Automatic Speech Emotion Recognition: A Survey

Purnima Chandrasekar$, Santosh Chapaneri#, Dr. Deepak Jayaswal*
M.E. Student$, Assistant Professor#, Professor*,
Department of Electronics and Telecommunication Engineering,
St. Francis Institute of Technology, University of Mumbai
santoshchapaneri@gmail.com

Abstract—**The area of Automatic Speech Emotion Recognition (ASER) has garnered a lot of interest among researchers. The framework of ASER typically includes three steps viz. speech feature extraction, dimensionality reduction and feature classification. At the base of this framework lies the design and recording of the database of emotional states through which the most popular set of emotions-happiness, sadness, anger, fear, disgust, boredom (which are typically called as 'archetypal emotions') and neutral among others have been obtained. This paper surveys the extent of work done in this field especially highlighting the three steps of the ASER framework. Starting with the different languages that have been explored till date for creating the databases, this paper attempts to categorize the features that have been typically extracted, enlist the dimensionality reduction techniques that have been chosen and discuss the pros and cons, if any, of the feature classifiers that have been modelled.**

Keywords— feature extraction; dimensionality reduction; feature classification

I. INTRODUCTION

Emotion, which is the most natural way to express, is a common instinct for human beings. Understanding the emotional state of the speaker can help the listener detect the real meaning of speech hidden between words [1]. The research on Automatic Speech Emotion Recognition (ASER), which started around 1970, has since then come a long way to include diverse applications ranging from automatic remote call centers, humanoid robots, e-learning, medical field etc. The basic framework for ASER typically includes three stages: Feature Extraction, Dimensionality Reduction, and Feature Classification.

Emotion recognition is a statistical pattern classification problem. While the theory of pattern classification is well developed, the extraction of features for emotion recognition is a highly empirical issue and depends on the specific application and database. The goal of feature extraction is to select an equivalent parametric representation of speech by extracting distinctive features of the speech that contributes to accurate detection of emotions [2]. The complexity of the system depends on the dimensions of input features and thus for reduced memory and computation, reducing the dimensionality of the extracted features becomes a crucial aspect of the ASER system. To achieve the desired result of recognizing the emotion appropriately from a given utterance, feature classification as an inherent technique to ASER has been considered. The basis of ASER is a speech database for which a record of already available speech data collections that is close to real-life specifications has helped in extensive research with regards to ASER. These databases have recorded diverse emotions starting with emotions being labelled as negative/non-negative in early research to the more recent approach to specific labelling of emotions in the name of 'archetypal emotions' which include the emotions of happiness, sadness, anger, fear, disgust, boredom etc.

Section II discusses the various databases that have been most frequently used in research related to ASER. Section III describes the various feature extraction techniques followed by the explicit mention of dimensionality reduction and feature classification techniques in Section IV. In Section V, a narrowed down discussion has been made after having taken a cue from the previously discussed sections.

II. EMOTION SPEECH DATABASES

The databases that have been used in literature for the purpose of ASER can be broadly classified into acted, naturalistic and induced database. Most importantly, the basic requirements that were met for building the existing databases included a large number of speakers (both male and female) of different age groups with varied accent, recording quality, balanced distribution of class of emotions etc. [3]. The different languages explored for creating the different existing databases in the research of ASER are as follows:

A. German

The most commonly used database in ASER is the Berlin database of emotional speech (Emo-DB) that has about 500 utterances spoken by actors expressing the archetypal emotions whose complete description is given in [4] and used in [5-7]. In [8], along with Emo-DB, using the SmartKom corpus, a naturalistic database was recorded. Extending the recording scenarios to materials from television, one of the many databases that were explored was the Vera-Am Mittag that consisted of recordings taken from a German TV talk show 'Vera am Mittag' [5]. Another German database that involved children for recording their spontaneously emotional reactions in interaction with SONY's robot AIBO was the FAU AIBO emotion corpus [9] that considered the emotions ranging from joy, surprise to anger, boredom and irritation.

