Semi-supervised Learning Techniques for Speech Emotion Recognition

K Remya\(^1\), B S Shajee Mohan\(^2\) and K V Ahammed Muneer\(^3\)

\(^1\) M.Tech Signal Processing, Applied Electronics and Instr. Engineering Dept., Govt. Engineering College Kozhikode, Kerala, India, 673005
\(^2\) Applied Electronics and Instr. Engineering Dept., Govt. Engineering College Kozhikode, WestHill, Kerala, 673005
\(^3\) Applied Electronics and Instr. Engineering Dept., Govt. Engineering College Kozhikode, WestHill, Kerala, 673005

E-mail: remyakrajagopalan@gmail.com, smohan@geckkd.ac.in, ahammedcet@gmail.com

Abstract. The core objective of this paper is to explore semi-supervised learning methods for recognizing emotions contained in speech. Semi-supervised learning techniques are used when the availability of labeled examples are sparse. There are different methods that are used in the semi-supervised settings. These techniques include generative model, graph based methods, metric based methods etc. The speech emotion data is considered for this experiment. Speech signal contains emotion specific data also. Emotion dependent features are extracted from the speech. This paper aim to enhance some existing techniques for semi-supervised learning that are used for speech emotion recognition.

Keywords: MFCC, SVM, semi-supervised learning, PCA, SS-DML, distance metric learning

1. Introduction

Speech is the most popular means for communication. The objective of this work is to identify human emotions from speech signals. Capturing the emotions from speech is difficult, the proper feature selection in relation with both the time and frequency domains together is necessary to produce optimal results. This work mainly concentrate on recognizing seven emotions: happy, fear, sad, anger, disgust, boredom and neutral from speech. For this, features like Mel-Frequency Cepstral Coefficients (MFCC), Energy, Linear prediction cepstrum coefficient (LPCC), Pitch etc., are used. Semi-supervised support vector machine (SVMs), can be used for recognizing emotions contained in the speech. Berlin dataset (German emotional speech database) is used for the experiments.

This paper is organized in 7 Sections. Section 1 gives the introduction and Section 2 gives an overview of speech emotion recognition using semi-supervised techniques, Section 3 provides the related works in speech emotion recognition systems, Section 4 discusses the methodology of the proposed system for speech emotion recognition, Section 5 explains the experimental setup including data set description, Section 6 gives the results and observations and in Section 7 gives the concluding remarks of the work and the future scope.
2. Overview of Speech Emotion Recognition using Semi-supervised Techniques
The aim of speech emotion recognition system is to recognize the emotion of a human by extracting emotion dependent features from the voice of the speaker and to take the appropriate methods to identify emotions from speech. Speech emotion recognition system mainly includes the following elements: recording of speech input, feature extraction, feature selection and optimization, classifier and recognized output (decision). There are many speech emotion recognition systems, each speech emotion recognition system differ from the other by its classification methods used. Initially features extraction is performed in speech emotion recognition system, then appropriate features are selected and applied to the classifier, classifier recognizes the emotion within the given speech input. Various types of classifiers have been used for speech emotion recognition. Artificial Neural Network (ANN), Hidden Markov Model (HMM), $k$-nearest neighbors ($k$NN), Gaussian mixture model (GMM) and Support Vector Machine (SVM) etc., based methods are utilized as the classifiers. Each classifier has its own limitations and advantages.

3. Related works
Two main semi-supervised learning approaches are proposed for the emotion recognition in the literature: Co-training and self training. The work done in [2] shows an enhanced co-training algorithm that utilizes a huge number of unlabeled speech input for a semi-supervised learning system. This framework makes 9.0% improvement on female model, 7.4% on male model in normal precision. It reduces noise that is produced during classification by error labeling in the semi-supervised learning. The work in [4] proposes semi-supervised autoencoders to improve emotion recognition. It employs combination of labeled data and unlabeled input information. The model proposed in this work expands a prevalent unsupervised autoencoder by connecting a supervised learning objective. They uses INTERSPEECH 2009 emotion challenge database for their work. The work in [3] proposes a ladder network based semi-supervised learning to create an effective representation of features suitable for emotion recognition from speech. Normalized static acoustic features are utilized as the input to ladder network and these features are mapped to high level hidden representations. The system train the model to concurrently minimize the cost functions of supervised and unsupervised learning using back propagation. The extricated hidden representations are utilized as emotional features in SVM model for emotion recognition from speech. The test done on IEMOCAP database, gives 2.6% performance improvement over denoising autoencoder and 5.3% performance improvement over the inactive acoustic features.

