Speech Emotion Recognition: A Review

Rahul. B. Lanjewar, D. S. Chaudhari

Abstract -The man-machine relation has demanded the smart trends that machines have to react after considering the human emotional levels. The technology boost improved the machine intelligence that it gained the capability to identify human emotions at expected level. Harnessing the approaches of signal processing and pattern recognition algorithms a smart and emotions specific man-machine interaction can be achieved with the tremendous scope in the field of automated home as well as commercial applications. This paper reviews the aspects of speech prosody in the form of pitch, intensity, speaking rate at the same the contribution of Mel Frequency Cepstrum Coefficients based speech features in speech emotion recognition implementation. The impact of incorporating fusion techniques, wavelet domain analysis and the classifier models on the recognition rate in the identification of six emotional categories namely happy, angry, neutral, surprised, fearful and sad from the standard speech database is emphasized with intend to improve recognition fidelity.

Keywords: - Features, Emotion, MFCC, HMM, Classifier, Database, Fusion.

I. INTRODUCTION

The dynamic requirements of automated systems have pushed the extent of recognition system to consider the precise way of command rather to run only on command templates. The idea correlates itself with the speaker identification at the same time recognizing the emotions of speaker. The acoustic processing field not only can identify 'who' the speaker is but also tell 'how' it is spoken to achieve the maximum natural interaction. This can also be used in the spoken dialogue system e.g. at call centre applications where the support staff can handle the conversation in a more adjusting manner if the emotion of the caller is identified earlier. The human instinct recognizes emotions by observing both psycho-visual appearances and voice. Machines may not exactly emulate this natural tendency as it is but still they are not behind to replicate this human ability if speech processing is employed. Earlier investigations on speech open the doors to exploit the acoustic properties that deal with the emotions. At the other hand the signal processing tools like MATLAB and pattern recognition researcher's community developed the variety of algorithms (e.g. HMM, SVM) which completes needed resources to achieve the goal of recognizing emotions from speech.

Manuscript received on March, 2013

Rahul. B. Lanjewar, M.Tech III[rd] Sem.Student: Electronic System & Communication Government College of Engineering Amravati Amravati, India.

Dr. D. S. Chaudhari, Head: Department of Electronics & Telecommunication Government College of Engineering Amravati Amravati, India.

This paper focuses on technical challenges that arise when equipping human-computer interface to recognize the user vocal emotions. Starting with the system overview, the acoustic properties of voice, their features extraction and selection based on the emotion relevance which is identified by the earlier studies is reviewed and later the previous work by the speech processing community dealt with emotions is discussed in details.

II. SPEECH EMOTION RECOGNITION SYSTEM

In a generalized way, a speech emotion recognition system is an application of speech processing in which the patterns of derived speech features (MFCC, pitch) are mapped by the classifier (HMM) during the training and testing session using pattern recognition algorithms to detect the emotions from each of their corresponding patterns. The technique is synonymous to speaker recognition system but its different approach to detect emotions makes it intelligent and adds security to achieve better service in various applications.

