SPEAKER EMOTIONAL STATE CLASSIFICATION BY DPM MODELS WITH ANNEALED SMC SAMPLERS

Bilge Gunsel, Ozgun Cirakman and Jarek Krajewski+

Multimedia Signal Proc. and Pattern Rec. Group, Istanbul Technical University, Istanbul, Turkey
+ Institute of Safety Technology, University of Wuppertal, Wuppertal, Germany

ABSTRACT

We propose a speaker emotional state classification method that employs inference-based Bayesian networks to learn posterior density of emotional speech sequentially. We aim to alleviate difficulty in detecting medium-term states where the required monitoring time is longer compared to short-term emotional states that makes temporal content representation harder. Our inference algorithm takes advantage of the Sequential Monte Carlo (SMC) sampling and recursively approximates the Dirichlet Process Mixtures (DPM) model of the speaker state class density with unknown number of components. After learning the target posterior, classification of speaker states has been performed by a simple minimum distance classifier. Test results obtained on two different datasets demonstrate the proposed method highly reduces the training data length while providing comparable accuracy compared to the existing state-of-the-art techniques.

Index Terms— Graphical models, emotion classification, Dirichlet Process Mixtures model, perceptual audio features, HCI.

1. INTRODUCTION

Emotional state classification is a challenging problem in the design of human-machine interactive systems. In this paper, the targeted problem is online speaker state classification from short term as well as medium term emotional speech data. Short term states include well known discrete emotions, i.e., happy, angry, sad, etc., transformed to arousal valance space. As the medium term state classification problem, we deal with sleepiness detection which constitutes a medium term quasi-emotional speaker state. Much of the difficulty in classification comes from the ambiguity of representative features. Also design of an online system to satisfy requirements of applications, i.e., driver sleepiness level monitoring, mood detection in interactive games, etc. is a challenging problem. To achieve an online speaker emotional state classification with high fidelity, it is necessary to involve in a sequential learning scheme which is capable of tracking the dynamic nature of emotional content through time. It is also required to employ features extracted by sequentially processing the speech.

Conventional emotion detection systems make use of acoustic features which are originally proposed for speech recognition hence they may not fully model the speaker emotional states [1]. Consequently, a high performance detector could only be achieved by using very large feature sets (i.e., openEAR) [2] or considerably small feature sets in combination with highly complex classifiers [3,4]. Furthermore their supra-segmental feature extraction scheme prevents online labeling and reporting. Unlike the conventional methods that rely on the linguistics content of speech, we work with prosodic features extracted sequentially by psychoacoustic masking in spectral and temporal domain [5]. It is shown that we can capture discriminative features leading to significantly raised recall rates with a sparse dictionary learned by bag-of-words [6]. By using the same feature set, this work aims to employ inference-based Bayesian networks to learn mixture density of emotional speech sequentially.

It is common to use mixture models in emotion recognition because of their efficiency in modeling diverse statistics of data. Using standard expectation maximization (EM) techniques, the HTK toolkit is employed to build mixture distributions where each emotion is modeled by its own GMM [7, 8]. Unlike the supra-segmental modeling of openEAR, HTK's frame-level modeling is suitable to sequential learning. However specification of the number of mixture components that precisely represent the data is a vital problem in achieving high recall rates. Also GMMs have a serious shortcoming - they are statistically inefficient for modeling data that lie on or near a nonlinear manifold in the data space [9]. To overcome this problem, recently it has been demonstrated that deep neural networks (DNNs) can effectively generate discriminative features that approximate the complex nonlinear dependencies between features hence improve the emotion recognition performance [9]. In [10] Generalized Discriminant Analysis (GerDA) based on DNNs is proposed for learning low dimension discriminative features from a large set of acoustic features for emotion recognition. Currently, the biggest disadvantage of DNNs compared with GMMs is that it is much harder to make good

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

use of large cluster machines to train them on massive data sets. Better ways of parallelizing the fine-tuning of DNNs is still a major issue [9, 11]. In order to alleviate these problems, the Bayesian learning scheme proposed in this work utilizes a Dirichlet Process Mixture (DPM) model in order to estimate each emotional class density where the number of mixture components is unknown. Similar to DNNs we aim to overcome difficulties arise from EM while taking the advantages of employing mixture densities, but unlike the DNNs the introduced method uses a directed graphical model with small number of parameters. As of our knowledge this work is the first attempt for adopting the DPM models to emotion recognition problem. Recall rates reported on EMO-DB [12] and Sleepy Language Corpus (SLC) [13] demonstrate that the proposed method provides comparable accuracy compared to the existing state-of-the-art techniques.

