Gaussian Mixture Hidden Conditional Random Fields for Emotional Speech Classification

La The Vinh

Hanoi University of Science and Technology, No. 1, Dai Co Viet, Hai Ba Trung, Hanoi, Viet Nam

Received: January 08, 2015; accepted: April 24, 2016

Abstract

In this study, we investigate in the use of hidden conditional random fields model to classify emotional speech. We introduce a novel hidden conditional random fields model, which is able to approximate complex distributions using a mixture of full covariance Gaussian density functions. In our experiments, we extracted Mel-frequency cepstral coefficients (MFCC) features from the well-known Berlin emotional speech dataset and eNTERFACE 2005 dataset. After that, we used the 10-fold cross validation rule to train, evaluate and compare our proposed model with the conventional learning method, hidden Markov model (HMM) and the existing hidden conditional random fields model, which can only utilize diagonal covariance Gaussian distributions. The experiments show that our method achieves significant improvement (p-value < 0.05) regarding the classification accuracy.

Keywords: Emotion classification, Conditional Random Fields, HMM, GMM

1. Introduction

Emotion is a mental state that arises spontaneously. In daily life, emotion is not only an effective way to convey our intention in communication but also a good indicator of our mental health. That is the reason why automatic detection of human emotions is an important factor to enhance the quality of the service provided by the computer such as human-computer interaction (HCI) [1, 2], lifestyle monitoring in ubiquitous health care systems [3]. While human emotion can be expressed by a variety of physiological changes such as speech, blood pressure, heart rate, facial expression, etc; many researchers prefer acoustic speech as a source of emotion [4, 5, 6, 7] because speech signal is the most commonly used and most natural method of human communication.

A speech-based emotion classification system comprises of two stages: a signal processing unit that extracts features from the input speech data, and a classifier that decides the emotion label of the corresponding input. Regarding feature extraction techniques, there is quite a large number of speech features that has been proposed for emotion recognition. These features can be divided into four main categories namely continuous speech features (such as pitch, formant, energy) [1, 7, 8], voice quality features [1, 9] (such as harsh, tense, breathy), spectral features [8, 10] (such as linear prediction coefficients, Mel-frequency cepstrum coefficients), and Teager energy operator (TEO) [11]. In a comprehensive survey of speech-based emotion recognition systems [4], the authors recommended that the choice of proper features highly depends on the classification task being considered. The authors also pointed out that Mel-frequency cepstrum coefficients (MFCC) are promising features for speech representation. We utilize existing methods to extract the spectral features (MFCC) to be used with our proposed classifier.

While there is quite a large number of recent researchers focusing on improving the feature extraction stage [2, 5, 6, 8, 9], almost all the proposed speech emotion classification systems utilize conventional learning methods [1] such as hidden Markov model (HMM), Gaussian mixture model (GMM), support vector machine (SVM), artificial neural networks (ANN), etc. Among these classifiers, HMM is pointed out by several studies [10, 12-18] to be the most commonly used method. However, some recent research in other areas such as speech recognition [19], gesture recognition [20, 21], showed that HMM, which is a generative learning model, is less accurate than its discriminative counterpart, the hidden conditional random fields model (HCRF).

Motivated by the lack of improvement in learning model, we investigated in the use of HCRF for the speech-based emotion classification problem. Furthermore, in our work we showed that the existing HCRF model is limited by independence assumptions, which may decrease the classification accuracy. Therefore, we proposed a novel algorithm to relax the assumption making our model capable of

* Corresponding author: Tel.: (+84) 985.260.681
Email: vinh.lathe@hust.edu.vn

using full-covariance distribution. In the rest of our paper, we present our proposed method in section 2. Experiments and discussions are presented in section 3. Finally, we conclude our paper in section 4.

