Loss-Scaled Large-Margin Gaussian Mixture Models for Speech Emotion Classification

Sungrack Yun, *Student Member, IEEE*, and Chang D. Yoo, *Member, IEEE*

***Abstract*—This paper considers a learning framework for speech emotion classification using a discriminant function based on Gaussian mixture models (GMMs). The GMM parameter set is estimated by margin scaling with a loss function to reduce the risk of predicting emotions with high loss. Here, the loss function is defined as a function of a distance metric using the Watson and Tellegen's emotion model. Margin scaling is known to have good generalization ability and can be considered appropriate for emotion modeling where the parameter set is likely to be over-fitted to the training data set whose characteristics may differ from those of the testing data set. Our learning framework is formulated as a constrained optimization problem which is solved using semi-definite programming. Three tasks were evaluated: acted emotion classification, natural emotion classification, and cross database emotion classification. In each task, four loss functions were evaluated. In all experiments, results consistently show that margin scaling improves the classification accuracy over other learning frameworks based on the maximum-likelihood, maximum mutual information and max-margin framework without margin scaling. Experiment results also show that margin scaling substantially reduces the overall loss compared to the max-margin framework without margin scaling.**

***Index Terms*—Gaussian mixture models (GMMs), margin scaling, speech emotion classification, Watson and Tellegen's model.**

I. Introduction

THERE has been growing research interest in the field of human–computer intelligent interaction (HCII) [1]. Currently, such interactions when writing an e-mail, searching a file, or running a utility do not involve any intelligence on the part of the computer: a computer simply takes a user's input and displays characters, images, and videos corresponding to the input. Unsatisfied with this type of interactions (computer responding passively to a user's input), many researchers are looking into ways to interact in a more intelligent manner such that a more friendly and intuitive interaction is possible. In HCII, a computer is supposed to understand, cognize, and interpret a user's inherent intentions and take actions more intelligently, e.g., correcting the input to what a user intended, displaying additional information by predicting a user's need from the input history and suggesting web contents related to a user's preference. Recognizing human emotional state could give way for intelligent interaction. Certainly, understanding the emotional state of others helps us to communicate with others more naturally. For these reasons, recognizing and processing human emotions have become important tasks in the HCII and affective computing [2], [3].

For various reasons, research in human emotion classification is conducted mostly based on speech. There are various modalities for emotion classification: facial expressions and speech [4]–[6], gesture and body language [7], and bio information such as electrocardiogram (ECG), electromyography (EMG), electrodermal activity, skin temperature, galvanic resistance, blood volume pulse (BVP), and respiration [8], [9]. Compared to other modalities, speech is readily accessible: microphones are more affordable and inexpensive than the ECG, EMG, and BVP sensors and are less cumbersome to use. For this reason, a wide range of HCII applications are based on speech emotion classification. According to the user's emotion, a service robot takes appropriate actions, a media player changes music or movie and a computer game controls the game status. An audio response system of a call center detects customer's emotion and connects to an expert when the customer becomes angry.

Various results in speech emotion classification have been reported [6], [10]–[17]. As emotional features, fundamental frequency, log energy, mel-frequency cepstral coefficients (MFCCs), zero-crossing rate, linear prediction coefficients (LPCs), pitch, duration of voiced/unvoiced part and Teager energy cepstrum coefficients are widely used. In [10]–[12], feature selection algorithms such as sequential forward floating selection and genetic algorithms have been proposed. In [13] and [14], to improve the classification performance, jitter and shimmer are added to MFCCs, and gender information is used, respectively. In [15], various emotional features are classified into groups, and the most relevant features from different feature groups are selected for better understanding of speaker-independent emotion classification. In [6] and [12], various graphical models such as hidden Markov model (HMM) and its variations were used to model speech emotion. In [16] and [17], various classifiers such as support vector machine (SVM) classifiers, HMM-based classifiers, linear discriminant analysis, quadratic discriminant analysis, neural networks, and k-nearest neighbors were used to compare the performance of the classifiers.

Almost all learning algorithms for modeling speech emotion are liable to the over-fitting problem due to the limited database

Manuscript received November 12, 2010; revised April 05, 2011 and June 20, 2011; accepted June 20, 2011. Date of publication July 18, 2011; date of current version December 14, 2011. This work was performed for the Intelligent Robotics Development Program, one of the Frontier R&D Programs funded by the Ministry of Knowledge Economy (MKE). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Steve Renals.

The authors are with Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: yunsungrack@kaist.ac.kr; cdyoo@ee.kaist.ac.kr).

Digital Object Identifier 10.1109/TASL.2011.2162405

size: collecting many speech samples which cover all variations is very expensive. As a result, the classification performance on characteristics not included in the training data set may not be satisfactory. To improve the classification accuracy, a speaker adaptation algorithm [18] can be used for speakers not included in the training data set; however, the model adapts to the emotion of the speakers involved in the adaptation data and generally does not yield speaker independency. For speaker-independent emotion classification, various techniques [15], [19]–[21] have been reported with good testing data performance.

This paper considers loss-scaled large margin Gaussian mixture models (GMMs) for speaker-independent emotion classification. Each emotion is modeled using one GMM, and each GMM parameter set is estimated by maximizing the margin which is the minimum separation between the GMMs of the correct emotion and of other competing emotions. Here, the margin is scaled according to a loss function (a measurement of recognition accuracy). This learning framework is known as margin scaling [22], [23]. The GMMs estimated by margin scaling can provide good generalization ability [24], [25]; thus, it performs well in emotion classification where there is substantial statistical mismatch between training and testing data set due to the difference in speakers and speaking styles. Also, margin scaling can be easily applied to emotion modeling where the number of emotions to classify is small: we do not need a method to reduce the number of margin constraints in the learning framework.

We evaluate four loss functions in scaling the margin: Hamming loss function and three loss functions based on the Watson and Tellegen's emotion model (WTM). The Hamming loss function, defined as the number of mismatched labels, is usually used in margin scaling [22], [26], [27]. In emotion classification where an utterance is expressed with only one emotion as opposed to a sequence of emotions, the Hamming loss becomes the zero-one loss, and the margin is not scaled. We define other loss functions using a distance metric between two emotions based on the WTM [28]–[30] which describes an emotion as two major coordinates: positive affect and negative affect. To compare the performance, we use three loss functions based on the distance metric: linear function, log function, and exponential function to compare the classification performance with different rate of increase in loss for the same distance. By scaling the margin with the WTM-based loss, we can separate the emotion which is very different from the correct emotion and thus reduce the risk of misclassifying the emotion with high loss.

We use the MFCCs, log energy, pitch, zero-crossing rate and the corresponding delta and acceleration coefficients as emotional features: the focus of this paper is in the understanding of the considered learning framework in emotion classification, and does not investigate any feature selection algorithms as suggested in [10], [12]. Thus, any performance improvement will be mainly due to the learning framework. To classify speech emotion, the maximum *a posteriori* (MAP) criterion is used. In the experiment, we show that loss-scaled large-margin GMMs obtained by margin scaling using the Hamming loss function (MSH) and the margin scaling using the linear (MSN), log (MSL), and exponential function (MSE) based on the WTM outperform both GMMs estimated by the maximum-likelihood (ML) and the maximum mutual information (MMI) criteria. Compared to the MSH, MSN reduces the classification error and the risk of misclassifying a very different emotion. Also, three loss functions were compared and analyzed. In our previous work [31], we observed that the MSN improves the classification accuracy over the ML and the MMI using the Berlin database of emotional speech (EMO-DB). In this paper, we evaluate and analyze loss-scaled large margin GMMs by performing three classification tasks: acted emotion classification, natural emotion classification and cross database classification. In acted emotion classification, the EMO-DB, speech under simulated and actual stress (SUSAS) and Danish emotional speech (DES) were used. We chose these databases to see the effectiveness of the proposed learning criterion with different training data size. In natural emotion classification, Vera am Mittag German (VAM) database including spontaneous speech was used. In cross database classification, EMO-DB and SUSAS were used as training data set, and DES was used as testing data set.

The outline of the paper is as follows. Section II introduces emotion classification using GMM with two different conventional training criteria. Section III describes margin scaling based on the WTM. Section IV compares the performances of considered learning framework to the learning frameworks using conventional training criteria on three tasks. Section V concludes and summarizes the paper.

II. Background

A. Emotion Classification Using GMM

This paper uses GMM (single-state HMM) for emotion classification. In emotion classification, a label $\mathbf{y}^*$ representing one of M emotions in $\mathcal{Y}$ is predicted from a given speech feature $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ where each element is in the D-dimensional vector space $\mathcal{X}$ such that

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

where F is a discriminant function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ which is parameterized by θ. The feature vector $\mathbf{x}_t$ is extracted from each speech frame, $1 \le t \le T$. When using GMM, the conditional distribution $\log p_\theta(\mathcal{Y}|\mathcal{X})$ is the discriminant function, and the decision criterion becomes the MAP decoding rule as

$$\begin{aligned}\mathbf{y}^* &= \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{X}, \mathbf{y}; \theta) \\ &= \arg\max_{\mathbf{y}\in\mathcal{Y}} \log p_\theta(\mathbf{y}|\mathbf{X}) \\ &= \arg\max_{\mathbf{y}\in\mathcal{Y}} [\log p_\theta(\mathbf{X}|\mathbf{y}) p(\mathbf{y})]\end{aligned} \tag{2}$$

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

We also assume that an utterance is expressed with a single emotion (not a sequence of different emotions). Thus, the emotional statistics of $\mathbf{X}$ does not change over time, and each emotion is modeled using one GMM. The distribution of K Gaussian mixture components is expressed as

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

where w_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Lambda}_k$ are respectively the mixture weight, mean vector and covariance matrix of the kth Gaussian distribution which is given by

$$\mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)=\frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|\boldsymbol{\Lambda}_k|}}\times\exp\left[-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu}_k)_t'\boldsymbol{\Lambda}_k^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_k)\right].$$