4. Methodology of the Proposed System

![Diagram](image)

**Figure 1.** The block schematic of the proposed system for speech emotion recognition
The block diagram in the Figure 1 represents the methodology of the semi-supervised speech emotion recognition system. In speech emotion recognition system, features extracted from the speech such as pitch, energy, MFCC, LPCC are given to a classifier such as kNN, SVM, BPNN etc. The system mainly contains five modules: input database (speech), feature extraction, feature selection and optimization, classifier and recognized output (decision). The system make use of the study of the production component of speech signal there by extraction of emotion dependent features. Initially emotional relevant features are selected from the extracted speech features. Finally, classifier detects the emotions using semi-supervised algorithm. The evaluation of the system is done using various performance measures.

4.1. Preprocessing

Noise removal is a preprocessing step to remove the unwanted signals in the input. Mainly STFT based method is used for noise removal. The extracted features may have different units, some of them may be biased and hence their magnitude may have different order, this can result adverse effects while training the classifiers. Thus feature normalization is necessary in these cases. Min-max normalization is considered in the study.

4.1.1. STFT based speech noise reduction

STFT can be used to reduce noise in a speech signal. It consists of three steps:

(i) Compute the STFT of noisy signal

\[ S(m,w) = STFTx(n) \] (1)

(ii) Threshold the STFT

\[ S_2(m,w) = g(S(m,w)) \] (2)

where \( g(a) \) is a thresholding function

(iii) Compute the inverse STFT

\[ y(n) = STFT^{-1}S_2(m,w) \] (3)

STFT-thresholding is a non-linear noise reduction technique because the thresholding operation is nonlinear.

4.2. Feature extraction

Large number of parameters which contains the emotional characteristics are present in speech signals. For emotion recognition system such features are used. Two main categories of speech features, that is long term and short term features [6]. Emotions in speech is mainly indicated by the features like Mel frequency cepstral coefficient (MFCC), linear prediction cepstral coefficient (LPCC), pitch, duration, formant and energy [6]. With the different emotions in the speech, the extracted features from the speech reflect these variability.

4.2.1. Mel-Frequency cepstral coefficients (MFCC)

MFCC are popularly utilized for representing the spectral properties of speech signals. MFCC is most commonly used feature for speech emotion recognition because it takes human discernment affectability with respect to frequencies into thought. The Fourier transform can be computed for each frame and are mapped into the mel-frequency scale. The discrete cosine transform (DCT) of the mel log energies were computed and therefore the primary 12 DCT coefficients provide the MFCC values which are utilized within the process of classification. The schematic of the MFCC calculation is demonstrated in Figure 2 [10] . MFCC is extensively employed in speech emotion recognition systems and also the rate of recognition using the MFCC is extremely sensible. Within the
low frequency region, higher frequency resolution and robustness to noise may possibly be accomplished with the help of MFCC than for high frequency. MFCC represents the short term power spectrum of sound.

4.2.2. Linear prediction cepstrum coefficient (LPCC) : LPCC gives the characteristics of explicit channel of any person and this channel characteristic will get changed in accordance with various emotions, therefore by exploitation of these features one can extract the emotions in speech. The advantages of LPCC is that, it have less computational overhead, is more effective and it might depict the vowels in a much superior way.