2. PERCEPTUAL FEATURE SET

The low level features used in speaker state detection are computed in the Bark scale as well as in Hz. The feature set is referred as perceptual because in order to model physiological and the perceptual effects of the human ear we apply the outer ear masking on spectrograms prior to feature extraction. An additional psychoacoustic masking is applied before the feature extraction in the Bark scale. Six out of nine of our features are computed in the Bark scale. Table 1 lists the features and gives a brief description of each where more explanation can be found in [6]. Section 3, 4 and 5 present details of the sequential target density estimation and model selection that constitute the main subject of this paper.

3. DIRICHLET PROCESS MIXTURES

This section gives the background on DPM. We define an explicit *time* index *n* and denote the observation sequence by $y_n = \{y_{n,1}, \dots, y_{n,n}\}$. Each observation $y_{n,i}$ $(i = 1, \dots, n)$ is initially assigned to a cluster where $z_{n,i} \in \{1, \dots, k_n\}$ is the corresponding cluster label, and $k_n \in \{1, \dots, n\}$ represents the number of existing clusters at time *n*. The vector of cluster variables is defined as $z_n = \{z_{n,1}, \dots, z_{n,n}\}$ and corresponding cluster parameters are represented with the parameter vector $\theta_n = \{\theta_{n,1}, \dots, \theta_{n,k}\}$.

The DPM model assumes that the cluster parameters are independently drawn from the prior $\pi(\theta_n)$ and the observations are independent of each other conditional on the assignment variable $z_{n,i}$. Hence the DPM posterior density $\pi(x_n)$ is,

$$\pi_n(x_n) = p(z_n, \theta_n \mid y_n) \propto p(z_n) \prod_{j=1}^{k_n} p(\theta_{n,j}) \prod_{i=1}^{n} p(y_{n,i} \mid \theta_{n,z_{n,i}}) \quad (1)$$

where $x_n = \{z_n, \theta_n\}$. The prior on clustering variable vector z_n is formulated by (2) in a recursive way,

LOW LEVEL DESCRIPTORS CALCULATED IN HZ	
Average harmonics structure magnitude(AHSM)	Average of the fundamental frequencies estimated from the log spectrum of the correlations of emotional differences.
10dB perceptual bandwidth (BW1)	The highest frequency component which exceeds the noise floor by at least 10 dB.
5dB perceptual bandwidth (BW2)	The highest frequency component which exceeds the noise floor by at least 5 dB.
LOW LEVEL DESCRIPTORS CALCULATED IN BARK	
Average number of emotional blocks (ANSB)	Expected number of emotional blocks within a time interval (i.e. 1sec in our work).
Normalized emotional level difference (NSD)	Average of the masked variations between the pitch patterns of the audio frame and the reference frame computed over the Bark scales.
Normalized Spectral Envelope Difference (NSED1)	Normalized envelope variations of the unsmeared pitch patterns within the successive frames for each critical band.
NSED2	Average of NSED1 over all critical bands
NSED3	The temporal average of NSED1 through successive Y audio frames.
Overall loudness of the frames (OLF)	Sum across all critical bands of all outer ear weighted loudness values of an audio frame.

Table 1. Features extracted for emotional speaker state modeling

$$p(z_{n,i+1} = j \mid z_{n,\{1:i\}}) = \begin{cases} \frac{l_j}{i+\kappa}, & j = 1, \dots, k_i \\ \frac{\kappa}{i+\kappa}, & j = k_i + 1 \end{cases} \quad (2)$$

where k_i is the number of clusters in the assignment $z_{n,\{1:i\}}$. In (2) l_i is the number of observations that $z_{n,\{1:i\}}$ assigns to cluster *j* and κ is a 'novelty' parameter [14].

