REGULARIZED ESTIMATION OF GAUSSIAN MIXTURE MODELS FOR SVM BASED SPEAKER RECOGNITION

BY

KAIZHI QIAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science in Electrical and Computer Engineering
in the Undergraduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Mark Hasegawa-Johnson

Abstract

Speaker adaptation based on the Universal Background Model (UBM) has become a standard approach for speaker recognition. A GMM supervector is constructed by normalizing and stacking the means of the adapted mixture components, which provides a compact representation of the speaker-dependent model in speaker recognition tasks. The estimation of the unknown GMM parameters is usually obtained by the method of maximum a posteriori estimation (MAP), which can be regularized to increase the model interpretability with insufficient training data. In this thesis, the speaker-adapted models are estimated using the MAP with L1-regularization, referred to as the elastic net, based on the assumption that the distinctions between any two speakers are sparse. Experiments on the NIST2008 speaker recognition evaluation task show error rate reduction with the elastic net.

Subject Keywords: speaker recognition; elastic net; sparsity; Gaussian mixture model; supervector

Contents

1. Introduction

Our focus in this thesis is regularization of the Gaussian Mixture Models (GMMs) supervectors for Support Vector Machines (SVMs) based text-independent speaker recognition. That is, given a test utterance and a claim of identity, determine if the claim is true or false [1][2]. A standard approach is to produce GMM supervectors for speaker utterances, which are classified using SVMs [3]. Given a speaker utterance, the Universal Background Model (UBM) is updated using Maximum A Posteriori (MAP) adaptation, which is essentially equivalent to the idea of Ridge regression [4]. However, Ridge regression does not produce a highly interpretable model, because it keeps all the predictors in the model [5]. Lasso regression [6] can produce a parsimonious model with high interpretability, but it has certain limitations [5]. A promising technique combining the ridge and the lasso, called the elastic net [5], overcomes the limitations of the lasso while improving the interpretability of the model. In our case, a variable selection process seems appropriate, because the number of model is much greater than the number of observations. As such, we applied the elastic net to model estimation by adding the lasso to the traditional MAP adaptation, assuming a sparse representation for the text-independent speech feature.

The reminder of this thesis is organized as follows, chapter 1 introduces the background basics in three sections, chapter 2 presents review of prior publications in speaker recognition and regularization techniques, chapter 3 derives equations, chapters 4 and 5 describe the experiments setup and results, and chapter 6 discusses the problems with the setup and possible future improvements. We conclude in chapter 7 that elastic net increased the model interpretability and thus reduced the error rate for speaker recognition tasks.

2. Literature Review

2.1 GMM Based Speaker Recognition

GMM modeling in text-independent speaker recognition tasks was first introduce by Reynolds in 1992, and later became a dominant approach for numerous years [7]. In particular, since 1996, the GMM-UBM speaker recognition system developed by the MIT Lincoln Laboratory, has been the basics of the state-of-art systems [7]. The GMM-UBM system employs Bayesian adaptation of models from a universal background model based on a binary hypothesis test [7].

A universal background model (UBM) is a Gaussian mixture model (GMM) trained with large amount of speaker data to represent the general information of the speakers [8]. The basic form of the GMM acoustic model is expressed as

$$p(x) = \sum_{i=1}^{N} \lambda_i \mathcal{N}(x; m_i, \Sigma_i) \tag{2.1}$$

where λ_i is the mixture weight, $\mathcal{N}(.)$ is the multivariate Gaussian distribution, and m_i, Σ_i are the mean and covariance of the Gaussian respectively.

Given a speaker utterance Y and a hypothesized speaker identity S, the single-speaker verification task can be modeled as a simple binary hypothesis test between

$$H_0 : \text{Y is from the hypothesized speaker S}$$

and

$$H_1 : \text{Y is NOT from the hypothsized speaker S}.$$

The optimum method to solve the above problem is given by a likelihood ratio test:

$$\frac{p(Y|H_0)}{P(Y|H_1)} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{reject } H_0 \end{cases} \tag{2.2}$$

where $p(Y|H_i), i = 0,1$, is the likelihood of the hypothesis H_i given the speaker utterance Y, and θ is the decision threshold.

