Egyptian Informatics Journal 22 (2021) 45–51

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

An improved gaussian mixture hidden conditional random fields model for audio-based emotions classification

Muhammad Hameed Siddiqi

Department of Computer Science, Jouf University, Sakaka, Saudi Arabia

ARTICLE INFO

Article history:
Received 28 November 2019
Revised 23 January 2020
Accepted 29 March 2020
Available online 15 April 2020

Keywords:
Emotion classification
Conditional random fields
Hidden markov model
Gaussian mixture model

ABSTRACT

The analysis of human emotions plays a significant role in providing sufficient information about patients in monitoring their feelings for better management of their diseases. Audio-based emotions recognition has become a fascinating research interest for such domains during the last decade. Mostly, audio-based emotions systems depend on the recognition stage. The existing model has a common issue called objectivity suppositions problem, which might decrease the recognition rate. Therefore, this study investigates the improved version of a classifier that is based on hidden conditional random fields (HCRFs) model to classify emotional speech. In this model, we introduced a novel methodology that will incorporate multifaceted dissemination with the help of employing a combination of complete covariance Gaussian concreteness function. Due to this incorporation, the proposed model tackle most of the limitations of existing classifiers. Some of the well-known features like Mel-frequency cepstral coefficients (MFCC) are extracted in our experiments. The proposed model has been validated and evaluated on two publicly available datasets likes Berlin Database of Emotional Speech (Emo-DB) and the eNTER FACE'05 Audio-Visual Emotion dataset. For validation and comparison against the existing techniques, we utilized 10-fold cross validation scheme. The proposed method achieved significant improvement under the p-value <0.03 for classification. Moreover, we also prove that computational wise, our computation technique is less expensive against state of the art works.

1. Introduction

Commonly, emotion is a psychological status that impulsively performed and raised. Emotion is not only a good pointer to determine the mental status of a human but also an effective source of conveying our intentions in daily conversations. This is because the automatic recognition of human sentiments is a fascinating parameter in order to improve the superiority of the facilities delivered by the computer like human–computer interaction [1,2], daily life observing in omnipresent health care systems [3]. There are certain physiological variations like talking, blood density, heart signal, facial expressions, etc. that express the human sentiments (emotions). Among these variations, most of the researchers point out that audio speech is the main source of emotions [4–10]. Because audio signals are the most widely and naturalistic method for human to human communication.

Generally, there are two steps in a typical audio-based recognition system: First step extracts the most prominent features from the input data; while, the second step decides the appropriate label for incoming input data. In audio-based emotion recognition system, many methods have been proposed for the feature extraction stage to extract the most significant features. There are four categories for such features (named continuous speech features) like pitch of sound, formant, vitality [11,1,10], voice quality features [1,12] (e.g harsh, tense, breathy), spectral features [11,13] (e.g undeviating extrapolation measurements, Mel-frequency cepstrum coefficients), and Teager energy operator (TEO) [14]. In literature of speech classification [7], certain systems [15] suggested that some suitable feature selection are highly dependent on the recognition task; therefore, it should be considered. Moreover, they pointed out that for speech demonstration, the MFCC features are the most significant features. As a new feature extraction method is not in the scope of this study; therefore, we are utilizing an existing technique to excerpt the MFCC features that further will be employed in the proposed model.

Although there is relatively an enormous number of latest research concentrating on refining the classification phase

E-mail address: mhsiddiqi@ju.edu.sa
Peer review under responsibility of Faculty of Computers and Information, Cairo University.

https://doi.org/10.1016/j.eij.2020.03.001

[11,8,12,9,2]. Mostly, in the existing audio-based emotion recognition systems, the authors utilized conformist learning model [1] like Gaussian mixture model, hidden Markov model, artificial neural networks, support vector machines, etc. Certain studies [16–18,13,19–22] highlighted that HMM is one the most significant model used in audio-based emotion recognition system. Furthermore, other areas like speech classification [23], posture classification [24,25] pointed out that HMM is a multiplicative learning method, due to which it is least precise than its discriminative corresponding part such as HCRFs. The following contributions have been made in this study.

- The previous HCRF technique is insufficient through unconventionality norms that might decrease the recognition rate. Therefore, in this work, a recognition model has been presented that diminish the supposition by utilizing the full covariance distribution, which is the first objective of this study.
- The second objective of this study is to show that the proposed model significantly reduced the complexity against the previous techniques. In this method, the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) technique has been utilized in order to find the optimal point, which is the alternative goal of this model to determine certain factors at the training stage to extend the conditional probability of the training data. So, in order to calculate the conditional probability, we only employed the forward and backward methods that further used for computing the gradients. Due to which the computational time has thoroughly condensed.
- In order to show the significance of the proposed model, a set of experiments is performed that showed the best performance compared to the latest works.

2. Related work

Recently, there has been growing research work in the field of emotion recognition based on speech data to increase the accuracy of such systems [1]. However, very few of these attempts truly add to the efficiency of learning models for speech data. Authors of [1] have rightly pointed out that, although numerous classification methods [26–29] have been tested and used to improve the efficiency of learning model, HMM still remains as the most common and efficient method. The accuracies of HMM on various datasets as compared to GMM, ANN, and SVM etc. is still comparable. Moreover, HMM presents the advantage and ability to process sequential data e.g. processing of frame-level features whereas GMM, ANN, and SVM lack this and cannot process sequence of feature vectors.

