**Using voice biometrics to authenticate and identify the speaker under the influence of emotions.**

**Wiktoria Brzęczek eng.**

**ABSTRACT**

**Artificial Intelligence project for speaker recognition was implemented in order to identify the speaker using emotionally colored statements. The project was created using the Python programming language in Machine Learning and Neural Networks with TensorFlow and Keras framework. The Speech Emotion Recognition tool developed as part of the work consists of two programs based on different technologies. The Human Speech Emotion Recognition Program was created in accordance with the neural network methodology. The Speaker Recognition Program, on the other hand, is a machine learning classifier using a mixture model of Gaussian distributions. Finally, the implemented tool achieved the accuracy of recognizing speakers under the influence of emotions at the level of 79.34%. Three main human emotions were observed, i.e. sadness, joy and anger. Additional unit samples were neutral standards. To train and test the tool, emotions expressed by actors were used primarily in the CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) database. The impostor database consisted of a combination of audio clips from the JL Corpus, EMOVO, and ESD databases. The hypothesis of the work was to state that a classifier trained on a sample of a given emotion will achieve better speaker recognition results when the test sample comes from the same category. Based on the performed research, the hypothesis was confirmed. The tool shows lower percentage compatibility when samples are exposed to a classifier trained on a different emotion or a mixed classifier. In all cases, the best-performing classifiers were the appropriate classifiers for the emotion. The second place was always occupied by mixed classifiers, whose recognition percentage value was comparable to the classifiers of a given emotional tone. The least effective were the classifiers of emotions opposite to the analyzed audio sample. The worst recognition result in the entire testing phase of the tool was achieved by the anger classifier for sadness samples with a score of 13.27%. On the other hand, the neutral classifier analyzing samples of neutral emotions was the most effective (95.60%).**

1. **INTRODUCTION**

Voice biometrics technology is critical to speaker verification and identification where it is widely used. Biometric voice recognition uses the human voice to uniquely identify biological characteristics to authenticate a person. This is significantly different from passwords, which require physical entry. With advancing technology, many environments have abandoned orthodox authentication methods in favor of the most modern biometric solutions and their components. Fingerprint, face and palm recognition remain the most common forms, while biometric voice recognition has seen a significant increase. Security systems such as speaker identification and authentication have become increasingly popular in recent decades. They are widely used in effective security management by reducing the need for human resources. Access control systems are implemented in most of them. In particular, many research efforts have focused on the problem of speaker recognition. The assessment of the people involved depends to a large extent on many strategic aspects. Human voice analysis is a simple way to verify identity. In fact, voice recognition systems represent biometric systems that enable quick and non-intrusive access control, limiting collaboration with people to a minimum. Huge progress in the field of neural networks in recent years has enabled the development of faster and more accurate voice biometric algorithms, and at the same time increased the demand for modern solutions needed in many progressive market technologies.

The purpose of project was to implement and test an artificial intelligence model for speaker recognition. It makes it possible to identify the speaker using emotionally colored statements. The motivation to develop the subject of the work was the huge development of neural networks in recent years, which increased the demand for modern solutions as well. The constantly developing area of voice biometrics also leaves a lot of space for action and development. Adapted to the requirement of constant verification of the speaker due to the changing environment or emotional state of the speaker, it encourages the definition of new solutions and the creation of innovative tools. The hypothesis of the work was to state that a classifier trained on a sample of a given emotion will achieve better speaker recognition results when the test sample comes from the same category. The program, analyzing the received voice sample, assigns the predicted emotion. Then the sample is interpreted by the appropriate speaker recognition classifier. The tool is designed to show a lower percentage accordance when samples are subjected to a different emotion classifier or a mixed classifier.

1. **TOOL CONCEPT**

The project presents an approach to speaker recognition under the influence of emotions based on machine learning classifiers and neural networks. The tool was entirely created using the Python programming language. First of all, the CREMA-D database set was used in 4 output classes of emotions: neutrality, anger, sadness and joy. The tool consists of two interconnected programs using different concepts of methodology. The first stage is speech emotion recognition using a recursive neural network (RNN) classifier operating on the LSTM layer. The second program was speaker identification, using machine learning and mixed Gaussian models. The frameworks used to create the tool were TensorFlow, Keras, Sklearn and Librosa.

