Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 1011–1015
DOI: 10.15439/2023F4485
ISSN 2300-5963 ACSIS, Vol. 35

Gender-aware speaker's emotion recognition based on 1-D and 2-D features

Włodzimierz Kasprzak, Mateusz Hryciów
0000-0002-4840-8860, 0000-0000-0000-0000
Warsaw University of Technology, Institute of Control and Computation Eng.
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland
Email: wlodzimierz.kasprzak@pw.edu.pl

***Abstract*—An approach to speaker's emotion recognition based on several acoustic feature types and 1D convolutional neural networks is described. The focus is on selecting the best speech features, improving the baseline model configuration and integrating in the solution a gender classification network. Features include a Mel-scale spectrogram and MFCC-, Chroma-, prosodic- and pitch-related features. Especially, the question whether to use 2-D maps of features or reduce them to 1-D vectors by averaging, is experimentally resolved. Well–known speech datasets RAVDESS, Tess, Crema-D and Savee are used in experiments. It appeared, that the best performing model consists of two convolutional networks for gender-aware classification and one gender classifier. The Chroma features have been found to be obsolete, and even disturbing, given other speech features. The f1 accuracy of proposed solution reached 73.2% on the RAVDESS dataset and 66.5% on all four datasets combined, improving the baseline model by 7.8% and 3%, respectively. This approach is a serious alternative to other proposed models, which reported accuracy scores of 60% - 71% on the RAVDESS dataset.**

I. Introduction

A HUMAN is able to effectively recognize the emotions of the speaker, but attempts to create automatic systems of this type (SER - speaker's emotion recognition) give quite limited results – their intensive development is still carried out [1]. SER systems can find a variety of applications: detecting the emotions of mobile phone users, call center operators and customers, car drivers and other participants of human-machine communication [2]. In some situations, this would allow computer-generated characters to be used to have natural conversations by appealing to human character. Only with this ability is it possible to achieve a fully meaningful dialogue between man and machine.

Human emotionality includes personality, character, temperament, and inspiration as the main psychological parameters that drive human emotions." It can therefore be concluded that there are different sources of communication through which a person expresses his emotions. These include, among others, facial expressions, gestures, speech and writing. On the basis of each of these methods, models can be constructed that will enable the recognition of emotions. One can expect, that a reliable emotion recognition system will require the use of various information modalities, like face videos synchronized with speech and wearable sensor recordings [3].

In the theory of basic emotions [4], it is assumed that people have a limited number of emotions (e.g. joy, anger, fear) that are biologically and psychologically basic. Each of them manifests in the majority of society in an organized, repeating pattern of related behavioral components [5]. Thanks to this, it is possible to label them. However, for this purpose, the number of possible classes must be specified. One of the theories was developed by Robert Plutchik. He distinguished eight basic emotions: joy, acceptance, fear, surprise, sadness, anger, disgust, and anticipation [6].

Despite the relatively easy task of classifying emotions into several categories, it can be problematic even for people if they are not exaggerated. It turns out that in the case when they need to determine the emotional state of an unknown person, the recognition rates are about 60% [7]. Errors in labelling can also affect the quality of automatic classification systems at a later stage.

Currently developed systems for speaker's emotion classification are based on acoustic modeling used in speaker recognition (speaker identification and verification) systems [8]. The basic classic machine learning methodologies used for this problem are UBM-GMM (Universal Background Model - Gaussian Mixture Model) and "i-vectors" [9]. An early solution based on deep neural networks is the "x-vector" [10] network.

In section 2, related work and database issues are introduced in more details. The implemented system SER (speaker's emotion recognition) is presented in section 3. The main experiments and emotion classification results are shown in section 4. At the end, in section 5, we conclude the work.

II. Related work

A. Speech features and recognition techniques

The development of SER systems is inseparably connected with the use of machine learning techniques. It turns out that the vast majority of solutions used to recognize emotions based on speech are based on these solutions. These include [11]: neural networks (NN), convolutional neural networks (CNN), deep neural networks (DNNs), hidden Markov models (HMM), support vector machines (SVM) [12], decision trees and random forests [13].

This work was supported by "Narodowe Centrum Badań i Rozwoju", Warszawa, Poland, grant No. CYBERSECIDENT/455132/III/NCBR/2020. The publication was funded by Warsaw University of Technology.