HMM, however, presents some limitations as pointed out in [23,25,24]. The main limitations of HMM are due to its propagative nature and the objectivity hypothesis among its states and the interpretations. To address the limitations of HMM, a technique called maximum entropy Markov model was proposed. This model shows better results for certain operations/tasks including part-of-speech tagging (POS) [30], evidence abstraction [31], and the recognition of automatic speeches [32]. However, maximum entropy Markov model itself suffers from the weakness/problem of label biasness [33]. Label biasness in MEMM is mainly caused by its per-state stabilization of transitional notches that implies a notch conservation at every notch.

To address the label biasness in MEMM, conditional random field [33] and hidden conditional random field (HCRF) [23,25] were proposed. CRF and HCRF are generalizations of MEMM are generalized forms of MEMM and hence inherit its properties. Both, CRF and HCRF use global normalization in contrast to MEMM's per-state normalization, and additionally, HCRF maintains hidden states to be capable to absorb unknown construction of successive records. As result, CRF and HCRF can work with weighted scores making the set of parameters used comparatively larger as compared to MEMM and HMM. We refer the reader for a comprehensive and detailed analysis of HCRF and its limitations to [34].

There have been certain approaches that utilized HCRF model and showed good results. These are explained and presented in [35,36]. However, these systems do not address the limitations of HCRF. In [34], the authors argue that only HCRF might be utilized diagonal (slanting) covariance Gaussian dissemination. Particularly, the variables are considered pairwise autonomous. From now, this model is referred diagonal covariance Gaussian Mixture hidden conditional random field. Additionally, the authors stated, through a particular set of fixed prices, the solidity of observations at every state converges to the Gaussian procedure. However, this assumptions is not supported by a training algorithm, and hence these assumptions may be counter productive i.e. decreasing the efficiency of the model. For more in-depth study, we refer the reader to [37,38].

To address the limitations of HCRF and other learning models for emotion recognition on speech data, in the following section, we present our novel approach based on HCRF method, which has the ability to overtly exploit combination of full covariance Gaussian dissemination. Our method gets the benefits from existing HCRF. We apply and test our model on speech data to recognize emotions and compare its results with those obtained by HMM and HCRF with diagonal covariance Gaussian functions.

3. The proposed method

As presented in the previous section, the current version of HCRF model didn't utilize full covariance matrix and cannot assure the convergence of parameters. This leads the existing HCRF model not being able to generate a set of values, where the provisional probability is modeled as a mixture of usual solidity functions. Therefore, in the feature functions, we include combinations of Gaussian dissemination in order to solve the above problem. Our function (feature) are given by the following equations:

$$\overset{Prior}{f}_{k}(Z,\overline{K},Y) = \delta(k_1 = k).z, y \forall k, \tag{1}$$

$$f_{k\,k'}^{Transition}(Z,\overline{K},Y) = \sum_{t=1}^{T}\delta(k_{t-1} = k)\,\delta(k_t = k').z, y \forall k, k', \tag{2}$$

$$f_{k}^{Observe}(Z,\overline{K},Y) = \sum_{t=1}^{T}\log\left(\sum_{l=1}^{L}\Gamma_{k,l}^{Obsr}N(y_t^2,u_{k,l},\Sigma_{k,l})\right).\delta(k_t = k).z, \tag{3}$$

then,

$$N(y_t^2,u_{k,l},\Sigma_{k,l}) = \frac{1}{(2\pi)^{\frac{Dim}{2}}|\Sigma_{k,l}|^{\frac{1}{2}}}\exp\begin{pmatrix} -\frac{1}{2}(y_t^2-u_{k,l})'. \\ \sum_{k,l}^{-1}(y_t^2-u_{k,l}) \end{pmatrix}, \tag{4}$$

Here, L represents the amount of solidity functions, Dim shows the dimension of observations, $\Gamma_{k,l}^{Obsr}$ indicates the fraternization mass of the n^{th} factor having average $u_{k,l}$, and surrounding substance (covariance matrices) $\sum_{k,l}$. We might modify Γ, u and $\sum$ in order to build any fusion of standard solidities through investi-

gating against Eq. (3). Hence, the conforming reflection weightiness $\Lambda_{k,l}^{Obsr}$ does not necessarily require to be rationalized throughout the training phase, and this allows us to fix

$$\Lambda_{k,l}^{Obsr} = \forall k \tag{5}$$

Consequently, we can write the conditional probability as follows:

$$Posterior\left(Z|Y;\sum,u,\Gamma,\Lambda\right) = \frac{\sum\limits_{\overline{K}} \exp\left(\begin{array}{l}\sum\limits_{k}\Lambda_{k}^{Prior} f_{k}^{Prior}(Z,\overline{K},Y)+\\ \sum\limits_{kk'}\Lambda_{kk'}^{Transition} f_{kk'}^{Transition}(Z,\overline{K},Y)\\ +\sum\limits_{k} f_{k}^{Observe}(Z,\overline{K},X)\end{array}\right)}{Norm(Y,\sum,u,\Gamma,\Lambda)}, \tag{6}$$

$$Posterior\left(Z|Y;\sum,u,\Gamma,\Lambda\right) = \frac{\sum\limits_{\bar{K}=k_1,k_2,\ldots,k_T} \exp\left(\begin{array}{l}\Lambda_{k_1}^{Prior}+\sum\limits_{t=1}^{T}\left(\Lambda_{k_{t-1},k_t}^{Transition}\right)+\\ \log\left(\sum\limits_{l=1}^{L}\Gamma_{k_t,l}^{Obsr} N(y_t^2,u_{k_t,l},\Sigma_{k_t,l})\right)\end{array}\right)}{Norm(Y,\sum,u,\Gamma,\Lambda)} \tag{7}$$

$$Posterior\left(Z|Y;\sum,u,\Gamma,\Lambda\right) = \frac{fScore(Z|Y;\sum,u,\Gamma,\Lambda)}{Norm(Y;\sum,u,\Gamma,\Lambda)} \tag{8}$$

where $Norm(Y;\sum,u,\Gamma,\Lambda)$ is the normalization factor.

