Speech emotion recognition using machine learning

Ravnik Šimen

University of Ljubljana, Faculty of Computer and Information Science

Student at the Faculty of Computer and Information Science, University of Ljubljana
E-mail: sr8905@student.uni-lj.si

Abstract. In this research our goal was to predict human emotion based on speech records. We collected the speech records dataset from a well known RAVDESS database. From the voice samples we first extracted the features using three methods, which are Mel-Spectrogram, MFCC and Chroma. From the extracted features we constructed feature vectors upon which we trained our three models. Before that we split the dataset into train and test to properly evaluate our results. The models we chose for the classification were SVM, MLP ,logistic regression and Random forest. We first used 10-fold cross validation on each model to validate their accuracy. After that we also estimated the accuracy together with precision and recall on the test dataset. The best model for our problem turns out to be Random forest with 70.78% accuracy.

Ključne besede: Speech emotion recognition

1 Introduction

In the last couple of years, speech recognition has played a vital role in a variety of applications. Many large corporations are using speech recognition as a new way of using their products. For example Apple, Google, Tesla are using speech recognition as an alternative or in some cases the main way of interaction. With speech emotion application evolving, new branches in this field are also being discovered. Among those is also speech emotion recognition which allows us to recognise people's emotion based on their voice.

Speech emotion recognition is a field in machine learning that uses machine learning models to classify human emotions based on their voice. Emotions have many modalities which we can determine based on a human's face, physiological signals and speech. In contrast to other modalities, voice signals are considered to be more generalized and can be easily captured. [1] However, collecting voice signals can lead to large biases since everything is closely related to the speaker that we are recording. In this research we used records of actors that were trying to mimic the emotions using different speech techniques. Records were then collected into a single dataset, where each of the actors repeat the same sentence in different tones to imitate different emotions.

In this research our goal was to use machine learning models to classify the speaker’s emotion. To achieve this goal we used different approaches which will be presented in the continuation. But before that a lot of preprocessing had to be done in order to prepare voice signals into computer readable format. We can imagine that machine learning models expect numerical data as their input, whereas here we are dealing with voice signals. The core of our research is therefore feature extraction, meaning how to translate speech signals into model readable format.

After the preprocessing step, we used several different machine learning models to classify speakers. To solve this kind of problems, modern machine learning approaches include mainly neural networks, logistic regression and support vector machines algorithms. And those algorithms were also the core in our research.

2 Related work

There have been a lot of studies related to speech emotion recognition. In general we can seperate speech emotion recognition pipeline into 4 segments, which are speech analysis, feature extraction, modeling and testing. [2] The first segment speech analysis is used in the first step, where we want to extract the information that is connected to the speaker's behavior. These include the source of arousal, the vocal tract and behavior characteristics that can be viewed as the identity of the speaker. Next, the feature extraction segment, which is used to convert signal data into parametric representations, which can be used for classification and analysis. The modeling segment is used to construct a machine learning model that maps

the feature vector into the output of the model. And the last, testing segment is used to evaluate how well our model performs and what is its accuracy.

Different research teams tackle different segments for their improvements. The most common deviation is in the feature extraction segment. [3] We found three different research papers in the context of speech emotion recognition that were interesting for us. In the first paper we discovered that the research team used temporal representation as a way of presenting features of voice signal. In their studies they used the German speaking population and they are successfully recognising seven emotions including anger, anxiety, disgust, fear, pleasure, boredom and neutral. In the second paper a famous Ryerson Audio-Visual Database was used where english speaking records are collected and is extremely popular for speech emotion recognition. They used MFCC, which is the most popular and well known method for feature extraction in speech recognition algorithms, and combined that with convolution neural networks (CNN) and long short-term memory (LSTM), and achieved 80% accuracy. [6, 7, 8] In the third interesting article, the authors used statistical representations for classifying the emotions. [4] In their studies they use the Indonesian database (I-SpeED) to build models to predict speech emotion based on 3420 voice records. [5] The used features like average amplitude, frequency, volume and duration as a feature vector. As a classification algorithm they used support vector machines (SVM) and artificial neural network (ANN), which gave them accuracy 76% and 66 %, respectively.

