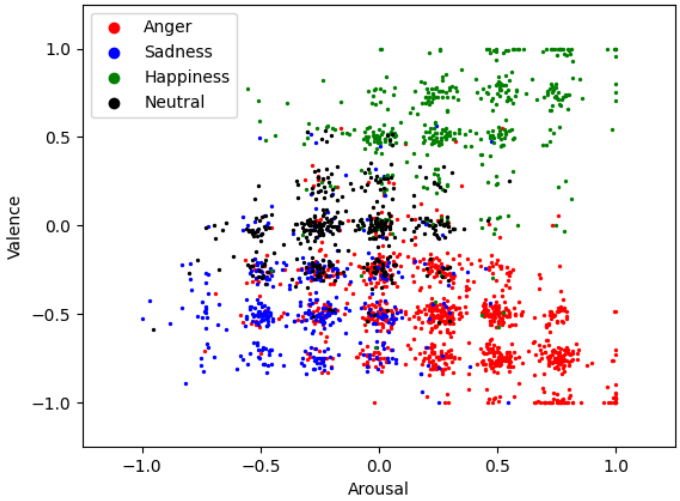

Fig. 2. IEMOCAP data distribution within the affect space, with added white Gaussian noise.

3.3. Results and discussion

We used the scikit-learn framework for Python to implement and fit the discussed models. For KMM and GMM mapping, we tested native initialization (based on IEMOCAP data) and WKB initialization (based on WKB data) for the class centroids (means). For the SVM model, WKB initialization cannot be used. We tested linear, RBF, and 3rd order polynomial kernels, with several values for the regularization parameter C, using the OvO strategy. In all experiments, we used 5-fold cross-validation, reserving one session for testing (20% of the data).

Results are given using both unweighted accuracy (UA) and weighted accuracy (WA), as per (1) and (2), where K is the number of classes, N_i and H_i are, respectively, the number of samples and of correctly made predictions for class i, and N is the size of the entire dataset. In general, UA is more relevant when classes are unequally represented.

$$\mathrm{UA} = \frac{1}{K}\sum_{i=1}^{K}\frac{H_i}{N_i} \tag{1}$$

$$\mathrm{WA} = \frac{1}{N}\sum_{i=1}^{K} H_i \tag{2}$$

When using the majority and ranked majority voting schemes (instead of unanimity), average performance was reduced by 10% and by an 15%, respectively, illustrating the less reliable nature of the extended dataset. Since the quality of the data is a most important factor, we kept and reported only the results obtained using the unanimous voting scheme. In Table 1, we present the results for the KMM and GMM mappings, which are very similar, as well as the direct WKB model.

Table 1

Results for KMM, GMM, and direct WKB model mappings

Model	Centroid init.	Metrics
KMM	Native	UA = 75.2%, WA = 71.6%
	WKB	UA = 74.3%, WA = 71.2%
GMM	Native	UA = 75.4%, WA = 72.3%
	WKB	**UA = 74.3%, WA = 72.5%**
WKB	–	UA = 73.4%, WA = 73.1%

Table 2

Results for SVM mapping

Model	Kernel	C	Metrics
SVM	Linear	1	UA = 76.5%, WA = 75.5%
		0.1	UA = 76.4%, WA = 75.1%
		0.01	UA = 75.1%, WA = 71.8%
	RBF	1	UA = 77.0%, WA = 75.0%
		0.1	**UA = 77.3%, WA = 75.7%**
		0.01	UA = 76.3%, WA = 74.1%
	3rd ord. poly.	1	UA = 75.5%, WA = 71.5%
		0.1	UA = 73.3%, WA = 69.6%
		0.01	UA = 60.4%, WA = 56.3%

Whereas Table 2 includes the SVM model metrics. Expectedly, the GMM's unconstrained covariance matrix allows for better data fitting than the KMM, while the SVM model with RBF kernel and C of 0.1 outperforms both.

By contrast, as discussed previously, the direct WKB model theoretically offers the best generalization thanks to the increased reliability of the WKB corpus but is a poorer fit for the IEMOCAP data. Thus, we consider the best compromise between model accuracy and generalization is achieved by the GMM dimensional model with WKB initialization. The KMM dimensional model is illustrated in Fig. 3. The direct WKB model differs only in terms of centroid positions, and, as such, has not been included.