2. The proposed method

In the area of pattern classification, conditional random fields (CRF) [22] and hidden conditional random fields (HCRF) [19, 21], generalizations of maximum-entropy Markov model (MEMM), are proposed to fully take the advantages of MEMM and to solve the *label bias problem*. HCRF extends the capability of CRF with hidden states making it able to learn hidden structure of sequential data. Both of them use global normalization instead of per-state normalization. Thus, they allow weighted scores, making the parameter spaces larger than those of MEMM and HMM. In the following paragraphs, we will briefly review the background theory of HCRF and clearly analyze the limitations in the existing implementations.

We consider a task of mapping from inputs X to labels $Y \in \Gamma$ in a speech-based emotion classification problem. Each input X is a sequence of T feature vectors, $X = \{x_1,x_2,...,x_T\}$. Our training set contains N pairs $(X_i,Y_i), i = 1,2,...,N$. In a Q-state HCRF, the conditional probability of a class label Y given input X and the model's parameter set Λ will be computed by

$$p(Y \mid X;\Lambda) = \frac{\sum_{\bar{S}} exp\{\Lambda.f(Y,\bar{S},X)\}}{z(X,\Lambda)} \quad (1),$$

where $$z(X,\Lambda) = \sum_{Y'} p(Y' \mid X;\Lambda) \quad (2),$$

is the normalization factor to guarantee the sum-to-one rule of the conditional probability. In (1) and (2), $\bar{S} = \{s_1,s_2,...,s_T\}$ is a sequence of hidden states, each s_i, $i = 1,2,...,T$, can have an integer value from 1 to Q, the number of states, Λ is the parameter vector and $f(Y,S,X)$ is a feature vector which decides what statistics will be learnt by the model. The choice of the feature vector determines the dependencies of the HCRF model. For example the below selections will form a Markov chain HCRF with a Gaussian distribution at each state.

$$f_s^{Prior}(Y,\bar{S},X) = \delta(s_1 = s)\forall s \quad (3),$$

$$f_{ss'}^{Transition}(Y,\bar{S},X) = \sum_{t=1}^{T}\delta(s_{t-1} = s)\delta(s_t = s')\forall s,s' \quad (4),$$

$$f_s^{Occurence}(Y,\bar{S},X) = \sum_{t=1}^{T}\delta(s_t = s)\forall s \quad (5),$$

$$f_s^{M1}(Y,\bar{S},X) = \sum_{t=1}^{T}\delta(s_t = s)x_t\forall s \quad (6),$$

$$f_s^{M2}(Y,\bar{S},X) = \sum_{t=1}^{T}\delta(s_t = s)x_t^2\forall s \quad (7),$$

where $\delta(A)$ is one if the A is true and zero otherwise.

Although, there has been other existing work that utilizes the above HCRF model and shows good results [23, 24]; they, however, did not address and overcome the limitations of the model. As we can see in the above equations, x^2_t is a per-component squares of x_t; it means that the model can only utilize diagonal-covariance Gaussian distribution. In another word, the variables (columns of x_i, $i = 1,2,..,N$) are assumed to be pair-wise independent. Hereafter, we call this model *diagonal covariance Gaussian mixture hidden conditional random fields* (DCGM-HCRF). In addition, it can be seen that with a particular set of value, the observation density at each state will converge to Gaussian form. Unfortunately, there is not any training algorithm designed to guarantee this convergence. Therefore, those assumptions may result an incorrect measurement of accuracy. In order to inherit the advantages of the HCRF model and completely tackle the limitations of the existing work, we propose our novel HCRF algorithm which is able to explicitly utilize mixture of *full-covariance Gaussian distributions* (FCGM-HCRF) by explicitly including a mixture of Gaussian distributions in the feature functions, thus our feature functions are described in the following forms:

$$f_s^{Observation}(Y,\bar{S},X) = \sum_{t=1}^{T} log\left(\sum_{m=1}^{M}\Gamma_{s,m}^{Obs}N(x_t,\mu_{s,m},\Sigma_{s,m})\right)\delta(s_t = s) \quad (8),$$

$$N(x,\mu_{s,m},\Sigma_{s,m}) = \frac{1}{(2\pi)^{\frac{D}{2}}\mid\Sigma_{s,m}\mid^{\frac{1}{2}}}\times exp\left(-\frac{1}{2}(x-\mu_{s,m})'\Sigma_{s,m}^{-1}(x-\mu_{s,m})\right) \quad (9),$$

where N is the density function of the multivariate Gaussian distribution [28], M is the number of density functions, D is the dimension of the observation, Γ^{Obs} is the mixing weight of the m^{th} component with mean $\mu_{s,m}$ and covariance matrix $\Sigma_{s,m}$.