1. **DATA PREPROCESSING**

Obraz zawierający tekst, diagram, Plan, zrzut ekranu

Opis wygenerowany automatycznie

Figure 1 Preprocessing for a single audio file; Source: Own elaboration

The following data is extracted from each sample in the database:

1. Emotion representation obtained by parsing the name of the audio file. They contain the beginnings of the English name, which are crucial for every emotion. Accordingly, for the emotion of sadness it is "SAD", for anger - "ANG", happiness written as "HAP", while neutral samples can be recognized by the presence of "NEU". The keyword is placed in the third segment of the file name,
2. Sample rate, which provides information about the number of audio samples per second. The databases were recorded at a frequency of 22.5 kHz.
3. Audio processing is done in the following order:
   1. Audio Segment responsible for loading an audio file using the AudioSegment library,
   2. Normalization meaning each loaded object is normalized to +5dBFS,
   3. Converting the object to an array of samples is critical for the rest of the preprocessing. The operation allows to unify the analyzed samples and the functions that use them in further operation,
   4. Cleaning the audio file from silence at the beginning and at the end allows to get rid of excessive data,
   5. Filling each audio file to equalize the length of the analyzed data,
   6. Noise reduction allowing for more effective performance of the further operation of the tool.
4. **COLLECTION OF DATASET**

The review and study of the literature on the subject of the work was the first step in creating the tool. During this stage, many study aids related to the field were studied. Using a properly completed database is crucial to achieve the best system accuracy. Using the wrong databases leads to incorrect conclusions. The database used in this work was created by combining four databases: CREMA-D, ESD, JL Corpus, EMOVO. The CREMA database was used the most in the project. That dataset consists of 7442 original audio clips from 91 actors [1]. Among them are 48 men and 43 women aged 20 to 74, representing various races and ethnicities (African-American, Asian, Caucasian, Hispanic and Unspecified). The actors presented 12 sentences from six different emotions (Anger, Disgust, Fear, Joy, Neutral and Sad) and four levels of emotion (Low, Medium, High and Indeterminate).

1. **TOOL IMPLEMENTATION**

5.1 - Speech Emotion Recognition Program

The program aims to explore and implement an artificial intelligence (AI) model that will analyze input audio files, identify and present the emotions expressed in them. The implemented classification model was developed in the RNN method, i.e. Recurrent Neural Network (RNN) [2]. The recursive network was chosen because of its precise predictive results and favorable performance in terms of model training time and its stability in various types of environments. The chosen advanced model for time series analysis was Long Short-Term Memory (LSTM), which performs very well in continuous speech recognition with a large vocabulary due to its impressive learning capacity. The accuracy of speaker recognition using the proposed model also ensures low losses in the identification process by ignoring unnecessary data in the training phase. Emotions expressed by actors from the databases described above were used to train the model. Selected functions extracted using the Librosa library for the speech emotion recognition model are: Root Mean Square [3], Zero Crossed Rate [4], MFCCs [5].

5.2 Model Definition and Training

The model is executed with the Keras library, using 2 hidden LSTM layers with 16 nodes and an output (dense) layer with 4 nodes each for one emotion using "Softmax" activation function [6]. The optimizer that gave the best results was "RMSProp" with default parameters [7]. There are 339 values defined in the input for each feature. The database was divided into the necessary sets using the train\_test\_split function from the sklearn library. The distribution of samples between the training and testing samples was 0.225. The validation samples were then extracted from the test samples, leaving 0.304 of the initial value in the second set.

The numeric values of the respective pre-database splits are as follows:

1. Samples used to train the model: 3787,
2. Samples used for initial model testing: 435,
3. Validation samples: 874.

The selected Batch size was set to 1, which is the greatest common divisor for all samples in the sets. The training of the model was divided into five stages depending on the number of epochs. The results of the accuracy of predicted emotions for the test and validation sets for 100 epochs have been presented on Figure 2. There is a clear improvement in percentage accuracy with increasing number of training epochs. The smallest difference in the results was obtained between 80 and 100 epochs. It was decided that the final model used for the experiments would be a one hundred epoch-trained model. The accuracy of the model validation set reached 79.34% and the accuracy of the test set 77.92%. The training uses a model checkpoint that saves the best weights according to the accuracy of the model, thus avoiding overfitting.

