Speech emotion recognition using kernel sparse representation based classifier

Pulkit Sharma, Vinayak Abrol, Abhijeet Sachdev, A. D. Dileep, Anil Kumar Sao
IIT Mandi, India
{pulkit_s,vinayak_abrol, abhijeet_sachdev}@students.iitmandi.ac.in, {addileep,anil}@iitmandi.ac.in

Abstract—**In this paper, we propose to use a kernel sparse representation based classifier (KSRC) for the task of speech emotion recognition. Further, the recognition performance using the KSRC is improved by imposing a group sparsity constraint. The speech utterances with same emotion may have different duration, but the frame sequence information does not play a crucial role in this task. Hence, in this work, we propose to use dynamic kernels which explicitly models the variability in duration of speech signals. Experimental results demonstrate that, given a suitable kernel, KSRC with group sparsity constraint performs better as compared to the state-of-the-art support vector machines (SVM) based classifiers.**

Index Terms—**Kernel sparse representations, group sparsity, speech emotion recognition.**

I. Introduction

Automatic emotion recognition from speech, is an active research subject in the speech processing area where the goal is to classify the input utterance into various categories of emotion, e.g. neutral, happiness, anger, fear, disgust, and sadness [1]–[5]. During the past decades, many classification approaches based on the Gaussian mixture model (GMM), the hidden Markov model, the support vector machine (SVM) etc. have been successfully applied to speech emotion recognition [6], [7].

Existing works in literature had applied sparse representation based signal processing for various signal processing and pattern classification tasks [8]–[15]. For instance, work in [11], [14] used SR based features for tasks in speech recognition, [10] proposed a greedy dictionary for efficient speech signal reconstruction using SR, [13] proposed to use CS with LP based dictionary for voiced/nonvoiced detection and [12] used SR for face recognition using sparse representation-based classifier (SRC). In SRC, for a testing example, a sparse vector is computed using a dictionary. Here, dictionary used can be single, where all the training examples of all the classes are used as basis, or class specific, where individual dictionaries consisting of training examples of individual class, are used [12]. In case of a single dictionary, the weights (in the sparse vector) corresponding to the true class have high amplitude as compared to weights corresponding to rest of the classes which is used to find the identity of test example. In case of multiple dictionaries, the test example is classified to the class giving minimum reconstruction error. Although SRC performed better than some of the existing pattern classification methods, its classification ability degrades with data having the same directional distribution [16], [17]. This means the efficiency of SRC decreases if vector direction of a test sample is same as that of training samples belonging to two or more classes. To address this problem, kernel based methods are proposed, where a nonlinear mapping is used to map input data space to a high dimensional kernel feature space. Kernel trick allows the use of linear classification methods (in high dimensional space) to the corresponding nonlinear data (in input space). This kernel trick is previously applied in support vector machines (SVM) [18], [19], along with principal component analysis, Fisher discriminant analysis and sparse representation known as kernel principal component analysis (KPCA) [20], kernel Fisher discriminant analysis (KFDA) [21] and kernel sparse representation based classifier (KSRC) [16], respectively.

In this paper, a novel speech emotion recognition method using KSRC is proposed, where its effectiveness is studied using various kernels. In addition, we propose to use group sparsity constraint in KSRC, which improves the performance by estimating more discriminative and accurate weights. This is achieved by considering the cooperation among training samples of same class while estimating the sparse vector. Static kernels can't be used for our task as speech utterances are represented as varying length sets of feature vectors [22]. Hence dynamic kernels namely, Gaussian mixture model-based intermediate matching kernel (GMM-IMK) [22] and example-specific density based matching kernel (ESDMK) [23] are employed to address the issue of classifying varying length sequential emotion pattern.

The organization of the paper is as follows : Section II describes the proposed KSRC technique for speech emotion classification. Kernels used in this work are described in section III. Experimental results are discussed in section IV and finally the paper is concluded in section V.