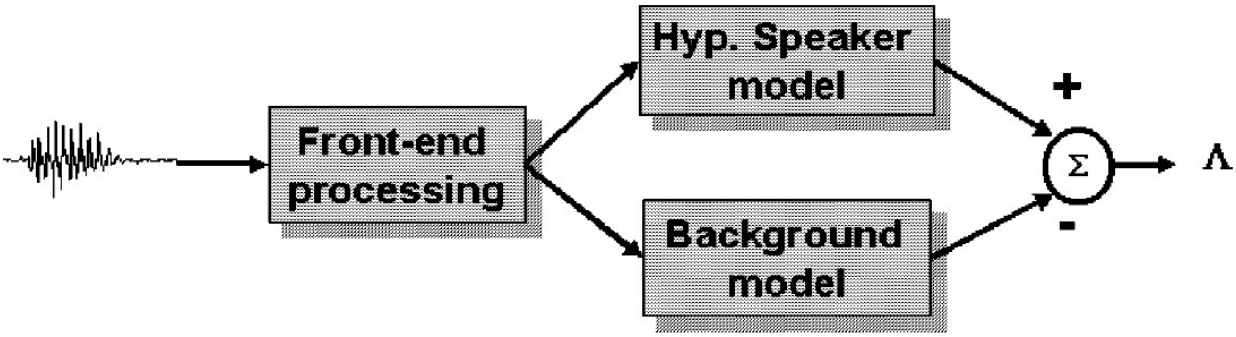

Figure 1. Likelihood ratio test for speaker verification [7]

From figure 1, the background model λ_{UBM} is a speaker independent model trained using thousands of utterances from different speakers, and the speaker-dependent model λ_{spk} is produced by MAP adaptation from the background model [7].

A score Λ is calculated by taking the log-likelihood ratio of the given utterance against the background model and speaker model respectively, as given by

$$\Lambda(Y) = \frac{1}{T}\sum_{i=1}^{T}\left\{\lg\frac{p(Y_i|\lambda_{spk})}{p(Y_i|\lambda_{UBM})}\right\} \tag{2.3}$$

If Λ is larger than a certain predetermined threshold θ, we decide that the utterance belongs to the hypothesized speaker model [7].

2.2 SVM Based GMM Supervector Method

An SVM [9] is a binary classifier constructed from the sum of kernels $K(x_1, x_2)$. Given any vector x, the output of the SVM is

$$f(x) = \sum_{i=1}^{N}\alpha_i t_i K(x, x_i) + d \tag{2.4}$$

where t_i is the ideal output such that $\sum_{i=1}^{N}\alpha_i t_i = 0$ and $\alpha_i > 0$, x_i is the support vector, and $K(.)$ is the kernel function given by

$$K(x, y) = \mathbf{b}^T(x)\mathbf{b}(y) \tag{2.5}$$

under the Mercer condition [9].

The ideal output is -1 or 1, depending on whether the corresponding support vector is in class 0 or class 1, respectively [2]. Training the SVM is a process during which an optimum hyperplane is constructed to achieve the maximum margin after all inputs are mapped into a high-dimensional space, as shown in figure 2. The key idea is that some nonlinearly separable low-dimensional inputs become separable in a higher-dimensional space. The generalization performance of the SVM is better when the margin is larger. The data points lying on the boundaries are the support vectors.

