RESEARCH REVIEW JOURNALS
Volume-04
Issue-05
May-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[UGC Listed Journal]

Consonant-Vowel Recognition in Emotional Environment using Trajectory based Stochastic Feature Mapping

[1]Sushil Kumar Gupta, [2]Siddharth Srivastava & [3]Dr. Jainath Yadav

[1]*Assistant Professor, Faculty of Engineering,Lucknow University(India)*
[2]*Assistant Professor,Goel Institute of Technology, AKTU Lucknow (India)*
[3]*Assistant Professor,Department of Computer Science, Central University of South Bihar(India)*

ARTICLE DETAILS

Article History
Published Online: 25 May 2019

Keywords
Consonant-Vowel (CV), Mel Frequency Cepstrum coefficient (MFCC), Hidden Markov Model (HMM), Support Vector Machine (SVM), Trajectory based Stochastic Feature Mapping (TSFM).

***Corresponding Author**
Email:sushil.mpec[at]gmail.com

ABSTRACT

The characteristics of consonant-vowel (CV) units differ from one emotion to other emotions. Therefore, the existing CV recognition systems fail to recognize CV units in the emotional environments. The effective way of conveying messages by human beings is by expressing their emotions during conversations effectively. Therefore, in this regard, we propose the CV recognition method in the emotional environments, adaptable to varying emotional moods of the speakers. CV recognition system has been explored to transform from emotional MFCC features to neutral MFCC features. We have proposed method for increasing the accuracy of consonant-vowel (CV) in Indian languages for the emotional speech. In this work, we have developed a mapping method based on MFCC feature transformation framework for developing CV recognition system in the emotion environments. In the proposed method, we are using trajectory based stochastic feature mapping (TSFM) method which is used to map emotional MFCC (Mel Frequency Cepstrum coefficient) features to neutral MFCC features. In the proposed method, we have recognized consonant-vowel in two stages. In the first stage, we have recognized vowels using HMM, while in the second stage, consonants are recognized using Support Vector Machines (SVM). After, normalized performance scores from HMM and SVM are merged for CV recognition, the average performance of (HMM+SVM) is increased by using TSFM from 55.84% to 63.1% for female speaker and from 53.31% to 61.47% for male speaker in three emotions (anger, sadness and neutral) for CV units.

1. Introduction

Robust speech recognition systems in the emotional environments have attained more attraction in the present time in order to implement in CV recognition. Emotional environments are better described as the environmental speech that is produced by the speakers under the impact of emotional states such as sadness, anger and neutral. The main motivation comes from the eagerness to build a machine that is more flexible and susceptible to a speaker identity in the emotional environments. The main work of capable human-machine interaction is to empower a computer with the intuitive computing ability so that a computer can recognize the identity of the users in such environments for many different applications. The CV unit is an utterance of the speech which is varied according to the emotions. The Indian languages contain more number of CV units [1]. The main problem in the recognition of CV unit under emotional environments is due to the fact that it contains more number of classes and similarity among the CV units.

In the literature, there are three models used to recognized CV units; Support Vector Machines (SVM), Hidden Markov models (HMM) and Multi-Layer Feed Forward Neural Network (MLFFNN) [2, 15]. From the literature survey, it is observed that SVM models and MLFFNN models are more effective for recognition of CV units compared to commonly used HMM models. However, if there is more number of classes, then HMM is working better than the MLFFNN. SVM gives better performance for consonant regions and HMM gives better performance for vowel regions of CV units [3]. This is due to HMMs are fine at accrue the sequence of vocal tract shapes corresponds to the state sequence. For each vowel, sequences of vocal tract shapes are unique. But, in case of emotional CV unit the variations of vocal tract shapes vary in different emotions. Due to this reason, HMM and SVM doesn't give better performance in presence of emotional CV units. Therefore, we proposed a trajectory based stochastic feature mapping (TSFM) method which is capable of mapping emotional MFCC features to neutral MFCC features of consonant-vowel units by using joint-probability and minimum mean square error method. This method minimizes the variation of CV unit in different emotions by transforming it into corresponding neutral CV unit. Therefore, proposed method increases the performance and robustness of the model. We have used two level approaches for CV recognition in the emotion environments for the Indian language. In the first level, vowel sections of the CV unit classes are recognized, and in the second level consonant sections are recognized. After the recognized CV unit class, normalized scores from HMM and SVM models are merged for enhancing the recognition accuracy in the CV unit forthe emotional environments. HMMs and SVMs are built using two different learning approaches. HMM uses maximum likelihood approach and SVM usesdiscriminant learning approach. From the different type of learning approaches, it is provided with more accuracy for large number of CV unit classes in the emotional environments.

This paper is organized as follows. The emotional speech corpus of CV unit has been described in section II. The Details of the proposed method has been given in section III. Section IV elaborates proposed TSFM method. Experimental setup used for CV recognition in the emotion environmentisexplained

in section V. Section VI contains the summary and conclusions of this paper.

