Speech Emotion Recognition Using Derived Features from Speech Segment and Kernel Principal Component Analysis

Matee Charoendee, Atiwong Suchato, and Proadpran Punyabukkana
Department of Computer Engineering
Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
matee.c@student.chula.ac.th, atiwong.s@chula.ac.th, proadpran.p@chula.ac.th

Abstract—**Speech emotion recognition is a challenging problem, with identifying efficient features being of particular concern. This paper has two components. First, it presents an empirical study that evaluated four feature reduction methods, chi-square, gain ratio, RELIEF-F, and kernel principal component analysis (KPCA), on utterance level using a support vector machine (SVM) as a classifier. KPCA had the highest F-score when its F-score was compared with the average F-score of the other methods. Using KPCA is more effective than classifying without using feature reduction methods up to 5.73%. The paper also presents an application of statistical functions to raw features from the segment level to derive global features. The features were then reduced using KPCA and classified with SVM. Subsequently, we conducted a majority vote to determine the emotion for the entire utterance. The results demonstrate that this approach outperformed the baseline approaches, which used features from the utterance level, the utterance level with KPCA, the segment level, the segment level with KPCA, and the segment level with the application of statistical functions without KPCA. This yielded a higher F-score at 13.16%, 7.03%, 5.13%, 4.92% and 11.04%, respectively.**

Keywords-Speech emotion recognition; Support vector machine; Kernel principal component analysis

I. Introduction

Interaction between humans and computers occurs in many forms. Humans communicate with computers through inputs, such as the mouse, keyboard, voice, images, and various modern sensors. Some are more difficult to process than the others. In this work, we focus on the voice interface, particularly on emotions derived from speech. The understanding of emotion from speech can be highly beneficial to applications, such as monitoring customer satisfaction during a phone conversation in a call center business and in the area of robotics.

In this paper, we used a feature set that is provided in the INTERSPEECH 2013 ComParE Challenge [1]. The set consists of energy, spectral, mel-frequency cepstral coefficients (MFCC), and voicing-related low-level descriptors (LLDs) as well as other LLDs, including jitter, shimmer, logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness.

It is inevitable that the more the features are used in speech emotion recognition, the greater the likelihood of the occurrence of the "curse of dimensionality" problem [2]. When such a problem happens, it naturally leads to lower accuracy in the classification. This problem may be reduced by optimal feature reduction. However, we believe that thus far, there are no theory-supported formulas for choosing optimal methods based on information about the dataset. Researchers tend to rely on empirical studies and follow previous research with similar datasets and learning tasks [3].

Some of the advantages of feature selection using the filter method include its scalability to very high-dimensional datasets and its classifier-independent property. For these reasons, many studies, including those in speech emotion recognition, favor this method. Examples include the use of chi-square with principal component analysis (PCA) on SEMAINE [4], the use of the gain ratio on a Berlin dataset [5], and the use of RELIEF-F to reduce features in song emotion recognition [6].

Apart from the feature selection method, feature extraction using PCA, kernel PCA (KPCA) is commonly used in speech recognition. For example, KPCA was used to reduce features in the detection of mild Alzheimer's disease from elderly speech [7], and PCA was used to reduced features in gender dependent emotion recognition [8].

Feature extraction in speech emotion recognition studies are extracted from three main levels [9]: the frame level (local features), utterance level (global features), and segment level. Most researchers have agreed that global features are more beneficial than local features in regards to classification accuracy and classification time. However, some researchers have a different view, that global features' efficiency is limited, as they are only able to differentiate between high-arousal (e.g. anger, fear, and joy) and low-arousal (e.g. sadness) emotions. On the other hand, the global features seem to fail to classify similar-arousal; yet temporal information can be lost in speech signals. Segment level feature extraction can be constructed from underlying phonemes, voiced speech segment or frame stacking [10]. In our proposed method, we use features from the segment level, because of using global features will yield fewer training vectors. This may be insufficient for reliably estimating model

978-1-5090-4834-2/17/$31.00 ©2017 IEEE

parameters, especially when using complex classifiers, such as the support vector machine (SVM).

