ANKARA SCIENCE UNIVERSITY, RESEARCHER
Vol. 3, No. 2, Decmber 2023
e-ISSN: 2717-9494
Research Article/Araştırma Makalesi
Doi: https://doi.org/10.55185/researcher.1192370

A Literature Review on Emotion Recognition in Speech

Ö. Çağrı DALA[1*]
[1]Ankara Science University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, Ankara, Türkiye; ORCID: 0000-0001-8202-7802
* Corresponding Author: cagri.dala@ankarabilim.edu.tr

Received: 20 October 2022; Accepted: 24 September 2023

Reference/Atıf: Ö.Ç. Dala, "A Literature Review on Emotion Recognition in Speech", Researcher, vol.3, no.2, pp.1-7, December 2023, doi: 10.55185/researcher.1192370

Abstract

In our age, we are bombarded with multimedia content daily. Although, face-to-face communication always outgrows the potential factors of healthy assessment of our peers through recorded content or live media interaction, (be it text, video, images, speech) new approaches to render us able to understand and discern between emotions of our peers on multimedia content are getting more and more popular and more complex. Two robust topics in this regard are generally named as sentiment analysis and emotion detection. The advent and exponential growth of social networks and for instance, the employment of speech bots have made it a necessity to particularly address the problem of healthy emotion recognition outside face-to-face, everyday conversations or interactions. Machines' capability to perform the set of tasks through Machine Learning approaches, namely consisting of detecting, expressing, and understanding emotions is collectively known as, as in humans, emotional intelligence. Different modes of input as human behavior like those taken from audio, image, video sources and signal interpretations processed through Electro-encephalography (EEG), related brain wave measurements are used in emotion recognition. My study aim is intended to be the examination and review of recent study approaches in Emotion Detection in Speech, possibly establishing links or differences between recent study publishes because each study paper focuses on a single or set of Machine Learning approaches which are employed in Emotion Detection in Speech. This paper tries to examine various relevant research involving methods of Machine Learning which were studied and tested under these research respective to Speech Emotion Recognition (SER). Effectiveness of the involved methods and databases are discussed while commenting on the studies and expressed in the form of their findings. Improvements throughout these studies are, though not chronologically, compared using simple tables which show independent accuracies of several Machine Learning classifier combinations.

Keywords: speech emotion recognition, emotion detection in speech, machine learning classifier, machine learning approaches

1. Introduction

Human emotion detection in speech is a subset of emotion recognition. Emotion recognition focuses on other modes of input as well such as video captures, images, signal interpretations processed through Electro-encephalography (EEG), related brain wave measurements; all collected where a human being is the subject/source. Human emotion can be identified by many physical attributes like responses, body gesture, heart rate, body temperature etc. It can also be detected without a physical contact such as using speech [1]. Speech features such as intensity, intonation, pitch which are prosodic features are extracted from human audio sources. Spectral features are represented by further processing signals from these audio sources. Speech emotion recognition (SER) is benefitting many applications in recent years. Examples include tutoring system used in distance learning which can detect the bored or not interested users and can allow changing the style and the level of study material provided [1][2]. There are three important factors under consideration when detecting emotion from speech:

1. Contents: "what is said"

2. Style/way: "how it was said"

3. Human: It refers to male/female/actor "who says it".

The main steps in SER are speech database (preprocessing if needed), the process of feature Extraction and classifier used to detect the emotions [1].