The superscript $'$ denotes vector transpose. The parameter set θ is the vector comprising mixture weights, mean vectors and covariance matrices for all GMMs, and the mixture weights satisfy the following constraint:

$$\sum_{k=1}^{K}w_k=1. \tag{4}$$

Assuming $\mathbf{x}_t$ is independent and identically distributed, we express the discriminant function as

$$\begin{aligned}F(\mathbf{X},\mathbf{y};\theta)&=\log p_\theta(\mathbf{X}|\mathbf{y})p(\mathbf{y})\\&=\log\left[\prod_{t=1}^{T}p_\theta(\mathbf{x}_t|\mathbf{y})\cdot\frac{1}{M}\right]\\&=\log\left[\frac{1}{M}\prod_{t=1}^{T}\sum_{k=1}^{K}w_k\mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)\right].\end{aligned} \tag{5}$$

B. Conventional Training Criteria for GMM: ML and MMI

In this section, we will briefly review two conventional training criteria for GMM namely ML and MMI. The training goal is to find θ using a set of input–output pairs $(\mathbf{X}_n,\mathbf{y}_n), n=1,\ldots,N$ such that the decision criterion leads to the minimum prediction error. The ML criterion finds θ by maximizing the following objective function [32]:

$$F_{ML}(\theta)=\sum_{n=1}^{N}\log p_\theta(\mathbf{X}_n|\mathbf{y}_n). \tag{6}$$

The ML is the most widely used training criterion for GMM, and θ is usually obtained using the Baum–Welch algorithm [32]. However, it does not consider the relationship between labels. The parameter set of $\mathbf{y}_n$ is estimated only using the corresponding input $\mathbf{X}_n$. Thus, it does not discriminate one label from another.

The MMI criterion, also referred to as conditional maximum-likelihood criterion, finds θ by maximizing the following objective function [33], [34]:

$$\begin{aligned}F_{MMI}(\theta)&=\sum_{n=1}^{N}\log p_\theta(\mathbf{y}_n|\mathbf{X}_n)\\&=\sum_{n=1}^{N}\log\frac{p_\theta(\mathbf{X}_n|\mathbf{y}_n)p(\mathbf{y}_n)}{\sum_{\mathbf{y}\in\mathcal{Y}}p_\theta(\mathbf{X}_n|\mathbf{y})p(\mathbf{y})}.\end{aligned} \tag{7}$$

It maximizes the discriminant function used in the MAP decoding rule. The parameter set corresponding to label $\mathbf{y}_n$ is estimated by considering all other labels $\mathbf{y}\in\mathcal{Y}$ as in (7). The MMI estimation can be implemented using the extended Baum–Welch algorithm [35] or approximated MMI algorithm [34]. The criterion attempts to minimize the prediction error on the training data set. For this reason, the framework based on the MMI criterion performs better than that based on the ML criterion when there is considerable discrepancy between the probabilistic model and the data [36].

C. Generalization Ability

In the cases where we have sufficiently large number of data such that the statistical characteristics of training data set are equivalent to those of the testing data set, the performance on the testing data set is as good as that on the training data set. However, given a small number of data, the statistical characteristics of training data set may not match well with those of testing data set. For example, due to the difference in speakers, microphones and other environmental factors, there may be a mismatch between training and testing data set. In such cases, the model parameters are over-fitted to the training data set, and the performance on the testing data set is poor even though the performance on the training data set is good.

Margin scaling has better generalization ability on testing data than the learning frameworks based on the ML and MMI criteria when there is a mismatch between training and testing data [24], [25]. Maximizing the margin leads to minimizing the upper bound of the error on the testing data set [23]. As a result, we obtain the parameter set with good classification accuracy even when there is a mismatch between testing and training data set.

III. Margin Scaling for Emotion Classification

This section describes a method to estimate the GMM parameter set by margin scaling. Based on the Watson and Tellegen's model, a distance metric between emotions is defined, and the loss is computed as a function of the distance metric.

A. Formulation

The learning framework based on the margin scaling finds parameter set θ by maximizing the margin ρ and simultaneously minimizing the sum of the slack variables $\boldsymbol{\xi}=\{\xi_1,\ldots,\xi_N\}$ such that the difference between the discriminant function given the correct emotion $\mathbf{y}_n$ and the discriminant function given the incorrect emotion $\mathbf{y},\mathbf{y}\neq\mathbf{y}_n$ is greater than or equal to $\rho\Delta(\mathbf{y}_n,\mathbf{y})-\xi_n$ for all $n=1,\ldots,N$ [22], [23], [37]:

$$\begin{aligned}\min_{\rho,\boldsymbol{\xi},\theta:\|\theta\|=\gamma}\quad&-\rho+\frac{C}{N}\sum_{n=1}^{N}\xi_n\\\text{subject to}\quad&d_n(\mathbf{y};\theta)\geq\rho\Delta(\mathbf{y}_n,\mathbf{y})-\xi_n,\quad\rho\geq0,\ \xi_n\geq0\\&\forall n,\ \forall\mathbf{y}\in\mathcal{Y}\setminus\mathbf{y}_n\end{aligned} \tag{8}$$

where

$$d_n(\mathbf{y};\theta)=F(\mathbf{X}_n,\mathbf{y}_n;\theta)-F(\mathbf{X}_n,\mathbf{y};\theta) \tag{9}$$

and $\Delta(\mathbf{y}_n,\mathbf{y})$ is a loss function that quantifies the risk of classifying $\mathbf{X}_n$ into $\mathbf{y}$ given the correct label $\mathbf{y}_n$. The balance coefficient C controls the tradeoff between the maximization of the margin and the minimization of sum of slack variables. We make the problem well-posed by restricting the L_2-norm of θ to be a positive constant γ [23].

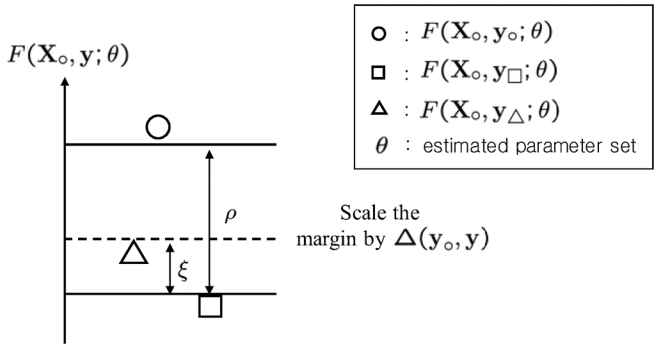

Fig. 1. The circle, rectangle, and triangle represent values of the discriminant function given the correct label $\mathbf{y}_\circ$, incorrect label $\mathbf{y}_\triangle$, and another incorrect label $\mathbf{y}_\square$, respectively. The rectangle is separated from the circle by the separation margin ρ while the triangle is ξ closer to the circle. It is assumed that the loss between $\mathbf{y}_\circ$ and $\mathbf{y}_\square$ is greater than that between $\mathbf{y}_\circ$ and $\mathbf{y}_\triangle$. When θ is estimated by margin scaling, the rectangle with high loss is moved further away from the circle than the triangle which has a lower loss.

The margin is scaled by $\Delta(\mathbf{y}_n, \mathbf{y})$ to distance the discriminant function of the correct label $\mathbf{y}_n$ to be further away from that of the label with high loss than that of the label with low loss. This is illustrated in the Fig. 1. The discriminant functions of true label and other two incorrect labels are denoted by circle, rectangle, and triangle, respectively. Let the loss between circle and rectangle be larger than that between circle and triangle. By scaling the separation margin with a loss, the rectangle is placed further away from the circle than the placement of the triangle with respect to the circle. Thus, we reduce the risk of predicting the rectangle which has high loss.

Collecting "infinite emotional data" such that there is no mismatch between training and testing data (training data that covers all variations in speech emotion) is very difficult. For this reason, the model parameter set of an emotion model is over-fitted to the speakers' statistics in the training data set whose characteristics may differ from other speaker characteristics in the testing data set. Under this condition, margin scaling performs well due to its good generalization ability [24].

Also, margin scaling can be easily implemented. We do not need to consider a method such as the sub-gradient method [38] and the cutting-plane method [23] to reduce the number of constraints as in (8) since there are only a few number of emotions to classify.

B. Loss Function for Emotion Recognition

To scale the margin, a loss function needs to be defined. Most learning methods adopting margin scaling use the Hamming loss [22], [26], [27] which is defined as the number of mismatched positions. In this paper, we assume that the utterance is expressed with only one emotion. Then, the Hamming loss becomes the zero-one loss: if $\mathbf{y}_n \neq \mathbf{y}$, $\Delta(\mathbf{y}_n, \mathbf{y}) = 1$, and otherwise, $\Delta(\mathbf{y}_n, \mathbf{y}) = 0$. Thus, if we use the Hamming loss, $\Delta(\mathbf{y}_n, \mathbf{y}) = 1$ for all constraints in (8). This means that the separation margin is not scaled with different values.

