Emotion Intensity and Gender Detection via Speech and Facial Expressions

Elahe Bagheri[1], Oliver Roesler[2], Hoang-Long Cao[1], and Bram Vanderborght[1]

Robotics and Multibody Mechanics Research Group, Vrije Universiteit Brussel and Flanders Make, Brussels, Belgium.
Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium.
elahe.bagheri@vub.be, oliver@roesler.co.uk, hoang.long.cao@vub.be,
bram.vanderborght@vub.be

Abstract. Human emotion detection has received increasing attention over the last decades for a variety of applications and systems. However, detecting the intensity of the expressed emotion has not been investigated as much as detecting the type of the expressed emotion. To fill this gap, we investigate the utility of different facial and speech features for emotion intensity detection. To this end, we designed different Deep Neural Network based models and applied them to the RAVDESS dataset. Obtained results show that speech signal features are better indicators of emotion intensity than facial features. However, in the absence of speech signals, finding emotion intensity by facial expressions is more accurate for males in comparison to females.
The difference between the accuracy of emotion intensity detection for two genders motivated us to use speech signals for the gender detection task. The obtained results confirm that the proposed model achieves higher accuracy in emotion intensity detection and is more robust in gender detection than the state-of-the-art.

Keywords: Emotion Intensity Detection · Gender Detection · Deep Learning.

1 INTRODUCTION

Detecting human emotions is crucial in developing cognitive and adaptive behaviors for artificial intelligent systems, robots, and (virtual) agents. Emotion detection is the ability to recognize another's affective state, which typically involves the integration and analysis of human expressions through different modalities, like facial expression, speech, body movements, and gestures [5]. Mehrabian [21] showed 55% of human emotions are conveyed through facial expression and 38% through speech, therefore, facial and speech emotion recognition received significant attention during the last decades. Although finding the type of expressed emotion is essential to adapt to a user's affective state, it is not enough, and a difference in intensity has been proven to be important to distinguish different emotional states [13]. For instance, a polite smile versus embarrassed smile [1] and posed versus spontaneous smile are separable by differences in their expression intensities [7]. Since there is not much research on emotion intensity detection, in the following sections, the state-of-the-art in speech emotion recognition and facial emotion recognition are discussed.

1.1 Speech Emotion Recognition (SER)

The effect of emotions can be seen in both acoustic characteristics and lexical content of speech. Some examples of acoustic features are Mel-Frequency Cepstral Coefficients (MFCC), energy, jitter, and shimmer, which are known as Low-Level Descriptors (LLDs) [14] [1]. Some examples of lexical features are the presence/absence of word stems, and bag-of-words sentiment categories [28]. However, when the linguistic content is not emotionally rich, recognizing emotion from the transcript is very difficult [28], thus, in this study, our focus is on the acoustic characteristics for recognizing the emotion intensity.
In traditional SER methods, the acoustic features are first extracted and then different machine learning algorithms like Support Vector Machine (SVM) [23], K-nearest neighbor [15] and Hidden Markov model [26] are applied to the obtained features to classify them into considered emotion classes. To obtain these features from utterance-level, each signal is broken into shorter frames of 20 to 50 milliseconds, and their features, i.e., frame-level features, are extracted. However, emotional contents are not in static values of these features but are in their temporal variations. Thus, different statistical functions, e.g., minimum, maximum, mean, variance, linear regression coefficients, etc., are applied to these

[1] Some other investigated acoustic characteristics, i.e., LLDs, are zero-crossing rate, duration, and higher-order formants, Mel-filterbank features, spectral features, formant locations/bandwidths, perceptual linear prediction, fundamental frequency, and pitch.

features to illustrate their temporal variations and contours. The obtained results, afterward, are unified in a vector to achieve utterance-level features [22].

