Emotion Recognition Through Analysis of Speech – A Review

Rasim Atakan Poyraz[1], Prajyot Suvarna[1], Alexander I. Iliev[1, 2[0000-0002-4220-3637]]

[1] SRH University of Applied Sciences, Ernst-Reuter-Platz 10, 10587, Berlin, Germany
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. Georgi Bonchev Str., Block 8, 1113, Sofia, Bulgaria
atakanpoyraz1@gmail.com, prajyotsuvarna10@gmail.com, ailiev@berkeley.edu

Abstract. The feature extraction is very important for emotion recognition through speech. There are several approaches when dealing with emotion recognition. In this paper, we present different feature extraction approaches as well as different models used to differentiate between a neutral speech versus an emotional speech sample. This research is instrumental for the digitization and preservation of cultural heritage, as it allows us to capture and analyze the emotional nuances in historical audio recordings, ensuring their accurate representation for future generations. We have selected two works consisting of a total of four different methods for emotion recognition. In the first paper by Jacob (2017), we look at Decision tree and Logistic Regression. Decision tree attains an 84.45% accuracy on the test class whereas logistic regression is able to achieve an accuracy of 66.85% after stepwise regression. These methods contribute to the digital archiving of cultural heritage by providing robust tools for analyzing and preserving the emotional content of spoken artifacts. In another paper by Bhatti et all. (2004), sequential forward selection (SFS) was used to create subsets from the given features and relevance of the subsets of features. General regression neural network was used to evaluate the accuracy which was found to be 80.69%. As a complementary purpose, modular neural network was performed with an accuracy of 83.31% with the same dataset. These techniques enhance our ability to maintain the integrity and emotional depth of cultural heritage recordings in digital archives.

Keywords: Emotion Recognition, Decision Trees, Logistic Regression.

1 Introduction

For decades, machine learning and artificial intelligence have become integral parts of human life. They are used in various areas, and the developments are still ongoing. Despite all these advancements, machine learning still faces the challenge of adapting to human life with an emotional direction. With the continuous development of the new generation of human-machine interaction technology and the increasing demands for

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 14, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2024. ISSN: 1314-4006, eISSN: 2535-0366

emotional intelligence in smart home services, human-machine interaction technology based on speech emotion recognition has attracted wide research attention. This technology is also pivotal for the digitization and preservation of cultural heritage, as it enables the accurate interpretation and archiving of emotional expressions in historical audio recordings. In this paper, we analyze two research papers on the topic of emotion recognition through speech and their varying approaches. We dive into their methods and results to compare them, highlighting their contributions to preserving the emotional richness of cultural heritage in digital form.

2 Methods

In the work done by Jacob (2017 in his paper "Modelling speech emotion recognition using logistic regression and decision trees" aims for Speech Emotion Recognition (SER) in Malayalam, the native language of Kerala, the most literate state in India. He refrains from using the hidden Markov models (HMM) as the process of selection of features is challenging. To use the HMM model, the features must fit into the HMM structure.

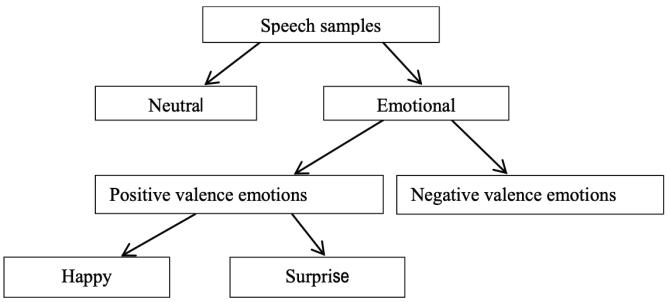

Fig. 1. Schematic of the intuitive modelling for the binary classifications of emotions.

In building intuitive models for Speech Emotion Recognition (SER), the following binary classifications were considered as relevant as seen in Figure 1:

- Classification of speech samples as emotional versus neutral.
- Classification of emotional speech as positive valence vs negative valence
- Classification of positive valence emotional speech as happy vs surprise.