We consider another loss function based on the WTM [28]–[30] illustrated in Fig. 2. The model shows a 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 or strong engagement and pleasantness. We choose positive affect and negative affect as coordinates: e.g., happy is a combination of high positive and low negative affect and sad is a combination of high negative affect and low positive affect. Using the WTM, we consider a distance between emotions: e.g., 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). In [28], a measurement of the positive and negative affectivity for each emotion is evaluated by a self-report. In this paper, a simple measurement of the positive and negative affectivity for each emotion is defined. As illustrated in the Fig. 2, we define eight emotion groups $(\mathrm{G1}, \ldots, \mathrm{G8})$ assign a measurement $\mathbf{l} = (l_1, l_2)$ for each group. The measurement of positive affectivity and the measurement of negative affectivity are denoted by l_1 and l_2, respectively. Table I shows the measurement for each emotion group. For example, all emotions in the G2 are represented by (0.5,0.5). Using the measurement, we define a distance metric between $\mathbf{y}_n$ and $\mathbf{y}$ as $d(\mathbf{y}_n, \mathbf{y}) = \|\mathbf{l}_{\mathbf{y}_n} - \mathbf{l}_{\mathbf{y}}\|_1$ where $\mathbf{l}_{\mathbf{y}_n}$ is the measurement of $\mathbf{y}_n$, $\mathbf{l}_{\mathbf{y}}$ is the measurement of $\mathbf{y}$, and $\|\cdot\|_1$ is the L_1-norm of a vector. Finally, we define three loss functions using the distance metric

$$\Delta_N(\mathbf{y}_n, \mathbf{y}) = \begin{cases} \alpha d(\mathbf{y}_n, \mathbf{y}) + \beta, & \mathbf{y}_n \neq \mathbf{y} \\ 0, & \mathbf{y}_n = \mathbf{y} \end{cases} \tag{10}$$

$$\Delta_L(\mathbf{y}_n, \mathbf{y}) = \begin{cases} \alpha \log\left[d(\mathbf{y}_n, \mathbf{y})\right] + \beta, & \mathbf{y}_n \neq \mathbf{y} \\ 0, & \mathbf{y}_n = \mathbf{y} \end{cases} \tag{11}$$

$$\text{and } \Delta_E(\mathbf{y}_n, \mathbf{y}) = \begin{cases} \alpha \exp\left[d(\mathbf{y}_n, \mathbf{y})\right] + \beta, & \mathbf{y}_n \neq \mathbf{y} \\ 0, & \mathbf{y}_n = \mathbf{y} \end{cases} \tag{12}$$

where α and β are real-valued constants. To scale the separation margin and also to compare the performance, we use the above loss functions which yield different loss values for the same distance difference.

C. Implementation

We optimize the problem (8) using a semi-definite programming (SDP) solver. In [39] and [40], the optimization problem of large-margin HMM for the separable case is solved using the DSDP [41] which is a SDP solver. Based on the procedure in [39], we express the optimization problem of margin scaling as a SDP. For simplicity, we only estimate the mean vector of GMM: i.e., the mixture weight and covariance matrix are obtained by the ML estimation. The detail implementation procedure is described in the Appendix.[1]

IV. Experiment

We conducted experiments using the EMO-DB [42], SUSAS [43], DES [44], and VAM [45]. Emotion feature vectors were extracted from speech frames: the frame size and rate was set to 25 and 10 ms, respectively. A feature vector consisting of 12 MFCCs, log energy and the corresponding delta and acceleration coefficients was used. In the cross database experiment, we also used pitch and zero crossing rate. Each emotion was modeled by one GMM with different number of Gaussian mixture

[1]An implemented code of the proposed learning algorithm is available in http://slsp.kaist.ac.kr/xe/software.

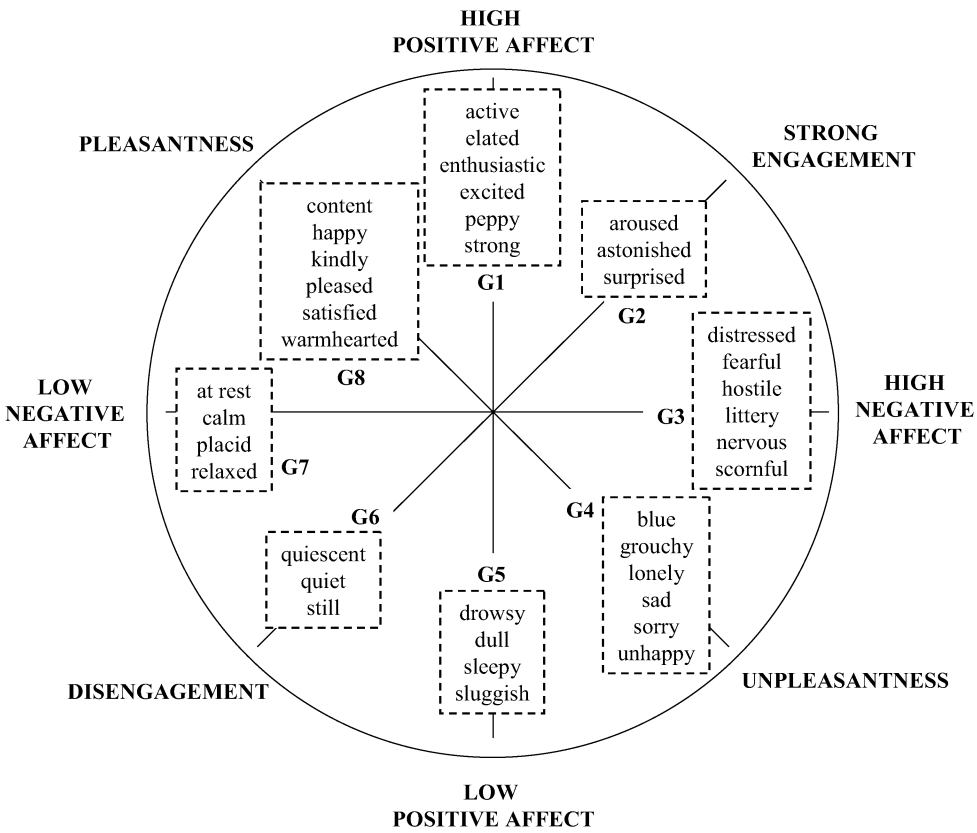

Fig. 2. Watson and Tellegen's model (WTM). It shows the trait or tendency of a person in expressing an emotion. Emotions are classified into eight groups: $G1, G2, \ldots, G8$. The model assumes each emotion is a combination of two major coordinates: positive affect and negative affect or strong engagement and pleasantness.

TABLE I
MEASUREMENT OF POSITIVE AND NEGATIVE AFFECTIVITY FOR EACH EMOTION GROUP

	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

components. A diagonal covariance matrix rather than a full covariance matrix is used for three reasons. First, the MFCCs are considered uncorrelated [46], [47]. Second, a small training data size can lead to inaccurate estimate of the full covariance matrix [32], [48]: the covariance estimate can be over-fitted to the small training data leading to performance degradation. Finally, for a D-dimensional feature vector, the computational complexity of estimating the diagonal covariance matrix of feature vector is reduced by a factor D: a full covariance matrix has $D(D+1)/2$ parameters while a diagonal covariance matrix has D parameters. Thus, we can approximate a full covariance matrix to a diagonal covariance matrix. We increase the number of Gaussian mixture components until no more improvement is observed. If we use a too large number of Gaussian mixture components, over-fitting occurs due to the large number of parameters compared to the limited number of training data.

The database was split into three data sets: training data set, testing data set, and development data set. Using the training data set, we obtained six different model parameter sets for performance comparisons. First, the baseline model was obtained by the ML estimation [32] using HTK 3.2 [49]. Based on the ML models, model parameter set θ was updated by the MMI estimation [34], MSH, MSN, MSL, and MSE. Given the preset value range, $0 < C \leq 1$, $0 < \gamma \leq 5$, $0 < \alpha < 1$ and $0 < \beta < 1$, the preset values were determined based on best performance using the development data set. Finally, speech segments were classified by (2) using the testing data set.

A. EMO-DB

The EMO-DB was collected from five male and female German actors expressing seven 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 utterances: seven emotions $\cdot$ ten actors $\cdot$ ten sentences+some second versions. A perception test using 20 subjects was performed in [42], and 494 utterances which are judged as natural by more than 60% human listeners and can be classified correctly more than 80% accuracy were chosen for the experiment.

The database was equally split into five folds. Let s_i be the ith speaker's data. EMO-DB contains ten speakers (male speakers: $s_0, \ldots, s_4$ and female speakers $s_5, \ldots, s_9$), and each fold holds the following pair of speakers' data: Fold 1-(s_0, s_5), Fold 2-(s_1, s_6), Fold 3-(s_2, s_7), Fold 4-(s_3, s_8), and Fold

TABLE II
AVERAGE ACCURACY(%) OF CORRECT CLASSIFICATION ON THE TESTING DATA SET OF THE EMO-DB

	ML	ML→MMI	ML→MSH	ML→MSN	ML→MSL	ML→MSE
1-mix	44.55	51.68	57.86	63.79	62.91	61.85
2-mix	54.66	60.13	65.70	69.54	70.03	69.57
4-mix	64.91	70.45	72.81	76.32	74.63	74.17
8-mix	70.96	72.73	74.83	78.99	78.51	78.05
16-mix	76.37	77.27	80.23	83.30	82.66	81.68
32-mix	78.46	81.17	83.24	87.80	86.36	85.45

5-(s_4, s_9). Four folds (four male and four female speakers' data) were used as the training data set. Remaining one fold (one male and one female speaker's data) was used as testing and development data sets. There are two ways to divide the remaining one fold (two speakers' data) into testing data set (one speaker's data) and the development data set (the other one speaker's data). In each experiment, we consider both ways. The training, testing, and development data set were completely disjoint. Ten experiments based on 5×2 possible combinations were performed with different folds for the training, testing and development data set, and the average results across all trials were computed.

Seven emotions of the EMO-DB were assigned to one of eight groups as in Fig. 2: anger (A) to G3, disgust (D) to G3, fear (F) to G3, sadness (S) to G4, boredom (B) to G5, neutral (N) to G7 and happiness (H) to G8. Although anger and disgust are not displayed in the Fig. 2, we assumed that they have high negative affectivity: in other words, anger and disgust were considered in G3. Then, we computed the distance metric and loss function between emotions.