Obraz zawierający stół

Opis wygenerowany automatycznie

Figure 2 Accuracy of predicted emotions for 100 training epochs; Source: Own elaboration

5.3 - Speaker Recognition Program

The voice recognition program is mainly divided into two parts: speaker verification and speaker identification. Speaker identification determines which registered speaker is making a given speech out of a set of known speakers. Speaker verification accepts or rejects the speaker's declaration of identity.

5.4 - Feature extraction

The purpose of the implemented tool is to correctly identify the speaker using the Gaussian mixture model. The first step in working with an audio sample is to extract features from it, i.e. to identify the components of the audio signal. The cepstral frequency coefficient Mel (MFCC) was used to isolate the features. Selected functions extracted using the Librosa library for the speech emotion recognition model are: Spectrogram, Mel-Spectrogram [8], Chroma features [9]. The MFCC coefficient with a tuned parameter as the basic feature and delta MFCC, also known as differential and acceleration coefficients, are considered, which are used to deal with speech information that is related to the dynamics and the calculation of the MFCC trajectory over time.

5.5 Model Definition and Training

The Gaussian Mixture Model (GMM) is one of the most popular models used to train audio data [10]. The use of the combined features of MFCC and GMM leads to the achievement of the goal of correct speaker identification. GMM is used to train the model on previously extracted functions (MFCC and Delta).

5.6 Model Testing

GMM models will be used to calculate feature scores for all models. The speaker model with the maximum score is predicted as the identified test speech speaker.

1. **MODEL EVALUATION**

The model was evaluated using the following factors:

1. Visualization of loss value trend and categorical accuracy during the training proces [11].

Figure 3 is an example of a proper fit of the model. The values of validation and training losses are low, which significantly reduced the number of errors affecting the effectiveness of the system, which meant permission to stop further training. The categorical accuracy trend for one hundred epochs of training is shown in Figure 4. In this case, the requirements met by the model are confirmed. The continuous learning and improvement of the model over time can be noticed on the plot.

Obraz zawierający tekst, linia, Wykres, diagram

Opis wygenerowany automatycznie

Figure 3 Loss trend for 100 model training epochs; Source: Own elaboration

Obraz zawierający tekst, zrzut ekranu, Wykres, diagram

Opis wygenerowany automatycznie

Figure 4 Categorical accuracy trend for 100 model training epochs; Source: Own elaboration

2- Confusion Matrix to visualize the number of successful predictions of each emotion: for validation and test sets [12].

Confusion matrices show exactly which samples have been misclassified. For the validation set (Figure 5) the largest false classification turned out to be anger samples considered as neutral samples. The best results, which means the least misclassification, go to the opposite emotion samples. The least false matches were given to sadness samples for joy samples and anger samples Relatively low fit was achieved by neutral samples matched with happiness samples and anger samples matched with sadness samples. Worth noting is that the results obtained for the test set matrix (Figure 6) are very similar to the results of the validation set. The largest number of correctly identified samples is once again attributed to the emotion of joy. There are no falsely recognized sadness samples for the anger and joy samples for both stages of model training. The highest false classification was found in neutral samples considered as sad samples and joy samples matched with anger samples.

Obraz zawierający zrzut ekranu, Prostokąt, kwadrat, diagram

Opis wygenerowany automatycznie

Figure 5 Validation set confusion matrix for 100 model training epochs; Source: Own elaboration

*Obraz zawierający kwadrat, Prostokąt, zrzut ekranu, wzór

Opis wygenerowany automatycznie*

Figure 6 Test set confusion matrix for 100 model training epochs; Source: Own elaboration

3- Model prediction accuracy rates for each emotion: for validation and test sets [13] which was described in paragraph 5.2.

1. **CALCULATING ERR**

7.1 Defining the FAR threshold

In order to ensure the security of the biometric system, it is important to keep the FAR low [14]. Organizations should regularly test their biometric systems to ensure that they perform as expected and that the FAR is within acceptable limits. In addition, organizations should take steps to improve the security of their biometric systems, such as using strong encryption and access controls.