In the training phase, our goal is to find the parameters (Λ, Γ, μ, and Σ) to maximize the conditional probability of the training data. In our work, we utilize L-BFGS method implemented in [25] to search for the optimal point. Because the optimization algorithm is beyond the scope of this paper, we do not mention the detail of L-BFGS which

is handled by [25]. What we contribute here is how the full-covariance Gaussian mixture model is embedded using our feature functions.

3. Results and discussions

In our experiments, we use two well-known datasets namely *Berlin emotional speech* dataset [26] and *eNTERFACE 2005* multi-modal emotion dataset [27]. With each dataset, Mel-frequency cepstral coefficients (MFCCs) are extracted, then the training and testing data are built based on the 10-fold cross validation rule. We perform the classification experiments with our own algorithm (FCGM-HCRF) and two others including the hidden Markov model (HMM), and the existing hidden conditional random fields model which uses diagonal-covariance Gaussian mixtures (DCGM-HCRF). Then we utilize the paired t-test to calculate p-values in order to compare our algorithm with the others. Details of the experiments will be presented in the following paragraphs.

Berlin dataset contains emotional utterances produced by 10 German actors (5 males and 5 females) reading ten pre-defined sentences in one of seven emotion states namely anger, joy, sadness, fear, disgust, boredom, and neutral. Each recording was evaluated by 20 judges, and only those recognized by at least 80% of the listeners were kept.

We first run the experiments using HMM with different numbers of states and Gaussian mixtures. Table 1 shows the average classification rates of 10 folds for each pair of state number and mixture number. From the table, we can see that a HMM which has 2 states and 6 mixtures produces the highest accuracy. We apply those values to train and evaluate the other two algorithms (FCGM-HCRF, and DCGM-HCRF) then compare the results with the baseline method HMM as depicted in Fig. 1.

Table 1. 10-fold average accuracy (%) of hidden Markov model with different state number (rows) and mixture number (columns)

	2	4	6	8
2	70.57	69.99	**73.78**	72.66
4	71.08	72.09	72.16	67.22
6	71.70	71.15	63.64	61.58
8	71.88	67.22	63.49	61.00