Utilizing equations(7) and (8), re-evaluate the conditional probability using forward and backward algorithms, as shown below step by step:

$$a_\tau = \sum_{\overline{K}=k_1,k_2,\ldots,\{k_\tau=k\}} \exp\left(\begin{array}{l}\Lambda_{k_1}^{Prior}+\sum\limits_{t=1}^{\tau}\left(\Lambda_{k_{t-1},k_t}^{Transition}\right)+\\ \log\left(\sum\limits_{l=1}^{L}\Gamma_{k_t,l}^{Obsr} N.(y_t^2,u_{k_t,l},\Sigma_{k_t,l})\right)\end{array}\right)$$

$$= \sum_{k'} a_{\tau-1}(k')\exp\left(\Lambda_{k'k}^{Transition}+\log\left(\sum_{l=1}^{L}\Gamma_{k,l}^{Obsr} N.(y_\tau,u_{k,l},\Sigma_{k,l})\right)\right), \tag{9}$$

$$\beta_\tau(k) = \sum_{\bar{K}=\{k_\tau=k\},k_{\tau+1},\ldots,k_T} \exp\left(\begin{array}{l}\Lambda_{k_1}^{Prior}+\sum\limits_{t=\tau}^{T}\left(\Lambda_{k_{t-1},k_t}^{Transition}\right)+\\ \log\left(\sum\limits_{l=1}^{L}\Gamma_{k_t,l}^{Obsr} N.(y_t^2,u_{k_t,l},\Sigma_{k_t,l})\right)\end{array}\right)$$

$$= \sum_{k'} \beta_{\tau+1}(k')\exp\left(\Lambda_{kk'}^{Transition}+\log\left(\sum_{l=1}^{L}\Gamma_{k,l}^{Obsr} N.(y_\tau,u_{k,l},\Sigma_{k,l})\right)\right), \tag{10}$$

$$fScore\left(Z|Y;\sum,u,\Gamma,\Lambda\right) = \sum_k a_T(k) = \sum_k \beta_1(k). \tag{11}$$

So far, in the initial (training) phase, we were focused on finding the parameters $(\sum,u,\Gamma,\Lambda)$ of the training data, which have the capability to make best use of conditional probability.

We now are more interested in utilizing one of the existing well known techniques in the proposed model to look the optimum fact. Nevertheless, to compute this, we used only the forward method and backward method when computing the conditional probability (as utilized by other existing works [23]). Furthermore, we re-used its resultant value for computing the inclines. This expectedly reduces computational cost significantly. In the following, we show the gradient computation method briefly:

Let us denote

$$\begin{aligned} c(Z,\bar{K},Y) &= \sum_k \Lambda_k^{Prior} f_k^{Prior}(Z,\bar{K},Y) \\ &+ \sum_{kk'} \Lambda_{kk'}^{Transition} f_{kk'}^{Transition}(Z,\overline{K},Y) \\ &+ \sum_k \Lambda_k^{Observe} f_k^{Observe}(Z,\bar{K},Y) \end{aligned} \tag{12}$$

From Eqs. 6,8,12 we derive

$$\begin{aligned} \frac{dScore(Z|Y;\sum,u,\Gamma,\Lambda)}{d\Lambda_k^{Prior}} &= \sum_{\overline{K}} \frac{dc(Z,\overline{K},Y)}{d\Lambda_k^{Prior}} \exp\left(c(Z,\overline{K},Y)\right) \\ &= \sum_{\overline{K}} f_s^{Prior}(Z,\overline{K},Y)\exp\left(c(Z,\overline{K},Y)\right) \\ &= \beta_1(k) \end{aligned} \tag{13}$$

$$\begin{aligned} \frac{dScore(Z|Y;\sum,u,\Gamma,\Lambda)}{d\Lambda_k^{Transition}} &= \sum_{\overline{K}} \frac{dc(Z,\overline{K},Y)}{d\Lambda_{kk'}^{Transition}} \exp\left(c(Z,\overline{K},Y)\right) \\ &= \sum_{\overline{K}} f_{kk'}^{Transition}(Z,\overline{K},Y)\exp\left(c(Z,\overline{K},Y)\right) \\ &= \sum_{t=1}^{T} a(t,k)\beta(t+1,k') \end{aligned} \tag{14}$$

