Automatic Speech Emotion Detection System using Multi-domain Acoustic Feature Selection and Classification Models

Nancy Semwal
Bhabha Atomic Research Centre
Visakhapatnam, India
nancys@barc.gov.in

Abhijeet Kumar, Sakthivel Narayanan
Bhabha Atomic Research Centre
Visakhapatnam, India
abhijeetk@barc.gov.in, svel@barc.gov.in

Abstract

Emotions exhibited by a speaker can be detected by analyzing his/her speech, facial expressions and gestures or by combining these properties. This paper concentrates on determining the emotional state from speech signals. Various acoustic features such as energy, zero crossing rate(ZCR), fundamental frequency, Mel Frequency Cepstral Coefficients (MFCCs), etc are extracted for short term, overlapping frames derived from the speech signal. A feature vector for every utterance is then constructed by analyzing the global statistics (mean, median, etc) of the extracted features over all frames. To select a subset of useful features from the full candidate feature vector, sequential backward selection (SBS) method is used with k-fold cross validation. Detection of emotion in the samples is done by classifying their respective feature vectors into classes, using either a pre-trained Support Vector Machine (SVM) model or Linear Discriminant Analysis (LDA) classifier. This approach is tested with two acted emotional databases – Berlin Database of Emotional Speech (EmoDB), and RML Emotion Database (RED). For multi class classification, accuracy of 80% for EmoDB and 73% for RED is achieved which are higher than or comparable to previous works on both the databases.

1. Introduction

With the rapid advancements being made in artificial intelligence, one of the foremost motives being to enrich the experience of human machine interactions, there has been a surge of interest in the field of emotion detection [1]. It has not only been an active area of research in academic fields such as psychology, neuroscience, psychiatry, cognitive sciences etc but also finds use in numerous practical applications, viz. call centers, gaming industry, medical fields, etc [1] [2] [3].

Speech made by a speaker conveys its linguistic meaning as well as the feeling with which the speech is delivered. As such, it is an efficient medium to detect the emotions. Emotion detection is often done by using both facial features and speech signals [2] [4]. Speech energy, speech rate, loudness, pitch, tone etc are some of the various prosodic characteristics which vary with our emotions and hence are strong cues to differentiate between them. This paper focuses on deducing the emotion in a speech sample by analyzing low level descriptors (LLD) that represent the emotional state of the speaker.

In the literature, a speech based emotion recognition system encompasses three major stages [1] [2] [5]-[8]: *Feature extraction* stage is required to derive such features from the speech signals which are representative of the emotional content present in the speech. Along with various temporal and spectral characteristics, MFCCs are the most utilized feature for emotion recognition [2] [6].A majority of previous works have utilized various softwares, like PRAAT, OpenSMILE, etc for the purpose of extracting features from speech files [5] [7] [9] [10]; *Feature selection* techniques such as sequential feature selection, genetic algorithms, principal component analysis, etc pick up the most discriminatory features from the exhaustive list of extracted features. The WEKA toolkit has been widely used in many implementations for picking best features [6] [8] [10]; *Classifier modeling and prediction* stage learns the subspace of the emotions from the features of the training dataset and performs classification on the samples in the test dataset based on its learning. Related research works have used a variety of learning techniques such as k-Means, principal component analysis (PCA), k-Nearest Neighbors (KNN), SVM, multilayer perceptron (MLP), naïve Bayes classifier, etc for modeling [1]-[3],[5]-[11].

This paper follows the described approach and implements the three stages to build an automatic system for emotion detection from speech. Acoustic features have been extracted from the speeches and SBS has been utilized for elimination of irrelevant features. For model construction, two techniques have been explored — LDA and SVM. The approach has been tested by exercising the repeated 10-fold cross validation method. On the first database, EmoDB [12], the accuracy figures reported by this implementation are comparable to [2] [6] [7]. For RED [13], the classification results are better than [4] and [14]. Particular attention has been given to the capability of the approach for multiclass classification.