The FAR calculation process was as follows:

1. Loading the database of impostors, which consists of 30 people who are not in the basic database. Each single person base consists of the same number of samples as in the base dataset,
2. Performing normalization during pre-processing and extracting features,
3. Loading the model,
4. The model is tested against the impostor's dataset. It is known that the dataset does not contain the original models, so all predictions that will be made are wrong. The tests aim to get a prediction of how many cheaters exceed a given threshold,
5. The threshold has been set from 0 to 100,
6. A loop has been created where predictions are made for all impostor samples,
7. In the next loop, for the defined threshold, the number of impostors exceeding each threshold value was determined.

The FAR curve represents the case where a fraudster can be identified as genuine and gets into the system. At a threshold of 0, all cheaters will be qualified, and by increasing the threshold, they are less likely to be accepted correctly. The false acceptance rate for the model is shown in Figure 7.

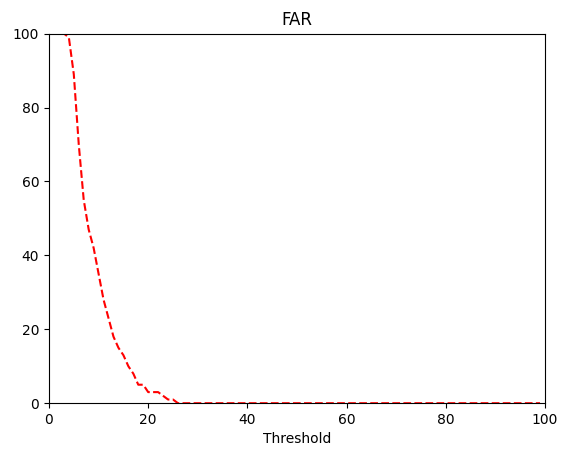
****

Figure 7 False acceptance rate for the model; Source: Own elaboration

* 1. Defining the FRR threshold

False rejection rate is the percentage of valid users who are incorrectly rejected by the system. FRR is an important metric for evaluating the performance of an authentication system as it measures the system's ability to correctly identify legitimate users [15]. It is therefore a measure of the usability of the implemented system.

The process of calculating the FRR value was as follows:

1. Loading the test database, which is the original set of 30 data models in the basic database,
2. Performing normalization during pre-processing and extracting features,
3. Loading the model,
4. The model is tested on a set of test data. The tests aim to obtain a prediction that the model correctly predicts the input samples,
5. The threshold has been set from 0 to 100,
6. A loop has been created in which predictions are made for all test samples,
7. In the next loop, for the defined threshold, the number of test samples rejected for each threshold value was determined.

The FRR plot represents the case where the original sample is rejected and by increasing the threshold, the original chance of rejection increases as well. At a threshold of 0, the original will exceed 100%. The false rejection rate for the model is shown in Figure 8.

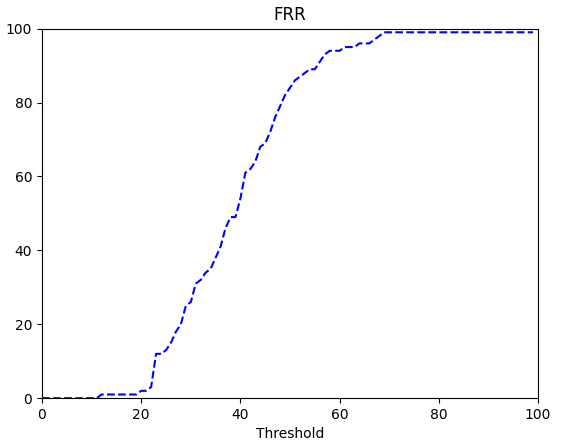


Figure 8 False rejection rate for the model; Source: Own elaboration

7.3 Defining the FRR threshold

The Equal Error Rate is the point where the False Rejection Rate (FRR) and the False Acceptance Rate (FAR) are equal. EER is a used measure of accuracy in biometric systems because it provides a single value representing the overall accuracy of the system [16]. The EER point represents the best threshold for choosing to minimize both error rates simultaneously, as it is the point where the FAR and FRR meet. In the case of the developed model, the EER value is equal to 2.56% (Figure 9).

