A Study of Support Vector Machines for Emotional Speech Recognition

Nattapong Kurpukdee *†, Sawit Kasuriya †, Vataya Chunwijitra †,
Chai Wutiwiwatchai † and Poonlap Lamsrichan *

* TAIST Tokyo Tech, ICTES Program, Kasetsart University, Thailand,
† NECTEC, National Science and Technology Development Agency (NSTDA),
112 Pahonyothin Road, Pathumthani, 12120, Thailand
Email: {nattapong.kurpukdee, sawit.kasuriya, vataya.chunwijitra, chai.wutiwiwatchai}@nectec.or.th,
fengpll@ku.ac.th

***Abstract*—In this paper, efficiency comparison of Support Vector Machines (SVM) and Binary Support Vector Machines (BSVM) techniques in utterance-based emotion recognition is studied. Acoustic features including energy, Mel-frequency cepstral coefficients (MFCC), Perceptual linear predictive (PLP), Filter bank (FBANK*), pitch, their first and second derivatives are used as frame-based features. Four basic emotions including anger, happiness, neutral and sadness in Interactive Emotional Dyadic Motion Capture (IEMOCAP) database are selected for training and evaluating in our experiments. The best accuracy of emotional speech recognition is 58.40% in average from SVM with polynomial kernel. Energy features combination with FBANK, pitch and their first and second derivatives features are the most suitable for computing utterance feature. Binary Support Vector Machines (BSVM) techniques show accuracy improvement in some emotions, such as sadness and happiness emotion.**

Keywords—Emotional Speech Recognition (ESR) and Classification, Utterance features, SVM (Support Vector Machines), Binary Support Vector Machines (BSVM)

I. INTRODUCTION

Speech is obviously one of the most natural, convenient, and important ways of human communication. Speech can easily convey not only the linguistic information but also the non-linguistic one such as emotions. It would be helpful if a computer could recognize the emotion implied in a given utterance. Consequently, speech emotion recognition (SER), which aims to recognize emotion states from speech signals, is one of the efforts which have been widely developed in these decades. Emotion recognizer can be applied in a wide range of applications. For example, the household robot can recognize the emotion from the human speech for friendly interaction with humans [1]. Customer care interactions can also use emotion recognition systems to assess customer satisfaction and quality of service [2]. In e-learning system, the emotion data of the person can be analyzed, and hence the learning contents are automatically adjusted for improving the learning efficiency [3]. All these applications benefit from the understanding the emotion of a speaker. However, we are now far from having a natural communication because the machine does not understand the emotional state of the speaker.

There are two major factors in the speech emotion recognition process; feature extraction and emotion classification. Feature extraction is the important part to extract the characteristic of the emotional state of speech. Many research aim to intercept the important feature of speech such as pitch [4], energy [4], Zero-crossing rate (ZCR) [4], Mel-frequency cepstral coefficients (MFCC) [4] [5], formant frequency [6], Perceptual Linear Prediction (PLP) [7]. In [8], the combination of acoustic features with lexical features are proposed and also demonstrated the accuracy improvement. However, these features are sequential short-term estimates of emotion states extracted using the frame-based methods. In fact, the remarkable characteristics of emotion patterns are the long-term utterance-level dynamics rather than the short-term fluctuations. Consequently, the global utterance-level feature is an attractive pattern for analyzing and representing the emotion in speech.

In emotion classification, many researchers have explored several methods [9] [10] [11], such as K-nearest neighbor (k-NN) [12], Naive Bayes [4], support vector machines (SVM) [13], Neural network (NN) [14] [15], Gaussian Mixture Model (GMM) [16], Hidden Markov model (HMM) [17]. There are some research works have utilized hierarchical tree structure in performing emotion recognition tasks [18]. The key idea behind this proposed emotion recognition framework is the use of binary classifiers in a hierarchical tree structure. There are many well-established states of the art classifiers that can be readily implemented to work with binary classification problems, e.g., logistic regression, Support Vector Machines, Fisher discriminant analysis, etc. SVM is a famous classifier and has been used in several studies to detect emotion from speech [5] [6] [13] [18] [19]. Therefore, it is interesting to use SVM as the classifier in the binary tree structure that called binary SVM (BSVM).

In this paper, we investigate the efficiency of BSVM with the utterance-based feature. In our first step of the framework, we perform a preprocessing step in the input speech signal by separating voiced and unvoiced part. Then, only the speech data is extracted to the utterance-based feature. After that, the feature is flowed down each node in the binary SVM tree to classify between pairs of emotion classes. Finally, each speech utterance is labeled in one of emotions.