4. PROPOSED MULTIVARIATE DENSITY MODEL

In this section we introduce a multi-dimensional Gaussian mixture density model that we learn for each emotional state class. To achieve this we transformed the univariate density learning proposed in [15] to the multidimensional feature space. Furthermore we present our approach to multivariate conjugate prior selection for the emotional speaker state class. The model assumes that observations are drawn from a multivariate Gaussian distribution with unknown mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \Sigma\}$, where the number of mixtures are unknown. As given in Section 3, we deal with the conjugate DPM model that enables us to estimate the mixture parameters given the labeling vector z_n and compute the proposal kernels [18] in closed form. Hence we utilize a Normal-inverse Wishart prior for the parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \Sigma\}$ where,

$$p(\boldsymbol{\mu}, \Sigma) \equiv NIW(\tau_0, w_0, \Lambda_0, \upsilon_0). \quad (3)$$

The joint pdf $NIW(\tau_0, w_0, \Lambda_0, \upsilon_0)$, can be factorized into product of the distributions of the covariance matrix, $p(\boldsymbol{\Sigma})$, and the mean vector, $p(\boldsymbol{\mu}|\Sigma)$ as in (4). The covariance matrix is inverse Wishart distributed, $\Sigma \sim IW(\Lambda_0^{-1}, \upsilon_0)$ where Λ_0, υ_0 are the inverse scale matrix and degrees of freedom respectively, and conditioned on the covariance matrix, Σ, the mean vector, $\boldsymbol{\mu}$, is Normal distributed according to $\sim \mathcal{N}(\tau_0, \Sigma/\mathrm{w}_0)$.

$$p(\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{\left[\frac{\upsilon_0 + d}{2}\right]+1} \times \exp\left(-\frac{1}{2} tr(\Lambda_0 \Sigma^{-1}) - \frac{w_0}{2}(\boldsymbol{\mu} - \tau_0)^T \Sigma^{-1} (\boldsymbol{\mu} - \tau_0)\right) \quad (4)$$

In our model the parameters of each component are denoted by $\theta_{n,j} = \{\boldsymbol{\mu}_j, \Sigma_\mathrm{j}\}$ for notational simplicity, where $j \in \{1, \dots, k\}$ is the index to the number of components, and the overall parameter vector is $\boldsymbol{\theta}_n = \{\theta_{n,1}, \dots, \theta_{n,k_n}\}$. Each mixture component parameter $\boldsymbol{\theta}_j$, $j \in \{1, \dots, k_n\}$ is distributed according to $\boldsymbol{\theta}_j \sim NIW(\tau_0, w_0, \Lambda_0, \upsilon_0)$, where covariance matrix and mean vector are distributed according to, $\Sigma_\mathrm{j} \sim IW(\Lambda_0^{-1}, \upsilon_0)$ and $\boldsymbol{\mu}_j \sim \mathcal{N}(\tau_0, \Sigma/\mathrm{w}_0)$, respectively.

Conditional on assignment z_n the joint posterior distribution of the j'th cluster parameter $\{\boldsymbol{\mu}_j, \Sigma_\mathrm{j}\}$ can be represented by an inverse Wishart prior $NIW(\boldsymbol{\mu}_j, \Sigma_\mathrm{j}|\tau_j, w_j, \Lambda_j, \upsilon_j)$, where the parameters are,

$$\Lambda_j = \Lambda_0 + \sum_{i=1}^{n_j} (\boldsymbol{y}_{j,i} - \bar{\boldsymbol{y}}_j)(\boldsymbol{y}_{j,i} - \bar{\boldsymbol{y}}_j)^T + \frac{w_0 n_j}{w_0 + n_j}(\bar{\boldsymbol{y}}_j - \tau_0)(\bar{\boldsymbol{y}}_j - \tau_0)^T. \quad (5)$$

$$w_j = w_0 + n_j, \quad \tau_j = \frac{w_0 \tau_0 + n_j \bar{\boldsymbol{y}}_j}{w_0 + n_j}, \quad \upsilon_j = \upsilon_0 + n_j.$$