2. Emotional Speech Corpus

We have collected the speech data form All India Radio, Varanasi. One female and one male speaker were selected for recording emotional speech. Speakers were given instruction to read sentences in neutral style. For recording the emotions, different sentences have been used for both male and female in three distinct emotions (anger, neutral and sadness). We have recorded 2000 sentences for the selected male and female. Each sentence contains 7-10 different words. The total duration of the database is around 10 hours. The speech signal was sampled at 16 kHz and represented as 16-bit numbers. The total number of emotion CV unit that have been extracted from the continuous speech utterances are 12,367. Each CV unit is classified according to their vowel classes /a/, /e/, /i/, /o/ and/u/ for both male and female in all the three emotions. The vowel categories /a/, /e/, /i/, /o/ and/u/contain 35, 17, 28, 9and 18 consonant classes respectively. Table 1 depicts all the CV units corresponding to each subgroup in only one emotion. Same classes, subclasses and CV units have been used in all the three emotions for both male and female speakers. All the CV unit classes are shown in the Table 1.

TABLE I: List of 104 CV units

Subgroup	CV units
/a/	/a/, /aai/, /aap/, /ak/, /an/ /ba/, /baa/, /bha/, /cha/, /da/, /dha/, /ga/, /gaa/ /gha/, /ha/, /da/, /haa/, /kha/, /la/, /ma/, /na/, /naa/ /pa/, /pha/, /pra/, /ra/, /naa/, /sa/, /sha/ /ta/, /tha/, /va/, /vaa/, /ya/, /yaa/
/e/	/be/, /che/, /de/, /e/, /ek/, /ge/, /he/ /jhe/, /ke/, /me/, /ne/, /re/, /se/ /te/, /the/, /ve/, /ye/
/i/	/bhi/, /bi/, /chi/, /di/, /hai/, /hi/ /i/, /ii/, /is/, /jii/, /khi/, /ki/, /kii/, /li/ /mi/, /ni/, /phi/, /pi/, /rii/, /shi/ /si/, /thi/, /ti/, /vii/, /yi/
/o/	/bo/, /do/, /ho/, /jo/, /kho/ /o/, /or/, /ro/, /yo/
/u/	/bu/, /chu/, /du/, /ghu/, /gu/, /hu/ /ju/, /ku/, /mu/, /pu/, /puu/ /ru/, /ruu/, /su/, /u/, /un/, /us/, /yuu/

3. Proposed Method

In the proposed method, 104 CV units are classified into five subclasses according to vowel class (/a/, /e/, /i/, /o/ and/u/). In the first step, we calculate MFCC feature vectors for both emotional and neutral CV unit. MFCC feature vectors are calculated for three emotions for both male and female speakers. We provided emotional feature vectors as the input to the trajectory based stochastic feature mapping (TSFM) method. The proposed method is used to map the emotional MFCC feature vectors to the corresponding neutral MFCC feature vectors using joint probability density and minimum mean square error method. TSFM method is described in section IV. This method is applied for each vowel class (/a/, /e/, /i/, /o/ and/u/), in all three emotions (anger, sadness and neutral) for mapping emotional MFCC feature vectors to the neutral MFCC feature vectors. Using these neutral MFCC feature vectors, we first build the HMM model using HTK toolkit. Vowel section of the emotional unit is recognized by using HMM model in two parts. In the first part, we trained the HMM. In the second part, we tested the HMM trained model using half of CV units which are used for the training the HMM. SVM model is built using SVMTORCH. Consonant parts of the neutral CV units are recognized by using SVM. SVM model also proceeds in two parts. In the first part, we trained the model and in the second part, we tested the model. After building both the HMM and SVM models, we combined their scores.

In the proposed method, the obtained normalized scores form HMM and SVM are combined as follows:

Class (A) = max (Z_1*S(A) + Z_2*H(A))

Where S(A) is the score obtained from SVM and H(A) is the score obtained from HMM. Z_1 weight given for SVM and Z_2 is the weight given for HMM. In our study, Z_1 is varied from 0to 1; and Z_2 = 1- Z_1

A. Hidden Markov Model (HMM)

Statistical pattern recognition approach for speech recognition is mainly based on HMMs [4]. HMM is a statistical model in which the system being modeled is assumed to be Markov process with hidden states and unknown parameters. Elements of state of HMM model are given below:

1. N, is denoted as the number of state in model q = q_1, q_2, q_3…..q_N
2. M, is number of distinct observation symbols in each state,
o = o_1; o_2; o_3…….o_M

3. Transition probability distribution state is given by B =b_{ij}
Where b_{ij} = P [S_{t+1} = q_j | S_t = q_i]; 1≤i; j≤ N

4. The observation symbol probability distribution, A = $a_j(k)$ where
$a_j(k)$ = P [o_t = oj | s_t = q_j]; 1 ≤I, j ≤ N; 1≤k≤M
defines symbol distribution in state j.

5. Starting state distribution ∏=$∏_j$
$∏_j$ = P [s_1 = q_i], 1≤i≤N

HMM is indicated by the notation λ = (B, A, $∏_j$)

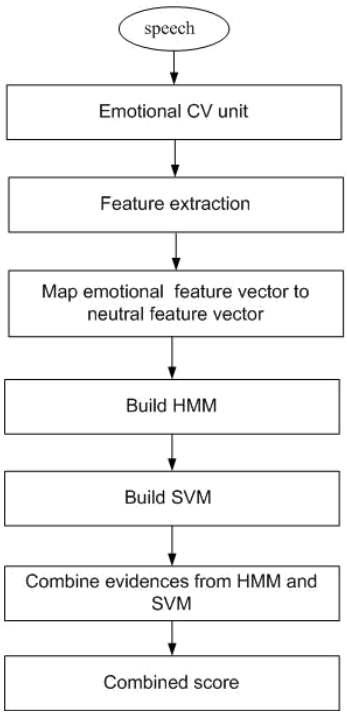

Fig. 1: Flow diagram of CV recognition model using proposed Trajectory based Stochastic Feature Mapping (TSFM) method.

