AN INVESTIGATION INTO SUBSPACE RAPID SPEAKER ADAPTATION FOR VERIFICATION

Simon Lucey and Tsuhan Chen

Advanced Multimedia Processing Laboratory
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh PA 15213, USA
slucey@ieee.org, tsuhan@cmu.edu

ABSTRACT

Rapid speaker adaptation is becoming more important in emerging applications where storage, computation and training utterances are at a premium (e.g. PDAs, cell phones). Effective adaptation can be achieved for the task of speaker verification, based on a maximum a posteriori (MAP) learning framework, by restricting the client's parametric model to be a linear combination of parameters estimated from training observations and a speaker independent "world" model (i.e. relevance adaptation (RA)). Subspace adaptation (SA) attempts to restrict a client's parametric representation to a pre-defined subspace during estimation. In this paper we elucidate where SA outperforms RA, demonstrate where and why SA is sometimes not as effective and give insights into what cost criteria should be used to construct the adaptation parametric subspace. Results are presented on the acoustic portion of the XM2VTS database for the task of Gaussian mixture model (GMM) based text-independent speaker verification.

1. INTRODUCTION

The *rapid* adaptation of speaker models for the purposes of speaker verification in emerging technologies such as mobile applications (i.e. cell phones, PDAs), where memory and computational capacity is at a premium, is a topic of great importance as these technologies cement themselves into our everyday lives. Unlike mobile speech recognition applications, feasible mobile speaker verification applications, due to security and cost constraints, require both the evaluation and estimation of speaker models on the system. This paper addresses the latter problem of estimating robust speaker models from a modest amount of training observations.

Rapid adaptation, borrowed from Kuhn et. al [1], refers to the ability of a system to estimate robust and accurate speaker models, whilst avoiding the need for "unacceptably large" amounts of adaptation observations for each speaker. The definition of "unacceptably large" varies from application to application but for the task of speaker verification on a low power, mobile device, with a simple lexicon (i.e. numerical digits) training good speaker models from *only* several seconds of speech (eg. five to ten) is ideal in terms of storage, computation and most importantly the creation of an ascetically pleasing device (i.e. not requiring the user to train the device with minutes of speech).

The generic task of automated speaker verification is considered a mature topic, with current state of the art speaker verification systems based on hidden Markov models (HMMs) for text-dependent tasks or Gaussian mixture models (GMMs) for text-independent tasks [2]. Training of these models is performed using the EM-algorithm [3], typically differences in adaptation techniques lie in how the parameters are constrained during the update portion of the EM-algorithm. For the task of speaker verification it has been demonstrated [2] good performance can be achieved, using a maximum a posteriori (MAP) learning paradigm, that restricts the client's parametric model to be a linear combination of parameters estimated from training observations and a speaker independent "world" model during the learning process. Although commonly referred to as MAP adaptation [2] we have decided to refer to the technique as *relevance adaptation* (RA) as the concept of MAP/Bayesian adaptation does *not* necessarily have to be restricted soley to this implementation.

Kuhn et. al [1] recently proposed a new adaptation technique, referred to as "eigenvoices", that restricts the client's parametric model to vary within a pre-defined subspace defined by preserving the major modes of variation from a development set of speakers. In its original form the eigenvoice technique employs principal component analysis [4] (PCA), however theoretically any criterion for defining a subspace can be used; this more general scenario we shall refer to as *subspace adaptation* (SA). In this paper we have investigated the possible use of linear discriminant analysis [4] (LDA), which uses a criterion of class separation to define its subspace, to restrict a client's parametric representation; as there has been some speculation [5] over the validity of reconstruction error as a criterion for forming the subspace. LDA has been previously used by Thyes et. al [6] for SA based speaker verification where it was compared to PCA. Our paper extends this initial work as we elucidate upon where and when PCA is better than LDA for the task of SA speaker verification.

Experimental results received indicate that criterion, other than reconstruction error with respect to the type of models being trained for (i.e. text-independent models), receives poorer performance during SA. Additionally, we have shown that SA, where there is extremely modest amounts of training observations, can marginally outperform RA. Finally we give some insights into how SA aids in model adaptation and possible avenues for further performance improvement. All experiments were conducted on the acoustic portion of the XM2VTS [7] database for the task of Gaussian mixture model (GMM) based text-independent speaker verification.

2. GAUSSIAN MIXTURE MODELS

Gaussian mixture models (GMMs) have been shown [2] empirically to be the classifier of choice for the task of text-independent speaker verification. A GMM models the probability distribution of a d dimensional statistical variable $\mathbf{o}$ as the sum of M multivariate Gaussian functions, A GMM models the probability distribution of a d dimensional statistical variable $\mathbf{o}$ as the sum of M multivariate Gaussian functions,

$$f(\mathbf{o}|\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{1}$$

where $\mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the evaluation of a normal distribution for observation $\mathbf{o}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The weighting of each mixture component is denoted by w_m and must sum to unity across all mixture components. The parameters of the model $\boldsymbol{\lambda} = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1}^{M}$ can be estimated using the Expectation Maximization (EM) algorithm [3] based on either a maximum likelihood (ML) or maximum a posteriori (MAP). K-means clustering was used to provide initial estimates of these parameters.

