JSc EUSL(2017) Vol. 8 No. 1, p 13-22

ISSN 1391-586X; © 2017 published by Faculty of Science, Eastern University, Sri Lanka

Speech Emotion Classification Analysis using Short-term Features

Thirukumaran, S* and Archana, A.F.C

Department of Physical Science, Vavuniya Campus of the University of Jaffna

Abstract

Speech is an auditory signal produced from the human speech production system used to express ourselves. In this era, speech signals are also used in biometric identification technologies and interacting with machines, so that it can give different response. Emotion recognition is not a new topic and researches and applications exist using different methods to extract specific features from the speech signals. This paper presents a classification analysis of emotional human speech only with short term processing features of the speech signals using artificial neural network based approach. Speech rate, pitch and energy are the most basic features of speech signal but they still have significant differences between emotions such as angry and sad. The most common way to analyze the speech emotion is to extract important features which are related to different emotion states from the voice signal. In the speech pre-processing phase, the samples of four basic types of emotional speeches sad, angry, happy, and neutral are used. Then feed those extracted short term features into the input end of the classifier and obtained different emotions at the output end. 23 short term audio signal features of 40 samples of two frames are selected and extracted from the speech signals to analyze the human emotions. These derived data along with their related emotion target matrix are fed to test and design the classifier using artificial neural network pattern recognition algorithm. The confusion matrix is generated to analyze the performance results. The overall correctly classified results for two times trained network is 73.8 %, while increasing the training times to ten, 95 % of the emotions are correctly classified. The accuracy of the neural network system is improved by multiple times of training. The overall system provides a reliable performance and correctly classifies more than 85 % for the new non-trained dataset.

Keywords: Confusion Matrix, Neural Network, Short-term Features, Speech Emotions

*Corresponding author: thiruvks@gmail.com

Introduction

In human interaction, emotions play important role. Human beings possess and express emotions in everyday interactions with others. When we talk about communication, it is striking that *what* we are talking, but it is more consequential that *how* we are expressing. There may be different types of sign that indicate emotions. In communication between human–human, emotions can be expressed in terms of verbal or facial. Speech signals contain different types of information including not only the information about message but also speaker's identification, emotions identification and identification of language and so on.

One important aspect of human-computer interaction is to train the system to understand human emotions through voice. People can use their voice to order commands to many electrical devices such as car, smart phone, computer, etc. Hence make the devices understand human emotions and give a better experience of interaction. Typically, the most common way to recognize speech emotion is to first extract important features that are related to different emotion states from the voice signal (e.g.: energy is an important feature to distinguish happy and sad), then feed those features to the input end of a classifier and obtain different emotions at the output end.

Speech analysis can be done either in time domain or in frequency domain using the short term or mid-term processing of speech. Short-term processing features divide the speech signals into short analysis segments which are isolated and processed with fixed properties. Mid-term processing features divide the audio signal into mid-term segments and which are used to compute the statistical values. The important problems in this emotion classification analysis is only using the short term features of the speech signals and analyze the performance of the neural network classifier.

Review of previous work

There have been many studies about speech recognition in recent years, different features as well as different classification methods have been used, i.e. Nogueiras *et al*., used Hidden Markov Models (HMM) to recognize emotions from the features pitch and energy where the accuracy is 80 % [1]. B. Schuller *et al*., used four different classification methods to compare their performance and to recognizing emotions [2] using HMM [3, 4], Naive Bayes Classifier [5], and Decision Tree Classifier [6]. Alexandros Georgogiannis, Vassilis Digalakis introduced one of speech's features *Teager MFCC*, which can also work in noisy environment [7]. Singh *et al*., described database

development for Hindi Hybrid word and main focus is to analyze database using end point detection [8]. Mina *et. al*. reported an effort towards automatic recognition of emotional states from continuous Persian speech by building database of emotional speech in Persian. The resulting average accuracy was about 78 % [9]. Valery A. Petrushin described an experimental study on emotion recognition by developing 140 utterances per emotional state. Each utterance was recorded using close talk microphone. Vocal energy, frequency, formats were used for feature extraction using neural network. He presented result with accuracy of 61.4 % for happiness, 72.2 % for anger, 68.3 % for normal [10].

