SPEECH EMOTION RECOGNITION VIA A MAX-MARGIN FRAMEWORK INCORPORATING A LOSS FUNCTION BASED ON THE WATSON AND TELLEGEN'S EMOTION MODEL

Sungrack Yun and Chang D. Yoo

Korea Advanced Institute of Science and Technology
Divison of Electrical Engineering, School of Electrical Engineering & Computer Science
2106, LG Semicon Hall, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea
email: yunsungrack@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper considers a method for speech emotion recognition by a max-margin framework incorporating a loss function based on a well-known model called the Watson and Tellegen's emotion model. Each emotion is modeled by a single-state hidden Markov model (HMM) that is trained by maximizing the minimum separation margin between emotions, and the margin is scaled by a loss function. The framework is optimized by the semi-definite programming. Experiments were performed to evaluate the framework using the Berlin database of emotional speech. The framework performed better than other conventional training criteria for HMM such as maximum likelihood estimation and maximum mutual information estimation.

Index Terms— Speech emotion recognition, max-margin framework, Watson and Tellegen's emotion model

1. INTRODUCTION

Research in the recognition of human emotion is one of growing research fields in human-machine interface (HMI) and affective computing [1]. Emotion recognition can be achieved by analyzing various modalities: speech and facial expressions [2], [3], gesture and body language [4] and bio information such as electrocardiogram, electromyography, electrodermal activity, skin temperature, blood volume pulse and respiration [5]. Compared to other modalities, speech signal can be obtained more easily and inexpensively. For this reason, it has a wider range of HMI applications: a service robot that responds to the owner's emotion, a computer game that controls the game status by game-player's emotion, and an audio response system of the call center that automatically connects the customer to the expert counsellor if the customer is angry.

A number of methods to recognize the speech emotion have been presented. Most methods recognize the emotion by extracting features such as fundamental frequency, log energy, mel-frequency cepstral coefficients (MFCCs), pitch and duration and recognizing selected features with various classifiers: support vector machine (SVM), hidden Markov models (HMMs), linear discriminant analysis, quadratic discriminant analysis and k-nearest neighbors [2], [6], [7].

This paper considers a max-margin framework incorporating a loss function, called margin scaling [8], [9] for emotion recognition. Most methods do not consider a loss function between emotions which quantifies the risk for predicting a label given the correct label. We adopt margin scaling to scale the minimum separation margin by the loss function. The distance metric between emotions is defined by the Watson and Tellegen's emotion model (WTM) [10], and the loss function is computed by the 1-norm of the distance metric. The max-margin framework is known to have a good generalization ability [11], [12] thus it performs well in many classification problems where there is a statistical mismatch between training and testing data set.

We represent each emotion by a single-state HMM and use the MFCCs for emotional features. The HMMs are trained by the margin scaling which maximizes the minimum separation margin scaled by a loss function. In the experiment, we show that our method performs better than the other conventional HMM training criteria such as maximum likelihood (ML) and maximum mutual information (MMI).

The outline of the paper is as follows. First, we introduce the emotion recognition using the HMMs in Section 2. Then, in Section 3, we explain the max-margin framework and the loss function based on the WTM. In Section 4, the performance comparisons are evaluated on the Berlin database of emotional speech (EMO-DB). Finally, we conclude and summarize the paper in Section 5.

2. EMOTION RECOGNITION USING HMMS

Emotion recognition is a classification that predicts a label $\mathbf{y}*$ from a given speech feature $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\} \in \mathcal{X}$ such that

