COMPARISON OF ADAPTATION METHODS FOR GMM-SVM BASED SPEECH EMOTION RECOGNITION

Jianbo Jiang[1,2], *Zhiyong Wu*[1,2], *Mingxing Xu*[2], *Jia Jia*[2], *Lianhong Cai*[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2] Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
jjb10@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, {xumx, jjia, clh-dcs}@tsinghua.edu.cn

ABSTRACT

The required length of the utterance is one of the key factors affecting the performance of automatic emotion recognition. To gain the accuracy rate of emotion distinction, adaptation algorithms that can be manipulated on short utterances are highly essential. Regarding this, this paper compares two classical model adaptation methods, maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR), in GMM-SVM based emotion recognition, and tries to find which method can perform better on different length of the enrollment of the utterances. Experiment results show that MLLR adaptation performs better for very short enrollment utterances (with the length shorter than 2s) while MAP adaptation is more effective for longer utterances.

Index Terms**—**emotion recognition, GMM supervector based SVM, MAP adaptation, MLLR adaptation

1. INTRODUCTION

Expressive speech with emotions plays important role in the communications between humans. Automatic recognition of emotion in speech has been one of the latest challenges in the field of speech processing. It has gained great interests in many practical application areas, such as detection of lies, security system, psychiatric aids and interactive games. It is also being integrated into a broad area of researches, such as human-computer interaction, emotional speech or speaker recognition, emotion recognition on visual speeches [1].

To recognize emotion from speech, a large number of acoustic features have been put forward. These features can be classified as prosodic features, spectral features and voice quality features. Prosodic features consist of statistics that are derived from the fundamental frequency (f0) and energy contours. Spectral feature mainly include features derived from Mel frequencies, such as Mel-frequency cepstral coefficients (MFCCs). Statistics of jitter, shimmer and harmonic-to-noise ratio (HNR) belong to voice quality features [1]. We highly concentrate on acoustic features commonly used in different literatures. Hu et al. provides MFCCs as features for emotional classification, or rather, 13-dimensional MFCC plus energy, together with their delta and acceleration coefficients, 42 dimensions altogether [2].

In terms of model representation, hidden Markov model (HMM) and Gaussian mixture model (GMM) have achieved outstanding results on speech emotion recognition. Schuller et al. brought forward the feasibility of emotion recognition with hidden Markov models [3]. In [4], a method using phoneme-class dependent HMM classifiers with short-term spectral features was proposed. Luengo et al. believed that traditional GMM based emotional speech classifiers using spectral features could achieve a high accuracy [5].

Support vector machines (SVMs) have been proved to be able to achieve better performance for solving problems in classification, regression and novelty detection than many other classifiers. As the dimension of spectral features extracted from the speech utterances with various lengths is not fixed, a GMM supervector based SVM with spectral features was brought forward in [2]. The main idea is, the MFCCs extracted from each emotional speech utterance are used to train a GMM, and then the GMM supervector is constructed by concatenating the mean vectors of all the Gaussian mixtures in the GMM, and this GMM supervector is served as the input feature for SVM [2].

In our approach, we also adopt GMM supervector based SVM with spectral features for speech emotion recognition. In training the GMM, we use the adaptation technology that is widely used in speaker recognition to adapt a universal background model (UBM) to derive the final GMM for each emotion category. It is found that the length of the speech utterances used for adaptation is one of the key factors that affect the performance of the adapted GMM for recognizing emotions. Considering the requirement of the applications where only short speech utterances (with the length shorter than 5 seconds) are available (e.g. interactive dialog system), the adaptation algorithms that can be manipulated on short utterances are highly essential. Regarding this, in this paper, we compare two classical model adaptation methods the maximum a posteriori (MAP) and the maximum likelihood linear regression (MLLR) for GMM-SVM based emotion recognition, and try to find which method can perform better on different length of enrollment of speech utterances.

The rest of this paper is organized as follows. In Section 2, the GMM supervector based SVM system for emotion recognition is characterized. Then two different adaptation methods, MLLR and MAP are described in Section 3. Experiments and results are presented in Section 4. Finally, Section 5 gives the conclusions.

2. GMM-SVM BASED EMOTION RECOGNITION

In this work, GMM supervector based SVM (GMM-SVM) with spectral features is adopted for emotion recognition of speech. Details of the method are described as follows.

2.1. GMM supervector based SVM system

Assume that an utterance *Y* with only one kind of emotion is chosen, and the hypothesized emotion category is *X*, the task of speech emotion recognition is to determine if *Y* is of the emotion category *X*. Then the core function of emotion recognition can be described as a basic hypothesis test to evaluate which of the following two statements is true:

- ➢ S_0: *Y* is of the hypothesized emotion *X*, and
- ➢ S_1: *Y* is not of the hypothesized emotion *X*.