In various studies, it has been shown that reducing the number of features has a positive effect on classification. It increases the generalization abilities of individual models, and the time of their training decreases. What's more, it turned out that reducing the number of features did not negatively affect the accuracy of emotion prediction, and sometimes even led to its improvement. For example, using the SVM algorithm, reducing the number of attributes from 276 to 75 resulted in an increase in the recognition rate by 5% [12]. In other studies, using the random forest method, selecting 16 traits out of 84 dropped the accuracy by 5%[13].

From the analysis of existing methods we can conclude the following motivation for our research work:

1) Acoustic features of man and women usually differ. For example, the fundamental frequency (pitch) of women voice is usually higher than man.
2) Prosodic features are useful in emotion recognition. The correlation between prosodic features and emotional states of has been already demonstrated [14]. Thus, our work will take such features into account.
3) The use of pitch classes for emotion recognition need to be evaluated. A pitch class is a set of notes with a given halftone pitch from all octaves. Chroma features associated with pitch classes are most often used to recognize emotions in musical works, but some works indicate their correlation with emotional states expressed by speech [15].
4) As the problem is similar to the classification of speaker groups, the solution can well be based on speaker identification methods, like GMM-UBM, JFA, i-vectors or x-vectors. However, we focus on lightweight neural network models using network types, like MLP and CNN.

B. Datasets

Four databases of annotated speech recordings were used for training and testing of the proposed speaker's emotion recognition (SER) system - the "Ryerson Audio-Visual Database of Emotional Speech and Song" (RAVDESS) [16], the "Toronto emotional speech set" (TESS) [17], the "Crowd-sourced Emotional Multimodal Actors" dataset (Crema-D) [15], and the "Surrey Audio-Visual Expressed Emotion" dataset (Savee) [18]. We use the RAVDESS and TESS databases in initial experiments, dealing with the selection of feature sets and the tuning of network models. An overall evaluation of the proposed solution is also given on a combination of the four datasets. The RAVDESS set is most important for us, as it contains recordings of 24 actors (12 Male and 12 Female voices) and annotates the full number of 8 emotional states: *angry, disgust, fear, happy, neutral, sad, surprise* and *calm*. The TESS database is larger, but uses only the first 6 emotion classes and contains samples of 2 female actors only. Savee holds samples of 7 emotion classes (no *calm* class) from 4 male speakers and Crema-D – samples of first 6 classes only from 91 actors (48 male and 43 female actors). In RAVDESS, every actor delivered 60 sentences with the content "Kids are talking by the door" or "Dogs are sitting by the door", giving a total of 1440 recordings with an average recording length of approximately 3.7 s.

III. System SER

A. Structure

The SER (System for Emotion Recognition) solution was designed with a general structure shown in Figure 1. There are three basic stages of processing:

1) Signal segmentation and detection of acoustic features;
2) Gender classification;
3) Two emotion classifiers - trained separately for male and female speakers.

The results of studies of models with different configurations indicate that it is useful to pre-classify the gender of the speaker and train separate models for male and female voices. Hence, the target configuration of the emotion classifier has three networks - one gender model for binary classification into Male and Female speaker and two emotion models for male and female emotion classification.

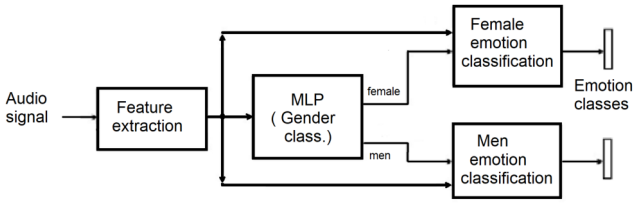

Fig. 1. Structure of the SER solution

B. Acoustic features

The feature vector in our solution can contain features of six types: zero-crossing rate (ZCR), Chroma, RMS value of the signal (root mean square), MFCC–based features, Mel-spectrogram coefficients, and prosodic features (e.g., mean and variance of the fundamental frequency). To determine the above features, the audio signal is processed by functions from the librosa library [19]. Different combinations of the above characteristics were investigated as well as individual types of features. In a final chosen solution, based on 1D convolutional layers, the features are averaged over time (over signal frames) to give a 1D input data – each recording may be represented by a vector of up to 162 features. We also experimented with solutions based on true 2D CNNs and LSTM networks – the feature averaging was omitted to provide a map of 2D features as input to the neural networks (Fig. 2).

