International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 17 (2018) pp. 13251-13257

Performance Analysis of Acoustic Features in Telugu Speech Emotion Recognition

[1,*]N. Ratna Kanth and [2]Dr. S. Saraswathi

[1]Department of Computer Science and Engineering, Pondicherry Engineering College, Pillaichavadi, Puducherry, 605014, India.

[2]Department of Information Technology, Pondicherry Engineering College, Pillaichavadi, Puducherry, 605014, India.

Abstract

This paper investigates the performance of various speech features in recognizing emotions from Telugu speech. A Telugu emotional speech corpus containing 497 samples with 7 emotional classes -anger, disgust, fear, happy, neutral, sad and surprise- is used for the study. Short term MFCC, Energy, Energy Entropy, Fundamental Frequency, Zero Crossing Rate, Spectral centroid, spectral rolloff, spectral entropy, spectral flux and harmonic features are extracted. These features are grouped into three categories – MFCC, Prosodic and Spectral-Harmonic. Phrase level features are computed from the short term features and used for the recognition task. Support vector machines (SVMs) are used for classification task. The recognition accuracies of each of the above groups of features and their combinations are evaluated using two standard approaches for multiclass classification task- One-vs-One (OVO) and Directed Acyclic Graph (DAG). Leave one speaker out (LOSO) cross validation is used and the accuracy is measured using precision and recall. Experimental results of selected feature groups and their combinations clearly indicate that maximum recognition rate can be achieved by using the combination of MFCC, Prosodic, Spectral and Harmonic features.

Keywords: Speech Emotion Recognition, Emotional Speech Corpus, Support Vector Machines, Cross Validation, One-vs-One and Directed Acyclic Graph.

INTRODUCTION

Speech is the most natural and fastest way of communication among human beings. In our daily communication we take the emotional state of the person with whom we are communicating into consideration and correspondingly we deal with them. In case of Man-Machine interaction this requires that the machine should have sufficient intelligence to recognize human voices and able to understand the emotional state of the person as well. From nineteen fifty's, there is a lot of research on speech recognition, and in spite of the great progress made in the area of speech recognition, we are still far away from a natural way of interaction between man and machine. This is because machine is unable to understand the emotional state of the speaker. This resulted in one of the recent research areas, namely Speech Emotion Recognition (SER). Speech Emotion Recognition is the process of automatic inferring of the emotional state of a person from his speech [1], [12].

There are several reasons which make speech emotion recognition very difficult and challenging. First, there is no clarity regarding which speech features are best suited to recognize emotions. Second, because of different sentences, speakers, styles of speaking, and rate of speaking acoustic variability is introduced. This is another obstacle because this is going to affect most commonly extracted features from speech such as pitch, and energy contours [1], [2]. Another challenging issue is cultural and environmental differences play an important role in how a particular emotion is expressed.

Speech emotion recognition is useful in a number of applications such as web movies, computer tutorial applications, in-car board safety systems [3], as a diagnostic tool for therapists [4], in automatic translation systems and in aircraft cockpits [5]. Speech emotion recognition has also been used in call centre applications and mobile communication [6].

The rest of the paper is organized as follows: Section 2 gives literature survey, Section 3 describes emotional speech corpuses and section 4 explains feature extraction and phrase level feature computation. Section 5 discusses implementation of OVO and DAG approaches, section 6 presents the results with discussion and section 7 gives conclusion.

LITERATURE SURVEY

One of the important issues in speech emotion recognition is selecting a set of important emotions that are to be considered for automatic emotion classification. Most of the emotions encountered in our lives are given by Linguists. A typical set containing 300 emotional states is given by Schubiger [1], [7] and O'Connor and Arnold [8]. But classifying such a large number of emotions is very difficult. There is a general agreement among most of the researchers that any emotion can be decomposed into primary emotions, just in the similar way that any color is a combination of some basic colors. Anger, Disgust, Fear, Joy, Sadness, and Surprise are considered as Primary emotions [9]. These are the most obvious and distinct emotions in our life. Another important issue in designing a speech emotion recognition system is extracting suitable features that can efficiently characterize

different emotions [1]. Proper selection of features significantly affects the classification performance.