4.3. Feature selection & optimization
Feature selection and optimization are exceptionally critical steps in speech emotion recognition. The need to utilize an algorithm for feature selection is that there are numerous features contained in a speech and there’s no fixed set of features to show the emotions. The classifiers faces the problem of increased dimensionality, increase in training time and over-fitting that will influence the rate of prediction when a huge sum of features are included within the process. Feature selection is that the method of selecting the foremost important and useful subset of features from the extricated set of features. The insignificant qualities are distinguished and expelled to increase the accuracy of the predictive model. Before classification, feature selection, also referred to as variable selection or feature reduction, is commonly utilized in speech emotion recognition so as to accelerate the training process. One of the popular feature selection methods is principal component analysis (PCA).

4.3.1. Principal component analysis : The data vector $y_i$ from a space of $m$ variables is mapped to a new space of $m$ factors which are uncorrelated over the dataset by the transformation $U = YW$. The truncated transformation, $U_K = YW_K$ is given by keeping the first $K$ principal components, that are produced by using only the first $K$ Eigen vectors. Here the matrix $U_K$ has $K$ columns and $m$ rows. PCA learns a linear transformation $u = W^U y, y \in \mathbb{R}^m, u \in \mathbb{R}^K$, here the columns of $m \times K$ matrix $W$ form an orthogonal basis for the $K$ features that are decorrelated. The transformed data matrices have $K$ columns the scatter matrix maximises the variance in the preserved original data while minimises the total squared reconstructed error that is given as:

$$\|UW^U - U_K W_K^U\|^2_2 \quad \text{or} \quad \|Y - Y_K\|^2_2$$

4.4. Classification
There are various types of classifiers which can be utilized for speech emotion recognition systems, some of them are based on HMM, GMM, SVM, ANN, kNN etc., In fact, an optimal classifier can be selected that is suitable for the emotion recognition task. It is also noted that every classifier has its own limitations and advantages. During this study SVM based classifier is
used for the classification. Semi-supervised learning technique is employed to train the classifiers using limited number of labels.

4.4.1. Semi-supervised learning: Mainly, there are 2 types of algorithms in machine learning, they are

- Unsupervised learning algorithm (no labeled data)
- Supervised learning algorithm (labeled data)

Table 1. Comparison of characteristics of supervised and unsupervised learning methods

| Parameters       | Supervised     | Unsupervised   |
|------------------|----------------|----------------|
| Input data       | Labeled data input | Unlabeled data input |
| Computational complexity | less complex | Complexity is high |
| Accuracy         | Highly accurate | Less accurate |

A brief comparison of unsupervised and supervised learning is given in Table 1. To overcome the demerits of supervised and unsupervised learning alogorithms, semi-supervised learning is utilized. It is an approach to machine learning that make use of a few labeled information with a large quantity of unlabelled information throughout training. Various types of semi-supervised techniques are generative models, graph based models, self training, co-training and label propogation. Self training and label propogation methods are utilized in this proposed system.

Self training:
Self-training is a popularly utilized technique for semi-supervised learning. In this technique a classifier is prepared with accessible set of labeled data in the first step, then the classifier is utilized to classify the unlabeled examples. The foremost guaranteed unlabeled data, along with their anticipated labels, are included to the training set. The new training set is used to re-train the classifier and thus the strategy is repeated. This method in which the classifier is re-trained by itself is known as self-training or bootstrapping.

Label propagation:
A popular approach to semi-supervised learning is to create a graph that connects examples in the training dataset and propagate known labels through the edges of the graph to label unlabeled examples. An example to this approach to semi-supervised learning is the label propogation algorithm for classification predictive modeling[8]. Label propagation algorithm (LPA) is used to label previously unlabeled data using a semi-supervised machine learning algorithm. It is assumed that a small subset of data examples have labels during the initialization of the algorithm.