For speech emotion recognition, there are numbers of appropriate classifiers, such as the hidden Markov model (HMM), Gaussian mixture model (GMM), SVM, and artificial neural networks (ANN). However, there is no clear evidence as which classifier is generally the most superior. Therefore, we selected SVM for our work, because it has advantages over GMM and HMM in terms of the global optimality of the training algorithm. Moreover, it is insensitive to model initialization [11].

This paper has two components, the first part is an empirical study that addresses emotion classification using a feature set extracted from the utterance level by comparing the classification results between feature reduction methods on emotional tagged corpus on "Lakorn" (EMOLA). In the second part, we proposed a speech emotion recognition model by using features from the segment level that is motivated by [10]. Then, we transformed the features by applying statistical functions and making the reduction using KPCA.

The rest of this paper is organized as follows. In Section II, we introduce four feature reduction methods and the SVM as a classifier. Section III describes the methodology, which reveals data preprocessing. We present the empirical study of four feature reduction methods in experiment I, as well as describe the proposed method in details in experiment II. Section IV includes the experimental results. Lastly, Section V contains the conclusion and future work.

II. Background Knowledge

In this paper, we review feature reduction methods and the SVM as a classifier.

A. Feature reduction methods

Feature reduction can be categorized into feature selection and feature extraction (feature transformation).

1) Feature selection methods

Feature selection is the identification subset of relevant features. We review three feature selection methods: chi-square, gain ratio, and RELIEF-F. Each method assigns a score to the feature. The larger scores indicate that the feature is more discriminative.

a) Chi-square (CHI)

The objective of the chi-square method is to identify dependability between two variables. In our case, the two variables are the feature and the class. The null hypothesis (H0) states that the two variables are independent. It requires that the data be on nominal or ordinal scale [12]. The equation for chi-square is shown below.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \tag{1}$$

In this equation, r represents the number of distinct features, while c is the number of classes. $O_{i,j}$ is the number of instances with i value in class j. Finally, $E_{i,j}$ is the expected number of instances with i value in class j.

$E_{i,j}$ is derived from *(p□q)/n*, where p is the number of instances with feature i, q is the number of instances with class j, and n is the number of all instances.

The larger the chi-square value, the more relevance there is between the feature and the class.

b) Gain ratio (GR)

Information gain stems from information theory. It posits that information value is dependent on the probability of the data measurable in bits. However, it may introduce bias toward features with many values. To reconcile such bias, the "split information" concept was introduced.

The gain ratio is a ratio between the information gain and the split information, which can be calculated with the following equations [13].

$$Gain\ Ratio(S, A) = \frac{Gain(S,A)}{Split\ Information(S,A)} \tag{2}$$

$$Gain(S, A) = Ent(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|}\, Ent(S_v) \tag{3}$$

$$Ent(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{4}$$

$$Split\ Information(S, A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{5}$$

In the above equations, S is the dataset, S_v is the subset of S for which attribute A has value v, S_i is the instances with class i, p_i is the ratio of the instances with class i, and c is the number of classes.

c) RELIEF-F (RF)

RELIEF-F was developed from RELIEF which can handle multiclass classification [14]. RELIEF-F randomly selects instance R_i and its k nearest instances in the same class, which is referred to as nearest hits, H_j. Then, it selects k nearest instances from each of the different classes, or nearest miss $M_j(C)$. The default value of the nearby neighbor's number k is 10.