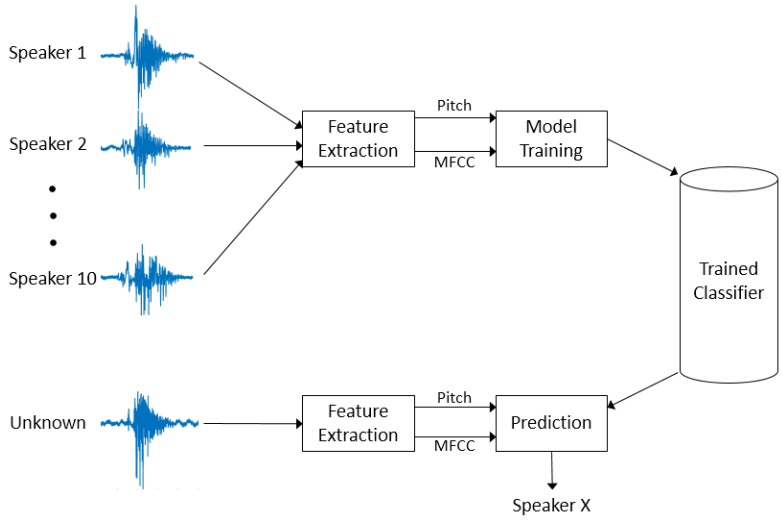

Figure 1 – Classification Process

Speech database collection is important for emotion recognition from speech. A lot of work has been done on collection techniques and evaluation of speech databases. There are many criteria that are used to evaluate the suitability of speech databases for different ML Classification methods.

Feature selection and extraction is the second and one of the important steps in SER activities.

Extraction of features is a critical step in SER, determining the final stage performance which ultimately is the accuracy of Machine Learning classifiers. In the case of relevant features, combination of features can also be relevant in improving the emotion recognition [1] Prior to this final stage which is Classification, the ML model or models in question must be trained with the help of a portion of the dataset. (TRAIN_SET comprising usually %70-80 of the whole samples belonging to the dataset).

Genetic algorithms are originally a breed of optimization over search problem approaches. If we assume going through weighted nodes or edges to have made out a pattern, in consistently lesser durations of time, is refined closer to the goal at hand, search problem solution improvements can be met as shown in many previous studies. One stretched assumption which shows us that application of GAs on MLs is feasible is that in a credit risk assessment setting when a priori information about the potentially attractive areas is available, then the initial population of the GA can be generated in such a way that the attractive areas of the feasible region must be covered with a set of points and the dimensionality of the problem can be reduced to those features that form attractive areas. In this way, the attractive area (the region of attraction) of a global minimum is defined as the largest set of points, such that for any starting point from that set, the infinitely small step of the gradient descent algorithm will converge to the global minimum. Feature selection is the problem of choosing a small subset of features that ideally is necessary and sufficient for describing the target concept. (Kira & Rendell, 1992) [3]

A more recent study comparing performance of heuristic optimization among options applicable to problems on making decisions shown to be critical by UAV flight range limitations can be found in [4]. These options include improved GA and Nearest Neighbors heuristics.

Parallel computation that starts each optimization process from a different starting point (vertex) and works with diversified vertices can be considered as a mechanism to prevent stagnation during the steps of search optimization techniques. [5]

As an additional remark due to the following part quoted from a study in Machine Learning in [6], "we show that with proper regularization, distributional bias of the data can push the solution towards the global minima."

2 Literature Review

Classifiers are used for recognition of subsets of human emotion. Several classifiers can be implemented for SER and their performance depends on the database design and features extracted from the speech [1]. There are many Machine Learning classifiers to classify emotions from speech input. Some of the popular ML classifiers are mentioned in Table 1 [1]:

TABLE 1: ML classifiers

Classifiers	Description
Incomplete Sparse Least Square Regression (ISLRS)	It is a model which uses least square regression to find emotion labels based on linear relationship between speech features [1][8]. It is a statistical method which depends on sparseness of the model itself and incompleteness of the dataset. There are two critical reasons why sparse linear models might be preferable to non-sparse models. First, sparse models often produce lower variance predictions, and hence better generalization. Also, models with a reduced number of nonzero coefficients tend to represent only strong effects of the data, thus eliminating details that may be unimportant to a further analysis.
Extreme learning machine (ELM)	It is a feed forward neural network which uses only one hidden layer. ELM is very efficient and effective when the training set is small. Unlike conventional NNs whose weights need to be tuned using the backpropagation algorithm, in ELM the weights between the input layer and the hidden layer are randomly assigned and then recalculated. The weights between the hidden layer and the output layer can be analytically determined through a simple generalized inverse operation of the hidden layer output matrices. [1][9].
Gaussian Mixture Models (GMMs)	GMMs perform better when data is normally distributed. The classification accuracy of GMMs also depends on factors like number of Gaussians in each class, size of the dataset, distribution of data and so on. GMM is a probabilistic model and is considered as a state-of-art model and is used mostly for identification and verification of speaker. This model is the most appropriate model in case of global feature [1] [10-16].
Support Vector Machine (SVM)	SVM is more popular in case of non-linear features as opposed to ISLRS (Kernel function). SVM is used in case of speaker independent application and gives better results in such a goal as compared to other classifiers [1][16].
Artificial Neural Networks (ANN)	ANN is another classifier which is popular for its non-linear features (Kernel function) [1]. This classifier gives better results compared to GMM and HMM in case of smaller number of training samples [1] After ANN training, the information may produce output even with inadequate data. The loss of performance here relies upon the significance of missing data.
Vector Quantization (VQ)	In VQ, a vector of fixed size is created with fixed dimensions. [1] In a Learning Vector Quantization setup, the neural network in question has a first competitive layer and a second linear layer. The competitive transfer function produces a 1 for output element a_i^1 corresponding to i^*, the winning neuron. All other output elements in a^1 are 0. Neurons close to the winning neuron are updated along with the winning neuron. The linear layer transforms the competitive layer's classes into target classifications defined by the user.
k-Nearest Neighbors (k-NN)	k-NN assigns the class label of most of the K-nearest patterns in data space. For this sake, we must be able to define a similarity measure in data space. The choice of K defines the locality of KNN. For K = 1, little neighborhoods arise in regions, where patterns from different classes are scattered. For larger neighborhood sizes, e.g. K = 20, patterns with labels in the minority are ignored.
Hidden Markov Model (HMM)	In case of automated emotion recognition, HMM is very popular and gives accurate results. This model is especially used when samples are subjectively considered to be little in number. [1] Forced alignment can be used to represent audio features as HMM states. HMM can be incorporated along with a Gaussian Mixture Model.