One more important issue is what type of classifier or learning algorithm is used for speech emotion recognition. Artificial Neural Networks (ANNs), Decision Trees, Support Vector Machines (SVMs), Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) are the most preferred methods used for Speech Emotion Recognition.

Speech Emotion Recognition based on Hidden Markov Model has been presented in [10]. Language-independent emotion recognition system based on Modular Neural Network approach is used in [11] producing an overall classification accuracy of 83.31% on the test set.

Detecting emotion from nonverbal features of speech and applying the method to assess the public speaking skills using Support Vector Machines is proposed in [14]. Pair-wise classifiers are constructed for nine emotional classes and the average cross-validation accuracy achieved is 89% for the pair-wise machines while for the fused machine the accuracy is 86%.

Emotion recognition based on multiple classifiers using Acoustic-Prosodic (AP) information and Semantic Labels (SLs) is proposed in [15]. Emotion recognition performance based on MDT achieved 80%, which is better than each individual classifier, while average recognition accuracy of 80.92% is obtained for SL-based recognition. Combining acoustic-prosodic information and semantic labels achieved 83.55%, which is superior to either AP-based or SL-based approaches.

Global and local prosodic features extracted from sentence, word and syllables are used for speech emotion in [16], [17]. Duration, pitch, and energy values are used to represent the prosodic information, and studies are carried out on simulated Telugu emotion speech corpus (IITKGP-SESC). Support vector machines are used for developing the emotion models. Score level combination of energy, pitch and duration features, improved the emotion recognition performance to 64 %.

MFCC, Chroma, ZCR and prosodic features are used in [13]. Binary Support Vector Machines were used on German Emotional Speech Corpus EmoDB. An average accuracy of 92.25% for the Binary SVMs and 77.07% for the Multiclass SVM is achieved on the test set. On the same test set using the fused model has achieved an overall accuracy of 87.86%.

EMOTIONAL SPEECH CORPUSES

Douglas-Cowie et al [18]. and Ververidis and Kotropoulos [19] listed several emotional speech corpuses that are used by many researchers working on speech emotion recognition. Two of the most popular speech databases are Danish emotional speech database (DES) [20] and the German 'Berlin' [21] database, which are publicly available for research use. But there is a growing attention towards handling more spontaneous speech. One of the ways to collect emotional speech data spontaneously is recording of speakers dialogues during their interaction with other agents, where the interaction may elicit certain emotions. For example, Kismet [22] is an infant- and robot-directed database of speech, and BabyEars [23] contains recordings of parents expressing different emotions such as approval, attention, and prohibition which are elicited while they are talking to their children. Hansen and Bou-Ghazale [24] collected the speech under simulated and actual stress (SUSAS) database contained isolated-word sounds by 32 speakers produced in the conditions of stress and emotion, including both spontaneous and acted speech.

Collection of emotional speech in more naturally interacting settings is done by some researchers. For instance, Lee and Narayanan[25] recorded calls of customers in a live conversation, in a commercially installed call centre setting, where customer frustration may lead to negative emotions. Working with this type of data is especially difficult as spontaneous emotions are hard to distinguish and the proportion of data displaying particular emotions can be sparse with respect to the whole database. Another consideration is that these databases are not usually publicly available because of commercial and copyright issues. As well as limiting the possibility for researchers to use them, this also renders the results obtained problematic because they cannot easily be replicated and validated by others.