B. English

English as a language for recording the emotions has been explored in [10], in which classification of emotions into

978-1-4799-2494-3/14/$31.00 ©2014 IEEE

'negative' and 'non-negative' was obtained from spoken dialogs with a machine agent over the telephone from a call center application. In [11], under the DARPA Communicator project, telephone recordings taken from travel agencies were an integral part of the database formation. Deviating from the typically recorded archetypal emotions, here other emotions like tired, amused etc. were also explored. Further in [12], database material was obtained from the ISL Meeting Corpus that consisted of recordings of 18 meetings of an average duration of 35 minutes each.

C. *Swedish*

This language was first explored in [13] by A. Abelin and J. Allwood in 2000 by choosing a non-professional Swedish speaker to utter phrases while expressing different emotions of anger, disgust, dominance, fear, joy, sadness, shyness and surprise. This language was further used in [12] by considering voice-controlled information recorded with respect to airlines, ferry, traffic and postal services by a Swedish company, Voice Provider.

D. *Mandarin*

Among the first to record a Mandarin corpus were L. Yang and N. Campbell in 2000 whose intentions were to study complex emotions in spontaneous speech [14] and whose usage was also found in [15]. Thereafter, a lot of self-recorded databases have been developed for this language. In [16], using six native Mandarin speakers, archetypal emotions were recorded. In [17], an acted speech and a natural speech corpus were recorded both in Mandarin and Cantonese that explored negative, positive and neutral emotions.

E. *Indian*

For Indian languages, relatively lesser work has been done for recognizing emotions in native speech. K. Rao and S. Koolagudi [18] attempted to explore the Hindi language and its various dialects for the purpose of emotion recognition as well as for dialect identification. Also, in [19], A. Kandali, A. Routray and T. Basu created a database that explores five native languages of Assam viz. Assamese, Bodo, Dimasa, Karbi and Mishing. With around 22 official Indian languages existing today, exploring them with the intention of creating an emotion recognition system that is multi-linguistically diverse will improve on the versatility of the ASER system.

III. FEATURE EXTRACTION

Various features can be extracted from speech and with an attempt of categorizing these features, classification into vocal tract spectrum features, prosodic features, non-linear features among others have been made with the intention of extracting equivalent parametric values of speech that is utilized for further processing in the ASER system.

A. *Vocal tract spectrum features*

For emotion recognition, the most valuable information is contained in the spectral shape of the vocal tract. Processing the speech signal into blocks called as 'frames' is a common procedure over which the statistical properties of the speech waveform are observed to be relatively constant. This is done owing to the slowly varying nature of the speech signal [20]. Features extracted from these frames are referred to as 'segmental' features or typically spectral features. Some of the spectral features that have been popularly extracted include Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC), Mel Energy Spectrum Dynamic Coefficients (MEDC), Zero Crossing Rate (ZCR), jitter, shimmer etc.

In [16], LFPC was used to represent the speech signal into respective feature vectors and along with F_0 i.e. fundamental frequency, it was used as a representative spectral feature in [21]. MFCC [2] was initially used in [22] and finds detailed description in [23]. As a spectral feature, MFCC has also been extracted in [8] and [12]. The delta and double delta MFCC features also help in improving the accuracy since these considers the temporal variations of the spoken signal. As compared to MFCC, a new set of features in which the first delta operation is performed in the spectral domain instead of the cepstral domain in order to suppress the slowly-changing noisy components are the Delta-Spectral Cepstral Coefficients (DSCC) [24, 25]. An analysis of some of the features are mentioned in Table I.

TABLE I. ANALYSIS OF VOCAL TRACT SPECTRUM FEATURES

Extracted Features	Pros	Cons
MFCC	• Popular feature • Helps capture the phonetic perception of human ear • Including the delta and double-delta values improves the recognition accuracy	• Not robust to noise
LPCC	• Popular choice of features that represents the phonetic content of speech • LPCC values are observed to be decorrelated, hence a diagonal covariance of its values could be used as input to the feature classifier [26]	• Overlap of coefficient values is typically observed for different emotions (especially, for anger and sadness)
ZCR	• Simplicity in calculation • Time-domain feature	• ZCR values tend to vary significantly, depending on amount of noise present
Jitter, Shimmer	• Compared to using them individually, combining them with other spectral features such as MFCC helps improve the classification accuracy substantially	• Certain emotions like anger and disgust tend to show similar jitter and shimmer values [27, 28]
LFPC	• Performs well over the traditional features of MFCC and LPCC [16]	• Most study groups have only compared LFPC with MFCC and LPCC
DSCC	• Robust to noise • Performs better than MFCC [24]	• Does not consider the non-linear variations in the speech signal

