

Comparison of Emotion Recognition Models in Spoken Dialogs

Preedhi Garg; Smriti Sehgal

Department of Computer Science and Engineering, Amity University, Noida

preedhigarg@gmail.com; smriti1486@gmail.com

ABSTRACT

Recognition of accurate emotions in human speech, while conversing with a spoken dialog system is the biggest challenge for all the emotion recognition models. As spoken dialog systems are more prevailing now-a-days, their usage is bounded to simple information exchange only. To ensure a compatible understanding between these systems and human user, definite detection of emotions in speech can give fine boon for the design of more natural human-machine speech interfaces. Recently, a lot of such models were proposed for identifying emotions in human's voice by extracting various types of features from speech. This paper focuses on comparing these models on the basis of features extracted and classifiers used. So that it will be easier for the programmers to analyse which approach is better to use while programming a spoken dialog system

Keywords: Classifiers, Emotion Recognition Model, Features Extracted, Spoken Dialog.

1. INTRODUCTION

Emotions play an important role in human life specially while communicating with others. An emotion is feelings with which a human speak, perceive and communicate with others but a machine dialog system always perform all activities with constant rigid and unchanged voice. In human machine interaction, a human have tendency to interact emotionally and machine ignores human's negative emotions which are vital to detect so that emotion recognition systems could provide users with improved services by being adaptive to their emotions.

Emotional awareness is important in various applications where one need to process the information received by human machine

interaction, and those are: Alert System, to detect urgency in speaker's voice; Simulation, to enhance natural interface; Broadcast, like emoticons used in e-mail communication; Problematic dialog detection, to detect problem at speaker's end; Negative emotion detection, to detect anger level of speaker [6].

Main focus is given on negative emotion detection; example says anger detection in customer care call process primarily focuses on transferring calls to human agent before customer hangs up, measure customer satisfaction and evaluates system as faulty system can make the customer more angry. Basic approach applied in system includes these steps: information given, expressions of customer's emotion; Perceive emotions, it can be denial or distrust; Take actions, and machine will yield its behavior.

There are various components of emotion recognition system and those are depicted in the figure below:

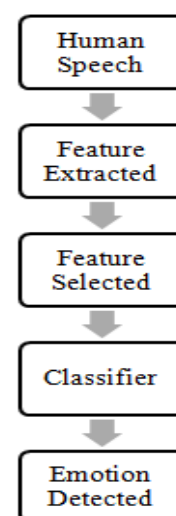


Fig 1: Components of Speech Emotion Recognition System

Among the components given in the figure, feature extraction and classifier are vital to study. There are a variety of features extracted and those are: acoustic features, prosodic features, lexical features, user and system performance features, paralinguistic features, etc. Selecting features as input for any classifier is an important task as it decides or fluctuates the efficiency of the classifier. Correct identification of emotions depends on the type of features taken out from speech. Various features are extracted from human voice which are further selected by Forward Feature Selection Algorithm and are used as inputs to classification algorithms. There are various algorithms which various researchers have applied like Nearest K Neighbor algorithm, Hidden Markov Model, Bays Classification, Kernel Regression, Support Vector Machines, and Neural Network etc.

The algorithms applied are based on three basic steps and which are: Data Acquisition, Data Labeling and Results which is accompanied by verifying the results whether there is satisfaction of user in dialog recordings. Classifiers mentioned above can be differentiated on the basis of efficiency they generate and the type of emotions they detect.

2. FEATURES EXTRACTED

In any of the Emotion Detection Systems, the first most important component is Feature Extraction. Human voice is the input for the emotion detection system and is so versatile that it conveys almost 15 types of emotions and further a numerous types of features can be extracted easily, which are categorized as follows:

2.1 Acoustic Features

Acoustic features of speech can be recorded, analyzed and experimentally observed, as its fundamental frequency, format structure. Volume, pitch, and duration are the most distinct acoustic features. Various other features can be extracted from these broader categories which are max, median, mean, standard deviation, ratio, etc. The Variations in distinct acoustic features helps in detection of different types of emotions. Duration features are the least important while pitch and energy features are moderately important [12].

2.2 Prosodic Features

Rhythm, stress and intonation of speech are the prosodic features. Prosody reflects many features of the speaker or the utterance: speaker's emotional state; the morphs of the utterance (statement, question, or command); the presence of irony, sarcasm; emphasis, contrast, and focus.

2.3 Lexical Features

In order to understand emotion in spoken dialog the machine must be able to identify the sequence of words in the speech stream and to access a range of different kinds of information about each of those words from the mental lexicon, allowing them to be interpreted and combined with other words to build up a representation of an emotion. The words people say play a part in their emotional state, although they may not be the only indicators. For Example: in speech call centre, words that indicate that the caller wishes to be transferred to a human operator are ("person", "human", "speak", "talking", "machine").