Due to the success of deep learning in different fields like image, video, and natural language processing, the interest in applying Deep Neural Networks (DNN) for speech emotion recognition also increased. Authors in [11] and [17] used deep feed-forward and recurrent neural networks to learn the frame-level acoustic features, and used extreme learning machines for the utterance-level aggregation. Mirsamadi et al. [22] used Rectified Linear Unit (ReLU) dense layers to learn frame-level features, and Bidirectional Long Short-Term Memory (BiLSTM) recurrent layers to learn the temporal aggregation. Neumann et al. [25] used a Convolutional Neural Network (CNN) with one convolutional layer and one pooling layer, to learn the representation of the audio signal, and an attention layer to compute the weighted sum of all the information extracted from different parts of the input. Lim et al. [18] transformed the speech signal into 2D representation using Short Time Fourier Transform and sent them to concatenated CNN and LSTM architectures without using any traditional hand-crafted features. Trigeorgis et al. [32] applied two BiLSTM layers to balance the frame-level characterization and utterance-level aggregation and transform frame-level convolutional features directly into continuous arousal and valence output so that the model learned direct mapping from time-domain speech signals into the continuous model of emotion. The temporal model proposed in [12] used BiLSTM to represent forward/backward contextual information of temporal dynamics of the speech signal and conducted a CNN and a capsule net to learn temporal clusters and classify the extracted patterns. Mustaqeem et al. [24] obtained the spectrogram of signals and then used CNN and LSTM to classify the speech.

1.2 Facial Emotion Recognition (FER)

Ekman [8] showed six basic emotions[2] are expressed universally the same through facial muscles. He introduced the Action Unit (AU) to indicate fundamental movements of a single or group of muscles through the facial expression of a special emotion [3]. He also defined the Facial Action Coding System (FACS) to encode the movements of these AUs [10]. One way to recognize a facially expressed emotion is detecting the status of all individual AUs and then analyzing the combination of the activated AUs to obtain the expressed emotion. On the other hand, promising results of DNN based approaches in comparison with classical machine learning algorithms lead to the proposal of numerous DNN based FER methods in the research community. For instance, Bagheri et al. [2] used facial muscle activities as raw input for a Stacked Auto Encoder (SAE). The applied SAE returns the best combination of muscles in describing a particular emotion, which is then sent to a Softmax layer to fulfill the multi-classification task. Liu et al [19] proposed a sign-based DNN architecture to investigate the effect of AUs in emotion recognition. The proposed model consists of three sequential modules, where the first module generates a complete representation of all expression-specific appearance variations by a convolution layer stacked by a max-pooling layer. The second module finds the best simulation of the combination of the AUs and the last module learns hierarchical features by Restricted Boltzmann Machines (RBM). Finally, a linear SVM classifier is used to recognize the six basic emotions. However, AU-aware layers, in the second module, are not able to detect all FACS in images. Pitaloka et al [27] used CNN to extract features from an input image, which is then passed to a max-pooling layer to reduce the image size. A fully connected layer, in the end, classifies the input image into one of the six basic emotions. However, the performance of the proposed algorithm decreases when the dimension of the input image increases regarding the complexity of the high dimensional images.

Research on emotion intensity detection has been focused on the estimation of the intensity of Action Units (AU), e.g., [6, 13] and FERA 2015[4], however, there is no conclusion about the intensity of the expressed emotion, thus, the goal of this study is developing a model by which the intensity of the expressed emotion in a given image, speech signal or video can be identified.

The remainder of this paper is structured as follows: the applied models, dataset, and extracted features are explained in Section 2. Section 3 demonstrates the conducted experiments and obtained results. Finally, Section 4 concludes this paper.

[2] Happiness, sadness, fear, anger, surprise, and disgust.

[3] `https://imotions.com/blog/facial-action-coding-system/`

[4] Facial Expression Recognition and Analysis challenge

Table 1
The architectures of the proposed DNN based models.