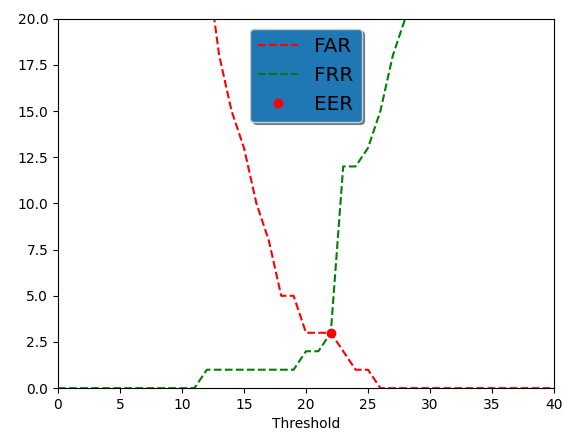
****

Figure 9 Equal error rate for the model; Source: Own elabaration

1. **TOOL ARCHITECTURE**

**Obraz zawierający tekst, diagram, Plan, Rysunek techniczny

Opis wygenerowany automatycznie**

Figure 70 Speech Emotion Recognition tool architecture; Source: Own elaboration

The architecture of the Speech Emotion Recognition tool developed as part of the work (Figure 10) is a graphical representation of the combination of two programs to create a speech recognition system. The operation of the system begins with the receipt of a test sample at the input of the program. The sample then goes through the steps described in the previous section. After initial processing and feature extraction, it goes to the emotion classifier, where it is analyzed in terms of four emotions. The classifier calls a previously trained neural network model with the LSTM layer and provides one of five possible answers - Neutral, Angry, Sad, Happy, Unknown. Based on the received qualification, the sample is transferred to the appropriate machine learning classifier based on the Gaussian mixture. The sample then undergoes feature extractions to be compared in a loop with the saved Gaussian models from the database of the corresponding emotion. The function uses the method of greatest similarity, thanks to which the system returns the answer in the form of a matched sample of the speaker from the database.

1. **TESTING THE EFFECTIVENESS OF THE TOOL BASED ON CONDUCTED EXPERIMENTS AND EVALUATIONS**

The purpose of testing is to compare the speaker models stored in the database with the objects in the set, which are test instances. Appropriate emotion classifiers analyzing input samples will be used for this task. The experimental testing phases were divided into two basic stages: testing the type of input samples and testing the type of classifier. The first is to pass audio samples to each of the classifiers in order to observe the speaker recognition percentage. The second phase is to check the efficiency of classifiers by introducing appropriate sequences of emotion samples. Testing begins with the initial phase of entering audio files into the corresponding classifiers.

9.1 Initial testing phase

Figure 11 shows the basic phase of system testing. It consists in extracting test samples from the initial database and entering them into the appropriate emotion classifier. You are expected to receive a percentage of correctly identified speakers. The action is performed for each of the four emotions. Tests were performed for each of the 14 emotion samples for 91 people from the base of the model training base.

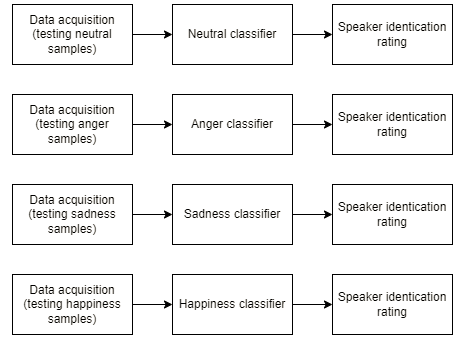


Figure 11 Basic phase of testing the implemented model; Source: Own elaboration

Average results of correct recognition:

* Anger samples in the anger classifier: 88.15%,
* Sadness samples in the sadness classifier: 88.30%,
* Happiness samples in the joy classifier: 87.28%,
* Samples neutral in the neutral classifier: 92.39%.

Figure 12 shows the correctly recognized speakers for the samples in the corresponding emotion classifiers.

