Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 475810, 10 pages
http://dx.doi.org/10.1155/2015/475810

Research Article

A Robust Method for Speech Emotion Recognition Based on Infinite Student's t-Mixture Model

Xinran Zhang, Huawei Tao, Cheng Zha, Xinzhou Xu, and Li Zhao

School of Information Science and Engineering, Southeast University, Nanjing 210096, China

Correspondence should be addressed to Xinran Zhang; zxrzxr87324@126.com

Received 27 April 2015; Revised 18 September 2015; Accepted 27 September 2015

Academic Editor: Chih-Cheng Hung

Speech emotion classification method, proposed in this paper, is based on Student's t-mixture model with infinite component number (iSMM) and can directly conduct effective recognition for various kinds of speech emotion samples. Compared with the traditional GMM (Gaussian mixture model), speech emotion model based on Student's t-mixture can effectively handle speech sample outliers that exist in the emotion feature space. Moreover, t-mixture model could keep robust to atypical emotion test data. In allusion to the high data complexity caused by high-dimensional space and the problem of insufficient training samples, a global latent space is joined to emotion model. Such an approach makes the number of components divided infinite and forms an iSMM emotion model, which can automatically determine the best number of components with lower complexity to complete various kinds of emotion characteristics data classification. Conducted over one spontaneous (FAU Aibo Emotion Corpus) and two acting (DES and EMO-DB) universal speech emotion databases which have high-dimensional feature samples and diversiform data distributions, the iSMM maintains better recognition performance than the comparisons. Thus, the effectiveness and generalization to the high-dimensional data and the outliers are verified. Hereby, the iSMM emotion model is verified as a robust method with the validity and generalization to outliers and high-dimensional emotion characters.

1. Introduction

In naturalistic human-computer interaction (HCI), speech emotion recognition (SER) is becoming increasingly important in various applications. In the literature, the studies on speech emotion recognition may be categorized into the following: (a) the searching for robust speech emotion features, (b) the study on emotion classification models (Gaussian mixture model (GMM) [1] and regression are adopted, resp., for discrete emotion model and continual dimensional model [2–4]), (c) the study towards naturalistic emotional behavior, and (d) the approaches that consider the complex environment factors, for example, context dependent [5], cross-language, and gender dependent speech emotion recognition [6].

On account of simple structures and easy implementation, classifier models with Gaussian kernel are commonly used in speech emotion recognition. However, the division performance in space of Gaussian kernel largely depends on the characterization of characters in the training samples themselves. If a certain proportion of outliers (singular points) arise in the actual emotion feature samples, it may lead to the surge in number of divisory components, which would seriously affect the recognition performance. Moreover, in previous research on emotion models [7–9], the model component numbers usually need to be gained through a plenty of training on data. Then, when the underfitting of emotion characteristics leads to mismatching between test samples and training samples, the model component number division will be affected, which results in the performance degradation of classifiers. To solve the above problems, this paper introduced feature classification method of speech emotion based on Student's t-mixture model with infinite component numbers (iSMM). First of all, based on nonparametric Bayesian statistical distribution [10], Student's t-distribution emotion model is set up for high-dimensional space modeling of speech emotion features, while according to the appropriate numbers of components characteristics classification proceeds effectively. Then, the iSMM models are based on normal distributions [11]. Compared to other

models, their unique "long-tail" structures make them more robust when handling outliers in speech emotion features and monitoring data with underfitting sample spaces. At last, profiting from latent variables with the hybrid model distributions, the iSMM emotion model in this paper can automatically determine the best composition numbers to conduct the optimal partition of feature space. Meanwhile, outliers in various kinds of speech emotion characters can be effectively dealt with and the problem of model size selection could be solved simultaneously.

In order to verify our proposed method, we adopt three different databases in two different languages: Danish and German. The cross-language speech emotion recognition is still a challenge. Although the emotion expression has a universal character in different nations and cultures, the acoustic features are greatly impacted by different languages. Hence, some of the emotion recognition algorithms may achieve good result in one language while having poor performance in another one.

The remainder of this paper is organized as follows. Section 2 gives the introduction of Student's t-mixture model. Section 3 provides the detailed establishment of emotion model based on the iSMM. Section 4 gives the experimental results on different databases and different languages. Finally, the conclusion is given in Section 5.

2. Speech Emotion Feature Classifiers Based on Statistical Models

FA (factor analysis) is a statistical model to process high-dimensional observed data of speech emotion. FA consists of low-dimensional latent factors, generally, used for dimensionality reduction of local offline data, such as GMM. The combination of a series of factor analyses can form finite local MFA (mixture factor analysis). However, during the process of MFA applied, standardization may usually lead to distribution errors on several factor components, which can further result in the decline of performance with the presence of underfitting data and outliers. The common solution is adopting an independent dimensionality reduction strategy instead of MFA, such as LDA (Linear Discriminant Analysis) and PCA (Principal Component Analysis). However, not only does this increase the complexity of the classification scheme, but processing the singular values or outliers of speech emotion features is also very difficult.

As shown in Figure 1, Student's t-distribution is introduced to replace the MFA in normal distribution, constructing the finite Student's t-distribution mixture model (SMM) [12]. Specifically, the degree of freedom is integrated into Student's t-distribution as a new parameter, bringing the model distribution a "long tail." So, to outliers and atypical feature data, emotion model could keep robust.

In allusion to speech emotion classifiers based on distribution models, the finite mixture model is applied to assess observed emotion sample data. The density expression of a finite mixture model is formed as

$$\delta(\mathbf{x} \mid \vartheta) = \sum_{g=1}^{G} \tau_g \delta_g\left(\mathbf{x} \mid \boldsymbol{\theta}_g\right), \tag{1}$$

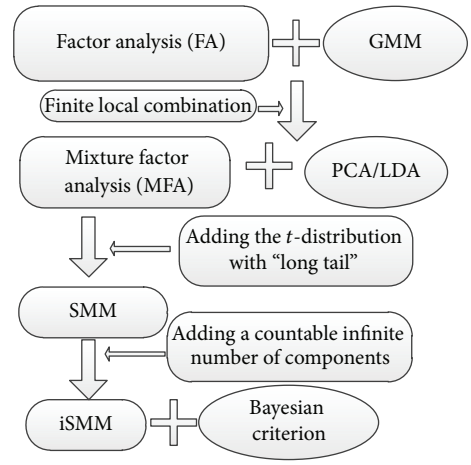


Figure 1: Statistical models for emotion data space.

where $\tau_g > 0$ and the mixture probability $\sum_{g=1}^{G} \tau_g = 1$ is satisfied. $\vartheta = (\tau_1, \ldots, \tau_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters while $\delta_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the density of gth component in the model. In the former research of setting up emotion models for SER [9, 13], multivariate Gaussian distributions are often used in the analytical formula of classifiers to express the component densities, which is shown in Figure 2(a). However, recent studies have consciously focused on mixture models based on non-Gaussian frameworks which were used for classification or clustering in high-dimensional data for SER or other fields [14–17].