Speech signals are produced from a time varying vocal tract system with time varying excitation. Due to this reason, the speech signals are non-stationary in nature but, in blocks of short time speech signals are viewed as a stationary [11]. Spectral and prosodic features are used for speech emotion recognition because both of these features contain the emotional information. Fundamental frequency, loudness, pitch and speech intensity and glottal parameters are the prosodic features used to model the different emotions. The Mel-Frequency Cepstrum Coefficients (MFCC) is an accurate representation of short time power spectrum of a sound [12]. The audio signals are broken into possibly overlapping frames and a set of features is computed per frame. These short analysis segments are called analysis frames and overlap in one another [13].

Materials And Methods

To this classification, samples of recorded English speech signals of four emotions are used from the Emotional Prosody speech and Transcripts in the Linguistic Data Consortium (LDC) Dataset, in which actors and actresses perform different emotions. The speech samples for four emotions categories in the dataset contain both male and female speakers. Samples are taken from the speech and the analog signals are converted to digital signals. Each speech sentence is normalized to ensure that all the sentences are in the same volume range. At the last process uses the segmentation to separate the signal into frames so that the speech signal can maintain its characteristics in short duration. Each sample is between one second in length and separates each sample into two overlapping frames with 30 ms segments. Usually the speech signal properties change slowly with time, hence allowing the examination of short time window of speech to extract parameters. In general, the time-domain short-term audio features are extracted directly from the samples of the audio signal. Typical examples of the short-term features

are the short-term energy, short-term Zero-Crossing Rate (ZCR), short-term entropy of energy, short-term harmony, Mel-Frequency Cepstrum Coefficients (MFCC), Spectral entropy, Spectral flux, Spectral entropy, Spectral centroid and the spectral spread [14].

These features are extracted from the speech signals to create and load input data and target data. A 23×80 matrix is used to create input data which indicates 23 features of 40 samples of two frames. Here, 13 short-term feature values for MFCC, two feature values for spectral centroid and one different value for each of the other eight features are extracted from the audio signals and the values are stored in a vector as an input data. Target data is 4×80 matrix which indicates the four emotion states for these 40 samples of two frames. After importing those data, next step is to randomly divide the percentage of input data into three categories namely training, validation and testing. The training set is used to fit the parameters of the classifier i.e., to find the optimal weights for each of the features. The validation set is used to tune the parameters of a classifier that is to determine a stop point for training set. Finally, the test set is used to test the final model and estimate the error rate.

The input vectors and target vectors are randomly divided into three data sets as follows:

(i) 70 % is used for training.
(ii) 15 % is used for validation to measure network generalization, and to halt training when generalization stops improving.
(iii) The last 15 % is used for testing and it has no effect on training and provides an independent measure of network performance during and after training.

A total of 80 sample data have been split by 56 of which are used in the training session, the 12 for the validation and 12 for the testing. The training, validation and test data sets are mutually exclusive in each run.

The back propagation neural network model is selected to classify the emotions since it is the most significantly used model for emotion classification and back propagation is better than the other neural network models. We can infer that when handling noise and multiple inputs of data, back propagation performs better than the pattern recognition method SOM. Another method called LVQ is an excellent for classification, but when handling noise, it is a little bit worse than back propagation method [15]. Finally, train the system to classify the emotions according to the input and target matrices. Let the system trains several times and after that Cross-entropy together with error rate would indicate how good the results are.

Results and Discussion

The emotions used in the samples are *happy*, *sadness*, *angry* and *neutral*. The below sections contain the corresponding classification results.

Classifier

The network used in the experiment is composed of three layers: the input layer, the hidden layer and the output layer. The input layer takes the 23 feature values for 40 samples of two frames. The hidden layer has 30 nodes and uses a sigmoid transfer function. The number of nodes in the output layer depends on how many emotional categories to recognize. For this research a resilient back propagation training algorithm in the network is used. The advantage of this training algorithm is that it can eliminate harmful effects of the magnitudes of the partial derivatives.

Performance analysis

The below description is the classification result for the trained Neural Network classifier. Two times trained ANN emotions classification shown in Figure 1 and Figure 2. Four emotions are listed together with the error rate for each row and column representing for target class and output class respectively.