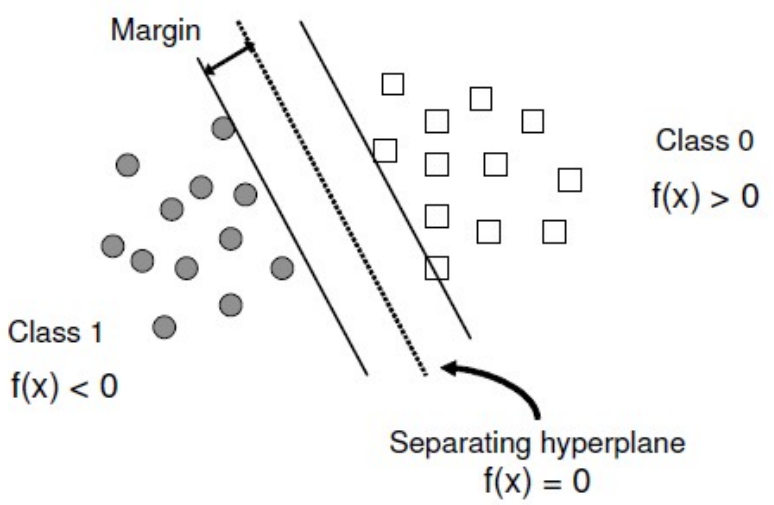

Figure 2. Support Vector Machine concept [2]

For classification, a class decision is based upon whether the value of $f(x)$ is above or below a threshold [2].

A GMM supervector is constructed for each speaker utterance by stacking the means of the adapted UBM, which can be thought of as a mapping between a low-dimensional utterance to a high-dimensional vector [1], as shown in figure 3.

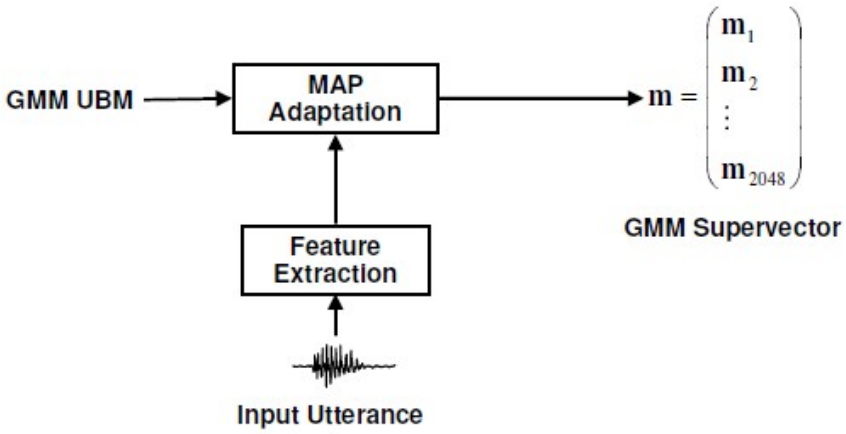

Figure 3. GMM supervector concept [1]

According to the kernel function given in Equation (2.5), the expansion vector $\mathbf{b}(x)$ is a vector of all monomials of the input feature vector [2] for the traditional SVM speaker recognition method. The traditional SVM speaker recognition shares the same idea with the binary hypothesis test, because the SVM is essentially a binary classifier. Unlike the GMM method described in section 2.1, the scores are computed using a sequence kernel function instead of the conditional probability density function. Experiments show SVMs achieved comparable performance as the GMM method [2].

The idea of stacking the means of the GMM to form a supervector inspired the application of supervector in SVM classifier [1], which was introduced by Campbell in 2006. A supervector is a compact representation for the text-independent speech features. The GMM supervector representation fits well with the idea of an SVM sequence kernel [1]. Referring to Equation (2.5), $\mathbf{b}(x)$ is done by the GMM supervector mapping that provides equal dimensional vectors for computing the inner product. The monomial feature expansion in the traditional SVM based speaker recognition is replaced by the supervector representation for the SVM based GMM supervector method. The new system has shown state-of-the-art performance in the NIST 2005 speaker recognition tasks [1].

2.3 Regularization Techniques

Acoustic model estimation often needs to address the problem of data sparsity and model complexity [10]. Given insufficient training data, over-fitting makes the estimated parameters fails to capture the general feature of the data. In the presence of over-fitting, the model describes more noise instead of the underlying relations between the model and the features in low signal-to-noise-ratio (SNR) acoustic scenarios.

Let $\mathbf{y} = (y_1, \cdots, y_n)^T$ be the response and $\mathbf{X} = (\mathbf{x}_1 | \cdots | \mathbf{x}_p)$ be the model matrix, where $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})^T$, $j = 1, \cdots, p$, are the predictors. The response and predictors are assumed to be properly normalized [5].