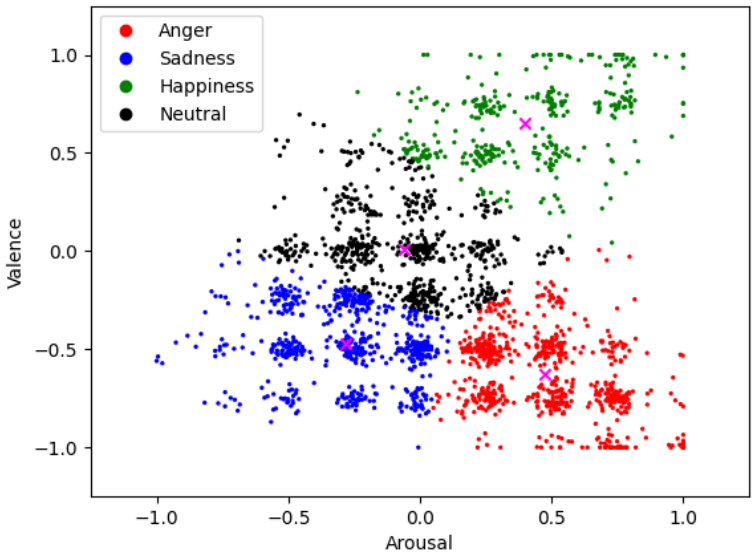

Fig. 3. KMM mapping (dimensional model) with WKB initialization. Cluster centroids are marked with magenta crosses. Data points are colored according to the model output.

Fig. 4 and Fig. 5 represent the GMM and SVM dimensional models, respectively. For the KMM, the boundaries between classes are linear, while those of the GMM and SVM are non-linear. The biggest difference between the illustrated mappings concerns the happiness and neutral classes. The GMM and SVM models confine the neutral class to a central subdomain of the affect space, whilst the KMM model extends it towards low *arousal* and high *valence*, which is not valid. On the other hand, the KMM and GMM models relegate happiness only towards medium and high *arousal*, whilst the SVM model does not offer a valid boundary for this class. The other two classes (anger and sadness) are always associated with correct subdomains within the affect space (e.g., high *arousal* and negative *valence* for anger). These arguments further indicate GMM mapping to be the best compromise.

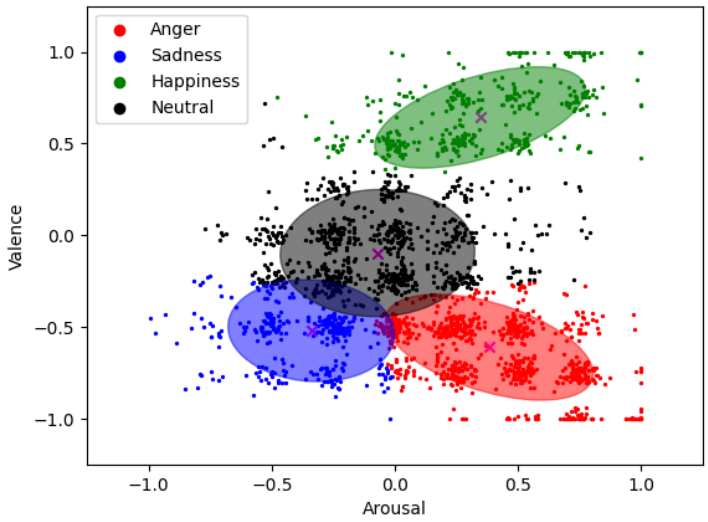

Fig. 4. GMM mapping (dimensional model) with WKB initialization. Cluster centroids are marked with magenta crosses. The 50% probability contours for each class component are also drawn. Data points are colored according to the model output.