B. Prosodic features

Also called as acoustic features, these are extracted over regions longer than a typical frame and are thus also referred to as 'suprasegmental' features. Commonly extracted prosodic features include pitch, energy, articulation rate, pause, spectral tilt features, duration etc. The contours of prosodic features (that indicate plateau, rising or falling slopes) derived in the research of ASER include in general minimum, maximum, median and interquartile range. Pitch and energy as prosodic features have been extracted in [8, 29]. Some of the attributes of the prosodic features are mentioned in Table II.

TABLE II. ATTRIBUTES OF PROSODIC FEATURES

Extracted Features	Attributes
Pitch	• An apt choice for emotion recognition as it depends on the tension created on the vocal cords • Finds utility in the application of emotional colouring [30]
Energy	• Effortful speech contains relatively greater energy in the lower frequency bands making hearing of such speech relatively pronounced • Contours of energy rely on broad class of emotions, thus making it independent of the spoken language and its content [31]
Spectral Tilt	• Is distinct for different emotions typically the ones that have high arousal rates like anger or joy

C. Non-linear features

Typically the vocal tract shape is considered as a linear filter that filters the excitation signal to produce the speech signal [32]; however under stressful conditions, as researched by Teager in the 80's, additional excitation signals rose across the spectrum as harmonics because of non-linear air flow in the vocal tract due to a stressfully expressed or an angry speech [30]. The non-linear features that are typically extracted from stressed speech are called as Teager-energy operators and have been extracted in [33, 30] with its attributes mentioned in Table III.

TABLE III. ATTRIBUTES OF TEAGER-ENERGY OPERATOR

Extracted Features	Attributes
Teager-energy operator	• Useful for stress indication and analysis through measure of non-linear flow of energy • With robustness to noise, one of its application is recognizing the emotions of the driver in presence of car noise

D. Other features

Other features relevant to emotions that are extracted from speech include i) linguistic information of speech whose representative form includes the Bag-of-Words (BOW) [34], ii) discourse information that includes emotionally salient features [35], iii) Wavelets which are considered as an alternate to the Fourier transform and works on the Teager energy operator, and iv) modulation spectral features that capture both spectral and temporal characteristics of the speech signal by frequency analysis of the amplitude modulations across multiple acoustic bins [36]. In Table IV, the highlighting attributes of extracted features falling in this category are given.

TABLE IV. ATTRIBUTES OF OTHER EXTRACTED FEATURES

Extracted Features	Attributes
Bag-of-Words	• Neglects the negation constructions and syntactical relations in sentences [37]
Emotional Salience	• Helps obtain emotional keywords in the speech signal • Quantifies the information provided by a word that points to a given emotion
Wavelets	• Supports time-frequency localization • Good time resolution at high frequencies and good frequency resolution at low frequencies
Epoch	• Effective model for vocal fold's closure [38] • Contours of epochs helps in improving classification performance

IV. DIMENSIONALITY REDUCTION AND FEATURE CLASSIFICATION

A popularly used dimensionality reduction technique is Principal Component Analysis (PCA). However, the biggest drawback with this technique is that PCA involves obtaining a set of transformed features instead of a subset of the original features which does not even consider any class information (e.g. emotional class) while making the transformation [39]. This may pose a problem while attempting to differentiate between the different emotions based on the PCA output thus encouraging only unsupervised learning and not supervised learning. Yet another technique that is used for dimensionality reduction is Greedy Feature Selection (GFS) [40], which involves selecting the most representative feature among the input features iteratively and then eliminating its effect from the reduced input feature vector. Research says that GFS helps overcome overfitting and an extensions to this are Sequential Floating Forward Selection (SFFS) and Sequential Floating Backward Selection (SFBS). C. Lee and S. Narayan in their work [10, 35] have used the technique of Forward Selection (FS) along with PCA. Some of the other techniques for dimensionality reduction include elastic net [41] in which the number of input values are higher than the number of output values resulting in a novel shrinkage and selection method which like a 'stretchable fishing net' retains 'all the big fish' i.e. only those features that are relevant and important for feature classification. Also, fast Correlation-based filter [42] is used which employs the entropy based correlation that defines a symmetric uncertainty as a goodness measure and the selected relevant features are arranged in decreasing order.

