Extraction of Novel Features Based on Histograms of MFCCs Used in Emotion Classification from Generated Original Speech Dataset

Muhammet Pakyurek[1, *], Mahir Atmis[2], Selman Kulac[1], Umut Uludag[3]
[1]Department of Electrical - Electronics Engineering, Faculty of Engineering, Duzce University, Duzce, Turkey
[2]Department of Computer Engineering, Faculty of Engineering, Ozyegin University, Cekmekoy/Istanbul, Turkey
[3]TUBITAK BILGEM,
Baris Mah. Dr. Zeki Acar Cad. No:1, 41470 Gebze/Kocaeli, Turkey
mpak85@hotmail.com

***Abstract*—This paper introduces two significant contributions: one is a new feature based on histograms of MFCC (Mel-Frequency Cepstral Coefficients) extracted from the audio files that can be used in emotion classification from speech signals, and the other – our new multi-lingual and multi-personal speech database, which has three emotions. In this study, Berlin Database (BD) (in German) and our custom PAU database (in English) created from YouTube videos and popular TV shows are employed to train and evaluate the test results. Experimental results show that our proposed features lead to better classification of results than the current state-of-the-art approaches with Support Vector Machine (SVM) from the literature. Thanks to our novel feature, this study can outperform a number of MFCC features and SVM classifier based studies, including recent researches. Due to the lack of our novel feature based approaches, one of the most common MFCC and SVM framework is implemented and one of the most common database Berlin DB is used to compare our novel approach with these kind of approaches.**

***Index Terms*—Emotion classification; MFCC; SVM; Speech signal.**

I. INTRODUCTION

Human-computer interaction systems have been drawing attention increasingly in recent years. Understanding the emotions of humans plays a significant role in these systems, since human feelings provide a better understanding of human behaviours. Furthermore, in order to increase the accuracy of recognition of the words spoken by human, many of the state-of-the-art automatic speech recognition systems are dedicated to natural language understanding. Emotion classification has a key role in performance improvements for natural language understanding. The other areas, in which an emotion classification system can be used are as follows: voice search tagging, word search with specific emotions, and emotion based advertisement placement [1].

Manuscript received 30 April, 2019; accepted 12 October, 2019.

In this study, MFCCs are calculated for all audio files in both of the utilized databases. Then, these are classified based on the type of emotions. In [2], Plutchik claims that emotions are categorized as the Primary Emotions and Secondary Emotions. Primary emotions are anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. In this study, emotions of sadness, happiness, and neutral can be recognized by our designed system. We focused only on these three emotions as the amount of the train data is generally not large enough for the remaining ones to arrive at statistically robust conclusions. There are two main contributions in this study. One is our novel feature, which is MFCCs representation based on their histograms and other contribution is PAU speech data, whose emotions are labelled and cross-checked by PhD students.

Section II covers academic studies related to this paper. In Section III, experimental framework and its steps are elaborated. Section IV mentions our novel feature and classical MFCCs feature of academic literature in detail. Section V describes speech data and their characteristics. Finally, Section VI exhibits the experimental results and Section VII draws conclusions.

II. LITERATURE SURVEY

Various types of classifiers have been used for the task of speech emotion classification: Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Artificial Neural Networks (ANN), k-Nearest Neighbors (k-NN), and many others. In fact, there has been no agreement, which classifier is the most suitable one for emotion classification. It also seems that each classifier has its own advantages and limitations.

Many recent studies show that DNN based approaches outperforms SVM in many areas, such as image, speech, and text studies within abundant data [3]. In recent papers [4]–[6], these two R&D groups independently have established closely related DNN architectures with multi-

task learning capabilities for multilingual speech recognition. On the other hand, although the conventional deep learning-based method can outperform the SVM classifier, it requires plenty of training samples to construct models of DNN [7], [8]. Therefore, we cannot implement DNN due to the limited data.

In study [9], the authors have leveraged MFCC for extraction of features and multiple Support Vector Machine (SVM) as a number of classifiers. Their extensive experiments are based on happiness, anger, sadness, disgust, surprise, and neutral emotion sound database. Performance analysis of multiple SVM reveales that non-linear kernel SVM achieves greater accuracy than linear SVM [10]. As the authors mention, their best performance on Berlin DB is 75 % accuracy.