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{X}, \mathbf{y}; \theta) \qquad (1)$$

where the label $\mathbf{y}$ represents one of M emotions $\mathcal{Y} = \{y_1, ..., y_M\}$, and F is a discriminant function $F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ which is parameterized by θ. The feature $\mathbf{X}$ consisting of T feature vectors which are in the D-dimensional vector space $\mathcal{X}$ is extracted from an utterance. Emotion recognition using HMMs is based on the framework of probabilistic graphical models where the conditional distribution $\log p_\theta(\mathcal{Y}|\mathcal{X})$ is a discriminant function. Thus, the decision criterion in Eq. (1) becomes

$$\begin{aligned} \mathbf{y}^* &= \arg\max_{\mathbf{y}\in\mathcal{Y}} \log p_\theta(\mathbf{y}|\mathbf{X}) \\ &= \arg\max_{\mathbf{y}\in\mathcal{Y}} \log p_\theta(\mathbf{X}|\mathbf{y})p(\mathbf{y}) \end{aligned} \quad (2)$$

where $p(\mathbf{y})$ is the prior probability of an emotion. We assume equal prior probability $1/M$ for all emotions.

We also assume that $\mathbf{X}$ represents a single emotion (not a sequence of emotions) and the statistical characteristic of $\mathbf{X}$ does not change over time. Thus, each emotion $\mathbf{y}$ is modeled by a single-state HMM. The state is modeled by Gaussian mixtures with its output probability for $\mathbf{x_t}$ given by

$$p_\theta(\mathbf{x}_t|\mathbf{y}) = \sum_{k=1}^{K} w_k N(\mathbf{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \quad (3)$$

where w_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ are the weight, mean vector and covariance matrix of the Gaussian k, respectively. The number of Gaussians is denoted by K, and $\sum_{k=1}^{K} w_k = 1$. Thus, the discriminant function can be expressed as

$$\begin{aligned} F(\mathbf{X}, \mathbf{y}; \theta) &= \log p_\theta(\mathbf{X}|\mathbf{y})p(\mathbf{y}) \quad (4) \\ &= \log \left[\frac{1}{M} \prod_{t=1}^{T} \sum_{k=1}^{K} w_k N(\mathbf{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \right] \end{aligned}$$

with the assumption that $\mathbf{x}_t$ is independent and identically distributed.

3. A MAX-MARGIN FRAMEWORK FOR EMOTION RECOGNITION

This section describes the method to estimate HMMs by a max-margin framework incorporating a loss function, called margin scaling. The estimation goal is to find θ using a set of input-output pairs $(\mathbf{X}_n, \mathbf{y}_n),\ n = 1, ..., N$ so that the decision criterion leads to the minimum prediction error. The parameter θ is a vector whose elements are the parameters of the Gaussian mixtures for all emotions. We adopt margin scaling to consider the loss function between emotions.

3.1. Formulation

The margin scaling finds the parameter vector θ so that the minimum separation margin ρ is maximized, and the sum of the slack variables ξ_n is minimized under the constraints that the difference between $F(\mathbf{X}_n, \mathbf{y}_n; \theta)$ given the correct label $\mathbf{y}_n$ and $F(\mathbf{X}_n, \mathbf{y}; \theta)$ given the incorrect label $\mathbf{y}$ is at least larger than the scaled margin subtracted by the slack variable for all $n = 1, ..., N$ as follows [8], [9], [13]:

$$\begin{aligned} \min_{\rho, \boldsymbol{\xi}, \theta : ||\theta|| = \gamma} \quad & -\rho + \frac{C}{N} \sum_{n=1}^{N} \xi_n \\ \text{subject to} \quad & d(\mathbf{X}_n, \mathbf{y}; \theta) \geq \rho\Delta(\mathbf{y}_n, \mathbf{y}) - \xi_n,\ \forall n \\ & \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}_n,\ \rho \geq 0,\ \xi_n \geq 0,\ \forall n \end{aligned} \quad (5)$$

where $\Delta(\mathbf{y}_n, \mathbf{y})$ is the loss function that quantifies the risk for predicting $\mathbf{y}$ given the correct label $\mathbf{y}_n$ and $d(\mathbf{X}_n, \mathbf{y}; \theta) = F(\mathbf{X}_n, \mathbf{y}_n; \theta) - F(\mathbf{X}_n, \mathbf{y}; \theta)$. Slack variables $\boldsymbol{\xi} = \{\xi_1, ..., \xi_N\}$ are introduced to allow errors in training data set, and the balance coefficient C controls the trade-off between margin maximization and training error minimization. To make the problem well-posed [9], we restrict the L_2 norm of $||\theta||$ to $\gamma,\ (\gamma > 0)$. The margin is scaled by $\Delta(\mathbf{y}_n, \mathbf{y})$ to separate the discriminant function of the true label $\mathbf{y}_n$ more from that of labels far from $\mathbf{y}_n$ than that of labels close to $\mathbf{y}_n$.

