Classifiers for speech emotion recognition: A Review

[1]Shubham Sarwade, [2] VV Yerigeri
[1]M.Tech Student, [2]Professor
Department of PG

MBESs College of Engineering, Ambajogai, MS, India

Abstract: Automatic speech emotion recognition can fail to a certain extent when confronted with emotionally distorted speech. Great efforts have been spent so far to cope with noise conditions or speaker's characteristics. Yet, adaptation to the emotional condition of the speaker could help to further improve the overall performance. In this paper we reviewed different classifiers for speech emotion recognition, with performance parameters and a robust and reliable recognition of the speaker's emotional state by acoustic features only prior to speech recognition itself. Thereby we can load according emotional speech models. In this work we introduce an optimal feature set for this task selected by Sequential Floating Search Methods. The set comprises high-level prosodic features extraction and resembling utterance-wise statistic analysis of low-level contours as pitch, higher order formants, energy, and spectral development.

Index Terms **- Emotion Speech Recognition, Features Extraction, Classifiers**

• Introduction

Affective Emotion classification is a fundamental task for humans in order to interpret social interactions. Although emotions are expressed at various levels (e.g., behavioral, physiological), vocal and verbal communication of emotions is a central domain of communication research [14]. Classification accuracy is essential in order to be ensured of the validity and reliability of emotional constructs used in psychological research. Given the importance of accurately classifying emotions to understanding human interactions, many researchers have developed automatic emotion classification computer systems**.**

Distinguishing between different types of audio is an amazing aptitude held by human beings without conscious effort and without being considered complex. People can instantly differentiate between different human voices, they can evaluate the tempo and mood of a piece and also compare and contrast pieces of audio. Conversely, both classical and modern day computer systems merely comprehend an audio piece as a sequence of extracted parameters and have used a wide variety of algorithms to tackle the problem of emotion classification from a piece of audio. Emotion samples, especially spontaneous ones, are hard to obtain. This is especially true when aiming at a high number of evenly distributed samples among emotions of diverse speakers. Having such relatively small training sample sizes compared to the dimensionality of the data, a high danger of bias due to variances in the corpus is present. In order to improve instable classifiers as neural nets or decision trees a solution besides regularization or noise injection is construction of many such weak classifiers and combination within so called ensembles. Two of the most popular methods are bagging and boosting, firstly introduced in emotion recognition in [7]. Within the first random bootstrap replicates of the training set are built for learning with several instances of the same classifier. A simple majority vote is fulfilled in the final decision process [15].

In Boosting the classifiers are constructed iteratively on weighted versions of the training set. Thereby erroneously classified objects achieve larger weights to concentrate on hardly separable instances. Also a majority vote, but based on the weights, leads to the final result. However, these methods both use only instances of the same classifier. If we strive to combine advantages of diverse classifiers Stacking is an alternative. Hereby several outputs of diverse instances are combined. In [8] Stacking C as improved variant is introduced, which includes classifier confidences e.g. by Maximum Linear Regression. It is further shown that by Stacking C most ensemble learning schemes can be simulated, making it the most general and powerful ensemble learning scheme. One major question however remains the choice of right base classifiers.

In [8] an optimal set with four classifiers is introduced. We use a slightly changed variant of their set, which delivered better results in our case. Accuracy obtained with various base-classifiers and constructed ensembles are shown in the following table. The major drawback of the firstly selected well known base classifier Naïve-Bayes (NB) is the basing assumptions that features are independent given class, and no latent features influence the result. Another rather trivial variant is a nearest distance classifier based on entropy calculation (K*) [9]. Support Vector Machines (SVM) show a high generalization capability due to their structural risk minimization oriented training [16-18].

• SPEECH EMOTION RECOGNITION

The motivation of Speech Emotion Recognition (SER) is to provide more flexibility in human computer interaction, and make more intelligent machines. The interest behind this study is also beneficial to better speech, speaker recognition systems and more responsible human robot interactions. Our SER system should be able to recognize the different emotional patterns in an efficient and faster way. Since the emotion recognition is a hard problem for humans, the task would not be easy for machines [1-2].