It is important to design a classification model in such a manner that complexity is kept at bay so that the problem of 'overfitting' is not encountered frequently, i.e. the model should not be so simple that it cannot explain the differences between the categories (of emotions) yet not so complex as to give poor classification. Given the limited number of recorded speech signals, the entire database is split so that part of it is used for training the classifier and the remaining is used for evaluating its performance (which is rightfully called as 'validation'). Averaging the validation values over several splits leads to cross-validation and in general, the performance of a feature classifier can be evaluated using cross-validation. Some of the classification models that have been explored include Hidden Markov Model (HMM), Gaussian Mixture Models (GMM), Support Vector Machines (SVM), Naïve Bayes Classifier, Dynamic Time Warping (DTW), Linear Discriminant Classifier (LDC), etc.

T. Nwe et al in their work [16] used HMM as the classifier model and were of the opinion that HMM with four states delivered the best optimal performance. Their observation was supported by T. Pao et al in 2005 [43] with the intention of detecting emotions in Mandarin speech. This observation has not been challenged by anyone till date for the reason that a small number of states forces the HMM to group together observations that may introduce a substantial variance as against having too many states using which the HMM model may not be able to estimate the desired parameters of each state correctly. Further, T. Pao et al used LOO (Leave-One-Out) Cross Validation (LOO CV) to calculate the recognition rates. However, it has been observed that LOO gives high variance leading to unreliable estimates.

Given some speaker dependent vocal tract configuration, it is observed that the spectral shape (of the vocal tract) can be represented by the mean vector μ_i while variations can be represented by the covariance matrix Σ_i both of which characterize the GMM along with mixture weights w_i for a given emotional class *i*. GMM which is a single state HMM has one differentiating characteristic over HMM. Unlike HMM, GMM models the temporal structure of the training data. This may seem useful if the recognition of emotions is sequence dependent; however, this may not seem important given that emotion recognition from speech typically is text-independent and non-linguistic. This observation was supported by [44] and [45].

Support Vector Machines (SVM) involves recognising the desired emotions by defining a separating hyperplane in which the 'support vectors' are the training samples that are typically close to the hyperplane and are the patterns most useful for the classification. When compared with GMM, it has been observed that SVM learns the exact parameters for the global optimum something that GMM, which incorporates the Expectation Maximization (EM) algorithm, is unable to estimate. For creating a multi-class SVM, combination of binary SVM's are required; however the performance of multi-class SVM thereby formed has shown to be comparatively less as compared to its binary counterpart. This is where a new approach in the form of SVM ensemble has been proposed in [46].

Other classification models that have been proposed in the literature survey of ASER include Naive Bayes Classifier, Dynamic Time Warping (DTW), Linear Discriminant Classifier (LDC) etc. In Naive Bayes Classifier [43], given a class label *C*, it involves learning from the training data the conditional probability of each attribute A_i. However, it works under the assumption that every attribute is independent from the rest which could pose a problem when dealing with emotions from the same category (e.g. non-negative emotional category or negative emotional category). Similarly, DTW is a technique of non-linear warping alignment of a given test signal over the training signal to determine the similarities between them. This is one technique of feature classification that may appear to be time consuming given that every signal has to be compared with every other signal from a given database to check out for the minimum distance optimal path. LDC that finds usage in [10] and [35] is loosely related to GMM and Naïve Bayes Classifier as it classifies between the (emotional) classes using the Bayes rule of maximum probability applied on the estimated parameters of Gaussian distribution of mean and variance.