The framework can be easily implemented in emotion recognition. A sequential classification problem with large size of $\mathcal{Y}$ will have many constraints, and thus it requires a method to reduce the constraints. However, the size of $\mathcal{Y}$ in emotion recognition is small, and thus the constraints do not need to be reduced. Also, many speech emotion databases are collected from a small number of actors. Thus, for such a small sized database, the possibility of statistical mismatch between training and testing data set is considerable. The max-margin framework is known to perform well in this environment [11], [12].

We use the semi-definite programming (SDP) in implementing the margin scaling. The detail implementation procedure of the SDP for the max-margin framework using HMMs is described in [14].

3.2. Loss function for emotion recognition

We need a loss function that scales the separation margin in Eq. (5). In margin scaling, the Hamming loss function which is defined as the number of positions for which the corresponding labels are different is widely used [8], [9]. However, in emotion recognition, the Hamming loss becomes one for all constraints since the label $\mathbf{y}$ represents only one emotion, i.e. $\Delta(\mathbf{y}_n, \mathbf{y}) = 1,\ \mathbf{y}_n \neq \mathbf{y},\ \forall \mathbf{y},\ \forall n$. Thus, the separation margin is not scaled.

We use the WTM [10], illustrated in Fig. 1, for a loss function between emotions. The model shows the trait or tendency of a person in expressing an emotion and assumes that each emotion is a combination of two major coordinates: positive affect and negative affect. For example, happy is a combination of high positive and low negative affect. From this emotion model, we consider the distance between two emotions; happy (mixture of high positive and low negative) is further away from sad (mixture of low positive and high negative) than surprised (mixture of high positive and high negative).

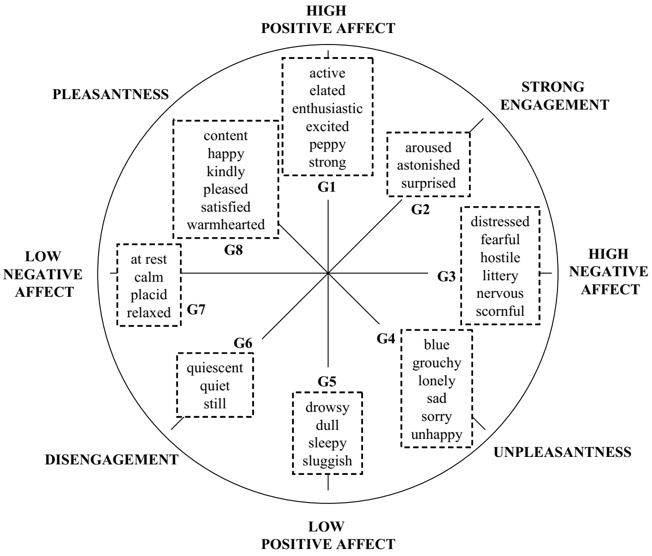

Fig. 1. Watson and Tellegen's model (WTM).

A self-report measurement of the positive and negative affectivity for each emotion is presented in [10]. In this paper, we define a simple measurement for each emotion based on the WTM. We classify each emotion into one of 8 groups (G1,...,G8) as in the Fig. 1 and assign the measurement $\mathbf{l}_{\mathbf{y}} = (l_1, l_2)$ to each group where l_1 is the measurement of positive affectivity and l_2 is the measurement of negative affectivity for emotion $\mathbf{y}$. In Table 1, the measurement for each emotion group is shown. For example, happiness which is in the group 8 is represented by $(-0.5, -0.5)$. Based on this table, the distance metric is defined as $\mathbf{d}(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{l}_{\mathbf{y}_1} - \mathbf{l}_{\mathbf{y}_2}$. We simply define the loss function as a linear function of the 1-norm of the distance metric:

$$\Delta(\mathbf{y}_1, \mathbf{y}_2) = \alpha ||\mathbf{d}(\mathbf{y}_1, \mathbf{y}_2)||_1 + \beta \qquad (6)$$

where α and β are positive real constants. We use this loss function to scale the separation margin in Eq. (5).