In human emotion recognition from speech, most essential part is to identify relevant features and classifiers. From the literature, the accuracy of the above-mentioned classifiers with different speech databases are summarized in Table 2 as follows:

TABLE 2: Literature Survey of Emotion Recognition from Speech

Ref.	Classifier used	Database	Accuracy (%)
[1]	Autoencoder based Adaptation model	ABC, GeWEC, EMODB, SUSAS	63.30 as an autoencoder is a type of artificial neural network used to learn efficient codings of unlabeled data (unsupervised learning) which is suitable for use at the preprocessing step. An autoencoder learns two functions: an encoding function that transforms the input data, and a decoding function that recreates the input data from the encoded representation.
[1][8]	ISLRS	eNTERFACE DB, FAUAIBO	69.33 and 60.50 as we can compare linear and logistic models' performance by using root mean squared error (RMSE) and the coefficient of determination (R^2 score), the linear models being of the nature having more applicability to prediction purposes
[1][9]	ELM	IEMOCAP	54.30 as Extreme Learning Machine approach used when in its primal development level is moderately efficient
[1]	GMM	Real life data	75 A further improvement which can be applied on the class margin (in terms of distance between mixture means) is critical for convergence of self training to useful models
[1][10]	GMM	Basque	84.7 as real-life data can be considered randomly generated to a degree except the language in which it is recorded, performance in Basque was more uniform due to setting
[1][11]	ANN and GMM	IITKGP-SESC	56
[1][12]	GMM and SVM	IITKGP-SESC	60.6 where a Support Vector Machine is representative of a more accurate hyperplane among features
[1][13]	SMO (Sequential Minimal Optimization Algorithm) with RBF (Radial Basis Function)	Berlin (EMO-DB) English (SUSAS)	78 as RBF approach includes mapping your point into the function being a Gaussian distribution where data is represented ultimately in a linear fashion. As SMO is fastest for linear SVMs and sparse data sets, this SMO + RBF as part of a SVM approach is successful with the EMO-DB dataset being comprised of the synthesis of data with emotion-simulating speech.
[1][14]	HMM with 4- mixture GMMs	Human speech and animal voice	69, 82.7 as in a general Gaussian Mixture Model, the mixture is brought together as each model of utterances is a set of model parameters with estimated mean vector, covariance matrix and mixture weight. Each of parameter models is trained as part of unsupervised classification by using the expectation maximization property (EM) in mixture GMMs. A HMM is also employed where for speech recognition, the conditional probability distribution of a state in HMM is classically modelled using GMM.
[1][15]	GMM	Swedish VP and English ISL meeting corpus Real life data	79
[1][16]	SVM	Real Life data	74.75 where SVMs provide new efficient separability of non-linear regions that use "kernel functions": generalization of 'similarity' to new kinds of similarity measures based on dot products. SVM brings an option of use of quadratic optimization problem to avoid 'local minimum' issues with neural nets. SVMs have higher accuracy on average than most other approaches.
[1][10]	SVM	Berlin (EMO-DB), VAM DB	91.6
[1][7]	SVM	Danish	88
[1]	SVM	English speaking people, English movies	71.62
[1][20]	Fuzzy Set	Malay English	60 as when fuzzy sets are implemented by their own in classification purpose, can provide lower accuracy in spite of the setting at the beginning but they can be used in cooperation with Extreme Learning Machine approaches.
[1]	kNN	English, Chineseurdu and Indonesian	80.69 where the value of k is crucial, and one needs to choose it wisely to prevent overfitting or underfitting the model.
[1][23]	HMM	Semi-professional female actress	88 as the accuracy is better than a regular GMM. A HMM can be thought of as a general mixture model plus a transition matrix, where each component in the general Mixture model corresponds to a node in the hidden Markov model, and the transition matrix informs the probability that adjacent symbols in the sequence transition from being generated from one component to another.
[1][24]	HMM	German	64.7