B. Support Vector Machines (SVM)

The main idea of a SVM is to construct a hyper plane as the decision surface in such a way that separationbetween positive and negative examples (margin) is maximized [5], [6]. The support vectors comprise a small subgroup of the training data obtained using the support vectors learning method. The division into the hyper plane and the nearest data point is called the margin of separation. The goal of SVM is to determine an accurate hyper plane.

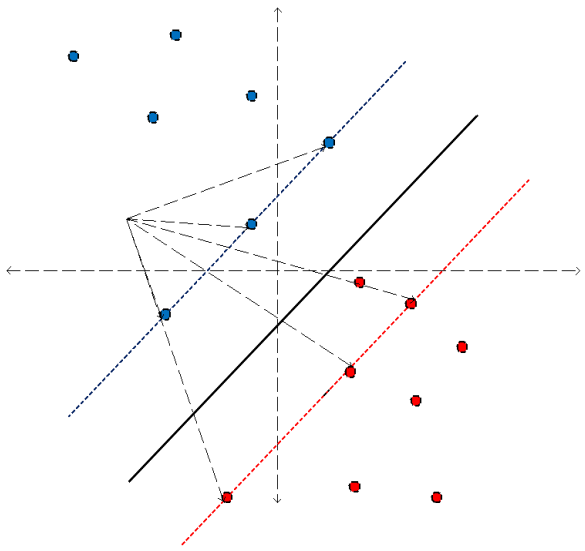

Fig. 2: Graphical representation of SVM

4. Proposed Trajectory Based Stochastic Feature Mapping (TSFM) For Emotional CV Unit

Stochastic feature compensation (SFC) techniques have been used for robust speech recognition tasks. These techniques do not assume any form of mathematical structure for emotionaleffect. The effect of emotions may be represented as additive terms to the mean vectors and covariance matrices of the neutral speech GMM. y_t is the given emotional test feature vector, $\hat{x}_t$ neutral vector is estimated based on a minimum mean squared error using following equation:

$$\hat{x}_t = E[x \mid y_t] = \int_x xp(x|yt)dx \quad (1)$$

Neutral feature vectors are represented by using x, where x is a random variable and the conditional probability distribution function (pdf) $p(x|y_t)$ of x given y_t. There are mainly two approaches for feature compensation, joint probability modeling and independent probability modeling. The independent probability modeling methods makes independent GMMs for neutral and emotional data. Additive term to the mean vectors and covariance matrices of the GMMs are used for representing the outcome of the emotion. The conditional pdf is obtained according to arithmetic approximations using the additive terms. Three standard independent probability model based SFC techniques are multivariate Gaussian-basedcepstral normalization (RATZ), stereo piece-wise linear compensation for environment (SPLICE) [7] and multivariate model based cepstral normalization (MMCN) [8]–[10]. One of the drawbacks of this approach is that the additive terms which may produce false effect in the emotional environments. Joint probability models assume the availability of emotional and neutral feature vectors of the stereo data. Stereo data is pair of speech utterances recorded in the neutral environments and the other corresponding data in the emotion environment. During the training, a single GMM is built from emotional and neutral feature vectors of the stereo data. During testing with the given emotional speech vectors, corresponding neutral speech vectors are estimated using conditional pdf. Joint probability modeling methods perform better than independent probability modeling methods, but they require more training data compared to independent probability modeling methods.

Figure 3 shows the independent probability model and trajectory based stochastic feature mapping (TSFM) method. Steps are given below:

1) The emotional and neutral MFCC training vectors are concatenated to obtain joint vectors (Z).
2) Single GMM is used to model the joint vectors which show the joint pdf.
3) The parameters of the joint pdf are used to obtain conditional pdf
4) Given an emotional test vector Y_t', a neutral vector X_t' is derived based on MMSE or maximum likelihood estimate (MLE).

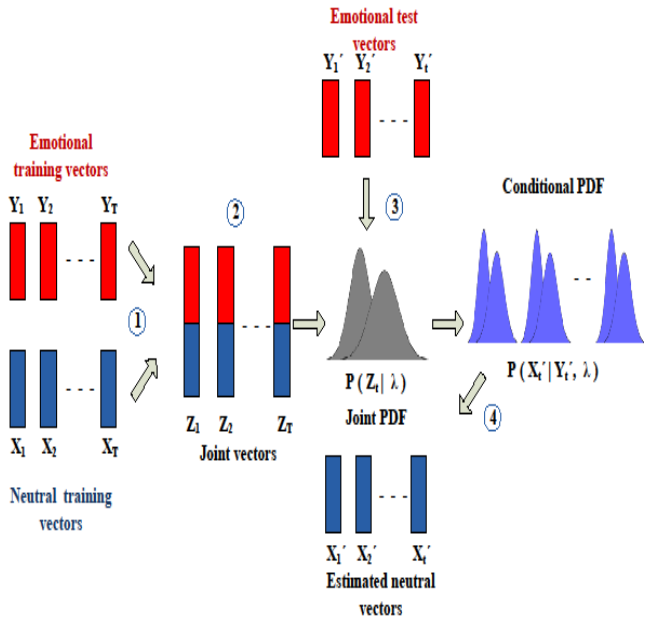

Fig. 3: The block diagram of the proposed Trajectory based Stochastic Feature Mapping (TSFM) method for emotional CV units.