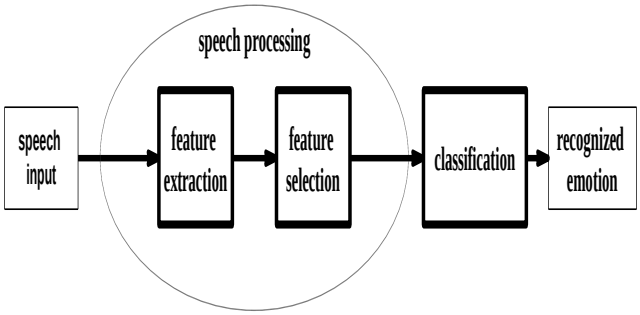

Figure.1 Basic Speech Emotion Recognition System

After getting the prior knowledge from previous studies we can draw the basic modular flow of the system processes as shown in Figure 1. Since the emotion is to be detected from the input speech signal the whole signal processing revolves around the speech signal for the extraction and selection of speech features correspond to emotions. The next is generating a database for training and testing of extracted speech features followed by the last stage of emotion detection by the classifier section using pattern recognition algorithms [AI 2012]. Firstly the speech signal is pre-processed for the removal of noise and d.c components to pass it further for features extraction and selection. The speech features are the acoustics information usually derived from the analysis of speech in both time as well as frequency domain. The extracted features are then selected in terms of emotion relevance and also to control the dimensionality of combined features which will further be classified to determine the emotion in speech as discussed in details in the further section.

A. Extraction and Selection of speech features

The extraction of speech features involves potential audio segmentation followed by acoustic pre-processing like

filtering to form their meaningful units. The purpose of the audio segmentation is to segment a speech signal into units that are representative for emotions. These are usual linguistically motivated middle-length time intervals such as words or utterances. The next step is the extraction of relevant features by finding the properties of the digitised and pre-processed acoustic signal which characteristically deals with emotions and further represent them into n-dimensional feature vector. In the early methods the features sets consisted originally mainly of pitch and energy related features which later gained the prominence in emotion detection. The pitch related statistics, Formants and Mel Frequency Cepstral Coefficients (MFCC) are also frequently found to contribute as feature vectors. The parametric representations of Spectral measures other than MFCCs are also common in research including the features of Wavelets, Teager energy operator (TEO) based features, Log Frequency Power Coefficients (LFPC) and Linear Prediction Cepstral Coefficients (LPCC). These extracted features are stored by training the database for the classification stage. The larger the features used the more improved will be the classification process but practically the feature space suffers the phenomenon of 'curse of dimensionality'. Thus the feature selection process is employed to select only those features which carry relevant emotion information to improve the classification process [SK 2007] [TB 2005] [KH 2003].

B. Database for training and testing

A good database is as important as the desired results. There are different databases created by speech processing community with the help of professional actors which is widely used in research work. The results are uncompromising though the emotions are acted rather than spontaneous or natural. The famous databases are The Danish Emotional Speech Database (DES), and The Berlin Emotional Speech Database (BES), as well as The Speech under Simulated and Actual Stress (SUSAS) Database. DES and BES are representative for the early databases in the nineties but still serve as exemplars for acted emotional databases. For English, there is the 2002 Emotional Prosody Speech and Transcripts acted database available.

C. Classifiers to detect emotions

The selected feature vectors are stored in the database and later fed to classifier to detect the emotions by comparing the vectors from the trained data and test data vectors. There are various classifiers available meant for their specific usage based on types of features to be classified. If feature vectors belong to the global statistics SVM (Support Vector Machine), Neural Networks, Decision trees are employed and for the vectors of short-term features HMM (Hidden Markov Model) is used for its dynamic performance [AI 2012].

III. DISCUSSION

Busso *et al*. conducted two experiments to analyse the salient aspects of pitch. In their first experiment the emotional pitch contours were compared with neutral speech while in the second experiment the measure of discriminative power of pitch features for the emotions was derived lead to conclude that sentence level pitch features outperforms the voiced-level pitch statistics both in accuracy and robustness [5]. Eva Nava and I. Luengo studied the role of semantics and prosodic features to build the emotion recognition system. Their study reveals that prosodic features specially the pitch and energy, if considered individually are the appropriate indicators of emotions but not for all the emotions while the semantic information can detect disgust emotion efficiently [6]. Their work indicates that long term spectral envelop provides larger separation among emotion than prosodic features by using unsupervised clustering [7]. More diverse results can be obtained in terms of robustness when mathematical Teager Energy Operator (TEO) were employed on speech to calculate their energy that reflects the stress in the airflow structure of speech for both text dependent and text independent models [8].