The average classification accuracy of each training method for different number of Gaussian mixture components is summarized in Table II. For one-Gaussian mixture component, the accuracy on the testing data set was 44.55% in the ML estimation. However, the accuracy on the training data set using the same ML model was 52.94%. The large difference in accuracy between two data sets is mainly due to the small number of training utterances and large difference in speaker characteristics: characteristics between female and male speakers are quite different, and for each speaker, we have about seven utterances per emotion. In this situation, the max-margin framework performs better than other frameworks. For one-Gaussian mixture component, the MSH and MSN yielded, respectively, 29.88% and 43.19% relative performance improvements from the ML while the MMI yielded 16.00% relative performance improvement. Also, for other mixture components, Table II shows that the MSH and MSN yield better result than the ML and MMI. It means that the MSH and MSN produce less speaker-dependent models and thus have better generalization ability than the ML and MMI. Also, the MSN considerably improves the classification accuracy over the MSH. As the number of Gaussian mixture components increases, the classification accuracy also increases. The highest accuracy of 87.80% (higher than the result in [50]) was obtained using the MSN with 32-mixture components.

We computed the confusion matrix which shows the classification accuracies between emotions where the first column indicates the true emotion labels that a speaker expressed, and the first row indicates various emotion labels recognized. The confusion matrices of the EMO-DB using the MSH and MSN model for 32-Gaussian mixture components are shown in Table III and Table IV, respectively. Table IV shows that the classification accuracy of anger, disgust, sadness, boredom, and neutral are over 90%. We can see the effect of the margin scaling from these two confusion matrices. For example, given the emotion fear (in G3), the loss of neutral (in G7) is higher than that of anger (in G3) or boredom (in G5). Thus, given the true emotion fear, the rate of predicting neutral is further reduced (42.42%→36.37%) than that of predicting boredom (18.18%→15.15%) or anger (6.07%→3.03%) as shown in Table III and IV. For other emotion pairs, for example, the classification error of sadness (in G4) given boredom (in G5) is more reduced than that of fear (in G3) given happiness (in G8) since the classification error of boredom is less than that of happiness.

TABLE III
CONFUSION MATRIX BETWEEN SEVEN EMOTIONS OF THE EMO-DB USING THE MSH MODEL FOR 32-GAUSSIAN MIXTURE COMPONENTS

	A	D	F	S	B	N	H
A	100	0	0	0	0	0	0
D	0	100	0	0	0	0	0
F	6.07	0	33.33	0	18.18	42.42	0
S	0	0	0	100	0	0	0
B	0	0	0	16.13	83.87	0	0
N	0	0	0	0	0	100	0
H	8.57	5.71	22.86	0	0	0	62.86

TABLE IV
CONFUSION MATRIX BETWEEN SEVEN EMOTIONS OF THE EMO-DB USING THE MSN MODEL FOR 32-GAUSSIAN MIXTURE COMPONENTS

	A	D	F	S	B	N	H
A	100	0	0	0	0	0	0
D	0	100	0	0	0	0	0
F	3.03	0	45.45	0	15.15	36.37	0
S	0	0	0	100	0	0	0
B	0	0	0	3.23	96.77	0	0
N	0	0	0	0	0	100	0
H	8.57	0	20.00	0	0	0	71.43

Overall, using (10) for all testing data, we observed that the MSN yields 24.63% relative loss reduction[2] from the MSH for 32-mixture components. This means that we could greatly reduce the overall loss in the EMO-DB by reducing the risk of predicting the label with high loss more than the label with low loss.

[2]Relative loss reduction is computed by (previous loss value—reduced loss value)/previous loss value.

TABLE V
AVERAGE ACCURACY(%) OF CORRECT CLASSIFICATION ON THE TESTING DATA SET OF THE SUSAS

	ML	ML→MMI	ML→MSH	ML→MSN	ML→MSL	ML→MSE
1-mix	53.68	54.76	53.84	53.94	53.81	53.94
2-mix	54.60	55.11	55.77	57.08	57.24	56.60
4-mix	62.67	64.06	65.65	66.67	67.59	66.83
8-mix	64.79	65.17	66.89	68.38	70.41	69.71
16-mix	67.75	70.45	71.56	71.87	72.51	71.94
32-mix	69.78	72.44	72.92	73.65	73.78	72.38

We compared the accuracy obtained by three loss functions. In all Gaussian mixture components (except the two-Gaussian mixture components), MSN outperformed MSL and MSE. Also, MSL outperformed MSE. In the EMO-DB, seven emotions were well separated when the linear loss function was used. The exponential loss function leads to loss values that are too large for small distance difference, and large margin between two emotions was not observed (leading to smaller performance improvement compared to the MSN).

B. SUSAS

The SUSAS was created by the Robust Speech Processing Laboratory at the University of Colorado–Boulder. The database is partitioned into two conditions: actual and simulated. The utterances in the actual condition were not used: we consider only acted or simulated emotion. The simulated utterances were collected from nine male speakers expressing eleven speaking styles: angry, clear, cond50, cond70, fast, Lombard, loud, neutral, question, slow, and soft. Each speaker produced two utterances for 35 words per speaking style. The number of all utterances was 3150. We only selected five of the eleven speaking styles which can be mapped to the WTM: angry, clear, cond50, loud, and soft. Fast, Lombard, question and slow were not considered to be related to emotional expression. Soft and neutral are almost identical in speaking style (identical emotion), and only soft was included in the experiment.

The database was equally split into three folds. Let s_i be the ith speaker's data. SUSAS contains nine speakers (male speakers: $s_0, \ldots, s_9$), and each fold holds the following speakers' data: Fold 1-(s_0, s_1, s_2), Fold 2-(s_3, s_4, s_5), and Fold 3-(s_6, s_7, s_8). Two folds (six speakers' data) were used as the training data set. Remaining one fold (three speakers' data) was used as testing and development data sets. There are three ways to divide the remaining fold (three speakers' data) into testing data set (two speakers' data) and the development data set (the other one speaker's data). In each experiment, we consider three ways. The training, testing, and development data set were completely disjoint. Nine experiments based on 3×3 possible combinations were performed with different folds for the training, testing and development data set, and the average results across all trials were computed.

Five speaking styles of the SUSAS were assigned to one of eight groups: angry(A) to G3, clear(C) to G1, cond50(Co) to G3, loud(L) to G1, and soft(So) to G6. Although clear, cond50, loud, and soft are not displayed in the Fig. 2, we assumed that they have, respectively, high positive affectivity, high negative affectivity, high positive affectivity, and high disengagement: in other words, clear, cond50, loud, and soft were considered in G1, G3, G1, and G6.

TABLE VI
CONFUSION MATRIX BETWEEN FIVE SPEAKING STYLES OF THE SUSAS USING THE MSH MODEL FOR 32-GAUSSIAN MIXTURE COMPONENTS

	A	Co	L	So	C
A	68.41	5.40	19.05	0.16	6.98
Co	0.79	80.00	1.59	6.83	10.79
L	20.32	1.11	76.67	0	1.90
So	0	10.63	0	86.83	2.54
C	3.81	34.60	7.94	0.95	52.70

TABLE VII
CONFUSION MATRIX BETWEEN FIVE SPEAKING STYLES OF THE SUSAS USING THE MSN MODEL FOR 32-GAUSSIAN MIXTURE COMPONENTS

	A	Co	L	So	C
A	69.84	4.92	19.05	0	6.19
Co	0.79	80.16	1.90	5.40	11.75
L	19.84	1.11	77.62	0	1.43
So	0	10.63	0	86.51	2.86
C	3.49	33.17	8.57	0.63	54.13

The average classification accuracy of each training method for different number of Gaussian mixture components is summarized in Table V. For one-Gaussian mixture component, the accuracy on the testing data set was 53.68% in the ML estimation. In this case, the accuracy on the training data set using the same ML model was 54.11%. The small difference in accuracy between two data sets is mainly due to the large number of training utterances and small difference in speaker characteristics: there are only male speakers in the SUSAS. In this situation, the MMI can yield better result, and the performance improvement of the MSH and MSN is less than that in the EMO-DB. For one-Gaussian mixture component, the accuracy of the MMI is better than that of the MSH and MSN: the MMI yielded 2% relative performance improvement from the ML while the MSH and MSN yielded 0.3% and 0.5% relative performance improvements. For other mixture components, the MSH and MSN yields slightly better result than the MMI. The improvements of the MSH and MSN over the ML and MMI are not considerable compared to those in the EMO-DB. This means that the over-fitting is not considerable in the SUSAS. Also, in this case, the MSN improves the classification accuracy over the MSH. As the number of Gaussian mixture components increases, the classification accuracy also increases. The highest accuracy of 73.78% (higher than the result in [50]) was obtained using the MSL with 32-mixture components.

The confusion matrices of the SUSAS using the MSH and MSN model for 32-Gaussian mixture components are shown in Table VI and Table VII, respectively. Table VII shows that the classification accuracy of cond50, loud, and soft are over 75%.

TABLE VIII
AVERAGE ACCURACY(%) OF CORRECT CLASSIFICATION ON THE TESTING DATA SET OF THE DES

	ML	ML→MMI	ML→MSH	ML→MSN	ML→MSL	ML→MSE
1-mix	26.40	29.55	35.82	41.41	39.76	40.66
2-mix	43.57	48.80	57.35	61.49	61.14	61.45
4-mix	50.38	56.05	65.34	67.60	65.94	66.27

We can see the effect of the margin scaling from these two confusion matrices. For example, given the cond50 (in G3), the loss of soft (in G6) is higher than that of clear (in G1) or loud (in G1). Thus, given the cond50, the rate of predicting soft is further reduced (6.83%→5.40%) while that of predicting clear (10.79%→11.75%) or loud (1.59%→1.90%) increases as shown in Table VI and VII. Overall, using (10) for all testing data, we observed that the MSN yields 2.85% relative loss reduction from the MSH for 32-mixture components. In the SUSAS, we could slightly reduce the overall loss.