For any fixed non-negative λ_1 and λ_2, the naïve elastic net can be expressed as [5]

$$L(\lambda_1, \lambda_2, \beta) = |\mathbf{y} - \mathbf{X}\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1 \tag{2.6}$$

where β is the weight of each predictor, and the second term and the third term of Equation (2.6) are called L2-regularization and L1-regualraiztion, respectively, as expressed specifically below in Equation (2.7) and (2.8).

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2 \tag{2.7}$$

$$|\beta|_1 = \sum_{j=1}^{p} |\beta_j| \tag{2.8}$$

The first term of (2.6) alone is the ordinary least square (OLS), the first term plus the second term is the ridge regression, and the first term plus the third term is the lasso regression.

The ordinary linear regression (OLS) is the simplest model for data fitting and prediction. However, the OLS suffers from low prediction ability and low parsimony when the number of predictors becomes large [5]. Regularization must be applied to address the above issues. Thus, Ridge regression was proposed by Hoerl and Kennard in 1988, which gains performance through a variance-bias trade-off subject to a bound on the L2-norm [5]. However, the Ridge regression does not produce a parsimonious model, because it tries to spread weights on all the predictors, which results in smaller weights on the principle predictors. In order to address the problem of model parsimony, Tibshirani (1996) introduced a promising variable selection method called Lasso [6]. According to Equation (2.8), Lasso minimizes the OLS subject to a bound on L1-norm of the weight coefficients. The Lasso does continuous shrinkage and automatic variable selection at the same time, due to the property of the L1-norm [5]. Both Ridge and Lasso overcomes over-fitting by performing an element-wise shrinkage of weight coefficients toward zero in the absence of an opposing data-driven force [10][11]. Unlike Ridge, the variable selection process of Lasso can produce a parsimonious model by driving some small non-principle weight

coefficients toward zero, which can increase the model interpretability with insufficient training data. This variable selection property of the Lasso gains popularity in sparse coding applications, and inspires its application in acoustic modeling as well, which has led to several interesting publications in speech signal processing [10].

However, the Lasso does have pronounced limitations in some situations. First of all, when there are more predictors than observations, the Lasso selects variables at most equal to the number of observations before it saturates due to the nature the convex optimization problem [5]. Second, when a group of variables has strong correlations among each other, the Lasso tends to only select one variable from the group [5]. As such, the Lasso loses the group information of the data due to the above limitations. Finally, the performance of the Lasso is dominated by the Ridge when there are more observations than predictors, meaning there are sufficient training data [5]. Therefore, it is desirable to further improve the performance of the Lasso.

An intuitive inspiration would be to find a regularization method to achieve a compromise between the performance of the two. In year 2003, Zou and Hastie proposed a novel regularization technique called the elastic net [5], which is a linear combination of the Ridge and the Lasso. According to Equation (2.6) ,the elastic net mimics the Ridge or the Lasso depending on whichever performs the best. Experiments have shown that the elastic net out performs the Lasso in many situations, while achieving a similar parsimony as the Lasso [5]. In our case, the speaker-dependent models are adapted using MAP estimation, which is essentially the same as the Ridge regression from a Bayesian point of view [4]. The Lasso can be added to the original MAP estimation to form the elastic net. In this thesis, we investigate the effect of elastic net by adding an L1-regularization to the traditional MAP speaker adaptations.

3. L-1 Regularized MAP Adaptation

Let $\mathbf{x} = \left(x_1, \cdots, x_T \right)$ be a sample of T i.i.d. drawn from a K components GMM. The joint distribution for each k is specified by the equation

$$p(\mathbf{x} \left| \mathbf{m}_k, \Sigma_k \right.) = \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{x}_t \left| \mathbf{m}_k, \Sigma_k \right. \right)^{z_{kt}} \tag{3.1}$$

