ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

Speaker Independent and text Independent Emotion Recognition System Based on Random Forest Classifier

May Mon Lynn[1], Chaw Su[2]

Assistant Lecturer, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), PyinOoLwin, Myanmar[1]

Associate Professor, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), PyinOoLwin, Myanmar[2]

ABSTRACT: Recently, attention of the emotional speech signals research has been boosted in human machine interfaces due to availability of high computation capability. There are many systems proposed in the literature to identify the emotional state through speech. Selection of suitable feature sets, design of a proper classifications methods and prepare an appropriate dataset are the main key issues of speech emotion recognition systems. This paper focuses on feature selection method, which aims to extract effective acoustics, features to improve the performance of emotion recognition. For the extracted features of speech signal, the combination of MFCC (Mels Frequency Cepstral Coefficients) with Berouti Spectral Subtraction is used to obtain compressed feature vectors without losing much information. Random Forest classifier is used to classify the emotions according to the feature database. The modelling technique is speaker-independent and text-independent. This system is carried out with Matlab2016a and the overall system accuracy is about 68.3% corresponding to the created dataset.

KEYWORDS: Emotion Recognition, Database, MFCCs, Random Forest.

I. INTRODUCTION

The recognition of the human emotional plays an important role in the field of interpersonal relationships [1]. Although human can sense emotion easily, the machine cannot do like that. Therefore, emotion recognition system is intended to improve machine and human communication using knowledge-based emotion [2].

Speech emotion recognition is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics and medical analysis [3]. Recognizing of emotional conditions in speech signals are so challengeable area for several reason. First issue of all speech emotional methods is selecting the best features, which is powerful enough to distinguish between different emotions. The presence of various language, accent, sentences, speaking style, speakers also add another difficulty because these characteristics directly change most of the extracted features include pitch, energy [4].

Furthermore, it is possible to have a more than one specific emotion at the same in the same speech signal, each emotion correlate with a different part of speech signals. Therefore, defines the boundaries between parts of emotion in very challenging task. The majority of works are concentrated on monolingual emotion recognition, and making a presumption that there are no cultural diversity between utterers. However, the multi-lingual emotion classification process have been considered in some research [5].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

II. RELATED WORKS

Many researcher have been proposed the emotion recognition method. The following table I shows a few of lists.

TABLE I. LIST OF RELATED WORKS

Ref:No	Emotion Types	Modelling Techniques	Feature Extraction Method	Classifier	Accuracy	Remark
6	happy, sad, angry and neutral	speaker dependent and text independent case.	MFCC	GMM	44.04% for male 43.385% for female	the database taken is small consisting of only 120 sentences from 10 different speakers. Recognition rate can be higher if we work with a larger database.
7	Angr, Disgust, Fear, Happy, Sad, Surprise, Neural	(i)Speaker dependent but text independent (ii) Speaker independent and text independent	MFCC	GMM	76.2% 69.6%	the average mean-success-scores of experiment (ii) in all cases remained lower than their corresponding values in experiment (i).
8	Angry, happy, sad, neutral		MFCC	ANN	85	increasing training data can improve the accuracy but also increase processing time
9	Happy, Sad, Anger, Afraid and Surprise	-	MFCC	LDA	-	not able to recognize the happy emotion with better accuracy
10	Anger, Disgust,Fear, Joy, Sadness andSurprise	Speaker dependent but text independent	LFPC	hidden Markov models	79.9% (Burmese) 76.4% (Mandarin)	A total of 12 speakers contribute 720 emotion utterances. The system performance was tested on unseen text.
Prposed method	Angry,disgust,fear, happy,sad and surprise	Speaker independent and text independent	MFCC	Random Forest	68.3 %	Dataset is created from myanamr movies and the system performance was also tested on unseen text.

Table Ishows the utilized models along with the number of emotion types, modelling techniques, their recognition rates and significant remarks. And also present the results of my current research.

This paper presents about the speech emotion recognition from speech using acoustic features. The three methods; Berouti spectral subtraction, MFCC and Random Forest are used for speech enhancement, feature extraction and classification respectively.

This paper is organized as follows, the introduction about the system and related works are presented in section I. Section II describes Emotion Recognition System. Speech enhancement method, feature extraction method and classification method are represented in Section III, Section IV and Section V respectively. In Section VI, the overview of the data collection and experimental results is described and the conclusion will be followed.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

III. EMOTION RECOGNITION SYSTEM

The speech emotion recognition system contains five main modules emotional speech input, feature extraction, feature selection, classification, and recognized emotional output. The structure of the speech emotion recognition is as shown in Figure 1.