$$\begin{aligned} \frac{dScore(Z|Y;\sum,u,\Gamma,\Lambda)}{d\Gamma_{k,l}^{Obsr}} &= \sum_{\overline{K}} \frac{dc\left(Y,\overline{S},X\right)}{d\Gamma_{k,l}^{Obsr}} \exp\left(c(Z,\overline{K},Y)\right) \\ &= \frac{\sum\limits_{\overline{K}} f_k^{Observe}(Z,\overline{K},Y)}{d\Gamma_{k,l}^{Obsr}} \exp\left(c(Z,\overline{K},Y)\right) \\ &= \sum_{\overline{K}}\sum_{t=1}^{T} \frac{N(y_t,u_{k,l},\Sigma_{k,l})}{\sum\limits_{l=1}^{L}\Gamma_{k,l}^{Obsr} N(y_t,u_{k,l},\Sigma_{k,l})} \delta(l_t=l)\exp\left(c(Z,\overline{K},Y)\right) \\ &= \sum_{t=1}^{T} \frac{N(y_t,u_{k,l},\Sigma_{k,l})}{\sum\limits_{l=1}^{L}\Gamma_{k,l}^{Obsr} N(y_t,u_{k,l},\Sigma_{k,l})} \alpha(t,k)\gamma(t+1) \end{aligned} \tag{15}$$

Table 1
Average recognition rate accuracy for the proposed model along with diverse number of states and mixtures under 10-fold cross validation scheme against Emo-DB dataset (%).

	1 Mixture	2 Mixtures	3 Mixtures	4 Mixtures	5 Mixtures	6 Mixtures	7 Mixtures	8 Mixtures
1 State	57.72	65.66	67.93	71.71	70.57	73.78	73.93	72.64
2 States	57.18	70.57	**79.31**	69.99	72.98	73.78	73.38	72.66
3 States	61.49	71.51	71.50	73.78	73.34	72.24	70.93	70.79
4 States	64.69	71.08	73.46	72.09	73.26	72.16	70.55	67.22
5 States	65.83	71.95	73.51	71.93	70.43	68.70	67.37	63.09
6 States	63.21	71.70	74.89	71.15	65.65	63.64	64.36	61.58
7 States	69.47	72.85	72.88	69.64	64.63	65.73	60.82	60.27
8 States	66.75	71.88	72.95	67.22	68.89	63.49	61.76	61.00

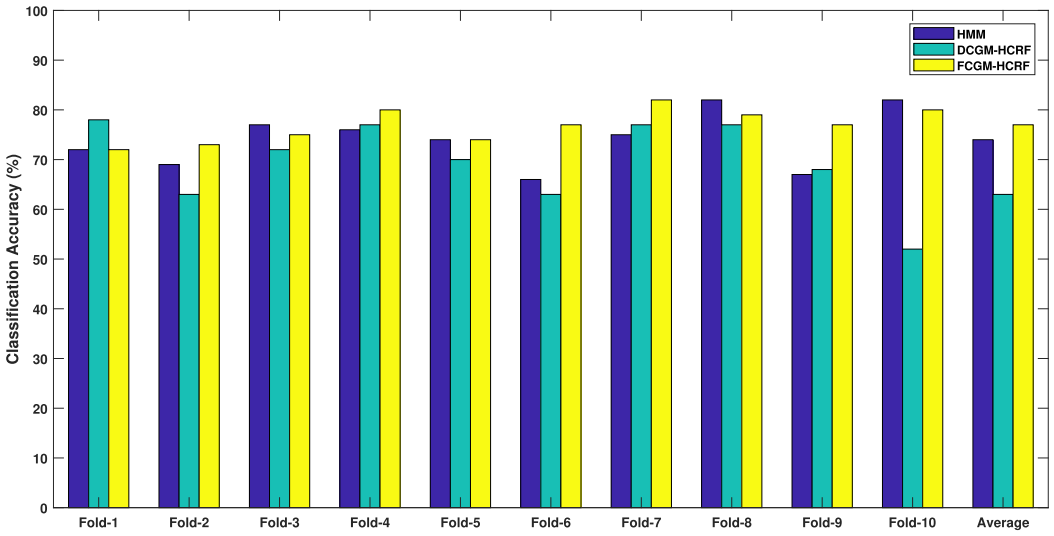

Fig. 1. Recognition rates of three techniques with 2 states and 3 mixtures against Emo-DB dataset.

Table 2
Average recognition rate accuracy for the proposed model model along with diverse number of states and mixtures under 10-fold cross validation scheme against eNTER FACE'05 dataset (%).

	1 Mixture	2 Mixtures	3 Mixtures	4 Mixtures	5 Mixtures	6 Mixtures	7 Mixtures	8 Mixtures
1 State	51.00	52.76	50.49	61.89	60.22	59.95	56.88	54.75
2 States	53.27	60.82	63.51	64.63	50.64	58.88	62.85	57.47
3 States	61.58	57.36	63.64	55.65	61.15	54.89	61.70	63.21
4 States	63.09	57.37	58.70	60.43	61.93	**65.73**	60.95	53.83
5 States	57.22	49.55	52.16	63.26	62.09	63.46	59.08	56.69
6 States	60.79	59.93	64.24	63.34	53.78	61.50	59.51	61.49
7 States	62.66	63.38	58.78	62.98	59.99	49.31	50.57	57.18
8 States	57.64	53.93	63.78	60.57	61.71	63.93	55.66	57.72