5. Experimental setup
The .wav files from Berlin emotion dataset containing emotional speech utterences is the input to the system. Then feature extraction is carried out. In feature extraction process, MFCC and mean of MFCC features are extracted. At that point the extracted features and their respective class labels are provided as input to LibSVM classifier. The classifier yield could be a name of particular emotion class. There are add up to seven feeling classes: disgust, boredom, angry, sad, happy, neutral and fear. Each label represents individual emotion class.
5.1. Feature extraction
The proposed system for speech emotion recognition is implemented using LibSVM tool with the features MFCC, mean of the MFCC etc. For the MFCC extraction, the audio signals are framed as short frames. For each frames 13 MFCC coefficients are computed. Number of frames for each audio signals depends on the length of the audio signals. MFCC of the each of the seven emotions speeches are concatenated separately. Mean of the MFCC are computed from the concatenated MFCC. Here, the dataset contains 62 sad, 80 neutral, 71 happy, 69 fear, 46 disgust, 81 boredom and 126 anger examples. Dimension of MFCC based feature vector showing number of frames and MFCC per frame for a speech utterance is shown in Table 2.

|       | Dimension of MFCC per frame |
|-------|-----------------------------|
| Sad   | 392×13                      |
| Neutral | 389×13                      |
| Happy | 257×13                      |
| Fear  | 321×13                      |
| Disgust | 341×13                      |
| Boredom | 341×13                      |
| Anger | 251×13                      |

Now, MFCC and mean of the MFCC frames are used for the classification of emotions. More features will be included in the future to obtain better results than obtained with MFCC and mean of the MFCC.

5.2. Dataset description
It is very important to have an appropriate and adequate dataset for training the speech emotion recognition framework and then to assess by the training dataset and performance and robustness using appropriate performance metrics. Systems are influenced in the event that they are not well trained, with the most appropriate dataset performance may degrade. Berlin emotional dataset is considered for this experiment.

5.2.1. Berlin dataset
The Berlin emotion dataset [9] is popularly used in emotion recognition from speech. It consists of a total of 535 utterances in 7 recreated emotions (neutral, boredom, fear, disgust, sadness, happy and anger) that is spoken by ten actors (five female, five male) [9]. The merits of Berlin dataset are: i) good quality recordings ii) public and popular dataset.

5.3. Classifier
In the first stage of this work, SVM classifier is utilized using supervised learning for emotion classification. Then semi-supervised learning for emotion recognition using a SS-SVM classifier is considered. SVM classifiers are useful technique for data classification [7]. The classification task typically includes the separation of data into training and testing sets. Every instance is associated with one class labels and multiple features or variables observed in the training set. The purpose of the SVM is to generate a model which is based on training set and to predicts
the class labels of test data, given as it were the test data attributes. Consider \((y_i, z_i), i = 1, \ldots, m\) where \(y_i \in \mathbb{R}^m\) and \(z \in 1, 1^m\) with a training set of example-label pairs, SVM solves the following optimization problem:

\[
\min_{u, b, \xi} \frac{1}{2} u^T u + C \sum_{i=1}^{m} \xi_i
\]

subject to

\[
z_i (u^T \phi(y_i) + b) \geq 1 - \xi_i \\
\xi_i \geq 0
\]

Training vectors \(y_i\) are mapped into a higher dimensional space by the \(\phi\) function. SVM finds a straight isolating hyperplane with the maximum margin in this higher dimensional feature space. The penalty parameter of the error term is \(C > 0\) and the slack variables \(\xi_i\) are considered to be positive.

Further, \(K(z_i, z_j) \equiv \phi(z_i)^T \phi(z_j)\) is called the kernel function. The four basic kernels are: [7]

- Linear kernel: \(K(u_i, u_j) = u_i^T u_j\)
- Polynomial kernel: \(K(u_i, u_j) = (\gamma u_i^T u_j + r)^d, \gamma > 0\)
- Radial basis function (RBF) kernel: \(K(u_i, u_j) = \exp(\gamma ||u_i - u_j||^2), \gamma > 0\)

Here \(r, \gamma\) and \(d\) are the kernel parameters.

LibSVM may be a most popularly used tool for SVM classification and regression.