Fig.1. Emotion Recognition System

The overview of the proposed system design is presented in Fig. 2. Here is the step-by-step evaluation of all of the proposed methods: enhancement method, feature extraction method, and classification method.

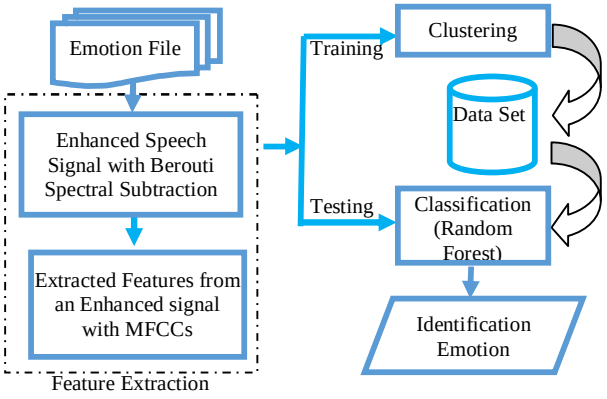

Fig.2. Overview of the Proposed System Design

IV. THEORETICAL BACKGROUND FOR PROPOSED SYSTEM

Step 1: The purpose of the speech enhancement algorithm is to enhance the quality of the noisy speech signal using various types of enhancement algorithms. Among enhancement algorithms, the spectral subtraction method is ease to estimate noise [11].

There are many spectral subtraction algorithms normally used for noise removal like magnitude spectral subtraction, power spectral subtraction, Berouti spectral subtraction, multiband spectral subtraction and so on. Berouti improves the noise reduction compared to the basic spectral subtraction, so this method is used in this system.

The values of the input signal are described by the continuous number. These values can change corresponding to the types of emotions. Therefore the extraction of features from these values are an important factor in speech emotion recognition system [12]. References [13, 14] indicated that pitch, power, interval, formant, Mel frequency cepstrum coefficient (MFCC), and linear prediction cepstrum coefficient (LPCC) are the important features. The reference [15] argued that statistics relating to MFCCs also carry emotional information.

Therefore, MFCC is used to extract the features for this recognition system. In real-world applications, the performance of MFCC degrades rapidly because of the noise [16]. Since 1980, notable efforts have been carried out to enhance MFCC feature in noisy environments. Therefore, the purpose of the system is to overcome the weaknesses of MFCCs in the noisy speech that the input signal is enhanced with speech enhancement algorithm firstly.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

A. *Spectral Subtraction*

The basic principle of spectral subtraction is to subtract an estimate of the average noise spectrum from noisy speech magnitude spectrum [17]. In equation (1) show how to evaluate the basic spectral subtraction.

$$x(n)= s(n)+d(n) \quad \text{eq. (1)}$$

where, x(n)= noise-corrupted signal, d(n)=additive noise and S(n)= clean speech signal. Fig.2 shows a general form of basic spectral subtraction [18].

Taking the DTFT on both sides of equation 1 we get

$$X(w) = S(w) + D(w) \quad \text{eq. (2)}$$

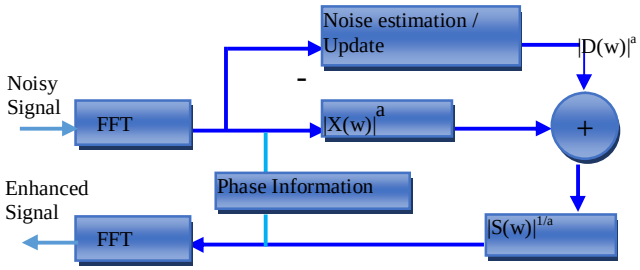

Fig.3. The general form of Spectral Subtraction

The polar form of equation 3 is given as

$$X(w) = |X(w)|e^{j\theta x(w)} \quad \text{eq. (3)}$$

Where X(w) is the magnitude spectrum and $\theta w(x)$ is the phase of noisy signal. Similarly, for noise spectrum, the polar form is defined as below

$$D(w) = |D(w)|e^{j\theta d(w)} \quad \text{eq. (4)}$$

|D(w)| is a magnitude of noise spectrum generally unknown but it can be replaced with an average value computed during non-speech activity. Applying these assumptions to equation 2 we get an estimate of the clean signal spectrum.

$$\hat{S}(w) = (|X(w)| - |\hat{D}(w)|)e^{j\theta x(w)} \quad \text{eq. (5)}$$

$|\hat{D}(w)|$ is the estimated average magnitude spectrum of noise measured during non-speech activity. The enhanced speech signal can be obtained by taking IDFTof $\hat{S}(w)$ [18].