The models that are most commonly used in speech emotion recognition are SVM, multilayer perceptron (MLP) and logistic regression. The most popular among those is SVM since it is often described as the most efficient when it comes to speech recognition. The second is MLP which in some cases performs even better in the newer studies. Deep neural networks are still being applied in this field and are already giving us even better results if the right parameters are used.

Different research teams also use different datasets, which can also lead to variety when it comes to picking the best model. Some use datasets from talk shows or movies, others are collecting voice records directly from people.[4] Therefore the unbalanced datasets are often produced, which can be solved with synthetic minority over-sampling technique (SMOTE). The pear to pear method also has the advantage that the emotions are more transparent.

3 Methodology

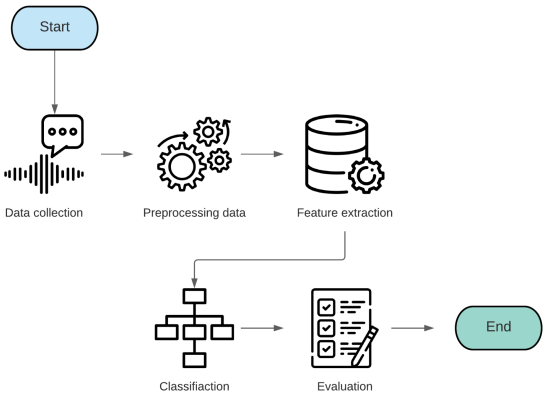

Figure 1. Speech emotion recognition pipeline.

According to the Figure 1, this research is focused on 5 different segments:

1. **Data collection segment** - collecting dataset upon which the algorithms will be performed and it should be prelabeled
2. **Data preprocessing segment** - preparing the data to be processed and the data quality analysis
3. **Feature extraction segment** - use Mel-frequency, MFCC and Croma to extract features from voice records
4. **Classification segment** - use different classification approaches from SVM, MLP to Logistic regression and Random forest
5. **Evaluation segment** - evaluate how well the models perform

3.1 Data collection segment

For our research we used the RAVDESS dataset [9]; this is the Ryerson Audio-Visual Database of Emotional Speech and Song dataset, and is free to download. The dataset contains 7356 files that were rated 247 people on emotional validity, intensity, and genuineness. The dataset contains speech records from 24 actors where each recorded 60 records which give us 1440 records in total. And there are 8 emotions that were captured; neutral, calm, happy, sad, angry, fearful, disgust, surprised. In our case for prediction we were focusing mostly on emotions calm, happy, sad, angry, since those are the most common emotions and probably the most important ones when it comes to recognising people's mood.

3.2 Data preprocessing segment

The whole RAVDESS dataset is of size 24.8GB. The dataset sample rate was lowered in order to reduce the size of the dataset. The final size was 173.8MB.

3.3 Feature extraction segment

As already mentioned, feature extraction play a vital role in speech emotion recognition. To convert signal data into model readable data we use various approaches. In our case, we used Mel-Spectrogram, MFCC and Chroma for the feature extraction step.

1. **Mel-Spectrogram** is a spectrogram, where the frequencies are converted into mel scale. The emotions over time represent the voice signal. The voice signal is transformed into frequency domain using fast Fourier Transform and then the shifted frequencies and the amplitude combined forms the spectrogram. [10]
2. **MFCC** is a method that is most commonly used in speech processing. It extracts the phonetical characteristics of speech. MFCC features represent phonemes (distinct units of sound) as the shape of the vocal tract (which is responsible for sound generation) is manifest in them. [11]
3. **Chroma** is being used for vocal content representation in audio files. As it traverses the helix, it also defines the angle of pitch rotation. [7]

We combined the extracted features from all three methods into one feature vector for each record. Next, we used different classification models upon which we used our feature vectors to train the models.

3.4 Classification segment

In this segment we used four different machine learning models. We used models that are most used in these files and those are SVM, MLP and logistic regression. Additionally we also used the Random forest prediction model since it is known to have great results in the big data field. In order to verify our results we splitted the dataset into training (80%) and test (20%) subsets. For an even better estimate of our models' performance we also used 10-fold cross validation. Of course the 10-fold cross validation was performed only on the train dataset where on each step we splitted the set into train and validation sets.