The standard joint probability modeling is stereo-based stochastic mapping (SSM) method [11]. The basic idea of SSM is to train GMM for joint distribution of neutral and emotion features from stereo data. During testing, neutral feature corresponding to given emotion feature are estimated using MAP and MMSE estimators. MAP and MMSE estimators use posterior weighting based on the joint and marginal distributions, respectively. Both estimators are mixture of linear transformation weighted by component posteriors. The parameters of linear transformation are derived from the joint distribution of emotional and neutral features. A drawback of SSM based mapping is that it performs mapping frame by frame. Although, SSM based mapping works well, there still remain two major issues, the over-smoothing effect and the time-independent mapping. Trajectory based stochastic feature mapping (TSFM) [12] method addressed this problem by considering the feature correlation between frames. Instead of mapping frame by frame, entire frame of an utterance are simultaneously transformed. This approach works well for both emotion-synthesis and voice conversion applications [12]. Mathematical details of the trajectory based stochastic feature mapping (TSFM) algorithm is given below.

Emotional and neutral MFCC vector sequences, X and Yare written as

$$X = [\, X_1^T, X_2^T \ldots\ldots X_T^T, \,]^T \quad (2)$$

$$Y = [\, Y_1^T, Y_2^T \ldots\ldots Y_T^T, \,]^T \quad (3)$$

where T represents the number of frames in the sequence. X_t and Y_t are 39-dimensional (13 static MFCC + 13 delta +13 acceleration coefficients) emotional and neutral MFCC feature vectors at the t^{th} frame given as

$$X_t = [\, x_1^T, \Delta x_2^T, \Delta^2 x_T^T]^T \quad (4)$$

$$Y_t = [\, y_1^T, \Delta y_2^T, \Delta^2 y_T^T]^T \quad (5)$$

A joint vector sequence Z is constructed by concatenating the pair of emotional and neutral MFCC vector sequences X and Y given as

$$Z = [Z_1^T, Z_2^T, \ldots . Z_T^T \,]^T \quad (6)$$

Where $z_t = [y_t^T, x_t^T]^T$ is a 78- dimensional joint feature vector. Joint vector can be modeled using a GMM $\lambda^{(z)}$ as

$$p(z_t) = \sum_{j=1}^{M} \omega_z(j) \mathcal{N}(z_t; \mu_{z(j)}, \Sigma_z(j)) \quad (7)$$

Where

$$\mu_z(j) = \begin{bmatrix} \mu_y(j) \\ \mu_x(j) \end{bmatrix}, \Sigma_z(j) = \begin{bmatrix} \Sigma_{yy}(j) & \Sigma_{yx}(j) \\ \Sigma_{xy}(j) & \Sigma_{xx}(j) \end{bmatrix} \quad (8)$$

To train the GMM $\lambda^{(z)}$of the joint pdf $p(Z_t|\lambda^{(Z)})$, the stereo training data Z_t is utilized.

The transformation of emotional MFCC vector sequence to its neutral equivalent is performed by exaggerating following likelihood function

$$p(X|Y, \lambda^{(Z)}) = \sum_j p(j|Y, \lambda^{(Z)}) p(X|Y, j, \lambda^Z) \quad (9)$$

$$= \prod_{t=1}^{T} \sum_{j=1}^{M} p(j|Yt, \lambda^{(z)}), \; p(Xt|Yt, j,)\lambda^{(z)})$$

Where j = (j_1, j_2…j_T) shows the component sequence of mixture. The conditional pdf at every frame is designed asGMM. The j^{th} mixture component weight $p(j|Y_t; {}^{(Z)})$, at frame t and the j^{th} conditional probability distribution $P(X_t|Y_t, j,\lambda^{(Z)})$ are given by the following equation

$$p(j|Yt, \lambda^{(z)}) = \frac{w_j^y ; \mathcal{N}(y_t; \mu_j^y \;, \; \Sigma_j^{yy})}{\sum_{j=1}^{M} w_j^y w \;\; \mathcal{N}(y_t; \mu_j^y \;, \; \Sigma_j^{yy})} \quad (10)$$

$$p(x_t / y_t, i, \lambda^{(z)}) = \mathcal{N}(x_t; E_{j,t}^x \;\; D_j^x) \quad (11)$$

Where $E_{j,t}^X = \mu_j^X + \sum_j XY \left(\sum_j YY\right)(y_t - \mu_j^y)$ (12)

$$D_j^x = \Sigma_j^{xx} - \Sigma_j^{xy} (\Sigma_j^{yy})^{-1} \Sigma_j^{yx} \quad (13)$$

In Equations (12) and (13) we use similar notations for the conditional mean and conditional covariance matrix transformation from emotional feature vector sequence $\hat{X}$ to corresponding neutral feature vector sequence Yis performed in two steps:

1) Conversion of emotion feature vector sequence to static MFCC vector sequence $\hat{x}$ using GMM parameter generation algorithm [14].