- *Berouti Spectral Subtraction*: The main problem of basic spectral subtraction is the preservation of musical noise after subtraction. This musical noise is harder to hear that the original noise. Berouti improves noise suppression compared to basic spectral subtraction. It introduces an over-subtraction factor (α) and spectral floor parameter (β) and it is defined as [16].

$$|\hat{X}(w)|^2 = \begin{cases} |\hat{Y}(w)|^2 - \alpha|\hat{D}(w)|^2 & \text{if } |\hat{Y}(w)|^2 > (\alpha+\beta)|\hat{D}(w)|^2 \\ \beta|\hat{D}(w)|^2 & \text{else} \end{cases} \quad \text{eq. (6)}$$

β controls the amount of remaining residual noise and the amount of perceived musical noise.

B. *Mel Frequency Cepstral Coefficient (MFCCs)*

The second purpose of investigating spectral features is using mel frequency cepstrum coefficients (MFCC). It has superior performance compared to other methods and it is widely used for speech recognition [12].

The integration of berouti spectral subtraction and MFCC makes the feature vector robust and compact. The MFCCs

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

feature extraction process including the following step [19].

1) Pre-emphasis:

$$Y[n] = X[n] - 0.95\,X[n-1] \quad \text{eq. (7)}$$

2) Framing: The frame is segmented into small speech samples frames 20 to 40 milliseconds in length.

3) Hamming Windowing: If the window is defined as W (n), $0 \le n \le N-1$

$$Y[n] = X(n) \times W(n) \quad \text{eq. (8)}$$

Where, N = number of samples in each frame, Y[n] = output signal, X (n) = input signal, W (n) = Hamming window

4) Fast Fourier Transform (FFT): FFT is used to convert each frame of N samples from time domain to frequency domain.

$$Y[w] = FFT[h(t) * X(t)] = H(w) * X(w) \quad \text{eq. (9)}$$

5) Mel Filter Bank Processing

$$F(Mel) = [2595 * \log 10 [1 + f]700] \quad \text{eq. (10)}$$

6) Discrete Cosine Transform: DCT is a process of transforming a log-mel spectrum into the time domain using a discrete cosine transform (DCT).

In the speech emotion recognition system, the best features are provided to the classifier after the calculation of the features. The tasks of the classifier recognize the emotion in the speaker's speech utterance. There are many types of classifiers proposed for speech emotion recognition purposes. Each classifier has several advantages and disadvantages compared to other classifiers.

C. *Random Forest*

In the speech emotion recognition system, the best features are provided to the classifier after the calculation of the features. The tasks of classifier recognizes the emotion in the speaker's speech utterance. There are many types of classifiers proposed for speech emotion recognition purposes. K-nearest neighbors (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Gaussian Mixtures Model (GMM), Artificial Neural Network (ANN), etc. are the classifiers used in the speech emotion recognition system. Each classifier has several advantages and disadvantages compared to other classifiers.

The random forest classifier is a combination of decision tree. Each tree is generated using a random vector sampled by the input vector regardless of the training set.

Leo Breiman [20] proposed that random forest is a group of raw classifications or regression trees formed from random selection of samples in training data and a random process is selected in the derivation process. Prediction is made in majority voting to classify the predictions of ensemble. Random Forest Classifier is carried out with the following step [21].

1) Random Record Selection : We apply K iteration of bagging to create total K number of trees.

2) Random Variable Selection: Now for each of the K sample training set, we apply the attribute bagging and learn the decision tree said that the variable from any new node is the best variable among the extracted random subspace.

Several prediction variables (eg, m) are randomly chosen from all prediction variables and the best division for this m is used to divide the nodes. By default, m is the square root of the total number of all predictor variables in the classification.

Each tree is grown as described in [22,23]:

•By randomly sample N, If the replacement N is the number of cases in the training set, from the original data. This sample will be used as the training set for growing the tree.

•When dividing a node, for variable number of input variables, m << M is specified at each node and variable m is chosen such that m variables are arbitrarily selected from M, node division. when the forest growing, the value of m is kept constant.

•Each tree is as large as possible, pruning is not used.

The general random forest algorithm shows a significant performance improvement over singletree classifier such as C4.5. The generalization error rate is advantageous over Adaboost, but it is more robust to noise.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

V. DATABASE AND EXPERIMENTAL RESULTS

A. *Database*

Databases play a vital role for automatic emotion recognition as the rest statistical methods are learned using examples. The databases used till know in the research acted, Elicited and real life or natural emotions. The most popular examples of acted database are Berlin Database of emotional speech which comprised of 5 male and 5 female actresses and the Danish Emotional speech corpus (DES)[24].