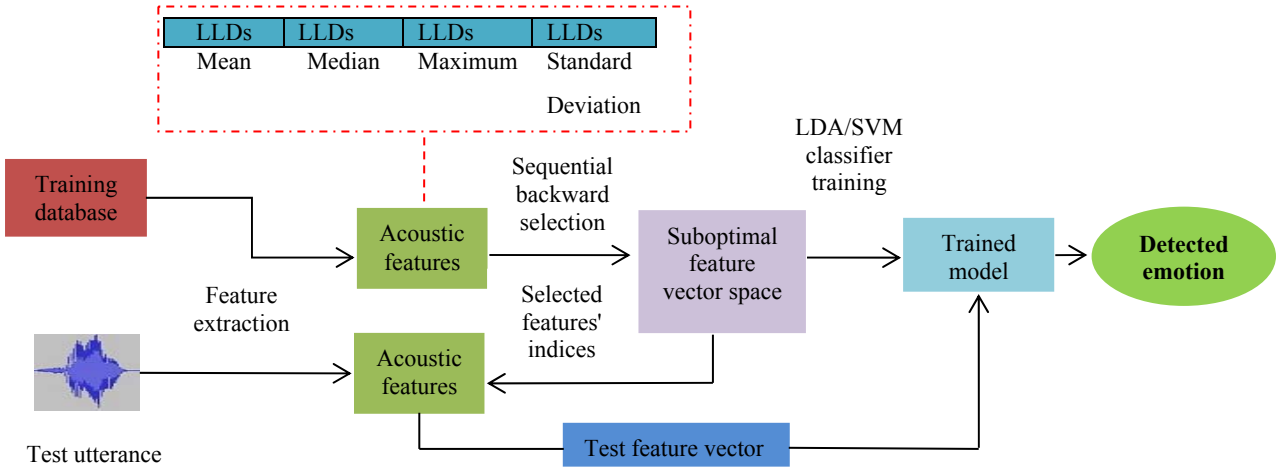

Figure1. Flow diagram of the implemented approach for emotion detection system

The paper is organized as follows. Section 2 presents the proposed methodology for emotion recognition from speech; Section 3 describes the experiments performed to measure the accuracy of the system; the experimental results have been discussed in Section 4; Section 5 provides the conclusion and discusses future work.

2. Proposed Methodology

Figure 1 presents an overview of the adopted approach in our implementation.

2.1. Front End Processing

The front end processing comprises of 3 stages in order to build feature vector space. The process has been discussed below.

2.1.1 Preprocessing

Preprocessing is a crucial step in almost every speech processing application, like speaker recognition, speech recognition, etc. It increases the efficiency of subsequent feature extraction and classification stages and therefore to improve the overall recognition performance. Voice activity detection (VAD) is a widely performed preprocessing step in speech applications.

VAD detects speech segments while removes silences or non-speech fragments from the signal. This greatly enhances the quality of the speech signal and leads to better feature extraction [6]. In this implementation, we explored a two-pass, segment based unsupervised technique for VAD, presented in [15]. In the first pass of the method, high energy segments are detected and if no pitch is detected within any of these segments, that segment is considered to be a high energy noise segment. In the next pass, noise reduction is applied to the speech sample and SNR weighted energy difference is then applied to the noiseless speech for VAD. Thus we obtain speech containing only speech information without any noise. It is to be noted that VAD was performed in this study only on RED, since it was observed that the speech samples in this database contain silences in the beginning and at the end.

2.1.2 Feature Extraction

Feature extraction derives various low level descriptors (LLDs), typically from time, spectral and cepstral domain, for short speech frames of the speech signal. It is found that gross statistics such as mean, variance, maximum, etc are always more useful in emotion analysis than the low level features themselves [5] [16]. Thus, feature extraction determines global descriptors (GD) over the LLDs across all the short frames in the signal. Aggregation of the GDs provides us with a feature vector for each signal.

Table 1. Extracted Low level descriptors

Temporal LLDs	Spectral LLDs	Cepstral LLDs
Zero crossing rate	Spectral centroid and spread	17 Mel frequency cepstral coeffs. (MFCCs)
Frame energy	Spectral entropy	1st derivative (deltas) of MFCCs
Frame entropy	Spectral flux	
	Spectral roll off	
	Fundamental freq.	
	Chroma vector	
	Harmonic	

Table 1 shown above gives a domain-wise list of the LLDs utilized in our implementation. For the GDs, mean, median, maximum and standard deviation were empirically determined to be the best statistics to be used. Unlike

numerous previous works which extract features by use of toolkits, we intend to extract features on the fly in order to build a fully automatic system. Usage of various features has been explored with reference to [17].