3.4.1 Support vector machine

Support vector machine is one of the supervised machine learning algorithms that is usually used to solve binary classification problems. The aim of the model is to create a hyperplane that classifies all vectors training vectors. At each step the algorithm calculates the maximum margin between the data point and the hyperplane which we call support vectors. The formula for hyperplane calculation can be seen in Equation 1 and the SVM example in Figure 2.

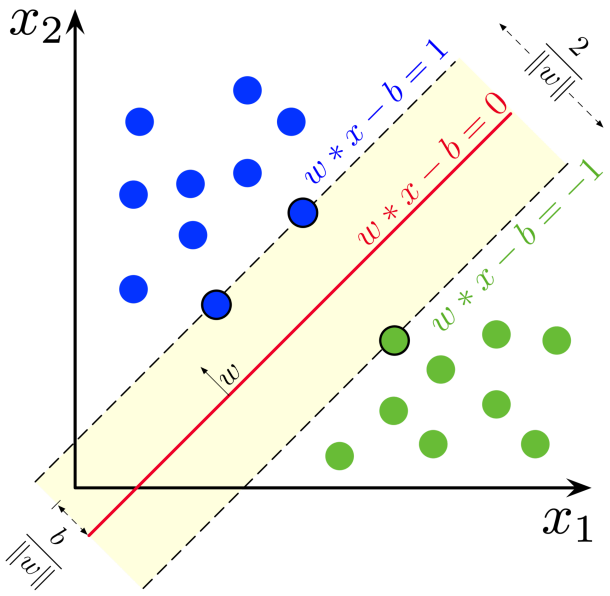

Figure 2. SVM visualization in two dimensional space. (source https://en.wikipedia.org/wiki/File:SVM_margin.png)

$$g(\vec{x}) = \vec{w}^T \vec{x} + \omega_0 \quad (1)$$

From the equation above, a hyperplane can be formed that classifies each data point into 2 classes. If g(x) > 1 data point belongs to class 1 and if g(x) < -1, data point belongs to class 2.

3.4.2 Multilayer perceptron

Multilayer perceptron (MLP) belongs to the feedforward artificial networks, and is constructed out of input and output layers that are connected to each other. It can also contain hidden layers which can be seen on Figure 3.

Each of the nodes of the neural network is constructed out of activation function (usually sigmoid function or ReLu), which determines whether a certain node has to be activated or not. The nodes are connected between each other with edges that contain weights. The

weights represent how important a certain node is for the classification. During the training step, the algorithm is correcting the weights to minimize the classification error. We call this process back propagation.

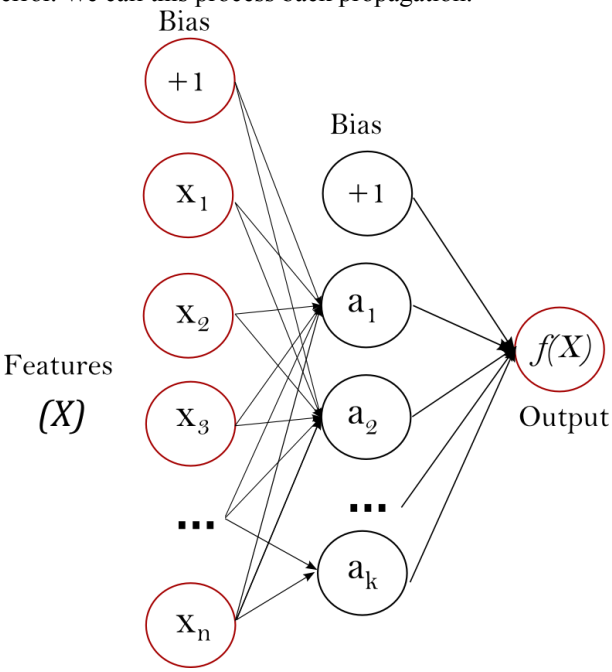

Figure 3. Multi-Layer Perceptron example. (source https://scikit-learn.org/stable/_images/multilayerperceptron_network.png)

Neural networks help us find hidden non linear patterns inside data. which are often missed in other models.