3. RELEVANCE ADAPTATION

MAP adaptation, or Bayesian adaptation as it is commonly referred to, is a technique for learning based on employing a priori knowledge of the parametric distribution $g(\boldsymbol{\lambda})$. An explicit form of MAP adaptation, which we refer to as relevance adaptation, has been shown [2, 8] to greatly improve automatic speaker verification performance over traditional ML training. Generic MAP adaptation attempts to incorporate the a priori knowledge in $g(\boldsymbol{\lambda})$ into the learning process, which results in trying to estimate an $\boldsymbol{\lambda}_{MAP}$ that satisfies,

$$\boldsymbol{\lambda}_{MAP} = \arg\max_{\boldsymbol{\lambda}} f(\mathbf{O}|\boldsymbol{\lambda}) g(\boldsymbol{\lambda}) \tag{2}$$

There are a variety of ways to gain a priori information about the distribution $g(\boldsymbol{\lambda})$. In speaker verification, the employment of a *world*, or *universal background model* as it is sometimes referred to [2], has been shown empirically to greatly improve speaker verification process. A world model is simply a single model trained from a large number of speakers representative of the population of speakers expected during verification, and usually has been estimated from a training set independent of the client to be adapted. This world model is typically trained using the ML criterion (i.e. no informative prior).

Given a world model $\boldsymbol{\lambda}_w = \{w_{w_m}, \boldsymbol{\mu}_{w_m}, \boldsymbol{\Sigma}_{w_m}\}_{m=1}^{M}$ and training observations from a single client, $\mathbf{O} = [\mathbf{o}_1, \ldots, \mathbf{o}_R]$, using the iterative EM-algorithm one can obtain update equations that incorporate the a prior knowledge in world model, to maximize the parametric representation of an GMM. Typically in GMM based speaker verification adaptation is only applied to the means of the mixture components, for RA this results in the following update equation,

$$\boldsymbol{\mu}_{c_m} = (1 - \alpha_m)\boldsymbol{\mu}_{w_m} + \alpha_m \frac{\sum_{r=1}^{R} \gamma_m(\mathbf{o}_r)\mathbf{o}_r}{\sum_{r=1}^{R} \gamma_m(\mathbf{o}_r)} \tag{3}$$

where $\gamma_m(\mathbf{o})$ is the occupation probability for mixture m and α_m is a weight used to tune the relative importance of the prior and is calculated via a relevance factor τ in,

$$\alpha_m = \frac{\sum_{r=1}^{R} \gamma_m(\mathbf{o}_r)}{\tau + \sum_{r=1}^{R} \gamma_m(\mathbf{o}_r)} \tag{4}$$

for our experiments an $\tau = 16$ received good results. The total number of parameters per client is $M \times d$.

4. SUBSPACE ADAPTATION

Kuhn et. al [1] recently developed a new approach for adaptation that preserves most of the variations between class models, but in a smaller parametric subspace $K << M \times d$. The main advantage of such an approach is the decrease in the number of free parameters needing to be found, allowing for the estimation of better trained models using less observations. A client model can be expressed as,

$$\boldsymbol{\mu}_c = \mathbf{V}\mathbf{x} \tag{5}$$

where $\mathbf{V} = \{\mathbf{v}_k\}_{k=1}^{K}$ is the concatenated matrix of the K eigenvectors/voices $\mathbf{v}_k$ corresponding to the K largest eigenvalues, $\boldsymbol{\mu}_c$ is the concatenated vector of M mixture component means $\boldsymbol{\mu}_{c_m}$ and $\mathbf{x}$ is the parameter vector of client c within the subspace. This parameter vector can be learned using a modified EM-algorithm as described by [1].

4.1. Criteria for generating a subspace

There are a myriad of techniques available for generating subspaces depending on specific cost criteria. When little is known about the nature of the problem (i.e. number of classes, how each class varies), a criteria of reconstruction error (PCA) often receives good performance. In this paper we were fortunate to have additional knowledge of how the parametric representation of clients vary between different digits, providing an ideal situation for investigating LDA as a plausible technique for creating a parametric subspace. Additionally we employed a randomly generated subspace, to evaluate the importance of choosing an effective cost criteria

PCA: attempts to generate a subspace based on selecting the eigenvectors of the scatter matrix $\mathbf{S}$ that corresponds to the K largest eigenvalues. In this paper PCA was used in two scenarios. First, a subspace was found using client models trained from across all digits to model solely how text-independent models $\boldsymbol{\mu}_c$ vary; which we will refer to as PCA text-independent (PCA-TI) where,

$$\mathbf{S}_{PCA-TI} = \sum_{c=1}^{C} \boldsymbol{\mu}_c \boldsymbol{\mu}_c' \tag{6}$$