The following likelihood ratio (LR) is calculated to examine the statements:

$$\frac{p(Y \mid S_0)}{p(Y \mid S_1)} \begin{cases} \geq \theta, \text{accept } S_0, \\ < \theta, \text{ accept } \mathrm{S}_1. \end{cases} \tag{1}$$

where $p(Y|S_i)$, i=0,1 is the probability density function of the statement S_i evaluated for the speech utterance *Y*, θ is the given decision threshold for accepting or rejecting S_i.

The basic goal of a speech emotion recognition system is to determine techniques to compute values for the two likelihoods, $p(Y|S_0)$ and $p(Y|S_1)$. In this work, we adopt the GMM supervector based SVM (GMM-SVM) approach. The framework of the proposed emotion recognition system is illustrated in Fig.1. The input utterance is first processed by the front-end processing module to construct the GMM supervector, which is then used to compute the likelihoods of S_0 and S_1, where S_0 is represented by an SVM model λ_{hyp} which characterizes the hypothesized emotion *X* and S_1 is represented by another SVM model λ_{UBM} characterizing the universal backgrounds. The likelihood ratio is computed as follows and used to test against the threshold θ to make the final accept/reject decision.

$$\Lambda = \frac{p(Y \mid \lambda_{hyp})}{p(Y \mid \lambda_{UBM})} \tag{2}$$

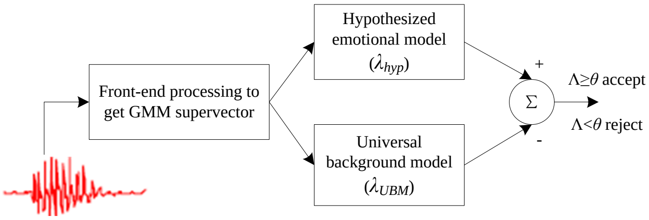
Hypothesized emotional model (λ_{hyp})
Front-end processing to get GMM supervector
Universal background model (λ_{UBM})
+
Σ
-
Λ≥θ accept
Λ<θ reject

Fig.1. Likelihood ratio based emotion recognition system

2.2. Constructing GMM supervector

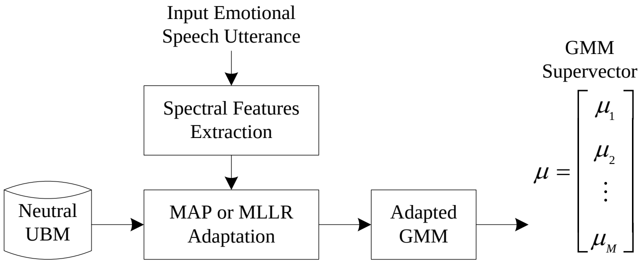
Input Emotional Speech Utterance
Spectral Features Extraction
Neutral UBM
MAP or MLLR Adaptation
Adapted GMM
GMM Supervector
$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_M \end{bmatrix}$

Fig.2. Constructing GMM supervector from an utterance

The process of constructing the GMM supervector from an input emotional speech utterance is shown in Fig.2.

The density function of a GMM is defined as following:

$$p(x) = \sum_{i=1}^{M} w_i N(x; \mu_i, \Sigma_i) \tag{3}$$

where $\boldsymbol{N}(;,)$ is the is the Gaussian density function, *M* is the number of Gaussian mixtures, w_i, μ_i and Σ_i are the weight, mean and covariance matrix of the ***i***-th Gaussian mixture respectively. The supervector of a GMM is defined by concatenating the mean of each Gaussian mixture, which can be thought of as a mapping between an utterance and a high-dimensional vector:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_M \end{bmatrix} \tag{4}$$

Given an input emotional speech utterance, the spectral features are first extracted and used to adapt the GMM from a neutral universal background model (UBM). In this work, the UBM is a GMM that is trained using neutral speeches from a large number of speakers. The MAP or MLLR adaptation algorithm is used to adapt the GMM from neutral UBM for the input utterance, and during adaptation, only the mean vector μ_i of each Gaussian mixture is adapted. From the adapted GMM, the final GMM supervector is constructed as the representation of the input utterance. Details of the adaptation methods will be elaborated later in the next section.