Dahake *et al.* [11] has two main contributions: one is feature extraction using pitch, formants, and MFCC, and the other is to improve speaker dependent SER by comparing the results with different kernels of SVM classifier [12]. The highest accuracy is obtained with the feature combination of MFCC +Pitch+ Energy on both Malayalam emotional database (95.83 %) and Berlin emotional database (75 %), tested using SVM with linear kernel.

In [13], three emotional states are recognized: happiness, sadness, and neutral. Explored features include: energy, pitch, Linear Prediction Cepstral Coefficients (LPCC), MFCC, and Mel-Energy spectrum Dynamic Coefficients (MEDC). Berlin Database and self- built Chinese emotional databases are used for training the specified classifiers.

In [14], the basic emotion comparing speech features are being recognised. The authors use similar methodology with the study in this paper to recognize emotions. However, their database and features for recognition are quite different from ours.

In order to combine the merits of several classifiers, aggregating a group of them has also been recently employed [15], [16]. Based on several studies [17]–[22], we can conclude that SVM is one of the most popular classifiers in emotion classification probably because it had been widely used in almost all speech applications up to 2012. As shown in Table I [23], the average success rate of SVM for speech emotion classification is in the range of 75.45–81.29 %.

In [24], Kamruzzaman and Karim report on speaker identification for authentication and verification in security areas. This kind of identification is mainly divided into text-dependent and text-independent approaches. Even if many studies utilize the text-dependent approach based on a variety of predefined certain utterances, this study employs a text-independent methodology. Basically, the implementation part of this study is composed of feature generation and classification. MFCC coefficients are calculated as a foundation of our informative features and SVM utilizes these features in order to classify the speech data.

In [25], Demircan and Kahramanli extract MFCC's from the speech data obtained from Berlin Database [26] (Berlin Database of Emotional Speech, 2014). Seven statistical values are calculated from the MFCC: minimum value, maximum value, means, variance, median, skewness, and kurtosis. Using those values, k-Nearest Neighbor algorithm is used to classify the data. Their contribution is to reduce the dimension of the data to 7 different values.

TABLE I. CLASSIFICATION PERFORMANCE OF POPULAR CLASSIFIERS FOR THE SPEECH EMOTION CLASSIFICATION [23].

Classifier	**HMM**	**GMM**	**ANN**	**SVM**	**Classifier**
Average classification accuracy (%)	75.5–78.5	74.83–81.94	51.19–52.82	75.45–81.29	75.5–78.5
Average training time	Small	Smallest	Back-propagation: large	Large	Small
Sensitivity to model initialization	Sensitive	Sensitive	Sensitive	In-sensitive	Sensitive

III. EXPERIMENTAL FRAMEWORK

In order to carry out various experiments to show the performance of our novel emotion classification feature, we elaborate a framework with details. The steps of this emotion classification framework (Fig. 1) are as follows sequentially.

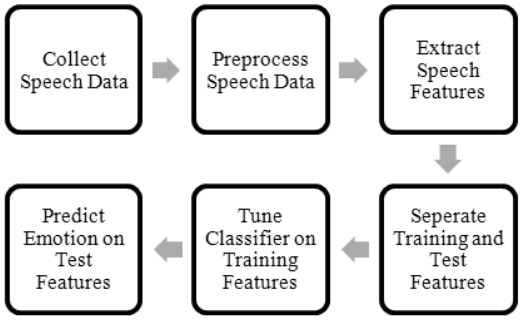

Fig. 1. Process flow of our emotion classification Framework.

A. Collect Speech Data

Collecting speech data plays a significant role in speech recognition studies due to the lack of comprehensive speech data. Therefore, speech data collection constitutes a major part of this study. The details of data properties and how to generate them are explained in Section V-C.

B. Preprocessing

Due to the fact that noise in speech breaks down speech data, removing outliers plays a significant role in the state-of-the-art classification system. In order to filter them out, Interquartile range method of John Tukey [27] is employed. Furthermore, min-max normalization is employed in feature wise for the sake of removing out the high variance sensitivity on features.

C. Feature Extraction

The extraction of suitable features that efficiently represent different emotions is one of the most important issues in the design of a speech emotion classification system. A proper group of features significantly affects the classification results, since pattern recognition techniques are rarely independent of the problem domain. In this study, MFCCs are selected as a group of features. More specifically, in the first feature, the first and second derivation of average MFCCs and the average of them are calculated. As the second feature, which is our novelty, weighted values of MFCCs combining MFCCs values and

their corresponding Probability Density Function's (PDF) values. In the third feature, concatenation of the first and second features is leveraged to get higher performance.