Russian database consists of ten pronounces sentences from61 speakers (12 male 49 female) of age group 16-28years expressing six emotions viz., Happy, sad, angry, fear, neutral and disgust. The example of Induced database is SmartKom corpus and the German Aibo emotion corpus without knowing the people that their emotions are being recorded. The call center communication by Devillers and et al. is obtained from live recordings and is an example of real emotional database. Other examples includes Surrey Audio Visual Expressed Emotion (SAVEE) which comprised of 4 male actors expressing 7 different emotions. The Speech Under Simulate and Actual Stress (SUSA) database of 32 speakers where speech was recorded in both simulated stress and actual stress.

This paper used Myanmar speech dataset created by ourselves. Data collection is carried out from Myanmar movies by using Sony Sound Forge Pro 10. The emotion speech file is recorded in WAV file (16 bits, mono) with 44.1 kHz sampling rate. Its average length is five seconds. It takes about eight months to get all recording files (totally 1100 files).

B. *Experimental Results*

A dataset of 1100 audios from different Myanmar movies, each size about five seconds is used to analyze the system performance. The input speech file is firstly enhanced with Berouti Spectral Subtraction. According to the SNR values comparison, the SNR values of Berouti are higher than that of basic spectral subtraction. Therefore, the enhanced signal from Berouti spectral subtraction algorithm is chosen as input to extract the feature.

In Berouti, the parameter α and β are used to justify the values of noise in the speech signal. After testing the various value of the parameter, the noise free output signal can be obtained by using (α=5) and (β=0.005).

And then the significant features are extracted using Mel Frequency Spectral Coefficient (MFCC). In MFCC, framing and overlapping factor are 0.38 milliseconds and 0.4 milliseconds are used for best feature extraction results.

For classification types of emotion, Random Forest classifier is used. In the process of Random Forest classifier, the number of trees are needed to consider. Therefore, the system overall accuracy is tested with different number of tree. If the number of trees is between 20 to 29, the overall accuracy is 47%. Alternatively, if the number of trees is 30, the overall accuracy is about 68%. However, when the number of trees is 31 to 35, the overall accuracy is about 49%. According to the above results, it can be seen that the achievement score is better than the other is if the number of tree is 30.

The following table II shows the comparison results of recognition rate with enhanced features and direct features. It can be seen thatour proposed Berouti shows better results than basic spectral subtraction on created dataset with Random Forest Classifier.

TABLE II. SURVEY ON ANALYSIS RESULT FOR RANDOM FOREST CLASSIFIER WITH EHNANCED FEATURE AND DIRECT FEATURE

Emotion Types	Recognition Rate(%) with Enhanced Features	Recognition Rate(%) with direct Features
Angry	86.7	45
Disgust	80	60.5
Fear	73.3	40
Happy	73.3	40
Sad	80	55
Surprise	76.7	45

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

TABLE III shows the confusion matrix of emotional dataset based on created own dataset.

TABLE III. SURVEY ON CONFUSION MATRIC FOR RANDOM FOREST CLASSIFIER

Emotion Types	Emotion Recognition (%)					
	Angry	Disgust	Happy	Fear	Sad	Surprise
Angry	76	6.7	4	-	13.3	-
Disgust	19.3	70.7	3.3	-	3.3	3.3
Happy	6.7	16.7	68.3	-	8.3	-
Fear	7.7	15.4	-	61.5	15.4	-
Sad	13.3	13.3	3.3	-	70	-
Surprise	5	5	15	5	5	65

The following bar chart shows the recognition rate of random forest classifier. Its overall accuracy is about 68.3%.

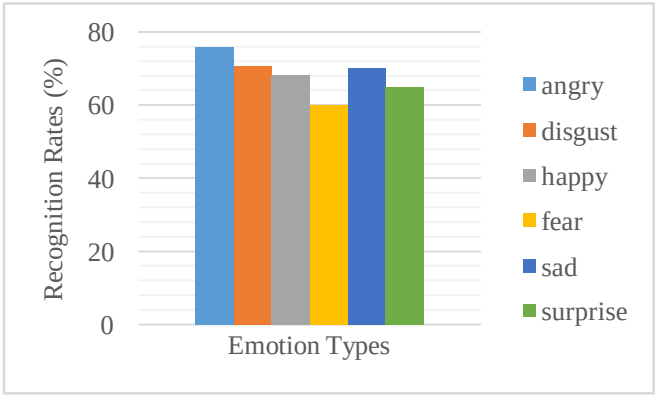

Fig.4. Survey on Recognition Rate of Random Forest Classifier

VI. CONCLUSION AND FUTURE WORK

This system refers to the area of automatic emotion recognition from speech signals. In this system, an emotional speech database from Myanmar movies is employed. Therefore the noise reduction method is needed and used to remove background music. The advantage of this system can tolerate the noise and control the real world application of emotion detection system. And then, the system intend to address any domain of emotion recognition system by modeling speaker independent and text independent system.