LSTM		**BiLSTM**		**CNN**		
Simple	**Attention**	**Simple**	**Attention**	**Simple**	**BiLSTM/LSTM**	**BiLSTM/LSTM+Attention**
LSTM	LSTM	BiLSTM	BiLSTM	CNN	CNN	CNN
Dropout	Attention	Dropout	Attention	CNN	CNN	MaxPooling
Dense	Dropout	Dense	Dropout	Dropout	MaxPooling	CNN
	Dense		Dense	MaxPooling	Flatten	MaxPooling
				Flatten	BiLSTM/LSTM	Flatten
				Dense	Dense	BiLSTM/LSTM
				Dense		Attention
						Dropout
						Dense

2 METHODOLOGY

2.1 Applied Models

Table 1 shows the number, type, and order of layers of proposed models that are applied to fulfill the emotion intensity detection task. The parameter settings are as follows: convolution layers are all 1D and have 64 filters and kernel size of three. ReLU activation function is applied for adding non-linearity. Dropout layers are used as regularizers and their ratio is set to 0.1. 1D max-pooling layers, with a kernel size of four are used to introduce sparsity in the network parameters and to learn deep feature representations. Dense layers are used with the activation functions of sigmoid for finding the predicted binary distribution of the target class. The number of epochs is selected as 250 and the batch size is set to 128. The number of units in applied LSTM and BiLSTM networks is five.

2.2 Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [20] is used to train and test the proposed models. RAVDESS contains videos which provide both facial and speech features. Each video lasts approximately three seconds and contains frontal face poses of twelve female and twelve male, all north American actors and actresses, speaking and singing two lexically-matched sentences while expressing six basic emotions plus calmness, and neutral. In this study, we only used speaking records. Further, expressed emotions in RAVDESS are categorized into two levels of normal and strong intensities, which makes it a good option for emotion intensity detection.

The dataset is partitioned, in a subject independent manner, into the train and test sets, i.e., the videos related to the first eighteen actors (nine female, nine male) are used for training, and the videos of the next six subjects (three female, three male) are used for testing. Since videos are recorded with 30 fps, 30 images are extracted per second for facial analysis. However, as each video starts from a neutral state, reaches an apex, and goes back to a neutral state, the first and last twenty obtained frames per video are discarded.

Since the expressions of neutral are not categorized as normal or strong, this emotional state is not considered for emotion intensity detection. Additionally, the normal and strong expressions of calmness are only marginally different, therefore, they are omitted from the emotion intensity detection.

2.3 Features and Data Pre-processing

To obtain facial and speech features, two open-source toolkits, i.e., OpenFace [3] and openEAR [9], are used. OpenFace is able to return different features including facial landmarks, head pose, facial action units activity, and eye-gaze from both video and image inputs. The applied features in this study are facial landmarks, facial action units activity, and face rigid and non-rigid shape parameters leading to a vector of 378 elements [5]. The obtained feature values are normalized between zero and one.

[5] Other features provided by OpenFace were also investigated, however, the mentioned features resulted in the highest classification accuracy.

Table 2
The obtained accuracies for emotion intensity detection by proposed models on RAVDESS dataset based on facial and speech features (without neutral and calmness expressions). The results were obtained over ten repetitions.

(a) Facial features.

Data	LSTM		BiLSTM		CNN			CNN	
	Simple	Attention	Simple	Attention	Simple	LSTM	BiLSTM	LSTM+Attention	BiLSTM+Attention
Female and Male	53.34	52.46	53.27	54.13	55.96	55.06	54.79	55.31	**56.24**
Female	51.67	50.31	50.14	51.12	52.50	53.52	51.72	50.63	**54.72**
Male	54.24	53.72	52.81	53.66	58.38	56.45	54.26	56.97	**58.31**

(b) Speech features.

Data	LSTM		BiLSTM		CNN			CNN	
	Simple	Attention	Simple	Attention	Simple	LSTM	BiLSTM	LSTM+Attention	BiLSTM+Attention
Female and Male	63.5	61.70	67.65	69.03	69.45	68.54	69.46	60.53	**73.53**
Female	68.51	67.62	67.41	69.52	72.61	72.45	72.43	59.83	**75.67**
Male	57.36	55.30	55.97	55.21	59.73	59.18	59.72	58.87	**65.5**

openEAR is the open-source toolkit that is used for speech feature extraction. It analyses the speech signals and returns three different sets of features based on the applied configuration, i.e., INTERSPEECH 2009, emobase, and INTERSPEECH 2013. In this study, the INTERSPEECH 2009 (emo-IS09) [29] configuration is used, which leads to 384 features including minimum, maximum, and mean values of different speech features. In this study, only the MFCC and PCM set of features are used, which lead to a vector of 156 elements [6]. The obtained feature values are normalized between zero and one.