1. Mel-Frequency Cepstrum Coefficients (MFCC)

MFCCs are calculated based on the known variation of the human ear's critical bandwidths with frequency. The main point to understand speech is that the sounds generated by a human are filtered by the shape of the vocal tract, including tongue, teeth, etc. This shape determines what sound comes out. If the shape is accurately determined, this should result in an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the purpose of MFCCs is to represent this envelope accurately [1].

In order to get a statistically stationary mean of data, the audio signal is divided into 25 ms of frames. If the frame is too short, it may not be possible to have enough samples to get a reliable spectral estimate. If it is too long, the signal changes too much throughout the frame. Each frame can be converted into 12 MFCCs plus a normalized energy parameter. The first and second derivatives (Delta and Delta-Delta, respectively) of MFCCs and energy can be calculated as extra features resulting in 39 numbers representing each frame. However, the derivation of the MFCC parameters is generally implemented when the original MFCC does not provide the necessary amount of information that leads to a good classification.

The MFCC algorithm steps are shown in Figure 2.

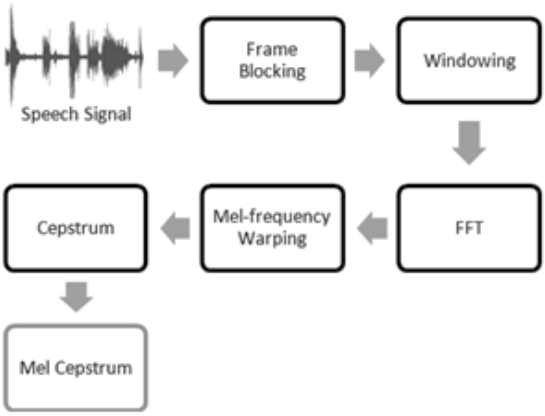

Fig. 2. Block diagram of the MFCC Algorithm [1].

D. Classification

A speech emotion classification system consists of two stages: (1) feature extraction from the available (speech) data and (2) classification of the emotion in the speech utterance. In fact, the most recent researches in speech emotion classification have focused on this step. A number of advanced machine learning algorithms have been developed for many different research areas. On the other hand, traditional classifiers have been used in almost all proposed speech emotion classification systems [23]. In this study, SVM is used to classify speech utterances by optimizing and training data set and presenting performance results on the test sets.

SVM is a supervised machine learning classifier technique used primarily for large databases to categorize new samples. The algorithm searches for the optimal hyperplane, which separates different classes with maximum margin between them. The libSVM [17], a scholarly accepted support vector library, is used to train and test the dataset. The data is separated into two parts – 90 % for training and 10 % for testing. On the training part, the validation sets for each fold are generated using 10-fold cross validation methodology. A Gaussian radial base function kernel is used to classify data, since it gives better approximations on data. The best SVM parameters C and gamma (γ) are obtained using 10-fold cross-validations on train dataset with validation data. Those parameters are determined using a mesh-grid search over the values suggested by [28].

E. Software Toolbox

LibSVM [28] library for Matlab is used for SVM routines. Matlab's TreeBagger class is utilized for RF classification. MFCC library of Wojcicki for Matlab [29] is used to calculate MFCC.

F. Algorithm

In the main part, firstly all datasets are acquired to calculate the MFCCs for each individual file. Then, the first, second, and third features of each file are calculated using MFCCs values for each element. More elaborately, each file is divided into the number of 25 ms. of frames. Then, MFCCs are calculated for each frame. After calculating the MFCCs, average value and their corresponding the first and second derivatives are counted. Then, a histogram of each MFCCs is created dividing to 10 equal distant bin for each MFCCs in min-max range. These counts of histograms are divided by total count to get the PDF of MFCCs. Then, in order to leverage PDF value and corresponding MFCC value, these two values are multiplied for each of the MFCCs PDF. Finally, all MFCCs values, their average, and first and second derivatives of each MFCCs are stored for each frame. At the end of the file, the histogram and PDF are calculated using each frame of MFCCs. The Covariance matrix and a label vector for the output emotion classes are generated by SVM. After the SVM analysis, Accuracy and Confusion Matrices are calculated as a mean value for all iterations.