2.3. SVM

An SVM is a non-probabilistic binary linear classifier, used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

SVM performs a non-linear mapping from an input space to a high-dimensional space through a kernel function $K(,)$. The SVM classifier is constructed from sums of the kernel function:

$$f(\mathbf{x}) = \sum_{i=1}^{N} a_i t_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{5}$$

where t_i is the ideal output with value either 1 or -1, depending on whether the corresponding support vector is in class 0 or class 1, respectively. $\sum_{i=1}^{N} a_i t_i = 0$, and $a_i>0$. The vectors $\mathbf{x}_i$ are support vectors. The classification result is depended on if $f(\mathbf{x})$ is above or below a given threshold.

For simplicity, the linear kernel is selected in the GMM supervector based SVM for speech emotion recognition.

3. ADAPTATION METHODS

Two classical model adaptation methods are investigated to adapt the GMM from neutral UBM for the input utterance, namely maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP).

3.1. Maximum Likelihood Linear Regression

Maximum likelihood linear regression (MLLR) computes a set of transformations that reduce the mismatch between an initial model set and the adaptation data. The model is adapted using a set of linear transformations for the mean (and variance) parameters of a Gaussian mixture model, estimated in a maximum likelihood fashion from the adaptation data. The effect of these transformations is to shift the means (and alter the variances) of each Gaussian mixture in the initial model so that the adapted Gaussian mixture is more likely to generate the adaptation data.

The transformation matrix used to give a new estimate of the adapted mean is found by

$$\hat{\mu} = W\xi \tag{6}$$

where W is the $n\times(n+1)$ transformation matrix (where n is the dimensionality of the data) and ξ is the extended mean vector,

$$\xi = \begin{bmatrix} w & \mu_1 & \mu_2 & \cdots & \mu_n \end{bmatrix} \tag{7}$$

where w represents a bias offset whose value is fixed at 1.

Simply W can be decomposed into

$$W = \begin{bmatrix} b & A \end{bmatrix} \tag{8}$$

where b is a bias vector on the mean and A represents an $n\times n$ transformation matrix that may be full, block diagonal or diagonal.

The target is to find the transformation matrix W which maximizes the likelihood of the adaptation data. W can be obtained by solving a maximization problem using the Expectation-Maximization (EM) algorithm, which is also used to compute the variance transformation matrix. In this work, only the mean vector of the GMM is adapted.

3.2. Maximum A Posteriori

Model adaptation can also be accomplished by maximum a posteriori (MAP) approach, which is sometimes referred to as Bayesian adaptation. MAP adaptation involves the use of prior knowledge about the model parameter distribution. If we know what the parameters of the model are likely to be (before observing any adaptation data) using the prior knowledge, we might be able to make good use of the limited adaptation data, to obtain a decent MAP estimate. This type of prior is often termed an informative prior. It should be noted if the prior distribution indicates no preference as to what the model parameters are likely to be (a non-informative prior), the MAP estimate obtained will be identical to that obtained using a maximum likelihood approach.

Assume that we want to estimate an unobserved parameter μ on the basis of observation x. Let f be the sampling distribution of x, so that $f(x|\mu)$ is the probability of x when the underlying population parameter is μ. Then assume that a prior distribution g exists. Hence μ can be treated as a random variable as in Bayesian statistics. The method of maximum a posteriori estimation then estimates μ as the mode of the posterior distribution of this random variable:

$$\hat{\mu}_{MAP} = \arg\max_{\mu} f(x \mid \mu) g(\mu) \tag{9}$$

For mathematical tractability conjugate priors are used, which results in a simple adaptation formula. The update formula for the i-th Gaussian mixture is:

$$\hat{\mu}_{i,MAP} = \beta\overline{\mu}_i + (1-\beta)\mu_i \tag{10}$$

where μ_i is the mean of one emotion and $\overline{\mu}_i$ is the mean of the observed adaptation data, or universal background. The universal background model (UBM) is trained from the database using Expectation–maximization (EM) algorithm. And β is the adjusting coefficient given by prior knowledge or calculated during the adaptation procedure.

3.3. Comparison between MLLR and MAP

One obvious drawback of MAP adaptation is that it requires more adaptation data when compared to MLLR, because MAP adaptation is specifically defined at the Gaussian mixture level. When larger amounts of adaptation training data become available, MAP begins to perform better than MLLR, due to this detailed update of each Gaussian mixture (rather than the pooled Gaussian transformation approach of MLLR). In fact the two adaptation processes can be combined to improve performance further, by using the MLLR transformed means as the priors for MAP adaptation. In this case Gaussian mixtures that have low occupation likelihood in the adaptation data (and hence would not change much using MAP alone) have been adapted using a regression class transform in MLLR.