RELIEF-F updates $W[A]$ weight for all features A using this equation.

$$W[A] = W[A] - \sum_{j=1}^{k} \frac{diff(A, R_i, H_j)}{m \cdot k} + Z \tag{6}$$

$$Z = \sum_{C \neq class(R_i)} \frac{\left[\frac{P(C)}{1 - P(class(R_i))} \sum_{j=1}^{k} diff(A, R_i, M_j(C))\right]}{m \cdot k} \tag{7}$$

The distance between instances X_1 and X_2 can be calculated using this equation.

$$Distance(X_1, X_2) = \sum_{j=1}^{j=n} diff(A_j, X_1, X_2) \tag{8}$$

Parameter m represents the number of instances for estimate the feature weights. A larger m implies more reliable approximation but it cannot exceed the number of training instances.

2) Feature extraction method

Feature extraction is the process of reducing the number of features by mapping from the original feature into another feature space while preserving variability in the data as much as possible.

a) Kernel Principal Component Analysis (KPCA)

PCA is a statistical algorithm that uses an orthogonal transformation to transform a group of features, for which some of the variables may be correlated with one another into a smaller number of linearly uncorrelated variables (principal components). The number of principal components is less than or equal to the number of original features or the number of instances depending on which number is lower.

However, most data are naturally in a nonlinear form, and PCA can be incapable to convert the data efficiently. One of the methods for managing nonlinear data is the use of KPCA, which involves mapping from d-dimensional space into a higher n-dimensional space by using a nonlinear kernel before using PCA on the extended space.

B. Support vector machine

In our work, we use the SVM, which is a supervised classifier. The data are separated by using the linear hyperplane. When there are several hyperplanes with suitable planes, the plane that yields the least amount of error and has a large margin, maximum marginal hyperplane (MMH) will be selected [15].

The SVM utilizes the kernel functions to nonlinearly map the original features to a high-dimensional space so that the data can be separated with a linear classifier. However, it cannot be guaranteed that the classification will be successful after the transformation. Though the separation may not be perfect in this case, the advantage of this method is the minimization of the over-fitting problem [9].

III. Methodology

In this paper, we begin with an empirical study to find the optimal feature reduction method which is shown in experiment I. Then, we apply it in our proposed method, which is detailed in experiment II. We use the same data preprocessing method in both experiments.

A. Data preprocessing

This work uses the data from the EMOLA corpus from Thailand's National Electronics and Computer Technology Center (NECTEC), where "Lakorn" means Thai TV shows. This particular corpus has an education version, which is a collection of 2,908 utterances in 16-bit, stereo, with a sampling rate of 48 KHz. We filtered out unusable files, such as those that were incomplete or had two simultaneous talkers. We also eliminated files with incorrect or missing subtitles, because these files do not have class labels. Files with noise or background music were also excluded, because we wanted to focus on emotion classification from clean speech without having to consider background noise and music. The final set was a corpus with 551 utterances that comprised four emotions: anger, happiness, sadness, and neutral. There were 170 utterances for anger, 211 for happiness, 52 for sadness, and 118 for neutral. We then converted the files into mono-channel using the SoX program and remove the silent portions at the beginning and the ending of the files using Praat [16].

All of the available data were divided into two sets using stratified random sampling to maintain the ratio of data for each of the emotion. The ratio for the training set to the test set was 80% to 20%, which resulted in 440 and 111 files, respectively. The details of the number of utterances by emotion are shown in TABLE I.

TABLE I. Number of utterances in the training and test sets

	Emotional classes				
	Anger	***Happiness***	***Sadness***	***Neutral***	***Total***
Training	136	168	41	95	440
Testing	34	43	11	23	111
All	170	211	52	118	511

B. Experiment I: Empirical study

This experiment shows a comparison between the use of various feature reduction methods: chi-squared, gain ratio, RELIEF-F, and KPCA. All features are from the utterance level.

We extracted the features from the converted files using openSMILE [17] and by relying on the template provided at the INTERSPEECH 2013 ComParE Challenge. There were 6,373 features, which are the statistical data from LLDs. These features cover prosodic, spectral, and voice quality. We then normalized the data using z-scores.

We computed feature importance values using chi-square, gain ratio, and RELIEF-F, and sorted the scores in descending order before selecting those features from 10% to 100%.