6. Results and Discussion
   6.1. Classification results using supervised SVM

| Kernel       | Kernel parameter | Classification accuracy (in %) |
|--------------|------------------|-------------------------------|
| Gaussian     | \(\sigma = 0.01\) | 58.071 ± 3.6521               |
|              | \(\sigma = 0.02\) | 60.1677 ± 7.128               |
|              | \(\sigma = 1/dim\) | 56.6037 ± 2.5660              |
| Polynomial   | Degree = 2       | 59.11 ± 1.2327                |
|              | Degree = 3       | 57.444 ± 6.052                |
| Linear       |                  | 44.444 ± 2.9630               |

LibSVM is used for the speech emotion classification, accuracy of 56.6% is obtained while using Gaussian kernel with sigma=1/dimension, accuracy of 58.07% is obtained for Gaussian kernel with sigma 0.02 and an accuracy of 60% is obtained for Gaussian kernel with sigma 0.01 are obtained. An accuracy of 59% and 58% are obtained for the polynomial kernel of degree 2 and for degree 3 respectively. For the linear kernel the obtained accuracy is 44%. The maximum accuracy obtained is of 60% for Gaussian kernel with sigma value of 0.02.
The obtained accuracy of supervised speech emotion recognition system for all the above cases are tabulated in Table 3. Confusion matrices for the above cases are also obtained and observed, it is observed that in most of the all cases the emotion sad is classified almost accurately which shows that sad is easy to classify and the emotion disgust is very hard to classify as in most cases it has been misclassified.

6.2. Distance metric learning based semi-supervised SVM
This section discusses distance metric learning (DML). A DML based semi-supervised SVM is proposed in [13]. This technique make use of a method that is used to learn a kernel gram matrix that is optimum for the given task. The optimum kernel gram matrix is for the complete examples with only a few training examples whose label data is known. The learning strategy uses Bregman projection that enables the learning of kernel gram matrix for the whole training data containing pairwise similarity details for a few training examples. Let the training dataset $T_R$ is divided into two sets, as a small labeled set $L$ and a large unlabeled set $U$. $\hat{L}$ is considered as the set containing labeled examples incrementally available after every $m^{th}$ iteration step so that

$$\hat{L}_0 = \phi, \hat{L}_i \cap \hat{L}_j = \phi, \forall i \neq j$$

The set S and set D initially consists of pairwise constraints that are taken from the set L. The base kernel gram matrix $K_0$ is computed for the complete training set. The next step (i.e., $m=0$) is to learn an optimal kernel gram matrix $K_0^*$ for the complete training data using the base kernel $K_0$ and the pairwise constraints. This is done using DML methods. Then $K_0^*$ is used in the SVM classifier that is based on posterior probability. Here the SVM classifier makes the advantage of a confidence measure that is computed from the posterior probabilities of the class label of an input test data sample. Here the examples in the set $U$ is considered as the test data. The examples contained in the set $U$ are taken as the test dataset for this SVM. For every $m^{th}$ step, an optimum kernel gram matrix $K_{m-1}^*$ is learnt using the $K_{m-1}$ obtained in the $(m-1)^{th}$ step [13]. The cyclic process of assigning labels to the examples in the the set $U$ is repeated

![Figure 3. Illustration of the semi-supervised DML based SVM classifier that uses SVM that uses a confidence value to label unlabeled training data [14](Image)](image)
until all the examples contained in the set $U$ are labeled. The entire process is explained in the illustration in Figure 3. The classification accuracies are determined for the speech emotion data using SVM classifier that make use of SS-DML kernels that are learnt from two base kernels namely polynomial kernel (degree 2) and Gaussian kernel (sigma = 1/dim). The results are given in the Table 4. It is observed that the SVM classifier that is based on SS-DML technique performs well even with only a considerably small amount of training examples are labeled.