3 EXPERIMENTAL SCENARIOS AND OBTAINED RESULTS

3.1 Experiment I: Emotion Intensity Detection

As the main goal of this study is to identify the intensity of an expressed emotion by a user, the facial and speech related features of all subjects are extracted and pre-processed (as explained in Section 2.3) and are given to the proposed models (Table 1). The obtained results in Table 2 show that speech features lead to higher accuracy in emotion intensity detection than facial features. Further, Table 1 shows the combination of convolutional layer with the BiLSTM and Attention layers (CNN+BiLSTM+Att) achieves the highest performance, i.e., 73.53%, which is higher than state-of-the-art, i.e., 70.4% [12] (Table 4).

Although speech features lead to higher accuracy than facial features, Table 2.a shows that the accuracy of the models in identifying the intensity of the expressed emotions by males is higher than expressed emotions by females when facial features are used, i.e., 58.31% vs 54.72%. In comparison, when speech features are used, the obtained accuracy for females is higher than for males, i.e., 75.67% vs 65.5% (Table 2.b). La Mura [16] showed some of the speech features related to emotion recognition are related to the subject's gender. Thus, one explanation can be that females convey more details about the intensity of their emotions through their speech. To verify this hypothesis, the CNN+BiLSTM+Att model is separately applied to both the facial and speech features of each individual subject. The obtained results (Table 3.a) show that for males obtaining the emotion intensity by facial expressions is more accurate than for females, i.e., the minimum and maximum accuracies for males are 63.92% and 78.79%, respectively, while the corresponding values for females are 58.26% and 71.15%. On the other hand, Table 3.b shows finding emotion intensity via speech features for females is more accurate than males, i.e., the minimum and maximum accuracies for females are 71.49% and 95.83% while the corresponding values for males are 59.29% and 85.71%, respectively.

[6] Different combinations of features were used, however, the highest accuracy was obtained for the the applied feature set, thus, we did not use the other features. In addition, since openEAR is able to analyze a file, we did not trim the videos into smaller intervals, which reduced the running time remarkably.

Table 3
Accuracy of emotion intensity detection based on facial and speech data for males and females. The results obtained over ten repetitions.

(a) Facial features.

Male	Acc	STD	Female	Acc	STD
Sub#1	64.97	2.3	Sub#2	68.36	1.6
Sub#3	77.53	2.5	Sub#4	66.05	2.5
Sub#5	78.79	1.3	Sub#6	69.46	1.4
Sub#7	75.85	2.6	Sub#8	62.17	5.3
Sub#9	72.62	1.8	Sub#10	67.55	1.2
Sub#11	72.71	2.1	Sub#12	68.42	4.7
Sub#13	63.92	3.1	Sub#14	69.19	2.2
Sub#15	66.03	2.5	Sub#16	63.11	3.6
Sub#17	67.78	2.1	Sub#18	71.15	2.1
Sub#19	73.23	2.3	Sub#20	69.32	2.8
Sub#21	74.91	1.9	Sub#22	61.62	1.4
Sub#23	73.84	2.7	Sub#24	58.26	1.9

(b) Speech features.

Male	Acc	STD	Female	Acc	STD
Sub#1	82.86	8.3	Sub#2	88.46	4.2
Sub#3	84.21	3.0	Sub#4	87.85	4.8
Sub#5	59.29	4.8	Sub#6	85.57	4.1
Sub#7	80.71	5.8	Sub#8	75.49	4.9
Sub#9	85.71	6.3	Sub#10	71.49	3.3
Sub#11	70.49	7.3	Sub#12	78.57	6.7
Sub#13	70.01	11.0	Sub#14	74.29	3.6
Sub#15	60.00	7.6	Sub#16	79.23	5.1
Sub#17	85.71	6.7	Sub#18	95.83	4.3
Sub#19	74.29	6.9	Sub#20	67.17	8.8
Sub#21	67.50	7.3	Sub#22	72.5	9.6
Sub#23	83.50	9.8	Sub#24	95.38	5.3

Table 4
Comparison between the proposed model and the state-of-the-art for emotion intensity detection over RAVDESS on speech features in a subject independent manner.