2.4 Paralinguistic Features

The recognition of the complete set of variations in the voice's feature's dynamics: loudness, tempo, pitch fluctuation, continuity, etc. When someone is angry and more excited, he or she will tend to speak in high pitch and with more loud voice. Paralinguistic features are the ones which usually are the combinations of prosodic, spectral, disfluences, etc.

2.5 User and System Performance Features

User and system performance features are the one which are extracted for the applications like tutoring dialogs. While discussing any topic if a student ask or talk about sub topic then it indicates a positive emotion or positive response otherwise a confused or negative emotion.

2.6 Contextual Features

When the past evidence of the user activity is used with present one to help and inform the emotion classification of the present user turn [4].

These are the various types of features extracted and further it depends on the programmer which features to extract and feed to the classifier. Single feature type or combination of various features can be used to detect the emotion.

3. COMPARISON OF FEATURES EXTRACTED

The various features extracted can be differentiated on the basis of performance, accuracy and emotion they detect and this difference can be used to analyze which features to use while modeling an emotion detection system.

Table 1 summarizes the difference between lexical and paralinguistic features when both used with unigram model and SVM respectively. Further it has been given that while using lexical features there is confusion between (fear, sadness) and (fear, anger) [10].

Table 1. Difference between Lexical and Paralinguistic features

Lexical Features	Paralinguistic Features
Gives 78% accuracy. Relief (90%) and Fear (86%) correct detected. It uses unigram model. Sadness recognition is low.	Gives 60% accuracy. Fear is best detected with paralinguistic features. It uses SVM classifier. Anger recognition is low.

Emotion detection on the basis of acoustic features and prosodic features is faster whereas lexical features are slow when it comes to detect emotion in speech.

A combination of lexical, prosodic, dialog acts and contextual features can also be used which shows the following accuracy as shown in the table 2.

Table 2[4]: Classification accuracy of user emotional state given different feature sets as well as relative performance improvement over the baseline

Features Set used	Accuracy	Improvement over BASELINE
BASELINE	73.1%	0.0%

LEX+PROS	76.1%	4.1%
LEX+PROS+DA	77.0%	5.3%
LEX+PROS+DA+CONTEXT	79.0%	8.1%

4. MODELS (CLASSIFIERS) USED

The performance of emotion recognition system relies heavily on the type of classifier used for classification of the features selected after extracting it. The field of artificial Intelligence provides various types of classifiers which can be used as a component in emotion recognition system. K Nearest Neighbor, Kernel Regression, Support Vector Machine, Neural Network, Bayes Classification, Decision Trees and many more.

All the classifiers can be used alone as a component or can be combined with another classifier to form a component of emotion recognition system. The classifiers have different working style and different performance is given by each classifier.

4.1 Support Vector Machine

One of the important classifiers is the support vector machine. SVM are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. SVM classifiers are mainly based on the use of kernel functions to nonlinearly map the original features to a high dimensional space where data can be well classified using a linear classifier. SVM classifiers are widely used in many pattern recognition applications and shown to outperform other well-known classifiers. SVM has shown to have better generalization performance than traditional techniques in solving classification problems. The accuracy of the SVM for the speaker independent and dependent classification are 75% and above 80% respectively [11].

4.2 K Nearest Neighbor

The k-NN classifier compares a given target instance with the k training instances that are the most similar or closest to it. There are a variety of

metrics used to measure similarity and the Euclidian distance metric is frequently used. The target instance is assigned to the class to which the majority of these nearest neighbors belong. There is no consensus on which value of k should be used in the case of emotion recognition from speech. Different researchers proposes values from $k = 1$ to $k = 20$. Usually the value of k is found by trial and error, different values are taken and the performance of these classifiers is compared.

4.3 Artificial Neural Network

Another common classifier, used for many pattern recognition applications is the artificial neural network (ANN). An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. They are known to be more effective in modelling nonlinear mappings. Also, their classification performance is usually better than HMM and GMM when the number of training examples is relatively low. Almost all ANNs can be categorized into three main basic types: MLP, recurrent neural networks (RNN), and radial basis functions (RBF) network. The classification accuracy of ANN is fairly low compared to other classifiers. The ANN based classifiers may achieve a correct classification rate of 51.19% in speaker dependent recognition, and that of 52.87% for speaker independent recognition.

4.4 Hidden Markov Models

An HMM has been studied long time by researchers for speech emotion recognition, as has advantage on dynamic time warping capability. Moreover, it has been proved useful in dealing with the statistical and sequential aspects of the speech signal for emotion recognition. However, the classify property of HMM is not satisfactory. Hidden Markov models (HMMs) are the best known for its temporal classification. They have various applications in areas such as speech, handwriting and gesture recognition. A hidden Markov model is a slight variant of a finite state machine but HMM is not deterministic and they do both transition and emit under a probabilistic model.