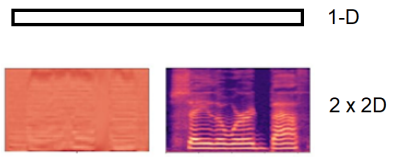

Fig. 2. Alternative input features of SER: a 1D vector of features averaged over time, two 2D maps of features

Regarding a prosody feature, we tracked the basic frequency (”pitch", F0) over time. In general, there are useful prosody features, like: waveform F0 (mean, minimum, maximum, variance), average energy of the voiced and voiceless parts, tempo of speech (inverse of the average time of voiced parts in a statement).

C. Gender model

For gender classification, we use an MLP network with 1D input (Figure 3). The model has three fully connected hidden layers containing 400 neurons each followed by a 0.1 dropout for each layer.

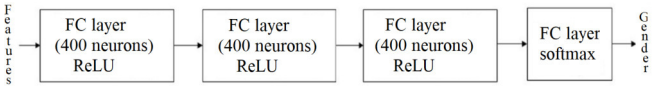

Fig. 3. Structure of the MLP network for gender classification

D. 1D-based SER

For emotion classification, we use a baseline solution built from 1D convolutional layers [20]. There are four layers with 1x5 masks followed by ”MaxPooling1D” layers. After the flattening layer there are two Dense layers - one with ReLu activation and the other with softmax. During experiments with various feature sets and model parameters, under a gender-aware policy, the performance of this baseline model was improved by several percent. The final architecture of our emotion classifier (used both for man and female emotions) differs from the base network mainly by the final two layers and the input vector (Table I). The last convolutional layer has 96 filters (replacing 64), while the fully connected layer has 128 neurons (replacing 32). The input data consists of 150 features, as the Chroma features have been found useless for this network configuration.

TABLE I
THE MODEL ”LAYER 128-96” FOR EMOTION RECOGNITION

Layer (type)	Output (shape)	Param. number
conv1d (Conv1D)	(None, 150, 256)	1536
max_pooling1d (MaxPooling1D)	(None, 75, 256)	0
conv1d_1 (Conv1D)	(None, 75, 256)	327936
max_pooling1d_1 (MaxPooling1D)	(None, 38, 256)	0
conv1d_2 (Conv1D)	(None, 38, 128)	163968
max_pooling1d_2 (MaxPooling1D)	(None, 19, 128)	0
dropout (Dropout)	(None, 19, 128)	0
conv1d_3 (Conv1D)	(None, 19, 96)	61536
max_pooling1d_3 (MaxPooling1D)	(None, 10, 96)	0
flatten (Flatten)	(None, 960)	0
dense (Dense)	(None, 128)	123008
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 8)	1032
Total params: 679016 Trainable params: 679016 Non-trainable params: 0		

E. 2-D based SER

For a 2-D data input, models based on CNN networks were studied. We considered two processing streams - one for the MFCC-based feature map and one for the Mel-spectrogram. Each 2D map is processed by a CNN model with convolutional layers and 3 maxpooling layers. The outputs of both models are concatenated and processed by a fully connected layer with softmax activation.

F. Data augmentation

A well-known approach in machine learning is data augmentation [21]. For automatic increase of the number of annotated recordings, there will be synthetic secondary recordings generated from existing recordings by following operations: adding noise, stretching or compressing the signal over time, and changing the frequency of the basic tone (F0). The operation of adding noise does not change the envelope of the signal, so it also does not change the type of emotion in the recording. Other operations can change the class, so the scope of changes has been limited.

G. Evaluation metrics

The metrics used to evaluate a classification system is typically based on the following counts of model prediction: TN – True Negative, TP – True Positive, FN – False Negative, FP – False Positive. The appropriate relationships of the above results lead to evaluation metrics of the system:

- Precision – the correctness degree of the positive prediction result of a given class: $Precision = TP/(TP + FP)$ Observe that ”false positive rate”: $FPR = 1 - Precision$.
- Recall = TPR (true positive rate) – the degree of correct prediction of examples of a given class: $Recall = TP/(TP + FN) = TPR$
- F1 score – the degree of correct positive predictions: $F1 = 2 \cdot (Recall \cdot Precision)/(Recall + Precision)$
- ROC (receiver operating characteristic) – a curve, $TPR = f(FPR)$ that relates true positive rate (TPR) versus false positive rate (FPR) (where such pairs of values are obtained for the same decision thresholds).
- AUC (area under the ROC curve) – it determines the probability that the classifier will rank a random example of a positive class higher than a random example of a negative class. Ideally, its value is 1.

