SPEAKER STATES RECOGNITION USING LATENT FACTOR ANALYSIS BASED EIGENCHANNEL FACTOR VECTOR MODELING

Ming Li, Angeliki Metallinou, Daniel Bone, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering, University of Southern California, Los Angeles, USA

mingli@usc.edu, metallin@usc.edu, dbone@usc.edu, shri@sipi.usc.edu

ABSTRACT

This paper presents an automatic speaker state recognition approach which models the factor vectors in the latent factor analysis framework improving upon the Gaussian Mixture Model (GMM) baseline performance. We investigate both intoxicated and affective speaker states. We consider the affective speech signal as the original normal average speech signal being corrupted by the affective channel effects. Rather than reducing the channel variability to enhance the robustness as in the speaker verification task, we directly model the speaker state on the channel factors under the factor analysis framework. In this work, the speaker state factor vectors are extracted and modeled by the latent factor analysis approach in the GMM modeling framework and support vector machine classification method. Experimental results show that the proposed speaker state factor vector modeling system achieved 5.34% and 1.49% unweighted accuracy improvement over the GMM baseline on the intoxicated speech detection task (Alcohol Language Corpus) and the emotion recognition task (IEMOCAP database), respectively.

Index Terms— Speaker state recognition, Emotion recognition, Latent factor analysis, Supervector modeling

1. INTRODUCTION

Automatic recognition of paralinguistic information (e.g., gender, age, emotional state), can guide human computer interaction systems to automatically adapt to different user needs. Identifying speaker state given a short speech utterance is a challenging task and has gained significant attention recently in the speaker emotion challenge[1], paralinguistic challenge[2], and speaker states challenge[3].

It has been shown in [4, 5, 6] that speaker state information can be modeled at various levels, such as phonetic, acoustic, and prosodic. Due to the different aspects of modeling, combining different classification methods can significantly improve the overall performance [4, 5]. Acoustic level modeling approaches, such as Hidden Markov Models (HMM) or Gaussian Mixture Models (GMM) operating on the mel-frequency cepstral coefficient (MFCC) features, play the most fundamental and important role among those various subsystems [5, 7] given short utterances. In this paper, we follow the GMM-MFCC framework and focus on further enhancing the performance of the intoxicated and affective speaker state recognition tasks.

Recently, latent factor analysis (LFA) [8, 9] has been successfully and widely used for the speaker verification task, in which session variability caused by different channels influences the system performance dramatically. However, systems that directly use LFA to remove the speaker variability factors in the speaker state recognition task have been shown to perform worse than a GMM baseline [7, 6]. This might indicate that the speaker variability is larger than the speaker state variability. To address this issue, we treat a paralinguistic speech signal as the normal average speech signal being corrupted by channel effects and consider the speaker states as the "channels". Thus we employ the Eigenchannel factors as a new kind of speaker state supervector and adopt SVM to model these factor vectors for the discriminative speaker state classification task.

The GMM LFA approach can be considered as a type of feature extraction frontend which summarizes the affective speaker states information into the low dimensional Eigenchannel factor vectors. Compared to the commonly used large dimensional feature vectors computed using statistical functionals [2], the proposed factor vector reduces the feature dimensionality dramatically; therefore, it is more efficient for the adaptive or online model training applications.

2. METHODS

An overview of the proposed method is displayed in Figure 1.

2.1. GMM baseline

The Gaussian Mixture Model (GMM) is adopted to model the MFCC features. In the proposed work, a universal background model (UBM) in conjunction with a maximum a posteriori (MAP) model adaptation approach [10] is used to model different speaker states in a supervised manner. Let the UBM be a N-components GMM model $\widetilde{\lambda}$:

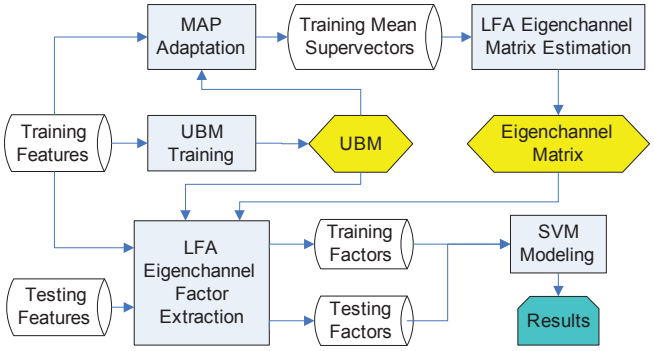

Fig. 1. The system overview

$\widetilde{\lambda} = \{p_i, \widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}}, \boldsymbol{\Sigma_i}\}, i = 1, \cdots, N$, where p_i and $\boldsymbol{\Sigma_i}$ are the i^{th} UBM mixture weight and diagonal covariance matrix, $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}}$ corresponds to the mean of the i^{th} Gaussian component of the UBM. For each target speech segment, a GMM was adapted from the UBM by the MAP adaptation [10]. As shown in (1), the GMMs were modeled with diagonal covariance matrices and only the means were adapted.

$$\boldsymbol{\mu_i} = \alpha_i \boldsymbol{E_i(x)} + (1-\alpha_i)\widetilde{\boldsymbol{\mu}}_{\boldsymbol{i}}, \alpha_i = \frac{\sum_{t=1}^{T} Pr(i|\boldsymbol{x_t})}{\beta + \sum_{t=1}^{T} Pr(i|\boldsymbol{x_t})} \quad (1)$$

where $Pr(i|\boldsymbol{x_t})$ denotes the occupancy probability of feature frame $\boldsymbol{x_t}(t = 1, \cdots, T)$ belongs to the i_{th} gaussian component and β is the constant relevance factor. Here the GMM mean supervector $\boldsymbol{M}$ is defined as a concatenation of the GMM mean vectors $\boldsymbol{M} = [\boldsymbol{\mu_1^t}, \cdots, \boldsymbol{\mu_i^t}, \cdots, \boldsymbol{\mu_N^t}]^t$ [9, 11]. Assume the feature vectors are D-dimensional, the GMM mean vector $\boldsymbol{M}$ is a ND dimensional vector.

2.2. Eigenchannel matrix estimation

In the GMM LFA framework for speaker verification [9], we can consider the paralinguistic speech as normal average speech being corrupted by affective channel variability. Let us denote $\boldsymbol{M_{k,c}}$ as the speaker and affective channel dependent mean supervector. Then $\boldsymbol{M_{k,c}}$ can be decomposed into speaker dependent mean supervector plus the channel variability $\boldsymbol{Uy}$, where $\boldsymbol{U}$ is the low rank Eigenchannel matrix learned from the Principle Component Analysis (PCA) on the pooled within speaker covariance matrix.

$$\boldsymbol{M_{k,c}} = \boldsymbol{M_k} + \boldsymbol{Uy} \quad (2)$$

$\boldsymbol{U}$ is a factor loading matrix and the components of $\boldsymbol{y}$ are the speaker state channel factors [9]. In order to train the Eigenchannel matrix, we need to use data from multiple speakers. Furthermore, for each speaker, there should be speech utterances from multiple speaker states. First, for each speaker $k, k = 1, \cdots, K$ and all his utterances $j = 1, \cdots, J_k$, UBM is adapted to obtain a supervector $\boldsymbol{M_{kj}}$. Second, the corresponding speaker true supervector is estimated by averaging all the supervectors from this speaker:

$$\widetilde{\boldsymbol{M_k}} = \sum_{j=1}^{J_k} \frac{\boldsymbol{M_{kj}}}{J_k}, \acute{\boldsymbol{M}}_{\boldsymbol{kj}} = \boldsymbol{M_{kj}} - \widetilde{\boldsymbol{M_k}}. \quad (3)$$