4. EXPERIMENT

Experiments were performed to evaluate the framework using EMO-DB [15]. The EMO-DB was collected from five male and female German actors expressing 7 emotions: anger, disgust, fear, sadness, boredom, neutral and happiness. Each actor produced ten utterances: five short and five long sentences. The database is comprised of 800 sentences: seven emotions · ten actors · ten sentences + some second versions. By the perception tests using 20 subjects, 494 sentences which are more than 60% natural and can be classified correctly with 80% accuracy were chosen for the experiment.

	negative affectivity (l_1)	positive affectivity (l_2)
Group 1 (G1)	0	1
Group 2 (G2)	0.5	0.5
Group 3 (G3)	1	0
Group 4 (G4)	0.5	-0.5
Group 5 (G5)	0	-1
Group 6 (G6)	-0.5	-0.5
Group 7 (G7)	-1	0
Group 8 (G8)	-0.5	0.5

Table 1. Measurement of positive and negative affectivity for each emotion group.

	ML	ML→MMI	ML→MS
1-mix	25.49	32.56	62.59
2-mix	54.66	60.13	69.28
4-mix	64.91	70.45	75.68
8-mix	70.96	72.73	78.67
16-mix	76.37	77.27	83.97
32-mix	78.46	81.17	86.32

Table 2. Average accuracy(%) of correct classification on the testing data for ML, MMI, and MS.

Seven emotions of EMO-DB were assigned to one of 8 groups as in the Fig. 1: anger(A) to G3, disgust(D) to G3, fear(F) to G3, sadness(S) to G4, boredom(B) to G5, neutral(N) to G6 and happiness(H) to G8. Although anger and disgust are not displayed in the Fig. 1, we assumed that they have high negative affectivity; in other words, they are in G3.

Emotion features consisted of 39 dimensions: 12 MFCCs, log energy and the corresponding delta and acceleration coefficients. Each emotion was modeled by a single-state HMM with different number of Gaussian mixture components. We assumed that each Gaussian mixture component has diagonal covariance matrices. The database was divided into five folds. Four male and female speakers were in training data set, and the remaining two speakers were in testing data set and the development data set which was used for tuning the parameters. The experiment for each fold was performed, and the average results across all trials were computed.

The baseline ML models were trained by the standard Baum-Welch algorithm using HMM toolkit 3.2 [16]. Based on the ML models, θ was updated by the MMI training algorithm [17] and margin scaling (MS) using the training data set. In the margin-scaling experiment, the parameter C, γ, α and β were manually tuned for the best performance using the development data set. The emotion recognition was performed by Eq. (2) using the testing data set.

The results of each training method for different number of Gaussian mixture components are summarized in Table 2. It shows that the margin scaling considerably improves the recognition accuracy compared to the ML and MMI. As the number of Gaussian mixture components increased, the recognition accuracy also increased. For small number of components such as 1-mix, 2-mix, 4-mix and 8-mix, improvements larger than 10% were observed. For larger number of components such as 16-mix and 32-mix, the MS still yielded about 7% improvement. The best accuracy was 86.32%.

We computed the confusion matrix which shows the accuracies between emotions where the first column indicates the true emotions that the speakers expressed, and the first row indicates the recognized emotions. In Table 3, the confusion matrix of the MS for 32-mix is shown. We can see the effect of the loss function from the confusion matrix. The label with high loss is separated from the correct label more than the label with low loss by scaling the separation margin with loss function. For example, the loss of happy (in G8) is higher than that of sadness (in G4) given boredom (in G5). Thus, given the true emotion of boredom, the rate of predicting happy (0%) was lower than that of predicting sadness (26.09%) as in Table 3. This means that we could reduce the possibility of predicting a label with high risk.