In SVM analysis part, train and test data are randomly selected. Then, 10-folds cross validation is performed on the train data. Accuracy results of SVM prediction are obtained by using the best parameters resulting from the cross validation.

IV. MFCCs Based Feature Vectors

In this study, 12 coefficients of MFCC + the energy of each frame are calculated for each individual's audio file [29]. The details of the MFCC are explained in Section III-C1. For the feature extraction, three features are generated using MFCC. These features are as follows.

1. Feature Set 1

Average of MFCCs, its *Delta* (first order derivative) and *Delta_Delta* (second order derivative): The average of MFCCs is calculated for all frames of each speech data. Delta and Delta-Delta (first and second derivatives) are calculated by subtracting the consecutive frames and consecutive Deltas correspondingly.

2. Feature Set 2

Weighted MFCCs values wrt (with respect to) their probability distribution: The PDF of each coefficient are calculated building the histogram of each of the MFCCs of all frames. During this calculation, different value interval for each MFCC is obtained considering min-max values of them. The second feature is calculated by the multiplication of values in this interval and corresponding PDF values:

$$c_i = [a_i, b_i],\ i = 1, 2 \ldots 13, \quad (1)$$

$$v_i = c_i * PDF(c_i), \quad (2)$$

where a_i and b_i are min and max values of each of MFCCs. In that case, c_i is the internal value within $[a_i, b_i]$. As shown in Fig. 3, c_i discrete feature value and pdf(c_i) are non-normalized probability values. In (2), "*" operation is the element-wise multiplication. In this case, we encode the histogram just multiplying these two values. So, only a number of bins data is used to represent the histogram. Otherwise, bin values and corresponding probability values must be used separately to describe the histogram. Thanks to this approach, the number of features is decreased, while the computational performance is increased because of halving the size of histogram representation.

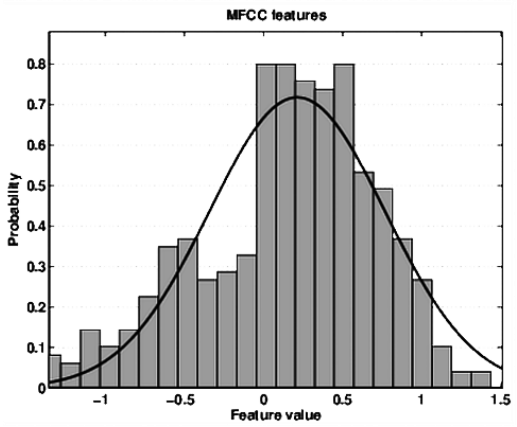

Fig. 3. PDF of one of MFCCs without normalization [30].

3. Feature Set 3

Concatenation of Feature Set 1 and Feature Set 2: In this feature set, Feature Set 1 and Feature Set 2 are assembled without any modification on both features.

V. MULTIPLE DATABASES

The details of databases utilized in this study are as follows:

1. The Berlin Database: This is a database frequently used by emotion classification researchers, which contains speech data in German language [23], [31]. Burkhardt *et al.* [26] show the details about the Berlin Database.
2. The PAU Database: We have collected English speech samples from YouTube video collections and videos of popular TV shows.

Figure 4 and Figure 5 illustrate histograms of the length of the audio files for the Berlin Database and our custom database, respectively. Bins of the histograms represent audio file length in seconds. Total number of files is 312 for the Berlin Database, and 320 for the PAU database. Total time for the Berlin database is 16 minutes, and for PAU - 10 minutes.

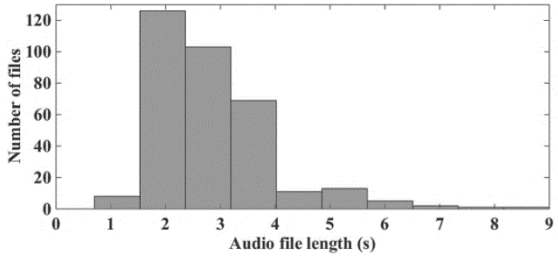

Fig. 4. Berlin Database file length histogram.

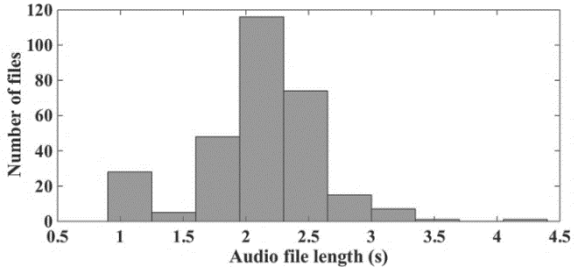

Fig. 5. PAU Database file length histogram.