4. EXPERIMENTS

4.1. Databases

Two emotional databases are used in the experiments of our work. The first one is the famous Berlin German emotional database [11], which contains about 500 utterances spoken by 10 actors in 7 emotional categories (i.e. joy, anger, fear, sadness, bored, disgust as well as neutral). The data were taken with the sampling rate of 48 kHz and downsampled to 16 kHz. The average length the speech recordings is 2.78s.

The second one is our homegrown Chinese emotional database, which contains 5 kinds of emotions, including four classic emotions (anger, fear, happiness and sadness) and neutral. 464 voice clips of the Chinese emotional speech database were interceptions from movies and TV series, with an average length of 3.16s. The data were digitized using a sampling rate of 16 kHz with 16-bit resolution, and saved in single channel wav files.

4.2. Experiments

The recordings from the databases are converted into 13-dimensional Mel frequency cepstral coefficients (MFCC) plus energy, together with their delta and acceleration coefficients, forming 42-dimentional acoustic features. The features are extracted every 10ms using the frame length of 25ms, with Hamming windowing and pre-emphasis factor of 0.97. The Gaussian mixture model (GMM) consists of 64 Gaussian mixtures. The neutral universal background model (UBM) is trained from the neutral speech recordings in the speech database as described above.

In the experiments, 5-fold cross validation is performed for error estimation. More precisely, each of the emotional databases described above is equally divided into 5 disjoint subsets, and the SVM classifiers are trained five times, each time with a different subset held out as a testing set.

To evaluate the performances of the two different adaptation methods and their relation with the length of the input speech utterances, three datasets of the utterances are created and used in the experiments. The first dataset consists of the original speech recordings. The other two datasets are created by cutting each utterance of the original speech recordings into two parts with shorter average length. For the Berlin German emotional database, the average lengths of the utterances of the two newly created datasets are 1.85s and 1.39s respectively. While for the homegrown Chinese emotional database, the average lengths are 2.11s and 1.58s respectively.

4.3. Results

The hit rate, the ratio of the number of utterances correctly recognized to the total number of all available utterances, is calculated for evaluating the experiment.

$$HitRate = \frac{\#\,of\ correctly\ recognized\ utterances}{\#\,of\ all\ utterances} \tag{11}$$

The hit rates of the experiments on three different datasets using MLLR or MAP adaptation are summarized in Table 1 and Table 2 for the Berlin German emotional database and the Chinese emotional database respectively.

Table 1. The hit rates (HitRate) over 3 datasets using MLLR or MAP adaptation for the Berlin German database

Dataset	1	2	3
Average length	2.78s	1.85s	1.39s
MLLR adaptation	76.6%	77.6%	70.1%
MAP adaptation	80.4%	72.0%	67.3%

Table 2. The hit rates (HitRate) over 3 datasets using MLLR or MAP adaptation for the Chinese database

Dataset	1	2	3
Average length	3.16s	2.11s	1.58s
MLLR adaptation	76.6%	74.5%	68.1%
MAP adaptation	84.0%	76.6%	58.5%

From the above tables, we could find that in the GMM-SVM based emotion recognition system, MAP adaptation performs well when the utterances are relatively long; while MLLR adaptation can achieve better performance when the input utterance is with short duration. For the Berlin German emotional database, MAP adaptation can achieve better performance than MLLR adaptation for the dataset with the average length of 2.78s. While for the Chinese emotional database, MAP adaptation outperforms MLLR adaptation for the dataset with the average length of both 3.16s and 2.11s.

To find the correlations between the performance of the two adaptation methods and the length of the input speech utterances, the hit rates on the utterances with different lengths varying from 1s to 3.5s are further evaluated for the two adaptation methods. The inferiors of the hit rates of the MLLR adaptation compared to MAP adaptation on the utterances with different lengths are shown in Fig.3, where each point indicates the inferior of the hit rate of the MLLR adaptation to that of the MAP adaptation. The inferior of the hit rates is computed as:

$$HitRate_{inferior} = HitRate_{MAP} - HitRate_{MLLR} \qquad (12)$$

And the hit rates for MAP and MLLR are computed on all utterances with the length within 0.25s around the indicated length. For example, the hit rate inferior -31.8% (at 1s) is computed using the utterances with the length from 0.75s to 1.25s. The hit rate inferior less than 0 indicates the performance of MLLR adaptation method is better than MAP method. The lower is the hit rate inferior; the better is the performance of the MLLR adaptation method. The results in Fig.3 indicate that MLLR adaptation achieve higher hit rate on very short-enrollment utterances, (e.g. the hit rate of the MLLR is 31.8% higher than those of the MAP for speech utterances of the length at around 1 second), while MAP performance better on long-enrollment utterances. It can also be observed that MAP adaptation performs better then MLLR adaptation when the utterances are longer than 2s.