Telugu Emotional Speech Corpus (TESC)

For our present work an acted Telugu Emotional Speech Corpus (TESC) was prepared using 4 (2 male and 2 female) native Telugu post graduate students. These students were given sufficient training with practice sessions by playing them carefully selected audio and visual recordings from Telugu movies, radio plays of All India Radio and speech samples from IITKGP-SESC and EmoDB as well. Ten emotionally neutral sentences were considered for studying the emotions. The number of words and syllables in the sentences were varying from 3–7 and 10–20 respectively. Seven emotions were considered: Anger, Disgust, Fear, Happy, Sad, Surprise and Neutral. Entire corpus was recorded in four sessions with duration of one week between each session. This duration is allowed to capture the emotional variability of the speakers resulting from their personal circumstances and mood changes. Each of the participants had to speak the 10 sentences in 7 given emotions in one session. The total number of utterances in the database was 1120 (4 speakers x 4 sessions x 7 emotions x 10 sentences). Each emotion had 160 utterances. Out of 1120 samples 497 samples were selected in which the emotion was properly expressed as judged by five other post graduate students. Recording was done in a quiet room using Ahuja microphone and Jet Audio is used for audio recording. Recordings were sampled at a rate of 16000 Hz and saved in WAV format.

Table 1. Emotional classes, number of speakers and total number of samples for TESC

Emotional Speech Corpus	Language	No of Speakers	No of Emotions	Emotions	Total No of Samples
TESC	Telugu	4 (2 male + 2 female)	7	Anger, Disgust, Fear, Happy, Sad, Surprise, Neutral	497

SPEECH FEATURES FOR EMOTION RECOGNITION

Pitch, intensity, speaking rate and voice quality are identified in the early and frequently cited work of Murray and Arnott [26], [27] as the acoustic features largely affected by emotions in speech. Among these features, prosody and voice quality are identified as the most important features to differentiate between various emotions as per the human perception. Murray and Arnott assumed that the emotion will remain constant over the whole utterance because the recordings were from actors, and so it was treated as a single unit. Most of the subsequent studies accepted this assumption as the basis. Whether this assumption is appropriate or not has been questioned by researchers working on spontaneous speech. In spontaneous speech, emotions tend to be transitory; hence features must be calculated over smaller units than the entire utterances. Some studies have [28], [29] tested this by dividing the whole utterance into parts containing only voiced speech; features are then taken from these voiced segments alone. Features calculated from segments are sometimes referred to as short-term whereas those obtained from the whole utterance are called long-term [30]. Results to date show that attempts to use segment level (short-term) features alone for emotion classification have not been as successful as using utterance level (long-term) features only. However, both Shami and Verhelst [28] and Casale [29] et al. got better results by combining the two levels, indicating that although long-term features by themselves do best, it is unrealistic to expect emotion to be constant over a whole utterance.

Different researchers have used many different sets of features for developing their systems. For example, Yang and Lugger [31] have argued that prosodic features can separate emotion classes in the arousal dimension whereas voice quality features are more effective in separating classes in the valence dimension. Similarly, Eyben et al [32] have found mel-cepstral features most effective in the valence dimension. Given this diversity, it is unclear exactly what the 'best' features are to use.

Feature extraction

In general, for speech emotion recognition features related to pitch, loudness, frequency, energy, etc are considered. For extracting the features from the audio samples MATLAB Audio Analysis Library [33] is used. Short term MFCC, Energy, Energy Entropy, Fundamental Frequency, Zero Crossing Rate, Spectral centroid, spectral rolloff, spectral entropy, spectral flux and harmonic features are extracted for each speech sample by dividing it into a frame size of 20ms with a frame shift of 10ms.

Phrase level feature computation

From the short term features extracted, phrase level or sentence level features are computed. Various Phrase level features computed are minimum, maximum; mean, median, differences of maximum and minimum, mean and minimum etc., quartiles, variance, standard deviation, standard deviation/mean, standard deviation/median. A listing of the Low Level Descriptors (LLDs) and statistical features used here is given in Table 2.