Before using the chi-square and gain ratio selection methods, we performed data discretization. To our knowledge, there is no known research that suggests a preferable method(s) for discretizing data for speech emotion recognition. Therefore, we selected class-attribute interdependence maximization (CAIM) [18], as it is one of the more popular methods [19].

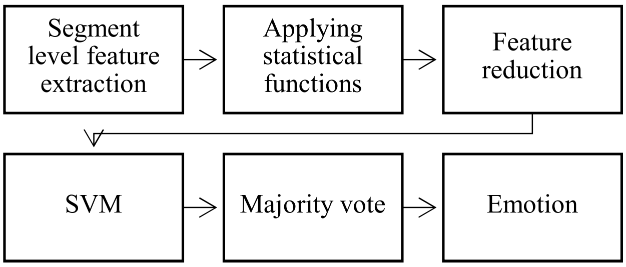
For KPCA, we used the sigmoid kernel, set the independent term to 2 and selected the first 72 principal components to retain 90% of the variance.

The training and testing tasks are carried out using the SVM with radial basis function (RBF) kernel. We set the class weight to balance the importance of each class while adjusting the parameters and set $\gamma$and C with a grid-search using 5-fold cross-validation. The parameters that we used were C=2^(-5), 2^(-3), …, 2^15 and γ=2^3, 2^1, …, 2^(-15). Finally, we selected those parameters that yielded the highest F-score from each of the subsets of the features.

Hence, as we have unbalanced classes and prefer to avoid the results from being biased given the imbalanced data distribution, we used macro-averaged F-scores to evaluate the classification performance.

C. Experiment II: Proposed method

Here, we will describe the details of our approach for the second part of experiment. Fig. 1 shows the overview of the approach.



Segment level feature extraction
Applying statistical functions
Feature reduction
SVM
Majority vote
Emotion

Figure 1. The proposed approach

It began with extracting the segment level features. Then, we transformed the features by applying statistical functions and making the reduction by using the feature reduction method from the previous experiment's results. After that, we classified the data with the SVM and use its results to perform a majority vote for determining the emotional state of the whole utterance.

In this second experiment, we again extracted the features from the converted files using openSMILE and the template provided at the INTERSPEECH 2013 ComParE Challenge. However, there is a small change in this extract, as it will be extracted using LLDs without the statistical functions. It is done at the frame level by extracting every 10 ms with 60 ms frames for voicing related LLDs and 25 ms frames for all other LLDs. At the end, there were a total of 130 features.

Next, we constructed features in the segment level by adapting the method from [10]. In each utterance, we stacked up features from each frame for 25 frames by setting 12 frames before and after. Thus, the total length of a segment was at least 10 ms x 25 + (25-10) ms = 265 ms. Some segments in an utterance may or may not contain emotional information. It is acceptable to presume that the highest energy segments contain the most eminent emotional information. Consequently, we selected the 10% of the segments that had a highest total value of root-mean-square signal frame energy in each utterance, both in the training and test sets. Then, we assigned a certain class label to every segment in each utterance, which led to data comprise of 130 x 25 = 3,250 features with the total instance of 19,654 and 4,654 for training and test set, respectively.

Due to the large number of features and a significant amount of data, using the feature reduction method directly may require a long computation time. Moreover, we needed to smoothen the data to reduce noise. In order to reduce the number of features and smoothen them before using the feature reduction technique, we transformed features by calculating the value of minimum, maximum, mean, and standard deviation of each LLD from every frame in each segment, as shown in equations (9) to (12). From such transformation, we obtained global features which were derived from the segment level instead of the utterance level.

$$f_1^k = \min_{j \in J} f_j^k \tag{9}$$

$$f_2^k = \max_{j \in J} f_j^k \tag{10}$$

$$f_3^k = \frac{1}{|J|} \sum_j^J f_j^k \tag{11}$$

$$f_4^k = \sqrt{\frac{1}{|J|} \sum_j^J \left(f_j^k - \mu^k\right)^2} \tag{12}$$

In the above equations, f_j^k is the feature (LLD) k of frame j, μ^k is the mean of feature k, and J is a collection of frames. After transforming, there remained the features of 130 x 4 = 520. We then normalized the data using z-scores and applied the feature reduction method before classifying with the SVM.