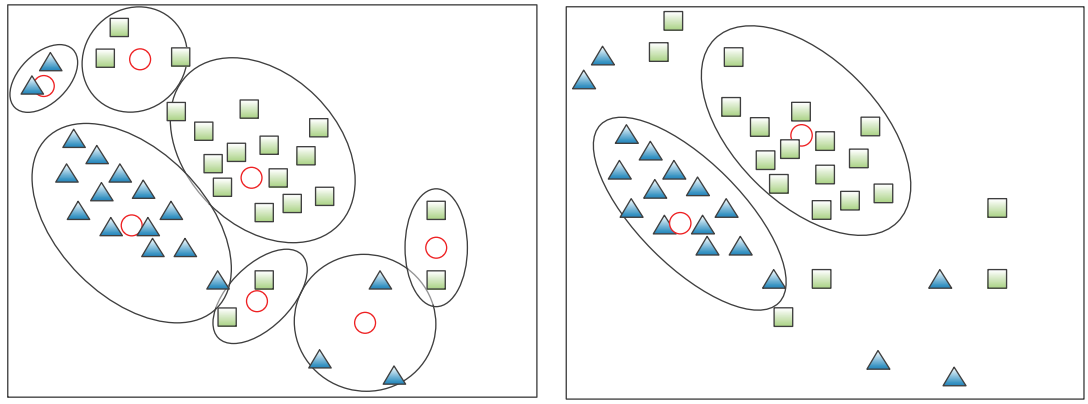
Being similar to Gaussian distribution and Skew-t distribution [18], Student's t-distribution is quite suitable for classifying high-dimensional data or clustering. The AECM (alternating expectation-conditional maximization) algorithm [19] is often used to estimate the model parameters. Meanwhile, the BIC (Bayesian information criterion) [20] may be introduced to determine the types of model.

3. Establishment of the Emotion Model Based on iSMM

In some cases such as exorbitant dimension of observed data (usually in the spontaneous speech emotion databases) or oversize number of components needed to be demarcated in the emotion feature spaces, the parameters of the SMM may be hard to control, thereby leading to the performance degradation of classifiers. Under this motivation, in this work, an infinite mixture countable number of components are introduced, constructing the iSMM model.

As shown in Figure 2(b), for each component, expectations of mean are denoted by "○" and expectations of covariance are represented as the shaded ellipse. So the iSMM propels monitoring data to match the model, rather than aiming at reducing the complexity through adjusting the structure of the models themselves.

3.1. The Emotion Model Based on Student's t-Distribution. The distribution of the emotional samples in the feature



(a) Component demarcation with GMM (b) Component demarcation with iSMM

Figure 2: Plot of fitting multicomponent GMM and iSMM to the emotion feature corpus.

space could be described by multiple superposed Student's t-functions. Theoretically, iSMM model may fit arbitrary density distribution function, as long as there are enough mixed Student's t-components. In our research, iSMM is applied to modeling various emotions. In detail, each emotion category corresponds to an iSMM and is judged through the maximum a posteriori probability criterion. Here, x_j is the jth utterance sample and c_k represents the emotion category where $k = 1, \ldots, K$. Then, the maximum a posteriori probability is

$$P\left(c_k \mid x_j\right) = \frac{P\left(x_j \middle| c_k\right) P\left(c_k\right)}{P\left(x_j\right)}. \tag{2}$$

In (2), $P(c_k|x_j)$ is obtained through each iSMM model. For a given statement sample, the probability of feature vector is a constant. Assume that each emotion possesses equal probability; p_k is the probability of the kth emotion class:

$$p_k = P\left(\mathbf{y} \mid \mathbf{Y}_k\right), \tag{3}$$

where $\mathbf{y}$ is the acoustic feature vector of corresponding speech utterance and $\mathbf{Y}_k$ is the feature distribution model which can be formulated by Student's t-mixture model with iSMM. Hence, identification of samples can be judged by

$$c_k = \arg\max\left\{p_k\right\}, \tag{4}$$

where c_k stands for the emotion class label in the emotion model.

Infinite Student's t-factor mixture distribution, in this paper, is introduced to structure the emotion classification model. Suppose that the number of infinite substructures in a model is countable, and then, according to the Dirichlet process, the mixture model suitable for emotion feature database could be obtained.

p-dimensional emotion characteristics data $\mathbf{y}_n$ $(n = 1, \ldots, N)$ can be described as

$$\mathbf{y}_n = \mathbf{A}\mathbf{u}_{ni} + \mathbf{e}_{ni}, \text{ with prob. } \pi_i. \tag{5}$$

Among them, the probability of feature data sample is set as π_i $(i = 1, \ldots, I)$, where $\sum_1^I \pi_i = 1$ and I is the number of the mixture compositions divided in the emotion feature space. The corresponding q-dimensional $(q < p)$ factor $\mathbf{u}_{ni}$ is distributed independently $t(\mathbf{u}_{ni} \mid \xi_i, \mathbf{\Omega}_i, \nu_i)$, where ν_i is called degrees of freedom, parameter ξ_i is q-dimensional vector, and $\mathbf{\Omega}_i$ is a $q \times q$ positive definite symmetric matrix, which represent the mean and variance of the ith emotion component in low-dimensional feature subspaces. Furthermore, $\mathbf{e}_{ni}$ represents component error of the emotion features, which obeys independent distribution $t(\mathbf{e}_{ni} \mid \mathbf{0}, \mathbf{D}, \nu_i)$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{A}$ is a $p \times q$ matrix shared by all divided emotion components. Therefore it is defined as the common emotion factor loading matrix. The t-distributions $t(\mathbf{u}_{ni} \mid \xi_i, \mathbf{\Omega}_i, \nu_i)$ and $t(\mathbf{e}_{ni} \mid \mathbf{0}, \mathbf{D}, \nu_i)$ can be regarded as average normal scale distributions $\mathcal{N}(\mathbf{u}_{ni} \mid \xi_i, \mathbf{\Omega}_i/w_{ni})$ and $\mathcal{N}(\mathbf{e}_{ni} \mid \mathbf{0}, \mathbf{D}/w_{ni})$ with the precision scalar w_{ni} Gamma distributions; that is,