Table 2. Description of Low Level Descriptors and statistical features derived

Feature Groups (6)	Low Level Descriptors(23)	Statistical features(15)
Cepstrum	MFCC 0-12	Minimum, Maximum, Mean,
Pitch	Fundamental	Median, Variance, Standard
Energy	Frequency (f_0)	Deviation,
	Log energy	minimum~maximum,
Spectral	Energy Entropy	minimum~mean,
	Spectral Centroid	minimum~median,
	Spectral Rolloff	maximum~mean,
	Spectral flux	maximum~median,
Zero	Spectral energy	quartile 1, quartile 3,
Crossing	ZCR	SD/Mean, SD/Median
Rate	Harmonic Ratio	
Harmonic		

Formation of feature groups

Phrase level features computed as explained in the previous subsection are categorized into three groups. First group MFCC containing all MFCC 0-12 features, second group Prosody containing pitch (fundamental frequency), log energy, energy entropy and zero crossing rate and third group Spectral-Harmonic consisting of spectral centroid, spectral flux, spectral roll of, spectral energy and harmonic ratio. Table 3 shows his division of features into three groups.

Table 3: Division of speech features into three groups

Group. No	Group Name	Features in Group
1.	MFCC	MFCC 0-12
2.	Prosody	Pitch or Fundamental Frequency (f_0) Log energy Energy Entropy Zero Crossing Rate (ZCR)
3.	Spectral-Harmonic	Spectral Centroid Spectral Rolloff Spectral flux Spectral energy Harmonic Ratio

METHODS USED FOR EMOTION RECOGNITION

There are two main approaches for multiclass classification problems: one considers all classes in one single optimization function and the other breaks the problem down into several binary classifications [36], [37]. In this work, we used the second approach. Fig. 1 & Fig. 2 show two standard approaches for the binary classifiers that are used in the present study. One of these is hierarchical and the other one is non-hierarchical.

Support vector machines were originally developed for binary classification; much subsequent work has been done to extend them to multiclass classification using schemes like those in Fig. 1. For instance, One-versus-One (OVO) classifier is described by Hsu and Lin [34] describe the, and the directed acyclic graph (DAG) by Platt et al. [35].

One-versus-One (OVO)

The OVO classifier uses a majority voting approach as shown in Fig. 1 which shows the case for four emotional classes. It builds m*(m-1)/2 binary classifiers for 'm' emotional classes, one for each pair of different classes possible. Data from only the i^{th} and j^{th} classes is used to train each binary classifier C_{ij}. To classify a given test sample x_i, if classifier C_{ij} predicts that it belongs to class i, then class i votes are increased by one; otherwise the class j votes are increased by one. At the end of this process, the majority voting approach assigns the test sample to the class with highest number of votes.

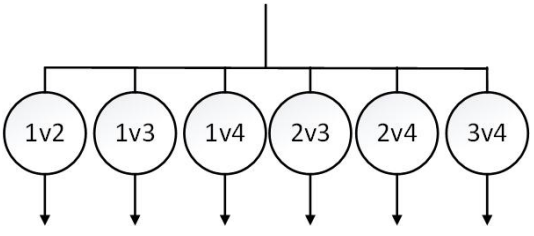

Figure 1. One-vs-One

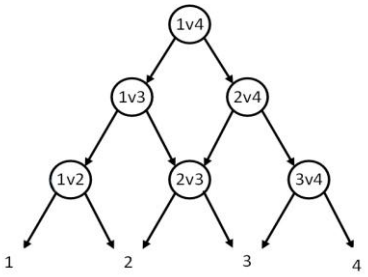

Figure 2. Directed Acyclic Graph

Directed Acyclic Graph (DAG)

The DAG method proposed by Platt et al. [35] also has to train m*(m-1)/2 binary classifiers for m classes. The training phase is the same as for OVO; however, in the testing phase, it uses a rooted binary directed acyclic graph with m*(m-1)/2 internal nodes and m leaves. Each node is a binary SVM classifier, C_{ij}, for i^{th} and j^{th} classes. For every test sample, starting at the root node, the sequence of binary decisions at each node determines a path to a leaf node that indicates the predicted class. Fig. 2 shows the DAG architecture for the classification of four classes. In this scheme, a test sample is correctly classified only if every single classification is correct.