2.1.3 Feature Selection Algorithm

Feature selection is required to pick the most discriminatory features from the total feature space. It is a general observation that on adding features beyond an optimum set, the features either become *redundant* (leading to no increase in system accuracy by their inclusion) or *deteriorating* (degrading accuracy by getting included). Feature selection finds this optimum set. This also reduces the dimensionality of the feature vector space, providing speed up in further processing. SBS is a sequential search algorithm for feature selection, which seeks to find the feature subset from full candidate feature set by maximizing the value of criterion in iterations. We choose initial criterion as the classification accuracy. This accuracy gets determined by the respective classifier being used (LDA or SVM), with k-cross fold validation exploiting the full candidate feature set. SBS keeps on removing features sequentially, unless criterion decreases upon a feature's removal.

```
   SBS (feature_set, accuracy)
1  features ← select full feature_set;
2  Initialize m ← 1;

3  while m <= number of features
4        remove m^th feature from feature_set;
5        determine criterion with remaining features;
6        if criteria >= accuracy
7              accuracy = criterion;
8              remove m^th feature permanently
9        m ← m + 1;
```

Figure 2. Sequential backward search algorithm

Figure 2 summarizes the key steps of SBS with the complete feature set *feature_set* and the classification rate *accuracy* on the whole set as inputs.

2.2. Classification Models

2.2.1 Linear Discriminant Analysis

LDA first projects a feature vector space into a new rotated space such that the between-class variance is maximized and the within-class variance is minimized. The axes of this rotated space are the eigen vectors (sorted in descending order) of the following general eigen values equation

$$S_B v = \lambda S_W v \quad (1)$$

, where λ is the diagonal matrix of eigen values and S_B and S_W are the between-class and within-class matrices respectively. Equation 2 and 3 give the expression for determining S_B and S_W.

$$S_B = \sum_{i=1}^{c} M_i (x_i - \mu)(x_i - \mu)^T \quad (2)$$

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \quad (3)$$

Here, c is the number of emotions; X_i is the set of speech samples, M_i is the number of samples and μ_i the mean vector of the samples belonging to the i^{th} emotion.

After determining the components that best separate the classes, the training samples are projected to reduced LDA space. The space is n – 1 dimensional, where n is the number of classes. LDA performs classification by assuming that different classes generate data based on different Gaussian distributions. Training the linear discriminant classifier is to estimate the parameters of the Gaussian distributions for each emotion class. A test sample x_k is assigned to the class with maximum likelihood f_i, calculated as

$$f_i = \mu_i C^{-1} x_k^T - \frac{1}{2} \mu_i C^{-1} \mu_i^T + \ln P_i \quad (4)$$

, where μ_i is the mean vector of i^{th} emotion, C is the pooled covariance matrix, and P_i is the prior probability for i^{th} emotion (equal for all emotions).

2.2.2 Support Vector Machines

An SVM transforms a set of feature vectors, each marked as belonging to one of the two categories, into a high dimensional space such that the vectors, now mapped into the new space, have been separated into their respective category by a gap as wide as possible. The transformation is done by using *kernel functions.* SVMs are highly effective even in the case of feature dimensions being higher than the number of samples, making them highly suitable in this study. SVMs perform classification for a test feature vector 'x' based on a decision function, *sgn(),* defined in Equation 5.

$$sgn(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + p) \quad (5)$$

Here, n is the number of training vectors, x_i is the i^{th} training vector, y_i ε [1,-1] is the class of the i^{th} vector, α is the regularization parameter, $K(x_i, x)$ is the kernel and p is an independent term. The *sgn()* function is actually determining the distance of the test vector from the boundaries of the classes that have been separated by a hyper plane. A positive value of the function for a class indicates that the test vector is predicted to be in that class; a negative value indicates otherwise. SVM, which is essentially a binary class classifier, performs multi class classification by using the "one-versus-one" or "one-versus-rest" technique. For this study, the SVM

implementation in LIBSVM library was used [18]. C-Support Vector Classifier (C-SVC) with linear kernel function was utilized to perform SVM classification. LIBSVM uses one-versus-one technique for multiclass classification.

3. Experimental Setup

3.1 Databases

The approach was tested on the following two databases:

1) Berlin Database of Emotional Speech (EmoDB) – EmoDB consists of 535 files from 7 emotions – anger, anxiety, boredom, disgust, happiness, neutral and sadness. The database is in German language and there are 10 speakers – 5 female and 5 male. Each speaker speaks the same 10 sentences in the 6 different emotional states. All the audio files have been sampled at 16 KHz. In this paper, we have excluded the emotion disgust from the experiments, thus taking a total of 489 files.