We compared the accuracy obtained by three loss functions. In all Gaussian mixture components (except the 1-Gaussian mixture component), MSL outperformed MSN and MSE, and MSN outperformed MSE. In the SUSAS, five speaking styles were well separated when the log loss function was used. The log loss function leads to small loss value given small distance difference, and we see that a small margin scaled by the log loss function is effective in the SUSAS.

C. DES

The DES was collected from two male and female actors expressing five emotions: anger, happiness, neutral, sadness, and surprise. Each speaker produced 13 utterances for each emotion: two single words, nine sentences and two passages of fluent speech. Additionally, 81 target utterances with neutral emotion were recorded. The database is comprised of 341 utterances: 175 from female speakers and 166 from male speakers.

The database was equally split into four folds. Let s_i be the ith speaker's data. DES contains four speakers (male speakers: s_0, s_1 and female speakers s_2, s_3), and each fold holds the following speakers' data: Fold 1-(s_0), Fold 2-(s_1), Fold 3-(s_2), and Fold 4-(s_3). Two folds (one male and one female speaker's data) were used as training data set, and the remaining two folds (one male and one female speaker's data) were used as testing and development data sets. There are two ways to divide the remaining one fold (two speakers' data) into testing data set (one speaker's data) and the development data set (the other one speaker's data). In each experiment, we consider both ways. The training, testing, and development data set were completely disjoint. Eight experiments based on 4×2 possible combinations were performed with different folds for the training, testing and development data set, and the average results across all trials were computed.

Five emotions of the DES were assigned to one of eight groups: anger(A) to G3, happiness(H) to G8, neutral(N) to G7, sadness(S) to G4, and surprise(Su) to G2.

The average classification accuracy of each training method for different number of Gaussian mixture components is summarized in Table VIII. For one-Gaussian mixture component, the accuracy on the testing data set was 26.40% in the ML estimation. However, the accuracy on the training data set using the same ML model was 59.93%. The large difference in accuracy between two data sets is mainly due to the small number of training utterances and large difference in speaker characteristics. For one-Gaussian mixture component, the MSH and MSN yielded respectively 35.68% and 56.86% relative performance improvements from the ML while the MMI yielded 11.93% relative performance improvement. The relative performance improvement is much larger than that in the EMO-DB and SUSAS due to the large difference in accuracy between two data sets. Also, for other mixture components, Table VIII shows that the MSH and MSN yield better result than the ML and MMI. Also, the MSN improves the classification accuracy over the MSH. As the number of Gaussian mixture components increases, the classification accuracy also increases. We increased the number of Gaussian mixture component only up to four due to the small number of training data. The highest accuracy of 67.60% (higher than the result in [50] was obtained using the MSN with four-mixture components.

TABLE IX
CONFUSION MATRIX BETWEEN FIVE EMOTIONS OF THE DES USING THE MSH MODEL FOR FOUR-GAUSSIAN MIXTURE COMPONENTS

	A	S	N	H	Su
A	57.69	0	19.23	9.62	13.46
S	0	51.92	40.39	1.92	5.77
N	0	10.67	88.77	0	0.56
H	15.38	11.54	17.31	44.23	11.54
Su	17.31	13.46	28.85	11.54	28.85

TABLE X
CONFUSION MATRIX BETWEEN FIVE EMOTIONS OF THE DES USING THE MSN MODEL FOR FOUR-GAUSSIAN MIXTURE COMPONENTS

	A	S	N	H	Su
A	63.46	0	13.46	3.85	19.23
S	0	51.92	40.39	1.92	5.77
N	0	10.67	88.77	0	0.56
H	15.38	9.62	13.46	50.00	11.54
Su	17.31	13.46	26.92	9.62	32.69

The confusion matrices of the DES using the MSH and MSN model for four-Gaussian mixture components are shown in Table IX and Table X, respectively. Table X shows that only neutral is well classified. We can see the effect of the margin scaling from these two confusion matrices. For example, given the anger (in G3), the loss of neutral (in G7) and happiness (in G8) are higher than that of surprised (in G2). Thus, given the anger, the rate of predicting neutral and happiness are further reduced (19.23%→13.46% and 9.62%→3.85%) while that of predicting surprised increases (13.46%→19.23%) as shown in Table IX and X. Overall, using (10) for all testing data, we observed that the MSN yields 6.96% relative loss reduction from the MSH for four-mixture components. Also, we could reduce the overall loss in the DES. The relative loss reduction in the DES is much smaller than that in the EMO-DB. This

TABLE XI
AVERAGE ACCURACY(%) OF BINARY EMOTION CLASSIFICATION ON THE TESTING DATA SET OF THE VAM

	ML	ML→MMI	ML→MSH	ML→MSN	ML→MSL	ML→MSE
1-mix	54.17	56.94	57.47	59.72	59.55	59.20
2-mix	57.47	59.90	61.98	64.05	63.54	63.02
4-mix	63.72	64.58	66.49	68.23	67.54	67.71

TABLE XII
AVERAGE ACCURACY(%) OF SIX-CLASS EMOTION CLASSIFICATION ON THE TESTING DATA SET OF THE VAM

	ML	ML→MMI	ML→MSH	ML→MSN	ML→MSL	ML→MSE
1-mix	31.94	32.99	34.03	34.86	34.72	34.55
2-mix	34.02	35.42	35.59	37.33	36.45	36.11
4-mix	35.07	36.28	36.63	38.54	37.50	36.81

is due to the high classification error between emotions. As shown in Table III, most emotions were well classified (except fear) in the EMO-DB while not in the DES.

We compared the accuracy obtained by three loss functions. In all Gaussian mixture components, MSN outperformed MSL and MSE, and MSE outperformed MSL. In the DES, five emotions were well separated when the linear loss function was used, and also it was observed that large margin scaled by the exponential loss function is effective in the DES.

D. VAM

The VAM database contains audio-visual data which were collected from the German TV show. We only used the audio part which consists of data from 47 speakers (11 male and 36 female, 947 utterances), and each utterance was evaluated in terms of the emotion primitives (valence, activation, and dominance) by 17 human listeners [45]. We performed two experiments: binary emotion classification and six-class emotion classification. In the binary emotion classification, to compare with the results in [51], we use the same experimental setup such that the database is divided using the merged evaluation results using the evaluator weighted estimator described in [52] and [53]. If the merged valence is between -0.25 and 0.25, the data were labeled as EMO class (grouping all the non neutral emotion states), and if the merged valence is greater than 0.25 or less then -0.25, the data were labeled as IDL class (consisting of neutral or neutral-like emotion states).

In addition, we performed six-class emotion classification. We divided the database into six emotions by k-means clustering using the three dimensional emotion primitives: valence, activation, and dominance.

The database was split such that 47 speakers' data were randomly assigned into three data sets. The training, testing, and development data set were completely disjoint: 28 speakers were used as training data set, 14 speakers were used as testing data set and 5 speakers were used as development data set. Eight experiments were performed with different speakers' data for the training, testing and development data set, and the average results across all trials were computed.

Emotions of the VAM were assigned to one of eight groups. In the binary emotion classifications, EMO to G1 and IDL to G5. In the six-class emotion classifications, each emotion, EMO1, ..., EMO6, was, respectively, assigned to the group G3, G1, G1, G1, G7, and G4 such that the measurement value of the group is the nearest to the center value (valence, activation) of the corresponding emotion cluster.

TABLE XIII
CONFUSION MATRIX BETWEEN SIX EMOTIONS OF THE VAM USING THE MSH MODEL FOR FOUR-GAUSSIAN MIXTURE COMPONENTS

	EMO1	EMO2	EMO3	EMO4	EMO5	EMO6
EMO1	27.12	18.64	5.08	25.42	22.05	1.69
EMO2	12.86	41.43	18.57	20.00	5.71	1.43
EMO3	2.86	48.56	34.29	14.29	0	0
EMO4	12.50	42.19	6.25	32.81	4.69	1.56
EMO5	35.90	30.77	0	17.95	15.38	0
EMO6	23.81	0	0	0	42.86	33.33

TABLE XIV
CONFUSION MATRIX BETWEEN SIX EMOTIONS OF THE VAM USING THE MSN MODEL FOR FOUR-GAUSSIAN MIXTURE COMPONENTS

	EMO1	EMO2	EMO3	EMO4	EMO5	EMO6
EMO1	44.07	23.73	3.39	22.03	6.78	0
EMO2	17.14	52.86	17.14	11.43	1.43	0
EMO3	2.86	39.99	40.00	14.29	2.86	0
EMO4	29.69	23.44	6.25	35.94	3.12	1.56
EMO5	17.95	25.64	0	7.69	41.03	7.69
EMO6	42.86	4.76	0	0	9.52	42.86

The average classification accuracy of each training method for different number of Gaussian mixture components is summarized in Table XI and Table XII. For one-Gaussian mixture component in the binary emotion classification, the MSH and MSN yielded, respectively, 6.09% and 10.24% relative performance improvements from the ML while the MMI yielded 5.11% relative performance improvement. For one-Gaussian mixture component in the six-class emotion classification, the MSH and MSN yielded respectively 6.54% and 9.14% relative performance improvements from the ML while the MMI yielded 3.28% relative performance improvement. We obtained that binary emotion classification yields better accuracy than six-class emotion classification since binary emotion classification has less number of competing labels than six-class emotion classification, and thus, the separation margin obtained in the binary classification is generally going to be larger than the margin obtained in the six-class emotion classification. Also, for other mixture components, Table XI and Table XII shows that the MSH and MSN yield better result than the ML and MMI. Also, the MSN improves the classification accuracy over the MSH. As the number of Gaussian mixture components increases, the classification accuracy also increases. We increased the number of Gaussian mixture component only up

TABLE XV
AVERAGE ACCURACY(%) OF CROSS-DB EMOTION CLASSIFICATION

	ML	ML→MMI	ML→MSH	ML→MSN	ML→MSL	ML→MSE
1-mix	27.56	29.55	31.82	33.03	32.68	32.25
2-mix	29.49	30.69	32.12	34.24	32.47	29.55
4-mix	36.36	38.64	41.67	40.91	41.23	40.20
8-mix	41.02	43.18	46.60	47.16	46.05	45.46
16-mix	43.18	45.46	47.10	48.69	48.18	47.04
32-mix	45.46	47.73	50.68	51.28	51.09	50.23

to four since no remarkable performance improvement was observed for larger Gaussian mixture component. The highest accuracy of 68.23% (higher than the result in [51]) and 38.54% was obtained using the MSN with four-mixture components in the binary emotion classification and six-class emotion classification respectively.