This paper is organized as follows: we first briefly introduce the feature extraction methods in Section 2. We present our

978-1-5090-4809-0/17/$31.00 ©2017 IEEE

process for recognizing the emotions in Section 3. In Section 4, we describe the experiments to evaluate the proposed methods in terms of the recognition performance. We finally conclude our work and discuss our future direction in Section 5.

II. Acoustic Features Extraction

Processes of data preprocessing and feature extraction are described in this section. Feature extraction consists of two steps: preprocessing and acoustic feature extraction as described in Fig. 1.

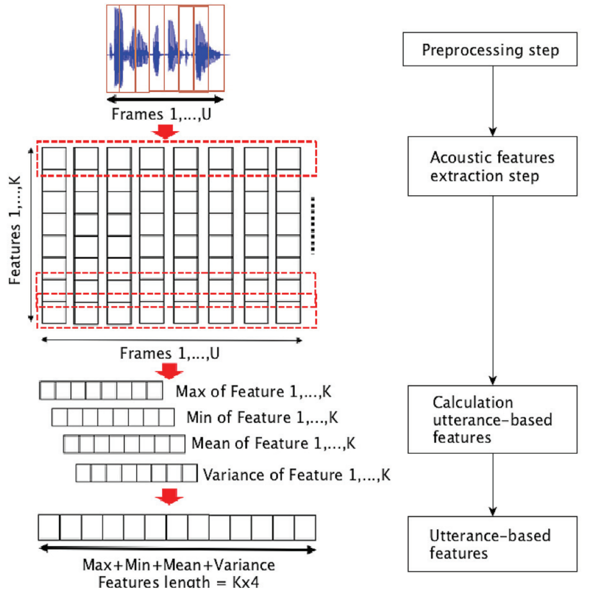

Fig. 1: Preprocessing and feature extraction

a) Preprocessing step: the input data in this research is utterances. We have to divide speech signal into frames. Then we compare each frame with phonemes label in the database and find the frame have a silence phoneme label and remove that frame. After that, we merge whole speech frames into utterances again. Silence is considered as an useless data in this research. After getting speech data, all of them are divided into frames. Each frame will be extracted features in below.

b) Acoustic feature extraction step: there are two kinds for feature extraction, which are frame-based feature and utterance-based feature. Features such as energy, pitch, Mel-frequency Cepstral coefficients (MFCC), Perceptual linear predictive (PLP), filter bank (FBank), and first and second derivatives of all features stated above are extracted as frame-based feature. While utterance-based features are calculated the statistical values like maximum, minimum, mean and variance of those frame-based features. In summary, there are four times from converting frame-based feature to utterance-based feature. All experiments in this paper are conducted only with utterance-based features on training and testing the classifiers.

All of utterance-based features are concatenated together, before calculating the first and second derivatives of them. For example, if we use 12 MFCCs as features, the features will be 36 (12+12+12). These 36 coeffiecients will be used to calculate utterance-based features (minimun, maximum, mean, and variance). The utterance-based features are computed statistics of the whole acoustic features in all frames of the utterance. Specifically, let $A_s(F_k)$ denote the acoustic features of the kth features for the frames s. We compute the utterance features for each utterance for all k = 1, …, K by following the 4 equation at below

$$f_1^k = \max_{s\in U} A_s(F_k)\ , \tag{2}$$

$$f_2^k = \min_{s\in U} A_s(F_k)\ , \tag{3}$$

$$f_3^k = \frac{1}{|U|}\sum\nolimits_{s\in U} A_s(F_k)\ , \tag{4}$$

$$f_4^k = \sqrt{\frac{1}{|U|}\sum\nolimits_{s\in U}\ A_s(F_k)^2}, \tag{5}$$

where U denotes the set of all frames. The utterance features f_1^k, f_2^k, f_3^k, and f_4^k corresponding to the maximum, minimum, mean and variance of the k^{th} acoustic features in whole frames, respectively. The size of utterance features vector is equal to the number of acoustic feature multiplied by four.

III. Emotion Recognition Framework

In this section, we describe the details of our framework that use two techniques of machine learning; Support Vector Machines (SVM) and Binary Support Vector Machines (BSVM). Both techniques are applied on this utterance-based emotion recognition.