• System Model

The proposed human emotion recognition system is of five components: input speech signal, pre-processing, feature extraction and selection, classification and finally emotions recognition. The Architecture of Emotional speech recognition system is as shown in figure 1. The Accuracy of the Emotional speech recognition system is based on the level of naturalness of the database which is used as an input to the speech emotion recognition system. The database as an input to the speech emotion recognition system may

contain the real world emotions or the acted ones. It is more practical to use database that is collected from the real life situations [8-11]. The emotion recognizer system identifies the emotion state of the input speech signal and displays the corresponding emotions of that particular speech. The speech processor, deals with the emotion features selection, speech preprocessing, and extraction algorithm. Determining emotion features is a vital issue in the emotion recognizer design. The emotion recognition result is strongly depending on the emotional features and which provided to various types of classifier that have been used to represent the emotion [6].

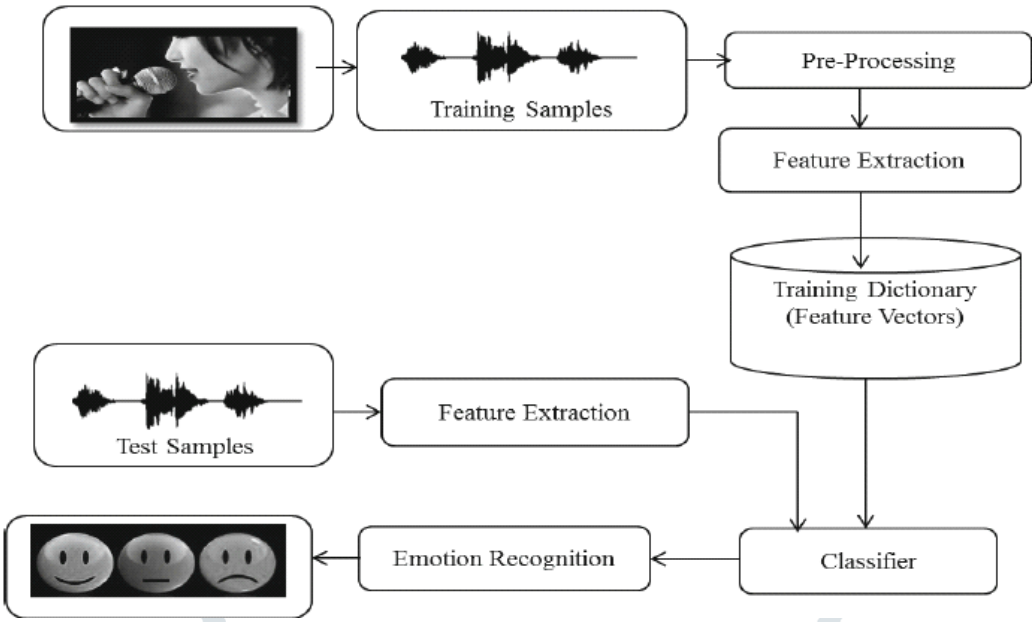

Figure 1: Speech Emotion Recognition System [1].

- **Features for Speech Recognition**

Most raw data, in particular audio data, can be extremely vast and can contain a large amount of redundancy. In any audio classification the system the first two stages consist of raw data being extracted in a feature extraction stage and then these features are reduced in a feature selection stage before being presented to a classification algorithm. This second stage of feature selection consists of choosing the important features to use for classification. There is much literature documenting the significant features necessary for audio and speech classification. Gerhard [2-3] outlines how there are many quantities such as frequency, spectral content, rhythm and formant location that can accurately describe spoken utterances. Gerhard was the first however accurately group these features. He describes how all audio features can either be classified as physical features or perceptual features. The physical features are described as being easier to recognize and to extract as they relate to the physical properties of the input signal itself, whereas perceptual features are related to the way humans consciously perceive sound and therefore rely on a certain amount of perceptual modeling [4].