where $\mathbf{m}_k$ and Σ_k are the mean vector and covariance matrix of each component. The covariance matrix is diagonal since it is assumed each frame of a speaker utterance is uncorrelated. The latent variable z_{ik} indicates the mixture component chosen by setting the corresponding entries to one while keeping the rest as zero. $\mathcal{N}\left(\mathbf{x}_t \left| \mathbf{m}_k, \Sigma_k \right. \right)$ is the k th Gaussian distribution given one frame of a speaker utterance, denoted by

$$\mathcal{N}\left(\mathbf{x}_t \left| \mathbf{m}_k, \Sigma_k \right. \right) \propto \left| \Sigma_k^{-1/2} \right| \exp\left(-\frac{1}{2} \left(\mathbf{x}_t - \mathbf{m}_k \right)^T \Sigma_k^{-1} \left(\mathbf{x}_t - \mathbf{m}_k \right) \right) \tag{3.2}$$

where $\mathbf{x}_t$ is a particular frame of one speaker utterance.

From the point of Bayesian regression, the traditional MAP estimation is equivalent to the maximum likelihood estimation with a Gaussian prior distribution on the weight coefficients [4]. The objective function equals to the likelihood function plus the prior distribution. The estimation result is more accurate with less data given more precise prior information. We would like to give a more precise prior information about speech features for more robust estimation. For real-world speech data, the empirical distribution is highly peaked at the center and has flatter tails than the Gaussian [12], which can be represented by a Laplacian distribution. In particular, we assume a sparse representation in model space for each speaker. Therefore, we set a prior as the weighted combination of a Gaussian and Laplacian distribution given by

$$p(\mathbf{m}_k \left| \mu_k, \Sigma_k \right.) = \mathcal{N}\left(\mathbf{m}_k \left| \mu_k, \Sigma_k \right. \right)^{\lambda_1} \exp\left(-\lambda_2 \left\| \Sigma_k^{-1/2} (\mathbf{m}_k - \mu_k) \right\|_1 \right) \tag{3.3}$$

where λ_1 and λ_2 are the weights of the Gaussian prior and the Laplacian prior respectively.

The optimization of the objective function given the new prior distribution is expressed as

$$\mathbf{m}_k = \underset{\mathbf{m}_k}{\operatorname{argmax}} \sum_{t=1}^{T} \gamma_k(t) \log \mathcal{N}(\mathbf{x}_t \,|\mathbf{m}_k, \Sigma_k) + \log p(\mathbf{m}_k \,|\mu_k, \Sigma_k) \tag{3.4}$$

where $\gamma_k(t)$ is the expectation of the latent variable z_{ik} derived as

$$\begin{aligned} \gamma_k(t) &= p\left(z_{kt} = 1 \middle| \mathbf{x}_t\right) = E\left[\#\text{ of times k is chosen} \middle| \mathbf{x}_t\right] \\ &= \frac{p(z_{kt}=1)\, p(\mathbf{x}_t \,|z_{kt}=1)}{\sum_{j=1}^{K} p(z_j = 1)\, p(\mathbf{x}_t \,|z_j = 1)} \\ &= \frac{\pi_{kt} \mathcal{N}\left(\mathbf{x}_t \,|\mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_{jt} \mathcal{N}\left(\mathbf{x}_t \,|\mu_j, \Sigma_j\right)} \end{aligned} \tag{3.5}$$

After computing the gradient, Equation (3.4) can be formulated as

$$\mathbf{m}_k = \underset{\mathbf{m}_k}{\operatorname{argmin}} \sum_{t=1}^{T} \frac{\gamma_k(t)}{2} \left\| \Sigma_k^{-1/2} (\mathbf{x}_t - \mathbf{m}_k) \right\|_2^2 + \frac{\lambda_1}{2} \left\| \Sigma_k^{-1/2} (\mathbf{m}_k - \mu_k) \right\|_2^2 + \lambda_2 \left\| \Sigma_k^{-1/2} (\mathbf{m}_k - \mu_k) \right\|_1 \tag{3.6}$$

which turns out to be the elastic net, where the L1-regularization and the L2-regualrization correspond to the Laplacian prior and the Gaussian prior respectively. The elastic net achieves a trade-off between the two regularizations as the L2-regularization keeps all the parameters in the model, while the L1-regularization selects as many parameters as the number of observations.