A. Support Vector Machines (SVM)

Support Vector Machines (SVM) is a set of supervised learning methods used for classification and regression. SVM is simple and effective in high dimensional spaces and one of the most popular learning methods. Although the training set is small, SVM can provide high performance. We decided to use SVM as a multi-classes classifier for four emotions like “anger”, “happiness”, “neutral” and “sadness”. There are two main stages in SVM, which are training and evaluation.

1) Training stage

The whole process of training is shown in Fig. 2. It starts with feature extraction, follows with data preparation, and training model at last. Feature extraction is already explained in previous section. After feature extraction, we relabeled all data with integer number in order to train SVM. In our case, E1 the represent of anger label, is replaced by “0”. E2 (a label for happiness) is replaced by “1”. While E3 and E4 are neutral and sadness, they are relabeled as “2” and “3” respectively.

2) Evaluation stage

There are four models for four emotions from training stage. They are used to evaluate the test set as shown in Fig. 3. In evaluation stage, all test data are processed in preprocessing and feature extraction in the same manner as training stage. The input of SVM classifier is an utterance, and the output is only one label of emotion. Three kernels of SVM including linear, Radial Basis Function (RBF), and polynomial kernel are investigated in our experiments. For the parameter setting in

RBF and polynomial kernel, we define γ (gamma) parameter equal to $\frac{1}{n\_features}$ where *n_features* is represents a number of features. For the polynomial kernel, the degree parameter is determined to be three.

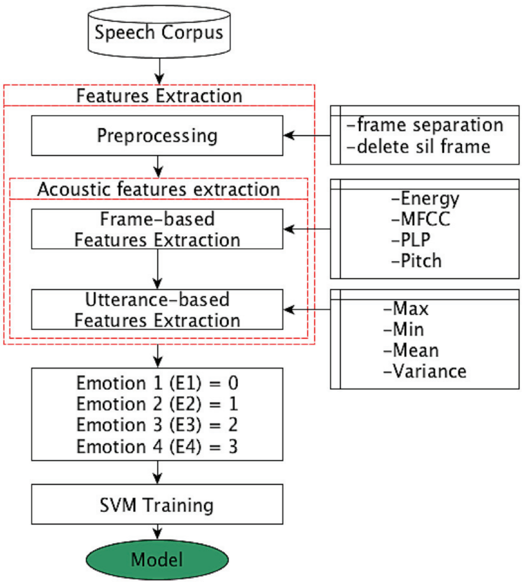

Fig. 2: Overview of training step of SVM

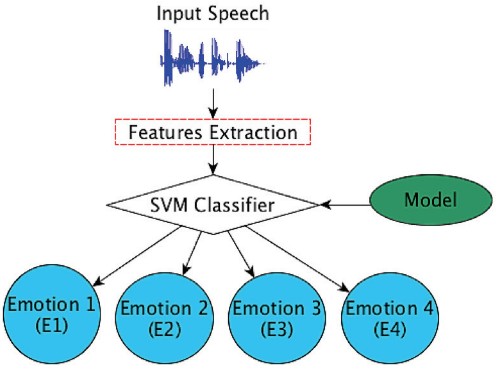

Fig. 3: Overview of testing step of SVM

B. *Binary Support Vector Machines (BSVM)*

Binary classification is one popular technique for classifying one of two classes. However, binary classification can support multiclass classification by combination more than one classifier. We applied that concept into our SVM classifier so called Binary Support Vector Machines (BSVM) in order to study and to compare the result with ordinary SVM. The details of BSVM are described into two stages as following.

1) Training stage

The overview of BSVM training process is shown in Fig. 4. The target emotion of recognition is to classify four emotions (E1, E2, E3, and E4), then three models are created in BSVM. In the first model (model 1), BSVM is designed to classify between the emotional state “anger” (E1) against the rest (E2, E3, and E4). In other words, all utterances in training set are trained BSVM in group 1 as depicted in Fig. 3. In group 2, only training data for “happiness” (E2), “neutral” (E3), and “sadness” (E4) are selected to train the second model (model 2) since this model is intended to divide happiness and non-happiness. Moreover, the last model is trained from utterances of E3 and E4. Totally our BSVM will have three models or three layers in binary classification.