V. DISCUSSION

As can be observed from the range of average accuracies obtained in Table V, an appropriate choice of feature extraction, dimensionality reduction and suitable feature classification can help obtain an achievable accuracy. The fluctuations seen in the resultant outputs can be dedicated to the fact that different study groups attempted to extract different features from the speech signal followed by their choice of dimensionality reduction and feature classification techniques. Some of the study groups considered a gender independent ASER system while some like in [15] attempted to bifurcate the database into male and female utterances over which the ASER system was worked upon to accordingly achieve an accuracy of 79.9% and 89.02% for female and male utterance respectively. The difference in accuracy can be supported by the fact that with the features of pitch and MFCC being extracted, a non-linear analysis of speech signal with male pitch being lower than female pitch is occurring. Higher the pitch, greater should be the physical difference between the two tones to be perceived differently and if this is not being achieved, the huge difference between the accuracies with respect to the male and female utterances cannot be accounted for. Also, T. Seehapoch et al in [7] have used the statistical method of reducing the dimensionality in which they have attempted to calculate different statistical parameters like mean, variance, median, maximum and minimum per frame that represents an equivalent form of the features that they have extracted. This deviates from all the other existing dimensionality reduction techniques enlisted, highlighting the fact that only the best combination of techniques can help achieve the desired accuracy. This however cannot overlook the remaining techniques explicitly mentioned in this paper as some measurable accuracy has also been achieved using them.

Some of the future scopes of ASER to name a few can include futuristic hotels deploying robotic waiters deployed with the ASER system to timely gauge the reactions of the customers in the waiting period between the order and arrival of the dish using the information of which, these robots can alert the chefs to quicken the task, a classroom environment where both the teacher and students are present (deviating from e-learning), thereby gauging the level of confidence among the students based on the kind of replies generated by them and so on. An ASER system should be able to perform its task of recognizing the emotional upheavals of human speech with an achievable precision so as to improve the human-machine interaction.

TABLE V. PERFORMANCE ANALYSIS OF ASER

Study Group	Year	Database	Features extracted	Dimensionality reduction	Feature classifier	Accuracy obtained
C. Lee et al [10]	2001	Real users engaged with a machine agent	1. Pitch 2. Energy	Forward selection followed by PCA	LDC among others	77%
T. Vogt et al [8]	2005	Emo-DB and SmartKom Corpus	1. Pitch and its contours 2. Energy and its contours 3. MFCC	Correlation based feature selection	Naïve Bayes Classifier	77.4% for reduced set as against
C. Lee et al [35]	2005	Real users engaged with a machine agent	1. Acoustic features 2. Discourse information	Feature selection followed by PCA	LDC among others	An increase of 40.7% for male and 36.4% for female
M. Lugger et al [47]	2007	Emo-DB	1. articulation rate 2. ZCR 3. Voice quality	SFFS	GMM	74.6%
X. Cheng et al [15]	2012	Mandarin emotional speech database	1. Pitch 2. MFCC	PCA	GMM	79.9% (female) 89.02% (male)
T. Seehapoch et al [7]	2013	Berlin, Japanese and Thai	1. F_0 2. Energy 3. ZCR 4. MFCC	Statistical method	SVM	89.8%

REFERENCES

[1] J. Rong, G. Li and Y. Chen, “Acoustic feature selection for automatic emotion recognition from speech”, Information Processing and Management, vol. 45, no. 3, pp.315-328, May 2009.

[2] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Trans. on Acoustics, Speech and Signal processing, vol. 28, no. 4, pp. 357-366, Aug 1980.

[3] A. Batliner et al, “The automatic recognition of emotions in speech”, Emotion-Oriented Systems, pp. 71-99, Springer Berlin Heidelberg, 2011.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, “A database of German emotional speech”, INTERSPEECH, pp. 1- 4, 2005.

[5] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, “Using multiple databases for training in emotion recognition: to unite or to vote”, Intl. Science Congress Association, pp. 1553-1556, Aug 2011.

[6] Y. Pan, P. Shen and L. Shen, “Speech emotion recognition using Support Vector Machine”, Intl. Jour. of Smart Home, vol. 6, no. 2, Apr 2012.