Christiansen *et al.* studied the low energy fluctuations below 16Hz in speech and distinguished it in the form of temporal and spectral properties. The temporal modulations are mainly due to syllabic structure of speech while the spectral modulations are due to harmonic and formants structure in speech. These spectro-temporal modulations play important role in sound perception [9]. Spectro-temporal processing of signal can enhance the speech affected by noise and reverberations at the same time restoring perceptual quality and intelligibility of speech. The temporal processing explicitly does not provide enhancement but the spectral processing have good impact in suppressing reverberations [10]. In another work by Ghosh *et al.* derived the two temporal features namely Average Level Crossing Rate (ALCR) and Extrema based Signal Track Length (ESTL) and found they have better ability of distinguishing those temporal features pattern which are different in amplitude and frequency. These features have better energy capturing ability than TEO [11]. In different domain approach by Farooq *et. al* analysed wavelet packet transform's multi-resolution capabilities to derived a new superior than MFCC feature sets and showed improvement in emotion recognition of unvoiced phonemes and stop statements [12]. The Bionic Wavelet transform (BWT) developed by Yao *et.al.* is a biomedical based approach that provides concentrated energy distribution to retain more energy and introduces active control mechanism in auditory system to adjust the wavelet transform. BWT has high sensitivity and selectivity [13]. Koolagudi *et.al* reviewed and concluded that entire speech region may not be necessary helps to recognize the underlying emotions. Linear Prediction Cepstrum Coefficients (LPCC), Mel Frequency Cepstrum Coefficients (MFCC) and formants represents the vocal tract information. MFCC are claimed to be robust of all the features for any speech tasks. The combination of formants and LPCC gave better emotion recognition performance specifically for both speaker and gender independent emotions. The consonant-vowel transition region also yields emotion recognition but when formants features combined with the basic spectral features have always improved the system performance [14].

In the new approach of features integration, the short-time feature vectors sequence are used to create a new single vector in large-time scale [15]. The integration of temporal features at two levels in the form of early integration and late integration can be more effective in classification performance [15] [16], the cross-spectral integration may improve speech intelligibility in acoustically challenging environments [17]. We can derive many features from the speech but the high dimensional data hampers the performance of the classifier, so the feature selection techniques are employed. Generally linguistics features

are avoided, the feature selection algorithm based on majority voting outperforms the ordinary feature selection techniques [18]. Wide work in recognition system gave a large variety of classifiers to map the data. The classification performance increases further when the linear network classifier is replaced by Hidden Markov Model (HMM) based classifier at the back end [14]. When the low-level descriptors were employed for classification the HMM gave higher accuracy up to 80% for the speaker dependent recognition [19]. The performance of neural network is also promising to detect seven emotions (neutral ,anger, boredom, disgust, fear, happiness, sadness)when used in the form of Multilayered Perceptron neural network (MLPNN) and Generalized feed forward neural network (GFFNN) [20]. Wu *et al.* used new approach of multiple classifiers in which they used three types of classifier models namely Gaussian Mixture Model (GMM), Support Vector Machine (SVM) and Multilayer Perceptron (MLP) as a base level classifier and the trio were fused with the help of Meta Decision Tree (MDT) to obtain the acoustic–prosodic information (AP) and semantic-label (SL) information for the four emotion states of (neutral, happy, angry and sad). Their result on speaker-independent experiment reveals that this technique can achieve emotion recognition up to 80% using MDT which is better than individual classifiers. The results are more satisfying if only SL type information is used making the recognition up to 80.92 while it is 83.55 % for only AP type of information but it reaches 85.79% when both AP and SL type of information were combined for the classification [21]. But the results of GMM are tending to be more discriminative that of HMM for Berlin Emotional Speech Database (BES) and it measure up to 76% than of 71% of HMM, 67% of k-NN and 55% of FFNN. They used GMM in two stages: first to classify high, low and neutral emotions, in second stage the emotions of same category were classified [22].