The confusion matrices of the VAM using the MSH and MSN model given four-Gaussian mixture components for six-class emotion classification are shown in Table XIII and Table XIV, respectively. We can see the effect of the margin scaling from these two confusion matrices. For example, given EMO1 (in G3), the loss of EMO5 (in G7) are higher than that of EMO3 (in G1). Thus, given the EMO1, the rate of predicting EMO5 is further reduced (22.05%→6.78%) while that of predicting EMO3 decreases (5.08%→3.39%) as shown in Table XIII and XIV. Overall, using (10) for all testing data, we observed that the MSN yields 6.35% relative loss reduction from the MSH for four-mixture components. Also, we could reduce the overall loss in the VAM.

We compared the accuracy obtained by three loss functions. In all Gaussian mixture components, MSN outperformed MSL and MSE, and MSL outperformed MSE. In the VAM, emotion models were well separated when the linear loss function was used.

E. Cross Database Experiment

In this section, we perform cross database experiment to show the generalization ability in the acoustically different environment: training using EMO-DB+SUSAS and testing using DES. In our experiments, We chose three overlapping emotions: happy, sad, and neutral. We extract and combine emotional features such as MFCCs, pitch, log energy and zero crossing rate. We use a 78-dimensional feature vector: 12 MFCCs, 12 pitches, log energy, zero crossing rate and corresponding delta and acceleration coefficients.

The average classification accuracy of each training method for different number of Gaussian mixture components is summarized in Table XV. For one-Gaussian mixture component, the MSH and MSN yielded respectively 15.46% and 19.84% relative performance improvements from the ML while the MMI yielded 7.22% relative performance improvement. Also, for other mixture components, Table XV shows that the MSH and MSN yield better result than the ML and MMI. Also, the MSN improves the classification accuracy over the MSH. As the number of Gaussian mixture components increases, the classification accuracy also increases. The highest accuracy of 51.28% (which is comparable to the result in [50]) was obtained using the MSN with 32-mixture components.

We compared the accuracy obtained by three loss functions. In all Gaussian mixture components, MSN outperformed MSL and MSE, and MSL outperformed MSE. In the cross database experiment, three emotions were well separated when the linear loss function was used.

V. CONCLUSION

We presented loss-scaled large margin GMMs for speech emotion classification. In the learning framework, the margin is scaled by the Hamming loss function and three loss functions using WTM-based distance metric. Margin scaling is known to have good generalization ability especially when over-fitting occurs due to the small size of the database as in the case for emotion modeling. We defined a distance metric based on the WTM, and the loss function was computed by a function of the distance metric. We used the MFCCs, log energy, pitch, and zero crossing rate and the corresponding delta and acceleration coefficients as emotional features. Each emotion was modeled using a GMM, and the GMM parameter set was estimated by six different learning criteria: the ML, MMI, MSH, MSN, MSL, and MSE.

All experiments including cross database experiment were conducted using the EMO-DB, SUSAS, DES, and VAM. In the experiments, we observed that the MSH and MSN have better generalization ability and yield less speaker-dependent models than the ML and MMI. In the EMO-DB and DES experiments, the MSH and MSN considerably improved the classification accuracy over the ML and MMI. Also, the MSN reduced the classification error and the risk of misclassifying into the emotions with high loss more than the MSH. However, in the SUSAS experiments, the MSH and MSN slightly improved the classification accuracy over the ML and MMI due to the small difference in speaker characteristics between training and testing data set. Still, the MSN reduced the risk of misclassifying into the emotions with high loss more than the MSH. Using the VAM, we performed two experiments: 1) binary emotion classification using only the valence value and 2) six-class emotion classification by k-means clustering using three dimensional primitives that include valence, activation, and dominance. Again, we observed that our learning framework is effective in improving the classification accuracy. The improvement from the baseline ML model is more effective in the binary emotion classification than in the six-class emotion classification. This is due to the fact that in binary emotion classification there are only one competing label in obtaining the separation margin as opposed to five in the six-class classification, and therefore, the separation margin obtained in the binary classification is generally going to be larger than the margin obtained in the six-class

emotion classification. In cross database experiment, we used EMO-DB+SUSAS as training data set and DES as testing data set. Given testing data set that is "acoustically different" (not only different in speakers but also in microphones and recording conditions) from the training data set, our learning framework also showed improvements in classification accuracy compared to other learning frameworks. In summary, the effect of the learning framework of margin scaling is remarkable in the cases where the parameter over-fitting is considerable due to the small number of data and large difference in speaker characteristics.

In this paper, we used utterances expressed with one emotion (not a sequence of different emotions) in the experiments and observed that the loss functions based on the WTM are more effective than the Hamming loss function in reducing the risk of misclassifying into the emotions which are very different from the true emotion.

APPENDIX
IMPLEMENTATION PROCEDURE FOR MARGIN SCALING USING THE SDP

In this section, we describe a procedure for solving the constrained optimization problem (8) using the SDP. In [40], a max-margin framework for separable case was expressed as a SDP. Based on the procedure, we extend it to the margin scaling case. The SDP is an optimization problem to find a symmetric positive semi-definite matrix $Z_j (j = 1, \ldots, J)$ such that a linear objective function is minimized under L linear constraints [41]:

$$\begin{aligned} \min_{Z_j} \quad & \sum_{j=1}^{J} \langle A_j, Z_j \rangle \\ \text{subject to} \quad & \sum_{j=1}^{J} \langle B_{ij}, Z_j \rangle = b_i, \quad i = 1, \ldots, L \end{aligned} \tag{13}$$

where A_j and B_{ij} are symmetric matrices, and b_i is real scalar. We define the inner product of A_j and Z_j as

$$\langle A_j, Z_j \rangle = \operatorname{tr}\left[A_j' Z_j\right] = \operatorname{tr}[A_j Z_j] \tag{14}$$

where $\operatorname{tr}[\cdot]$ is the matrix trace operator.

The objective is expressing the optimization problem (8) as a SDP problem (13). We explain the SDP expression of the optimization problem by seven steps.

1) Approximation of the discriminant function: we approximate the discriminant function by choosing the dominant mixture component.
2) Expression of the approximated discriminant function as a sum of traces of two matrices: we express the approximated discriminant function as a sum of traces of two matrices.
3) Expression of the approximated discriminant function as a sum of inner products of two symmetric matrices: we express the discriminant function as the form in (14).
4) Expression of the inequality constraint as a equality constraint by introducing slack variables: the SDP can solve the inequality-constrained problem by changing the inequality constraint into an equality constraint with a slack variable.
5) Expression of the L_2-norm restriction of θ: we express the L_2-norm restriction of θ as the form in (14).
6) Expression of the proposed formulation: the proposed optimization problem is expressed as the SDP.
7) Additional constraints: we explain the additional constraints for solving the proposed criterion by the SDP.

In the following, we will explain each step in more detail.

1) Approximation of the Discriminant Function: We approximate the sum of mixture components by the "Viterbi" approximation [40], [54]: choosing a dominant one as

$$\begin{aligned} F(\mathbf{X}_n, \mathbf{y}_n; \theta) &= \log\left[\frac{1}{M}\prod_{t=1}^{T}\sum_{k=1}^{K} w_{nk} N(\mathbf{x}_{nt}; \boldsymbol{\mu}_{nk}, \boldsymbol{\Lambda}_{nk})\right] \\ &\approx \log\left[\frac{1}{M}\prod_{t=1}^{T}\max_{k}\left[w_{nk} N(\mathbf{x}_{nt}; \boldsymbol{\mu}_{nk}, \boldsymbol{\Lambda}_{nk})\right]\right] \\ &= \log\left[\frac{1}{M}\prod_{t=1}^{T} w_{n_t}\mathcal{N}\left(\mathbf{x}_{nt}; \boldsymbol{\mu}_{n_t}, \boldsymbol{\Lambda}_{n_t}\right)\right] \end{aligned} \tag{15}$$

where w_{nk}, $\boldsymbol{\mu}_{nk}$, and $\boldsymbol{\Lambda}_{nk}$ are the mixture weight, mean vector, and covariance matrix of index of kth Gaussian mixture component for $\mathbf{y}_n$, respectively, and w_{n_t}, $\boldsymbol{\mu}_{n_t}$, and $\boldsymbol{\Lambda}_{n_t}$ are the mixture weight, mean vector, and covariance matrix of the dominant mixture component at t, respectively, and $\mathbf{x}_{nt}$ is the feature vector of tth frame for $\mathbf{X}_n$.

We can also approximate the discriminant function $F(\mathbf{X}_n, \mathbf{y}; \theta)$ for $\mathbf{y} \neq \mathbf{y}_n$:

$$F(\mathbf{X}_n, \mathbf{y}; \theta) \approx \log\left[\frac{1}{M}\prod_{t=1}^{T} w_{\widetilde{n}_t}\mathcal{N}\left(\mathbf{x}_{nt}; \boldsymbol{\mu}_{\widetilde{n}_t}, \boldsymbol{\Lambda}_{\widetilde{n}_t}\right)\right], \ \mathbf{y} \neq \mathbf{y}_n$$

where $\widetilde{n}_t$ is used for the dominant mixture component at t given $\mathbf{y}, \mathbf{y} \neq \mathbf{y}_n$.