2) RML Emotion Database (RED) – RED is an audio-visual database and it consists of 720 files from 6 emotions, each emotion having 120 files. The emotional states are anger, disgust, fear, happiness, sadness and surprise. 8 speakers, all male, have spoken 10 different sentences for each emotion. The languages in this database are English, Mandarin, Urdu, Punjabi, Persian and Italian. The audios have been extracted from the audio-visual files, with a sampling rate of 16 KHz. VAD has been performed on RED as explained in Section 2.1.1.

3.2 Feature vector space

Feature extraction was performed on the databases, followed by feature selection, as discussed in Section 2.1.2 and 2.1.3 respectively. From the total 224 features, feature selection stage picked 171 and 177 features for EmoDB and RED respectively.

3.3 Experiments

In order to evaluate the performance of the implemented approach, the following experiments were performed:

1) One versus one emotion classification by SVM.
2) Multi class classification by SVM and LDA.

Both LDA and SVM classifiers initially need to be trained to perform a correct classification and the trained classifiers subsequently need to be tested for accuracy measurement. Thus, we require partitioning the data training and testing datasets without any overlap between them. K-fold cross validation, as introduced in Section 2.1.3, is an efficient way to carry out this division.The accuracy of the experiments in our implementation has been measured by repeated 10-fold cross validation. This produces a more reliable estimate of the performance of the classifier than simply performing k-fold cross validation once [17].

4. Results and Discussions

We report the results of the experiments conducted for emotion detection task on both the databases.

4.1 Berlin Database (EmoDB)

Table 2 shows the accuracy percentages for one versus one emotion classification between 15 pairs of emotions. The classification has been carried out utilizing SVM classifier, performed on EmoDB. The emotion labels in the table are as follows -**A**ngry (127 files), An**X**iety (69 files), **B**oredom (81 files), **H**appiness (71 files), **N**eutral (79 files) and **S**adness (62 files).

Table 2. One versus one classification by SVM on EmoDB.

	A	**X**	**B**	**H**	**N**	**S**
A	-	93.87%	99.03%	76.26%	100%	100%
X	-	-	95.33%	91.43%	91.22%	99.24%
B	-	-	-	98.68%	85.62%	95.80%
H	-	-	-	-	96%	100%
N	-	-	-	-	-	95.74%

It is evident from Table 2 that the approach performs well in distinguishing emotions from one another. 100% accuracy is achieved in 3 pairs. These emotion pairs contain emotions which are exact opposite of one another in terms of arousal, which explains the high classification accuracy [18]. The only pair with a considerable misclassification is anger and happiness, with only 76.26% classification accuracy. Both of these emotions are high arousal emotions and both produce high excitation in the voice of the speaker, thus they get misclassified as one another [6] [19].

In our experiments, LDA classifier did not perform well for two class classification. One possible explanation for such behavior can be sparse amount of training data in such higher dimensional feature space which is not statistically significant for applying LDA. When such is the case, the classifier can get over fitted on the training samples, not giving a good generalization of the decision boundary, and as a result producing poor result for the new test samples. On the other hand, SVM eliminates this over fitting problem by regularizing the dimensions with the regularization parameter α (Equation 5).

Table 3 and Table 4 present the results of multi class classification by the LDA classifier and SVM classifier respectively. The columns of the matrix present the classification of the emotions, with the diagonal entries

representing the classification accuracy. For example, for the emotion anger, SVM correctly classified 107 files out of the total 127 files, making the positive rate or recall for the emotion 84.2%. By taking the sum of the diagonal elements of the confusion matrix as true positives, it can be seen that the detection rate of the system using LDA classifier is 78.11% while that of the system using SVM classifier stands at 80%. Considerable overlap is seen between emotions anger & happiness, and boredom & neutral in Figure 3, which shows the feature vectors for all emotions with the first two LDA components as axes**.** These LDA components are a linear combination of the extracted features. The confusion between the emotions in the latter pair can be associated to the fact that both are low arousal emotions and many features show similar behavior for these two [16].