### 6.3. Results for SVM classifiers that uses SS-DML kernels

| Dataset                     | Kernel                  | Classification accuracy(%) | Average training time |
|-----------------------------|-------------------------|----------------------------|-----------------------|
|                             |                         | 50% | 40% | 30% | 20% | 100% |                     |
| Berlin emotion dataset      | polynomial degree=2     | 59.02 | 50.82 | 45.90 | 39.34 | 60.65 | 488.15s            |
|                             |                         | 462.41s | 614.9s | 551.04s | 480.99s | 313.41s |                     |
|                             | Gaussian sigma = 1/dim  | 59.02 | 50.81 | 40.98 | 49.18 | 75.41 | 392.66s            |
|                             |                         | 467.39s | 421.91s | 399.07s | 399.07s | 173.29s |                     |

### 7. Conclusion

Berlin emotion dataset in German language is considered for the work. MFCC and mean of MFCC features are extracted from the input speech data in .wav format. The accuracy and confusion matrix are computed using LibSVM classifier that uses Gaussian kernel and polynomial kernel. Better results are obtained while changing parameters of the kernel functions for SVM based classifier. Semi-supervised classification using semi-supervised SVM classifiers is also experimented and obtained accuracies around 50% by providing only 20%, 30%, 40% and 50% of labeled data respectively for training the classifier. In future works, semi-supervised SVM for speech emotion classification can be explored further. Also to look forward to obtain more meaningful features that reflect the emotion contained in the speech so as to arrive on optimal representation so as to achieve better performance in emotion recognition from speech.

### References

[1] Z Zhang, F Ringeval 2017 Enhanced semi-supervised learning for multimodal emotion recognition IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Shanghai pp. 5185-5189

[2] J Liu, C Chen, J Bu, M You and J Tao 2017 Speech Emotion Recognition using an Enhanced Co-Training Algorithm IEEE International Conference on Multimedia and Expo Beijing pp. 999-1002

[3] J J Huang, Y Li, J Tao, Z Lian, M Niu and J Yi 2018 Speech Emotion Recognition Using Semi-supervised Learning with Ladder Networks First Asian Conference on Active Computing and Intelligent Interaction (ACIIAsia)Beijingpp. 1 – 5

[4] J Deng, X Xu, Z Zhang, S Fruhholz and B Schuller 2018 Semisupervised Autoencoders for Speech Emotion Recognition IEEE/ACM Transactions on Audio, Speech, and Language Processing vol. 26

[5] Parthasarathy and C Busso 2020 Semi-Supervised Speech Emotion Recognition With Ladder Networks IEEE/ACM Transactions on Audio, Speech, and Language Processing vol. 28 pp. 2697-2709

[6] S Basu, J Chakraborty, A Bag and M Aftabuddin 2017 A review on emotion recognition using speech International Conference on Inventive Communication and Computational Technologies (ICICCT) Coimbatore pp. 109-114
[7] Chih-Chung Chang and Chih-Jen Lin 2011 LIBSVM : A library for support vector machines ACM Transactions on Intelligent Systems and Technology, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[8] Zhu, Xiaojin and Zoubin Ghahramani 2003 Learning from labeled and unlabeled data with label propagation Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003) Washington DC

[9] Burkhardt, Felix Paeschke, Astrid Rolles, M Sendlmeier, Walter Weiss, Benjamin 2005 Database of German emotional speech 9th European Conference on Speech Communication and Technology pp.1517-1520

[10] Ananthi and Dhanalakshmi 2014 SVM and HMM Modeling Techniques for Speech Recognition Using LPCC and MFCC Features Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)

[11] P Shi 2018 Speech emotion recognition based on deep belief network IEEE 15th International Conference on Networking, Sensing and Control (ICNSC) Zhuhai pp. 1-5

[12] Leila Kerkeni, Yousef Serrestou, Mohamed Mbarki, Kosai Raoof and Mohamed Ali Mahjoub 2018 Speech Emotion Recognition: Methods and Cases Study Proc. International Conference on Agents and Artificial Intelligence (ICAART) Vol 2 pages 175-182

[13] B S Shajee Mohan and C C Sekhar 2017 Distance Metric Learnt Kernel based SVMs for Semi-Supervised Pattern Classification Ninth International Conference on Advances in Pattern Recognition (ICAPR) Bangalore India pp. 1-6

[14] Shajee Mohan B S 2018 Distance Metric Learning based Kernels for Some Pattern Analysis Tasks [Ph.D thesis Dept. of Computer science and engineering] Indian Institute of Technology Madras