Comparative Analysis of Emotion Recognition Using Facial Expressions and Speech

Niranjan Samudre[1], Chandansingh Rawat[2]

[1]*(Research Scholar, Pacific Academy of Higher Studies and Research University, Udaipur, Rajasthan*
[2]*(Associate Professor, VIvekanand Education Society's Institute of Technology, Mumbai*

Abstract: *The projected solution is aimed toward raising the interaction among humans and computers, therefore permitting effective human-computer intelligent interaction. The system is in a position to acknowledge six emotions (anger, boredom, disgust, fear, happiness and sadness). This set of emotional states is wide used for emotion recognition functions. In this paper we present the design of an artificially intelligent system capable of emotion recognition trough facial expressions. Three promising neural net-work architectures are customized, trained, and subjected to various classification tasks, after which the best performing network is further optimized. This paper analyzes the strengths and the limitations of systems based only on facial expressions or acoustic information.*

I. Introduction

Recently there has been growing interest to improve Human-computer interaction (HCI) means computers should interact to the humans in day to day life .In this context recognizing people emotional state and giving suitable feedback may play a crucial role. As a consequence, emotion recognition represents a hot analysis space in each industry and academic field. Usually emotion recognition based on facial or voice options. Most techniques process visual data and search for general patterns present in human faces. Face recognition can be used for surveillance purposes by law enforcers as well as in crowd management. However, the most promising applications involve the humanization of artificial intelligent systems. If computers are able to keep track of the mental state of the user, robots can react upon this and behave appropriately. Emotion recognition therefore plays a key-role in improving human-machine interaction.

This paper proposes a solution, designed to be employed in a smart phone Environment able to capture emotional state of a person starting from registration of speech signals in the surrounding obtained by mobile devices like smart phones. People can use their voice to give command to car, cell phone, computer TV and many electrical devices. Hence build the device perceives human feeling and provides a stronger experience of interaction becomes a very interesting challenge.

The body of work on detecting emotion in speech is kind of restricted. Currently, researcher's area unit still debating what options influence the popularity of feeling in speech. There is conjointly substantial uncertainty on the most effective algorithmic program for classifying emotion, and which emotions to class together.

In this paper we tend to attempt to address these problems. We use K-Means and Support Vector Machines (SVMs) to classify opposing emotions.

II. Literature survey

A suitable alternative of speech information (corpora) plays a very vital role within the field of have an effect on detection. A context rich emotional speech database is preferred for a good emotion recognition system. Mainly three types of corpora are used for developing a speech system they are1) Elicited emotional speech database: This type of corpora is collected from speaker by creating artificial emotional situation. Advantage of this kind of information is that it's very close to the natural information however there are some issues also. All emotions might not be accessible and if the speaker is alert to that they're being recorded, then the emotion expressed by him could also be artificial.2) Actor based speech database: This type of speech data set collected from professional and trained artists. Collecting of these form of data are very easy and a wide variety of emotion are accessible within the corpora .The main downside of this kind of information are it is episodic in nature and it's considerably artificial in nature.3) Natural speech database: this type of database created from real world data. These forms of information are completely natural and extremely helpful for real world emotion recognition. The problem is that, all emotions may not be present and it consists of background noise. Implementation of emotional speech database depends on objective of the analysis. For efficient affect detection system, it is important that the corpora must consist of real and natural emotional speech spoken by a large number of male and female persons. Though there are different corpora exists, there is no standard, globally approved speech database available for emotion recognition. In Indian context, there are different speech corpora for speaker recognition however there's a scarcity of corpora for emotion recognition.

A key feature in human interaction is the universality of facial expressions and body language. Already in the nineteenth century, Charles Darwin published upon globally shared facial expressions that play an important role in non-verbal communication [3]. In 1971, Ekman & Friesen declared that facial behaviors are universally associated with particular emotions [5]. Apparently humans, but also animals, develop similar muscular movements belonging to a certain mental state, despite their place of birth, race, and education. Hence, if properly modeled, this universality can be a very convenient feature in human-machine interaction: well trained systems can understand emotions, independent of who the subject is one should keep in mind that facial expressions are not necessarily directly translatable into emotions, nor vice versa. Facial expression is additionally a function of e.g. mental state, while emotions are also expressed via body language and voice [6]. More elaborate emotion recognition systems should therefore also include these latter two contributions.

Readers interested in research on emotion classification via speech recognition are referred to Nicholson et al. [14]. As a nal point of attention, emotions should not be con-fused with mood, since mood is considered to be a long-term mental state. Accordingly, mood recognition often involves longstanding analysis of some-one's behavior and expressions, and will therefore be omitted in this work.