II. Kernel sparse representation based classifier

Kernel sparse representation based classifier (KSRC) is a nonlinear extension of SRC [16]. Let us assume a k-class classification task. Consider the training set be $\{\mathbf{a}_i, b_i\}_{i=1}^{m}$, where $\mathbf{a}_i \in X \subset \mathbb{R}^N$, $b_i \in 1, 2, ..., k$ and m is total number of training examples. $\mathbf{a}_i$ represents the training sample of dimension N in input space X and b_i is the corresponding class label.

During testing, the goal is to assign label b to an arbitrary sample $\mathbf{a}$ in the input space X. In KSRC, a nonlinear mapping function $\Phi : \mathbb{R}^N \rightarrow \mathcal{S}$ is used to make nonlinearly separable

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

training samples separable in higher dimensional space such that $\Phi(\mathbf{a}) = [\phi_i(\mathbf{a}), \phi_2(\mathbf{a}), ..., \phi_C(\mathbf{a})]^T$, where $\Phi(\mathbf{a}) \in \mathbb{R}^C (C \gg N)$ is transformation of $\mathbf{a}$ into a higher dimensional space.

Similar to SRC, KSRC also uses sparse representation of signals, obtained using linear combination of class specific training data (after transformation to space $\mathcal{S}$) for signal reconstruction, and then classifies the test signal to the class giving minimum reconstruction error. This linear model can be expressed as [16]:

$$\Phi(\mathbf{a}) = \sum_{i=1}^{m} \Phi(\mathbf{a}_i)\beta_i = \boldsymbol{\Phi}\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_m]^T$ is the coefficient vector corresponding to $\boldsymbol{\Phi} = [\Phi(\mathbf{a}_1), \Phi(\mathbf{a}_2), ..., \Phi(\mathbf{a}_m)] \in \mathbb{R}^{C\times m}$ and β_i are the coefficients corresponding to $\Phi(\mathbf{a}_i)$. The problem of finding an estimate of $\boldsymbol{\beta}$ is formulated as:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ ||\boldsymbol{\beta}||_1 \quad \text{s.t.} \quad ||\Phi(\mathbf{a}) - \boldsymbol{\Phi}\boldsymbol{\beta}||_2^2 \leq \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon}$ is a small error term. The computational complexity of equation (2) is huge since the dimensionality of transformed feature space is large than that of input space ($C \gg N$). This issue can be addressed by using various dimensionality reduction techniques e.g., principal component analysis (PCA) [16], [20], [21].

One way to achieve this is to project the data from space $\mathcal{S}$ to a low dimensional subspace using a transformation matrix $\mathbf{M}^{C\times l}(l \ll C)$. Using $\mathbf{M}$, equation (1) can be modified as :

$$\mathbf{M}^T\Phi(\mathbf{a}) = \mathbf{M}^T\boldsymbol{\Phi}\boldsymbol{\beta}. \tag{3}$$

Here, each column M_j of $\mathbf{M} = [M_1, M_2, ..., M_l]$ is a linear combination of all the training signals in space $\mathcal{S}$, i.e.,

$$\mathbf{M} = \boldsymbol{\Phi}\boldsymbol{\Upsilon}, \tag{4}$$

where $\boldsymbol{\Upsilon} = [\gamma_1, \gamma_2, ..., \gamma_l]$ and $\boldsymbol{\gamma}_j = [\gamma_{j,1}, \gamma_{j,2}, ..., \gamma_{j,m}]^T$ is the pseudo-transformation vector corresponding to j^{th} transformation vector M_j. However, in most of the cases the transformation $\boldsymbol{\Phi}$ is not known, and the optimization of problem in equation (3) is infeasible using traditional methods. This issue can be addressed by using a kernel function $k(.,.)$, which avoids the explicit mapping of training signals to space $\mathcal{S}$, and helps in solving the problem in original space. Hence using equation (4), equation (3) can be rewritten as:

$$\boldsymbol{\Upsilon}^T\mathbf{k}(.,\mathbf{a}) = \boldsymbol{\Upsilon}^T\mathbf{K}\boldsymbol{\beta}, \tag{5}$$

where $\mathbf{K} = \boldsymbol{\Phi}^T\boldsymbol{\Phi} \in \mathbb{R}^{m\times m}$ is the kernel Gram matrix such that $K_{i,j} = k(\mathbf{a}_i, \mathbf{a}_j)$ and $\mathbf{k}(.,\mathbf{a}) = [k(\mathbf{a}_1, \mathbf{a}), k(\mathbf{a}_2, \mathbf{a}), ..., k(\mathbf{a}_m, \mathbf{a})]^T = \boldsymbol{\Phi}^T\Phi(\mathbf{a})$. Here some popular kernels e.g., linear, Gaussian and polynomial can be employed. The pseudo-transformation matrix $\boldsymbol{\Upsilon}$ can be obtained using both KPCA and KFDA as described in [16]. Thus the optimization problem in equation (2) can be replaced by a feasible optimization problem as :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ ||\boldsymbol{\beta}||_1 \quad \text{s.t.} \quad ||\boldsymbol{\Upsilon}^T\mathbf{k}(.,\mathbf{a}) - \boldsymbol{\Upsilon}^T\mathbf{K}\boldsymbol{\beta}||_2^2 \leq \boldsymbol{\epsilon}, \tag{6}$$

A. KSRC with group sparsity constraint

In KSRC, a linear combination is obtained with only a few non zero entries of dictionary (kernel Gram matrix) atoms. This is because the sparsity of linear coefficients is controlled using only l_1-norm regularization. In l_1-norm, all the training samples are treated equally and the cooperation of training samples from the same class is not considered [24]. A single sample from a group of correlated training samples is selected during l_1-norm and thus leads to misclassification when a test example has a similar training sample in different classes [24].

However, some additional structure (in the dictionary atoms belonging to same class) can be expected in the support of sparse representations. In order to find this structure in support of sparse representations, the group sparse classifier is proposed in [24]. The group sparsity constraint employed uses l_1-norm mixed l_2-norm. This regularization results in dense representations among the coefficients belonging to the same class but sparse representations among classes. These group sparse representations can be obtained by modifying the optimization problem in KSRC (equation (6)) as:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ \lambda \sum_{i=1}^{m} ||\boldsymbol{\beta}_{G_i}||_2 \ + \ ||\boldsymbol{\Upsilon}^T\mathbf{k}(.,\mathbf{a}) - \boldsymbol{\Upsilon}^T\mathbf{K}\boldsymbol{\beta}||_2^2, \tag{7}$$

where λ is a constant and the kernel Gram matrix $\mathbf{K}$ is partitioned into m disjoint groups (belonging to examples of m classes) $G_1, G_2, G_3, \ldots, G_m$ and $\boldsymbol{\beta}_{G_i}$ represents the group of weights corresponding to the group G_i. Term $\lambda \sum_{i=1}^{m} ||\boldsymbol{\beta}_{G_i}||_2$ in equation (7) can be viewed as a combination of both l_1 and l_2 norms. Weights within each group are obtained using the l_2 norm, whereas the results between groups are summed using l_1 norm. Group sparsity constraint will enable us to obtain more accurate and robust weights in each group along with the benefits of the sparsity.

Estimate of sparse vector $\boldsymbol{\beta}$ can be obtained either solving equation (6) or (7), which is then used to classify a test example $\mathbf{a}$ based on minimum reconstruction error computed as $min\ ||\boldsymbol{\Upsilon}^T\mathbf{k}(.,\mathbf{a}) - \boldsymbol{\Upsilon}^T\mathbf{K}\boldsymbol{\alpha}_i||_2$, where for a class i a characteristic function α_i is defined such that :

$$\alpha_i(\beta_j) = \begin{cases} \beta_j, & \text{if } b_j = i \\ 0, & otherwise \end{cases} \tag{8}$$

Pseudo code of the proposed classification method is explained in Algorithm 1.