[1][25]	VQ based HMM	Burmese	60.1, VQ allows each speech utterance under consideration to be represented as a time sequence. The dataset in this study should be evaluated for its suitability given the language hardships.
[1][26]	HMM	Spanish	70

From the literature review, it is observed that the accuracy of the human emotion detection from speech depends on feature extraction, feature vector size, classifier used and speech database.

3. Method

As the method of Research Study, Literature Review was chosen. A comparative display of simple tables including the approaches in Machine Learning and their applications on specific databases with the resulting accuracies were given depending on respective studies specified as references.
The research process was divided into three steps:
-Problem definition: A comprehensive reflection of methods used in Machine Learning for the purpose of Emotion detection in speech was to be given. To aid this goal, the following research questions were devised and examined:
a) What approaches in Machine Learning are used for emotion detection in speech?
b) What are efficient algorithms used for emotion detection in speech?
c) Were the approaches or combination of approaches sufficient in terms of accuracy or performance?
-Literature search: to display the paper which gave me further leads in Literature Review, I had to use specific keywords:

"Emotion recognition"
"speech database"
"feature extraction"
"classifier"

-Basic analysis: An elimination of references from the [1] referenced paper which were not immediately relevant to the Literature Review and the comparison of the approaches which were in the remaining studies given as references in [1] were carried out. The paper is intended to be a summarization of comparisons of such approaches in the field.

4. Results

The accuracy increases with combination of classifiers. From the survey, it is also observed that the accuracy of the human emotion detection is not only dependent on the classifier used to detect it, it also depends on the speech database, features and the size of the feature vector used [1]. It is also observed that mostly features are categorized into prosodic, spectral or combination of these. Choice of classifiers or the combination used also plays an important role in achieving more accuracy. In the case of database, size and quality of speech recordings are the important factors affecting the accuracy [1].

5. Discussions

If we were to assume that all the datasets used were of equal length and quality: the SVM approach is the most accurate in classifying Human emotions from speech with the HMM coming second where HMM is normally most efficiently used over a small number of samples. ELM is the least efficient approach possibly because it uses one hidden layer although being feed forward is a prospective feature for the approach. SMO (Sequential Minimal Optimization algorithm) with RBF (Radial Basis Function) are two linear optimizations which give out a considerably successful outcome. The database and relatedly, the language spoken by the human subject and the naturality of speech in these recordings should play a more important role in the evaluation of accuracy.

6. Conclusion

A Literature Review based on Emotion Recognition in Speech has been given in this study. Artificial features like noise have been found to impact the results of the given studies where resulting performance is affected. This problem can be resolved using speech analysis based on correlation. There are still challenges in selecting the right speech database, feature selection by the approach(es) and estimation of features, classifiers which can get accurate results.