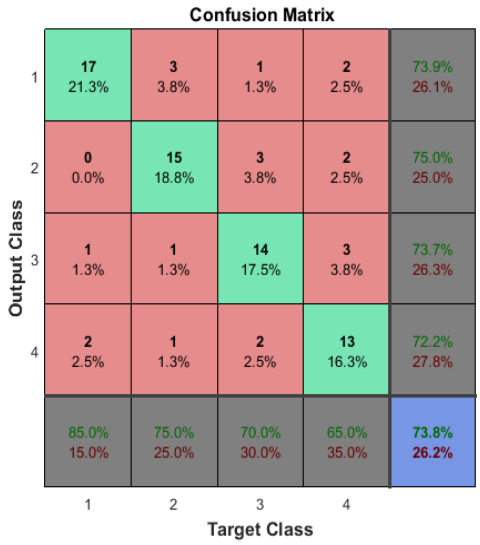

Figure 1: Over all Confusion matrix for two time trained ANN

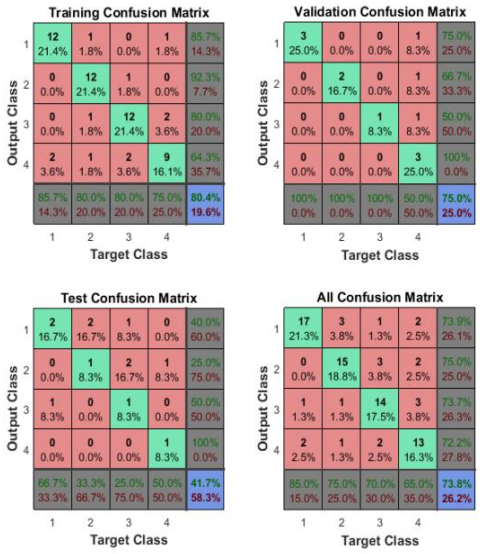

Figure 2: Three set of data Confusion matrix for two time trained ANN

Therefore, the number in cell "1" stands for how many sad speeches have been classified into the sad output. Cell "2" shows how many angry speeches have been misclassified into the class sad. The performance of classifier improved by increasing the number of training. After the training of the network into ten times, the overall emotions classification result shown as a confusion matrix in Figure 3 and the results of the three data sets shown in Figure 4. Classification results of all four sets of emotions and an overall result are given in tables 1 and 2 below. This gives a clear idea about the classification system.

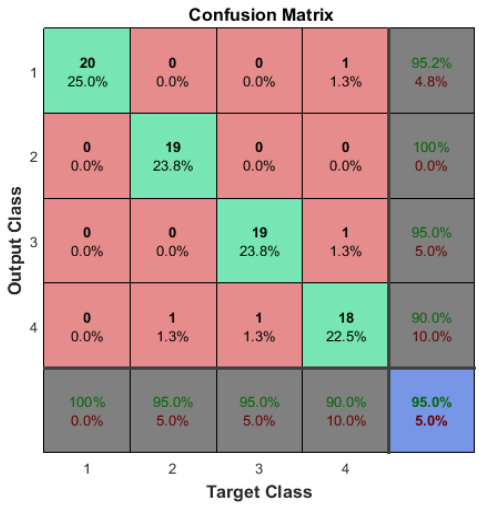

Figure 3: Over all Confusion matrix for ten time trained ANN

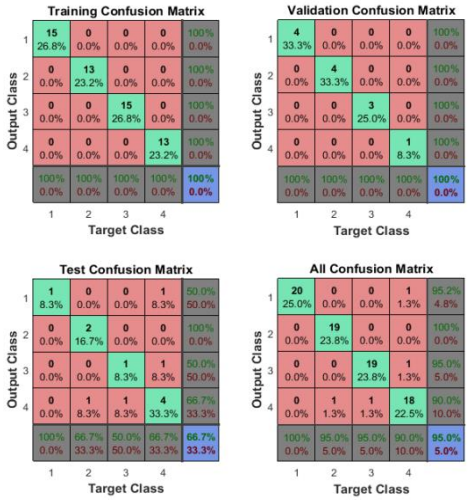

Figure 4: Three set of data Confusion matrix for ten time trained ANN

Overall matrix in Figure 1, 17 of sad speeches have been put into the correct output as sad, one sad speech is misclassified into happy speech and two of the sad speeches are misclassified into neutral speech. For the next class, 15 of the angry speeches are classified correctly. Three of angry speeches is misclassified into the sad output, one of them are misclassified into the happy output and one of them into neutral speech. For happy speeches, 14 of speeches are correctly classified and six of the speeches are misclassified output. At last, 13 nature speeches are correctly put into nature output and the rest are incorrect. Table 1 shows the result percentage of classified emotions for two times trained network.