Then we concatenate all the speakers' state variability supervectors $\acute{\boldsymbol{M}}_{\boldsymbol{kj}}$ into a variability supervector matrix $\boldsymbol{S}$ with ND rows and J columns ($J = \sum_{k=1}^{K} J_k$):

$$\boldsymbol{S} = [\acute{\boldsymbol{M}}_{\boldsymbol{11}}, \cdots, \acute{\boldsymbol{M}}_{\boldsymbol{1J_1}}, \cdots, \acute{\boldsymbol{M}}_{\boldsymbol{K1}}, \cdots, \acute{\boldsymbol{M}}_{\boldsymbol{KJ_K}}] \quad (4)$$

Finally, the Eigenchannel matrix $\boldsymbol{U}$ are given by R PCA eigenvectors of the within speaker covariance matrix $(1/J)\boldsymbol{SS^t}$ wihch corresponds to the R largest eigenvalues [9].

2.3. Eigenchannel factor extraction

Based on the latent factor analysis framework, the speaker states factor vector $\boldsymbol{y}$ is estimated as follows [8, 9]:

$$\boldsymbol{y} = (\boldsymbol{A} + \boldsymbol{E}^{-1})^{-1} \sum_{i=1}^{N} \boldsymbol{U_i'} \sum_{t=1}^{T} \gamma_i(t) \frac{\boldsymbol{x_t} - \boldsymbol{\mu_i}}{\boldsymbol{\Sigma_i}} \quad (5)$$

$$\boldsymbol{A} = \sum_{i=1}^{N} \frac{\boldsymbol{U_i'U_i}}{\boldsymbol{\Sigma_i}} \sum_{t=1}^{T} \gamma_i(t) \quad (6)$$

$\boldsymbol{U_i}$, $\gamma_i(t)$ and $\boldsymbol{x_t}$ denote the sub matrix of $\boldsymbol{U}$ corresponding to the i^{th} gaussian component ($D \times R$), occupancy probability of the t^{th} feature on the i^{th} gaussian component and the t^{th} feature vector, respectively. $\boldsymbol{U_i'}$ is the transpose of matrix $\boldsymbol{U_i}$. The diagonal covariance matrix $\boldsymbol{E}$ includes the R leading eigenvalues of the Eigenchannel matrix $\boldsymbol{U}$. The details about the Eigenchannel factor extraction are provided in [8, 9].

For each utterance, the acoustic MFCC features are mapped into the Eigenchannel factor vector $\boldsymbol{y}$. Then, a back end SVM classifier was trained using LIBSVM [12] to model the multi-class speaker states categories.

3. EXPERIMENTAL RESULTS

The proposed Eigenchannel factor vector modeling approach is evaluated on two speaker state recognition tasks, namely, intoxicated speech detection, in Section 3.1 and speaker emotion classification, in Section 3.2.

3.1. Intoxicated speech detection

The Alcohol Language Corpus (ALC) database [3] comprising 154 German speakers released in the 2011 speaker state challenge [3] was used to study the intoxicated speaker state recognition task. The two speaker states of interest are intoxicated (indicated by a blood-alcohol content above 0.5mg/L) and sober. First, in the GMM baseline system, a 512 component GMM was trained for each state on the 39 dimensional MFCC features in the training dataset. Second, in the

LFA framework, MAP adaptation from UBM model was performed for every utterance in both training and development dataset. The GMM mean supervectors were generated by concatenating the mean vectors of all components from the adapted GMM. Then, the Eigenchannel matrix was trained using the mean supervectors from the train set, and Eigenchannel factor vector extraction was performed to map each mean supervector into the factor vector. Speaker normalization [13] was adopted on top of these factor vectors. The GMM size N and Eigenchannel matrix rank R are 256 and 4, respectively. Finally, LIBSVM toolkit [12] was adopted to perform this binary classification task on the 4-dimensional Eigenchannel factor vectors, $\boldsymbol{y}$.

The classification accuracy on the development set are shown in Table 1. We can see that the proposed LFA Eigenchannel factor modeling approach outperformed the GMM baseline and achieved 3.94% and 5.34% improvement for weighted and unweighted accuracy, respectively.