A. Database Features

In this study, genders (male & female) of the associated individuals are noted as database metadata. Also, age categories are classified as "Young" (age between 12 and 30) and "Mature" (age between 31 and 60). Sadness, happiness, and neutrality are chosen as target emotions to predict. Audio files are in wav format and their duration varies from 1 to 9 seconds. Acted and neutral speech types are also available.

B. Labelling

Labelling the audio file plays a significant role in categorization of the data. In this study, all speech data are labelled with gender, emotion, and age data. Table II compares both databases according to their features.

C. PAU Database Generation

The PAU database is produced from the sources described in Table III by 4 (male) students, who are doing their PhDs in computer and electrical engineering departments. All speech data are inserted into the PAU database after the independent control steps. In this control step, each member checks other members' data sets also, which must be consistent with their corresponding label. It took nearly three months to collect and process the data, which is approximately 102 MB in size (the database files will be provided free of charge to the academic and research community).

TABLE II. COMPARISON BETWEEN THE BERLIN AND PAU EMOTION CLASSIFICATION SPEECH DATABASES.

Database	Emotion	Gender	Age Group	Language	Speech Type
Berlin	Sad, Happy, Neutral	5 Male, 5 Female	Young, Mature	German	Acted
PAU	Sad, Happy, Neutral	195 Male, 72 Female	Young, Mature	English	Acted, Natural

TABLE III. PAU EMOTION CLASSIFICATION SPEECH DATABASE DETAILS.

Source	Emotion	Gender	Age Group	Language	Speech Type
How I Met Your Mother	Sad, Happy, Neutral	16 Male 1 Female	Young, Mature	English	Acted
Sherlock Holmes	Sad, Happy, Neutral	2 Male	Young	English	Acted
Thrones Youtube	Sad	16 Male 8 Female	Young, Mature	English	Acted
YouTube Best Cry Videos	Sad	63 Male	Young, Mature	English	Natural
Shameless	Happy	1 Male 3 Female	Young	English	Acted
The Man From Uncle	Neutral	22 Male 7 Female	Young, Mature	English	Acted
Youtube News Compilation	Neutral	50 Male 9 Female	Young, Mature	English	Acted
Youtube Videos Compilation	Happy	25 Male 44 Female	Young, Mature	English	Natural

VI. EXPERIMENTAL RESULTS

The database consists of 632 audio samples in total. Experiments are conducted for the German Berlin database, PAU English database, and a combination of both. For each case, train and test data are selected from their own datasets.

The number of audio files per database is shown in Table IV.

TABLE IV. DISTRIBUTION OF EMOTIONS OF DATABASES.

	Emotions (number of audio files)			
Database	**Happiness**	**Neutral**	**Sadness**	**Total**
Berlin DB	105	103	104	312
PAU DB	107	105	108	320
Total	212	208	212	632

The accuracy results of SVM, shown in Table V, Table VI, and Table VII, are the average accuracy results of 60 runs. More specifically, all experiments are repeated 60 times. The peak (non-average) accuracy result obtained during the tests was 95 %. One of the models used in the paper [13] by Yixiong *et al*. consists of MFCC + MEDC + Energy triple.

That model has the highest accuracy rate (91.3043 %) among all their models on the Berlin Database, but it is not clear, whether that is a peak accuracy or a mean accuracy.

In [26], Burkhardt *et al.* did not mention how to separate train and test data. Their best neutral, happiness, and sadness recognition rates are 88.2 %, 83.7 %, and 80.7 %, respectively, while ours are 84.8 %, 85.29 %, 88.5 % for the third feature in the Berlin Database (in German). The results reveal that our features results in better performance for identifying emotions of happiness and sadness.

TABLE V. EXPERIMENTAL RESULTS FOR BERLIN DATABASE.

	First Feature			Second Feature			Third Feature		
Accuracy	**83.78 %**			**86.00 %**			**88.33 %**		
Confusion Matrix	0.28	0.04	0.01	0.28	0.05	0.00	0.29	0.05	0.00
	0.03	0.25	0.03	0.03	0.28	0.02	0.03	0.28	0.02
	0.01	0.04	0.31	0.00	0.04	0.31	0.00	0.04	0.31

TABLE VI. EXPERIMENTAL RESULTS FOR PAU DATABASE.