In (5) n_j is the number of observations in the j'th cluster, $\boldsymbol{y}_{j,i}$ indexes each observation in the j'th cluster, and $\bar{\boldsymbol{y}}_j$ is the mean vector of these observations. The marginal of the posterior distributions representing the mean vector $\boldsymbol{\mu}_j$ and the covariance matrix Σ_j can be computed in closed form by the inverse Wishart and student-t distributions respectively, as shown below,

$$p(\Sigma_\mathrm{j}|\boldsymbol{z}_n, \boldsymbol{y}_n) = IW(\Lambda_j^{-1}, \upsilon_j)$$
$$p(\boldsymbol{\mu}_j|\boldsymbol{z}_n, \boldsymbol{y}_n) = \boldsymbol{t}_{\upsilon_j - d + 1}\left(\tau_j, \frac{\Lambda_j}{w_j(\upsilon_j - d + 1)}\right) \quad (6)$$

Accordingly if the parameters $\boldsymbol{\theta}_j, j \in \{1, \dots, k_n\}$ are integrated out from (1) using the conjugacy property, the posterior probability $p(\boldsymbol{z}_n|\boldsymbol{y}_n)$ of the assignment $\boldsymbol{z}_n$ can be expressed up to a proportionality as follows,

$$p(\boldsymbol{z}_n|\boldsymbol{y}_n) \propto p(\boldsymbol{z}_n) \prod_{j=1}^{k_n} \frac{\Gamma_d(\upsilon_j/2) \Lambda_0^{\upsilon_0/2} \kappa_0^{d/2}}{\pi_j^{n_j d/2} \Gamma_d(\upsilon_0/2)\, \Lambda_j^{\upsilon_j/2} \kappa_j^{d/2}} \quad (7)$$

Where $p(\boldsymbol{z}_n)$ is the prior on clustering assignment vector $\boldsymbol{z}_n$, Γ_d is the multidimensional Gamma function and d is the dimension of the observation space.

The proposed model is capable of representing the multidimensional dependencies of emotional features with a reasonable complexity. This is achieved by defining the inverse Wishart distribution, which is the conjugate prior over the Gaussian likelihood function, as the prior over the mixture component parameters. This selection simplifies the sampling based inference scheme and reduces the variance of the SMC estimator. Moreover, unlike the conventional models that employs diagonal covariance matrices, the inverse Wishart distribution enables us to define a full covariance matrix that is much more appropriate to model correlations of features that yields precise learning of the number of components encountered in a particular emotional speaker state.

5. DENSITY ESTIMATION BY ANNEALED SMC SAMPLERS

In a sequential problem the posterior distribution changes over time and new modes of the posterior distribution may emerge as new observations are received. The algorithm must have a good mixing property to explore the modes of the time evolving posterior distribution and to achieve a good approximation to the true target posterior. The conventional approach applies Gibbs moves to each particle in order to obtain weighted samples from a sequence of target distributions denoted as $\pi_1(\boldsymbol{z}_1), \dots, \pi_n(\boldsymbol{z}_n)$. However, the Gibbs sampler may fail to represent the modes of the true target posterior due to the slow convergence property of the Gibbs samplers. This is particularly observable when the posterior distribution has a multi modal form where the modes are isolated [16]. To deal with this problem, we employ an annealing scheme to improve the efficiency of posterior estimation [15]. The annealing scheme is adopted to importance sampling to construct the proposal distribution suitable to sampling of the true target distribution [15,16].

To achieve our goal the annealed target posterior is defined as $\pi_n(\boldsymbol{z}_n) = \pi_n(\boldsymbol{z}_n|\kappa = \alpha_n)$ and as each new observation arrives the annealing is achieved by updating the novelty parameter of the underlying Dirichlet process which is set to α_n according to a geometric spacing function $\alpha_n = \alpha_{n-1} + c_\alpha(\kappa - \alpha_{n-1})$ where $\alpha_1 > 0$, $\alpha_n > \alpha_{n-1}$ and c_α is the common parameter that determines the amount of spacing at each time step. Note that α_n is a parameter of the prior distribution of number of components where a higher value yields higher number of mixtures. The annealed distributions can be interpreted as an underlying DPM model of which the parameters are relaxed in order to obtain a tractable annealed posterior which is easy to sample.