References

[1] "Survey on Human Emotion Recognition: Speech Database, Features and Classification" – Y.B. Singh, S. Goel

[2] "Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition", International Journal of Speech Technology, vol. 21, no. 1, pp. 167–183, 2018. – S.G. Koolagudi, Y.V.S. Murthy and S.P. Bhaskar

[3] "Genetic algorithm-based heuristic for feature selection in credit risk assessment" – S. Oreski, G. Oreski

[4] "An Efficient Genetic Algorithm for Routing Multiple UAVs under Flight Range and Service Time Window Constraints" – M. Karakaya, E. Sevinç

[5] "A novel parallel local search algorithm for the maximum vertex weight clique problem in large graphs" – E. Sevinç, T. Dökeroğlu

[6] "A Theoretical Characterization of Semi-supervised Learning with Self-training for Gaussian Mixture Models" – S. Oymak, T. C. Gülcü

[7] "Speech emotion recognition based on hmm and SVM", In *proceedings of International Conference on Machine Learning and Cybernetics,* pp. 4898–4901, 2005. Y.L. Lin and G. Wei

[8] "A novel speech emotion recognition method via incomplete sparse least square regression", *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014. – W. Zheng, M. Xin, X. Wang and B. Wang

[9] "Speech emotion recognition using deep neural network and extreme learning machine", In *proceedings of Fifteenth Annual Conference of the International Speech Communication Association,* pp. 223- 227, 2014. – K. Han, D. Yu and I. Tashev

[10] "Automatic emotion recognition using prosodic parameters", *INTERSPEECH*, pp. 493–496, 2005. – I. Luengo, E. Navas, I. Hernáez, J. Sánchez

[11] "Emotion recognition using LP residual", In *proceedings of IEEE Students' Technology Symposium (TechSym),* pp. 255–261, 2010. – A. Chauhan, S. G. Koolagudi, S. Kafley and K. S. Rao

[12] "Emotion recognition from speech signal using epoch parameters", In *proceedings of IEEE International Conference on Signal Processing and Communications,* pp. 1–5, 2010. – S. G. Koolagudi, R. Reddy and K. S. Rao

[13] "EmoVoice-A framework for online recognition of emotions from voice", *Perception in multimodal dialogue systems*, pp. 188–199, Springer, 2008. – T. Vogt, E. André and N. Bee

[14] "Stress and emotion classification using jitter and shimmer features", In *proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing,* vol. IV, pp. 1081-1084, 2007. – X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong and J. D. Newman

[15] "Emotion recognition in spontaneous speech using GMMs", *INTERSPEECH,* pp. 1-4, *2006*. – D. Neiberg, K. Elenius and K. Laskowski

[16] "Speech emotion recognition based on rough set and SVM", In *proceedings of Fifth IEEE International Conference on Cognitive Informatics,* vol. 1, pp. 53–61, 2006. – J. Zhou, G. Wang, Y. Yang and P. Chen

[17] "SPEECH EMOTION RECOGNITION BASED ON HMM AND SVM", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005 – Y. Lin, G. Wei

[18] "Automatic speech emotion recognition using modulation spectral features", *Speech communication,* vol. 53, no. 5, 768–785, 2011. – S. Wu, T. H. Falk, and W. Y. Chan

[19] "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles", In *proceedings ofNinth European Conference on Speech Communication and Technology,* pp.1-4, 2005. – B. Schuller, R. Müller, M. Lang and G. Rigoll

[20] "Comparison between fuzzy and NN method for speech emotion recognition", In *proceedings of Third* IEEE *International Conference on Information Technology and Applications,* pp. 297–302, 2005. – A. A. Razak, R. Komiya, M. Izani and Z. Abidin

[21] "A neural network approach for human emotion recognition in speech", In *proceedings of IEEE International Symposium on Circuits and Systems,* vol. II, pp. 181-184, 2004. – M. W. Bhatti, Y. Wang and L. Guan

[22] Dimensionality Reduction with Unsupervised Nearest Neighbors pp 13–23 Chapter K-Nearest Neighbors – O.Kramer

[23] "Emotion recognition based on phoneme classes", *INTERSPEECH*, pp. 205–211, 2004. – C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee and S. Narayanan

[24] “Hidden markovmodelbased speech emotion recognition”, In *proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, v*ol. II, pp. 1-4, 2003. – B. Schuller, G. Rigoll and M. Lang

[25] “Speech based emotion classification”, In *proceedings of IEEE region 10 International Conference on Electrical and Electronic Technology*, pp. 297–301, 2001. – T. L. Nwe, F. S. Wei and L. C. De Silva

[26] “Speech emotion recognition using hidden markov models”, *INTERSPEECH*, pp. 2679–2682,2001. – A. Nogueiras, A. Moreno, A. Bonafonte and J.B. Mariño, J. B.