However, there is no closed-form solutions for Equation (3.6) due to the discontinuity of the gradient of the L1-regularization at zero. As such, we introduce two auxiliary vectors $\mathbf{a}_k$ and $\mathbf{b}_k$ such that [10]:

$$\mathbf{m}_k - \mu_k = \mathbf{a}_k - \mathbf{b}_k, \qquad \mathbf{a}_k \geq 0, \mathbf{b}_k \geq 0 \tag{3.7}$$

where $\mathbf{a}_k = \left[\mathbf{m}_k - \mu_k\right]^+$, which takes the positive entries of $\left(\mathbf{m}_k - \mu_k\right)$ while keeping the rest as zero. Similarly, $\mathbf{b}_k = \left[-\left(\mathbf{m}_k - \mu_k\right)\right]^+$. Then, Equation (3.6) can be rewritten as a constrained optimization problem given by

$$\begin{aligned} \left\langle \mathbf{a}_k, \mathbf{b}_k \right\rangle = \underset{\mathbf{a}_k, \mathbf{b}_k}{\operatorname{argmin}} \sum_{t=1}^{T} \frac{\gamma_k(t)}{2} \left\| \Sigma_k^{-1/2} (\mathbf{x}_t - \mu_k - \mathbf{a}_k + \mathbf{b}_k) \right\|_2^2 + \frac{\lambda_1}{2} \left\| \Sigma_k^{-1/2} (\mathbf{a}_k - \mathbf{b}_k) \right\|_2^2 \\ + \lambda_2 \mathbf{1}_s^T \Sigma_k^{-1/2} \left(\mathbf{a}_k + \mathbf{b}_k\right) \end{aligned} \tag{3.8}$$

where $\mathbf{1}_s$ is an s-dimensional vector with all entries equal to one.

The search of global optimum is relatively easy for Equation (3.8), because the gradient is readily obtained. We can solve this problem numerically using the gradient projection method [13].

4. Experiments

We performed experiments on the NIST 2008 Speaker Recognition Corpus. We basically follow the setup described in [1]. All algorithms were implemented in MATLAB.

For feature extraction, the audio was sampled and silence was removed using energy-based VAD algorithm. Then, the samples were normalized to prevent possible overflow and underflow. The 13-dimension MFCCs vector were calculated from pre-emphasized speech every 10 ms using 20 ms Hamming window [1]. Cepstral mean normalization (CMN) and RASTA [14] filters were applied to mitigate channel effects. Finally, first and second order difference MFCCs were appended to the 13-dimension MFCC vector to form a 39-dimension MFCC vector.

The GMM-UBM consists of 2048 Gaussian mixtures, and was trained using Expectation Maximization (EM) from Switchboard-II corpus. The background model training set contains 2048 speakers with an average duration of 30 seconds per speaker.

A GMM supervector was formed by stacking the adapted means of the GMM-UBM using Equation (3.8) for each speaker utterance, as shown in figure 3. The final normalized GMM supervectors were given by

$$\mathbf{S}_k = C_k^{\frac{1}{2}} \Sigma_k^{-\frac{1}{2}} \left(\mathbf{m}_k - \mu_k \right) \tag{4.1}$$

stacked for all k, which gives a compact sparse representation of the speaker-dependent model, where $C_k = \sum_{t=1}^{T} \gamma_k(t)$ is the expected number of samples per mixture, and Σ_k is diagonal covariance matrix of mixture k. Only the mean vector was updated, because adapting the weights and covariance does not make any pronounced differences [7]. The relevance coefficient λ_1 was chosen to be 16.

The SVM background training involves 2048 imposter speakers' GMM supervectors. We created eight supervectors for each target speaker, and trained an SVM speaker-dependent model using the target supervectors and the background supervectors.

5. Results

We took the labeled testing speaker supervectors as inputs to the SVMs, and determined the accuracy of the results by matching the output labels and the input labels. The total error rate was defined as the number of the incorrect results over the number of tests.