We employ particle filtering as inference tool hence the annealed DPM target posterior density $\pi(\boldsymbol{x}_n)$ shown in (1) can be approximated at time $n-1$ with a set of weighted

particle $\{W_{n-1}^i, X_{1:n-1}^i\}_{i=1}^{N_p}$. At time n the path of each particle can be extended using a Markov kernel, $K_n(\boldsymbol{x}_{n-1}, \boldsymbol{x}_n)$ and the unnormalized importance weights $\tilde{\gamma}_n(\boldsymbol{x}_{1:n})/\eta_n(\boldsymbol{x}_{1:n})$, associated with the extended particles are calculated according to $w_n(x_{1:n}) = w_{n-1}(x_{1:n-1})v_{n-1}(x_{n-1}, x_n)$ where the incremental term of weight equation, $v_{n-1}(x_{n-1}, x_n)$ is,

$$v_{n-1}(x_{n-1}, x_n) = \frac{\gamma_n(\boldsymbol{x}_n)L_{n-1}(\boldsymbol{x}_n, \boldsymbol{x}_{n-1})}{\gamma_{n-1}(\boldsymbol{x}_{n-1})K_n(\boldsymbol{x}_{n-1}, \boldsymbol{x}_n)} \quad . \qquad (8)$$

The design of efficient sampling schema hinges on properly choosing the backward kernel L_{n-1}. Assuming K_n is an Monte Carlo Markov Chain (MCMC) kernel of invariant distribution π_n, an approximate backward kernel can be formulated as in [16] that yields to a good approximation for $\pi_{n-1} \approx \pi_n$ with the incremental weight, $v_{n-1}(x_{n-1}, x_n) = \gamma_n(x_{n-1})/\gamma_{n-1}(x_{n-1})$.

6. TEST RESULTS

In order to evaluate performance of the proposed method (DPM-P) we performed classification by the learned emotional class density functions and compared the achieved speaker state recognition rates with existing methods. For all test cases the number of particles is set to 100.

First we have used EMO-DB database to evaluate the short term emotional speaker state recognition capability. The six emotion classes included in EMO-DB are unemotional, disgusting, bored, angry, happy and fear. For benchmarking, we report the accuracy in continuous arousal-valance space by using the transformed classes provided in [8]. Table 2 reports the unweighted (UA) recall rates for *Arousal*, *Valance* and *All* categories as it is reported in [8] as benchmarking results. Recall rates, except the proposed DPM-P method and GMM-P (GMM classification by using our features), are copied from the related literature and references are given in the parenthesis. Particularly, SVM (openEAR) denote the unweighted recall rates achieved by openEAR features and are reported in [8]. SVM-P refers to our previous work that we have performed classification by using libSVM tool of WEKA. GMM/HMM (HTK) refer the recognition rates achieved by HTK as reported in [8]. GerDA denote the speaker state recognition rates obtained by the Generalized Discriminant Analysis based on DNNs [10]. All of the reported recall rates except DPM-P show the accuracy computed offline manner. Because all of the methods except DPM-P apply offline classification. Some (i.e., openEAR) apply supra-segmental feature extraction, some others (i.e., HTK) perform frame based feature extraction but apply chunk based or segmental classification. However unweighted recall rates listed for DPM-P illustrate that recognition rates reported over each test sample (for each sequentially processed 9-D feature vector) without any post processing or utterance based filtering. It can be concluded that the short time speaker state classification accuracy of the proposed online classification scheme is comparable to the existing methods. And also from the accuracy of GMM-P, SVM-P and DPM-P cases, it can be concluded that the representation capability of our perceptual features outperforms the conventional features.