Formants are the spectral maximum which results due to sound resonance in the vocal track. The first four formants (a) along with their respective bandwidth are considered as features for this model. The reason for this is that as emotions change, so does the intake of air which causes the formants and their respective bandwidth to fluctuate. Formants are computed by solving the roots of the linear predictive coding (LPC) polynomial.

Female speech is considered for emotion recognition as it is assumed to be more expressive of the two. This is not proven though, in a study conducted it showed that in the vocal channel the speaker's gender does not systematically affect the reliability of

emotional judgement (Lausen & Schacht, 2018). The database for the investigations consisted of 10 female native Malayalam speakers. Induced emotional speech was elicited from speakers in neutral and six basic emotions. The non-neutral content was used for their investigations. The database thus formed was validated by perpetual listening tests by 10 listeners not having any listening disabilities. Further, a five level Mean Opinion Score (MOS) was done and only the samples that qualified as 'Good' were proceeded with. In the end 2800 wav files were selected for feature extraction. For each speech file, the first four formants along with their bandwidths were extracted using Praat (Boersma & Weenink, 2009). Finally, the importance of this research has a pivotal role into the fields of Culture in Wellness and Healthcare, as cited in these studies (Ignatova, 2018; Ignatova, 2021; Ignatova, 2022; Ignatova, 2023a; Ignatova, 2023b; Ignatova, 2023c), which highlight the significant impact of emotional states on health outcomes, thereby supporting and broadening the scope of our current review.

2.1 Decision Trees

The modelling of the decision tree was done based on the standard Classification and Regression Tree (CART). Each feature was assessed algorithmically to select the ones necessary. Entire data were considered and binary splits on every predictor were performed and only the split with best optimization criteria was selected. Splitting was stopped when the node was pure (containing only one class). Pruning was also done, which is the process of turning some branch nodes into leaf nodes and removing the leaf node from the original branch. In an optimal pruning scheme, it first prunes branches which give the least improvement in error cost. The cost of each node was the classification error for the node multiplied by its probability. Matlab was used for modelling.

2.2 Logistic Regression

Regression is the second method used by (Jacob, 2017). Regression models the relationship between a response (Y) and predictors (x1,...,xk). We assume p to be the probability which is 1. Then we cannot simply write the regression function as p = β0 + β1x1 +⋯+ βkx as the right-hand side of the equation may evaluate to a value greater than 1 whereas we need it to be in the range of [0,1] since it is a probability. To obtain a valid equation, we need to convert this:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

which leads to,

$$\frac{p}{1-p} = \exp\left\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right\}$$

and

$$p = \frac{\exp\{\beta_0 + \beta_{1X_1} + \cdots + \beta_k x_k\}}{1 + \exp\{\beta_0 + \beta_{1X_1} + \cdots + \beta_k x_k\}}$$

Using the link function, the author guarantees that the logistic regression maps the interval (0,1) onto the whole real line. Estimating the values of coefficients β0, β1,...,βk and carrying out tests of significance on these values is known as Logistic regression. P-value for each coefficient tests the null hypothesis that the coefficient is equal to zero (no effect). A lower p-value would suggest that the feature is meaningful to the model.

This modelling technique also uses Stepwise regression where the most significant variable is systematically added, or the least significant one is removed. To validate the obtained results, certain criteria were checked:

- The categories of dependent variables need to be mutually exclusive.
- One or more independent variables should be continuous or nominal.
- There should be independence of observations to prevent repeated measures.
- There should be no multicollinearity.
- There should be no outliers.

Coefficients represent the mean change in the response for one unit of change in the predictor while other predictors in the model are constant. To assess the model fit, the data was grouped using their estimated probabilities from lowest to highest (in 10 groups by default) and performed a chi-squared test to check if the observed and expected frequencies are significantly different. If the p-value is lower than the chosen significance level, it is indicated that the predicted probabilities differ from observed probabilities.

Considering the probability that the event occurs as P(1), P (1) = exp (Y')/(1 + exp Y')), where Y' is the algebraic representation of the regression line.