2) Calculation of delta and acceleration coefficients from every static MFCC vector and then concatenate these coefficients with staticMFCC vector to obtain the outcome sequence $\hat{X}$. In contrast to the MMSE-based methods, the derivation of $\hat{x}$ is based on a maximum likelihood estimate (MLE)as follows:

$$\hat{x} = \arg\max \; p(X|Y, \lambda^{(Z)}) \quad (14)$$

where $\hat{x} = [\, \hat{x}_1^T, \hat{x}_2^T \ldots\ldots \hat{x}_T^T, \,]^T$is the order of predicated featurevectors feature vectors. A matrixWof dimension 3dT*dTis defined such that it transforms into static sequence $\hat{x}$tothe spreadedsequence $\hat{X}$ as following expression

$$\hat{X} = W\hat{x}$$

where $\hat{X}$ is the order of de-emotionalized MFCC vectors with dynamic (delta and acceleration) coefficients as defined in

Equations (3) and (5). The composition of the matrix W isdiscussed as follows:

$w = [w_1, w_2, \ldots..w_t, \ldots w_T]^T * I_{DXD}$ (16)

$w = [\, w_t^0, w_t^1, \, w_t^2]$ t = 1,2,….T (17)

$w_T^n = [0, (t - L_-^n)th, \, (t + L_+^n)th, \,$ (t)th, $w^{(n)}(-L_-^n), \; w^{(n)}(L_+^n),$ $w^{(n)}(0), \ldots.0]$ where n = 0, 1 , 2 (18)

In Equation (16), each sub-matrix W_t is of size T x3 and '*' depicts the Kronecker product. In equation (18),$w^{(n)}(\tau)$ denotes the weights needed for estimating the Δ^n MFCCcoefficient for the $(t+\tau)^{th}$ time frame. τ varies in a framespan of $[-L_-^n, L_+^n]$ describing using the equations givenbelow as –

$\Delta x_t = \sum_{-L_-^1}^{L_+^{(1)}} w(1)(\tau) x_{t+\tau}$ (19)

$\Delta^2 x_t = \sum_{-L_-^2}^{L_+^{(2)}} w(2)(\tau) x_{t+\tau}$ (20)

Equation (14) is used for estimate the MLE and it is determined by an EM algorithm which recursively maximizes an auxiliary function with respect to $\hat{x}$ as follows

$Q(X, \hat{X}) = \sum_j p\left(j/Y, X, \lambda^{(z)}\right) \log(p(\hat{X}, j/Y, \lambda^{(z)}))$ (21)

The sequence of vector $\hat{x}$ obtained as a solution of Equation (21) is given by

$\hat{x} = (w^T \overline{(D^x)^{-1}} w)^{-1} w^T \overline{(D^x)^{-1} E^x}$ (22)

Where

$\overline{(D^x)^{-1}} = diag[\overline{(D_1^x)^{-1}}, \overline{(D_2^x)^{-1}}, .. \overline{(D_t^x)^{-1}}, .. \overline{(D_T^x)^{-1}}$ (23)

$\overline{(D^x)^{-1} E^x} = \overline{(D_1)^{-1} E_1^x}^T, \overline{(D_2)^{-1} E_2^x}^T, \ldots\ldots$ (24)
$\overline{(D_t)^{-1} E_t^x}^T, .. \overline{(D_T)^{-1} E_T^x}^T$

$\overline{(D^x)^{-1}}$in equation (23) is a block diagonal matrix of size 3dT x 3dT while $\overline{(D^x)^{-1} E^x}$ shown in equation (24) is a vector of size 3dT x1.The independently composition of the matrices

i.e. $(D_t^X)^{-1}$and $(D_t^X)^{-1} E_t^x$ can be given by

$\overline{(D_t^X) - 1} = \sum_{j=1}^{M} \lambda_{j,t} (D_j^X)^{-1}$ (25)

$\overline{(D_t^X) - 1 E_t^x} = \sum_{j=1}^{M} \lambda_{j,t} (D_j^X)^{-1} E_{j,t}^x$ (26)

$\lambda_{j,t} = p\left(j/Yt, Xt, \lambda^{(z)}\right)$ (27)

On solving the equation (22) we get only static MFCC vectors i.e., a vector of size dTx1. The full sequence with delta and acceleration coefficients adjoining with the resultant vector can be determined by a simple linear operation $w\hat{x}$.