Table 3. Confusion matrix for multiclass classification by LDA on EmoDB

	A	**X**	**B**	**H**	**N**	**S**
A	**104**	2	0	20	0	0
X	8	**56**	2	6	5	0
B	0	0	**63**	0	14	2
H	14	7	2	**43**	2	0
N	1	3	10	2	**58**	2
S	0	1	4	0	2	**58**

Table 4. Confusion matrix for multiclass classification by SVM on EmoDB

	A	**X**	**B**	**H**	**N**	**S**
A	**107**	3	0	16	0	0
X	1	**59**	2	5	8	1
B	0	0	**67**	0	9	3
H	19	4	3	**48**	2	0
N	0	3	5	2	**59**	0
S	0	0	4	0	1	**58**

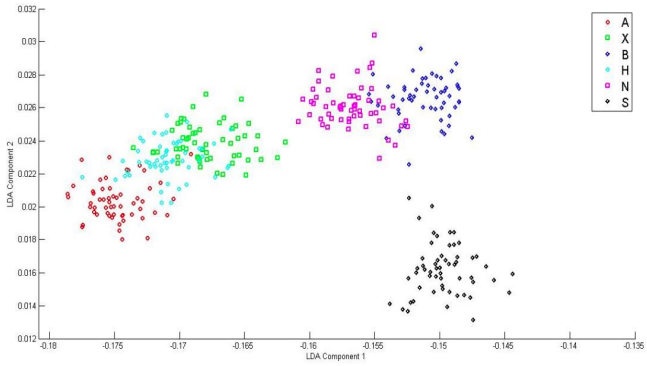

Figure 3. Feature vectors of the emotions in EmoDB shown by 2 LDA components as axes.

4.2 RML Emotion Database (RED)

The results of multiclass classification by LDA and SVM classifier on RED are shown in Table 5 and 6. The emotion labels in the table are as follows - Angry, Disgust, Fear, Happiness, Sadness and Surprise. Unlike EmoDB, every emotion has equal number of files in RED.

For RED, SVM classifier has an average accuracy of 73.33% while LDA classifier has an average accuracy of 71.81% with repeated K-cross fold validation on the database. Emotions disgust, fear and happiness overlap with each other despite having different values of arousal [19]. This misclassification for RED can be attributed to the presence of many short utterances (< 1 seconds) in the database for these emotions. The feature extracted for such small utterances may not be a correct representation of the emotion due to lack of sufficient statistics. Accuracy figures for emotions anger, surprise and sadness obtained are 82%, 81% and 81% respectively. Thus these three emotions are better recognized than the others. Figure 4 shows the feature vectors for all emotions with the first two LDA components as axes. These LDA components are a linear combination of the extracted features. As discussed in this section, well separated emotions (anger, surprised and sadness) and the overlapped emotions (disgust, fear and happiness) can be observed in this figure.

Table 5. Confusion matrix for multiclass classification by LDA on RED

	A	**D**	**F**	**H**	**S**	**Su**
A	**99**	2	10	7	0	11
D	1	**82**	14	17	4	3
F	8	13	**71**	8	14	5
H	6	13	8	**80**	14	2
S	0	9	17	8	**88**	2
Su	6	1	0	0	0	**97**

Table 6. Confusion matrix for multiclass classification by SVM on RED

	A	**D**	**F**	**H**	**S**	**Su**
A	**98**	3	10	10	0	9
D	5	**84**	14	17	5	4
F	5	11	**77**	9	12	3
H	3	13	7	**75**	6	5
S	0	7	12	8	**97**	2
Su	9	2	0	1	0	**97**

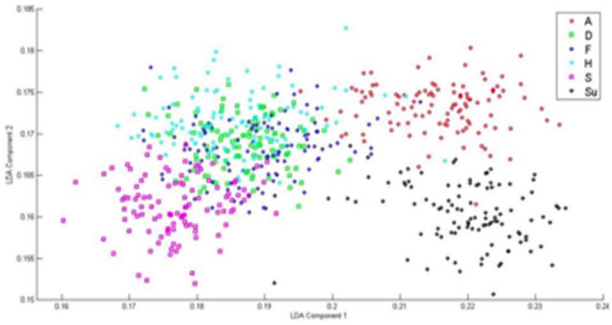

Figure 4. Feature vectors of the emotions in RED shown by 2 LDA components as axes.