In our work seven emotions are considered- anger, disgust, fear, happy, neutral, sad and surprise. So, both for OVO and DAG approaches 7*(7-1)/2 = 21 binary SVMs are constructed. Each binary SVM is trained with the features of speech samples belonging to the two emotional classes for which that SVM is built. For example, the binary SVM 2V3 is trained with the features of speech samples that belong to the emotional classes 2 and 3. To classify a test instance in OVO its feature vector is presented to all the 21 SVMs and their classifications are taken. A majority voting algorithm is used and the test instance is assigned the emotion that gets the highest number of votes. In DAG method to classify a test instance the rooted binary tree as shown is fig. 2 is traversed from the root node down to one of the leaf nodes by choosing the branch given by the classification of each node along the path.

PERFORMANCE EVALUATION

To measure the recognition accuracies of various feature groups considered and their combinations, two measures are used: Precision and Recall. A brief description of these two measures follows.

Precision and Recall

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. It is the number of positive predictions divided by the total number of positive class values predicted.

Precision = True Positives / (True Positives + False Positives)

Table 4. Precision and Recall of feature groups and their combinations

Feature Group	OVO		DAG	
	Precision	Recall	Precision	Recall
MFCC	51.70	50.31	51.47	52.11
Prosody	50.20	42.09	46.94	43.20
Spectral-Harmonic	43.66	46.31	43.35	46.61
MFCC +Prosody	51.47	53.35	54.46	55.29
MFCC + Spectral-Harmonic	50.62	52.67	50.44	53.71
Prosody + Spectral-Harmonic	45.27	44.24	45.11	42.72
MFCC + Prosody + Spectral-Harmonic	**54.70**	**54.35**	**57.61**	**57.13**

Recall is the number of true positives divided by the number of true positives plus the number of false negatives. Put another way it is the number of positive predictions divided by the number of positive class values in the test data.

Recall = True Positives / (True Positives + False Negatives)

Leave One Speaker Out (LOSO) cross validation

Leave one speaker out also known as speaker independent cross validation is used to measure the recognition accuracy. In this method all the speech samples of one speaker are set aside for testing while the models are trained with all the speech samples of the rest of the speakers. This process is repeated for all the speakers and the average values are taken. If there are 'n' speakers this corresponds to an n-fold cross validation. As the speech corpus employed in this work contains speech samples of 4 speakers, this corresponds to a 4-fold cross validation.

Table 5. Speaker wise Precision and Recall of feature groups and their combinations

Method	Feature Group	Speaker 1		Speaker 2		Speaker 3		Speaker 4	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
OVO	MFCC	56.39	53.06	38.72	37.85	42.49	50.46	69.22	59.87
	Prosody	37.73	32.51	56.70	45	48.76	45.29	57.62	45.58
	Spectral-Harmonic	46.58	49.08	46.78	46.42	35.50	40.01	45.78	49.74
	MFCC +Prosody	52.05	54.16	47.98	47.14	45.50	53.09	60.36	59.02
	MFCC + Spectral-Harmonic	57.92	61.35	50.28	44.28	43.65	50.30	50.62	54.74
	Prosody + Spectral-Harmonic	40.33	38.40	40.68	41.42	57.26	48.96	42.79	48.18
	MFCC + Prosody + Spectral-Harmonic	57.85	61.67	51.19	46.43	59.25	54.01	50.51	55.32
DAG	MFCC	55.62	53.06	41.17	41.42	44.33	51.35	64.76	62.59
	Prosody	37.65	32.51	58.30	45	33.75	44.27	58.05	51.03
	Spectral-Harmonic	46.30	48.96	49.20	48.57	32.64	39.17	45.25	49.74
	MFCC +Prosody	52.87	54.11	53.11	48.57	63.79	56.03	64.05	62.46
	MFCC + Spectral-Harmonic	39.63	37.51	40.61	41.42	56.52	42.33	43.68	49.61
	Prosody + Spectral-Harmonic	39.63	37.51	40.61	41.42	56.52	42.33	43.68	49.61
	MFCC + Prosody + Spectral-Harmonic	60.12	60.72	56.16	53.57	60.01	54.01	54.17	59.03

Results of the leave one speaker out cross validation are shown in table 4 and table 5. Table 4 shows overall Precision and Recall of feature groups and their combinations Table 5 shows speaker wise Precision and Recall of feature groups and their combinations. From table 4 it can be observed that out of the three groups MFCC, Prosody and Spectral-Harmonic, MFCC is giving highest accuracy. Among the three combinations of the groups- MFCC + Prosody, MFCC + Spectral-Harmonic, Prosody + Spectral-Harmonic- the combination of MFCC + Prosody is giving more accuracy which is better than MFCC alone. Finally the combination of MFCC + Prosody + Spectral-Harmonic performed well over all others with a precision of 54.7 %, recall of 54.35% for OVO approach and precision of 57.61%, recall of 57.13% for DAG approach.