The above discussed techniques are incomplete if the database is not strong, so a good and natural database can boost the results of recognition. The speaker specific information always plays an important role. The usage of same speaker for the training and testing the data makes the model to lack generality. Thus the databases with reasonably large speaker and text prompts can give discriminative results at the same time natural emotions database can outclass the simulated database for the real life challenges [14]. In a different master class review performed by Ververidis et al. on the available 32 database they derived three important results. First, not more than 50% classification can be achieved for the four basic emotions in automated emotion recognition system. Second, simulated emotions are easy to classify as compared to natural emotions. Third, the emotions in the descending order of their easier classification are anger, sadness, happiness, fear, disgust, joy, surprise and boredom [23].

IV. CONCLUSION

A new framework for the emotion recognition can be envisaged to demonstrate the feasibility of integrating the new features derived by wavelet decomposition with the spectro-temporal and the baseline features (MFCC, pitch) in the recognition of human emotional states. In case of simulated database the HMM based multi-classifier approach at the different levels of classification can turn more discriminative than previous results of emotion detection. The accuracy of emotion recognition and the robustness of the system deals with the extracted features set, classifier and the database. The impact of proposed system can also be checked on music to classify the genre if the regression model of integrating speech features is employed.

REFERENCES

[1] A. B. Ingale, D. S. Chaudhari,"Speech Emotion Recognition", *Int'l Journal of Soft Computing and Engineering*, vol-2, Issue-1, pp 235-238, Mar. 2012.
[2] S. Kim, P. Georgiou, S. Lee, S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", *Proc. of IEEE Multimedia Signal Processing Workshop*, Greece, 2007
[3] Kwon, Oh-Wook, Chan, K. Hao, J., Lee, Te-Won, "Emotion recognition by speech signals", EUROSPEECH - Geneva, pp 125–128, 2003.
[4] T. Bänziger, K. R. Scherer, "The role of intonation in emotional expression", *Proc. IEEE Int'l Conf. on Speech Communication*, vol.46, pp 252-267, 2005
[5] C. Busso, S. Lee and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection", *IEEE Trans. on Audio, Speech and Language processing*, vol. 17, no. 4, pp 582-596,May 2009
[6] E. Navas, I. Hernáez, I. Luengo , "An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp 490-501,Jul. 2006
[7] I. Luengo, E. Navas, I. Hernáez , "Feature Analysis and Evaluation for Automatic Emotion Identification in Speech", *IEEE Trans. on Multimedia*, vol. 12, no. 6,pp 1117-1127, Oct. 2010
[8] G. Zhou, J. L. Hansen, and J. F. Kaiser, "Methods for Stress Classification: Nonlinear Teo and Linear Speech Based Features, *Proc. IEEE Int'l Conf. Acoustics and Signal Processing*, pp. 2087-2090, 1999.
[9] T. U. Christiansen and S. Greenberg, "Distinguishing Spectral and Temporal Properties of Speech Using an Information-Theoretic Approach", *Proc. IEEE Int'l Conf. Acoustics and Signal Processing*, pp. 2087-2090, 1999.
[10] P. Krishnamoorty and S. R. Mahadeva Prasanna, "Temporal and spectral processing methods for processing of degraded speech: A Review" *IETE Technical Review*, vol 26, no.2,pp 137-148, Mar-Apr 2009
[11] P. K. Ghosh, A. Sarkar and T. V. Sreenivas, "ALCR and ESTL: Novel Temporal Features and their Application to Speech Segmentation" International Conference on Acoustics, Speech, and Signal Processing - ICASSP , vol. 4, pp. IV-1065-IV-1068, 2007.
[12] O. Farooq and S. Datta, "Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition", *IEEE Signal Processing Letters*, vol. 8, no. 7,pp 196-198, Jul. 2001
[13]Yao, Yuan-Ting Zhang, "Bionic Wavelet Transform: A New Time–Frequency Method Based on an Auditory", *IEEE Trans. on Biomedical Engg.*, vol. 48, no. 8, pp 856-863,Aug. 2001
[14] S. G. Koolagudi, K. Sreenivasa Rao, "Emotion recognition from speech: a review", Int'l Journal of Speech Technology,pp 99-117,2012
[15] C. Joder, S. Essid, and G. Richard "Temporal Integration for Audio Classification with Application to Musical Instrument Classification", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp 174-186, Jan. 2009.
[16] A. Meng, P. Ahrendt, J. Larsen and L. K. Hansen, "Temporal Feature Integration for Music Genre Classification", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 5, pp 1654-1664,Jul. 2007
[17] T. U. Christiansen and S. Greenberg, "Distinguishing Spectral and Temporal Properties of Speech Using an Information-Theoretic Approach", Centre for Applied Hearing Research, Technical University of Denmark, Kgs. Lyngby, and Denmark Springer.
[18] F. Dellaert, T. Polzin, A. Waibel, "Recognizing Emotion in Speech",*Pro.Int'l Conference on Spoken Language Processing*, pp 473-482, 1996.
[19] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño,"Speech Emotion Recognition Using Hidden Markov Models", *Pro. Int'l Conference INTERSPEECH,pp* 345-350,2001.