III. Dynamic kernels for speech emotion recognition

The duration of the speech signal varies from one utterance to another and hence, the number of frames also differs from one utterance to another. In the tasks such as speaker identification, spoken language identification, and speech emotion recognition, the duration of the data is long and preserving sequence information is not critical. However, since the duration of the speech utterances (with different/same emotion) is varying, thus dynamic kernels[1] are used in our work. Various

[1] Dynamic kernels are used for sets of varying length patterns.

Algorithm 1 KSRC algorithm for emotion speech classification.

```
Inputs: (i) Set of training examples $\{\mathbf{a}_i, b_i\}_{i=1}^m$, where $\mathbf{a}_i \in \mathbb{R}^N$, $b_i \in 1, 2, ..., k$ emotion classes.
(ii) Test example $\mathbf{a} \in \mathbb{R}^N$.
Outputs: class label $b$ corresponding to $\mathbf{a}$.
Select a kernel $k(.,.)$ and compute kernel Gram matrix $\mathbf{K}$ and vector $\mathbf{k}(., \mathbf{a})$.
Obtain the pseudo-transformation matrix $\boldsymbol{\Upsilon}$ and normalize columns of $\boldsymbol{\Upsilon}^T\mathbf{K}$ and $\boldsymbol{\Upsilon}^T\mathbf{k}(., \mathbf{a})$ to unit norm.
Solve equation (6) or (7) to obtain estimate of sparse vector $\hat{\boldsymbol{\beta}}$.
Compute residuals for all classes ($k$)
    $\rho_i(\mathbf{a}) = \|\boldsymbol{\Upsilon}^T\mathbf{k}(., \mathbf{a}) - \boldsymbol{\Upsilon}^T\mathbf{K}\alpha_i\|_2 \;, \;\; i = 1, 2, ...k$
Label $b$ for test example $\mathbf{a}$ is obtained as :
    $\hat{b} = \underset{i=1,...,k}{\operatorname{argmin}} \rho_i(\mathbf{a})$
```

kernels used for KSRC in this work are Gaussian mixture model-based intermediate matching kernel (GMM-IMK) [22] and example-specific density based matching kernel (ESDMK) [23]. The SVM-based classifier using ESDMK is shown to perform better for speech emotion recognition [23].

A. GMM-IMK

GMM-IMK uses components of class-independent Gaussian mixture model (CIGMM) as a representation for the set of virtual feature vectors [22]. For every component of CIGMM, a local feature vector each from the two sets of feature vectors that has the highest probability of belonging to that component is selected. These selected local feature vectors are then used to compute a base kernel. The sum of all the base kernels corresponding to different components of CIGMM is computed to obtain the IMK. The base kernel used in GMM-IMK is Gaussian kernel.

B. ESDMK

ESDMK is computed between the pair of examples, represented as sets of feature vectors, by matching the estimates of the example-specific densities computed at every feature vector in those two examples. The number of feature vectors of an example among the K nearest neighbors of a feature vector is considered as an estimate of the example-specific density. The minimum of the estimates of two example specific densities, one for each example, at a feature vector is considered as the matching score. The ESDMK is then computed as the sum of the matching scores computed at every feature vector in a pair of examples [23].

IV. Experimental details and datasets used for speech emotion recognition

In this section, effectiveness of the KSRC (with and without group sparsity constraint) is studied for speech emotion recognition task. A speech utterance is represented using a 39-dimensional feature vector derived on a frame by frame basis. The first 12 features of this 39 dimensional feature are the Mel frequency cepstral coefficients [25] and the 13th feature corresponds to the log energy. The delta and acceleration coefficients constitutes the remaining 26 features. For a given speech utterance the features are extracted using a frame size of 20 ms with a 10 ms shift. We compare the performance of proposed KSRC with the standard SVM-based classifiers, the maximum likelihood GMM-based system, the large margin GMM-based system and the adapted GMM-based systems. Two databases used for studies on speech emotion recognition are: (i) The Berlin emotional speech database (EMO-DB) [1], [26], and (ii) The German FAU Aibo emotion corpus (FAU-AEC) [3], [27]. The details of these datasets are given as:

1) EMO-DB: EMO-DB dataset consists of a total of 494 utterances corresponding to the following seven emotional categories: disgust (38), sadness (53), fear (55), happiness (64), boredom (79), neutral (78) and anger (127) (the number of utterances for each category are given in parentheses) [1], [26]. These speech utterances correspond to ten sentences in German language uttered by five female and five male speakers (actors). In this work, 80% of the utterances are used for training and the remaining 20% are used for testing. The speech emotion recognition accuracies presented in this work for the EMO-DB dataset are the average classification accuracies along with 95% confidence interval obtained for a 5-fold stratified cross-validation.

2) FAU-AEC: In the FAU-AEC dataset we have considered four super classes of emotions: (i) anger, (ii) emphatic, (iii) neutral, and (iv) motherese. An almost balanced subset of the corpus defined for these four classes by CEICES of the Network of Excellence HUMAINE funded by the European Union is used in this work [3], [27]. The classification is performed at the chunk (speech utterance) level in the Aibo chunk set. The speaker-independent speech emotion recognition accuracies presented in this paper for the FAU-AEC dataset is the average classification accuracies along with 95% confidence interval obtained for 3-fold stratified cross validation.

Classification accuracy for speech emotion recognition (SER) obtained using the proposed methods along with its comparison to the GMM-based classifiers and SVM-based classifiers with the state-of-the-art dynamic kernels is presented in Table I. Proposed classification results are also compared with SRC [12], incomplete sparse least square regression (ISLSR) [28] and SVM classifier using convolutional neural networks based features (SVMCNN) [29]. For SVMCNN, results reported are the best case results where same speaker data is used for both training and testing [29]. In this study, parameters of the GMMs are estimated using the maximum likelihood (ML) method (MLGMM), and the parameters of the UBM or CIGMM are adapted to the data of a class (adapted GMM) [30] are considered to build GMM-based classifiers. The accuracies presented in Table I are the best accuracies observed among the GMM-based classifiers and SVM-based classifiers with state-of-the-art dynamic kernels namely GMM based intermediate matching kernel (GMM-IMK) and example-specific density based matching kernel (ESDMK) [22], [23]. These classification results indicate that KSRC with group sparsity constraint not only outperforms KSRC, but also state-of-the-art SVM based classifiers, consistently.

Classification model		SER	
		EMO-DB	FAU-AEC
		$CA95\%CI$	$CA95\%CI$
MLGMM		66.81±0.44	60.00±0.13
Adapted GMM		79.48±0.31	61.09±0.12
SRC		67.25±0.16	47.64±0.27
ISLSR		78.03±0.14	60.50±0.18
SVMCNN		93.70±0.28	64.09±0.16
SVM using	GMM-IMK	85.62±0.29	62.48±0.07
	ESDMK	92.00±0.27	65.33±0.09
KSRC using	GMM-IMK	84.38±0.32	60.43±0.51
	ESDMK	90.17±0.47	64.08±0.16
$KSRC_G$ using	GMM-IMK	87.07±0.37	63.75±0.49
	ESDMK	**94.54±0.58**	**66.72±0.18**

TABLE I: Comparison of classification accuracies (CA) (in %) of the proposed KSRC and KSRC with group sparsity constraint (labeled as $KSRC_G$) with GMM-based classifiers and SVM-based classifiers using GMM-IMK and ESDMK for speech emotion recognition (SER). Here, $CA95\%CI$ indicates average classification accuracies along with 95% confidence interval.