IV. EXPERIMENTAL RESULTS

A. Gender classification

The gender classifier (Fig. 3) was tested on the RAVDESS set (24 speakers). The accuracy on the test set was 98.7% and the AUC value was 0.9993. These results show a high quality of the proposed model.

B. Effect of recording times

The effect of time (length) of the recording onto the emotion classification was tested using the MLP model as emotion classifier (with 8 outputs) and all the 165 features (Table II). For the two datasets (RAVDESS and TESS), the classification accuracy increased with increasing recording time and saturated between 4 and 6 seconds. In further experiments, sequences with a length of 4 s were chosen for analysis.

TABLE II
DIFFERENT LENGTHS OF RECORDINGS AND THEIR INFLUENCE ONTO CLASSIFICATION ACCURACY

Length	TESS		RAVDESS	
[s]	Test accuracy	AUC field	Test accuracy	AUC field
2	0,78	0,9725	0,43	0,8274
3	0,92	0,9945	0,55	0,9029
4	0,98	0,9998	0,60	0,9117
6	0,99	1,0000	0,62	0,9418

C. Effect of input features

In early experiments with the MLP-based emotion classification on the RAVDESS dataset, we studied the effect of using single types of features. The following test accuracies (F1 score) were observed: for prosodia features 0.49, MFCC 0.39, Mel-spectrogram 0.37, Chroma 0.20, all features 0.62. Thus, when combining all the features into an input vector for the 1D emotion classifier, the effect of cancelling Chroma features was tested. It turned out, that the modified baseline network, in every configuration performs better without the Chroma features (Table III). The presented results of training and test (validation) accuracies come after 50 or 100 training epochs, with batch size 64.

TABLE III
EFFECT OF MODIFYING THE FEATURE SET AND THE BASE MODEL ON EMOTION CLASSIFICATION ACCURACY

Model	All features (162)		No Chroma (150)	
	Avg. recall	F1	Avg. recall	F1
Baseline (100 ep.)	0.64	0.654	0.67	0.681
Layer 128-64 (100 ep.)	0.69	0.696	0.71	0.716
Layer 128-96 (100 ep.)	0.69	0.700	0.72	0.719
Baseline (50 ep.)	0.66	0.661	0.66	0.664
Layer 64-128 (50 ep.)	0.66	0.678	0.67	0.684
Layer 96-64 (50 ep.)	0.66	0.655	0.67	0.682
Layer 64-64 (50 ep.)	0.67	0.672	0.68	0.690
Layer 64-96 (50 ep.)	0.67	0.669	0.68	0.685
Layer 96-96 (50 ep.)	0.67	0.676	0.68	0.693

D. Effect of layer size modification

The experimental results, collected in Table III, also show, that by increasing the number of neurons in the FC layer to 128, and the number of filters in the last convolutional layer to 96, a significant increase of the performance can be achieved. This configuration is denoted as "Layer 128-96".

E. Gender-aware emotion classification

We complete the experiments on our 1D SER approach by training and testing two separate emotion classifiers – one for Males and one for Females. The results provided in Table IV are twofold. There is no significant increase of the performance of the baseline model when applied under perfect gender classification conditions. The practical observable performance is even slightly lower for the gender-aware (G-A) solution. A different effect is observed for our best modified model (Layer 128-96) used for separate gender emotion modelling – the F1 accuracy is now increased by 2.3% (in theory) and 1.3% (in practice) on the RAVDESS dataset.

TABLE IV
EFFECT OF GENDER-CONTROLLED EMOTION CLASSIFICATION

Model	All features (162)		No Chroma (150)	
	Avg. recall	F1	Avg. recall	F1
Baseline female	0.68	0.7019	0.73	0.7426
Baseline male	0.58	0.6019	0.61	0.6222
2x Baseline theoretic	0.63	0.6519	0.67	0.6824
2x Baseline real	0.622	0.645	0.664	0.675
Female Layer 128-96	0.69	0.7241	0.76	0.7796
Male Layer 128-96	0.66	0.6722	0.70	0.7037
G-A 2x Layer 128-96	0.675	0.6982	0.73	0.7417
Real G-A 2x Layer 128-96	0.667	0.689	0.722	0.732

F. Comparison on 4 datasets

Finally, the baseline solution has been compared with our best modified model (Layer 128-96, no Chroma) by training and testing both on the four available datasets: RAVDESS, TESS, Crema-D and Savee. Classification reports are given in Figure 4. Again, our modified model keeps an advantage of 2% in the F1 score.