From experimentation and results, it is proved that the extracted features after the enhancement process are successfully developed for the training and classification system.

Currently, the enhanced feature recognition rates on the created dataset are about 68.3% for Random Forest .As reported by previous evaluation results, the recognition rate of enhanced features are more accurate than only MFCCs features for each emotion.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

REFERENCES

[1] Vishal B. Waghmare, RatnadeepR.Deshmukh, PukhrajP.Shrishrimal, Ganesh B. Janvale,"Emotion recognition system from artificial marathi speech using MFCC and LDA techniques", International Conference on Adcances in communication, Network and Computing, Elsevier, 2014.
[2] I. Chiriacescu , "Automatic emotion analysis based on speech" , M.Sc. THESIS Delft University of Technology, 2009.
[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.
[4] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," Pers. Soc. Psychol, vol. 70, no. 3, pp. 572–587, 1996.
[5] V. Hozjan and Z. Kačič, "Context-Independent Multilingual Emotion Recognition from Speech Signals," Int. J. Speech Technol., vol. 6, no. 3, pp. 311–320, 2003.
[6] N.JyotiGogoi1, J. Kalita2. "Emotion Recognition from Acted Assamese Speech." International Journal of Innovative Research in Science,Engineering and Technology, Vol. 4, Issue 6, June 2015.
[7] Aditya Bihar Kandali,"Emotion recognition from Assamese speeches using MFCC features and GMM classifier" IEEE, and Tapan Kumar Basu, INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR, june 17, 2009.
[8] Ms. Swati Shinde, Prof. Mrs. Swati Shilaskar, " Speech Based Emotion Recognition Using MFCC and ANN" International Journal of Computer Application (Special issue- Issue 5, Volume 2 (January 2015)
[9] Vishal B. Waghmare, RatnadeepR.Deshmukh, PukhrajP.Shrishrimal, Ganesh B. Janvale,"Emotion Recognition System from Artificial Marathi Speech using MFCC and LDA Techniques", International Conference on Adcances in communication, Network and Computing, Elsevier, 2014.
[10] T. L. New, S. W. Foo and L. C. De Silva, "Speech emotion recognition using hidden Markov models", Speech Communication., vol. 41, pp. 603–623, 2003.
[11] Ravi Bolimera, Siva Prasad Nandyala, T.Kishorekumar," Speech Enhancement using spectral subtraction, Affine projection Algorithms and classical Adaptive filters", EEC 2012.
[12] MukeshRana,SaloniMiglani,"Performance analysis of MFCC and LPCC techniques in automatic speech recognition", Volume -3 Issue-8 August, 2014 ISSN:2319-7242
[13] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, "Speech emotion recognition using hidden markov model", Eurospeech, 2001.
[14] P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.
[15] S.Kim, P.Georgiou, S.Lee, S.Narayanan. "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece, 2007
[16] N. S. Nehe and R. S. Holambe, "Isolated Word Recognition Using Normalized Teager Energy Cepstral Features," in Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT '09. International Conference on, 2009, pp. 106-110.
[17] A.M.Kondoz, Digital Speech, 2nd Ed., Wiley India Pvt. Ltd., 2007.
[18] R. Martin, "Spectral Subtraction Based on Minimum Statistics", in Proc. EUSPICO'94, pp. 1181-1185, 1994
[19] Ms. Swati Shinde, Prof. Mrs. Swati Shilaskar,"Speech based emotion recognition using MFCC and ANN", International Journal of Computer Application, ISSN:2250-1797, Issue-5, Volume 2 (January 2015)
[20] Breiman, L., Random Forests, Machine Learning 45(1), 5-32,2001.
[21] http:// www.stat.berkeley.edu /~breiman/ RandomForests/ cc_home.htm# prox Symposium, volume 1, July, 2005
[22] https://www.youtube.com/watch?v=2Mg8QD0F1dQ
[23] http: //www.stat.berkeley.edu/~breiman/ Random-Forests/ cc_home.htm#prox Symposium, volume 1, July, 2005.
[24] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A Database of German Emotional Speech. INTERSPEECH.