	A	D	F	S	B	N	H
A	100	0	0	0	0	0	0
D	5.00	95.00	0	0	0	0	0
F	0	0	100	0	0	0	0
S	25.00	0	0	66.67	0	0	8.33
B	13.04	0	0	26.09	60.87	0	0
N	0	0	0	0	0	100	0
H	0	0	0	0	0	0	100

Table 3. Confusion matrix of 7 emotions: anger(A), disgust(D), fear(F), sadness(S), boredom(B), neutral(N) and happiness(H).

5. CONCLUSION

We presented a method for speech emotion recognition by a max-margin framework incorporating a loss function based on the WTM. We defined a distance metric between two emotions, and the loss function was computed by the 1-norm of the distance metric. Experiments were performed on the EMO-DB. Each emotion was modeled by a single-state HMM which was estimated by three training methods. The results showed that the max-margin framework incorporating the loss function, called margin scaling, considerably improved the recognition rate over other methods: the ML and the MMI.

6. ACKNOWLEDGEMENT

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

7. REFERENCES

[1] R. W. Picard, *Affective Computing*, MIT Press, 1997.

[2] S. Casale, A. Russo, and G. Scebba, “Speech emotion classification using machine learning algorithms,” in *Proc. IEEE-ICSC*, 2008, pp. 158–165.

[3] M. Song, C. Chen, and M. You, “Audio-visual based emotion recognition using tripled hidden markov model,” in *Proc. IEEE-ICASSP*, 2004, vol. 5, pp. 877–880.

[4] H. Gunes and M. Piccardi, “Bi-modal emotion recognition from expressive face and body gestures,” *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[5] A. Haag, S. Goronzy, P. Schaich, and J. Williams, “Emotion recognition using bio-sensors: First steps towards an automatic system,” *LNCS*, vol. 3068, pp. 36–48, 2004.

[6] O. Kwon, K. Chan, J. Hao, and T. Lee, “Emotion recognition by speech signals,” in *Proc. Eurospeech*, 2003, pp. 125–128.

[7] Y. Lin and G. Wei, “Speech emotion recognition based on HMM and SVM,” in *Proc. ICMLC*, 2005, vol. 8, pp. 18–21.

[8] B. Taskar, C. Guestrin, and D. Koller, “Max-margin Markov networks,” *Advances in Neural Information Processing Systems*, vol. 16, 2004.

[9] I. Tsochantaridis, T. Joachims, and T. Hofmann, “Large margin methods for structured and interdependent output variables,” *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005.

[10] A. Tellegen, D. Watson, and L.A. Clark, “On The Dimensional and Hierarchical Structure of Affect,” *Psychological Science*, vol. 10, no. 4, pp. 297–303, 1999.

[11] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2000.

[12] George Heigold, Thomas Deselaers, Ralf Schlüter, and Hermann Ney, “Modified MMI/MPE: A direct evaluation of the margin in speech recognition,” in *Proc. ICML*, 2008, pp. 384–391.

[13] K. Crammer and Y. Singer, “On the algorithmic implementation of multiclass kernel-based vector machines,” *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[14] Y.Yin and H. Jiang, “A compact semidefinite programming (SDP) formulation for large margin estimation of HMMS in speech recognition,” in *Proc. ASRU*, 2007, pp. 312–317.

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” in *Proc. ICSLP*, 2005, pp. 1517–1520.

[16] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Boook (for HTK version 3.2)*, Univ. Cambridge, Cambridge, U.K., 2002.

[17] A. B. Yishai and D. Burshtein, “A discriminative training algorithm for hidden Markov models,” *IEEE Transaction on Speech and Audio Processing*, vol. 12, no. 3, pp. 204–217, 2004.