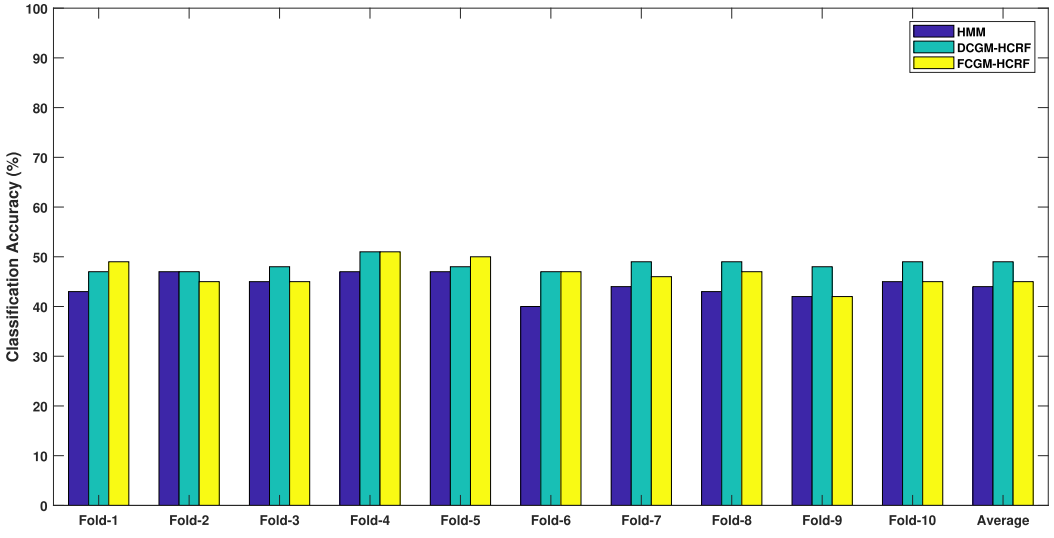

Fig. 2. Recognition rates of three techniques with 2 states and 6 mixtures against eNTER FACE'05 dataset.

$$\gamma(t) = \sum_{l} \beta(t,k) \tag{16}$$

Moreover, we get the gradients with respect to u, and $\sum$ similarly in the following:

$$\frac{dScore(Z|Y;\sum,u,\Gamma,\Lambda)}{du_{k,l}} = \sum_{t=1}^{T} \frac{\Gamma_{k,l}^{Obsr} \frac{cN(y_t,u_{k,l},\Sigma_{k,l})}{cu_{k,l}}}{\sum_{l=1}^{M} \Gamma_{k,l}^{Obsr} N(y_t,u_{k,l},\Sigma_{k,l})} \alpha(t,k)\gamma(t+1) \tag{17}$$