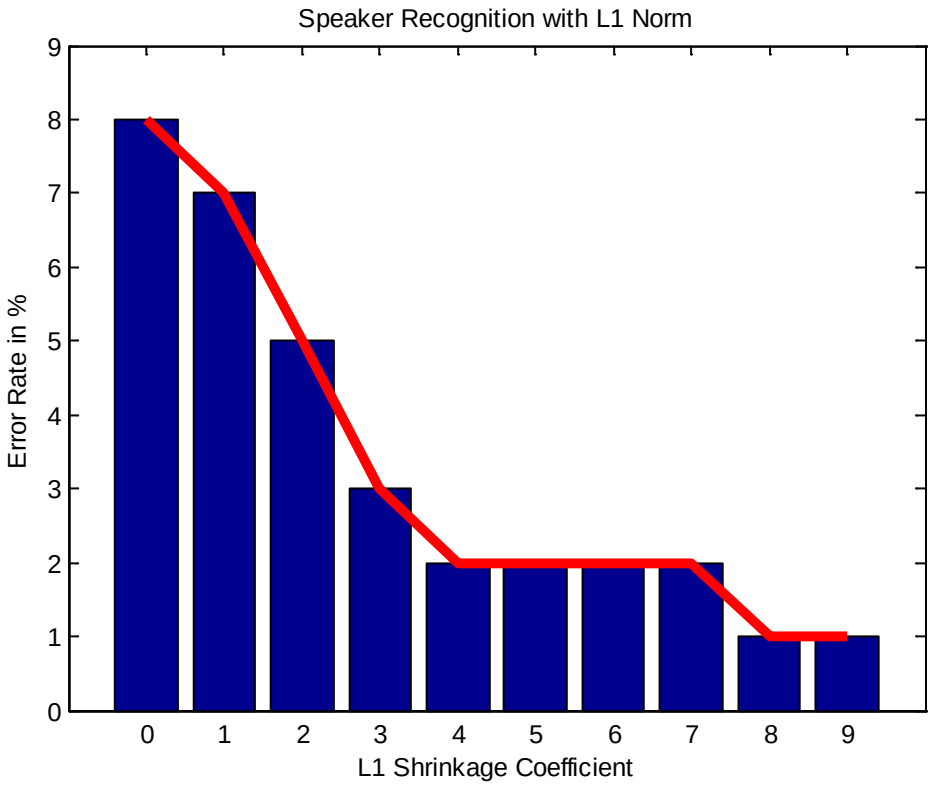

Figure 4. Error rate reduction with increasing La penalty coefficient

Figure 4 shows the experimental results on a subset of 100 test speakers using a fixed L2 penalty coefficient $\lambda_1 = 16$ and an increasing L1 penalty coefficient λ_2, where $\lambda_2 = 0$ corresponds to the baseline MAP estimation without the L1-regularization. We can see the total error rate reduced 7% for the optimal case.

However, as shown in Table 1, the total error rate reduction decreases when there are less overlap between the training set and the testing set for fixed number of test speakers and regularization parameters. The overlap is defined as the identical speech waveforms of the same speaker in both the training set and testing set. In other words, the error rate decreases when both the training and testing supervectors are adapted using the same utterances from one speaker.

Table 1. Error rate increases with less overlap

Amount of Overlap in %	Total Error Rate in %
75%	1%
50%	5%
25%	15%

An intuitive explanation for the above phenomenon is that the correlation between the training and testing features increases with the presence of data overlap, thus increases the recognition accuracy. Several experiments setup issues may have led to the accuracy reduction with less overlap. First, the GMM-UBM was not sufficiently trained in both data and numerical stability due to the limited computing power. Second, the channel variability was not mitigated, leading to channel mismatch issues. Although the experiments were not setup in more standard settings , the L-1 regularization effectively reduces the error rate by increasing the interpretability of the model as far as the sparse representation is concerned.