The overall correctly classified emotions are 73.8 % and error rate is 26.2 % as shown in Table 1, then the accuracy of the system needs to be improved.

Table 1: Description of two times trained Network classification result

Emotion	Sad	Angry	Happy	Neutral	Classified
Sad	17				85.0%
Angry		15			75.0%
Happy			14		70.0%
Neutral				13	65.0%
	Overall				73.8%

Table 2: Description of ten times trained Network classification result

Emotion	Sad	Angry	Happy	Neutral	Classified
Sad	20				100.0%
Angry		19			95.0%
Happy			19		95.0%
Neutral				18	90.0%
	Overall				95.0%

Then the performance is improved by increasing the training times to ten to let the system reaches an optimal result. In Table 2 shows the results after ten times trained. The overall correctly classified emotions are 95 % and the error rate is 5 % as shown in Figure 3 and Figure 4. The accuracy of the system is improved after increasing the number of training.

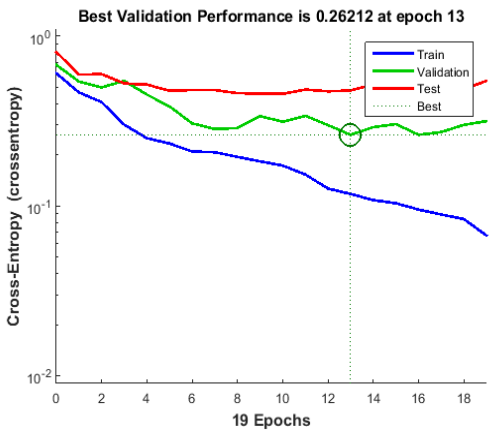

Figure 5: The performance of classifier after two times trained

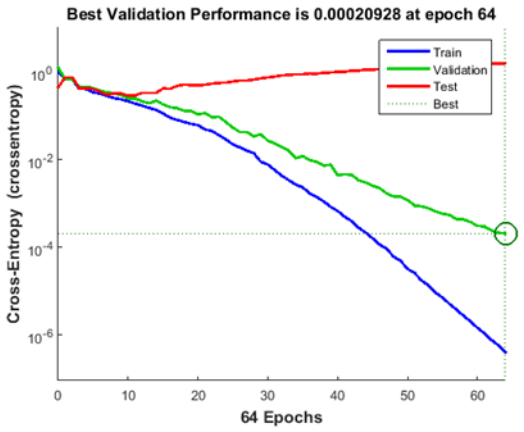

Figure 6: The performance of classifier after ten times trained

In Figure 5 shows the classifier reaches the best validation performance at epoch 13 with the value of 0.26212, where epoch means the number of times for all the training vectors used once to update the weights to the features.

In Figure 6 shows the validation performance of the classifier decreases to 0.00020928 from 0.26212 after the training ten times and the confusion matrix shows a lower error rate as shown in Figure 3 and Figure 4.

Lower values of Cross-entropy indicate that the classification is better. Zero Cross-Entropy means no error. For a new non-trained eight datasets, the classifier classifies the emotions with 87.5 % accuracy as shown in the Figure 7.

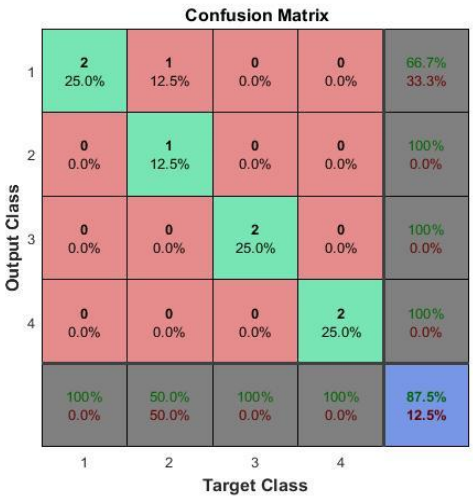

Figure 7: The performance of classifier for new data set

Furthermore, after a suitable number of training process with a low error rate, the neural network classifies completely new eight speech corps. The classification results in terms of error are shown in Figure 7 where two sad speeches are classified as correct; one angry speech is recognized as sad emotion; two happy speeches are correctly classified; two nature speeches are in the nature output and the total classification rate is 87.5 % for the new speech samples.