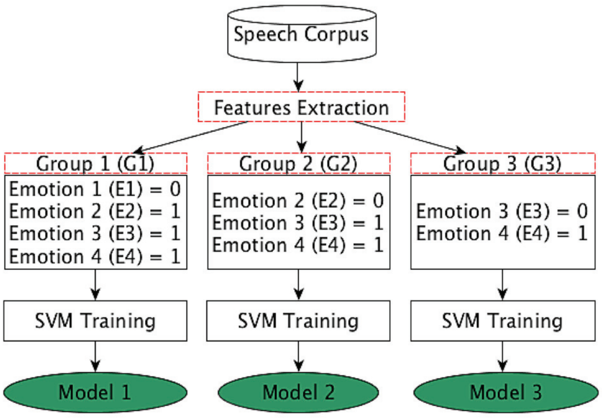

Fig. 4: Overview of training step of BSVM

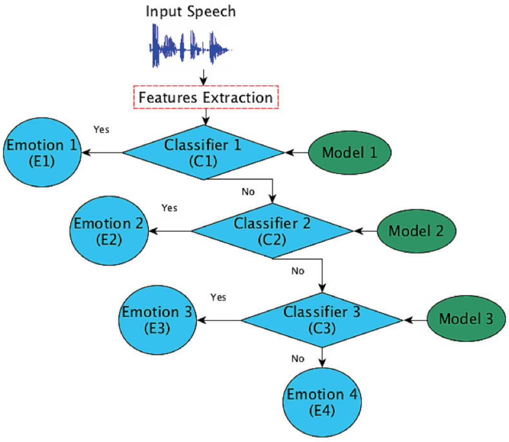

Fig. 5: Overview of testing step of BSVM

2) Evaluation stage

According to implementing BSVM in classification, there are three layers in this binary classification as explained in Fig. 5. All utterances in test set are extracted features as mentioned in previous section. All features are flowed down from first classifier (C1) until the last (C3). The results of each binary classifier will be collected as the answer of overall classification. In fact, each utterance will labeled in one of four emotions (E1 to E4). The accuracy of each BSVM will count on the final results of all three classifiers.

IV. Experiments

A. *Experimental setting*

We examined our approach on Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [20]. The database contains audiovisual data from 10 actors (5 males, and 5 females). There are 10 emotional categories such as neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited and other. Since three annotators were asked to label emotions in IEMOCAP, we applied majority rule to classify an emotion for an utterance. For example, annotator A and B labelled the utterance as “anger”, while annotator C gave “sadness”. That utterance will be considered as “anger”. However, we focused on four basic emotions (anger, happiness, sadness, and neutral) in this study. In fact, the number of happiness is less than other emotions significantly as shown in Table I. If we used only happiness for training and test set, the models will be trained from unbalanced data. Therefore, we have to increase the number of happiness by adding another emotion like excitement because we strongly believe that excitement and happiness mostly share more similarity characteristic in emotional definition than others [19].

TABLE I. Number of emotion utterances and durations per emotion category

Emotion	*Anger*	*Happiness*	*Excitement*	*Neutral*	*Sadness*	*Total*
Number	1103	595	1040	1708	1084	5530
		1635				
Duration (Hours)	1.38	0.71	1.38	1.85	1.65	6.89
		2.09				

A five-fold leave-one-out cross-validation (LOOCV) scheme was employed in all our experiments. Each number of utterance is divided into five groups. We use the first four groups from each emotion for training set and then use the last group for testing set. The ratio between training set and testing set is 80/20. For example, in anger emotion we have 1103 utterances and we divided the number of utterance into five groups. Each group consists of around 220 utterances. Then, we will have approximately 880 utterances for training and 220 utterances for testing. To perform fivefold LOOCV, we rotated all emotional data in five times. All results of each fold will be averaged at the end.

Kaldi Speech Recognition Toolkit [21] was used to extract acoustic features for each frame. The input signal is sampled at 16 kHz and converted into frames using a 25-ms window shifted by 10-ms. Acoustic features are in the form of feature vector composes of one energy, 12 MFCCs or 12 PLP or 40 FBANK and augmented with three pitches, their first and second derivatives. The maximum dimension of acoustic features is 204 values.

Skit-learn [22] machine learning in Python was used to train SVM and BSVM. The maximum, minimum, mean and variance of each frame in an utterance are used as the utterance-based features. The maximum dimension of utterance-based features is 816 values.

B. *Experimental result*

Three SVM kernels: Linear, RBF, and Polynomial are compared with three utterance-based features. Then we have 12 models to perform accuracies as shown in Table II.