6. Discussion

The result shown in Figure 4 was based on a small subset of only 100 speakers using the odd indexed samples for training and even indexed samples for testing. The error rate was expected to be smaller than the published results, because there are limited number of speakers and significant overlap between the training set and testing set, which is quite different from the published setups. As such, we must modify and improve the experiments in order to compare our results with the published standard results. As a prerequisite of robust model adaptation, the UBM should be trained using an average duration of two minutes for each speaker. Besides, we must use the equal error rate (EER) defined as the error rate at which the false alarm rate and miss rate are equal. Furthermore, several channel compensation techniques in both feature and model space can be applied to maximize speaker-relevant information. We would like to show performance comparisons of channel compensated systems.

Table 2. Comparison of speaker recognition systems [15]

Systems	EER in %
GMM-SVM	14.79
GMM-SVM+NAP	5.78
i-SVM+LDA+WCCN	4.40

Linear discriminant analysis (LDA), within class covariance normalization (WCCN) and nuisance attribute projection (NAP) have been the standard model space techniques to suppress channel variability and emphasize speaker-relevant information. The i-SVM is the i-vector based method, which represents the GMM supervector in a single total-variability space [16]. According to Table 2, in future we would like to include channel compensation techniques in sparse modeling for the speaker recognition tasks.

7. Conclusion

In this thesis, we primarily investigated the effect of the L1-regularized MAP estimation of the GMM-UBM for SVM based speaker recognition. The idea of adding the L1-regularization is based on the assumption that text-independent speech feature is sparse in model space. The combination of the L1-regularization and MAP estimation, referred to as the elastic net, produces a sparse representation for the speaker-dependent model. From our experiment results on a subset of SRE-2008 speaker recognition evaluation task, we have observed error rate reduction and improved model robustness. However, further experiments have shown the error rate reduction decreased when there is less overlap between the training and testing data sets. This may be due to the increased numerical stability of the updates caused by insufficient training of the UBM as well as the channel variance. The performance could be improved by more rigorous training of the UBM and preprocessing of the feature space. In the future, we plan to study the effect of L1-regularization or elastic net on the standard speaker recognition systems.

References

[1] W. M. Campbell, D. E. Sturim and D. A. Reynolds, A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol.1, pp. 14-19 May 2006.

[2] W. M. Campbell, J. R. Campbell, D. A. Reynolds, E. Singer and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, pp. 210 - 229 2006.

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

[4] C. M. Bishop and M. E. Tipping, "Bayesian regression and classification," *Advances in Learning Theory: Methods, Models and Applications,* J. Suykens et al., eds., vol. 190, pp. 267-285, IOS Press, NATO Science Series III: Computer and Systems Sciences, 2003.

[5] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Roy. Statist. Soc.*, vol. 67, no. 2, pp. 301 -320 2005.

[6] R. Tibshirani, "Regression Shrinkage and Selection via the LASSO," *J. Royal Statistical Soc. B,* vol. 58, no. 1, pp. 267-288, 1996.

[7] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Dig. Signal Process.*, vol. 10, no. 1-3, pp. 19 -41 2000.

[8] T. Hasan and J. H. L. Hasen, "A study on universal background model training in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 1890-1899, Sept. 2011.

[9] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines*, Cambridge University Press, Cambridge, 2000.

[10] L. Liang, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for speech recognition," *Signal Processing Letters, IEEE*, vol. 18, no. 7, pp. 419-422, July 2011.

[11] T. Hastie, R. Tibshirani and J. Friedman *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2005: Springer

[12] T. Eltoft, T. Kim and T. Lee, "On the multivariate Laplace distribution," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 300-303, May 2006.

[13] M. Figueiredo, R. Nowak and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586-597 2007.

[14] F. Liu, R. Stern, X. Huang and A. Acero, "Efficient cepstral normalization for robust speech recognition," *Proceedings of ARPA Human Language Technology Workshop*, 1993.

[15] J. Kua, J. Epps and E. Ambikairajah, "I-vector with sparse representation classification for speaker verification," *Speech Communication., vol. 55, no. 5,* pp. 702-720 2013.

[16] A. Kanagasundaram , D. Dean, S. Sridharan, M. McLaren and R. Vogt, "I-vector based speaker recognition using advanced channel compensation techniques," *Computer Speech & Language*., vol. 28, no. 1, pp. 121-140 2014.