- **Speech Emotion Classifiers**

For multiclass emotion classification systems, commonly used are [11-20],

Neural Networks: Neural Networks also known as an Artificial Neural Network (ANN) is a popular classification approach The Neural Network they used was trained by back propagation and consisted of a 33 input neuron input layer, a hidden layer with 100 neurons and an output layer with 7 neurons each representing Ekman's six basic emotions plus an added neutral class. Dai et al. (2008) also use a Neural Network in their study on the Emotional Prosody Speech and Transcripts corpus obtaining an accuracy of over 90% in hot anger in neutral utterances and over 80 % distinguishing between happiness and sadness. Tato et al. (2002) also utilize a Neural Network in their study using prosodic features derived from pitch, loudness and quality features.

Bayesian Networks: Bayesian Networks are popular models used in classification problems based on Bayes Theorem The novelty of Lugger and Yang's approach lies in the combination of the prosodic and voice quality features they use; they test on short acted utterances from the Berlin emotional database. They find that their approach using the combined prosodic and voice quality features is a good baseline and use this as a basis to propose an interesting cascading approach. The cascading approach outlined stems from the fact that prosodic features tended to be useful in distinguishing between emotions of high and low activation. Lugger and Yang define emotions of high activation as being happiness, anger and anxiety and those of low activation to be neutral boredom and sadness.

Decision Trees: Decision Trees are another approach which many researchers employ in tackling the problem of automatic affect recognition. This is a simplistic classifier which makes observations on data and maps these observations to decisions on class ownership. It functions by constantly querying a test instance to gain more information about which class it may belong through a combination of if-then rules. This study experimented with a number of data pre-processing approaches on prosodic features from speech samples achieving up the 83.33% accuracy for this classifier.

Gaussian Mixture Models and Support Vector Machines: GMMs are a type of probabilistic mixture model which assumes all points in a feature space are generated from a number of Gaussian distributions. The mixture model is learned from training data and then a test instance is classified by the class label of the Gaussian distribution which is the most probabilistic. They are another widely used classification algorithm used to tackle SER problems. SVMs are another classification approach which functions by analyzing a feature space and attempting to construct a hyperplane to separate data points belonging to different classes. They operate by mapping data onto a higher dimensional space using a kernel function and defining the hyperplane there. Although

SVMs are inherently binary classifiers they can be modified for multiclass problems by using pair wise classification, which tackles a problem as a series of binary problems.

K-Nearest Neighbor: K-Nearest Neighbor is a non-parametric lazy learning algorithm which classifies test instances based on computed distances measures to labeled training instances. This algorithm simply records training data and then uses a distance measure from a test instance to known training instances to predict which class the test instance should belong to by examining its nearest neighbors.

Linear Discriminate Analysis: Linear Discriminate Analysis is a common feature reduction technique but is also used as a linear classifier. It functions by constructing a linear combination of input features which offers the best class discrimination or separability. This approach used features such as functionals of the fundamental frequency (F0) and energy i.e. mean, median, standard deviation, maximum, minimum, range (max–min) and linear regression coefficients.

• CONCLUSION

In this paper we discussed the novel block diagram of SER system. For emotion classification in real scenarios, noise is a factor that inevitably needs to be considered for performance evaluations. We discuss several noisy scenarios that may require speech-based emotion classification. Experimental results show the impact of noise on the emotion classification performance. For interaction designers and HCI participants designing interactive systems using users' emotion from speech, the noise effect should be taken into consideration as an important factor. To more effectively classify emotion on noisy data, the system should be trained using noisy data instead of clean data. To reduce the influence of noise on system reliability, we can adapt the system to different noise levels. For example, we can choose to increase the confidence score threshold used in the SVM thresholding fusion for very noisy scenarios, and only classify emotions when the confidence score is relatively high. Additionally, we can sample the user's speech multiple times within a short period of time, and derive the user's emotions.