TABLE II. Accuracy Of SVM Technique

Features	*Accuracy (%) of SVM*		
	linear	*RBF*	*poly*
MFCC	54.20%	31.00%	49.40%
PLP	55.00%	53.20%	49.40%
FBANK	56.40%	57.60%	**58.40%**
MFCC+PLP+FBANK	56.40%	31.00%	52.60%

Where MFCC features are represented to statistical features as utterance features $f_1^k, f_2^k, f_3^k and\ f_4^k$ respectively and compute utterance features by using energy, MFCC, pitch and their first and second derivatives features. We can see from the Table II, the best accuracy is 58.40% and most suitable kernel is polynomial kernel and most suitable features is utterance-based feature of FBANK. When we analyze the result of SVM with polynomial kernel and use FBANK as features for all emotions, the accuracies of all emotions are shown in Table III.

TABLE III. A Confusion Matrix Of SVM With Polynomial Kernel And FBANK Utterance Features

		Output (Accuracy %)			
	Emotion	*Anger*	*Happiness*	*Neutral*	*Sadness*
Input	***Anger***	**62.82%**	21.75%	12.51%	2.90%
	Happiness	15.65%	**50.88%**	24.58%	8.86%
	Neutral	6.67%	20.43%	**60.59%**	12.29%
	Sadness	3.41%	12.17%	22.50%	**61.90%**

From table III, the best accuracy is 62.82% in anger emotion follow by sadness, neutral and happiness emotion. Since, BSVM combination with SVM technique shown its significance for efficient emotional speech recognition [18], it is very interesting to apply with this work. In order to improve accuracy by using BSVM, the order of classifier is most importance as show by Fig 5. Thus, if we can design the order of BSVM beforehand, it might help us improving the performance of ESR in short time. In this study, we design the order of BSVM by analyzing the performance of each emotion as shown in Table III. We then design the order of BSVM by focused on conflict between each emotion. Thus, the first classifier (C1) must start with the highest accuracy, it is anger (A) vs other emotions. Because anger is lowest conflict emotion. The second classifier (C2) is a classifier between neutral (N) vs the rest emotions (sadness and happiness). The last classifier (C3) classifies between sadness (S) vs happiness (H). We believe this design might give the best accuracy in

BSVM. However, we create all possible twelve cases of BSVM to compare the accuracies as shown in Table IV.

TABLE IV. ACCURACY TWELVE CASE OF BSVM WITH POLYNOMIAL KERNEL AND FBANK UTTERANCE FEATURES

No.	BSVM Classifier			*Accuracy (%)*
	C1	**C2**	**C3**	***SVM-POLY***
1	A vs H, N, S	H vs N, S	N vs S	57.77%
2	A vs N, S, H	N vs S, H	S vs H	**58.24%**
3	A vs S, N, H	S vs N, H	N vs H	57.85%
4	H vs A, S, N	A vs S, N	S vs N	57.41%
5	H vs N, S, A	N vs S, A	S vs A	55.93%
6	H vs S, N, A	S vs N, A	N vs A	57.07%
7	N vs A, S, H	A vs S, H	S vs H	57.75%
8	N vs H, S, A	H vs S, A	S vs A	56.63%
9	N vs S, H, A	S vs H, A	H vs A	56.91%
10	S vs A, N, H	A vs N, H	N vs H	58.10%
11	S vs H, N, A	H vs N, A	N vs A	57.17%
12	S vs N, A, H	N vs A, H	A vs H	57.43%

From Table IV, the best accuracy is 58.24% and occurred at case number 2. This confirms that our assumption and explanation at beforehand is right. We analyze the details of results in case number 2 and show accuracies of each emotion in Table V. The best accuracy is 66.32% for sadness and follow by anger, neutral and happiness. However, the accuracy of BSVM in each emotion of case number 2 can show some improvement from SVM such as happiness and sadness emotion. Especially, sadness can be improved by 4.42% (from 61.90% to 66.32%) and happiness can be improved by 0.74% (from 50.88% to 51.62%).

TABLE V. A CONFUSION MATRIX OF BSVM WITH POLYNOMIAL KERNEL AND FBANK UTTERANCE FEATURES OF CASE NO 2.