3.2. Speaker emotion classification

In the speaker emotion study, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [14]. This database contains approximately 12 hours of audio-visual data from five mixed-gender pairs of actors[14]. IEMOCAP contains detailed face and head information obtained from motion capture as well as video, audio and transcripts of each session. Two act types were used; scripted and improvisation of hypothetical scenarios. The goal was to elicit emotional displays that resemble natural emotional expression. Dyadic sessions of approximately 5 minute length were recorded and were later manually segmented into utterances. Each utterance was annotated by at least 3 annotators into categorical labels (anger, happiness, neutrality, etc). We examine all 10 available speakers and use only speech modality signals. We examine classes of anger, happiness, excitation, neutrality, and sadness where there was majority consensus across the three annotators. We have merged the classes of happiness and excitation into a single class which we will refer to as happiness.

We organize our emotion recognition experiments using 10-fold leave-one-speaker-out cross validation. The mean and standard deviation of the number of test utterances across the folds is: 62 ± 28, angry; 87 ± 26, happy; 58 ± 23, neutral; and 65 ± 25, sad. The 30 dimensional feature set is composed of 13 MFCC features, energy, pitch and their first order derivatives. The GMM size is 512 and the Eigenchannel matrix rank is 26. In Table 2, the speaker-independent emotion classification results averaged over the 10 folds are presented. We can observe that only moderate improvements are achieved (1.77% WA and 1.49% UA). This might be because the emotional states are not stable and vary dynamically both within and across utterances. It is shown in [15] that the factor analysis based speaker factor vectors can be used for the

Table 1. Unweighted accuracy (UA) and weighted accuracy (WA) [3] on the development set of ALC database in the 2011 speaker state challenge.

	WA	UA
GMM baseline	65.33%	65.05%
LFA Eigenchannel Factor Modeling	69.27%	70.39%

Table 2. Unweighted accuracy (UA) and weighted accuracy (WA) per utterance for 10-fold leave-one speaker out cross validations on the IEMOCAP database.

	WA	UA
GMM baseline	54.11%	54.35%
LFA Eigenchannel Factor Modeling	55.88%	55.84%

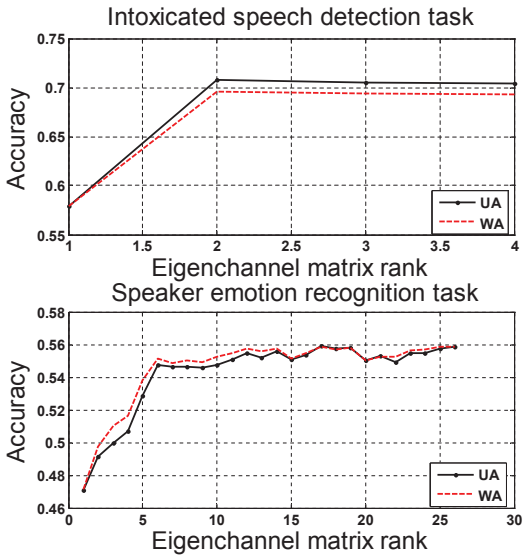

Fig. 2. Accuracy against Eigenchannel matrix rank

speaker change point detection task and achieved promising results. Therefore, our further work will focus on analyzing and tracking the proposed speaker states factor vectors along the entire speech conversation or dialog using the sliding window framework which may have great potential in the speaker states change point detection and tracking tasks.

The SVM classification results against the Eigenchannel matrix rank R are shown for both tasks in Fig.2. We can see that the results are not sensitive to the Eigenchannel matrix rank and that even small rank ($R < 5$) can achieve competitive results. This property validates that the proposed Eigenchannel factor vector is highly informative and effective in terms of the speaker state feature. Furthermore, the matrix rank in the intoxicated speaker state recognition task is smaller than in the speaker emotion classification task which might be due to smaller speaker state categories.