In the case of Neutral vs Emotional speech, the stepwise regression eliminated the fourth bandwidth (eighth predictor) upon determining its contribution as insignificant. The new regression equation obtained was as follows:

$$Y's = -4.07 - 0.00393\ F1 + 0.001854\ F2 + 0.00323\ F3 - 0.001990\ F4 + 0.00489\ F5 + 0.003452\ F6 - 0.000995\ F7.$$

whereas the regression model containing all predictors was obtained as:

$$Y' = 4.20 + 0.00380\ F1 - 0.001839\ F2 - 0.00328\ F3 + 0.001971\ F4 - 0.00487\ F5 - 0.003482\ F6 + 0.000987\ F7 + 0.000305\ F8$$

In the case of Positive valence vs Negative valence, the stepwise regression obtained with the most significant predictors F were as follows:

$$Y's = 17.58 - 0.001403\ F2 - 0.004967\ F3 - 0.001418\ F5 + 0.001138\ F6 - 0.000795\ F7$$

whereas the model containing all predictors was as follows:

$$Y' = 19.03 - 0.001172\ F1 - 0.001314\ F2 - 0.004697\ F3 - 0.000441\ F4 - 0.001486\ F5 + 0.001132\ F6 - 0.000814\ F7 + 0.000240\ F8$$

For the case where the direction of positive valence emotion was tested for happy vs surprise. The HL statistic was indicative of a poor fit. The equation obtained was as follows:

$$Y's = 16.13 - 0.00246\ F3 - 0.00208\ F4 - 0.002774\ F6$$

whereas the equation with all the predictors is as follows:

$$Y' = 15.67 + 0.00439\ F1 - 0.00260\ F2 - 0.00153\ F3 - 0.00200\ F4 - 0.00004\ F5 - 0.00290\ F6 + 0.00033\ F7 - 0.000848\ F8$$

The classification accuracies and Hosmer Lemeshow statistic (HLS) obtained are provided in Table 4.

2.3 General Regression Neural Network

In the paper of Bhatti et al. (2004) clearly points out that collecting a few features is the goal. They first started to explore several features for classifying the speaker effect which are phoneme and silent duration, short-time energy, and pitch statistics. To be able to focus on the emotional state, they have chosen 17 prosodic features (also known as acoustic features) by analyzing the speech spectrogram which can be seen in "Table 1".

Table 1. Feature Description.

Feature	Description
1.	Pitch range (normalized)
2.	Pitch mean (normalized)
3.	Pitch standard deviation (normalized)
4.	Pitch median (normalized)
5.	Rising pitch slope maximum
6.	Rising pitch slope mean
7.	Falling pitch slope maximum
8.	Falling pitch slope mean
9.	Overall pitch slope mean
10.	Overall pitch slope standard deviation
11.	Overall pitch slope median
12.	Amplitude range (normalized)
13.	Amplitude mean (normalized)
14.	Amplitude standard deviation (normalized)
15.	Amplitude median (normalized)
16.	Mean pause length
17.	Speaking rate

They proposed an efficient one pass selection procedure which is called sequential forward selection (SFS) (Kittler, 1978). It incrementally constructs a sequence of subsets from features by adding relevant features to previously selected ones. To evaluate the relevance of the subsets of features, they have chosen to use the general regression

neural network (GRNN) (Specht, 1991). GRNN is based on the estimation of a probability density function. Unlike the conventional multilayer feed-forward neural networks, it requires many iterations in training to achieve the desired result. In mathematical terms, if we have a vector random variable x, a scalar random variable y, let X be a particular measured value of x, then the conditional mean of y given X can be represented as:

$$\hat{Y}(X) = \frac{\sum_{i=1}^{n} Y_i \exp(-\frac{D_i^2}{2\sigma^2})}{\sum_{i=1}^{n} \exp(-\frac{D_i^2}{2\sigma^2})}$$

However, it is not possible to compute the optimum value of σ since the underlying parent distribution is unknown. So, they decided to find the value on an empirical basis. Therefore, a leave-one-out cross validation was used to determine the value of σ for minimum error.