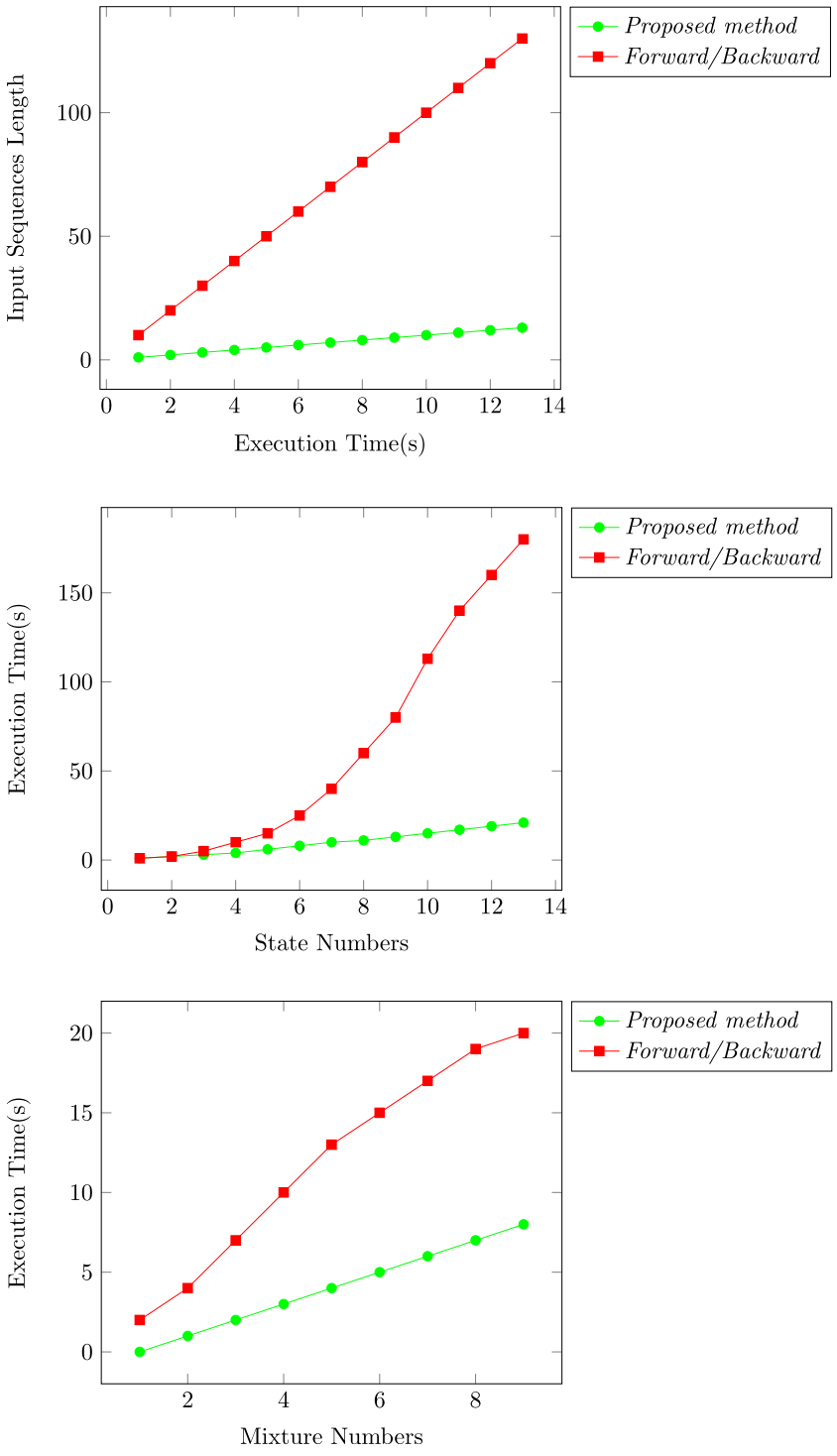

Fig. 3. Comparison gradient computation time of the proposed method against forward/backward on different mixture numbers, state numbers, and input sequence lengths (as shown in Eq. (17)).

$$\frac{dScore(Z|Y;\sum,u,\Gamma,\Lambda)}{d\Sigma_{k,l}}=\sum_{t=1}^{T}\frac{\Gamma_{k,l}^{Obsr}\frac{cN(y_t,u_{k,l},\Sigma_{k,l})}{c\Sigma_{k,l}}}{\sum_{l=1}^{L}\Gamma_{k,l}^{Obsr}N(y_t,u_{k,l},\Sigma_{k,l})}\alpha(t,k)\gamma(t+1) \tag{18}$$

Having the gradients computed as shown in Eqs. (13)–(18), we now make use of the L-BFGS algorithm and find a local maximum for the conditional probability. Since we cannot define or find a global maximum, we execute the algorithm several times with different initial values/points and collect the results of each run in a set. Among those results, we then choose the set of parameters. This method provides the best results as we show in the following section.

4. Results evaluation and discussion

This section presents the experimental results of the designed approach. Moreover, the experimental framework, the datasets, and the outcomes are also presented in this section.

To perform a fair assessment on publicly benchmark datasets, we use the Emo-DB dataset [39] and the eNTER FACE'05 dataset [40]. First, from each dataset, we extract the Mel-frequency cepstral coefficients (MFCCs). Then, 10-fold cross validation scheme has been employed to separate each of the above two datasets into a training part and a validation part. We then execute classification algorithms on these datasets and the algorithms are: 1) our proposed FCGM-HCRF, 2) HMM model and 3) the HCRFs model which uses diagonal covariance Gaussian mixtures (DCGM-HCRF). Next, for a fair comparison of our algorithm with others, we compute p-values using the paired t-testing. The evaluation results of those comparisons are presented in the following.

4.1. Berlin database of emotional speech – Emo-DB

The Emo-DB dataset consists of expressive exclamations of 10 actors and actresses from German. Of which 50% are male and 50% are female. The utterances (or the spoken sentences) are a set of predefined sentences in one of the 7 defined emotional states: 1) neutral, 2) boredom, 3) disgust, 4) fear, 5) sadness, 6) Joy and 7) anger. Each successful attempt by the actors and actresses have been evaluated by a group of 20 judges and the final concluding utterance is only selected if 80% of the listeners have correctly recognized.

In our experimentation for this dataset, we run HMM along with diverse numbers of states and Gaussian mixtures. In Table 1, for each pair of state and mixture numbers, we present the average classification rates of 10 folds.

From Table 1, it can be seen that the HMM with exactly two states and three Gaussian mixtures gives the highest accurate results. Therefore, we apply the initial values of this HMM to train and evaluate FCGM-HCRF and DCGM-HCRF algorithms. The results of these algorithm compared to HMM are shown in Fig. 1.

4.2. The eNTER FACE'05 Audio-Visual Emotion Dataset

We now present our evaluation on the eNTER FACE'05 dataset. This dataset consists of one thousand three hundred and twenty (1320) videos produced by 44 subjects/actors. Each of these actors in videos tries to simulate six different emotions: 1) anger, 2) disgust, 3) fear, 4) happiness, 5) sadness, and 6) surprise. These emotions are simulated while reading 5 different pre-defined sentences. First of all, we extract the audio from the original video files/data and then extract the MFCC coefficients. We then use these coefficients to generate the test and traning datasets similarly to what we did in the Emo-DB dataset. We also repeat the same process of computing the initial points and evaluating the proposed model as we did for Emo-DB dataset. Results of our benchmark algorithms compared to the proposed approach on eNTER FACE'05 dataset are presented in Table 2, and Fig. 2.

4.3. Computational complexity

In this section, we briefly discuss the computational complexity of the proposed approach as compared to others. The existing HCRF algorithm computes the gradients by a series of forward and backward algorithms. We however, execute the forward and backward algorithm once and cache the results in-memory for later use. Forward and backward algorithm has the complexity $O\left(TQ^2M\right)$ for input sequences of length T, states of size Q and the M number of mixtures. This complexity can be seen/derived from Eqs. (9) and (10). On the other hand, our proposed algorithm with caching has the complexity $O(TM)$ for computing gradients. The proof for these complexity theoretic results can be seen in Eqs. (13)–(18).