4.5 Decision Tree

Decision tree learning uses a decision tree as a prescient model which maps perceptions about a thing to decisions about the thing's target esteem. It is one of the prescient demonstrating methodologies utilized as a part of insights,

information mining and machine learning. More illustrative names for such tree models are arrangement trees or relapse trees. In these tree structures, leaves speak to class marks and extensions speak to conjunctions of peculiarities that prompt those class names. In choice investigation, a choice tree can be utilized to outwardly and unequivocally speak to choices and choice making. In information mining, a decision tree portrays information however not choices; rather the ensuing order tree can be a data for decision making.

4.6 Bayes Classification

A naive Bayes classifier accepts that the vicinity or nonappearance of a specific peculiarity is random to the vicinity or unlucky deficiency of some other gimmick, given the class variable. Case in point, an apples and oranges may be thought to be an apple on the off chance that it is red, round, and around 3" in width. A naive Bayes classifier considers each of these peculiarities to help freely to the likelihood that this tree grown foods is an apple, paying little mind to the vicinity or nonattendance of alternate gimmicks.

For a few sorts of likelihood models, naive Bayes classifiers can be prepared proficiently in an administered learning setting. In numerous functional applications, parameter estimation for naive Bayes models utilizes the system for greatest probability; at the end of the day, one can work with the naive Bayes model without tolerating Bayesian likelihood or utilizing any Bayesian systems.

4.7 Decision Tree

The kernel regression is a non parametric technique in statistics to estimate the conditional expectation of a random variable. The objective is to find a non-linear relation between a pair of random variables X and Y .

In any nonparametric regression, the conditional expectation of a variable Y relative to a variable X may be written as

$$E(Y|X) = m(X) \quad \dots (1)$$

where m is an unknown function.

5. FIGURES/CAPTIONS

In previous section various classifiers have been mentioned which can be used as component in Emotion Recognition System. This section further

talks about those classifiers highlighting the differences between performances of these classifiers and which emotion do they detect properly and further this section also tells when classifiers are used in combination then how do they perform and what emotion do they detect together.

While using neural network and support vector machine one notable difference is found and that is when there is sufficient training data then neural network performs better and support vector machine outperforms other classifiers in the scarcity of training data and the same can be verified from table 3.

Table 3[9]: Accuracy in recognizing hot anger versus neutral speech using 37 features

Classifier	7/1 actor-split	6/2 actor-split
Neural Net	94.00%	86.90%
SVM	90.90%	90.79%
3NN	87.88%	81.60%
C4.5	63.65%	76.32%

Further it was also noted that for neural network and 3NN less features are sufficient and for support vector machine performance degrade as features reduced.

The above mentioned differences were about detecting hot anger but when happiness is also included then there is confusion between happiness and anger and confusion between sadness and boredom. So as to eliminate this confusion groups can be formed in which one group consists of happiness and hot anger and another group consist of sadness and boredom. Then using support vector machine and decision tree gives different accuracy. Accuracy of 77% is achieved with support vector machine and accuracy of 81.8% is achieved with decision trees. This can be verified by [9].

This was all about when using single classifier as a component but a combination of classifiers can easily be used like when Bayes classification,

kernel regression and KNN is combined together then accuracy is achieved 60-65% and it can detect sadness, anger, happiness and fear. Another combination includes linear discrimination, KNN and support vector machine which can detect negative and non negative emotions with 75% accuracy [5].

Real time emotion recogniser using neural network can detect agitation and calm emotions with 77% accuracy [4]. Further support vector machine with input as prosodic features behaves as binary classifier gives 73% accuracy and can detect anger, happiness, sadness and neutral emotions.

6. RELATED WORK

The first studies that were conducted were not so much trying to get an efficient machine recognition device, but rather were searching for general qualitative acoustic correlates of emotion in speech. More recently, the increasing awareness that affective computing had an important industrial potential ([9]) pushed research towards the quest of performance in automatic recognition of emotions in speech.

Dellaert *et al.* [3] compared three classifiers: the maximum likelihood Bayes classification, kernel regression, and k nearest neighbor (K-NN) methods with particular interest in sadness, anger, happiness, and fear. They used features from the pitch contour. An accuracy of 60%-65% was achieved. Lee *et.al.* [3] used linear discrimination, k-NN classifiers, and support vector machines (SVM) to distinguish two emotions: negative and non-negative emotions where they reached a maximum accuracy of 75%.

Petrushin developed a real time emotion recognizer using neural networks for call center applications, and achieved 77% classification accuracy in two emotions ("agitation" and "calm") using eight features chosen by a feature selection algorithm. Tato *et.al.* [2] discussed techniques that exploit emotional dimension other than prosody. Their experiments showed how "quality features" (based on formant analysis) are used in addition to "prosody features" (pitch and energy) to improve the classification of multiple emotions. The quality features were mostly speaker-dependent and hence cannot be used in IVRs.