5. Conclusion

In this paper we presented an automatic system for emotion recognition from speech. Global features are derived on the fly by combining the mean, median, maximum and standard deviation of LLDs from temporal, spectral and cepstral domain. SBS algorithm is employed to select emotion conveying global features from the entire feature set. Emotion recognition is performed by using two different classifiers – LDA and SVM. With the experiments performed using LDA classifier, an accuracy of 78% for EmoDB and 71% for RED was achieved. Also, in LDA space, the emotions were represented in only 5 feature dimensions. Further, the results illustrated that SVM classifier performs well for both binary class as well as multi class classification. For binary class classification, high accuracy figures (>90%) were achieved for 13 out of 15 emotion pairs. We obtained an accuracy of 80% and 73% for EmoDB and RED respectively for multi class classification with SVM.

The approach can be enhanced in the future to perform cross dataset emotion recognition, wherein the training and testing files originate from entirely different databases. Such a system would be universal and independent of the emotion database. Also, the presented approach should be tested to recognize emotions from purely spontaneous speech samples instead of only acted emotions, to test the effective of the approach in natural emotional speech.

References

[1] V. Petrushin, "Emotion in speech: Recognition and application to call centers", in Proc. Artificial Neural Networks in Engg, vol. 710, St.Lious, MO, 1999, pp 7-10.

[2] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese. A.Sciarrone, "Gender-driven emotion recognition through speech signals for ambient intelligence application", IEEE Trans. Emerging Topics in Computing, vol. 1, no. 2, pp. 244-257, 2014

[3] A. Davletcharova, S. Sugathan, B. Abraham, A.P. James, "Detection and analysis of emotion from speech signals", Procedia Computer Science 58, pp. 91-96,2015.

[4] M.K. Aditia, G.K. Verma, "Spontaneous Affect Recognition from Audio-visual Cues using Multi-resolution Analysis", International Journal of Information & Computation Technology, vol. 4, no. 17, pp. 1739-1745, 2014.

[5] P.L. Otero, L.D. Fernandez, C.G. Mateo, "iVectors for continuous emotion recognition", in Proc. of Interspeech 2014, pages 31-40, 2014.

[6] M.Bhargava, T.Polzehl, "Improving automatic emotion recognition from speech using rhythm and temporal feature", 2013 [Online]. Available: https://arxiv.org/abs/1303.1761.

[7] E. Dimitrieva, K. Nikitin, "Design of Automatic Speech Emotion Recognition System", in Proc. of the International Workshop on Applications in Information Technology, Aizu-Wakamatsu, Japan, Oct.2015.

[8] T. Vogt, E. Andre, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation", in Proc. of Int'l Conf. of Language Resources and Evaluation, 2006.

[9] A. Mordokovich, K. Veit and D. Zilber, "Detecting Emotion in Human Speech", [online] Dec 16, 2011.

[10] T. Vogt, E. Andre, N. Bee, "Emovoice - a framework for online recognition of emotions from voice", in Proc. of IEEE PIT 2008, vol. 5078, pp. 188-199, June 2008

[11] S·. Hawlett, "Emotion Detection from Speech", 2007, [Online].Available:http://cs229.stanford.edu/proj2007/Shah Hewlett%20-%20Emotion%20Detection%20from%20Speech.pdf

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A Database of German Emotional Speech", in Proceedings of lnterspeech, 2005, pp1517-1520.

[13] Y. Wang, L. Guan, "Recognizing human emotion from audiovisual signals", IEEE Transactions on Multimedia, Vol. 10, No. 5, pp. 936-946, August 2008.

[14] Zhibing Xie, "Audiovisual Emotion Recognition Using Entropy-estimation-based Multimodal Information Fusion", Ph.D dissertation, Electrical and Computing Dept, Ryerson University, Ontario, Canada, 2015.

[15] Z.H. Tan, B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection", IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pp. 798-807,2010.

[16] C. Busso, S. Lee, S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection", IEEE Transactions on Audio, Speech, and Language processing, vol. 17, no. 4, may 2009.

[17] T. Giannakopoulos, A. Pikrakis, "Audio features" in Introduction to Audio Analysis - A MATLAB approach, 1st edition, AP, 2014, ch.4, pp.59-103.

[18] C.C. Chang, C.J. Lin, "LIBSVM – A library for support vector machines", 2001[Online]. Available: http://www.csie.ntu.edu.tw/-cjlin/libsvm/

[19] M.E. Mena, "Emotion Recognition from Speech Signals", Erasmus Exchange Project Work, Ljubljana, 2012