In our case, an MLPClassifier object from scikit learn library was used to perform classification. We used ReLu as an activation function and Adam as a classifier. The alpha was tested from 0.01 to 0.05 and the value 0.01 gave us the best results. We set the hidden layer size to 300. The other parameters used for classification are alfa: 0,01, batch_size: 256, activation: relu, solver: adam, hidden_layer_sizes: (300,), learning_rate: adaptive, max_iter: 500.

3.4.3 Logistic regression

Logistic regression is also a supervised machine learning algorithm that predicts the probability that a certain data point belongs to a certain class. Mostly is used for solving binary classification problems but in our case we can use one-vs-rest technique which allows us to solve multiclass classification problems.

Logistic regression uses the Sigmoid function for the classification which maps into values between zero and one. The Sigmoid function is written in Equation 2 and also displayed in Figure 4.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{2}$$

Theta in the above equation represents the vector of coefficients of length m (m is the number of characteristics or parameters), and e represents the Euler number. In the training step we are again trying to minimize the classification error. Therefore we must define the loss function, which will tell us how well the Sigmoid function fits the data. The loss in logistic regression is defined in Equation 3.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \, log \, h_\theta(x_i) \; + \; (1{-}y_i) \, log \, (1 - h_\theta(x_i))] \tag{3}$$

To optimize the classification accuracy, our goal is to minimize the loss function. For this process we use stochastic gradient descent where we are slowly (with learning rate alpha) moving toward the minimum of the function.

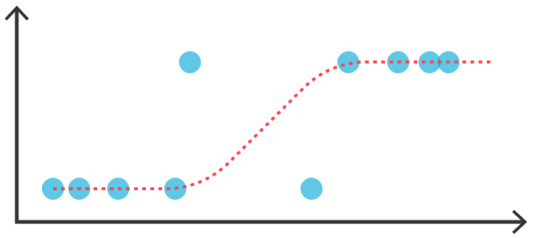

Figure 4. The sigmoid function used in logistic regression. (source https://www.tibco.com/sites/tibco/files/media_entity/2020-09/logistic-regression-diagram.svg)

3.4.4 Random forest

Random forest is one of the supervised machine learning algorithms that is used both for regression and classification problems. It is one of the most robust and relatively simple machine learning algorithms. It consists of several decision trees that are connected to the forest and are used to predict unseen cases. The idea of the algorithm is based on the fact that the single decision tree can be very unstable, meaning small changes in data can cause large differences in tree construction. Next, a single decision tree is often overfitted to the problem, which we are solving with pruning the tree branches. This step is not necessary in Random Forest since our final prediction is the most common prediction of all constructed trees.

To construct prediction trees in the random forest, we randomly select n subsets upon which we construct the trees. When predicting unseen cases, we predict the outcome with each of the constructed trees and take the majority vote for our final prediction. This gives us stability and robustness. The model is thus resilient to small changes of data and often do not have problems with overfitting.

For better illustration random forest visualization can be seen in figure 5.

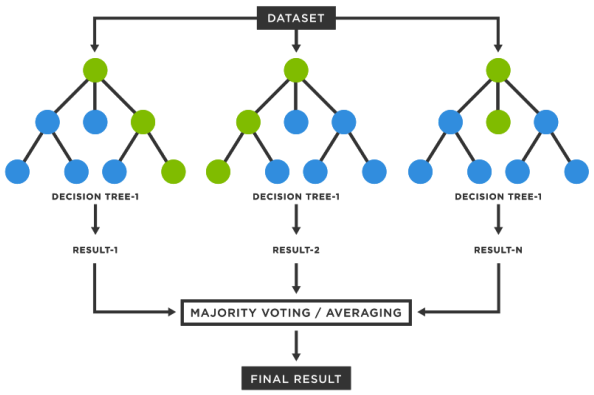

Figure 5. Random forest visualization. (source https://www.tibco.com/sites/tibco/files/media_entity/2021-05/random-forest-diagram.svg)