$$\begin{aligned} &t\left(\mathbf{u}_{ni} \mid \xi_i, \mathbf{\Omega}_i, \nu_i\right) \\ &\quad = \int \mathcal{N}\left(\mathbf{u}_{ni} \mid \xi_i, \frac{\mathbf{\Omega}_i}{w_{ni}}\right) \mathcal{G}\left(w_{ni} \mid \frac{\nu_i}{2}, \frac{\nu_i}{2}\right) \mathrm{d}w_{ni}, \\ &t\left(\mathbf{e}_{ni} \mid \mathbf{0}, \mathbf{D}, \nu_i\right) \\ &\quad = \int \mathcal{N}\left(\mathbf{e}_{ni} \mid \mathbf{0}, \frac{\mathbf{D}}{w_{ni}}\right) \mathcal{G}\left(w_{ni} \mid \frac{\nu_i}{2}, \frac{\nu_i}{2}\right) \mathrm{d}w_{ni}. \end{aligned} \tag{6}$$

Here, data space model based on t-distribution of speech emotion features could be gained, where the parameter set of emotion model, Θ, contains the parameters ξ_i, $\mathbf{\Omega}_i$, ν_i $(i = 1, \ldots, I)$, $\mathbf{A}$, $\mathbf{D}$, and π_i $(i = 1, \ldots, I-1)$.

3.2. Probabilistic Distribution Modeling of Emotion Features. Because the topic of SER this paper discussed belongs to supervised machine learning, in order to accomplish

the mission of emotion classification, the focus on the posterior distribution of stochastic variables of the emotion feature parameters is unnecessary while the research about the emotion characteristics data $\mathbf{y}_n'$ for testing holds the main part in our work.

For this purpose, the distribution of emotion characteristics data tested is evaluated as $P(\mathbf{y}_n' \mid \mathbf{Y})$. According to the principle of the Bayesian framework, it is known that the distribution above can be obtained through the calculation of marginal likelihood about $P(\mathbf{y}_n' \mid \Theta)$ and posterior distribution $P(\mathbf{Y}_k \mid \Theta)$, as the following formula:

$$\begin{aligned}\log p\left(\mathbf{y}_n' \mid \mathbf{Y}\right) &= \log \int \mathrm{d}\Theta p\left(\mathbf{y}_n' \mid \Theta\right) p\left(\Theta \mid \mathbf{Y}\right) \\ &\approx \log \int \mathrm{d}\mathbf{C} q\left(\mathbf{C}\right) \int \mathrm{d}\mathbf{A} q\left(\mathbf{A}\right) \int \mathrm{d}\xi \mathrm{d}\mathbf{\Lambda} q\left(\xi, \mathbf{\Lambda}\right) \\ &\cdot \left\{\sum_{i=1}^{T} p\left(\mathbf{z}_{ni} \mid \pi_i\left(\mathbf{C}\right)\right) p\left(\mathbf{y}_n' \mid \mathbf{z}_{ni}, \mathbf{A}, \xi_i, \mathbf{\Lambda}_i, \mathbf{D}, \mathbf{v}_i\right)\right\},\end{aligned} \tag{7}$$

where the evaluation on posterior distribution of emotion samples is obtained from training data of each type of emotions. In formula (7), $q(\mathbf{C})$, $q(\mathbf{A})$, and $q(\xi_i, \mathbf{\Lambda}_i)$ and parameters $\mathbf{D}$ and $\mathbf{v}$ are gotten from inference prediction of corresponding emotion categories in the infinite Student's t-distribution.

Furthermore, through applying step "E" in expectation-maximization (EM) algorithm [21] to the test data $\mathbf{y}_n'$, the posterior probability of latent variables z_n, which is derivation from emotion characteristics, could be obtained. The iSMM which is composed of several latent weights could approximate the real distribution of samples by the Student's t-model. The estimate of iSMM model includes three parameters: the mixed weight of latent weight, the mean $\boldsymbol{\mu}$, and variance $\sum$ of each t-distribution function. Hereby, the weight of latent variables z_n is

$$a_{z_n} = \frac{1}{N} \sum_{n}^{N} P\left(z_n \mid \mathbf{y}_n, \Theta\right). \tag{8}$$

The recursive formula of mean and variance is

$$\boldsymbol{\mu}_{z_n} = \frac{\sum_n^N \mathbf{y}_n P\left(z_n \mid \mathbf{y}_n, \Theta\right)}{\sum_n^N P\left(z_n \mid \mathbf{y}_n, \Theta\right)}, \tag{9}$$

$$\sum_{z_n} = \frac{\sum_n^N P\left(z_n \mid \mathbf{y}_n, \Theta\right)\left(\mathbf{y}_n - \boldsymbol{\mu}_{z_n}\right)\left(\mathbf{y}_n - \boldsymbol{\mu}_{z_n}\right)^T}{\sum_n^N P\left(z_n \mid \mathbf{y}_n, \Theta\right)}, \tag{10}$$

$$P\left(z_n \mid \mathbf{y}_n, \Theta\right) = \frac{a_{z_n} P_{z_n}\left(\mathbf{y}_n \mid \boldsymbol{\theta}_{z_n}\right)}{\sum_{k=1}^{K} a_k P_k\left(\mathbf{y} \mid \boldsymbol{\theta}_k\right)}, \tag{11}$$

where $P_{z_n}(\mathbf{y} \mid \boldsymbol{\theta})$ and $P_k(\mathbf{y} \mid \boldsymbol{\theta})$ are both Student's t-distribution.

At last, the forecast distribution $\mathbf{y}_n'$ of characteristic in each type of emotions can be calculated according to the analyses above.