References

- Na Yang, Jianbo Yuan, Yun Zhou, "Enhanced multiclass SVM with thresholding fusion for speech based emotion classification", Springer journal, Oct-2016.
- Jung M., Hwang K., and Choi S., 2011, "Interference Minimization Approach to Pre-coding Scheme in MIMO-Based Cognitive Radio Networks", IEEE Communications Letters, 15(8), pp. 789-91.
- Lee K.J., and Lee I., 2011, "MMSE Based Block Diagonalization for Cognitive Radio MIMO Broadcast Channel", IEEE Transactions on Wireless Communications, 10(10), pp. 3139–3144.
- Kim, J., Choi, W., Nam, S. and Han, Y., 2014, "An Efficient Pre-whitening Scheme for MIMO Cognitive Radio Systems", IEEE Transactions on Vehicular Technology, 63(4), pp.1934-1939.
- Krondorf, M. and Fettweis, G., 2009, "Numerical Performance Evaluation for Alamouti Space Time Coded OFDM under Receiver Impairments", IEEE Transactions on Wireless Communications, 8(3), pp.1446-1455.
- Jin, Y. and Dai, F.F., 2012, "Impact of Transceiver RFIC Impairments on MIMO System Performance", IEEE Transactions on Industrial Electronics, 59(1), pp.538-549.
- Dai, X., Zou, R., Sun, S. and Wang, Y., 2013, "Transceiver Impairments on the Performance of the LMMSE-PIC Iterative Receiver and its Mitigation", IEEE Communications Letters, 17(8), pp.1536-1539.
- Boulogeorgos, A.A.A., Salameh, H.A.B. and Karagiannidis, G.K., 2017, "Spectrum Sensing in Full-Duplex Cognitive Radio Networks under Hardware Imperfections", IEEE Transactions on Vehicular Technology, 66(3), pp.2072-2084.
- Chang T H., Chiang W C., Peter Hong Y W., and Chi C Y., 2010, "Training Sequence Design for Discriminatory Channel Estimation in Wireless MIMO Systems", IEEE Transactions on Signal Processing, 58(12), pp-6223-6237.
- Chiang C T., and Fung C C., 2011, "Robust Training Sequence Design for Spatially Correlated MIMO Channel Estimation", IEEE Transactions on vehicular technology, 60(7), pp. 2882-2894.
- Nasir A A., Mehrpouyan H., Durrani S., Blostein S D., Kennedy R A., and Ottersten B., 2013, "Optimal Training Sequences for Joint Timing Synchronization and Channel Estimation in Distributed Communication Networks", IEEE Transactions on communications, 61(7), pp. 3002-3015.
- Afifi, W. and Krunz, M., 2015, "Incorporating Self-Interference Suppression for Full-Duplex Operation in Opportunistic Spectrum Access Systems," IEEE Transactions on Wireless Communications, 14(4), pp.2180- 2191.
- Liao Y, Wang T, Song L, Han Z, 2017, "Listen and Talk: Protocol Design and Analysis for Full Duplex Cognitive Radio Networks", IEEE Transactions on vehicular technology, 66(1), pp. 656-66.
- Saghiri, A.M. and Meybodi, M.R., 2016, "An Approach for Designing Cognitive Engines in Cognitive Peer-to-Peer Networks", Journal of Network and Computer Applications, 70, pp.17-40.
- Kustiawan, I., and Chi, K. H., 2015, "Handoff Decision Using A Kalman Filter and Fuzzy Logic in Heterogeneous Wireless Networks", IEEE Communications Letters, 19(12), pp. 2258–2261.
- Zhioua, G., Tabbane, N., Labiod, H., and Tabbane, S., 2015, "A Fuzzy Multi-Metric Qos Balancing Gateway Selection Algorithm in a Clustered VANET to LTE Advanced Hybrid Cellular Network", IEEE Transactions on vehicular technology, 64(2), pp. 804–817.
- Matinmikko, M., Del Ser, J., Rauma, T., and Mustonen, M., 2013, "FuzzyLogic Based Framework for Spectrum Availability Assessment in Cognitive Radio Systems", IEEE Journal on Selected Areas in Communications, 31(11), pp. 2173–2184.
- Bellegarda, J. R. (2013). Data-driven analysis of emotion in text using latent affective folding and embedding. Computational Intelligence, 29(3), 506–526.
- Bitouk, D., Ragini, V., & Ani, N. (2010). Class-level spectral features for emotion recognition. Journal of Speech Communication, 52(7–8), 613–625.
- Black, M. P., Katsamanis, A., Baucom, B. R., Lee, C. C., Lammert, A. C., & Christensen, A. (2013). Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. Speech communication, 55(1), 1–21.