	Arousal	*Valance*	*All*
DPM-P	83.5	82.8	81.7
GerDA [10]	97.6	82.2	79.1
GMM-P	92.1	91.9	92.5
HMM/GMM (HTK) [8]	91.5	78.0	73.2
SVM-P [6]	95.2	94.3	86.3
SVM (openEAR) [8]	96.8	87.0	84.6

Table 2. Comparison of unweighted recall rates with existing work on EMO-DB

Secondly the medium term speaker emotional state classification accuracy of DPM-P has been tested on SLC corpus [17] used in the Speaker State Challenge [17] to compare our performance with the existing systems. SLC ground truth is provided at 10 sleepiness levels on Karolinska Sleepiness Scales (KSS) where a level exceeding 7.5 is labeled as sleepy [17]. Since the SLC speech content is scattered to various KSS levels it is difficult to model the detection as a two-class classification problem, i.e., sleepy (SL) and non-sleepy (NSL). The SLC data includes 9089 utterances, which features 21 hours of speech recordings of 99 subjects. The sampling rate of speech is down-sampled to 16 kHz. According to the data used for training and test stages, test scenarios are named as *Train vs Develop* and *Train+Develop vs Test* as in [17]. Number of utterances used for the training and test are respectively 3366 and 2915 for *Train vs Develop*. For *Train+Develop vs Test,* we use 6281 and 2808 utterances, respectively. Note that we used exactly the same test data to compare performances of different methods (Table 3).The only difference is the sequential training procedure in DPM-P case. Therefore we have also reported the speaker emotional state classification accuracy for smaller training sets. This is achieved classification of the same test data by employing the target mixture density estimated at different training instants (Table 4). Note that the longer training data is obtained by extending the previous short data with sequentially included new observations.

Table 3 reports the sleepiness detection performance obtained by the proposed method DPM-P compared to the existing ones. Columns RR_{SL} and RR_{NSL} respectively denote the unweighted recall rates for sleepy (SL) and nonsleepy (NSL) speaker states. *Avg* corresponds to the arithmetic mean of RR_{SL} and RR_{NSL}. Rates achieved offline by the SVM with the same 9 features are reported as SVM-P [6] (we used Weka libSVM toolbox with RBF kernel, C:2). IS2011 Win refers

	Train vs Develop			*Train+Dev vs Test*		
	RR_{SL}	RR_{NSL}	Avg	RR_{SL}	RR_{NSL}	Avg
DPM-P	74.2	64.0	69.1	89.3	71.6	80.5
SVM-P [6]	89.1	97.2	93.2	79.9	80.1	80.0
IS2011 Win [3]	60.3	75.7	68.0	64.2	79.1	71.6
IS2011[17]	NA	NA	67.3	NA	NA	70.3

Table 3. Comparison with existing work on SLC

Training length (minute)	*SL-k_n*	*NSL-k_n*	*RR_{SL}*	*RR_{NSL}*	*Avg*
	Train+Dev vs Test				
1.9	6.5	10.2	82.9	60.3	71.6
18.98	15.0	17.0	95.0	64.0	79.5
75.91	16.0	17.0	95.4	65.7	80.6
189.77	22.0	21.0	89.3	71.6	80.4
Full: 872.18	23.0	21.0	77.3	79.1	78.2
	Train vs Develop				
1.9	7.0	11.0	67.3	61.4	64.4
18.98	15.1	16.3	68.2	61.3	64.7
75.91	18.0	18.0	66.4	63.5	65.0
189.77	24.0	21.0	74.2	64.0	69.1
Full: 453.55	24.0	21.0	63.1	73.0	68.0

Table 4. Recall rates vs training time for DPM-P method

the highest scores reported by the Interspeech 2011 Speaker States Challenge participants where the features of openEAR are used [3]. IS2011 SSC denote the highest baseline performance declared in [17] and the results are obtained by the openEAR features. It can be seen from Table 3 that classification accuracy of the proposed sequential method (DPM-P) in first case is lower than the SVM-P and comparable to the IS2011 Winner and IS2011 SSC. However in the second case DPM-P is comparable to SVM-P and clearly outperforms others. Since other methods perform offline supra-segmental feature extraction, it can be concluded that the introduced method DPM-P is promising for online applications.