3.3. Evaluation: Decision Function of SER Based on Student's t-Distribution. In speech emotion classification, label c_k represents a category of K emotions in data sample set $\mathbf{Y}$, where $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ is the sample set of emotion features, in which $\mathbf{y}_n$ $(n = 1, \ldots, N)$ is p-dimensional emotion data and category number k belongs to feature samples of speech for training. During the process of testing, each emotion data tested adopts the "E" step of EM algorithm to be derived. At the same time, each prediction distribution of features is calculated in emotion categories: $P(\mathbf{y}_n' \mid \mathbf{Y}_k)$. Based on formula (4), the criterion of speech emotion model can be obtained by Student's t-distribution as

$$c_k = \arg\max_{k=0}^{K} F\left(\mathbf{Y}_k, \mathbf{y}_n; \Theta\right), \tag{12}$$

where emotion feature vector $\mathbf{y}_n$ satisfying $1 \le n \le N$ is extracted from the frame of each speech on the database in use.

Based on the judgment criteria (12) and the conditional distribution $\log P(\mathbf{y}_n' \mid \mathbf{Y})$, the emotion model can be established by Student's t-distribution, and the decision function for feature recognition could be inferred as follows:

$$\begin{aligned} c_k &= \arg\max_{k=1}^{K} F\left(\mathbf{Y}_k, \mathbf{y}_n'; \Theta\right) \\ &= \arg\max_{k=1}^{K} \log P_\Theta\left(\mathbf{y}_n' \mid \mathbf{Y}_k\right) \\ &= \arg\max_{k=1}^{K} \left[\log \int \mathrm{d}\Theta P\left(\mathbf{y}_n' \mid \Theta\right) P\left(\Theta \mid \mathbf{Y}\right)\right]. \end{aligned} \tag{13}$$

At the beginning of test on emotion model with Student's t-distribution, step "E" for deriving model parameters is carried out on each test data $\mathbf{y}_n'$. Meanwhile, according to formula (11), the forecast distributions of each emotion category are calculated as $P(\mathbf{y}_n' \mid \mathbf{Y}_1), \ldots, P(\mathbf{y}_n' \mid \mathbf{Y}_K)$. Consequently obtained by formula (13), the emotion category c_k, which possesses the maximum prediction distribution, is the recognition results of emotion feature data tested.

3.4. Analysis on High-Dimensional Feature Space of Speech Emotion Based on iSMM. The emotion model, where variable and observed data is one-to-one correspondence, benefits from Bayesian criterion and the average field theory to obtain the relevant posterior distribution, which can be implemented to determine the model structure for emotion feature classification. For the reason of nonparametric Bayesian statistics principle, the iSMM adjusts component numbers on the basis of the distribution model established.

After the evaluation of decision function for SER accomplished, further analysis on the probability distributions of parameters in the emotion model is necessary. High-dimensional emotion feature samples generated by the iSMM model are $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. The iSMM is formed through joining an accompanying countable infinite number of components to the SMM, where the infinite dimensional latent stochastic variables $\mathbf{z}_n$ introduced are one-to-one correspondence with emotional features data $\mathbf{y}_n$ for experiment. The constraint conditions of the latent emotion feature elements

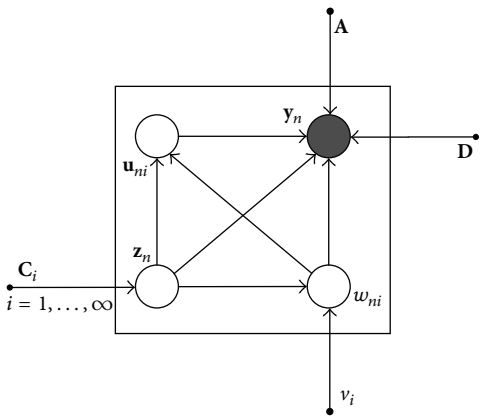


Figure 3: Probabilistic graphical model representation of the iSMM emotion model. Circles represent stochastic variables, and arrows describe conditional dependence between these variables.

in $\mathbf{z}_n$ are $\mathbf{z}_n \in \{0, 1\}$ and $\sum_i \mathbf{z}_n = 1$. In Student's t-distribution, because $\mathbf{y}_n$ is the mixed component of emotion features stemming from the ith partition on the sample space, thus the specific feature element here satisfies $\mathbf{z}_n = 1$ and other elements are all 0. Therefore, the entirety latent variables space of features could be set as $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. According to formula (3), the marginal distribution of $\mathbf{Z}$ conforms to the mixture probabilistic $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_\infty\}$, which is given as

$$P(\mathbf{Z} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{i=1}^{\infty} \pi_i^{z_{ni}}. \tag{14}$$

Then, based on the principle of nonparametric Bayesian statistics, $\boldsymbol{\pi}$ could be obtained by the stick-breaking [22] method.

With respect to the data space $\mathbf{y}'_n$ of emotion features tested, in which it is assumed that samples are obeying identically independent distribution, the decision function can be represented as

$$\begin{aligned} F\left(\mathbf{Y}_k, \mathbf{y}'_n; \Theta\right) &= \log P_\Theta\left(\mathbf{y}'_n \mid \mathbf{Y}_k\right) = \log q\left(\mathbf{u} \mid \mathbf{z}, \mathbf{w}\right) \\ &= \prod_{n=1}^{N} \prod_{i=1}^{T} \mathcal{N}\left(\mathbf{u}_{ni} \mid \widetilde{\boldsymbol{\mu}}_{\mathbf{u}_{ni}}, \frac{\widetilde{\sum}_{\mathbf{u}_{ni}}}{w_{ni}}\right)^{z_{ni}}, \end{aligned} \tag{15}$$

where the hyperparameters $\widetilde{\boldsymbol{\mu}}_{\mathbf{u}_{ni}}$ and $\widetilde{\sum}_{\mathbf{u}_{ni}}$ are inferred through the estimation of latent variable $\mathbf{z}_n$ by the convergence criterion standard [23]. The relationship between the variables in the iSMM emotion model is shown in Figure 3.