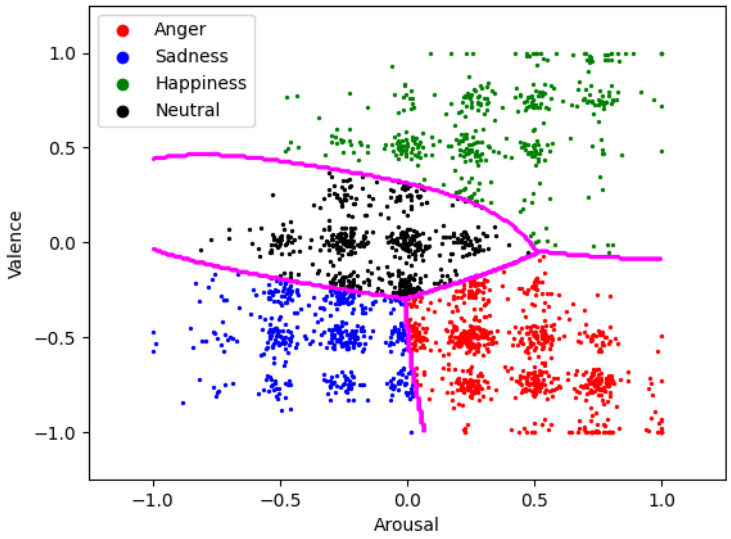

Fig. 5. SVM mapping (dimensional model) with RBF kernel. The magenta lines represent the decision boundaries for each class. Data points are colored according to the model output.

In Table 3, we compare our proposed approach to other works using standard classification systems for discrete emotions. As it can be seen, dimensional models can lead to best performance, as long as reliable affective dimension data exists; in other words, if the overall affect space coordinates of speech segments can be correctly predicted by a regression model.

Table 3

Comparison between the proposed approach maximum performance and existing results

Method	Best results
[7]	WA = 68.6%
[9]	UA = 61.0%
[10]	UA = 65.0%, WA = 66.1%
[11]	UA = 58.8%, WA = 63.5%
[13]	UA = 63.9%, WA = 70.4%
[15]	UA = 66.0%, WA = 70.5%
[16]	UA = 64.7%
[17]	UA = 67.0%, WA = 68.1%
Dimensional model mapping	**UA = 74.3%, WA = 72.5%**

In a fully implemented multidomain system (such as the examples proposed in Fig. 1 – middle and bottom), which would include automatic emotion recognition models, the higher the performance of the regression model, the closer the accuracy of the full system (the final classification) would approach the results reported in this work.

5. Conclusion

In this paper, we proposed unifying the two paradigms of emotion representation into multidomain systems, using dimensional models to map the discrete emotion classes within a continuous (*arousal-valence*) affect space. We used K-means clustering, GMMs, and SVMs to develop such dimensional models based on data from the IEMOCAP database, additionally using the WKB corpus to increase model generalization. Experiments yielded promising results and illustrated the viability of the approach for future work.

R E F E R E N C E S

[1] *B. Schuller*, "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks and Ongoing Trends," Communications of the ACM, vol. 61, no. 5, pp. 90-99, May 2018.

[2] *D. Schuller and B. Schuller*, "The Age of Artificial Emotional Intelligence," Computer, vol. 51, no. 9, Sep. 2018, pp. 38-46.

[3] *P. Ekman*, "An Argument for Basic Emotions," Cognition and Emotion, vol. 6, no. 3-4, May 1992, pp. 169-200.

[4] *J. A. Russell*, 'A Circumplex Model of Affect," Journal of Personality and Social Psychology, vol. 39, no. 6, Dec. 1980, pp. 1161-1178.

[5] *D. Watson, D. Wiese, J. Vaidya, and A. Tellegen*, "The Two General Activation Systems of Affect: Structural Findings, Evolutionary Considerations, and Psychobiological Evidence," Journal of Personality and Social Psychology, vol. 76, no. 5, May 1999, pp. 820-838.

[6] *D. C. Rubin and J. M. Talarico*, "A Comparison of Dimensional Models of Emotion: Evidence from Emotions, Prototypical Events, Autobiographical Memories, and Words," Memory, vol. 17, no. 8, Nov. 2009, pp. 802-808.

[7] *Q. Jin, C. Li, S. Chen, and H. Wu*, "Speech Emotion Recognition with Acoustic and Lexical Features," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Queensland, Australia, Apr. 2015, pp. 4749-4753.