		Output (Accuracy %)			
	Emotion	***Anger***	***Happiness***	***Neutral***	***Sadness***
Input	***Anger***	**59.65%**	21.75%	13.32%	5.25%
	Happiness	11.74%	**51.62%**	24.64%	11.98%
	Neutral	4.80%	20.55%	**58.54%**	16.10%
	Sadness	2.02%	12.99%	19.64%	**66.32%**

V. CONCLUSIONS

We studied the utilization of SVM and BSVM with utterance features by computing from five popular acoustic features in speech recognition like energy, pitch, MFCC, PLP and FBANK for predict emotional states. In experimental results, the best accuracy of SVM is 58.40% in average. The best suitable kernel is polynomial kernel and the best features is FBANK. Then we analyze the performance of SVM for design the order of BSVM and the best accuracy of BSVM is 58.24% in average. In this paper, BSVM technique shows improvement in happiness and sadness, especially for sadness. In future work, we will seek the way to apply other features and apply data selection for choosing utterance features improving the accuracy of emotional speech recognition.

VI. REFFERENCE

[1] Xu Huahu, Gao Jue and Yuan Jian, "Application of speech emotion recognition in intelligent household robot" in roceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010, vol. 1, 2010, pp. 537-541.

[2] Laurence Vidrascu and Laurence Devillers, "Detection of real-life emotions in call centers" in Interspeech 2005, 2005, pp. 1841-1844.

[3] Akputu K. Oryina and Abiodun O. Adedolapo, "Emotion Recognition for User Centred E-Learning" in Proceedings - International Computer Software and Applications Conference, vol. 2, 2016, pp. 509-514.

[4] Sagar K. Bhakre and Prof.Arti Bang, "Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier" in Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 2016, pp. 2363-2367.

[5] D. N. N. Shambhavi S. S and Dr.V. N Nitnaware,, "Emotion Speech Recognition using MFCC and SVM," in *International Journal of Engineering Research & Technology (IJERT)*, June-2015.

[6] Weishan Zhang, Xin Meng, Zhongwei Li, Qinghua Lu,, "Emotion Recognition in Speech using Multi-Classification SVM Weishan", in 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), 2015, pp. 1181-1186.

[7] T.-l. C. Y.-t. Y. J.-h. L. P.-j. Pao, "Mandarin Emotional Speech Recognition Based on SVM and NN" in 18th International Conference on Pattern Recognition (ICPR'06), 2006, pp. 1096-1100.

[8] Qin Jin, Chengxin Li, Shizhe Chen1 and Huimin Wu, "SPEECH EMOTION RECOGNITION WITH ACOUSTIC AND LEXICAL FEATURES", in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2015, pp. 4749-4753.

[9] M. K. M. S. K. F. El Ayadi, "Survey on speech emotion recognition: Features, classification schemes, and databases" in Pattern Recognition, vol. 44, 2011, pp. 572-587.

[10] S. Lugović, I. Dunđer and M. Horvat, "Techniques and Applications of Emotion Recognition in Speech",, 2016, pp. 1278-1283.

[11] D Satya Ganesh, Dr. Prasant Kumar Sahu,, “A Study on Automatic Speech Recognition Toolkits”, in International Conference on Microwave, Optical and Communcation Engineering,, India, December 18-20, 2015, p. IIT Bhubaneswar.

[12] S. A. M. R. R. R. P. Rieger, "Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers" in roceedings of the 9th International Symposium on Chinese Spoken Language Processing, ISCSLP 2014, 2014, pp. 589-593.

[13] H. Hu, M.-X. Xu, and W. Wu, , GMM supervector based SVM with spectral features for speech emotion recognition, vol. 4, in Proceedings of IEEE ICASSP 2007, IEEE, 2007, p. IV–413.

[14] H. L. M. C. L. Fayek, "Towards Real-time Speech Emotion Recognition using Deep Neural Networks" in 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), 2015, pp. 1-5.

[15] W. Q. Y. J. S. Z. Y. X. Zheng, "An experimental study of speech emotion recognition based on deep convolutional neural network" in 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 827-831.

[16] D. Neiberg, K. Elenius, and K. Laskowski,, "Emotion recognition in spontaneous speech using GMMs." in Proceedings of Inter- speech, 2006.

[17] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in Proceedings of IEEE ICASSP 2003, vol. 2, IEEE, 2003, p. II–1.

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

[19] Chen, Shizhe Jin, Qin Li, Xirong Yang, Gang Xu, Ji, "Speech Emotion Classification using Acoustic Features", 2014, pp. 579-583.

[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. M, IEMOCAP: Interactive emotional dyadic motion capture database, vol. 42, Language resources and evaluation, 2008, p. 335–359.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. , "The kaldi speech recognition toolkit," in *in IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society,* , Dec. 2011..

[22] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, 2011, pp. 2825-2830.