Emotional Speech Classification using Hidden Conditional Random Fields

La The Vinh
Dept. of Computer Engineering, Kyung Hee University
446-701, Seocheon-dong, Giheung-gu Yongin-si, Gyeonggi-do, Korea
vinhlt@oslab.khu.ac.kr

Sungyoung Lee
Dept. of Computer Engineering, Kyung Hee University
446-701, Seocheon-dong, Giheung-gu Yongin-si, Gyeonggi-do, Korea
sylee@oslab.khu.ac.kr

Young-Koo Lee
Dept. of Computer Engineering, Kyung Hee University
446-701, Seocheon-dong, Giheung-gu Yongin-si, Gyeonggi-do, Korea
yklee@khu.ac.kr

ABSTRACT

Although there have been a great number of papers in the area of emotional speech recognition, most of them contribute to the feature extraction phase. Regarding classification algorithm, hidden Markov model (HMM) is still the most commonly used method. Whereas HMM was pointed out to be less accurate than its discriminative counterpart, the hidden conditional random fields (HCRF) model, for example in phone classification or gesture recognition. Therefore in this study, we investigate the use of the HCRF model in emotional speech classification problem. In our experiments, we extracted Mel-frequency cepstral coefficients (MFCC) features from the well-known Berlin emotional speech dataset (EMO) and eNTERFACE 2005 dataset. After that, we used the 10-fold cross validation rule to train, evaluate and compare our result with that of HMM. The experiments show that HCRF achieves significant improvement (p-value ≤ 0.05) in classification accuracy. In addition, we speed up the training phase of the model by caching the gradient computation. Therefore our computation time is much less than that of the existing methods.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.5.1 [**Pattern Recognition**]: Statistical

General Terms

Emotion Classification

Keywords

HCRF, HMM, MFCC, Emotional Speech

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) [5] or lifestyle monitoring in ubiquitous health care systems [16]. 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 [1, 2, 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 are quite a large number of speech features that have been proposed for emotion recognition. In a comprehensive survey of speech-based emotion recognition systems [1], the authors recommended that the choice of proper features highly depends on the classification task being considered. The authors also pointed that Mel-frequency cepstrum coefficients (MFCC) are promising features for speech representation. Since feature extraction algorithms are beyond the scope of this paper, we utilize existing methods to extract the spectral features (MFCC) to be used with our proposed classifier.

While there are quite a large number of recent researches focusing on improving the feature extraction stage, almost all the proposed speech emotion classification systems utilize conventional learning methods [5] 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 [4, 11] to be the most commonly used method. However, some recent research in other areas such as speech recognition [6], gesture recognition [13], showed that HMM, which is a generative learning model, is less accurate than its discriminative counterpart, the hidden conditional random fields model (HCRF).

SoICT 2011, October 13-14, 2011, Hanoi, Vietnam.

Motivated by the lack of improvement in learning model, we have proposed our Gaussian mixture HCRF for the speech-based emotion classification problem. In the rest of our paper, we review some related work in section 2. Our proposed method is introduced in section 3. Experiments and discussions are presented in section 4. Finally, we conclude our paper in section 5.

2. RELATED WORK

In the recent years, a significant number of methods have been proposed to enhance the accuracy of speech-based emotion recognition systems [5]. Surprisingly, few of them contributed to the improvement of the learning model for the speech data. As pointed out in [5], even different kinds of classifiers have been applied to solve the problem, HMM is still the most commonly used method, and produces accuracies that are comparable to other well-known classifiers such as GMM, ANN, SVM, etc. In addition, HMM has its own advantage of handling sequential data when frame-level features are used. In such a case, other vector-based classifiers like GMM, ANN, SVM, are not able to learn the sequence of feature vectors.

However, the limitations of HMM are also clearly addressed in [6, 13], which are the generative nature and the independence assumption between states and observations. Maximum entropy Markov model (MEMM), a non-generative model, is proposed to overcome the limitations of HMM and shows good results for tasks such as part-of-speech (POS) tagging [14], information extraction [10], and automatic speech recognition (ASR) [8]. Nevertheless, MEMM has a commonly known weakness that is called *label bias problem* [9] because it uses per-state normalization of transition scores, implying a score conservation at each state.

Conditional random fields (CRF) [9] and hidden conditional random fields (HCRF) [6, 13], generalizations of MEMM, are then 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 certainly HMM.

However, the existing HCRF model [6] is capable of handling only diagonal-covariance Gaussian mixtures. Therefore, we overcome that limitation by proposing a full covariance Gaussian mixture HCRF model. Details of our method will be presented in the following section.