[8] *W. Rao et al.*, "Investigation of Fixed-dimensional Speech Representations for Real-time Speech Emotion Recognition System," in Proceedings of the International Conference on Orange Technologies (ICOT), Singapore, Dec. 2017, pp. 197-200.

[9] *S. Latif et al.*, "Augmenting Generative Adversarial Networks for Speech Emotion Recognition," in Proceedings of INTERSPEECH, Shanghai, China, Oct. 2020, pp. 521-525.

[10] *S. Liu et al.*, "Hierarchical Component-attention Based Speaker Turn Embedding for Emotion Recognition," in Proceedings of the International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, Jul. 2020, pp. 1-7.

[11] *S. Mirsamadi, E. Barsoum, and C. Zhang*, "Automatic Speech Emotion Recognition using Recurrent Neural Networks with Local Attention," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, Mar. 2017, pp. 2227-2231.

[12] *J. Han, Z. Zhang, F. Ringeval, and B. Schuller*, "Prediction-based Learning for Continuous Emotion Recognition in Speech," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, Mar. 2017, pp. 5005-5009.

[13] *Y. Zhang et al.*, "Attention Based Fully Convolutional Network for Speech Emotion Recognition," in Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, Nov. 2018, pp. 1771-1775.

[14] *D. Tang, P. Kuppens, L. Geurts, and T. Van Waterschoot*, "Adieu Recurrence? End-to-end Speech Emotion Recognition using a Context Stacking Dilated Convolutional Network," in Proceedings of the 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, Jan. 2021, pp. 1-5.

[15] *S. Fahad, A. Deepak, G. Pradhan, and J. Yadav*, "DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features," Circuits, Systems, and Signal Processing, Jul. 2020.

[16] *M. Chen, X. He, J. Yang, and H. Zhang*, "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition," IEEE Signal Processing Letters, vol. 25, no. 10, Oct. 2018, pp. 1440-1444.

[17] *Z. Zhao et al.*, "Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition," IEEE Access, vol. 7, Jul. 2019, pp. 97515-97525.

[18] *G. Trigeorgis et al.*, "Adieu Features? End-to-end Speech Emotion Recognition using a Deep Convolutional Neural Network," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016, pp. 5200-5204.

[19] *Z. Yao, Z. Wang, W. Liu, Y. Liu, J. Pan*, "Speech Emotion Recognition using Fusion of Three Multi-task Learning-based Classifiers: HSF-DNN, MS-CNN and LLD-RNN," Speech Communication, vol. 120, Jun. 2020, pp. 11-19.

[20] *C. Busso et al.*, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," Language Resources & Evaluation, vol. 42, no. 4, Nov. 2008, Art. no. 335 (2008).

[21] *J. C. Kim and M. A. Clements*, "Multimodal Affect Classification at Various Temporal Lengths," IEEE Transactions on Affective Computing, vol. 6, no. 4, Oct.-Dec. 2015, pp. 371-384.

[22] *G. S. Morrison, P. Rose, and C. Zhang*, "Protocol for the Collection of Databases of Recordings for Forensic-voice-comparison Research and Practice," Australian Journal of Forensic Sciences, vol. 44, no. 2, Jun. 2012, pp. 155-167.

[23] *A. B. Warriner, V. Kuperman, and M. Brysbaert*, "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas," Behavior Research Methods, vol. 45, no. 4, Dec. 2013, pp. 1191-1207.

[24] *C. Bishop*, Pattern Recognition and Machine Learning, 1st ed., New York, NY, USA: Springer-Verlag, 2006.

[25] *B. Schuller*, "Responding to Uncertainty in Emotion Recognition," Journal of Information, Communication and Ethics in Society, vol. 17, no. 3, Aug. 2019, pp. 299-303.

[26] *G. Rizos and B. Schuller*, "Average Jane, Where Art Thou? – Recent Avenues in Efficient Machine Learning Under Subjectivity Uncertainty," in M.-J. Lesot et al. (Eds.): Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Communications in Computer and Information Science, vol. 1237, Jun. 2020, pp. 42-55.

[27] *C. C. Aggarwal*, "Teaching Deep Learners to Generalize," in Neural Networks and Deep Learning, Cham, Switzerland: Springer International Publishing, 2018, ch. 4, pp. 169-216.