CONCLUSION

Performance of various speech features in recognizing emotions from Telugu speech is analyzed by grouping the features into three groups – MFCC, Prosodic and Spectral & Harmonic. The recognition accuracies of each of these groups of features and their combinations are evaluated using two standard approaches for multiclass classification task- One-vs-One (OVO) and Directed Acyclic Graph (DAG). Leave one speaker out (LOSO) cross validation is used and the accuracy is measured using precision and recall. Experimental results of selected feature groups and their combinations clearly indicate that maximum recognition rate can be achieved by using the combination of MFCC, Prosodic, Spectral and Harmonic features.

REFERENCES

[1] Moataz ElAyadi, MohamedS.Kamel, FakhriKarray, 2011, “Survey on speech emotion recognition: Features, classification schemes, and databases”, Pattern Recognition Vol. 44, pp. 572–587.

[2] R. Banse, K. Scherer, 1996, “Acoustic profiles in vocal emotion expression”, J. Pers. Soc. Psychol. Vol. 70 (3), pp. 614–636.

[3] B. Schuller, G. Rigoll, M. Lang, 2004, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture”, Proc. ICASSP 2004, Vol. 1, pp. 577–580.

[4] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, 2000, “Acoustical properties of speech as indicators of depression and suicidal risk”, IEEE Trans. Biomedical Eng. ,Vol. 47 (7), pp. 829–837.

[5] J. Hansen, D. Cairns, Icarus, 1995, “source generator based real-time recognition of speech in noisy stressful and Lombard effect environments”, Speech Commun., Vol. 16 (4) pp. 391–422.

[6] J. Ma, H. Jin, L. Yang, J. Tsai, 2006, “Ubiquitous Intelligence and Computing: Third International Conference, UIC 2006”, Proceedings (Lecture Notes in Computer Science), Springer-Verlag, NewYork, Inc., Secaucus, NJ, USA.

[7] M. Schubiger, 1958, “English intonation: its form and function”, Niemeyer, Tubingen, Germany.

[8] J. O’Connor, G. Arnold, 1973, “Intonation of Colloquial English”, Second ed., Longman, London, UK.

[9] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J. Taylor, 2001, “Emotion recognition in human–computer interaction”, IEEE Signal Process. Vol. 18, pp. 32–80.

[10] Björn Schuller, Gerhard Rigoll, and Manfred Lang, 2003, “Hidden Markov Model-Based Speech Emotion Recognition”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[11] Muhammed Waqas Bhatti, Yongjin Wang and Ling Guan, 2004, “A Neural Network Approach for Human Emotion Recognition in Speech”, International Symposium on Circuits and Systems (ISCAS), IEEE.

[12] N. Ratna Kanth, S. Saraswathi, 2014, "A Survey on Speech Emotion Recognition”, Advances in Computer Science and Information Technology (ACSIT), Vol. 1(3), pp. 135-139.

[13] N. Ratna Kanth, S. Saraswathi, 2015, "Efficient Speech Emotion Recognition Using Binary Support Vector Machines & Multiclass SVM", 2015 IEEE International Conference on Computational Intelligence and Computing Research Speech.

[14] Tomas Pfister and Peter Robinson, 2011, “Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis”, IEEE Transactions on Affective Computing, Vol. 2 (2), pp. 66-78.

[15] Chung-Hsien Wu, Wei-Bin Liang, 2011, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels”, IEEE Transactions on Affective Computing, Vol. 2 (1), pp. 10-21.