Research	Architecture	Accuracy
Jalal [12]	CNN + BiLSTM + CapsuleNet	70.4%
Proposed model	CNN + BiLSTM + Attention	**73.53%**

3.2 Experiment II: Gender Detection

Since the obtained accuracies for emotion intensity detection for males and females are noticeably different, in this experiment we investigated the speech and facial features for the task of gender detection. As the results of the proposed models (Table 1) by using facial features were not promising, a new model was designed for this experiment. The new proposed model uses raw images of 200 × 200 pixels as input and consists of four layers, wherein each a 2D convolutional layer is followed by a max-pooling and a dropout layer. The kernel size of the convolution layers is 3×3, with the same padding size, and ReLU is used as the activation function. The max-pooling layer is 2 × 2 and dropout rates in different layers are set to 0.6, 0.4, 0.2, and 0.2, respectively. The batch size during the train and test is set to 32. The first eighteen subjects are used for training and the last six subjects are used for testing (subject independent and gender balance). The obtained accuracy of this model is 70.46%.

Repeating the experiment with the speech features led to higher accuracy for gender detection via the proposed models in Table 1. More specifically, CNN+BiLSTM+Att model obtained an accuracy of 89.8% for gender detection using the MFCC and PCM feature sets, which is the highest obtained accuracy in comparison with the other proposed models in Table 1. Table 5 shows the obtained confusion matrix by the proposed model for gender detection. We noticed that 20 of the female samples that are wrongly predicted as male belong to one subject.

A straightforward comparison between the proposed model and the state-of-the-art for gender detection task, using speech signals of RAVDESS dataset, is difficult. For instance, Singh et al. [31] performed gender detection in each individual emotion class assuming the emotion class is known. Bansal et al. [4] used only four expressions of RAVDESS for gender detection and obtained an accuracy of 94.12%, and Shaqra et al. [30] considered six emotions and obtained an accuracy of 98.67%, while a gender detection model should be robust to various emotions. Thus, in this study, we used all expressed emotional states in RAVDESS dataset, i.e., eight emotional states, for the task of gender detection. Table 6 compares the obtained accuracy by the proposed model with the state-of-the-art. Although the proposed model could not beat the state-of-the-art, it is more robust since it considers more emotional states.

Table 5
Confusion matrix for gender detection.

	Predicted Female	Predicted Male
Actual Female	143	25
Actual Male	9	159

Table 6
Comparison between the proposed model and the state-of-the-art for gender detection over RAVDESS on speech features in a subject independent manner.

Research	Model	Accuracy
Bansal et al. [4] (four emotional states)	SVM	94.12%
Shaqra et al. [30] (six emotional states)	MLP	98.67%
Proposed model (eight emotional states)	CNN + BiLSTM + Attention	89.8%

4 CONCLUSION

In this study, we designed different deep neural network based models for emotion intensity and gender detection using features obtained by open-source toolkits. The RAVDESS dataset was used to evaluate the proposed models because it is, to the best of our knowledge, the only dataset that categorizes emotions based on their intensity.

The obtained results showed a difference between the obtained accuracy of emotion intensity detection for females and males based on the applied feature set, i.e., using facial features led to more accurate results for males than for females, while using speech features led to higher accuracy for females' emotion intensity detection. Additionally, the results showed that the MFCC and PCM feature sets led to higher accuracy than facial features in emotion intensity detection. Further, we used the proposed models for gender detection task using facial and speech features. The obtained results showed that gender detection is also more accurate by using speech features than facial features for the RAVDESS dataset. In addition, the obtained results showed that the proposed model is comparable with the state-of-the-art while it is more robust in terms of handling more emotional states.

ACKNOWLEDGEMENTS

The work leading to these results has received funding from Flanders Make Proud (PROgramming by User Demonstration) and the Flemish Government under the program Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen.