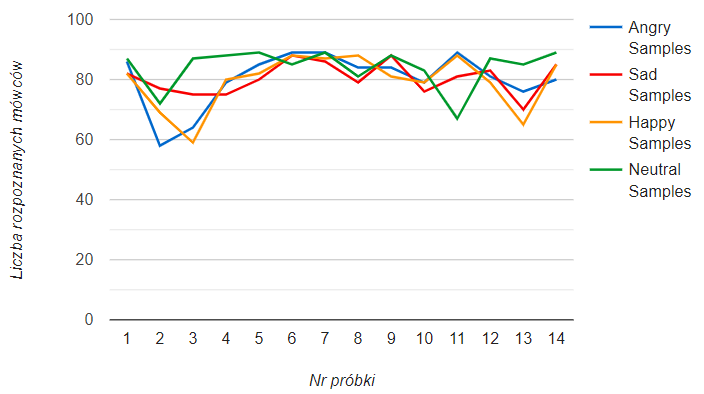


Figure 82 Graphical representation of correctly recognized speakers for each sample in the appropriate classifiers

9.2 Testing the type of input samples

The next stage consists in introducing successive audio samples to the classifiers of another emotion. In this way, the speaker recognition result is expected, which will allow to analyze the effectiveness of the relevant classifiers. Figure 13 shows an example diagram of how to perform testing for neutral samples.

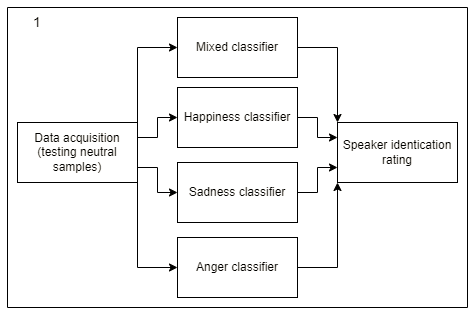


Figure 13 Experimental phase for Neutral emotions

9.3 Testing the type of classifiers

The stage consists in entering samples of each emotion into one of the classifiers. In this way, the result of speaker recognition is expected, which will allow to analyze which sample the properly tested classifiers perform with best. Figure 14 shows the percentage results of recognition of all samples of given emotions in the anger classifier.

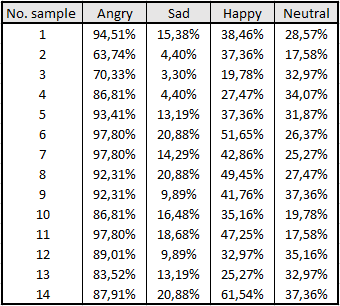


Figure 14 Percentage compability on speaker recognition for all samples in the anger classifier

Average results of correct recognition for the anger classifier:

* Anger samples: 88.15%,
* Sadness samples: 13.27%,
* Happiness samples: 39.17%,
* Neutral samples: 28.89%.

1. **SUMMARY AND CONCLUSIONS**

The Speech Emotion Recognition tool developed as part of the project consists of two programs based on different technologies. The human speech emotion recognition program was created in accordance with the neural network methodology. The speaker recognition program, on the other hand, is a machine learning classifier using a mixture model of Gaussian distributions. Finally, the implemented tool achieved the accuracy of recognizing speakers under the influence of emotions at the level of 79.34%. Three main human emotions, i.e. sadness, joy and anger, were observed. Additional unit samples were neutral standards. To train and test the tool, emotions expressed by actors were used primarily in the CREMA-D (Crowd-sourcedEmotionalMultimodalActorsDataset) database. The impostor database consisted of a combination of audio clips from the JL Corpus, EMOVO, and ESD databases.

Based on the performed research, the hypothesis was confirmed. The tool shows lower percentage agreement when samples are exposed to a classifier trained on a different emotion or a mixed classifier. In all cases, the best-performing classifiers were the appropriate classifiers for the emotion. The second place was always occupied by mixed classifiers, whose recognition percentage value was comparable to the classifiers of a given emotional tone. The least effective were the classifiers of emotions opposite to the analyzed audio sample. The worst recognition result in the entire testing phase of the tool was achieved by the anger classifier for sadness samples with a score of 13.27%. On the other hand, the neutral classifier analyzing samples of neutral emotions was the most effective (95.60%).

Taking into account the analysis of the basic phase of system testing, in which samples of emotions are tested and fed to their appropriate classifiers, it is possible to determine the coefficient of correct recognition of speakers. The issue of recognizing individuals by analyzing all samples of a given emotion in the appropriate classifiers returns the highest efficiency of the system for neutral samples. The results report the best recognition efficiency for neutral samples in the neutral classifier (92.39%), while the worst results belong to joy samples in the joy classifier (87.28%).