	First Feature			Second Feature			Third Feature		
Accuracy	**76.35 %**			**78.27 %**			**79.81 %**		
Confusion Matrix	0.28	0.02	0.04	0.31	0.01	0.02	0.25	0.03	0.06
	0.05	0.21	0.03	0.06	0.18	0.02	0.03	0.26	0.03
	0.05	0.04	0.27	0.06	0.05	0.29	0.04	0.02	0.29

TABLE VII. EXPERIMENTAL RESULTS FOR COMBINED DATABASE.

	First Feature			Second Feature			Third Feature		
Accuracy	**76.27 %**			**81.86 %**			**83.81 %**		
Confusion Matrix	0.26	0.05	0.03	0.27	0.04	0.02	0.27	0.04	0.02
	0.05	0.21	0.03	0.03	0.25	0.02	0.03	0.25	0.02
	0.04	0.03	0.30	0.03	0.03	0.30	0.02	0.03	0.30

VII. CONCLUSIONS

Even though DNN has better results (performs better) than SVM, in this study, SVM is carried out as a classifier because of the lack of huge size speech data.

Better results were obtained, because of distributions of all MFCCs have more information to represent the emotion rather than using only average of MFCCs. This novel feature provides smaller size of data for histogram representation and requires less computational power. We can clearly conclude that using this feature has two main advantages: feature representation size and computational cost.

Best results are achieved by the Berlin Database compared to PAU (English) database because the sentences for the speech in Berlin Data are the same for each individual and they are performed in the same framework as well (in studio environment). Procedural preferences during the speech, such as stressing words, mood, and mouth gesture, are almost the same.

As shown in Table V and Table VI, we have an approximately 8.5 % decrease of accuracy for the English

database (Table VI) compared to Berlin Database (Table V) because the sentences in every sample are quite different from one another for the former database. Furthermore, some additional noise resulting from the environment of speech has a great impact on audio files. All Berlin speech data are generated in indoor studio environment, while our database has different environment speech utterance. Therefore, the procedures of data generation are quite different from our methodology. As a conclusion, we should note that our framework for audio generation is more appropriate for the real-life conditions. Our study has better results than average classification accuracy of SVM for the speech emotion classification studies. The accuracy results obtained by SVM on PAU database for the first, second, and third feature are 70 %, 71 %, and 73 %, respectively. Those numbers are 75 %, 78 %, and 81 % for Berlin Database. The results obtained are the average accuracy results of 60 runs. Those results support that the third feature helps us to obtain a better classification result.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] R. Jang, *Audio Signal Processing and Recognition*, 2011. [Online]. Available: http://mirlab.org/jang/books/audioSignalProcessing/

[2] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots", *American Scientist*, vol. 89, no. 4, pp. 344–350, Jul.–Aug. 2001. DOI:10.1511/2001.4.344.

[3] L. Deng, D. Yu *et al*., "Deep learning: methods and applications", *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014. DOI: 10.1561/2000000039.

[4] L. Deng, J. Li, J. T. Huang, K. Yao, D. Yu, F. Seide, M. L. Seltzer, G. Zweig, X. He, J. D. Williams *et al*., "Recent advances in deep learning for speech research at Microsoft", in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 8604–8608. DOI: 10.1109/ICASSP.2013.6639345.

[5] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers", in *Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013, pp. 7304–7308. DOI: 10.1109/ICASSP.2013.6639081.

[6] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks", in *Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013, pp. 8619–8623. DOI: 10.1109/ICASSP.2013.6639348.

[7] P. Liu, K.-K. R. Choo, L. Wang, and F. Huang, "SVM or deep learning? A comparative study on remote sensing image classification", *Soft Computing*, vol. 21, no. 23, pp. 7053–7065, 2017. DOI: 10.1007/s00500-016-2247-2.

[8] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification", *Journal of Sensors*, vol. 2015, 2015. DOI: 10.1155/2015/258619.

[9] A. Sonawane, M. Inamdar, and K. B. Bhangale, "Sound based human emotion recognition using MFCC & multiple SVM", in *Proc. of 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, IEEE, 2017, pp. 1–4. DOI: 10.1109/ICOMICON.2017.8279046.