[7] T. Seehapoch and S. Wongthanavasu, “Speech emotion recognition using Support Vector Machines”, 5th IEEE Intl. Conf. on Knowledge and Smart Technology (KST), pp. 86-91, Jan 2013.

[8] T. Vogt and E. Andre, “Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition”, IEEE Intl. Conf. on Multimedia and Expo, pp. 474-477, Jul 2005.

[9] B. Schuller, S. Steidl and A. Batliner, “The INTERSPEECH 2009 emotion challenge”, INTERSPEECH, pp. 312-315, Sep 2009.

[10] C. Lee, S. Narayanan and R. Pieraccini, “Recognition of negative emotions from speech signal”, IEEE Workshop on Automatic Speech and Understanding, pp. 240-243, 2001.

[11] J. Ang, R. Dhillon, A. Krupski, E. Shriberg and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog”, INTERSPEECH, 2002.

[12] D. Neiberg, K. Elenius, I. Karlsson and K. Laskowski, “Emotion recognition in spontaneous speech”, Lund Working Papers in Linguistics, vol. 52, pp. 101-104, 2006.

[13] A. Abelin and J. Allwood, “Cross linguistic interpretation of emotional prosody”, ISCA Tutorial and Workshop (ITRW) on Speech and Emotion, pp. 1-18, 2000.

[14] L. Yang and N. Campbell, “Linking form to meaning: the expression and recognition of emotions through prosody”, 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, Aug 2001.

[15] X. Cheng and Q. Duan, “Speech emotion recognition using Gaussian Mixture Model”, 2nd Intl. Conf. on Computer Application and System Modeling, pp. 1222-1225, 2012.

[16] T. New, S. Foo and L. De Silva, “Speech emotion recognition using Hidden Markov models”, Elsevier Speech Commun. Jour., vol. 41, no. 4, pp. 603-623, Nov 2003.

[17] J. Rong, Y. Chen, M. Chowdury and G. Li, “Acoustic features extraction for emotion recognition”, 6th IEEE/ACIS Intl. Conf. on Computer and Information Science, pp. 419-424, Jul 2007.

[18] K. Rao and S. Koolagudi, “Identification of Hindi dialects and emotions using spectral and prosodic features of speech”, IJSCI: Intl. Jour. of Systemics, Cybernetics and Informatics, vol. 9, no. 4, pp. 24-33, 2011.

[19] A. Kandali, A. Routray and T. Basu, “Vocal emotion recognition in five native languages of Assam using new wavelet features”, Intl. Jour. of Speech Technology, vol. 12, no. 1, pp. 1-13, Mar 2009.

[20] L. Rabiner and R. Schafer, Introduction to digital speech processing, Foundations and trends in signal processing, vol. 1, no. 1, pp. 1-194, Jan 2007.

[21] I. Luengo and E. Navas, “Feature analysis and evaluation for automatic emotion identification in speech”, IEEE Trans.on Multimedia, vol. 12, no. 6, pp. 267-270, Oct 2010.

[22] D. Reynolds and R. Rose, “Robust text-independent identification using Gaussian Mixture speaker models”, IEEE Trans. on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, Jan 1995.

[23] S. Chapaneri, “Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping”, *Intl. Jour. of Computer Applications*, vol. 40, no. 3, pp. 6-12, Feb 2012.

[24] K. Kumar, C. Kim and R. Stern, “Delta-spectral Cepstral Co-efficients for robust speech recognition”, IEEE Intl. Conf on Acoustics, Speech and Signal Processing, pp. 4784-4787, May 2011.

[25] S. Chapaneri and D. Jayaswal, “Emotion recognition from speech using Teager based DSCC features”, IJCA Proceedings on Intl. Conf. on Commun. Technology, ICCT-No.1, pp. 15-20, Oct 2013.

[26] M. Xu, N. Maddage, C. Xu, M. Kankanhalli and Q. Tian, “Creating audio keywords for event detection in soccer video”, Intl. Conf. on Multimedia and Expo, vol. 2, pp. II-281- II-284, Jul 2003.

[27] C. Drioli,G. Tisato, P. Cosi and F. Tesser, “Emotions and voice quality:experiments with sinusoidal modelling”, ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis, pp. 127-132, 2003.