According to the emotion model based on infinite Student's t-distribution, in the feature space, the calculation of the posterior distribution $P(\Theta|\mathbf{Y})$ is essential. Therefore, the boundary likelihood probability of emotion feature space partitioned is necessary to be analyzed:

$$P(\mathbf{Y}) = \int P(\mathbf{Y}, \Theta)\, \mathrm{d}\Theta. \tag{16}$$

In detail, because there are interactions between multiple stochastic variables in $P(\Theta \mid \mathbf{Y})$, therefore, according to the method which is in accordance with the Kullback-Leibler (K-L) divergence theory described in previous research [24, 25], an arbitrary distribution $q(\Theta)$ is introduced to estimate the actual posterior distribution $P(\Theta \mid \mathbf{Y})$. On the premise of this assumption, the calculation of the boundary likelihood function of feature data in the emotion model could be realized by applying Student's t-distribution:

$$\begin{aligned} \log P(\mathbf{Y}) &= \mathcal{J}(q) + \mathrm{KL}(q \parallel p) \\ &= \int \mathrm{d}\boldsymbol{\Theta} q(\boldsymbol{\Theta}) \log \frac{P(\mathbf{Y}, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} + \mathrm{KL}(q \parallel p), \end{aligned} \tag{17}$$

where $\mathrm{KL}(q \parallel p)$ is defined for estimating the K-L divergence between the posterior distribution $q(\Theta)$ and the actual posterior distribution $P(\Theta \mid \mathbf{Y})$ [8]. In the above equation, for the K-L divergence is negative, $\mathcal{J}(q)$ is the lower bound of $\log P(\mathbf{Y})$. According to the mean field approximation principle [26], the estimation of the lower bound $\mathcal{J}(q)$ can be determined. Accordingly, predicted parameters of the emotion features in $q(\Theta)$ could be calculated.

Therefore, the emotion model could automatically determine the best number of components with lower complexity to complete various kinds of emotion characteristics data classification.

4. Experimental Results

4.1. Experimental Settings. In the experiment, we conducted three international general speech databases using the iSMM emotion model for SER. In the previous related research [27–29], the EMO-DB and Danish emotional speech (DES) are databases based on speech performance; the utterances included are obtained from the speakers under pressure of simulation and performance. These two databases are chosen to test the performance of the iSMM model in dealing with different amount of data and speech categories. Aibo database is composed of spontaneous speech, belonging to the natural speech emotion. We conduct Aibo in the experiment to evaluate the recognition performance of the iSMM emotion model on imparity properties of speech emotion samples.

We adopted the OpenEAR toolkit [30] to extract numerous low-level descriptors (LLD) including the delta and double-delta functions and achieved 1,582 dimensional features in total. After applying the LLD description criterion, the mapping static feature vectors could be obtained including original signal, the signal energy, pitch, timbre, frequency, and MFCCs. Particularly, the focus in our work is the performance of the iSMM emotion model proposed, which is applied to the SER, rather than research on the optimization algorithm about selecting features. Accordingly, any promotion in recognition performance could be given the credit to the emotion model proposed.

In this experiment, a set of $N = 1582$ p-dimensional speech feature data from various corpus set $\mathbf{y}_n$ is generated by the SMM [12], whose true underlying group structure is known: the number of components I is the number of emotion classes (5 or 7) and the dimension of factors q is 2. In each observed data $\mathbf{y}_n = (\mathbf{y}_{1n}^T, \mathbf{y}_{2n}^T)^T$, $\mathbf{y}_{1n}$ represents

$p = 10$ dimensional subvector containing the latent stochastic variables $\mathbf{z}_n$, while $\mathbf{y}_{2n}$ is a p_2-dimensional subvector of test variables with $p_2 = p - p_1$. Each observed data is obtained by both SMM and iSMM by substitution into formulas in Section 3:

$$\mathbf{y}_n = \left(\mathbf{A}_1^T, \mathbf{A}_2^T\right)^T \mathbf{u}_{ni} + \mathbf{e}_{ni} \text{ with prob. } \pi_i. \tag{18}$$

$p_1 \times p$ submatrix $\mathbf{A}_1$, the common factor loading matrix, is

$$\mathbf{A}_1^T = \begin{pmatrix} 0.5 & -0.9 & 0.3 & 0.6 & 0.2 & -0.7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & -0.7 & 0.5 & 0.6 & -0.4 & 0.3 & -0.5 \end{pmatrix}, \tag{19}$$

and all elements of $p_2 \times q$ submatrix $\mathbf{A}_2$ related to the singularities are zero. Assume that $I = 5$; then the mixing proportion vector is $\pi = (0.15 \;\; 0.2 \;\; 0.3 \;\; 0.15 \;\; 0.2)$. The mean vectors of $\mathbf{u}_{ni}$ are given by $\xi_1 = (0 \;\; 5)^T$, $\xi_2 = (-5 \;\; 0)^T$, $\xi_3 = (0 \;\; 0)^T$, $\xi_4 = (5 \;\; 0)^T$, and $\xi_5 = (0 \;\; -5)^T$ and their covariance matrices are specified as

$$\begin{aligned} \Omega_1 &= \begin{pmatrix} 0.1 & 0 \\ 0 & 0.5 \end{pmatrix}, \\ \Omega_2 &= \begin{pmatrix} 0.5 & 0 \\ 0 & 0.1 \end{pmatrix}, \\ \Omega_3 &= \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, \\ \Omega_4 &= \begin{pmatrix} 0.5 & 0 \\ 0 & 0.1 \end{pmatrix}, \\ \Omega_5 &= \begin{pmatrix} 0.1 & 0 \\ 0 & 0.5 \end{pmatrix}. \end{aligned} \tag{20}$$

The degrees of freedom are $\nu_1 = \nu_2 = \nu_4 = \nu_5 = 5$ and $\nu_3 = 3$. The covariance matrix $\mathbf{D}$ of $\mathbf{e}_{ni}$ is a diagonal matrix, whose diagonal elements are specified to be 0.2.

4.2. Experimental Analysis on Acting Corpus Using iSMM. We conduct SER on the three general databases using four emotion models: iSMM, SMM, GMM, and KNN (K-nearest neighbor) [31], respectively. The Berlin database, also known as EMO-DB, contains utterances spoken in German. EMO-DB is acquired, respectively, from phonetic performance of five male and female actors, including seven kinds of emotion data: anger, disgust, fear, sadness, boredom, neutral, and happiness. There are a total of 800 utterances in EMO-DB. In the literature [9], 20 respondents are invited to conduct actual test of the 494 utterances. Then, the results show that, in EMO-DB, the natural intelligible utterances reach 60% and accurate selection of emotion is above 80%. DES database is collected from 341 utterances within 5 kinds of emotion: anger, happiness, neutral, sadness, and surprise.