5. Experimental setup used for CV Recognition in theemotion environments

One male and one female professional artists from All India Radio, Varanasi were selected for recording emotional speech. We have considered three basic emotions (neutral, anger, sadness) for both male and female respectively using 2000 different sentence. After recording, we extract CV unit in three emotion using MATLAB program and categories in five classes (/a/, /e/, /i/, /o/ and /u/) in different syllable. Phonetic balanced unit database is collected at 16000 Hz sampling rate and it is down-sampled to 8000 Hz. In the experiment, we have taken 12,367 CV utterances in three emotions anger, sadness, and neutral. Out of the total units, 9,085 CV utterances are used for training and 3,272 CV utterances are used for testing. The number of CV utterances for female speaker for anger, sadness and neutral is 2176, 1913 and 2023 respectively. The number of CV utterances for male speaker for anger, sadness, and neutral is 2232, 1978 and 2045 respectively taken for the experiment. For building HMM model we uses HTK software toolkit. HTK toolkit has been developed by Carnegie Mellon University (CMU) and for SVM we have used SVMTORCH. SVM models are trained using one against the rest approach and HMM models are trained using maximum likelihood approach. Mel-frequency cepstral coefficients (MFCC) obtained from each 25 ms of CV segment with 5 ms shift are used for training and testing the acoustic models. For building SVM models a fixed dimension MFCC feature vector extracted using the formula:

p = (S * PL) / SL; where s = 0,1,…..(SL-1)

p = 0, 1…PL-1

PL is showing template size, and PL is segment length. Few frames are obtained, if the length of segment is larger than SL, If the PL is smaller than PL, a few frames of the segment are repeated.

6. Results and discussion

The recognition performance of emotional CV units using individual HMM and SVM and combined HMM+SVM have been shown in Tables II, III, IV and V for both male and female, with and without Trajectory based stochasticfeature mapping (TSFM) in five classes and three emotions (anger, sadness and neutral). As shown in Table II, among different HMM and SVM models, HMM for vowel recognition followed by SVM for consonant (Vowel based consonant subclasses /a/, /e/, /i/, /o/, /u/) recognition is giving optimal performance. From the results it is evident that HMM and SVM acoustic modelingapproach for emotional CV recognition gives 8-9 % improvement in performance compared to without using TSFM method. The recognition performance of CV units is analyzed by giving features extracted using MATLAB program. CV recognition models are trained with 9,085 CV utterances in three emotions (anger, sadness, and neutral) speech and it is tested with 3,272 CV utterances. For both male andfemale speakers, emotional CV units are recognized in different emotions (anger, sadness, and neutral) for each vowel class /a/, /e/, /i/, /o/, /u/. In Table II, we have shown that the average performance of all the vowel class is 60.07% in the neutralemotion for HMM and 54.51% in SVM.It is also observed that therecognition performance in anger and sad emotions is degraded than neutral emotion and it is 52.96%

and 47.67% respectively for HMM; 27.04% and 24.63 % respectively for SVM. From Table III, it is evident that the overall recognition performance for female speaker is increased by approximately 8% when TSFM approach is applied to map emotional feature vectors to neutral feature vectors.In Table IV, we have shown the average recognition performance for different emotions of male speaker without using TSFM. It is observed that the recognition performance for the vowel class is 59.15% in neutral emotion for HMM and53.17% in SVM. The recognition performance in anger and sadness is 51.12% and 45.41% respectively for HMM; 25.10% and 23.88 % respectively for SVM. The performance of anger and sadness emotion for female is 10%-12% less compared to the neutral emotion without using TSFM for male. After using TSFM method, the overall performance of emotional CV unit is enhanced by approximately 8% for male speaker (see Table V). The overall performance is enhanced to 63.18% after combining HMM+SVM for female speaker as shown in Fig. 4. Fig. 5 shows that the overall recognition performance for the male performance is enhanced to 55.79% and 54.74% for anger and sad emotions respectively for HMM; and 36.30% and 35.2% in anger and sad emotions for SVM respectively. After combining HMM+SVM result has raisedupto 61.47%, as shown in Fig. 6 and Fig. 7. Therefore, after using TSFM method the result is enhanced from 55.84% to 63.18% for male speaker and 53.31% to 61.47% in female speaker respectively for CV recognition in emotion environments.

TABLE II: Recognition Performance of emotional CV units for Female speaker without using TSFM

HMM					SVM				Result
Class	A	S	N	overall	A	S	N	Overall	Overall
/ a /	52.71	52.45	62.67	52.61	30.67	30.67	50.81	37.38	58.12
/ e /	52.45	50.21	57.21	52.63	24.45	25.37	53.22	34.34	54.97
/ i /	49.33	46.71	58.89	52.12	26.67	26.01	54.67	35.78	54.13
/ o /	52.33	43.21	59.81	51.81	28.11	21.21	58.89	36.07	53.11
/ u /	54.01	44.86	63.01	57.45	25.34	19.91	55.00	45.21	58.74
overall	52.96	47.67	60.07	53.69	27.04	24.63	54.51	37.75	55.84

A= Anger S= Sadness N= Neutral Result= HMM+SVM

TABLE III: Recognition Performance of emotional CV units for Female speaker by using TSFM

HMM					SVM				Result
Class	A	S	N	overall	A	S	N	Overall	Overall
/ a /	60.11	61.11	67.02	62.29	48.71	41.37	51.28	46.90	64.53
/ e /	54.21	53.21	58.10	56.05	35.21	37.67	52.53	42.00	62.07
/ i /	57.47	51.81	58.89	56.45	34.81	36.64	55.81	42.04	62.17
/ o /	58.81	57.53	66.01	58.64	39.21	33.11	59.78	43.73	63.02
/ u /	61.60	57.13	65.34	61.01	32.43	31.43	55.00	39.62	64.13
overall	58.44	56.23	63.52	58.88	38.07	36.04	54.88	42.85	63.18