3. THE PROPOSED APPROACH

We consider a task of mapping from inputs X to labels $Y \in \Gamma$, for example $\Gamma = \{angry, happy, neutral\}$ 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 and M-mixture HCRF, the conditional probability of a class label Y given input X and the model's parameter set Λ will be computed by

$$p(Y|X;\Lambda) = \frac{\sum_{\overline{S}} exp\left\{\Lambda f(Y,\overline{S},X)\right\}}{z(X,\Lambda)}, \quad (1)$$

where

$$z(X,\Lambda) = \sum_{Y'} P(Y'|X;\Lambda) \quad (2)$$

is the normalization factor to guarantee the sum-to-one rule of the conditional probability and M is the number of Gaussian density functions used in the mixture. In (1), $\overline{S} = \{s_1, s_2, ..., s_T\}$ 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 $f(Y, \overline{S}, X)$ is a feature vector which decides what statistics will be learnt by the model.

The choice the feature vector determines the dependencies of the HCRF model. To compare the performance of the HCRF model to that of HMM, we use the below selections to form a Markov chain HCRF with a Gaussian mixture distribution at each state.

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

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

$$f_s^{Observation}(Y,\overline{S},X) =$$

$$\sum_{t=1}^{T} log\left(\sum_{m=1}^{M} \Gamma_{s,m}^{Obs} N(x_t, \mu_{s,m}, \Sigma_{s,m})\right)\delta(s_t = s), \quad (5)$$

$$N(x, \mu_{s,m}, \Sigma_{s,m}) =$$

$$\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_{s,m}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu_{s,m})'\Sigma_{s,m}^{-1}(x-\mu_{s,m})\right), \quad (6)$$

where M is the number of density functions, D is the dimension of the observation, $\Gamma_{s,m}^{Obs}$ is the mixing weight of the m^{th} component with mean $\mu_{s,m}$ and covariance matrix $\Sigma_{s,m}$. As we can see in (5), by changing Γ, μ, and Σ we can create any mixture of the normal densities. Therefore, the corresponding observation weight (Λ_s^{Obs}) is not necessary to be updated during the training phase, hence we can set

$$\Lambda_s^{Obs} = 1\forall s. \quad (7)$$

As the result, the conditional probability can be rewritten as below

$$p(Y|X;\Lambda,\Gamma,\mu,\Sigma) = \frac{\sum_{\overline{S}} exp\left(P(\overline{S}) + T(\overline{S}) + O(\overline{S})\right)}{z(X,\Lambda,\Gamma,\mu,\Sigma)}, \quad (8)$$

where

$$P(\overline{S}) = \sum_s \Lambda_s^{Prior} f_s^{Prior}(Y,\overline{S},X), \quad (9)$$

$$T(\overline{S}) = \sum_{ss'} \Lambda_{ss'}^{Transition} f_{ss'}^{Transition}(Y,\overline{S},X), \quad (10)$$

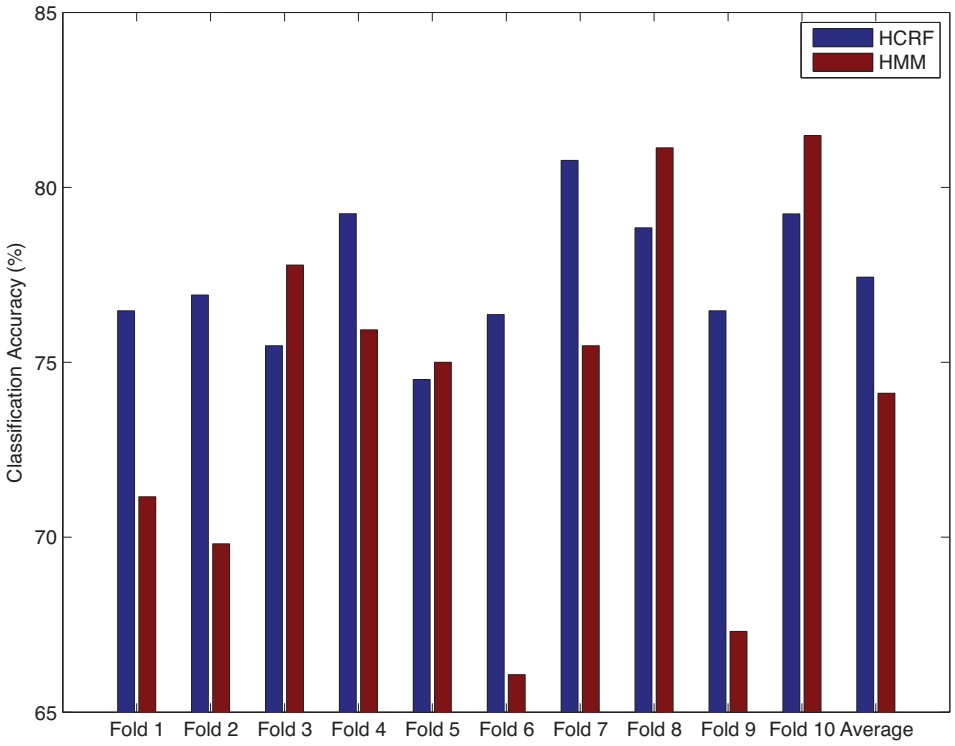

Figure 1: Classification accuracies of two methods (2 states, 3 mixtures) with EMO dataset