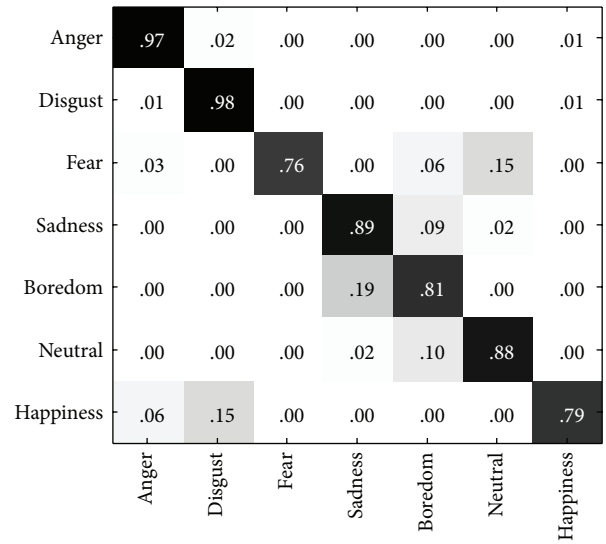


Figure 4: Confusion matrix between seven emotions of the EMO-DB using the iSMM model.

We analyze the results reflected in Figure 4, which demonstrates the normalized confusion matrix of SER experiment using the iSMM on the EMO-DB database. Compared with the improved GMM in literature [9], the iSMM proposed in our research has great improvement on SER performance. Specifically, the recognition rates of "angry" and "disgust" emotion are close to 100%. For the reason of Student's t-distribution, when dealing with EMO-DB, such database with deficiency speech samples conducts the EM algorithm to test emotion data $\mathbf{y}_n'$ for latent variables z_n of emotional features. Thus, the prediction distribution of emotion features in each category could be deduced as $p(\mathbf{y}_n' \mid \mathbf{Y}_k)$; as a result, influences of unstable recognition performance caused by underfitting feature space can be weakened.

However, analyzing the error result of "fear" and "boredom" emotions, which are almost classified to "calm" and "sad," it is shown that the correlations of emotion feature between them are relatively high. It should be noted that there are two types of errors: false positive and false negative. In our experiments, recognition accuracy is the ratio of correct judgment:

$$\text{Ac} = 1 - \text{FP}, \tag{21}$$

where Ac represents the accuracy and FP is the false positive. Here, the inaccuracy consists of leakage probabilistic and false drop rate.

Figure 5 reveals the result of SER on the Danish database by the iSMM emotion model. DES is the speech emotion database comparatively difficult to be recognized; on this account, experimental results show decline of the emotion recognition performance compared to Berlin database overall. In particular, the recognition rates of "fear" and "happy" are less than 80%. It is worth mentioning that, in most cases, some key dimensions of emotion feature in the speech emotion database possess the main information of

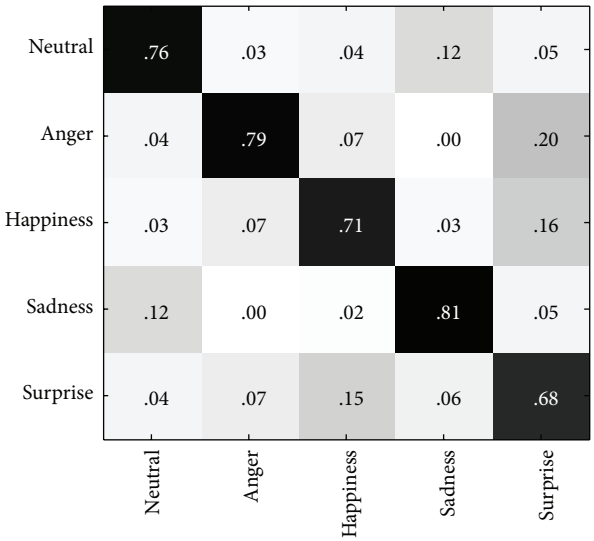


Figure 5: Confusion matrix between five emotions of the DES using the iSMM model.

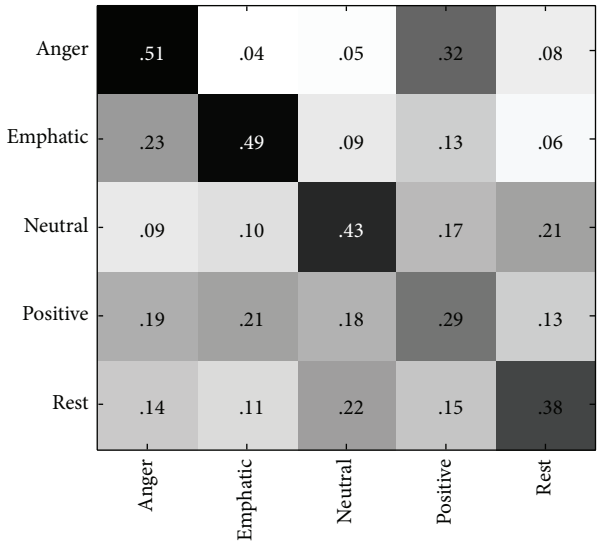


Figure 6: Confusion matrix between five emotions of the Aibo using the iSMM model.

their categories. Compared to many other dimensions, the necessity of their existence is much higher. Profiting from the unique "long-tail" structures, Student's t-distribution, compared with the GMM emotion model, is able to select the key dimensions with high feature information for component partition. Hence, under similar conditions, the SER method proposed in our work achieves obvious improvement of performance compared with the literature [32] which uses 3DEC hierarchical classifier.

4.3. Analysis of Contrast Experiment between iSMM and GMM on the Aibo Database. The Aibo speech database (Batliner et al., 2008 [33]; Steidl, 2009 [34]) was gathered from children accompanying pet robot Aibo developed by SONY. Then, ten kinds of emotion classes contained in the recordings, through machine learning, are mapped into the four categories: anger, emphatic, neutral, and positive, while the other samples are divided into to the fifth classes: rest. Thereafter, through the recognition test to nonspecific person by Björn et al. in 2009, the Aibo database is collated as the recent corpus: Interspeech 2009 emotion challenge [35].

This section focuses on the effect of the iSMM algorithm on emotion recognition from spontaneous speech database. It is revealed by Figure 6 that, compared with the acting speech databases, the recognition rates on the Aibo present striking descent.