In figure 3 we plot the first two dimensions of the eigen-

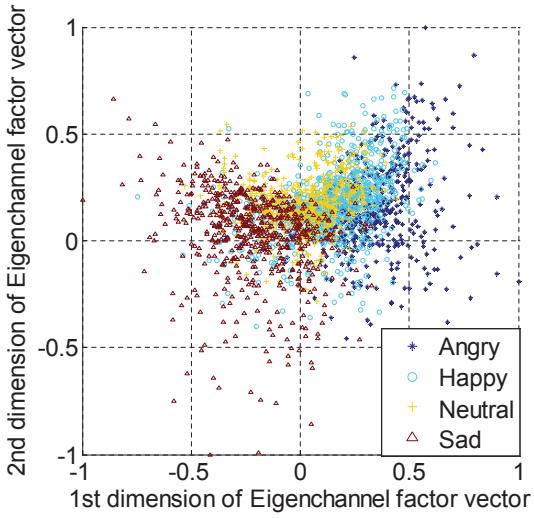

Fig. 3. First 2 dimension of Eigenchannel factor vector of fold 0 IEMOCAP training data

channel factor vector of the training data instances from the first fold (the plots are similar across folds). We observe that different emotions tend to occupy different, although overlapping, positions of the two-dimensional space, suggesting the discriminative ability of these two dimensions. Moreover the first dimension seems to carry activation-related information. Activation is an emotional attribute describing how active vs passive is an emotional state. Typical examples of highly activated emotions include anger, happiness and excitement, while neutrality and sadness are usually described by medium and low activation respectively [16]. Similar structure can be observed across the first dimension where angry and happy utterances tend to have high values while neutral and sad utterances have medium and low values respectively. This suggests that some of the computed Eigenchannel factors may carry some interpretable (here emotion-related) information.

Future work includes analyzing and tracking the proposed speaker state factor vectors along the entire speech conversation or dialog using a sliding window framework. Moreover, factor analysis based i-vectors [17] and lasso based s-vectors [18] in combination with various variability compensating methods, such as Within Class Covariance Normalization (WCCN), Probabilistic Linear Discriminant Analysis PLDA), may also be employed to model the speaker states.

4. CONCLUSIONS

In this work, a latent factor analysis based Eigenchannel factor vector modeling approach was proposed to recognize various speaker states. We consider the affective speech signal as the original average speech signal being corrupted by the paralinguistic channel effects. Rather than reducing the paralinguistic channel variability to enhance the robustness as in the speaker verification task, we directly model the speaker state on the paralinguistic Eigenchannel factors under the factor analysis framework. Experimental results show that the proposed method outperformed the GMM baseline on both the intoxicated speech detection and speaker emotion recognition tasks.

5. REFERENCES

[1] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. of the Interspeech*, 2009, pp. 312–315.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S.S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. of the Interspeech*, 2010, pp. 2794–2797.

[3] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Proc. of the Interspeech*, pp. 3201–3204.

[4] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J.G. Bauer, et al., "Comparison of four approaches to age and gender recognition for telephone applications," in *Proc. of the ICASSP*, 2007, vol. 4, pp. 1089–1092.

[5] M. Li, C.S. Jung, and K.J. Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition," in *Proc. of the Interspeech*, 2010, pp. 2826–2829.

[6] M. Li, K. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *submitted to Computer speech and language*.

[7] M. Kockmann, L. Burget, and J. Černockỳ, "Brno university of technology system for interspeech 2010 paralinguistic challenge," in *Proc. of the Interspeech*, 2010, pp. 2822–2825.

[8] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a gmm speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.

[9] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

[10] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[11] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[12] C.C. Chang and C.J. Lin, "Libsvm: a library for support vector machines," 2001.

[13] Daniel Bone, Matthew P. Black, Ming Li, Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan, "Intoxicated speech detection by fusion of speaker normalized hierarchical features and gmm supervectors," in *Proc. of the Interspeech*, Aug. 2011, pp. 3217–3220.

[14] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[15] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[16] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz, "What a neural net needs to know about emotion words," *Computational intelligence and applications*, pp. 109–114, 1999.

[17] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[18] M. Li, C. Lu, A. Wang, and S. Narayanan, "Speaker verification using lasso based sparse total variability supervector and probabilistic linear discriminant analysis," in *Proc. of the NIST speaker verification workshop*, 2011.