In 3, we present a comparative analysis of the total execution time when inclines are calculated using the forward method and backward method as compared to the proposed approach with caching. The reported execution times are measured using the Matlab R2013a running on an Intel machine with Duo 3.6 GHz processor and 4 GB of main memory. The proposed model has been compared with state of the art methods using. The corresponding recognition rates of the existing methods along with the proposed model on Berlin Database of Emotional Speech (Emo-DB) and eNTER FACE'05 Audio-Visual Emotion dataset are presented in Table 3. It is clear from the Table 3 that the proposed model showed better performance compared to other latest systems. This is because the proposed model utilized full-covariance distribution that considered most of the coefficients of the matrix, and that is one of the main reason to improve the performance. Moreover, this work showed that the existing HCRFs model has a common problem due to which it might decrease the recognition rate. This drawback is called objectivity suppositions problem. Therefore, full-covariance distribution based model is proposed to reduce the supposition that make the proposed model capable to consider the all coefficients of the matrix.

Table 3
Comparison result of the proposed model along with the existing methods under the two standard datasets (Unit: %).

State of the art methods	Accuracy	Standard Deviation
[41]	72.0	±1.2
[42]	70.5	±3.8
[43]	63.6	±2.2
[44]	77.0	±2.7
[45]	67.7	±1.5
[46]	78.3	±2.5
[47]	79.2	±1.1
[48]	72.3	±2.5
[49]	76.4	±2.5
Proposed Method	**79.3**	±3.8

5. Conclusions

Audio-based emotion recognition has received lots of attention over the past decade. Several audio-based emotion recognition systems have been proposed; however, still it is major issue for most of the systems to correctly classifying the emotions. There are some attributes which may degrade the accuracy, e.g extraction of the prominent features, and high similarity among different emotions that occurs in the presence of low between-class variance in the feature space.

Accordingly, we have presented a new version of the HCRF algorithm that uses full covariance Gaussian density functions. Then, we proved it theoretically and experimentally that the recognition rates of the proposed approach is comparatively precise than existing algorithms. We also proved that these improvements are statistically correct by using p-values for testing and comparisons. Moreover, our algorithm does not only add to the accuracy of recognition of emotions, it also has less theoretical complexity as compared to others in training the HCRFs model. As shown in previous section, our proposed approach has a linear complexity while the exiting methods are of quadratic complexity. This extends the functionality of HCRF and enables it to be used in more practical and scalable applications. Although the scope of this paper is restricted to audio-based emotion recognition, however, it is completely possible to extend it to other related ares of recognition including speech recognition, acoustic based context awareness, and gesture recognition among others.

References

[1] Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG. Emotion recognition in human-computer interaction. IEEE Signal Process Mag 2001;18(1):32–80.