3 Consistency-Based Feature Selection & Modular Neural Network

Consistency-based approach was used as complementary to evaluate the relevance of the features. The consistency measure formula where the distances are the space of features. It is important to get higher values of consistency measures for optimized features.

$$C = \frac{\text{mean inter-class distance}}{\text{mean intra-class distance}}$$

After computing the consistency measure formula for each feature, they found that feature 2 had the highest measure. The top 3 features' measures were the same as the GRNN and SFS which were 6, 7, and 15. Using the combined SFS/GRNN and consistency-based method, they got 12 features.

MNN can solve complex computational tasks by dividing the tasks into simple subtasks, and then combining their individual solutions. MNN also offers several advantages over a single neural network in terms of speed and capability. The architecture of MNN consists of 6 sub-networks, where each sub-network specializes in an emotion class. The hierarchical MNN architecture can be seen below in Figure 2.

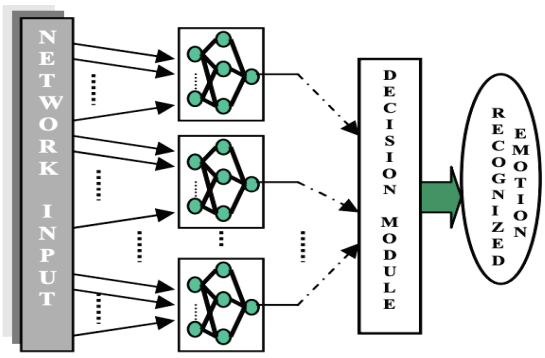

Fig. 2. Hierarchical MNN Architecture.

4 Consistency-Based Feature Selection & Modular Neural Network

4.1 Decision Tree

From Fig. 6, it was observed that the eight feature x8 present in the fourth bandwidth B4 is the least important predictor regarding neutral vs emotional binary classification. The root node classification was made on the value of the third formant x3. The subsequent classifications have been based on the second formant (x2), fourth formant (x4) and third bandwidth (x7). Higher accuracy was obtained with the subset of seven predictors rather than with the complete set.

Table 2. Coefficients indicating predictor performance.

Number of trees in ensemble	Coefficient values indicating predictor importance							
	F1	F2	F3	F4	B1	B2	B3	B4
50	0.0012	0.0019	0.0047	0.0080	0.0011	0.0019	0.0008	0
100	0.0008	0.0015	0.0036	0.0061	0.0009	0.0013	0.0008	0.0004

The SER accuracy obtained on train class was 89.45%.

Table 3. Confusion matrix for neutral/emotional classification of test class.

Percentage accuracies for SER		
Output/input	Neutral	Emotional
Neutral	85	15
Emotional	16.11	83.89

4.2 Logistic Regression Model

The predictors are fit using an iterative reweighted least squares algorithm to obtain maximum likelihood parameters for each of the three binary classification which are (i) Neutral vs Emotional (ii) Positive valence vs Negative valence (iii) Happy vs Surprise. Although the main objective of the regression procedure was for case (i) as mentioned above.

Table 4. Summary of results of various binary classifications for decision trees.

Percentage accuracies for SER						
Case	Class	Description	8 Features		7 Features	
			Resub	Cross val	Resub	Cross val
i	2	Neutral versus emotional	88.61	82.68	89.44	84.45
ii	2	Positive valence versus negative valence	90.80	87.76	91.51	87.67
iii	2	Happy versus surprise	92.69	84.19	93.63	87.31

Table 5. Prediction accuracies for various cases of binary logistic regression.

Sl. No.	Class	Description	Prediction accuracy using 8 features (%)			Prediction accuracy with features got by stepwise method (%)		
			Train	Test	HLS	Train	Test	HLS
1	2	Neutral versus emotional	70	68.06	0.692	66.9	66.85	0.692
2	2	Positive valence versus negative valence	69.22	68.13	0.881	70.87	69.49	0.631
3	2	Happy versus surprise	73	70	0.765	68	67.6	0.452

Through these results, we can summarize that the accuracy obtained when including only 7 features, i.e. emitting the fourth bandwidth B4 is higher than when all the data is considered.