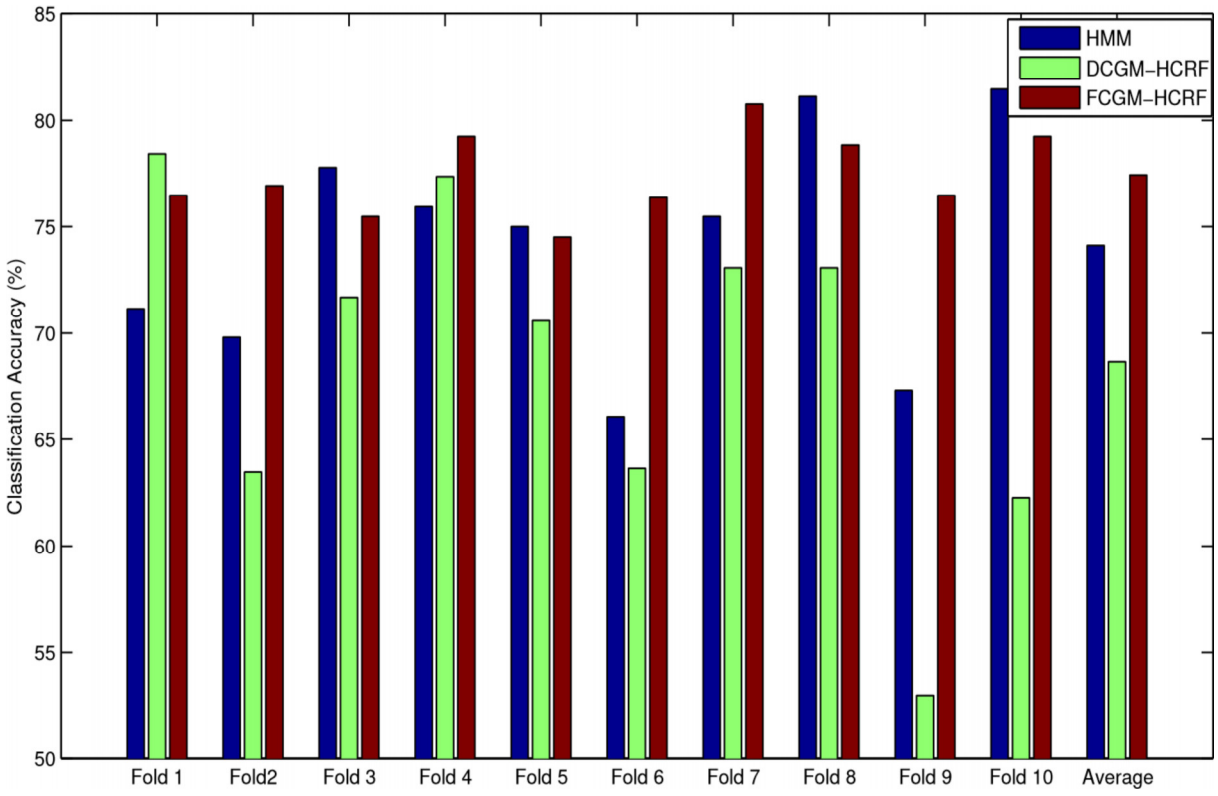

Fig. 1. Classification accuracies of three methods (2 states, 6 mixtures) with Berlin dataset

As can be seen, our method (FCGM-HCRF) achieves the highest average accuracy (77.43% compared to 74.11% and 68.65% of HMM and DCGM-HCRF respectively) and the lowest deviation (2% compared to 5% and 8% of the others). In addition, the p-values when we compare our algorithm with HMM and DCGM-HCRF are 0.05 and 0.01, respectively. It means that our method is significantly better than the other two. Table 2 shows the confusion matrix of the FCGM-HCRF method. From the table it can be seen that Neutral emotion is the best recognized, classification of emotions with strong expression like fear, anger, and happiness is also better than the others.

eNTERFACE 2005 emotional speech dataset eNTERFACE dataset contains 1320 videos produced by 44 subjects. Each subject simulated 6 emotion states (anger, disgust, fear, happiness, sadness, and surprise) by reading 5 predefined sentences. We separate audio data from those videos, then extract Mel-frequency cepstral coefficients (MFCCs) to construct training and testing data as what we did with Berlin dataset. All the above steps are repeated with eNTERFACE dataset, the result is depicted in Fig. 2. From Fig. 2 we can see that the average accuracy of HMM is the lowest (44.55%), DCGM-HCRF and FCGM-HCRF achieve 48.34% and 47.18%, respectively. Obviously the two HCRF methods are significantly better than HMM evidenced by p-values of 0.02 (compared to FCGM-HCRF) and 0.005 (compared to DCGM-HCRF). A p-value of 0.22 indicates an insignificant difference between the two HCRF algorithms.

4. Conclusion

From our experiment results, it is clear that by proposing to use full-covariance Gaussian density functions, our classification accuracy is higher than those of the others. We also prove that such improvement is significant from statistical point of view by showing p-values of the comparisons. Although in this research, we limit our application to audio-based emotion recognition, it is completely possible to use our proposed model in other similar areas like speech recognition, acoustic based context awareness, gesture recognition, etc.