[10] K. Aida-zade, A. Xocayev, and S. Rustamov, "Speech recognition using Support Vector Machines", in *Proc. of 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, 2016, pp. 1–4. DOI: 10.1109/ICAICT.2016.7991664.

[11] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and Support Vector Machine", in *Proc. of International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, IEEE, 2016, pp. 1080–1084. DOI: 10.1109/ICACDOT.2016.7877753.

[12] M. Sinith, E. Aswathi, T. Deepa, C. Shameema, and S. Rajan, "Emotion recognition from audio signals using Support Vector Machine", in *Proc. of 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, IEEE, 2015, pp. 139–144. DOI: 10.1109/RAICS.2015.7488403.

[13] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using Support Vector Machine", *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–107, 2012. DOI: 10.5120/431-636.

[14] S. S. Shambhavi, "Emotion speech recognition using MFCC and SVM", *International Journal of Engineering Research and Technology*, vol. 4, no. 6, pp. 1067–1070, 2015. DOI: 10.17577/IJERTV4IS060932.

[15] B. Schuller, M. Lang, and G. Rigoll, "Robust acoustic speech emotion recognition by ensembles of classifiers", in *Proc. of Jahrestagung für Akustik*, DAGA, 2005, vol. 31.

[16] M. Lugger, M. Janoir, and B. Yang, "Combining classifiers with diverse feature sets for robust speaker independent emotion recognition, in *Proc. of 2009 17th European Signal Processing Conference*, 2009. pp. 1225–1229.

[17] A. Shirani and A. R. N. Nilchi, "Speech emotion recognition based on SVM as both feature selector and classifier", *International Journal of Image, Graphics & Signal Processing*, vol. 8, no. 4, pp. 39–45, 2016. DOI: 10.5815/ijigsp.2016.04.05.

[18] J. Zhou, Y. Yang, P. Chen, and G. Wang, "Speech emotion recognition based on rough set and SVM", in *Proc. of 5th IEEE Int. Conf. on Cognitive Informatics (ICCI'06)*, 2006, pp. 53–61. DOI: 10.1109/COGINF.2006.365676.

[19] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture", in *Proc. of. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 577–580. DOI: 10.1109/ICASSP.2004.1326051.

[20] O. Pierre-Yves, "The production and recognition of emotions in speech: Features and algorithms", *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 157–183, 2003. DOI: 10.1016/S1071-5819(02)00141-6.

[21] C. M. Lee, S Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes", in *Proc. of. INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, 2004, pp. 889–892.

[22] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals", in *Proc. of 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003*, Geneva, Switzerland, 2003, pp. 125–128. DOI: 10.1.1.491.9261.

[23] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. DOI: 10.1016/j.patcog.2010.09.020.

[24] S. M. Kamruzzaman, A. N. M. Rezaul Karim, M. Saiful Islam, and M. Emdadul Haque, "Speaker identification using MFCC-domain Support Vector Machine", *International Journal of Electrical and Power Engineering*, vol. 1, no. 3, pp. 274–278, 2007. DOI: 10.3923/ijepe.2007.274.278.

[25] S. Demircan and H. Kahramanlı, "Feature extraction from speech data for emotion recognition", *Journal of Advances in Computer Networks*, vol. 2, pp. 2, 28–30, 2014. DOI: 10.7763/JACN.2014.V2.76.

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech", in *Proc. of 9th European Conference on Speech Communication and Technology*, 2005. DOI: 10.1.1.130.8506.

[27] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots", *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978. DOI; 10.1080/00031305.1978.10479236.

[28] Ch.-Ch. Chang, Ch.-J. Lin, "LIBSVM: A library for support vector machines", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27:1–27:27, 2011. DOI: 10.1145/1961189.1961199.

[29] K. Wojcicki, HTK MFCC MATLAB, MATLAB, File Exchange, 2011.

[30] A. N. Iyer, U. O. Ofoegbu, R. E. Yantorno, and S. J. Wenndt, "Speaker recognition in adverse conditions", in *Proc. of 2007 IEEE Aerospace Conference*, 2007, pp. 1–8. DOI: 10.1109/AERO.2007.352976.

[31] Y. Chavhan, M. L. Dhore, and P. Yesaware, "Speech emotion recognition using Support Vector Machine", *International Journal of Computer Applications*, vol. 1, pp. 8–11, 2010. DOI: 10.5120/431-636.