4.3 General Regression Neural Networks

By applying GRNN and SFS concepts, 17 features were used. The mean square error versus feature indexes were plotted as shown in Figure 3.

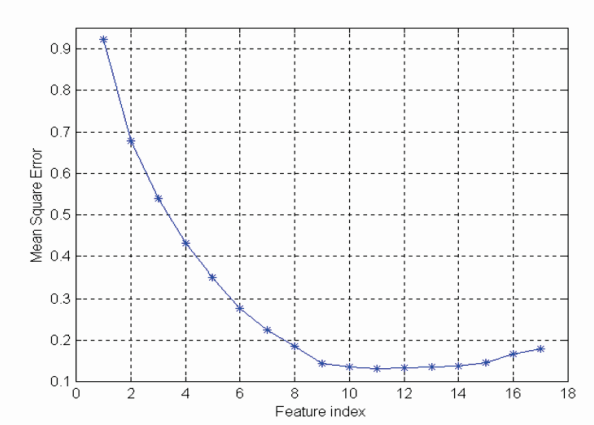

Fig. 3. Mean Square Error vs. Feature Index.

According to the graph above, the minimum error was achieved with 11 features but also it can be clearly seen that the selected feature number after 9, the error curve is almost flat which can be interpreted as the nature of the GRNN modeling process. This will make the network difficult to characterize the underlying mapping beyond a certain number of features due to the limited number of training samples. The experiments were performed on speech samples from 7 different subjects with a variety of four languages. 580 speech utterances delivered with one of the 6 emotions which was used for training and testing. The 6 emotional states were happiness, sadness, anger, fear, surprise and disgust. They chose 435 utterances for training and the rest were left for testing.

The GRNN was tested with an equal number of nodes and features and the output nodes were selected 6 since there were 6 emotional states to be detected. Each emotion was in association with 10 nodes. The overall correct recognition was 77.24% on 17 features and 80.69% was on 12 features.

4.4 Modular Neural Network

As an alternative approach to Bhatti et al. (2004) first approach, check the accuracy, a modular neural network (MNN) was tested. MNN is capable of mapping each set of input features to one of six emotional states. Even though they used GRNN for feature selection, it had the disadvantage of the computational complexity and therefore found that GRNN is not suitable for evaluation of new samples. Thus, they have applied the MNN based on the backpropagation for classification.

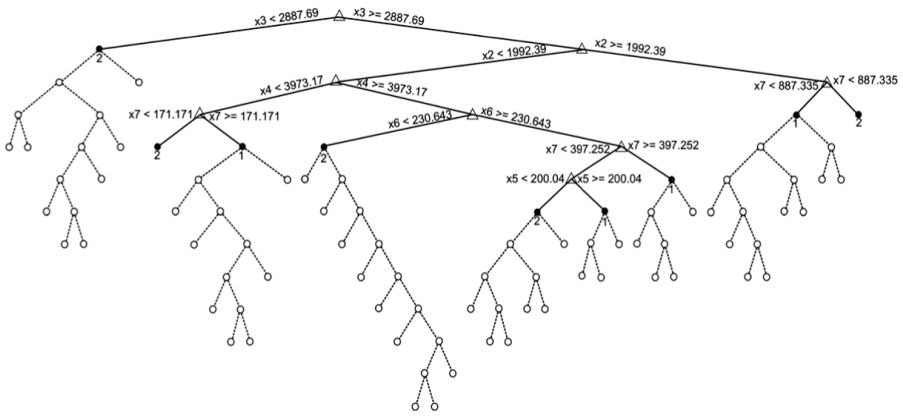

Fig. 4. Schematic of a pruned decision tree.

The same amount of training and testing data were used as the GRNN experiment. MNN was tested by using 12 features. Each subnet consisted of a 3-layered feed-forward neural network with one 12 element input vector. MNN had the most recognition rate of 83.31% with the fastest learning time.