2) Expression of the Approximated Discriminant Function as a Sum of Traces of Two Matrices: The right-hand side of above equation can be expressed as

$$c_{\widetilde{n}} - \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D}\left(\frac{x_{nt}^d - \mu_{\widetilde{n}_t}^d}{\lambda_{\widetilde{n}_t}^d}\right)^2 \tag{16}$$

where $x_{nt}^d, \mu_{\widetilde{n}_t}^d$, and $(\lambda_{\widetilde{n}_t}^d)^2$ are, respectively, dth element of the vector $\mathbf{x}_{nt}$, dth element of the vector $\boldsymbol{\mu}_{\widetilde{n}_t}$, and dth diagonal element of the matrix $\boldsymbol{\Lambda}_{\widetilde{n}_t}$, and

$$c_{\widetilde{n}} = \log\left[\frac{1}{M}\prod_{t=1}^{T} w_{\widetilde{n}_t}\right] + \log\frac{1}{(2\pi)^{D/2}\sqrt{\left|\boldsymbol{\Lambda}_{\widetilde{n}_t}\right|}}.$$

We define two vectors: let $\overline{\mathbf{x}}_t^{\widetilde{nn}}$ and $\overline{\boldsymbol{\mu}}_{\widetilde{n}_t}$ be

$$\overline{\mathbf{x}}_t^{\widetilde{nn}} := \left[\frac{x_{nt}^1}{\lambda_{\widetilde{n}_t}^1}; \cdots; \frac{x_{nt}^D}{\lambda_{\widetilde{n}_t}^D}\right]$$

$$\overline{\boldsymbol{\mu}}_{\widetilde{n}_t} := \left[\frac{\mu_{\widetilde{n}_t}^1}{\lambda_{\widetilde{n}_t}^1}; \cdots; \frac{\mu_{\widetilde{n}_t}^D}{\lambda_{\widetilde{n}_t}^D}\right].$$

Then,

$$F(\mathbf{X}_n,\mathbf{y};\theta) \approx c_{\widetilde{n}} - \frac{1}{2}\sum_{t=1}^{T}\sum_{d=1}^{D}\left(\frac{x_{nt}^d - \mu_{\widetilde{n}_t}^d}{\lambda_{\widetilde{n}_t}^d}\right)^2$$
$$= c_{\widetilde{n}} - \frac{1}{2}\sum_{t=1}^{T}\left(\overline{\mathbf{x}}_t^{n\widetilde{n}} - \overline{\boldsymbol{\mu}}_{\widetilde{n}_t}\right)'\left(\overline{\mathbf{x}}_t^{n\widetilde{n}} - \overline{\boldsymbol{\mu}}_{\widetilde{n}_t}\right)$$
$$= c_{\widetilde{n}} - \sum_{t=1}^{T}\mathrm{tr}\left[W_t^{n\widetilde{n}} Z_{\widetilde{n}_t}\right] \quad (17)$$

where

$$W_t^{n\widetilde{n}} = \frac{1}{2}\begin{bmatrix} \left(\overline{\mathbf{x}}_t^{n\widetilde{n}}\right)'\overline{\mathbf{x}}_t^{n\widetilde{n}} & -\left(\overline{\mathbf{x}}_t^{n\widetilde{n}}\right)' \\ -\left(\overline{\mathbf{x}}_t^{n\widetilde{n}}\right)' & I_D \end{bmatrix}, \quad (18)$$

$$Z_{\widetilde{n}_t} = \begin{bmatrix} 1 & \left(\overline{\boldsymbol{\mu}}_{\widetilde{n}_t}\right)' \\ \overline{\boldsymbol{\mu}}_{\widetilde{n}_t} & \overline{\boldsymbol{\mu}}_{\widetilde{n}_t}\left(\overline{\boldsymbol{\mu}}_{\widetilde{n}_t}\right)' \end{bmatrix} \quad (19)$$

and I_D is D-dimensional identity matrix.

3) Expression of the Approximated Discriminant Function as a Sum of Inner Products of Two Symmetric Matrices: Let G be the set of all Gaussian mixture components for all labels. We have $|G| = KM$ since there are K mixture components for all labels $\mathbf{y}_1,\ldots,\mathbf{y}_M$. Let $\mathcal{K} := KM$. Then, the dominant mixture component index $\widetilde{n}_t$ is in $(1,\ldots,\mathcal{K})$, and $\overline{\boldsymbol{\mu}}_{\widetilde{n}_t}$ is the mean vector of one elements in G. Thus, the summation in (17) can be expressed as

$$\sum_{t=1}^{T}\mathrm{tr}\left[W_t^{n\widetilde{n}} Z_{\widetilde{n}_t}\right]$$
$$= \mathrm{tr}\left[\sum_{t=1}^{T} W_t^{n\widetilde{n}} Z_{\widetilde{n}_t}\right]$$
$$= \mathrm{tr}\left[\sum_{t=1,\widetilde{n}_t=1}^{T} W_t^{n\widetilde{n}} Z_1 + \cdots \sum_{t=1,\widetilde{n}_t=\mathcal{K}}^{T} W_t^{n\widetilde{n}} Z_{\mathcal{K}}\right]$$
$$= \mathrm{tr}\left[\sum_{j=1}^{\mathcal{K}} V_j^{n\widetilde{n}} Z_j\right] = \sum_{j=1}^{\mathcal{K}}\mathrm{tr}\left[V_j^{n\widetilde{n}} Z_j\right] \quad (20)$$

where

$$V_j^{n\widetilde{n}} = \sum_{t=1,\widetilde{n}_t=j}^{T} W_t^{n\widetilde{n}}.$$

From (17) and (20), the discriminant function and the constraints in (8) are approximately

$$F(\mathbf{X}_n,\mathbf{y};\theta) \approx c_{\widetilde{n}} - \sum_{j=1}^{\mathcal{K}}\mathrm{tr}\left[V_j^{n\widetilde{n}} Z_j\right]$$
$$= c_{\widetilde{n}} - \sum_{j=1}^{\mathcal{K}}\left\langle V_j^{n\widetilde{n}}, Z_j\right\rangle, \quad \mathbf{y}\neq\mathbf{y}_n \quad (21)$$

and

$$d_n(\mathbf{y};\theta) = F(\mathbf{X}_n;\mathbf{y}_n;\theta) - F(\mathbf{X}_n;\mathbf{y};\theta)$$
$$\approx (c_n - c_{\widetilde{n}}) - \sum_{j=1}^{\mathcal{K}}\left\langle\left(V_j^{nn} - V_j^{n\widetilde{n}}\right), Z_j\right\rangle$$
$$\geq \rho\Delta(\mathbf{y}_n,\mathbf{y}) - \xi_n, \quad \forall n, \forall\mathbf{y}\in\mathcal{Y}\setminus\mathbf{y}_n. \quad (22)$$

4) Expression of the Inequality Constraint as a Equality Constraint by Introducing Slack Variables: The above inequality constraint can be converted into the equality constraint by introducing an additional positive slack variable $s_{n\widetilde{n}}$ as

$$\sum_{j=1}^{\mathcal{K}+1}\left\langle\left(V_j^{nn} - V_j^{n\widetilde{n}}\right), Z_j\right\rangle + \rho\Delta(\mathbf{y}_n,\mathbf{y}) - \xi_n + s_{n\widetilde{n}} = (c_n - c_{\widetilde{n}}). \quad (23)$$

5) Expression of the L_2-Norm Restriction of θ: In the proposed formulation, we restrict the L_2-norm of $\|\theta\|$ to γ, or equivalently $\|\theta\|^2 = \gamma^2$. The parameter set θ includes only $\boldsymbol{\mu}_j$ for $j = 1,\ldots,\mathcal{K}$ since we consider estimating only mean vectors of Gaussian mixture components. Thus,

$$\|\theta\|^2 = \sum_{j=1}^{\mathcal{K}}\boldsymbol{\mu}_j'\boldsymbol{\mu}_j$$
$$= \sum_{j=1}^{\mathcal{K}}\langle R_j, Z_j\rangle \quad (24)$$

where

$$R_j = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_j \end{bmatrix} \quad (25)$$

for $j = 1,\ldots,\mathcal{K}$.

6) Expression of the Proposed Formulation: From (22), (24) and (25), the constrained optimization problem in (8) can be expressed as

$$\min_{Z_j} \quad \sum_{j=0}^{\mathcal{K}+1}\langle A_j, Z_j\rangle$$
$$\text{subject to} \quad \sum_{j=0}^{\mathcal{K}+1}\left\langle B_j^{n\widetilde{n}}, Z_j\right\rangle = b_{n\widetilde{n}}, \quad \forall n,\ \forall\mathbf{y}\in\mathcal{Y}\setminus\mathbf{y}_n,$$
$$\sum_{j=0}^{\mathcal{K}+1}\left\langle H_j^i, Z_j\right\rangle = 1, \quad i = 1,\ldots,\mathcal{K},$$
$$\sum_{j=0}^{\mathcal{K}+1}\langle R_j, Z_j\rangle = \gamma^2 \quad (26)$$

where

$$A_j = \mathbf{0}, \quad (27)$$
$$B_j^{n\widetilde{n}} = V_j^{nn} - V_j^{n\widetilde{n}} \quad (28)$$

for $j = 1,\ldots,\mathcal{K}$ and $b_{n\widetilde{n}} = c_n - c_{\widetilde{n}}$.