V. Conclusions

In this paper, we proposed the application of KSRC for speech emotion recognition using dynamic kernels. Further a group sparsity constraint is employed to improve the classification performance of KSRC. This improvement in performance is attributed to the efficient estimation of sparse vector as all the examples of a group (class) are used to model the test signal. Dynamic kernels are used to model the varying duration of different speech utterances. Experimental results in this work confirm that the proposed speech emotion classification method outperforms the existing state-of-the-art SVM-based classifiers.

References

[1] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 emotion challenge,” in *INTERSPEECH*, pp. 312–315.

[2] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, “Whodunnit Searching for the most important feature types signalling emotion-related user states in speech,” *Computer Speech Language*, vol. 25, no. 1, pp. 4 – 28, 2011.

[3] A. Batliner, S. Steidl, and E. Nth, “Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus,” in *International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC)*, 2008, pp. 28–31.

[4] M. E. 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.

[5] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication*, vol. 53, no. 910, pp. 1062 – 1087, 2011.

[6] W. Zheng, M. Xin, X. Wang, and B. Wang, “A novel speech emotion recognition method via incomplete sparse least square regression,” *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014.

[7] Y. Jin, P. Song, W. Zheng, and L. Zhao, “A feature selection and feature fusion combination method for speaker-independent speech emotion recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4808–4812.

[8] V. Abrol, P. Sharma, and A. K. Sao, “Greedy dictionary learning for kernel sparse representation based classifier,” *Pattern Recognition Letters*, vol. 78, pp. 64 – 69, 2016.

[9] P. Sharma, V. Abrol, and A. K. Sao, “Compressed sensing for unit selection based speech synthesis,” in *European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1731–1735.

[10] M. Jafari and M. Plumbley, “Fast dictionary learning for sparse representations of speech signals,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1025–1031, Sept 2011.

[11] P. Sharma, V. Abrol, A. D. Dileep, and A. K. Sao, “Sparse coding based features for speech units classification,” in *INTERSPEECH*, Dresden, Germany, 2015, pp. 712–715.

[12] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.

[13] V. Abrol, P. Sharma, and A. K.Sao, “Voiced/nonvoiced detection in compressively sensed speech signals,” *Speech Communication*, vol. 72, pp. 194 – 207, 2015.

[14] T. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, “Exemplar-Based Sparse Representation Features: From TIMIT to LVCSR,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2598–2613, Nov 2011.

[15] V. Abrol, P. Sharma, and A. K. Sao, “Speech Enhancement Using Compressed Sensing,” in *INTERSPEECH*, Lyon, France, August 2013, pp. 3274–3278.

[16] L. Zhang, W.-D. Zhou, P.-C. Chang, J. Liu, Z. Yan, T. Wang, and F.-Z. Li, “Kernel sparse representation-based classifier,” *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1684–1695, April 2012.

[17] B. Wang, W. Li, N. Poh, and Q. Liao, “Kernel collaborative representation-based classifier for face recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 2877–2881.

[18] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[19] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[20] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[21] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, “Fisher discriminant analysis with kernels,” in *IEEE International Workshop on Neural Network and Signal Processing*, August 1999, pp. 41–48.

[22] A. Dileep and C. Sekhar, “GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1421–1432, 2014.

[23] A. Sachdev, A. D. Dileep, and V. Thenkanidiyoor, “Example-specific density based matching kernel for classification of varying length patterns of speech using support vector machines,” in *International Conference on Neural Information Processing (ICONIP)*, November 2015, pp. 177–184.

[24] A. Majumdar and R. Ward, “Classification via group sparsity promoting regularization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 861–864.

[25] B. Logan *et al.*, “Mel frequency cepstral coefficients for music modeling.” in *ISMIR*, 2000.

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech.” in *INTERSPEECH*, vol. 5, 2005, pp. 1517–1520.

[27] S. Steidl, *Automatic classification of emotion related user states in spontaneous children’s speech*. University of Erlangen-Nuremberg Germany, 2009.

[28] W. Zheng, M. Xin, X. Wang, and B. Wang, “A novel speech emotion recognition method via incomplete sparse least square regression,” *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 569–572, 2014.

[29] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[30] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.