A= Anger S= Sadness N= Neutral Result= HMM+SVM

TABLE IV: Recognition Performance of emotional CV units for Male speaker without using TSFM

HMM					SVM				Result
Class	A	S	N	overall	A	S	N	Overall	Overall
/ a /	52.01	50.44	64.67	53.70	29.27	29.80	50.81	36.39	55.53
/ e /	51.45	48.33	57.21	52.19	22.93	24.87	52.71	33.33	53.22
/ i /	48.01	43.41	55.11	48.84	24.80	25.00	53.76	34.50	51.78
/ o /	50.33	40.05	57.01	49.13	26.39	20.96	58.89	35.11	52.86
/ u /	52.81	42.86	62.15	52.60	22.08	18.0	54.12	36.67	53.19
overall	51.12	45.41	59.15	51.89	25.10	23.88	53.71	35.18	53.31

A= Anger S= Sadness N= Neutral Result= HMM+SVM

TABLE V: Recognition Performance of emotional CV units for Male speaker by using TSFM

HMM					SVM				Result
Class	A	S	N	overall	A	S	N	Overall	Overall
/ a /	57.11	59.61	64.67	60.46	47.21	40.41	58.11	48.57	63.12
/ e /	53.21	52.83	56.81	54.28	32.54	36.92	52.71	40.72	60.57
/ i /	50.81	53.11	55.2	53.01	33.81	35.47	53.71	40.00	60.01
/ o /	56.63	55.28	57.01	56.30	37.49	32.85	58.00	42.78	61.24
/ u /	61.21	52.74	62.15	58.70	30.47	30.44	54.56	38.30	62.45
overall	55.79	54.74	59.16	56.55	36.30	35.2	55.41	42.07	61.47

A= Anger S= Sadness N= Neutral Result= HMM+SVM

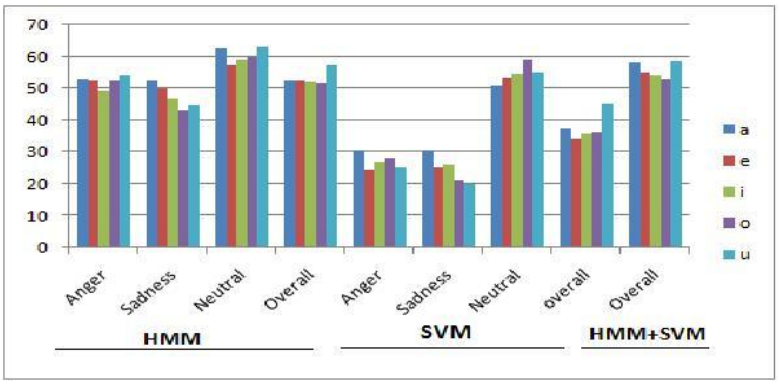

Fig. 4: Performance of female speaker in anger, sadness, and neutral for each CV unit in classes a, e, i, o, u without using TSFM

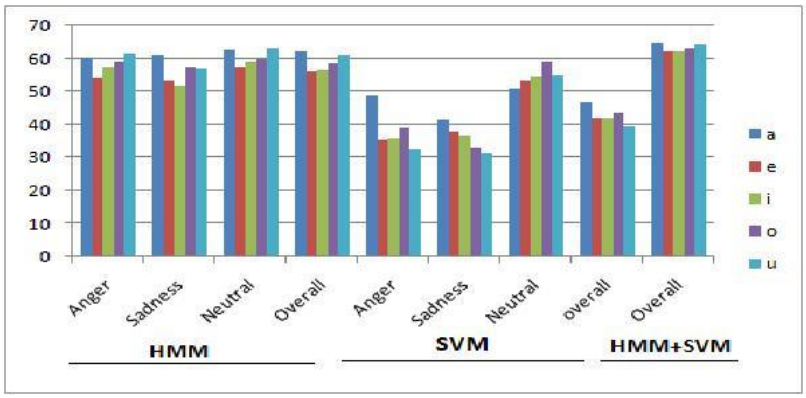

Fig. 5: Performance of female speaker in anger, sadness, and neutral for each CV unit in classes a, e, i, o, u by using TSFM.

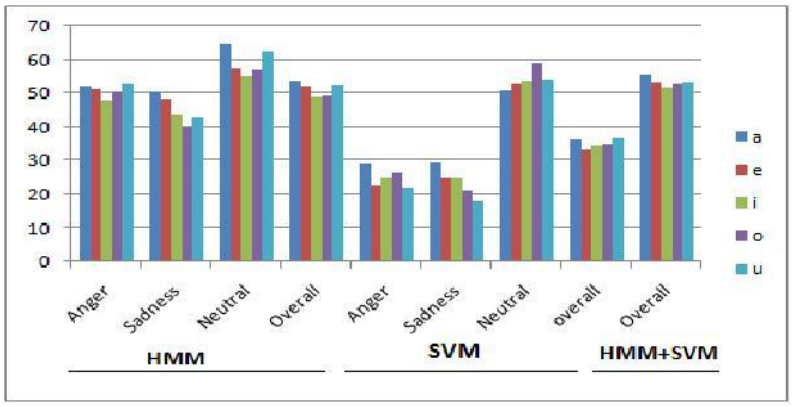

Fig.6: Performance of male speaker in anger, sadness, and neutral for each CV unit in classes a, e, i, o, u without using TSFM.