III. Proposed methodology

A) Emotion Recognition Using Speech Signals:

In this paper we intend to propose an effective method for speaker emotion recognition. The proposed technique has three main phases namely, feature extraction, feature selection and emotion recognition. Initially we select the input signal from the speech signal database. From the fetched input signal the features are extracted in the first phase. In second phase, the selected features are optimally selected by means of optimization algorithm. After feature selection, the selected features are fed to the emotion recognition technique for recognizing various speaker emotions such as Anger, Surprise, Fear, Happiness, Sadness, Disgust and Neutral state. The overall flow diagram is shown in fig.1

The overall procedure of the proposed work is classified into three important steps such as,

1. Feature Extraction
2. Feature Selection
3. Emotion Recognition

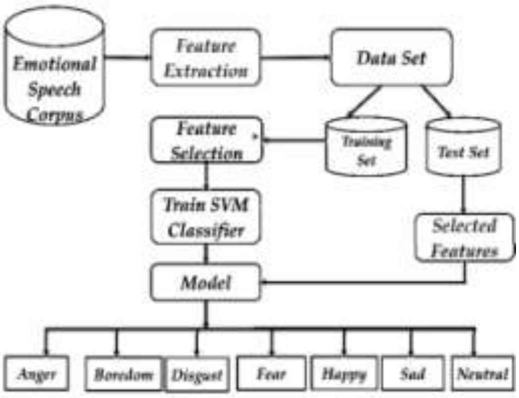

Figure1. Block Diagram of Speech Emotion Recognition

Speech recognition with support vector machines in a hybrid system:

While the temporal dynamics of speech are often represented very efficiently by Hidden markov Models (HMMs), the classification of speech into single speech units (phonemes) is sometimes done with Gaussian mixture models which do not discriminate well. Here, we use Support Vector Machines (SVMs) for classification by integrating this method in a HMM-based speech recognition system. In this hybrid SVM/HMM system we tend to translate the outputs of the SVM classifiers into conditional probabilities and use them as emission probabilities in an exceedingly HMM-based decoder. SVMs are terribly appealing because of their association with statistical learning theory. They have already shown excellent classification results indifferent fields of pattern recognition .We train and test our hybrid system on the darpa Resource Management (RM1) corpus. Our results show better performance than HMM-based decoder using Gaussian mixtures.

Processing of the derived features:

The features are freed of their mean value and normalized to their standard deviation. They are classified by single state HMMs (GMM), which are able to approximate the probability distribution, function of each derived feature by means of a mixture of Gaussian distributions. Up to four mixtures have been used. No further gain could be observed exploitation more than these. Each emotion is sculptured by one GMM in our approach. The maximum probability model are considered because the recognized emotion at a time throughout the recognition process.

Classification of System

In the speech emotion recognition system after calculation of the features, the most effective features are provided to the classifier. A classifier recognizes the emotion within the speaker's speech utterance. Various kinds of classifier have been proposed for the task of speech emotion recognition. Gaussian Mixtures Model (GMM), K-nearest neighbours (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN), etc. are the classifiers used in the speech emotion recognition system. Each classifier has some advantages and limitations over the others. In speech recognition system like isolated word recognition and speech emotion recognition, hidden markov model is generally used; the main reason is its physical relation with the speech signals production mechanism.

In speech emotion recognition system, HMM has achieved great success for modelling temporal information in the speech spectrum. The HMM is doubly stochastic process consist of first order markov chain whose states are buried from the observer. HMM has the important advantage that the temporal dynamics of speech features can be caught second accessibility of the well-established procedure for optimizing the recognition frame work. The main downside in building the HMM based mostly recognition model is that the features choice method. Because it's not enough that features carries info concerning the emotional states, however it should match the HMM structure further.

Usually, in the literature of the field, a Support Vector Machine (SVM) is used to classify sentences. SVM is a relatively new machine learning algorithm introduced by Vapnik [11] and derived from statistical learning theory in the 90s.The main idea is to transform the original input set into a high dimensional feature space by using a kernel function and then, to achieve optimum classification in this new feature space, where a clear separation among features obtained byte optimal placement of a separation hyper plane under the precondition of linear reparability. Differently from the previously proposed approaches two different classifiers, both kernel-based Support Vector Machines (SVMs), have been employed in this paper. Neural networks are chosen for the solution because a basic formula cannot be devised for the problem. The neural networks are also quick to respond which is a requirement as the emotion should be determined almost instantly. The training takes a long time but is irrelevant as the training is mostly done off-line. Deep learning has been applied to SER in prior work, as discussed. To the best of our knowledge, our work provides the first empirical exploration of various deep learning formulations and architectures applied to SER. As a result, we report state-of-the-art results on the popular Interactive Emotional Dyadic Motion Capture (IEMOCAP) database (Busso et al., 2008) for speaker independent SER

B) Emotion Recognition Using Facial Expressions:

Dataset

Neural networks and deep networks in particular, are known for their need for large amounts of training data. Moreover, the choice of images used for training is responsible for a big part of the performance of the eventual model. This implies the need for a both high qualitative and quantitative dataset. For emotion recognition, several datasets are available for research, varying from a few hundred high resolution photos to tens of thousands smaller images.