	precision	recall	f1-score	support
angry	0.79	0.74	0.76	1438
calm	0.77	0.74	0.76	137
disgust	0.57	0.50	0.54	1468
fear	0.58	0.60	0.59	1424
happy	0.61	0.61	0.61	1462
neutral	0.59	0.60	0.59	1310
sad	0.59	0.68	0.64	1400
surprise	0.82	0.84	0.83	483
accuracy			0.63	9122
macro avg	0.67	0.66	0.66	9122
weighted avg	0.64	0.63	0.63	9122

(a)

	precision	recall	f1-score	support
angry	0.78	0.79	0.78	1438
calm	0.71	0.80	0.75	137
disgust	0.60	0.57	0.58	1468
fear	0.62	0.59	0.60	1424
happy	0.62	0.62	0.62	1462
neutral	0.60	0.59	0.59	1310
sad	0.61	0.67	0.64	1400
surprise	0.84	0.86	0.85	483
accuracy			0.65	9122
macro avg	0.67	0.68	0.68	9122
weighted avg	0.65	0.65	0.65	9122

(b)

Fig. 4. Classification reports of (a) the baseline model and (b) our modified baseline model, when trained and evaluated on four emotion datasets

G. The 2D SER

Two-dimensional data is in the form of a feature map, where one axis represents discrete time (indexes of subsequent frames) and the other feature indexes for one frame. We studied maps representing the mel-spectral features and cepstral features of the MFCC. Each feature map was obtained from a recording with a duration of 4 seconds. We applied CNN or CNN+LSTM models for emotion classification. The best result was achieved for a Mel-spectrogram input – an accuracy of 37%, based on RAVDESS. For the MFCC feature map, an accuracy of 34% was reached. We also uses the well-known pretrained VGG16 convolutional model. Here the best result

was an accuracy of 43%, achieved on the Mel-spectrograms. The results of the classification based on the LSTM network have reached 34% only.

H. Comparison with other works

It should also be mentioned that for the RAVDESS set, the highest accuracy values achieved by different authors are 60% – 71%, but using much more complex models than ours [22], [23], [24], [25]. Some recent results, obtained on the RAVDESS dataset, are listed in Table V. Our best solutions — the single model "Layer 128-96" and the gender-aware configuration of three models – have outperformed other known solutions.

TABLE V
COMPARISON WITH OTHER WORKS EVALUATED ON THE RAVDESS DATASET

Model [ref]	Aver. Recall (%)	F1 (%)
CVT+SVM [22]	-	60.1
ResNet [23]	50.3	53.3
GResNet [23]	59.7	60.35
VGG16 [24]	71.0	-
Our "Layer 128-96"	72.0	71.86
Our "G-A 2x Layer 128-96"	72.2	73.2

V. CONCLUSIONS

In our research, we have confirmed a good performance of models processing a 1D feature vector. The approach to emotion classification based on 2D feature maps has failed. The use of a proper subset of speech features (without Chroma), a modification of the baseline network and the gender-aware approach have all contributed to a final result, that outperforms other known approaches validated on the RAVDESS dataset. The aim of our future research will be to explore more deeply prosodic features in emotion recognition.

REFERENCES

[1] M. Lech, M. Stolar, C. Best and R. Bolia, "Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding," *Frontiers in Computer Science*, vol. 2, 2020, article 14, https://doi.org/10.3389/fcomp.2020.00014.

[2] E. Andre, M. Rehm, W. Minker and D. Buthler, "Endowing spoken language dialogue systems with emotional intelligence," in *Affective Dialogue Systems. ADS 2004,* Lecture Notes in Computer Science, vol. 3068, 2004, pp. 178–187, Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-24842-2_17.

[3] C. Guo, K. Zhang, J. Chen, R. Xu and L.Gao, "Design and application of facial expression analysis system in empathy ability of children with autism spectrum disorder", *Proceedings of the 16th Conference on Computer Science and Intelligence Systems*, Annals of Computer Science and Information Systems, vol. 25, 2021, pp. 319–325, http://dx.doi.org/10.15439/2021F91.

[4] S. Gu, F. Wang, N. P. Patel, J. A. Bourgeois and J. H. Huang, „A Model for Basic Emotions Using Observations of Behavior in Drosophila," *Frontiers in Psychology*, vol. 10, 2019, article 781, https://doir.org/10.3389/fpsyg.2019.00781.