[2] Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1. IEEE; 2004. pp. I–577.
[3] Tacconi D, Mayora O, Lukowicz P, Arnrich B, Setz C, Troster G, Haring C. Activity and emotion recognition to support early diagnosis of psychiatric diseases. In: 2008 Second International Conference on Pervasive Computing Technologies for Healthcare. IEEE; 2008. p. 100–2.
[4] Rahman MA, Hossain MF, Hossain M, Ahmmed R. Employing pca and t-statistical approach for feature extraction and classification of emotion from multichannel eeg signal. Egypt Inf J.
[5] Alsayat A, Elmitwally N. A comprehensive study for arabic sentiment analysis (challenges and applications). Egypt Inf J.
[6] Nalini N, Palanivel S. Music emotion recognition: the combined evidence of mfcc and residual phase. Egypt Inf J 2016;17(1):1–10.
[7] El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn 2011;44 (3):572–87.
[8] Bitouk D, Verma R, Nenkova A. Class-level spectral features for emotion recognition. Speech Commun 2010;52(7–8):613–25.
[9] Iliev AI, Scordilis MS, Papa JP, Falcão AX. Spoken emotion recognition through optimum-path forest classification using glottal features. Comput Speech Language 2010;24(3):445–60.
[10] Lee CM, Narayanan SS, et al. Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 2005;13(2):293–303.
[11] Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. J Personality Soc Psychol 1996;70(3):614.
[12] Gobl C, Chasaide AN. The role of voice quality in communicating emotion, mood and attitude. Speech Commun 2003;40(1–2):189–212.
[13] Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden markov models. Speech Commun 2003;41(4):603–23.
[14] Teager H. Some observations on oral air flow during phonation. IEEE Trans Acoust Speech Signal Process 1980;28(5):599–601.
[15] Xue Y, Xue B, Zhang M. Self-adaptive particle swarm optimization for large-scale feature selection in classification. ACM Trans Knowledge Discovery Data (TKDD) 2019;13(5):50.
[16] Cairns DA, Hansen JH. Nonlinear analysis and classification of speech under stressed conditions. J Acoust Soc Am 1994;96(6):3392–400.
[17] Fu L, Mao X, Chen L. Speaker independent emotion recognition based on svm/hmms fusion system. In: 2008 International Conference on Audio, Language and Image Processing. IEEE; 2008. p. 61–5.
[18] Lee CM, Narayanan SS, Pieraccini R. Combining acoustic and language information for emotion recognition. In: Seventh International Conference on Spoken Language Processing.
[19] Otsuka T, Ohya J. Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences. In: Proceedings of International Conference on Image Processing, vol. 2, IEEE; 1997. p. 546–49.
[20] Schuller B, Rigoll G, Lang M. Hidden markov model-based speech emotion recognition. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03), vol. 2, IEEE; 2003, p. II–1.
[21] Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features, and methods. Speech Commun 2006;48(9):1162–81.
[22] Womack BD, Hansen JH. N-channel hidden markov models for combined stressed speech classification and recognition. IEEE Trans Speech Audio Process 1999;7(6):668–77.
[23] Gunawardana A, Mahajan M, Acero A, Platt JC. Hidden conditional random fields for phone classification. In: Ninth European Conference on Speech Communication and Technology.
[24] Wang SB, Quattoni A, Morency L-P, Demirdjian D, Darrell T. Hidden conditional random fields for gesture recognition. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2. IEEE; 2006. p. 1521–7.
[25] Quattoni A, Wang S, Morency L-P, Collins M, Darrell T. Hidden conditional random fields. IEEE Trans Pattern Anal Mach Intell 2007;10:1848–52.
[26] Farzaneh-Gord M, Mohseni-Gharyehsafa B, Arabkoohsar A, Ahmadi MH, Sheremet MA. Precise prediction of biogas thermodynamic properties by using ann algorithm. Renewable Energy 2020;147:179–91.
[27] Ramezanizadeh M, Ahmadi MH, Nazari MA, Sadeghzadeh M, Chen L. A review on the utilized machine learning approaches for modeling the dynamic viscosity of nanofluids. Renew Sustain Energy Rev 2019;114:109345.
[28] Kahani M, Ahmadi MH, Tatar A, Sadeghzadeh M. Development of multilayer perceptron artificial neural network (mlp-ann) and least square support vector machine (lssvm) models to predict nusselt number and pressure drop of tio2/water nanofluid flows through non-straight pathways. Numer Heat Transfer, Part A: Appl 2018;74(4):1190–206.
[29] Baghban A, Kahani M, Nazari MA, Ahmadi MH, Yan W-M. Sensitivity analysis and application of machine learning methods to predict the heat transfer performance of cnt/water nanofluid flows through coils. Int J Heat Mass Transf 2019;128:825–35.
[30] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging. In: Conference on Empirical Methods in Natural Language Processing.
[31] McCallum A, Freitag D, Pereira FC. Maximum entropy markov models for information extraction and segmentation. ICML 2000;17:591–8.
[32] Kuo H-KJ, Gao Y. Maximum entropy direct models for speech recognition. IEEE Trans Audio, Speech, Language Process 2006;14(3):873–81.
[33] Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
[34] Siddiqi MH, Ali R, Khan AM, Park Y-T, Lee S. Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. IEEE Trans Image Process 2015;24(4):1386–98.
[35] Reiter S, Schuller B, Rigoll G. Hidden conditional random fields for meeting segmentation. In: 2007 IEEE International Conference on Multimedia and Expo. IEEE; 2007. p. 639–42.
[36] Mahajan M, Gunawardana A, Acero A. Training algorithms for hidden conditional random fields. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1, IEEE; 2006. p. I–I.
[37] Siddiqi MH, Alruwaili M, Ali A, Alanazi S, Zeshan F. Human activity recognition using gaussian mixture hidden conditional random fields. Computat Intell Neurosci 2019.
[38] Lee S, Lee Y-K, et al. Emotional speech classification using hidden conditional random fields. In: Proceedings of the Second Symposium on Information and Communication Technology. ACM; 2011. p. 146–50.
[39] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of german emotional speech. In: Ninth European Conference on Speech Communication and Technology.
[40] Martin O, Adell J, Huerta A, Kotsia I, Savran A, Sebbe R. Multimodal caricatural mirror. In: eINTERFACE'05-Summer Workshop on Multimodal Interfaces; 2005.
[41] Lotz AF, Faller F, Siegert I, Wendemuth A. Emotion recognition from disturbed speech-towards affective computing in real-world in-car environments. Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018;2018:208–15.
[42] Zamil AAA, Hasan S, Baki SMJ, Adam JM, Zaman I. Emotion detection from speech signals using voting mechanism on classified frames. In: 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE; 2019. p. 281–5.
[43] Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Mahjoub MA. Speech emotion recognition: methods and cases study. In: ICAART (2); 2018. p. 175–182.
[44] Tursunov A, Kwon S, Pang H-S. Discriminating emotions in the valence dimension from speech using timbre features. Appl Sci 2019;9(12):2470.
[45] Choudhury AR, Ghosh A, Pandey R, Barman S. Emotion recognition from speech signals using excitation source and spectral features, in IEEE Applied Signal Processing Conference (ASPCON). IEEE 2018;2018:257–61.
[46] Bhavan A, Chauhan P, Shah RR, et al. Bagged support vector machines for emotion recognition from speech. Knowl-Based Syst 2019;184:104886.
[47] Avots E, Sapiński T, Bachmann M, Kamińska D. Audiovisual emotion recognition in wild. Mach Vis Appl 2019;30(5):975–85.
[48] Hajarolasvadi N, Demirel H. 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. Entropy 2019;21(5):479.
[49] Ma Y, Hao Y, Chen M, Chen J, Lu P, Košir A. Audio-visual emotion fusion (avef): a deep efficient weighted approach. Inf Fusion 2019;46:184–92.