Bibliography

[1] Ambadar, Z., Cohn, J.F., Reed, L.I.: All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. Journal of nonverbal behavior **33**(1), 17–34 (2009)

[2] Bagheri, E., Bagheri, A., Esteban, P.G., Vanderborgth, B.: A novel model for emotion detection from facial muscles activity. In: Iberian Robotics conference. pp. 237–249. Springer (2019)

[3] Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. pp. 59–66. IEEE (2018)

[4] Bansal, M., Sircar, P.: Phoneme based model for gender identification and adult-child classification. In: 2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS). pp. 1–7. IEEE (2019)

[5] Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., Karpouzis, K.: Multimodal emotion recognition from expressive faces, body gestures and speech. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 375–388. Springer (2007)

[6] Cohn, J.F., Kanade, T., Li, C.C.: Subtly different facial expression recognition and expression intensity estimation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 853–859 (1998)

[7] Cohn, J.F., Schmidt, K.: The timing of facial motion in posed and spontaneous smiles. In: Active Media Technology, pp. 57–69. World Scientific (2003)

[8] Ekman, P.: Facial action coding system (facs). A human face (2002)

[9] Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1459–1462 (2010)

[10] Hamm, J., Kohler, C.G., Gur, R.C., Verma, R.: Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. Journal of neuroscience methods **200**(2), 237–256 (2011)

[11] Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Fifteenth annual conference of the international speech communication association (2014)

[12] Jalal, M.A., Loweimi, E., Moore, R.K., Hain, T.: Learning temporal clusters using capsule routing for speech emotion recognition. In: Proc. Interspeech. vol. 2019, pp. 1701–1705 (2019)

[13] Jeni, L.A., Girard, J.M., Cohn, J.F., De La Torre, F.: Continuous au intensity estimation using localized, sparse facial feature space. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). pp. 1–7. IEEE (2013)

[14] Jin, Q., Li, C., Chen, S., Wu, H.: Speech emotion recognition with acoustic and lexical features. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 4749–4753. IEEE (2015)

[15] Kim, Y., Provost, E.M.: Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3677–3681. IEEE (2013)

[16] La Mura, M., Lamberti, P.: Human-machine interaction personalization: a review on gender and emotion recognition through speech analysis. In: 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT. pp. 319–323. IEEE (2020)

[17] Lee, J., Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: Sixteenth annual conference of the international speech communication association (2015)

[18] Lim, W., Jang, D., Lee, T.: Speech emotion recognition using convolutional and recurrent neural networks. In: 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA). pp. 1–4. IEEE (2016)

[19] Liu, M., Li, S., Shan, S., Chen, X.: Au-aware deep networks for facial expression recognition. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–6. IEEE (2013)

[20] Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one **13**(5), e0196391 (2018)

[21] Mehrabian, A.: Nonverbal communication. Routledge (2017)

[22] Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2227–2231. IEEE (2017)

[23] Mower, E., Mataric, M.J., Narayanan, S.: A framework for automatic human emotion classification using emotion profiles. IEEE Transactions on Audio, Speech, and Language Processing **19**(5), 1057–1070 (2010)

[24] Mustaqeem, M., Kwon, S., et al.: A cnn-assisted enhanced audio signal processing for speech emotion recognition. Sensors **20**(1), 183 (2020)

[25] Neumann, M., Vu, N.T.: Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. arXiv preprint arXiv:1706.00612 (2017)

[26] Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden markov models. Speech communication **41**(4), 603–623 (2003)

[27] Pitaloka, D.A., Wulandari, A., Basaruddin, T., Liliana, D.Y.: Enhancing cnn with preprocessing stage in automatic emotion recognition. Procedia computer science **116**, 523–529 (2017)

[28] Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Vembu, A.N., Prasad, R.: Emotion recognition using acoustic and lexical features. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)

[29] Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. In: Tenth Annual Conference of the International Speech Communication Association (2009)

[30] Shaqra, F.A., Duwairi, R., Al-Ayyoub, M.: Recognizing emotion from speech based on age and gender using hierarchical models. Procedia Computer Science **151**, 37–44 (2019)

[31] Singh, R., Puri, H., Aggarwal, N., Gupta, V.: An efficient language-independent acoustic emotion classification system. Arabian Journal for Science and Engineering **45**(4), 3111–3121 (2020)

[32] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5200–5204. IEEE (2016)