Table 2. Confusion matrix (in %) of the Berlin dataset classified by the FCGM - HCRF model

	A	B	D	F	H	S	N
Anger (A)	**81.2**	0.1	0.2	9.1	8.2	0.0	1.2
Boredom (B)	0.1	**69.0**	6.6	0.3	0.0	14.8	9.2
Disgust (D)	0.2	10.1	**73.3**	1.1	0.5	5.1	9.7
Fear (F)	8.1	0.0	0.9	**82.1**	6.7	0.1	2.1
Happiness (H)	9.9	0.2	0.5	7.4	**80.4**	0.0	1.6
Sadness (S)	0.3	13.0	7.6	0.0	0.0	**71.1**	8.0
Neutral (N)	0.0	4.1	4.3	0.0	0.0	6.7	**84.9**

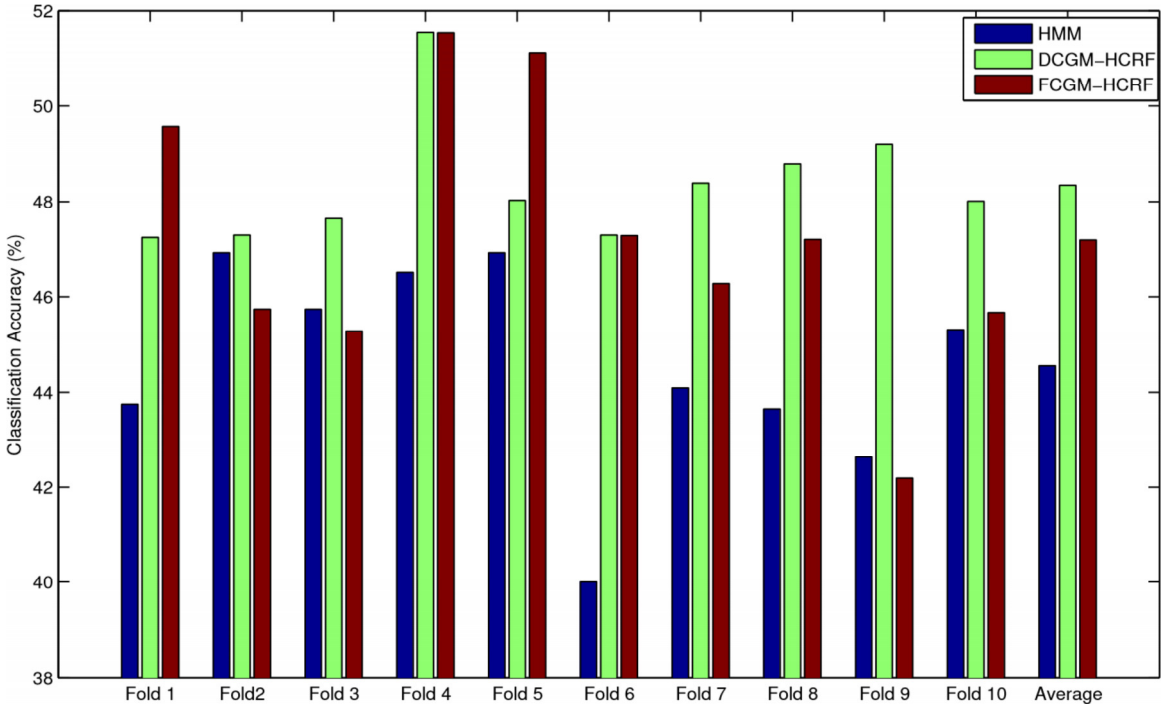

Fig. 2. Classification accuracies of three methods (2 states, 6 mixtures) with eNTERFACE dataset

References

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

[2] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, pp. 577 – 580, 2004.

[3] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troster, C. Haring, Activity and emotion recognition to support early diagnosis of psychiatric diseases, in: Proceedings of the Second International Conference on Pervasive Computing Technologies for Healthcare, pp. 100 – 102, 2008.

[4] M. E. Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (2011) 572 – 587.

[5] D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition, Speech Communication 52 (2010) 613 – 625.