[5] J. A. Russel, "Emotions are not modules," *Canadian Journal of Philosophy*, vol. 36, 2006, sup1, pp. 53–71, Routledge Publ. https://doi.org/10.1353/cjp.2007.0037.

[6] E.Y. Bann, "Discovering Basic Emotion Sets via Semantic Clustering on a Twitter Corpus," arXiv:1212.6527 [cs.AI], December 2012, https://doi.org/10.48550/arXiv.1212.6527.

[7] S. Lugovic, I. Dunder and M. Horvat, „Techniques and Applications of Emotion Recognition in Speech," *in 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* 2016, pp. 1278–1283, https://doi.org/10.1109/MIPRO.2016.7522336.

[8] U. Kamath, J. Liu and J. Whitaker, *Deep Learning for NLP and Speech Recognition,* Springer Nature Switzerland AG, Cham, 2019. https://doi.org/10.1007/978-3-030-14596-5.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, 2011, no. 4, pp. 788–798. http://dx.doi.org/10.1109/TASL.2010.2064307.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2018, pp. 5329–5333, http://dx.doi.org/10.1109/ICASSP.2018.8461375.

[11] B. J. Abbaschian, D. Sierra-Sosa and A. Elmaghraby, „Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models," *Sensors*, mdpi, 2021, 21(4), https://doi.org/10.3390/s21041249.

[12] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *IEEE International Conference on Multimedia and Expo,* 2005, https://doi.org/10.1109/icme.2005.1521560.

[13] J. Rong, G. Li and Y.-P. P. Chen, „Acoustic feature selection for automatic emotion recognition from speech," *Information Processing and Management,* Elsevier, vol. 45, 2009, issue 3, pp. 315–328, https://doi.org/10.1016/j.ipm.2008.09.003.

[14] M. B. Akcay and K. Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication,* Elsevier, vol. 116, January 2020, pp. 56–76, https://doi.org/10.1016/j.specom.2019.12.001.

[15] M. B. Er and I. B. Aydilek, "Music Emotion Recognition by Using Chroma Spectrogramand Deep Visual Features," *International Journal of Computational Intelligence Systems,* vol. 12, Issue 2, 2019, pp. 1622–1634, https://doi.org/163410.2991/ijcis.d.191216.001.

[16] S. R. Livingstone and F. A. Russo, "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, https://doi.org/10.1371/journal.pone.0196391.

[17] K. Dupuis and K. M. Pichora-Fuller, *Toronto emotional speech set (TESS),* University of Toronto, Psychology Department, Borealis data publisher, 2010, https://doi.org/10.5683/SP2/E8H2MF.

[18] S. Haq and P.J.B. Jackson, "Multimodal Emotion Recognition", in *W. Wang (ed), Machine Audition: Principles, Algorithms and Systems,* IGI Global Press, 2011, chapter 17, pp. 398–423, https://doi.org/10.4018/978-1-61520-919-4.

[19] LibRosa documentation, https://librosa.org/doc/.

[20] S. Burnval, *Speech emotion recognition*, (https://www.kaggle.com/shivamburnwal/speech-emotion-recognition)

[21] M. A. Kutlugün, Y. Sirin and M. A. Karakaya, "The Effects of Augmented Training Dataset on Performance of Convolutional Neural Networks in Face Recognition System", *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*, Annals of Computer Science and Information Systems, vol. 18, 2019, pp. 929–932, http://dx.doi.org/10.15439/2019F181.

[22] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition", in *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS),* Dec. 2016, IEEE, pp. 1–8. https://doi.org/10.1109/ICSPCS.2016.7843306.

[23] Y. Zeng, H. Mao, D. Peng and Z. Yi, "Spectrogram based multi-task audio classification", *Multimedia Tools and Applications,* vol. 78 (2019), no. 3, pp. 3705–3722, https://doi.org/10.1007/s11042-017-5539-3.

[24] A. S. Popova, A. Rassadin and A. Ponomarenko, "Emotion Recognition in Sound", *International Conference on Neuroinformatics,* vol. 736, 2018, pp. 117–124, 2018, http://dx.doi.org/10.1007/978-3-319-66604-4_18.

[25] D. Issa, F. M. Demirci and A. Yazici, "Speech emotion recognition with deep convolutional neural networks", *Biomedical Signal Processing and Control,* Elsevier, vol. 59, 2020, 101894, doi:10.1016/j.bspc.2020.101894.