In order to evaluate learning capability and speed of DPM-P, we also reported the speaker state labeling accuracy at different stages of training. Table 4 reports the performance at various training lengths. As can be seen from the table, performance tends to increase with the training time up to a level. But after a certain time performance of the system begins to drop because of overfitting. So rather than using full training data, DPM-P results reported at Table 3 are provided for 10k training observations which corresponds to 189.77 minutes of training. This is where the system performance has its peak on *Train vs Develop* case according to Table 4. $SL - k_n$ and $NSL - k_n$ are the observed number of mixture components learned by DPM-P at each training phase. Note that required training time is significantly less than a deep neural network based method but is still long enough for a fully online scheme. Currently we are working on to speed up the convergence by improving DPM-P resampling scheme. However, it should be noted that, even dramatically lower training durations such as 1.9 or 18.98 mins are used, performance of the DPM-P is still comparable with other methods. So it is possible to increase training speed as needed without serious performance loss.

7. CONCLUSIONS

We introduce an emotional speaker state classification framework that model the emotional speech with a mixture density. Reported performance is comparable to the state-of-the art methods that use deep neural networks, however choosing sensible values for hyper-parameters such as the learning rate schedule, the strength of the regularizer, the number of layers and the number of units per layer requires considerable skill and experience and is a major limitation of DNNs. The proposed DPM model with SMC samplers constitute a promising alternative to existing methods since it can be easily adopted to online applications that require sequential processing.

8. REFERENCES

[1] A Ayadi, M. Kamel, F. Karray, "Survey on speech emotion recognition; features, classification schemes, and databases," Pattern Recognition, vol:44(3), pp.572-587, 2011.

[2] F. Eyben, et al., openEAR– Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit, *in Proc. of* the Affective Computing and Intelligent Interaction , 2009.

[3] D. Huang, S. S. Ge, Z. Zhang,"Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines," INTERSPEECH, pp.3301-3304, 2011.

[4] J. Krajewski, S. Schnieder, M. Golz, A. Batliner, B. Schuller, "Applying multiple classifiers and nonlinear dynamics feature for detecting sleepiness from speech," Journal of Neurocomputing, vol:84, pp. 65-75, 2012.

[5] M. C. Sezgin, B. Gunsel, G. K. Kurt, "Perceptual Audio Features for Emotion Detection," *EURASIP J. on Audio, Speech, and Music Processing*, vol:16, 2012.

[6] M. C. Sezgin, B. Gunsel, J. Krajewski, "Medium term speaker state detection by perceptually masked spectral features,' Journal of Speech Communication, vol.67, pp. 26-41, 2015.

[7] S. Young, et al., The HTK book (v3.4). Cambridge, UK: Cambridge University Press, 2006.

[8] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," *in Proc. of* IEEE Automatic Speech Recognition and Understanding, pp.552-557, 2009.

[9] G. Hinton, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol.29(6), pp. 82-97, November 2012

[10] A. Stuhlsatz, et al, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in Proc. of IEEE ICASSP 2011, pp.5688-5691, 2011.

[11] J. Dean, G.S. Corrado, R. Monga, K. Chen, M. Devin, Q.V. Le, M.Z. Mao, M.A. Ranzato, A. Senior, P. Tucker, K. Yang, A. Y. Ng,"Large scale distributed deep networks," *NIPS* 2012.

[12] F. Burkhardt, et al, "A Database of German Emotional Speech," INTERSPEECH, pp. 1517–1520, 2005.

[13] The Center of Interdisciplinary Speech Science, Univ. of Wuppertal, Germany.

[14] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica* , vol.4, pp.639–650., 1994.

[15] Y. Ulker, B. Gunsel, A. T. Cemgil, "Annealed SMC Samplers for Nonparametric Bayesian Mixture Models," *IEEE Signal Processing Letters,* vol:(1), pp. 3-6, 2011.

[16] R. Neal, "Annealed importance sampling," *Statistical Computing*, vol.11, pp. 125–139, 2001.

[17] B. Schuller, et al, "Medium-term speaker states—A review on intoxication, sleepiness and the first challenge," Introduction to the Special Issue on Broadening the View on Speaker Analysis," J. of Computer Speech and Language,vol:28, pp.346-374, 2014.