[28] S. Patel, K. Scherer, E. Bjorkner and J. Sundberg, “Mapping emotions into acoustic space: the role of voice production”, Biological Psychology, vol. 87, pp. 93-98, 2011.

[29] C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha and L. Zhao, “Practical speech emotion recognition based on online learning: from acted data to elicited data”, Mathematical Problems in Engineering, Vol. 2013, pp. 1-9, Jun 2013.

[30] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: resources, features, and methods”, Speech Communication, vol. 48, no. 9, pp. 1162-1181, Apr 2006.

[31] B. Schuller, G. Rigoll and M. Lang, “Hidden Markov model-based speech emotion recognition”, Intl. Conf on Acoustics, Speech and Signal Processing, vol. 2, pp. II-1- II-4, Apr 2003.

[32] M. Kammaoun, D. Gargouri, M. Frikha and A. Hamida, “Cepstrum vs. LPC: a comparative study for speech formant frequencies estimation”, GESTS Intl. Trans. Commun and Signal Processing, vol. 9, no. 1, pp. 87-102, Oct 2006.

[33] G. Zhou, J. Hansen and J. Kaiser, “Nonlinear feature based classification of speech under stress”, IEEE Trans. on Speech and Audio Processing, vol. 9, no. 3, pp. 201-216, Mar 2001.

[34] B. Schuller, R. Muller, M. Lang and G. Rigoll, “Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles”, INTERSPEECH, pp. 805-808, 2005.

[35] C. Lee and S. Narayanan, “Towards detecting emotions in spoken dialogs”, IEEE Trans. on Speech and Audio Processing, vol. 13, no. 2, pp. 293-303, Mar 2005.

[36] S. Wu, T. Falk and W. Chan, “Automatic speech emotion recognition using modulation spectral features”, Speech Communication, vol. 53, no. 5, pp. 768-785, Sep 2010.

[37] A. Neviarouskaya, H. Prendinger and M. Ishizuka, “Textual affect sensing for sociable and expressive online communication”, Affective Computing and Intelligent Interaction, vol. 4738, pp. 220-231, 2007.

[38] S. Koolagudi, R. Reddy, S. Rao, "Emotion recognition from speech signal using epoch parameters," Intl. Conf. Signal Processing and Communications (SPCOM), pp.1-5, Jul 2010.

[39] J. Wagner, J. Kim and E. Andre, “From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification”, IEEE Intl. Conf. on Multimedia and Expo, pp. 940-943, Jul 2005.

[40] A. Farahat, A. Ghodsi and M. Kamel, “An efficient greedy method for unsupervised feature selection”, 11th IEEE Intl. Conf. on Data Mining, pp. 161-170, Dec 2011.

[41] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net”, Jour. of the Royal Statistical Society, vol. 67, no. 2, pp. 301-320, Apr 2005.

[42] L. Yu and H. Liu, “Feature selection for high-dimensional data: a fast correlation-based filter solution”, Proceedings of the 12th Intl. Conf. on Machine Learning, pp. 856-863, 2003.

[43] T. Pao, Y. Chen, J. Yeh and W. Liao, “Detecting emotions in Mandarin speech”, Computational Linguistics and Chinese Processing, vol. 10, no. 3, pp. 347-361, Sep 2005.

[44] E. Ayadi, M. Moataz, M. Kamel and F. Karray, “Speech emotion recognition using Gaussian Mixture vector autoregressive models”, IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, vol. 4, pp. IV- 957- IV- 960, Apr 2007.

[45] H. Tang, S. Chu, M. Johnson and T. Huang, “Emotion recognition from speech via boosted GMM”, IEEE Intl. Conf. on Multimedia and Expo, pp. 294-297, Jun 2009.

[46] H. Kim, S. Pang, H. Je, D. Kim and S. Bang, “Constructing Support Vector Machine ensemble”, Pattern Recognition, vol. 36, no. 12, pp. 2757-2767, 2003.

[47] M. Lugger and B. Yang,” An incremental analysis of different feature groups in speaker independent emotion recognition”, Intl. Conf. of Phonetic Sciences, 2007.