5 Conclusions

In this work we have analyzed four different algorithms and approaches from two papers that can be applied for SER. Although this problem can be approached in many ways, the feature selection process for any approach plays a crucial part in the accuracy of the prediction. This is particularly significant for the digitization and preservation of cultural heritage, where accurate emotion recognition can enhance the authenticity and richness of archived audio recordings. As we saw in the approach of the decision tree from Figure 4, where we see the Xs the splitting paths we take form each branch, with the accuracy obtained using the entire dataset and considering all the predictors was lesser compared to when the feature x8, i.e., the bandwidth of the fourth formant, was removed. The stepwise logistic regression also eliminated the fourth bandwidth (eighth

predictor) upon evaluating its contribution as insignificant. Thus, it is important to select features that are independent and mutually exclusive. A higher accuracy was observed using the decision tree model compared to the logistic regression model when the formants and their bandwidth were considered as features for SER. In contrast, when selecting features from "Table 1", higher accuracy in GRNN and MNN was obtained with a lower number of features (12). These findings underscore the importance of feature selection in developing effective SER models, which is essential for the reliable digitization and preservation of cultural heritage audio materials.

References

Bhatti, M. W., Wang, Y., & Guan, L. (2004). A neural network approach for human emotion recognition in speech. In *2004 IEEE International Symposium on Circuits and Systems (ISCAS), Vancouver, BC* (pp. II-181). IEEE Xplore. http://dx.doi.org/10.1109/ISCAS.2004.1329238

Boersma, P. & Weenink, D. (n.d.). *Praat: doing phonetics by computer*. https://www.fon.hum.uva.nl/praat/

Ignatova, D. (2018). The effects of swimming on preschool children with spinal abnormalities. In R. Penkova (Ed.) *17th International BASOPED Conference "Traditions and innovations in the education of the Balkan countries"* (pp. 207-212). Balkan Society for Pedagogy and Education.

Ignatova, D. (2021), Specificity of the motor potential for achieving Scholar Wellness. *Trakia Journal of Sciences, 19*(Suppl. 1), 867-873. https://doi.org/10.15547/tjs.2021.s.01.136.

Ignatova, D. (2022). Study the influence of yoga specialised practices on the formation of correct body posture and corrections of spinal deformities. *Smart Innovations in Recreative & Wellness Industry and Niche Tourism, 4*(1-2), 17-22. https://scjournal.globalwaterhealth.org/wp-content/uploads/2023/01/p.17-21_Ignatova_UK_V.4_Is.1-2_2022.pdf

Ignatova, D. (2023a). Implementation of motor complexes based on specialized application system blaze-pod trainer. *Strategies for policy in science and education, 31*(6), 653 - 667. https://doi.org/10.53656/str2023-6-6-imp

Ignatova, D. (2023b). Motor activity based on learning – contemporary trends in school wellness. *Smart Innovations in Recreative & Wellness Industry and Niche Tourism, 5*(1-2), 22-26. https://scjournal.globalwaterhealth.org/wp-content/uploads/2024/02/4.%E2%80%8CIGNATOVA__p.22-26-_V.5-Is.-1-2_2023.pdf

Ignatova, D. (2023c). Affirming wellness culture through innovative methodology related to Blaze-pod trainer system. *Strategies for policy in science and education, 31*(2), 212-225. https://doi.org/10.53656/str2023-2-7-aff

Jacob, A. (2017). Modelling speech emotion recognition using logistic regression and decision trees. *International Journal of Speech Technology, 20*(4), 897–905. https://doi.org/10.1007/s10772-017-9457-6

Lausen A., & Schacht, A. (2018). Gender Differences in the Recognition of Vocal Emotions. *Frontiers in Psychology, 9*, Article 882. https://doi.org/10.3389/fpsyg.2018.00882

Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks, 2*(6), 568-576. https://doi.org/10.1109/72.97934

Received: March 15, 2024
Reviewed: April 05, 2024
Finally Accepted: May 15, 2024