Analyzing the results, we are able to observe a similar effect of the mixed classifier for all emotion samples (Figure 15). For each type of emotion, it ranks second in recognition efficiency, right after the classifier appropriate for a given sample. That is the result of including all the speaker samples for the three extracted features. This gives a much higher probability of recognizing the speaker than in the case of classifiers of a completely different emotion.

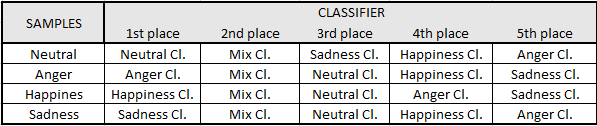


Figure 15 Sample recognition performance rating for all classifiers

There is a noticeable deterioration in the recognition efficiency for specific samples. For sample no. 13, all classifiers recorded a decrease in efficiency. Analyzing the file allowed the observation of a short and concise message, so the feature extraction functions were unable to collect the appropriate number of characteristic signatures for the speakers. The lower number of identifications in other single samples was dictated by the soft tone of the speaker (for the anger samples) and the fast pace of speaking (for the sadness samples).

The operation of classifiers is also conditioned by the characteristics of given emotions. The samples are recognized with the least efficiency in the opposing emotion classifiers (Figure 16). In both joy and anger classifiers, the samples of sadness emotions are characterized by the worst recognition efficiency. It depends on the characteristics of the analyzed emotion. A large amount of energy and a wide range of tones when expressing anger mean that the sadness classifier is not able to effectively identify the user. Moreover, the negative valence level, low level of energy activation and slow rate of expression of sadness emotions create problems in the recognition of samples in the anger classifier.

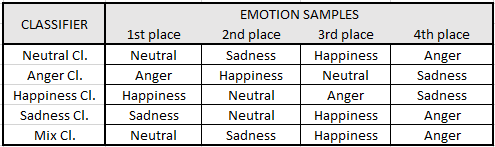


Figure 16 Classifier performance rating for all emotion samples

1. **FUTURE SCOPE**

The future expansion of the project is based primarily on the development of detecting the exact sampling frequency for each speaker. The condition can be met by using additional features such as speech rate, quality and stress. The rate of speech provides information about the imperfections of the model. It could also be a good practice to remove random silence from audio samples that appear during the speaker's spoken message. It is possible to add data effort to the input audio samples using the augmentation technique. The most well-known method is to modify the speed of the sound (acceleration/deceleration), as well as moving audio samples in time or searching for clips with more annotations. It is desirable to study new acoustic features and check their usefulness in the field of speech recognition.

The program is mainly based on MFCC functions, which could be extended with RAS-MFCC [17]. It is also possible to use different approaches, which are also intended for the research subject, such as Cepstrum harmonic [18], PLP [19] and LPCC [20]. An additional analysis is also following the approach based on the lexical features of speech emotion recognition using a set of lexical and acoustic models. This will improve the accuracy of the system as in some cases expressions of emotion are contextual rather than vocal.

It could also be effective to conduct additional tests based on the method of listing people. The program would return the specified number of people identified as the input sample. Rating more answers of the tool would allow to determine the position of the real speaker in the returned list. The analysis of the experiment would characterize more thoroughly the effectiveness of the implemented system. Verification of the speaker recognition by one of the three (Mel Spektrogram, Chroma, MFCC) features of the program would also be an appropriate assessment of the tool and its effectiveness. This would boil down to re-analysing the characteristics of each feature and determining the most optimal one for the speaker recognition program under the influence of emotions. During the evaluation of the system, there were also samples that achieved worse recognition in all classifiers. It would be a good practice to interpret the audio file and its metadata to find out what causes the lower percentage match.

**BIBLIOGRAPHY**

[1] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R.   
 (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE*  *transactions on affective computing*, *5*(4), 377-390.

[2] Grossberg, S. (2013). Recurrent neural networks. *Scholarpedia*, *8*(2), 1888.

[3] Yarlanki, S., Rajendran, B., & Hamann, H. (2012, May). Estimation of turbulence   
 closure coefficients for data centers using machine learning algorithms. In *13th*   *InterSociety Conference on Thermal and Thermomechanical Phenomena in*   *Electronic Systems* (pp. 38-42). IEEE.