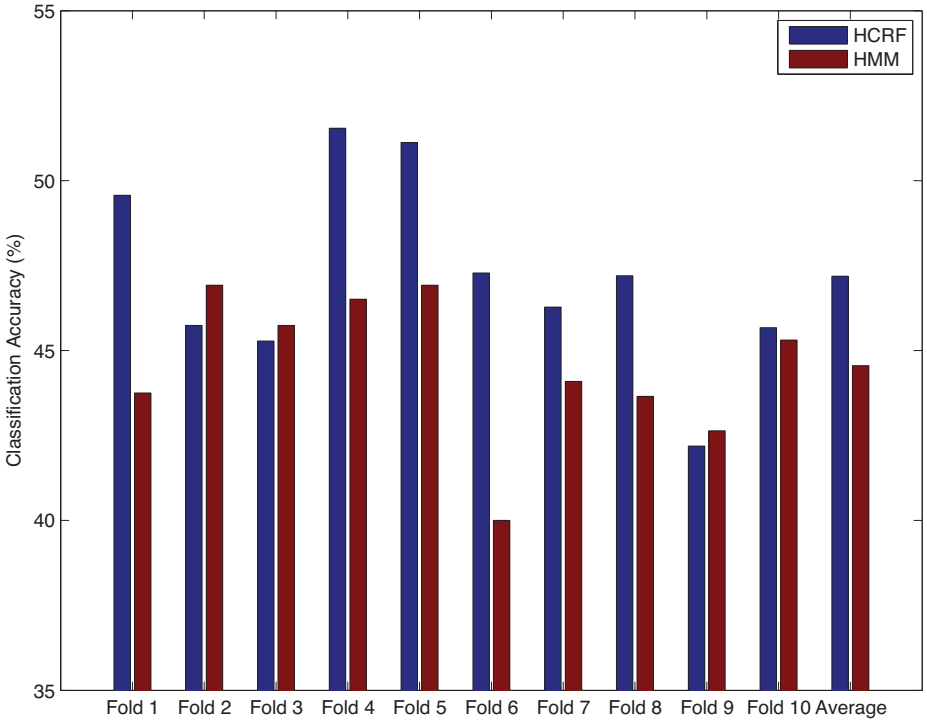

Figure 2: Classification accuracies of two methods (2 states, 6 mixtures) with eNTERFACE dataset

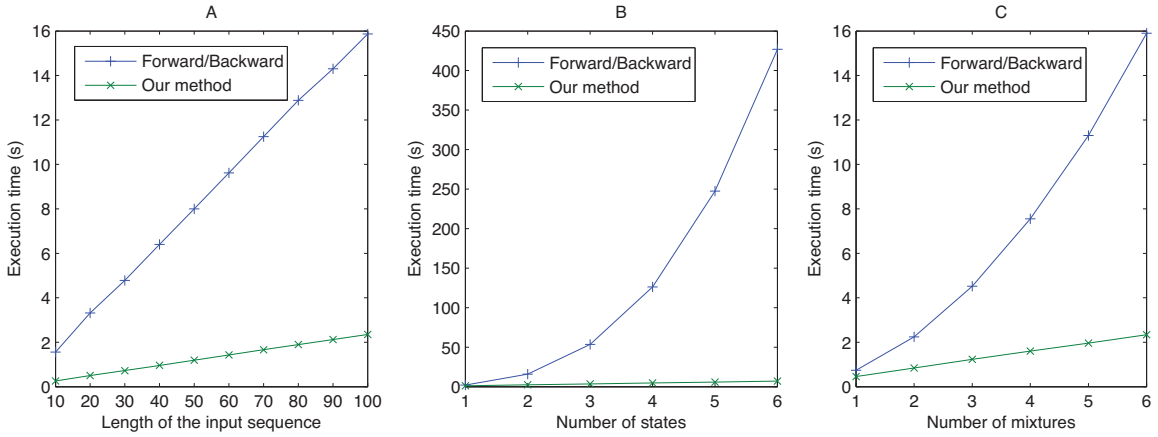

Figure 3: An example of the gradient computation time of the existing forward/backward and our method. Figure A, Q=2, M=6, T=10-100. Figure B, Q=1-6, M=6, T=100. Figure C, Q=2, M=1-6, T=100

	1	3	5	7
2	56.28	74.11	71.88	72.28
4	63.66	72.63	72.64	69.45
6	65.28	72.51	68.52	63.26
8	65.71	71.92	65.31	60.66

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

and

$$O(\overline{S}) = \sum_{s} f_s^{Observation}(Y, \overline{S}, X), \tag{11}$$

Based on equations (8), (9), (10) and (11) we can compute the conditional probability by using the well-known forward and backward algorithms.