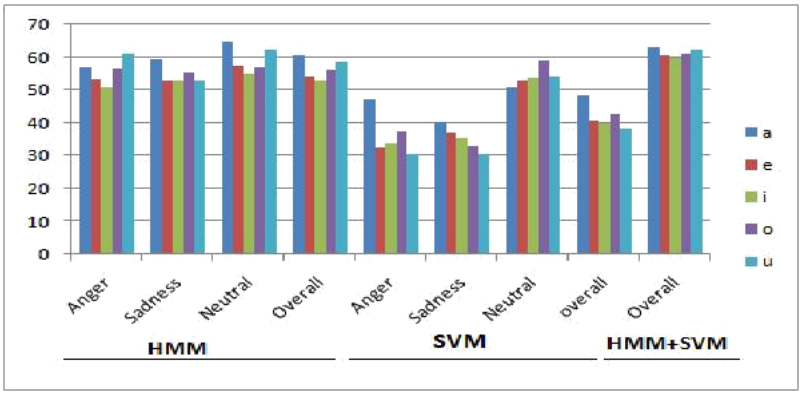

Fig. 7: Performance of male speaker in anger, sadness, and neutral for each CV unit in classes a, e, i, o, u by using TSFM.

7. Summary and Conclusions

In this paper, we have proposed a TSFM method for increasing the recognition performance of CV unit under emotional environments. In the TSFM method, we have used MFCC feature transformation mapping method, which transforms anger, sadness and emotional MFCC features to their corresponding neutral MFCC features for both male and female speakers so that anger, sad and emotional MFCC feature vectors are treated as neutral MFCC features. The mapped neutral CV unit is recognized in two steps. In the first step, vowel would be recognized using HMM model and in the second step consonant would be recognized using SVM model and combined their scores (HMM+SVM) for enhancing the performance. We found that average performance of HMM is 53.69% for neutral, sad and anger emotions. The average recognition performance using SVMis 40% in neutral, sadness and anger while it gives 50-58% only in neutral and after merging HMM and SVM the combined score yields 51% average performance without using TSFM method. In the TSFM method we map the emotional MFCC into neutral MFCC feature and then we build the HMM and SVM and combined their scores for both male andfemale speakers using mapped neutral MFCC. The performance of HMM are 58.88% average for neutral, sadness and anger. The SVM gives average performance 42.85% in neutral, sad and angry emotions. After combining HMM and SVM scores, the recognition performance is 63.12% for female and 61.47% for male. Therefore, the recognition performance of emotional CV unit is increased by approximately 8-9% using proposed TSFM method compared to the CV recognition performance without using TSFM.

References

1. K. Vuppala, K. S. Rao, S. Chakrabarti, P. Krishnamoorthy, and S. Prasanna, “Recognition of consonant-vowel (cv) units under back-ground noise using combined temporal and spectral preprocessing,” International Journal of Speech Technology, vol. 14, no. 3, pp. 259– 272, 2011.
2. K. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, “Vowel onset point detection for low bit rate coded speech,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 6, pp. 1894– 1903, 2012.
3. K. Vuppala, K. SreenivasaRao, and S. Chakrabarti, “Improved consonant–vowel recognition for low bit-rate coded speech,” Interna-tional Journal of Adaptive Control and Signal Processing, vol. 26, no. 4, pp. 333–349, 2012.
4. L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.
5. S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, “Spotting consonant-vowel units in continuous speech using alitoassociative neural networks and support vector machines,” in Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop. IEEE, 2004, pp. 401–410.
6. J. Burges, “A tutorial on support vector machines for pattern recognition,” Data mining and knowledge discovery, vol. 2, no. 2, pp. 121–167, 1998.
7. L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, “High-performance robust speech recognition using stereo training data,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’01), vol. 1, 2001, pp. 301–304.
8. L. Buera, E. Lleida, A. Miguel, and A. Ortega, “Multi-environment models based linear normalization for speech recognition in car conditions,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’04), 2004.
9. L. Buera, E. Lleida, A. Miguel, A. Ortega, and scar Saz, “Cepstral vector normalization based on stereo data for robust speech recogni-tion,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 3, pp. 1098–1113, 2007.
10. L. Buera, A. Miguel, scar Saz, A. Ortega, and E. Lleida, “Unsuper-vised data-driven feature vector normalization with acoustic model adaptation for robust speech recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 2, pp. 296– 309, 2010.
11. M. Afify, X. Cui, and Y. Gao, “Stereo-based stochastic mapping for robust speech recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 7, pp. 1325–1334, 2009.
12. T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.
13. H. Zen, Y. Nankaku, and K. Tokuda, “Stereo-based stochastic noise compensation based on trajectory GMMs,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’09), 2009.

14. K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, “An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features,” in Proceedings of the European Conference of Speech Communication Technology (EUROSPEECH ’95), Madrid, Spain, September 1995,pp.757–7

15. Jainath Yadav and K. SreenivasaRao. Neural network and GMM based feature mappings for consonant---vowel recognition in emotional environment. Int. J. Speech Technol. 21, 3 (September 2018), 421-433.