Analytically, due to strong continuous correlation between emotional categories in spontaneous speech databases, there exists a considerable amount of outliers and singular values leading to classified errors during the process of component demarcations by emotion models. These errors are more frequent in describing the categories of feature data, peculiarly, when conducting Gaussian distribution with equal means and variances. Benefiting from the long-tail structure of Student's t-distribution, the iSMM possesses more robustness to the atypical observation data (outliers).

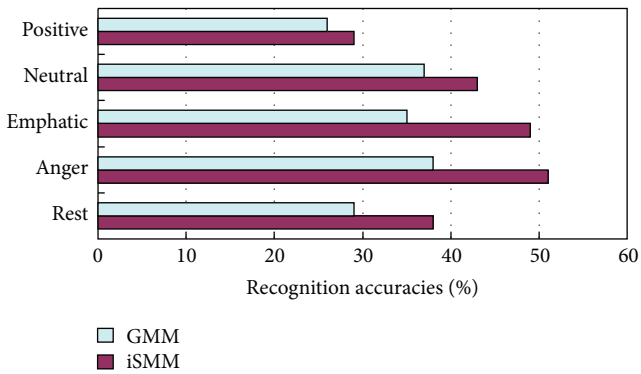


Figure 7: Comparison of SER accuracy between the iSMM and GMM classifiers on five emotions of the Aibo.

Thus, compared to other models, the iSMM has distinct superiority of SER performance on spontaneous speech emotion samples.

Regarding the emotion feature samples of the spontaneous database Aibo, the data distribution of space is discrete and there is lack of the training samples leading to underfitting, since the speech material is evoked. Therefore, this condition brings the emotion models a certain extent of difficulties with component partition. The analysis of Figure 7 shows that accuracies of correct classification on Aibo using the GMM model keep a considerable low level, especially on the two emotions: "positive" and "rest" are even lower than 30%. The iSMM, by contrast, possesses comprehensive advantages in SER. Particularly, on the three emotions: "emphatic", "anger," and "rest", there is a significant improvement. Since there is a given mass of emotion features with high redundancy distributing in the spontaneous database, in the process of component spaces partitioned, the GMM model conducts classification based on the principle of mean square error, which may result in the large redundancy

Table 1: Average accuracy (%) of correct classification on the testing data set of the three databases using four emotion models.

Database	Emotion model			
	iSMM	SMM	GMM	*K*NN
EMO-DB	86.9	85.4	82.7	81.2
DES	75.0	70.9	58.4	60.9
Aibo	42.8	38.2	30.8	31.4

Table 2: Time consumption (MS) of training features of "anger" emotion on EMO-DB using four models.

Time consumption of training			
Emotion model			
iSMM	SMM	GMM	*K*NN
335	562	3798	3976

of the singular points. In contrast, as the analysis above, the ability to deal with outliers brings Student's t-distribution favorable fitness to emotion features of the spontaneous speech. These aspects make the iSMM model keep a relatively good performance and robustness on the Aibo database.

4.4. Integrated Analysis on Various Databases Using the Four Emotion Models. The results of recognition experiment on EOM-DB in German and DES in Danish show that the iSMM model maintains good robustness for different language speech emotion databases. In the iSMM emotion model, each forecast distribution of component in emotion category $p(\mathbf{y}_n' \mid \mathbf{Y}_k)$ is a subset of the distribution of the initial component. Through the stick-breaking principle in the time domain, the number of infinite components on observed data can be inferred. Whereas the numbers of components in the GMM or *K*NN emotion models are changeless, as a result, the speech feature data may be overfitting.

As shown in Table 1, the iSMM holds higher average recognition rates on all databases than its comparison group. The recognition experiment results using GMM and *K*NN are almost flat, while SMM also has a significant advantage in comparison to them. Given this difference, we found that the iSMM and SMM approaches described here, which achieve better results than GMM and *K*NN, profit from the prediction distribution of emotion categories: $P(\mathbf{y}_n' \mid \mathbf{Y}_k)$. This inference of component partition in EM algorithm can effectively handle singular values and outliers. Therefore, in Student's t-distribution, the degradation of recognition performance caused by outliers' partition could be reduced. Furthermore, compared to SMM, iSMM due to the introduction of the infinite number of components has better adaptability to the high-dimensional and underfitting emotion feature data.

In further analysis, we can see that iSMM reaches the recognition rates of 86.9% and 75.0% on German database EMO-DB and Danish database DES, respectively. This means that our method is not significantly dependent on the language type. The variance in features caused by language differences can be modeled universally using infinite Student's t-distribution, which is an important character compared with GMM and other algorithms. Accordingly, we can see that the GMM method achieves a good recognition rate of 82.7% in EMO-DB, while giving a low recognition rate of 58.4% in the Danish database. These results show the iSMM's advantage in cross-language modeling. The relatively low recognition rates on Aibo database are mainly caused by naturalistic speech data, rather than the language factor.

According to the former description in Sections 2 and 3, we can obtain the total number of free parameters in iSMM, which is $(2I-1)+p+q\times(p+q)+0.5\times I\times q\times(q+1)-q^2$. In the FA + GMM, this number is $(2I-1)+2\times p\times I+I\times q\times p-0.5\times I\times q\times(q-1)$, while, in the SMM with common factor loading matrix, it is $(2I-1)+p\times I+q\times p-0.5\times q\times(q-1)+1+I\times(p-1)$. Therefore, in most cases, the iSMM has the smallest number of free parameters. When p is large and/or I is not small, the iSMM is a more feasible tool for modeling high-dimensional data.

As shown in Table 2, the time consumption experimental results are obtained by training features from the "anger" emotion containing 128 utterances on EMO-DB. It is obvious that the models based on Student's t-distribution achieve less time consumption than GMM and *K*NN. Particularly, the proposed iSMM obtains the optimal result. Moreover, the estimated means ξ_i, covariance Λ_i, and degrees of freedom v_i in the iSMM can be used to visualize the distributions of the factors in a low-dimensional latent space, enabling this approach to perform clustering or classification in the latent space.

It is worth mentioning that the independence of speaker, the independence of text, the number of emotional categories for classification, and the size of the database will all affect the actual recognition performance of the emotion mode. This paper puts forward a solution for SER on underfitting samples. Subsequently, the iSMM emotion model possessing good robustness is provided. However, as revealed in the experiments on the Aibo database, at present, by using the emotion models, the SER accuracy ratings on the spontaneous database are considerably low. Moreover, not only does the promotion of comprehensive recognition performance depend on the classifiers, but the optimization of selection on speech features also plays an important role.