Since the classification result is in the segment level, it must be transformed into an utterance by majority voting. The majority vote selects the class with the highest number of votes from all segments in an utterance. In some cases, the voting results in an equal value. Consequently, we calculated the average of the segment-level probabilities in each emotion from every segment of an utterance, then assigned the emotion that had the highest value. The reason why we focused on conducting a majority vote on predicted class instead of initially using probabilities is because the probability model from the SVM is created using cross validation, which yields results that are slightly different than those obtained by prediction [20].

IV. Results

A. Experiment I Result

Because of classifying without using feature reduction, we noticed that the values of precision, recall and F-score were at 0.490, 0.473, and 0.471, respectively. In the experiment, we observed that all feature selection techniques can improve classification performance when it is measured with the F-score. By using the gain ratio, the highest value among all methods is the F-score of 0.503 at the feature size of 30%, followed by the RELIEF-F with the F-score of 0.498 at the feature size of 40%, and the chi-square with the F-score of 0.484 at the feature size of 10% (see TABLE II).

TABLE III, shows the average and maximum precision, recall, and F-score values only where there is feature selection (10% to 90%). Bold values are the values of precision, recall and F-score that were equal to or higher than the data classification without feature selection.

It is noticeable that using the gain ratio, which is a univariate selection method, gives a better average F-score value than RELIEF-F, which is multivariate method. It is also important to note that the characteristics of features in the feature space are counted in order to improve the classification performance using the feature selection method.

TABLE II. Macro-averaged precision, recall, and F-score of the SVM classifier for different feature sizes with feature selection methods on the EMOLA

Feature Count		Precision	Recall	F-score
10% (638)	CHI	0.486	0.487	0.484
	GR	0.494	0.520	0.501
	RF	0.417	0.404	0.406
20% (1275)	CHI	0.443	0.464	0.445
	GR	0.443	0.441	0.441
	RF	0.461	0.470	0.462

Feature Count		Precision	Recall	F-score
30% (1912)	CHI	0.392	0.403	0.393
	GR	0.504	0.510	0.503
	RF	0.451	0.455	0.448
40% (2550)	CHI	0.420	0.437	0.425
	GR	0.475	0.489	0.480
	RF	0.492	0.514	0.498
50% (3187)	CHI	0.432	0.422	0.418
	GR	0.491	0.500	0.495
	RF	0.473	0.485	0.476
60% (3824)	CHI	0.477	0.487	0.480
	GR	0.481	0.483	0.482
	RF	0.465	0.466	0.463
70% (4462)	CHI	0.466	0.457	0.459
	GR	0.460	0.454	0.455
	RF	0.453	0.457	0.453
80% (5099)	CHI	0.450	0.449	0.445
	GR	0.467	0.432	0.430
	RF	0.475	0.478	0.475
90% (5736)	CHI	0.493	0.482	0.480
	GR	0.472	0.451	0.450
	RF	0.482	0.476	0.474
100% (6373)	CHI	0.490	0.473	0.471
	GR	0.490	0.473	0.471
	RF	0.490	0.473	0.471

TABLE III. AVERAGE AND MAXIMUM F-SCORE VALUES WITH FEATURE SELECTION METHODS

	Precision		Recall		F-score	
	Avg.	*Max.*	*Avg.*	*Max.*	*Avg.*	*Max.*
CHI	0.451	**0.493**	0.454	**0.487**	0.448	**0.484**
GR	0.476	**0.504**	**0.475**	**0.520**	**0.471**	**0.503**
RF	0.463	**0.492**	0.467	**0.514**	0.462	**0.498**

In addition to considering the classification performance of various filtering methods, we may also consider whether all three feature selection methods have the selection patterns that run in the same direction without considering the results from classification. This can be done by comparing the rating of each pair of feature selection by using Kendall's Tau, in which the correlation will be high when the selection methods have a similar rank.