[16] K. Sreenivasa Rao, Shashidhar G. Koolagudi, Ramu Reddy Vempada, 2013, “Emotion recognition from speech using global and local prosodic features”, Int J Speech Technology, 16, pp.143–160.

[17] Sreenivasa Rao Krothapalli, Shashidhar G. Koolagudi, 2013, “Characterization and recognition of emotions from speech using excitation source information”, Int J Speech Technology, 16, pp.181–201.

[18] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003, “Emotional speech: towards a new generation of databases”, Speech Communication 40 (2), pp.33–60.

[19] Ververidis, D., Kotropoulos, C., 2006, “Emotional speech recognition: resources, features, and methods”, Speech Communication 48 (9), pp.1162–1181.

[20] Engberg, I., Hansen, A., 1996, "Documentation of the Danish emotional speech database DES", Center for Person Kommunikation, Institute of Electronic Systems, Alborg University, Aalborg, Denmark.

[21] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005, "A database of German emotional speech", in Interspeech'05, Proceedings of 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, pp. 1517–1520.

[22] Breazeal, C., Aryananda, L., 2002, "Recognition of affective communicative intent in robot-directed speech", Autonomous Robots 12 (1), pp.83–104.

[23] Slaney, M., McRoberts, G., 2003, "BabyEars: a recognition system for affective vocalizations", Speech Communication 39 (3–4), pp.367–384.

[24] Hansen, J., Bou-Ghazale, S., 1997, "Getting started with SUSAS: a speech under simulated and actual stress database", in Proceedings of 5th European Conference on Speech Communication and Technology, Eurospeech'97, Rhodes, Greece, pp. 1743–1746.

[25] Lee, C.M., Narayanan, S.S., 2005, "Toward detecting emotions in spoken dialogs", IEEE Transactions on Speech and Audio Processing 13 (2), pp.293–303.

[26] Murray, I.R., Arnott, J.L., 1993, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", Journal of the Acoustical Society of America 93 (2), pp.1097–1108.

[27] Murray, I.R., Arnott, J.L., 2008, "Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech", Computer Speech and Language 22 (2), pp.107–129.

[28] Shami, M., Verhelst, W., 2007, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech", Speech Communication 49 (3), pp.201–212.

[29] Casale, S., Russo, A., Scebba, G., Serrano, S., 2008, "Speech emotion classification using machine learning algorithms", in IEEE International Conference on Semantic Computing, ICSC'08, Santa Clara, CA, pp. 158–165.

[30] Li, Y., Zhao, Y., 1998, "Recognizing emotions in speech using short-term and long-term features", in Proceedings of Fifth International Conference on Spoken Language Processing, ICSLP'98, Sydney, Australia, pp. 2255–2258.

[31] Yang, B., Lugger, M., 2010, "Emotion recognition from speech signals using new harmony features", Signal Processing 90 (5), pp.1415–1423.

[32] Eyben, F., Batliner, A., Schuller, B., Seppi, D., Steidl, S., 2010, "Cross corpus classification of realistic emotions: some pilot experiments", in 7th International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, pp. 77–82.

[33] Theodoros Giannakopoulos, Aggelos Pikrakis, 2014, "Introduction to Audio Analysis: A MatLab Approach", First edition, Academic Press, pp 77-95.

[34] Hsu, C., Lin, C., 2001, "A comparison of methods for multi-class support vector machines", IEEE Transactions on Neural Networks 13 (2), pp 415– 425.

[35] Platt, J., Cristianini, N., Shawe-Taylor, J., 2000, "Large margin DAGs for multiclass classification", in Proceedings of Neural Information Processing Systems, NIPS'99, Denver, CO, pp. 547–553.

[36] A. Hassan, R.I. Damper, 2012, "Classification of emotional speech using 3DEC hierarchical classifier", Speech Communication 54, pp 903–916.

[37] Chi-Chun Lee, Emily Mower et. al., 2011, "Emotion recognition using a hierarchical binary decision tree approach", Speech Communication 53, pp 1162–1171.