Networks

The networks are programmed with use of the TFLearn library on top of TensorFlow, running on Python. This environment lowers the complexity of the code, since only the neuron layers have to be created, instead of every neuron. The program also pro-vides real-time feedback on training progress and ac-curacy, and makes it easy to save and reuse the model after training. More details on this framework can be found in reference [6].

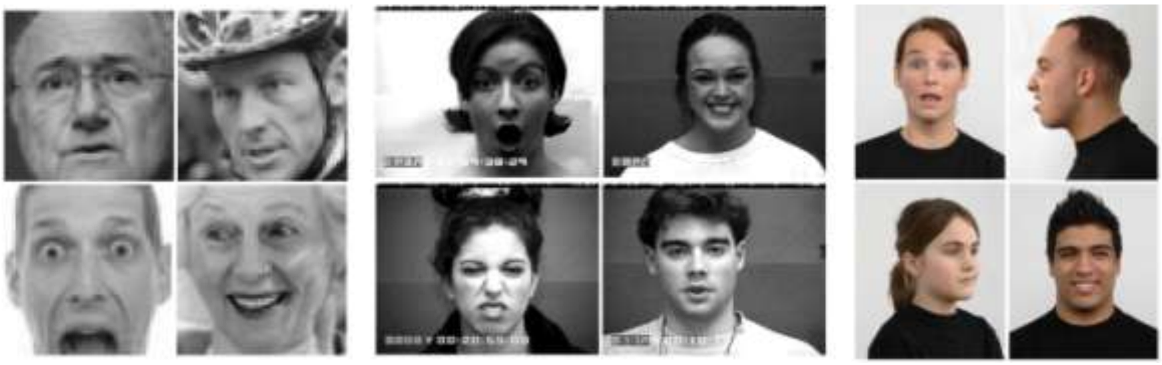

Figure 2: Samples from the FERC-2013 (left), CK+ (center), and RaFD (right) datasets

IV. Results

A) Emotion Recognition Using Speech Signals:

In this paper, an HMM based approach to emotion recognition has been presented. Results a good accuracy confirm both the usefulness of the approach and the convenience of the low level features used. Given the reduced scope of the scenario considered it may be arguable if the results can be generalized to other speakers and/or languages, yet we believe that the results achieved are encouraging: at least, they show the usefulness of the approach for multi-speaker emotion recognition besides, we believe that this is a good baseline for more ambitious tasks. Since the introduced approaches tend to strongly depend on the speaker, no cross-speaker evaluation results are presented.

Classification	Happiness	Neutral	Boredom	Sadness	Anger
Happiness	99.7	0	0	0	1.7
Neutral	2.9	90.5	1.5	0	0
Boredom	0	9.5	91.0	11.4	0
Sadness	0	0	6.0	88.6	0
Anger	4.5	0	0	0	90.1

B) Emotion Recognition Using Facial Expressions:

All networks are trained for 60 epochs with the data mentioned in section B and table 2 show various details of the training process and the nal model. For network A, the nal accuracy on the validation data is around 63%. Already after 10 epochs, the accuracy raised above 60%, indicating quick learning capabilities. Furthermore it is note-worthy that adjusting the filter dimension did not have a big influence on the accuracy, though it has on the processing time. This means that fast models can be made with very reasonable performance.

Table 2: Details of the trained networks

Network	Accuracy		Size
	Validation	RaFD	
A	63%	50%	Small
B	53%	46%	Large
C	63%	60%	Medium

Surprisingly, the second, much larger, network learns quickly as well, but converges to an accuracy of about 54%. Apparently reducing the network size breaks down the promising performance of the original network more than expected. Together with the much higher computational intensity, and therefore slower live performance, this model is not a worthy challenger of the other two architectures.

Final model

An overview of its architecture is shown in figure 2. The accuracy seems to still increase in the last epochs. We therefore will train the network for 100 epochs in the nal run, to make sure the accuracy converges to the optimum. In an attempt to improve the nal model even more, the network will be trained on a larger set than the one described previously. Instead of 9000 pictures, training will be done with 20000 pictures from the FERC-2013 dataset. Newly composed validation (2000 images) and test sets (1000 images) from the FERC-2013 dataset are used as well, together with the well-balanced RaFD test set from the previous experiment.