The results of Kendall's Tau for each pair of selection methods are shown in TABLE IV. It can be concluded that the pairs of the selection methods are not related (p-value greater than 0.05).

TABLE IV. CORRELATION AMONG EACH PAIR OF THE SELECTION METHODS

	Kendall's Tau	p-value
CHI-GR	-0.004	0.662
CHI-RF	0.006	0.494
GR-RF	0.000	0.994

According to feature reduction techniques, using KPCA yields the precision, recall, and F-score values at 0.507, 0.493, and 0.498, respectively. Its F-score is higher than the average F-score value from other feature selection methods and can efficiently reduce the number of features. Therefore, we selected KPCA for use in our proposed method.

B. *Experiment II Result*

We compared six approaches to the classification of whole utterances.

Approach A, utterance level: the features were calculated at the utterance level and yielded 6,373 features (see experiment I).

Approach B, utterance level using KPCA: the features were calculated at the utterance level, then KPCA was performed and yielded 72 features (see experiment I).

Approach C, segment level: constructed features at the segment level and obtained 3,250 features.

Approach D, segment level using KPCA: constructed features at the segment level, then performed KPCA by selecting the first 348 principal components to retain 90% of the variance.

Approach E, segment level using transformed features: constructed features at the segment level, then applied statistical functions to them, yielding 520 features.

Approach F (proposed method), segment level using transformed features and KPCA: similar to approach D but applied statistical functions to features before performing KPCA to reduce the number of features to 164.

For approaches C to F, we used a majority vote to transform the classification results into the utterance level. Approaches A to E were used as baselines.

As shown in TABLE V, approach A gave an F-score at 0.471, approach B at 0.498, approach C at 0.507, approach D at 0.508, approach E at 0.480, and approach F at 0.533. Approach F yielded a higher F-score than approach A by 13.16%, 7.03% for approach B, 5.13% for approach C, 4.92% for approach D and 11.04% for approach E. Bold letters show the higher value than the other approach.

TABLE VI shows the confusion matrix from classifying with approach F. The highest accuracy for anger is at 0.824, while happiness and neutral rank the worst at 0.488 and 0.391, respectively.

TABLE V. SHOWING PRECISION, RECALL, AND F-SCORE VALUES FROM CLASSIFYING ALL FOUR APPROACHES

	Precision	Recall	F-score
Approach A	0.490	0.473	0.471
Approach B	0.507	0.493	0.498
Approach C	0.523	0.509	0.507
Approach D	**0.565**	0.501	0.508
Approach E	0.548	0.477	0.480
Approach F	0.544	**0.562**	**0.533**

TABLE VI. CONFUSION MATRIX OF APPROACH F CLASSIFICATION

Actual class	**Predicted class**			
	Anger	***Happiness***	***Sadness***	***Neutral***
Anger	0.824	0.029	0.029	0.118
Happiness	0.256	0.488	0.163	0.093
Sadness	0.182	0.182	0.545	0.091
Neutral	0.217	0.174	0.217	0.391

TABLE VII shows the confusion matrix of classifying with approach D. It has a higher precision value than the other methods. The highest accuracy was obtained for anger at 0.824, similar to approach F, but it performed better when classifying happiness. Sadness and neutral had the lowest accuracy at 0.273 and 0.348, respectively. Sadness had significantly lower accuracy with this approach compared to approach F.