In this classification, using only short-term features, the rate is better than the result obtained in experimental study on emotion recognition developed 140 utterances per emotional state with both short-term and mid-term feature values [10]. Similarly, comparing with other approaches [1-9] to emotion recognition, the presented results provide higher accuracy with the selected short-term features.

Conclusion

The purpose of this work is to classify the four basic emotions in the speech signals using artificial neural network pattern recognition algorithm and analyze its performance. Artificial Neural Network is a powerful tool for pattern recognition and classification. The chosen short-term features of speech signals are loaded into the system and trained for the target emotions. After suitable number of times of training process, new test signals are loaded into the system for emotion classification and analysis. The selected 23 short-term features are proven to be good representations of emotions for speech signals with a desired accuracy of 87.5 % classification rate for the new data set.

Future Work

In future, the system could be improved by increasing the accuracy of extracted features to classify more complicated speech samples for multiple speakers and more emotions to increase the accuracy of the classifier and develop an automated emotion recognizer. This work could be developed for other spoken languages.

References

[1]. Nogueiras, A., Moreno, A. Bonafonte, A., and Marino, J. B., (2001). Speech Emotion Recognition Using Hidden Markov Models Seventh European Conference on Speech Communication and Technology.

[2]. Schuller, B., Rigoll, G., and Lang, M., (2004). Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture. *Proceedings of IEEE International Conference on Acoustics*, Speech, and Signal Processing (ICASSP).

[3]. Dellaert, F. Polzin, T., and Waibel, A., (1996). Recognizing Emotion in Speech, *Proceedings of Fourth International Conference on Spoken Language processing* (ICSLP), Philadelphia, 3: 1970-1973.

[4]. Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S., (2004). Emotion recognition based on phoneme classes, *Proceedings of Fourth International Conference on Spoken Language processing*(ICSLP), Jeju, Korea, 889-892.

[5]. Vogt, T., and Andre, E., (2006). Improving Automatic Emotion Recognition from Speech via Gender Differentiation, *Proceedings of Language Resources and Evaluation Conference*, Genoa, Italy, 1123-1126.

[6]. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A., (2002). Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog, *Proceedings of Fourth International Conference on Spoken Language processing*(ICSLP), Colorado, USA, 2037-2040.

[7]. Alexandros Georgogiannis., and Vassilis Digalakis., (2012). Speech Emotion Recognition Using Non-Linear Teager Energy Based Features in Noisy Environments. *Proceedings of the 20th European Signal Processing Conference* (EUSIPCO), IEEE, Romania.

[8]. Anand singh., Dinesh Kumar Rajoriya., and Vikash Singh., (2012). Database Development and Analysis of Spoken Hindi Hybrid Words Using Endpoint Detection, *International Journal of Electronics and Computer Science Engineering*, 1:(3),1623.

[9]. Mina Hamidi., and Muharram Mansoorizade., (2012). Emotion Recognition from Persian Speech with Neural Network, *International Journal of Artificial Intelligence & Applications* (IJAIA), 3: (5)

[10]. Valery A. Petrushin (2000). Emotion Recognition in Speech Signal: Experimental Study, Development, and Application, *Proceedings of the Sixth International Conference on Spoken Language Processing* (ICSLP), USA, 637-645.

[11]. Ronald W. Schafer., and Lawrence R. Rabiner., (1975). Digital Representations of Speech Signals, IEEE, 63:(4), 662.

[12]. Zhou y., Sun Y., Zhang J., and Yan Y., (2009). Speech Emotion Recognition using Both Spectral and Prosodic Features, IEEE, 23:(5), 545-549.

[13]. Lawrence R., Rabiner., and Ronald W.Schafer., (2007). Introduction to Digital Speech Processing, Foundations and Trends in Signal Processing, 1 :(1-2), 33-53. Now Publishers Inc. USA.

[14]. Nandhini, S., and Shenbagavalli, A., (2014). Voiced/Unvoiced Detection using Short Term Processing, *International Journal of Computer Applications* (0975-8887).

[15]. Amit Ganatra, Kosta, Y. P., Gaurang Panchal., and Chintan Gajjar., (2011). Initial Classification through Back Propagation In a Neural Network Following Optimization through GA to Evaluate the Fitness of an Algorithm: *International Journal of Computer Science & Information Technology* (IJCSIT), 3(1):98.