[4] Salman, S., Mir, J., Farooq, M. T., Malik, A. N., & Haleemdeen, R. (2021, January).   
 Machine learning inspired efficient audio drone detection using acoustic features.   
 In *2021 International Bhurban Conference on Applied Sciences and Technologies*  *(IBCAST)* (pp. 335-339). IEEE.

[5] Mahboob, T., Khanum, M., Khiyal, M. S. H., & Bibi, R. (2015). Speaker   
 identification using gmm with mfcc. *International Journal of Computer Science*  *Issues (IJCSI)*, *12*(2), 126.

[6] Banerjee, K., Gupta, R. R., Vyas, K., & Mishra, B. (2020). Exploring alternatives to   
 softmax function. *arXiv preprint arXiv:2011.11538*.

[7] Shi, N., & Li, D. (2021, April). Rmsprop converges with proper hyperparameter.   
 In *International conference on learning representation*.

[8] Hwang, Y., Cho, H., Yang, H., Won, D. O., Oh, I., & Lee, S. W. (2020). Mel-spectrogram   
 augmentation for sequence to sequence voice conversion. *arXiv preprint arXiv:2001.01401*.

[9] Garg, U., Agarwal, S., Gupta, S., Dutt, R., & Singh, D. (2020, September). Prediction   
 of emotions from the audio speech signals using MFCC, MEL and Chroma. In *2020*  *12th International Conference on Computational Intelligence and Communication*  *Networks (CICN)* (pp. 87-91). IEEE.

[10] Chen, W., Hong, Q., & Li, X. (2012, July). GMM-UBM for text-dependent speaker   
 recognition. In *2012 International Conference on Audio, Language and Image*  *Processing* (pp. 432-435). IEEE.

[11] Kim, H., Rossi, R., Sarma, A., Moritz, D., & Hullman, J. (2021). An automated   
 approach to reasoning about task-oriented insights in responsive visualization. *IEEE*   *transactions on visualization and computer graphics*, *28*(1), 129-139.

[12] Ahmad, S., Ansari, S. U., Haider, U., Javed, K., Rahman, J. U., & Anwar, S. (2022).   
 Confusion matrix-based modularity induction into pretrained CNN. *Multimedia*  *Tools and Applications*, *81*(16), 23311-23337.

[13] Koh, H. C. (1991). Model predictions and auditor assessments of going concern   
 status. *Accounting and Business Research*, *21*(84), 331-338.

[14] Kyoso, M. (2003, October). A technique for avoiding false acceptance in ECG   
 identification. In *IEEE EMBS Asian-Pacific Conference on Biomedical*  *Engineering, 2003.* (pp. 190-191). IEEE.

[15] Vatsa, M., Singh, R., & Noore, A. (2008). Reducing the false rejection rate of iris   
 recognition using textural and topological features. *International Journal of*   *Computer and Information Engineering*, *2*(7), 2533-2539

[16] Kaye, D. H. (2002). The error of equal error rates. *Law, Probability and Risk*, *1*(1),   
 3-8..

[17] Dev, A., & Bansal, P. (2010). Robust features for noisy speech recognition using mfcc   
 computation from magnitude spectrum of higher order autocorrelation   
 coefficients. *International Journal of Computer Applications*, *10*(8), 36-38.

[18] Niebudek-Bogusz, E., Strumiłło, P., Wiktorowicz, J., & Śliwińska-Kowalska, M.   
 (2013). Comparison of cepstral coefficients to other voice evaluation parameters in   
 patients with occupational dysphonia. *Medycyna Pracy*, *64*(6).

[19] Alam, M. J., Kinnunen, T., Kenny, P., Ouellet, P., & O’Shaughnessy, D. (2013).   
 Multitaper MFCC and PLP features for speaker verification using i-vectors. *Speech*  *communication*, *55*(2), 237-251.

[20] Gupta, H., & Gupta, D. (2016, January). LPC and LPCC method of feature extraction   
 in Speech Recognition System. In *2016 6th international conference-cloud system*  *and big data engineering (confluence)* (pp. 498-502). IEEE.