7) Additional Constraints: The matrices, A_0, Z_0 and $B_0^{\widetilde{nn}}$, are $(N+1)\times(N+1)$ diagonal matrices

$$A_0 = \text{diag}\left(-1; \frac{C}{N}; \cdots; \frac{C}{N}\right) \tag{29}$$

$$Z_0 = \text{diag}(\rho; \xi_1; \cdots; \xi_N) \tag{30}$$

$$B_0^{\widetilde{nn}} = \text{diag}\left(\Delta(\mathbf{y}_n, \mathbf{y}); -\boldsymbol{\mathcal{I}}_n\right), \quad \mathbf{y} \neq \mathbf{y}_n \tag{31}$$

where $-\boldsymbol{\mathcal{I}}_n$ has zero elements except that the nth element is 1. Remaining matrices, $A_{\mathcal{K}+1}$, $Z_{\mathcal{K}+1}$ and $B_{\mathcal{K}+1}^{\widetilde{nn}}$ are, respectively, the zero matrix, diagonal matrix with elements $s_{\widetilde{nn}}$, and the zero matrix except that the element at the corresponding position of $s_{\widetilde{nn}}$ is 1 so that $B_{\mathcal{K}+1}^{\widetilde{nn}} Z_{\mathcal{K}+1} = s_{\widetilde{nn}}$. The matrix H_j^i is introduced for the constraint where the element of the first diagonal element of Z_j is 1 as in (19). Thus,

$$H_j^i = \begin{cases} 1, & \text{at the first diagonal element} \\ 0, & \text{anywhere else} \end{cases} \tag{32}$$

for $i = j$. Otherwise, H_j^i is zero matrix. Remaining constraint on Z_j is

$$Z_j = \begin{bmatrix} 1 & (\overline{\boldsymbol{\mu}}_j)' \\ \overline{\boldsymbol{\mu}}_j & U \end{bmatrix} \tag{33}$$

where U must be $\overline{\boldsymbol{\mu}}_j(\overline{\boldsymbol{\mu}}_j)'$. Since the constraint is not linear, we relax it such that

$$U - \overline{\boldsymbol{\mu}}_j(\overline{\boldsymbol{\mu}}_j)' \succeq 0 \tag{34}$$

by the property [55] that

$$Z_j \succeq 0 \quad \Leftrightarrow \quad U - \overline{\boldsymbol{\mu}}_j(\overline{\boldsymbol{\mu}}_j)' \succeq 0 \tag{35}$$

where $Z_j \succeq 0$ denotes that Z_j is a positive semi-definite matrix.

As shown in (26), we express (8) as a SDP. Then, Z_j can be obtained using the DSDP which is a solver for SDP and update the mean vector of GMM parameter set from $[1(\overline{\boldsymbol{\mu}}_j)']$ (the first row of Z_j).

REFERENCES

[1] M. Lew, E. M. Bakker, N. Sebe, and T. S. Huang, "Human–computer intelligent interaction: A survey," *Lecture Note Comput. Sci.*, vol. 4796, pp. 1–5, 2007.

[2] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[3] J. Tao and T. Tan, "Affective computing: A review," *Lecture Note Comput. Sci.*, vol. 3784, pp. 981–995, 2005.

[4] I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition from facial expressions using multilevel HMM," in *Proc. Neural Inf. Process. Syst., Workshop Affective Comput.*, 2000.

[5] S. Casale, A. Russo, and G. Scebba, "Speech emotion classification using machine learning algorithms," in *Proc. IEEE Int. Conf. Semantic Comput.*, 2008, pp. 158–165.

[6] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition—A new approach," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2004, vol. 2, pp. 1020–1025.

[7] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *J. Netw. Comput. Applicat.*, vol. 30, no. 4, pp. 1334–1345, 2007.

[8] N. Garay, I. Cearreta, J. M. López, and I. Fajardo, "Assistive technology and affective mediation," *Assistive Technol.*, vol. 2, no. 1, pp. 55–83, 2006.

[9] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," *Lecture Note Comput. Sci.*, vol. 3068, pp. 36–48, 2004.

[10] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition," *Signal Process.*, vol. 88, no. 12, pp. 2869–3014, 2008.

[11] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, "Floating search method for feature selection with nonmonotonic criterion functions," *Pattern Recognit.*, vol. 2, pp. 279–283, 1994.

[12] S. Casale, A. Russo, and S. Serrano, "Multistyle classification of speech under stress using feature subset selection based on genetic algorithms," *Speech Commun.*, vol. 49, pp. 801–810, 2007.

[13] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, and K. M. Leong, "Stress and emotion classification using jitter and shimmer features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. 1081–1084.

[14] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proc. Eur. Signal Process. Conf.*, 2004, pp. 341–344.

[15] M. Lugger and B. Yang, "An incremental analysis of different feature groups in speaker independent emotion recognition," in *Proc. Int. Congr. Phonet. Sci.*, 2007, pp. 2149–2152.

[16] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2003, pp. 125–128.

[17] Y. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," in *Proc. IEEE Int. Conf. Mach. Learn. Cybern.*, 2005, vol. 8, pp. 18–21.

[18] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[19] L. Fu, X. Mao, and L. Chen, "Speaker independent emotion recognition based on SVM/HMMs fusion system," in *Proc. IEEE Int. Conf. Audio, Lang. Image Process.*, 2008, pp. 61–65.

[20] B. Schuller, S. Reiter, R. Muller, M. Al-Hamas, M. Lang, and G. Riqoll, "Speaker independent speech emotion recognition by ensemble classification," in *Proc. IEEE-ICME*, 2005, pp. 864–867.

[21] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 941–944.

[22] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, vol. 16.

[23] I. Tsochantaridis, T. Joachims, and T. Hofmann, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, 2005.

[24] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 2000.

[25] G. Heigold, T. Deselaers, R. Schlüter, and H. Ney, "Modified MMI/MPE: A direct evaluation of the margin in speech recognition," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 384–391.

[26] F. Sha and L. K. Saul, "Large margin hidden Markov models for automatic speech recognition," *Neural Inf. Process. Syst.*, vol. 19, pp. 1249–1256, 2007.

[27] S. Sarawagi and R. Gupta, "Accurate max-margin training for structured output spaces," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 888–895.

[28] A. Tellegen, D. Watson, and L. Clark, "On the dimensional and hierarchical structure of affect," *Psychol. Sci.*, vol. 10, no. 4, pp. 297–303, 1999.

[29] D. Watson and A. Tellegen, "Toward a consensual structure of mood," *Psychol. Bull.*, vol. 98, no. 2, pp. 219–35, 1985.

[30] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," *J. Personal. Soc. Psychol.*, vol. 54, no. 2, pp. 1063–1070, 1988.
[31] S. Yun and C. D. Yoo, "Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 4169–4172.
[32] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[33] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1986, vol. 1, pp. 49–52.
[34] A. B. Yishai and D. Burshtein, "A discriminative training algorithm for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 204–217, May 2004.
[35] Y. Normandin, R. Cardin, and R. D. Mori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 299–311, Apr. 1994.
[36] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional vesus conditional maximum likelihood," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 4, pp. 814–817, Aug. 1983.
[37] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, 2001.
[38] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "(Online) Subgradient methods for structured prediction," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2007.
[39] Y. Yin and H. Jiang, "A compact semidefinite programming (SDP) formulation for large margin estimation of HMMs in speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2007, pp. 312–317.
[40] X. Li and H. Jiang, "Solving large-margin hidden Markov model estimation via semidefinite programming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2383–2392, Nov. 2007.
[41] S. J. Benson and Y. Ye, "Algorithm 875: DSDP5—Software for semidefinite programming," *Res. J. Assoc. for Comput. Machinery Math. Software*, vol. 34, no. 3, pp. 16:1–16:20, 2008.
[42] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Int. Conf. Spoken Lang. Process.*, 2005, pp. 1517–1520.
[43] J. Hansen and S. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1997, vol. 4, pp. 1743–1746.
[44] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1997, pp. 1695–1698.
[45] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittang German audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008.
[46] Y. Qiao, M. Suzuki, and N. Minematsu, "Affine invariant features and their application to speech recognition," in *Proc. IEEE Int. Conf. Acousti., Speech, Signal Process.*, 2009, pp. 4629–4632.
[47] O. Kwon and T. Lee, "Phoneme recognition using ica-based feature extraction and transformation," *Signal Process.*, vol. 84, no. 6, pp. 1005–1019, 2004.
[48] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, 2008.
[49] 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)*. Cambridge, U.K.: Univ. of Cambridge, 2002.
[50] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affective Comput.*, vol. 1, no. 2, pp. 119–132, Jul.-Dec. 2010.
[51] S. Casale, A. Russo, and S. Serrano, "Analysis of robustness of attributes selection applied to speech emotion recognition," in *Proc. Eur. Signal Process. Conf.*, 2010, pp. 1174–1178.
[52] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2005, pp. 381–385.
[53] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10–11, pp. 787–800, 2007.
[54] K. Yao, K. K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," *Speech Commun.*, vol. 42, no. 1, pp. 5–23, 2004.
[55] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1994, vol. 15.

Sungrack Yun (S'06) received the B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2003. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, KAIST.

His research interest is in machine learning for signal processing.

Chang D. Yoo (S'92–M'96) received the B.S. degree in engineering and applied science from the California Institute of Technology, Pasadena, in 1986, the M.S. degree in electrical engineering from Cornell University, Ithaca, NY, in 1988, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996.

From January 1997 to March 1999, he worked at Korea Telecom as a Senior Researcher. He joined the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, in April 1999. From March 2005 to March 2006, he was with the Research Laboratory of Electronics, MIT. His current research interests are in the application of machine learning and digital signal processing in multimedia.

Dr. Yoo is a member of Tau Beta Pi and Sigma Xi. Prof. He currently serves on the Machine Learning for Signal Processing (MLSP) Technical Committee of the IEEE Signal Processing Society.