[20] K. B. Khanchandani, M. Hussain, "Emotion recognition using Multilayerd perception and Feedforward neural network",Journal of Scientific and Industrial Research vol.68, pp. 367-371, May 2009.
[21] Chung-Hsien Wu and Wei-Bin Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels", *IEEE Trans. on Affective Computing*, vol. 2, no. 1, pp 10-21,Jan-Mar 2011
[22] Moataz M. H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Speech Emotion Recognition Using Gaussian Mixture Vector Autoregressive Models", *IEEE Signal Processing Letters*, vol. 8, no. 7, pp 957-960,2007.

Rahul B. Lanjewar is a student of M.Tech in Electronics System and Communication at Government College of Engineering, Amravati. He received B.E in Electronics Engineering from University of Pune in 2010. His area of research includes MATLAB based Signal Processing, Wavelet analysis and Speech Emotion Recognition.

Dr. **D. S. Chaudhari** obtained BE, ME, from Marathwada University, Aurangabad and PhD from Indian Institute of Technology, Bombay, Powai, Mumbai. He has been engaged in teaching, research for period of about 25 years and worked on DST-SERC sponsored Fast Track Project for Young Scientists. He has worked as Head Electronics and Telecommunication, Instrumentation, Electrical, Research and incharge Principal at Government Engineering Colleges. Presently he is working as Head, Department of Electronics and Telecommunication Engineering at Government College of Engineering, Amravati. Dr. Chaudhari published research papers and presented papers in international conferences abroad at Seattle, USA and Austria, Europe. He worked as Chairman / Expert Member on different committees of All India Council for Technical Education, Directorate of Technical Education for Approval, Graduation, Inspection, Variation of Intake of diploma and degree Engineering Institutions. As a university recognized PhD research supervisor in Electronics and Computer Science Engineering he has been supervising research work since 2001. One research scholar received PhD under his supervision. He has worked as Chairman / Member on different university and college level committees like Examination, Academic, Senate, Board of Studies, etc. he chaired one of the Technical sessions of International Conference held at Nagpur. He is fellow of IE, IETE and life member of ISTE, BMESI and member of IEEE (2007). He is recipient of Best Engineering College Teacher Award of ISTE, New Delhi, Gold Medal Award of IETE, New Delhi, Engineering Achievement Award of IE (I), Nashik. He has organized various Continuing Education Programmes and delivered Expert Lectures on research at different places. He has also worked as ISTE Visiting Professor and visiting faculty member at Asian Institute of Technology, Bangkok, Thailand. His present research and teaching interests are in the field of Biomedical Engineering, Digital Signal Processing and Analogue Integrated Circuits.