Yu *et.al.* [1] used SVMs for emotion detection. They built classifiers for four emotions: anger,

happy, sadness, and neutral. Since SVMs are binary classifiers, their recognizers worked on detecting one emotion versus the rest. An average accuracy of 73% was reported.

[6] The recognition of emotional states from speech is a research topic with a long history as it is connected with the general research on the acoustical correlates of affective speech. In the following short review we concentrate on studies dealing with telephone data. Most classification algorithms for the detection of anger are based on a three-step approach: First, a set of acoustic, prosodic, or phonotactic features are calculated from the input speech signal. In a second step, different classification algorithms, e.g. Gaussian Mixture Models, Artificial Neural Networks, Support Vector Machines, , other vector clustering algorithms like k nearest neighbor, or linear discriminant analysis, are applied to derive a decision whether the current dialog turn is angry or not angry. Finally, post-processing technologies can be utilized for consideration of time dependencies of subsequent turns or for combination of the results of different classifiers. All these algorithms heavily depend on the availability of suitable acoustic training data that should be derived from the target application. With respect to the features that are used to classify the speech data, mainly prosodic features, often in conjunction with lexical based and/or dialog related features, were investigated, and while newer studies also include spectral features derived from Mel Frequency Cepstral Coefficients. Several studies have shown that the inclusion of dialog features can help to enhance the classification accuracy.

Walker et al. [6] Measure dissatisfaction with a questionnaire and train a prediction model for the user ratings by applying linear regression, using interaction parameters (e.g. dialog length) as predictors. In other studies, unsuccessful dialogs (e.g. because of user hang-up) are predicted from interaction parameters describing the first N dialog turns. Attempts were made to incorporate emotion recognition in such predictions. Direct relations between audio features and user ratings are analyzed.

7. CONCLUSION

In this survey paper, most recent work done on Emotion Recognition in Spoken dialog is

discussed. Various types of features that can be extracted are discussed and several types of classifier used are discussed and further a comparison is done between features extracted and classifiers used respectively. Success of emotion recognition is dependent on appropriate feature extraction as well as proper classifier selection from the sample emotional speech. It can be seen that Integration of various features can give the better recognition rate. Classifier performance is needed to be increased for recognition of speaker independent systems. The application area of emotion recognition from speech is expanding as it opens the new means of communication between human and machine. It is needed to model effective method of speech feature extraction so that it can even provide emotion recognition of real time speech.

8. REFERENCES

- [1] I. Shafran and M. Mohri, "A comparison of classifiers for detecting emotion from speech," In Proc. ICASSP, Philadelphia, 2005.
- [2] D. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelone, Spain, 2004.
- [3] P. Oudeyer, "Novel useful features and algorithms for the recognition of emotions in human speech," in *Proceedings of Speech Prosody*, Aix-en-Provence, France, 2002, pp. 547-550.
- [4] Liscombe, J., Riccardi, G., and Hakkani-Tür, D., "Using Context to Improve Emotion Detection in Spoken Dialogue Systems", in Proc. Interspeech, 2005.
- [5] Forbes-Riley, K. & Litman, D. (2004). Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. *Proceedings of HLT/NAACL*.
- [6] Felix Burkhardt, Markus van Ballegooy, Klaus-Peter Engelbrecht, Tim Polzehl and Joachim Stegmann, "Emotion Detection in Dialog Systems: Applications, Strategies and Challenges," *Affective Computing and Intelligent Interaction and Workshops*, 2009. ACII 2009.
- [7] Laurence Devillers, Lori Lamel and Ioana Vasilescu, "Emotion Detection In Task-

Oriented Spoken Dialogs,” in Proceedings of Multimedia and Expo, 2003. ICME '03.

- [8] *Laurence Devillers and Laurence Vidrascu*, “Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs,” *In proceeding of: INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006.*
- [9] *Sherif Yacoub, Steve Simske, Xiaofan Lin and John Burns*, “Recognition of Emotions in Interactive Voice Response Systems,” 8th European Conference on Speech Communication and Technology, 1-4 September 2003, Geneva, Switzerland.
- [10] *Hua Ai, Diane J. Litman, Kate Forbes-Riley, Mihai Rotaru, Joel Tetreault, Amruta Purandare*, “Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs,” *in proceedings of Interspeech 2006.*
- [11] *Dipti D. Joshi and Prof. M. B. Zalte*, “Speech Emotion Recognition: A Review,” in *International Journal of Speech Technology*, June 2012, Volume 15, Issue 2, pp 99-117.
- [12] *Dongrui Wu¹, Thomas D. Parsons¹ and Shrikanth S. Narayanan*, “Acoustic Feature Analysis in Speech Emotion Primitives Estimation,” *In proceeding of: INTERSPEECH 2010.*