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 to search for the optimal point. However, instead of repeating the forward and backward algorithms to compute the gradients as others did [6], we run the forward and backward algorithms only when calculating the conditional probability, then we reuse the result to compute the gradients. As a result, the computation time is significantly reduced.

4. EXPERIMENTS

In our experiments, we use two well-known datasets namely *Berlin emotional speech* dataset [3] and *eNTERFACE 2005* multi-modal emotion dataset [12]. 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 performed the classification experiments with HCRF and HMM. Then we utilized the paired t-test to calculate p-values [15] in order to show the significance level.

4.1 Berlin emotional speech dataset

The dataset contains emotional utterances produced by 10 German actors (5 males and 5 females) reading ten predefined 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.

In our experiments, we conducted the classification with HMM of different state number and Gaussian mixture number. From the result (depicted in Table 1), we selected the values that maximize the accuracy (2 states and 3 mixtures). After that, we performed the classification with the HCRF model of the same state number and mixture number. The classification result is depicted in figure 1.

As can be seen, the average accuracy of HMM (74.11%) is lower than that of HCRF (77.43%). In addition, HCRF achieves more stable accuracies since the standard deviation is only 2%, which is clearly smaller than 5% of the HMM. Obviously HCRF is significantly better than HMM evidenced by a p-value of 0.05.

4.2 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 figure 2. In this experiment, HMM's accuracy (44.55%) is still lower than that of the HCRF model (47.18%). The difference between the two methods are significant since p-value = 0.02, which is much smaller than the 0.05 threshold.

Furthermore, while the existing implementations of HCRF compute the gradients by repeating the forward and backward algorithms, our method executes them once and caches the result for the later use. Therefore, the execution time is significantly reduced as shown in Figure 3 (the execution time is measured by Matlab R2008a running on a computer

with Intel Duo 2.6GHz and 2GB RAM).

5. CONCLUSION

From our results, it is clear that HCRF's accuracy is significantly higher than that of HMM ($p - value \leq 0.05$). Moreover, our computation method strongly decreases the execution time for training the HCRF model. This achievement will extend the use of HCRF model to more scalable applications. In the classification phase, we use the same computation method as the others do, hence the complexity is not different from that of the existing work.

6. ACKNOWLEDGMENTS

This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement)" (IITA-2009-(C1090-0902-0002)). This work also, was supported by the Korea Science & Engineering Foundation(KOSEF) grant funded by the Korea government(MEST) (No. 2008-1342), and was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2009-0076798). In addition, this work was supported by the IT R&D program of MKE/KEIT, [10032105, Development of Realistic Multiverse Game Engine Technology].

7. REFERENCES

[1] M. E. Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587, 2011.

[2] D. Bitouk, R. Verma, and A. Nenkova. Class-level spectral features for emotion recognition. *Speech Communication*, 52(7-8):613 – 625, 2010.

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

[4] D. A. Cairns and J. H. L. Hansen. Nonlinear analysis and classification of speech under stressed conditions. *Journal of the Acoustical Society of America*, 96(6):3392 – 3400, 1994.

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

[6] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt. Hidden conditional random fields for phone classification. In *Proceedings of the International Conference on Speech Communication and Technology*, pages 1117 – 1120, 2005.

[7] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falcão. Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech and Language*, 24(3):445 – 460, 2010.

[8] H.-K. J. Kuo and Y. Gao. Maximum entropy direct models for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):873 – 881, may 2006.

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

[10] A. McCallum, D. Freitag, and F. C. N. Pereira. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591 – 598, 2000.

[11] T. L. Nwe, S. W. Foo, and L. C. D. Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41(4):603 – 623, 2003.

[12] M. O., A. J., H. A., K. I., S. A., and S. R. Multimodal caricatural mirror. In *Proceedings of the SIMILAR NoE Summer Workshop on Multimodal Interfaces*, pages 13 – 20, 2005.

[13] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1848 – 1852, oct. 2007.

[14] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Empirical Methods in Natural Language Processing*, pages 133 – 142, 1996.

[15] M. J. Schervish. P values: What they are and what they are not. *The American Statistician*, 50(3):203 – 206, August 1996.

[16] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troster, and 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*, pages 100 – 102, 2008.