[6] A. I. Iliev, M. S. Scordilis, J. P. Papa, A. X. Falco, Spoken emotion recognition through optimum-path forest classification using glottal features, Computer Speech and Language 24 (2010) 445 – 460.

[7] C. M. Lee, S. Narayanan, Towards detecting emotions in spoken dialogs, IEEE Transactions on Speech and Audio Processing 13 (2005) 293 – 303.

[8] K. R. Banse, R. Scherer, Acoustic profiles in vocal emotion expression, Journal of Personality and Social Psychology 70 (1996) 614 – 636.

[9] C. Gobl, A. N. Chasaide, The role of voice quality in communicating emotion, mood and attitude, Speech Communication 40 (2003) 189 –212.

[10] T. L. Nwe, S. W. Foo, L. C. D. Silva, Speech emotion recognition using hidden markov models, Speech Communication 41 (2003) 603 – 623.

[11] H. Teager, Some observations on oral air flow during phonation, IEEE Transactions on Acoustics, Speech and Signal Processing 28 (1980) 599 – 601.

[12] D. A. Cairns, J. H. L. Hansen, Nonlinear analysis and classification of speech under stressed conditions, Journal of the Acoustical Society of America 96 (1994) 3392 – 3400.

[13] L. Fu, X. Mao, L. Chen, Speaker independent emotion recognition based on svm/hmms fusion system, in: Proceedings of the International Conference on Audio, Language and Image Processing, pp. 61 – 65, 2008.

[14] C. M. Lee, S. S. Narayanan, R. Pieraccini, Combining acoustic and language information for emotion recognition, in: Proceedings of the International Conference on Spoken Language Processing, pp. 873 – 876, 2002.

[15] T. Otsuka, J. Ohya, Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences, in: Proceedings of the International Conference on Image Processing, volume 2, pp. 546 – 549, 1997.

[16] B. Schuller, G. Rigoll, M. Lang, Hidden markov model-based speech emotion recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pp. 1 – 4, 2003.

[17] D. Ververidis, C. Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech Communication 48 (2006) 1162 – 1181.

[18] B. Womack, J. Hansen, N-channel hidden markov models for combined stressed speech classification and recognition, IEEE Transactions on Speech and Audio Processing 7 (1999) 668 – 677.

[19] A. Gunawardana, M. Mahajan, A. Acero, J. C. Platt, Hidden conditional random fields for phone classification, in: Proceedings of the International Conference on Speech Communication and Technology, pp. 1117 – 1120, 2009.

[20] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, T. Darrell, Hidden conditional random fields for gesture recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 2, pp. 1521 – 1527, 2006.

[21] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, T. Darrell, Hidden conditional random fields, Pattern Analysis and Machine Intelligence, IEEE Transactions on 29 (2007) 1848 – 1852.

[22] J. Lafferty, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann, 2001, pp. 282 – 289.

[23] S. Reiter, B. Schuller, G. Rigoll, Hidden conditional random fields for meeting segmentation, in: Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 639 – 642.

[24] M. Mahajan, A. Gunawardana, A. Acero, Training algorithms for hidden conditional random fields, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume 1, p. I, 2006.

[25] M. Schmidt, minfunc: Matlab toolbox for unconstrained optimization of differentiable real-valued multivariate functions, http://www.di.ens.fr/mschmidt/software/minfunc.html .

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, A database of german emotional speech, in: Proceedings of the 9th European Conference on Speech Communication and Technology, pp. 1517 – 1520, 2005.

[27] M. O., A. J., H. A., K. I., S. A., S. R., Multimodal caricatural mirror, in: Proceedings of the SIMILAR NoE Summer Workshop on Multimodal Interfaces, pp. 13 – 20. Du, S.M. George, Thickness dependence of sensor response for CO gas sensing by tin oxide films grown using atomic layer deposition, Sens. Actuators B 135 (2008) 152-160.

[28] Walck, Christian, Handbook on statistical distributions for experimentalist, MLA Publisher, 2007.