5. Conclusions

The mixture model based on Student's t-distribution proposed in this paper, when conducted to speech emotion recognition, can process the outliers of feature samples by the "long-tail" distribution structure. Furthermore, to reduce the dependence of emotion model on the training samples, in our research, an infinite number of components are introduced constituting the iSMM emotion model. Given this solution, the emotion model, by the possibility of automatic partition of components on the feature space, realizes the self-adaptation to various database samples.

Emotion expression is related to many factors, such as personality, context, gender, age, and culture. These factors may create various problems for the speech emotion recognition. Emotions in speech are expressed and interpreted according to different circumstances by humans (e.g., context

dependent, personality specific, and gender dependent), which may be the key to bring speech emotion recognition to the real world applications. As a result, the further research will focus on how to obtain a more appropriate combination of the extractor of emotion feature and the classifier.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work has been supported by the National Natural Science Foundation of China (NSFC) under Grants nos. 61273266, 61231002, and 61375028.

References

[1] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Proceedings of the 13th Annual Neural Information Processing Systems Conference (NIPS '99)*, pp. 554–559, December 1999.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] C. Huang, Y. Jin, Y. Zhao, Y. Yu, and L. Zhao, "Speech emotion recognition based on re-composition of two-class classifiers," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII '09)*, pp. 1–3, IEEE, Amsterdam, The Netherlands, September 2009.

[4] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, and C. Cox, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies, INTERSPEECH," in *Proceedings of the 9th Annual Conference of the International-Speech-Communication-Association (INTERSPEECH '08)*, pp. 597–600, Brisbane, Australia, September 2008.

[5] A. Tawari and M. M. Trivedi, "Speech emotion analysis: exploring the role of context," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, 2010.

[6] Z. Wang, L. Zhao, and C. Zou, "Emotional speech recognition based on modified parameter and distance of statistical model of pitch," *Acta Acustica*, vol. 31, no. 1, pp. 28–34, 2006.

[7] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[8] Z. Xu, F. Yan, and Y. Qi, "Bayesian nonparametric models for multiway data analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 475–487, 2015.

[9] S. Yun and C. D. Yoo, "Loss-scaled large-margin Gaussian mixture models for speech emotion classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 585–598, 2012.

[10] W. Peng, H.-Z. Huang, M. Xie, Y. Yang, and Y. Liu, "A bayesian approach for system reliability analysis with multilevel pass-fail, lifetime and degradation data sets," *IEEE Transactions on Reliability*, vol. 62, no. 3, pp. 689–699, 2013.

[11] X. Wei and C. Li, "Bayesian mixtures of common factor analyzers: model, variational inference, and applications," *Signal Processing*, vol. 93, no. 11, pp. 2894–2905, 2013.

[12] T. M. Nguyen and Q. M. J. Wu, "Robust student's-t mixture model with spatial constraints and its application in medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 103–116, 2012.

[13] H. Muthusamy, K. Polat, and S. Yaacob, "Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals," *Mathematical Problems in Engineering*, vol. 2015, Article ID 394083, 13 pages, 2015.

[14] J. Zhou, Q. Wang, C.-C. Hung, and F. Yang, "Credibilistic clustering algorithms via alternating cluster estimation," *Journal of Intelligent Manufacturing*, pp. 1–12, 2014.

[15] C.-C. Hung, E. Casper, B.-C. Kuo, W. Liu, E. Jung, and M. Yang, "A quantum-modeled artificial bee colony clustering algorithm for remotely sensed multi-band image segmentation," in *Proceedings of the 33rd IEEE International Geoscience and Remote Sensing Symposium (IGARSS '13)*, pp. 2585–2588, Melbourne, Australia, July 2013.

[16] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[17] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '13)*, pp. 216–221, Olomouc, Czech Republic, December 2013.

[18] I. Vrbik and P. D. McNicholas, "Analytic calculations for the EM algorithm for multivariate skew-tt mixture models," *Statistics & Probability Letters*, vol. 82, no. 6, pp. 1169–1174, 2012.

[19] X.-L. Meng and D. Van Dyk, "The EM algorithm—an old folk-song sung to a fast new tune," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 59, no. 3, pp. 511–567, 1997.

[20] D. Posada and T. R. Buckley, "Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests," *Systematic Biology*, vol. 53, no. 5, pp. 793–808, 2004.

[21] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[22] J. Paisley and L. Carin, "Hidden markov models with stick-breaking priors," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3905–3917, 2009.

[23] J. L. Andrews and P. D. McNicholas, "Extending mixtures of multivariate t-factor analyzers," *Statistics and Computing*, vol. 21, no. 3, pp. 361–373, 2011.

[24] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2002.

[25] T. T. Georgiou and A. Lindquist, "Kullback-Leibler approximation of spectral density functions," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 2910–2917, 2003.

[26] N. Bali and A. Mohammad-Djafari, "Bayesian approach with Hidden Markov modeling and mean field approximation for hyperspectral data analysis," *IEEE Transactions on Image Processing*, vol. 17, no. 2, pp. 217–225, 2008.

[27] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 7398–7402, Vancouver, Canada, May 2013.

[28] M. Shah, L. Miao, C. Chakrabarti, and A. Spanias, "A speech emotion recognition framework based on latent Dirichlet allocation: Algorithm and FPGA implementation," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 2553–2557, Vancouver, Canada, May 2013.

[29] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[30] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR-Introducing the munich open-source emotion and affect recognition toolkit," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII '09)*, pp. 1–6, IEEE, Amsterdam, The Netherlands, September 2009.

[31] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15, no. 4, pp. 580–585, 1985.

[32] A. Hassan and R. I. Damper, "Classification of emotional speech using 3DEC hierarchical classifier," *Speech Communication*, vol. 54, no. 7, pp. 903–916, 2012.

[33] A. Batliner, S. Steidl, C. Hacker, and E. Nöth, "Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1-2, pp. 175–206, 2008.

[34] S. Steidl, *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*, University of Erlangen-Nuremberg, Erlangen, Germany, 2009.

[35] S. Björn, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, pp. 312–315, Brighton, UK, September 2009.