3.5 *Evaluation segment*

To evaluate our models for speech emotion recognition, 4 different measurements are needed, which are then constructed into a confusion matrix. Those values are TP, FN, FP, TP, from which accuracy, sensitivity, precision and F1-score are calculated. Those measurements allow us to transparently determine which or the model performs the best. Accuracy is the ratio between all positive predicted emotions and total predicted emotions. Precision is a measurement that calculates the amount of true predicted emotion divided by all that was predicted for that certain emotion. And sensitivity is a calculation that measures the amount of true predicted emotions divided by the total amount of specific emotion. We often combine precision and sensitivity into one measurement called F1-score, which is the harmonic mean of the two. It is calculated by Equation 4.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

4 RESULTS

For this research we used the four described models, which are SVM, MLP and logistic regression and Random forest. First we extracted the features from voice records using Mel-Spectrum, MFCC and Chroma, and split the constructed dataset into train and test sets. On the train set we used 10-fold cross validation to first evaluate the model and then used the test set to confirm our results.

4.1 *SVM model*

Table 1 is describing the results when using SVM model as classification model for predicting human emotions from speech. Our goal was to classify records into 4 classes, those are “angry”, “calm”, “happy” and “sad”. We first used 10-fold cross validation to evaluate the model using only train (and splitted validation) sets. The result of 10-fold cross validation was 66.7%. We then used the whole train dataset to construct the SVM model and predict the emotions on test data.

We can see that the model performed well. The final accuracy of the SVM model on test data is 69.48%, which is very good in the terms of speech recognition. We can also see the precision and recall scores and combine them into F1-score which turn out to be 69% for the SVM model. We can see the results in Table 2.

	precision	recall	f1-score
angry	0.77	0.79	0.78
calm	0.66	0.78	0.71
happy	0.69	0.63	0.66
sad	0.67	0.59	0.62

Table 1. Precision, recall and F1-scores for “angry”, “calm”, “happy” and “sad” using SVM.

	accuracy	f1-score
SVM	69.48%	69%

Table 2. Final accuracy and F1-score using SVM.

4.2 *MLP model*

Table 3 is describing the results we got using the MLP model. Again our goal was to predict the emotions “angry”, “calm”, “happy” and “sad”. We used the 10-fold cross validation to first evaluate the model. The result of 10-fold cross validation was 62.1% which is lower than the accuracy of the SVM model.

Then we trained the model on the whole train dataset and evaluated the results on the test dataset. The final accuracy was 64.85% which is again lower than the SVM model, and we can see that the MLP performed worse than SVM. From the precision and recall in Table 3 we can also see that those values are lower than in the previous table. The F1-Score is this time 65% which can be seen in Table 4.

	precision	recall	f1-score
angry	0.71	0.71	0.71
calm	0.73	0.47	0.58
happy	0.67	0.69	0.68
sad	0.56	0.73	0.63

Table 3. Precision, recall and F1-scores for “angry”, “calm”, “happy” and “sad” using Multi-Layer Perceptron.

	accuracy	f1-score
MLP	64.85%	65%

Table 4. Final accuracy and F1-score using Multi-Layer Perceptron.

4.3 Logistic regression model

The third model to evaluate was logistic regression. We performed 10-fold cross validation and we were predicting the same emotions as in the previous models. The 10-fold cross validation gave us a result of 61.7% accuracy which is lower than both SVM and MLP. Results in Table 5.

Again we constructed the model from the whole train dataset and evaluated it on the test dataset. The accuracy of the Logistic regression model was 61.69% and the F1-score 62% and we can see that the model performed the worse among the first three models (Table 6).

	precision	recall	f1-score
angry	0.74	0.68	0.71
calm	0.65	0.75	0.7
happy	0.61	0.54	0.58
sad	0.48	0.49	0.48

Table 5. Precision, recall and F1-scores for “angry”, “calm”, “happy” and “sad” using Logistic Regression.

	accuracy	f1-score
Log. regression	61.69%	62%

Table 6. Final accuracy and F1-score using Logistic Regression.