TABLE VII. CONFUSION MATRIX OF APPROACH D CLASSIFICATION

Actual class	**Predicted class**			
	Anger	***Happiness***	***Sadness***	***Neutral***
Anger	0.824	0.118	0	0.059
Happiness	0.279	0.558	0.047	0.116
Sadness	0.364	0.273	0.273	0.091
Neutral	0.261	0.391	0	0.348

V. CONCLUSION

In this paper, we evaluated various types of feature reduction methods in the first part of the experiment. We found that all four types of reduction method could help to improve the classification performance in an utterance level with the F-score indicator on EMOLA with the SVM as a classifier. Each technique is able to capture various characteristic features. Thus, the after-feature set can provide the better data. KPCA yielded the best result in classification when its F-score was compared to the average F-score of other selection methods. From the experiment of using Kendall’s Tau statistics, we noticed that the selection methods (CHI, GR, and RF) were not related to each other; in another hand each selection method has a different pattern.

For the proposed method, we made classifications using features from the segment level, and then we transformed the features by applying statistical functions and KPCA. The experimental results indicate that this yielded higher classification efficiency than other methods when measured with the F-score.

Our future research will focus on evaluating the proposed method on other speech emotion datasets that more closely to resemble real world situations. In addition, we may use speech features together with other information, such as word meanings instead of using speech features alone.

REFERENCES

[1] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.

[2] E. Keogh and A. Mueen, "Curse of Dimensionality," in *Encyclopedia of Machine Learning*, C. Sammut and G. Webb, Eds., ed: Springer US, 2010, pp. 257-258.

[3] Y. Aphinyanaphongs, L. D. Fu, Z. Li, E. R. Peskin, E. Efstathiadis, C. F. Aliferis, *et al.*, "A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization," *Journal of the Association for Information Science and Technology,* vol. 65, pp. 1964-1987, 2014.

[4] R. Calix, M. Khazaeli, L. Javadpour, and G. Knapp, "Dimensionality Reduction and Classification Analysis on the Audio Section of the SEMAINE Database," in *Affective Computing and Intelligent Interaction*. vol. 6975, S. D’Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 323-331.

[5] M. Bhargava and T. Polzehl, "Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature," *CoRR,* vol. abs/1303.1761, 2013.

[6] R. Panda, B. Rocha, and R. P. Paiva, "Music Emotion Recognition with Standard and Melodic Audio Features," *Applied Artificial Intelligence,* vol. 29, pp. 313-334, 2015/04/21 2015.

[7] S. Kato, A. Homma, T. Sakuma, and M. Nakamura, "Detection of mild Alzheimer's disease and mild cognitive impairment from elderly speech: Binary discrimination using logistic regression," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 5569-5572.

[8] S. J. Chaudhari and R. M. Kagalkar, "Automatic Speaker Age Estimation and Gender Dependent Emotion Recognition," *International Journal of Computer Applications,* vol. 117, pp. 5-10, 2015.

[9] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* vol. 44, pp. 572-587, 3// 2011.

[10] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech*, 2014, pp. 223-227.

[11] C. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, pp. 121-167, 1998/06/01 1998.

[12] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "Feature Selection with High-Dimensional Imbalanced Data," in *Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on*, 2009, pp. 507-514.

[13] T. M. Mitchell, *Machine Learning*: McGraw-Hill, 1997.

[14] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Machine Learning: ECML-94*. vol. 784, F. Bergadano and L. De Raedt, Eds., ed: Springer Berlin Heidelberg, 1994, pp. 171-182.

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Massachusetts: Morgan Kaufmann Publishers Inc., 2011.

[16] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International,* vol. 5, pp. 341-345, 2001.

[17] F. Eyben, F. Weninger, F. Gross, Bj, #246, and r. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," presented at the Proceedings of the 21st ACM international conference on Multimedia, Barcelona, Spain, 2013.

[18] L. A. Kurgan and K. J. Cios, "CAIM discretization algorithm," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 16, pp. 145-153, 2004.

[19] S. Garcia, J. Luengo, Sa, x, J. A. ez, Lo, *et al.*, "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 25, pp. 734-750, 2013.

[20] scikit-learn.org. (2016). *sklearn.svm.SVC Documentation* [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html