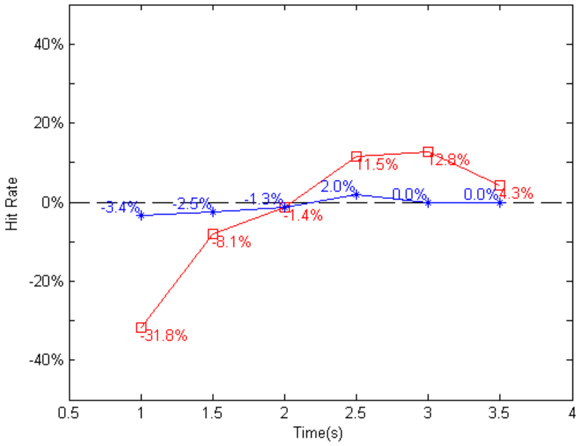

Fig.3. Hit rates on utterances with different lengths using MLLR adaptation inferior to MAP adaptation, where the asterisk indicates the hit rates on German database while square shows the hit rates on Chinese database

5. CONCLUSION

In the GMM supervector based SVM approach for speech emotion recognition, the adaptation technique that is widely used in speaker recognition has been adopted to adapt a UBM in delivering the final GMM for each emotion. It is found that the length of the speech utterances used for adaptation is one of the key factors that affect the performance of the adapted GMM for recognizing emotions. Considering the requirement of the applications where only short speech utterances (with the length shorter than 5 seconds) are available (e.g. interactive dialog system), the adaptation algorithms that can be manipulated on short utterances are highly essential. Regarding this, this paper compares two classical model adaptation methods, the maximum a posteriori (MAP) and the maximum likelihood linear regression (MLLR), for GMM-SVM based emotion recognition, and tries to find which method can perform better on different length of enrollment of speech utterances within 4 seconds. It is found that MLLR outperforms MAP adaptation when utterances were shorter than 2s, and MAP adaptation is a bit better than MLLR while the utterances are longer.

6. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (60928005, 60805008, 60931160443 and 61003094), the Ph.D. Programs Foundation of Ministry of Education of China (200800031015), the Upgrading Plan Project of Shenzhen Key Laboratory and the Science and Technology R&D Funding of the Shenzhen Municipal.

7. REFERENCES

[1] A. Tawari, “Speech emotion analysis: Exploring the role of context,” *IEEE Trans. Multimedia*, 12(6): 502-509, 2010.

[2] H. Hu, M.X. Xu, and W. Wu, “GMM supervector based SVM with spectral features for speech emotion recognition,” In: *Proc. ICASSP*, 413-416, 2007.

[3] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” In: *Proc. ICASSP*, II: 1-4, 2003.

[4] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z.G. Deng, S. Lee, and S. Narayanan, “Emotion recognition based on phoneme classes,” In: *Proc. ICSLP*, 2004.

[5] I. Luengo, E. Navas, I. Hernaez, and J. Sanchez, “Automatic emotion recognition using prosodic parameters,” In: *Proc. Interspeech*, 493-496, 2005.

[6] M.W. Mak, R. Hsiao, and B. Mak, “A comparison of various adaptation methods for speaker verification with limited enrollment data,” In: *Proc. ICASSP*, 929-932, 2006.

[7] J. Mariethoz, and S. Bengio, “A comparative study of adaptation methods for speaker verification,” In: *Proc. ICSLP*, 2002.

[8] B. Mak, R. Hsiao, “Robustness of several kernel-based fast adaptation methods on noisy LVCSR,” In: *Proc. Interspeech*, 266-269, 2007.

[9] A.B. Poore, B.J. Slocumb, B.J. Suchomel, F.H. Obermeyer, S.M. Herman, and S.M. Gadaleta, “Batch maximum likelihood (ML) and maximum a posteriori (MAP) estimation with process noise for tracking applications,” In: Proc. SPIE Signal, Data Processing of Small Targets, 2003.

[10] M. Dai, D. Yang, and M.X. Xu, “Research on the composition of UBM training set in speech emotion recognition,” In: *Proc. NCMMSC*, 2011 (in Chinese).

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” In: *Proc. Interspeech*, 2005.

[12] C.C. Chang, and C.J. Lin, “LIBSVM: a library for support vector machines,” *ACM Trans. Intelligent Systems and Technology*, 2(3), 2011.