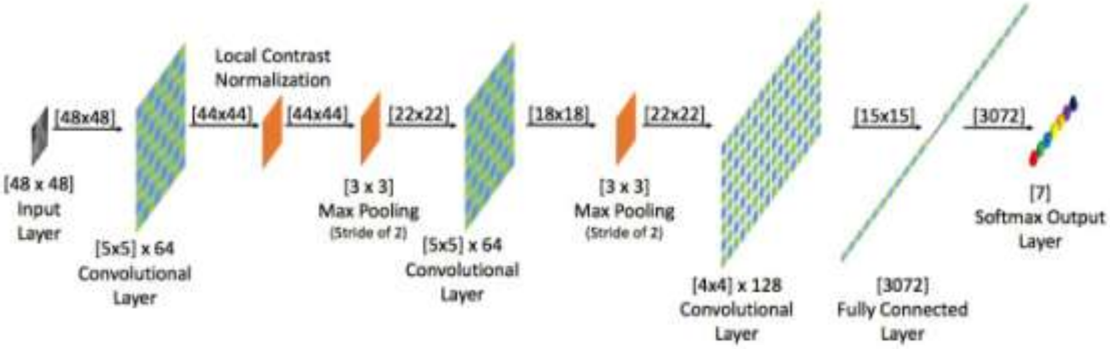
Local Contrast Normalization
[48x48]
[44x44]
[44x44]
[22x22]
[18x18]
[22x22]
[15x15]
[3072]
[48 x 48] Input Layer
[5x5] x 64 Convolutional Layer
[3 x 3] Max Pooling (Stride of 2)
[5x5] x 64 Convolutional Layer
[3 x 3] Max Pooling (Stride of 2)
[4x4] x 128 Convolutional Layer
[3072] Fully Connected Layer
[7] Softmax Output Layer

Figure 2: Overview of the network architecture of the nal model

The accuracy rates of the nal model are given in table 3. On all validation and test sets the accuracy was higher than during previous runs, underlining that more data and longer training can improve the performance of a network. Given that the state-of-the-art networks from previous research obtained about 67% on test sets, and keeping in mind our limited re-sources, the results are in fact pretty good. Notable is the accuracy on the RaFD test set, which contains completely different pictures than the training data. This illustrates the powerful generalizing capabilities of this nal model.

Table 3: Accuracy of the networks

Network	FERC-2013		RaFD
	Validation	Test	
A	63%		50%
B	53%		46%
C	63%		60%
Final	66%	63%	71%

To see how the model performs per emotion, a table is generated. Very high ac-curacy rates are obtained on happy (90%), neutral (80%), and surprised (77%). These are in fact the most distinguishable facial expressions according to humans as well. Sad, fearful, and angry are often misclassified as neutral too though. Apparently these emotions very look alike. The lowest accuracy is obtained on sad (28%) and fearful (37%). Finally it is noteworthy that even though the percentage of data with label disgusted in the training set is low, the classification rate is very reasonable.

V. Conclusion

The proposed system, able to recognize the emotional state of a person starting from audio signals registrations, is com-posed of two functional blocks: Gender Recognition (GR) and Emotion Recognition (ER). The former has been implemented by a Pitch Frequency Estimation method, the latter by two Support Vector Machine (SVM) classifiers (fed by properly selected audio features), which exploit the GR subsystem output. Various deep learning architectures were explored on a Speech Emotion Recognition (SER) task. Experiments conducted illuminate how feed-forward and recurrent neural network architectures and their variants could be employed for paralinguistic speech recognition, particularly emotion recognition. Convolution Neural Networks (Convents) demonstrated better discriminative performance compared to other architectures.

Reference

[1]. Kamran Soltani and Raja Noor Ainon. Speech emotion detection based on neural networks. In 9th International Symposium on Signal Processing and its Applications, 1 4244-0779-6/07, IEEE, 2007.

[2]. Jouni Pohjalainen and Paavo Alku. Multi-scale modulation filtering in automatic detection of emotions in telephone speech. International Conference on Acoustic, Speech and Signal Processing, 980-984, 2014.

[3]. https://www.irjet.net/archives/V3/i4/IRJET-V3I460.pdf

[4]. http://www.ijesit.com/Volume

[5]. https://www.sciencedirect.com/science/article/pii/S089360801730059X

[6]. TFlearn. T earn: Deep learning library featuring a higher-level api for tensorflow. URL http://tflearn.org/.