4.4 Random forest model

Our last model that we used was Random forest. Again we performed 10-fold cross validation and we were predicting the same emotions as in the previous models, those are “angry”, “calm”, “happy” and “sad”. The 10-fold cross validation gave us a result of 71.0% which was the best so far.

We constructed the Random forest model on the whole train dataset and evaluated it on our test dataset. The final accuracy of the model was 70.78% and we can see that random forest performed the best among all the models. The final accuracy can be seen in Table 7 and Table 8.

	precision	recall	f1-score
angry	0.88	0.74	0.8
calm	0.66	0.93	0.77
happy	0.59	0.69	0.63
sad	0.8	0.49	0.61

Table 5. Precision, recall and F1-scores for “angry”, “calm”, “happy” and “sad” using Random Forest.

	accuracy	f1-score
RF	70.78%	71%

Table 6. Final accuracy and F1-score using Random Forest.

4.5 Evaluation

The final evaluation of the research includes the performance of each of the prediction models. We can see that the model that gave us the best results is Random Forest with approximately 70.78% classification accuracy, the second and third best prediction models were SVM and Multi-Layer Perceptron with classification accuracy of 69.48% and 64.85% respectively and the worst performance gave us Logistic regression with 61.69% classification accuracy.

Based on the comparison of the models, our conclusion was that we take Random Forest as our best prediction model. The final accuracy is therefore 70.78% with F1-Score 71%.

5 CONCLUSION

In conclusion we have proven in this research that human emotions can be predicted using machine learning. Our best model for predicting emotions is Random Forest which has classification accuracy of 70.88% on the RAVDESS database. Since emotion recognition is in general an extremely complicated and difficult task, the accuracy we got is a positive surprise.

Emotion recognition is often used in applications to help indicate the emotion of the person we are talking to, upon which we can accommodate our way of speaking. With our model we can successfully predict the emotion of a speaker which means that it could be used for practical applications.

6 REFERENCES

[1] Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Ali Mahjoub M, Cleder C. Automatic Speech Emotion Recognition Using Machine Learning. In: Social Media and Machine Learning. IntechOpen; 2019. p. 1–16.

[2] Gupta K, Gupta D. An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system. Proc 2016 6th Int Conf - Cloud Syst Big Data Eng Conflu 2016. 2016;493–7.

[3] Fahmi, Jiwanggi MA, Adriani M. Speech-Emotion Detection in an Indonesian Movie. 2020;(May):185–93.

[4] Wunarso NB, Soelistio YE. Towards Indonesian speech-emotion automatic recognition (I-SpEAR). Proc 2017 4th Int Conf New Media Stud CONMEDIA 2017. 2017;2018-Janua:98–101.

[5] Basu S, Chakraborty J, Aftabuddin M. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. Proc 2nd Int Conf Commun Electron Syst ICCES 2017. 2018;2018-January(Icces):333–6.

[6] Motamed S, Setayeshi S, Rabiee A, Sharifi A. Speech emotion recognition based on fusion method. J Inf Syst Telecommun. 2017;5(1):50– 6.

[7] Suhail MSK, Guna Veerendra Kumar J, Mahesh Varma U, Vege HK, Kuchibhotla S. Mlp model for emotion recognition using acoustic features. Int J Emerg Trends Eng Res. 2020;8(5):1702–8.

[8] Winursito A, Hidayat R, Bejo A. Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition. 2018 Int Conf Inf Commun Technol ICOIACT 2018. 2018;2018-January:379–83.

[9] Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). PLoS ONE. 2018. 1–35 p.

[10] Dörfler M, Grill T, Bammer R, Flexer A. Basic filters for convolutional neural networks applied to music: Training or design? Neural Comput Appl. 2020;32(4):941–54.

[11] Bhuyan AK, Nirmal JH. Comparative study of voice conversion framework with line spectral frequency and Mel-Frequency Cepstral Coefficients as features using artficial neural networks. Proc - 2015 Int Conf Comput Commun Syst ICCCS 2015. 2016;230–5.

Šimen Ravnik, Student at the Faculty of Computer and Information Science, University of Ljubljana