Second, a subspace was found using conditional digit models of clients, as we are attempting to model both the variation in $\boldsymbol{\mu}_{c,d}$ from clients and text; which we will refer to as PCA-text-dependent (PCA-TD) where,

$$\mathbf{S}_{PCA-TD} = \sum_{c=1}^{C} \sum_{d=1}^{D} \boldsymbol{\mu}_{c,d} \boldsymbol{\mu}_{c,d}' \tag{7}$$

LDA: attempts to generate a subspace based on selecting the eigenvectors of $\mathbf{S}_b \mathbf{S}_w^{-1}$ that corresponds to the K largest

eigenvalues. Where,

$$\mathbf{S}_b = \sum_{c=1}^{C} (\bar{\boldsymbol{\mu}}_c - \bar{\boldsymbol{\mu}}_0)(\bar{\boldsymbol{\mu}}_c - \bar{\boldsymbol{\mu}}_0)' \tag{8}$$

$$\mathbf{S}_w = \sum_{c=1}^{C} \frac{1}{D} \sum_{d=1}^{D} (\boldsymbol{\mu}_{c,d} - \bar{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_{c,d} - \bar{\boldsymbol{\mu}}_c)' \tag{9}$$

It must be noted that the resulting eigenvectors, due to $\mathbf{S}_b\mathbf{S}_w^{-1}$ not being symmetric, must be found through simultaneous diagonalization [4] which does not ensure the resulting eigenvectors are orthonormal. However, the modified EM-algorithm in [1] requires each transform vector to be orthogonal; which if we assume $\mathbf{x}$ to be distributed according to a normal distribution equates to each element of $\mathbf{x}$ to be independent. To enforce this constraint, after preserving the eigenvectors corresponding to the K largest eigenvalues, a set of orthonormal vectors $\{\mathbf{v}_k\}_{k=1}^{K}$ were found that spanned the same subspace.

where C is the total number of clients and D is the total number of digits used to generate the subspaces, $\bar{\boldsymbol{\mu}}_c$ is the average parametric representation[1] of digits for client c and $\bar{\boldsymbol{\mu}}_0$ is the average representation between all clients and digits. Employing LDA and random based techniques as an alternate method to PCA for generating a subspace serves in two capacities. First, it gives one insights into how important the type of cost criteria is in obtaining good performance. Second, it demonstrates how useful class distinction is as a cost criteria for generating a compact subspace for adaptation.

5. FRONT-END PROCESSING

For feature extraction we used standard mel-frequency cepstral coefficients (MFCC) to generate 13 dimensional feature vector at 10ms intervals. Delta (first derivative) features were appended to this feature vector to create a 26 dimensional feature vector. Silence detection was performed using the bi-Gaussian method [5], where a two mode GMM is trained on a representative portion of the speech corpus with the hope that one Gaussian shall represent the speech features and the other Gaussian represent the silence features. Individual digit utterances were obtained for each speaker based on the length of the silence segments and the known digit order. Log energy and static MFCC coefficients were employed during the silence detection stage, with good segmentation results obtained.

6. EXPERIMENTS

Experiments were conducted on the digit acoustic portion of the XM2VTS [7] database, involving 16 repetitions of the digits 'zero' to 'nine' for each speaker taken over 4 recording sessions. The use of digits was chosen as this corresponds to a typical application scenario of speaker verification in a mobile application. The *Lausanne* 1 protocol [7] was used for our experiments with 200 speakers in the client set and 70 speakers in the test imposter set. Of the 16 digit sequence repetitions for each speaker, in the client set, 6 were used for training and 10 for testing. In total this resulted in 60 digit utterances of training observations for each client. For our experiments only a random subset of these training observations were ever used.

To ensure the separation of clients the first 100 speakers in the client set were used as a development set to train the world model $\boldsymbol{\lambda}_w$ and generate the subspace $\mathbf{V}$, with the remaining 100 speakers being used for testing. Each client model was tested using a randomly constructed sequence approximately 4 digits in length, as this was thought to be a typical in mobile applications (i.e. four digit security pin).

[1]Note $\bar{\boldsymbol{\mu}}_c$ is the average $\frac{1}{D}\sum_{d=1}^{D}\boldsymbol{\mu}_{c,d}$ *not* the model $\boldsymbol{\mu}_c$ trained across all digits.

7. SPEAKER VERIFICATION TASK

The speaker verification task is the binary process of accepting or rejecting the identity claim made by a subject under test. The verification process can be expressed simply as the decision rule,

$$\log f(\mathbf{O}|\boldsymbol{\lambda}_c) - \log f(\mathbf{O}|\boldsymbol{\lambda}_w) \underset{\text{accept}}{\overset{\text{reject}}{\lessgtr}} Th \tag{10}$$

where $f(\mathbf{O}|\boldsymbol{\lambda}) = \prod_{t=1}^{T} f(\mathbf{o}_t|\boldsymbol{\lambda})$ is the likelihood score describing how likely utterance $\mathbf{O} = [\mathbf{o}_1, \ldots, \mathbf{o}_T]$ belongs to the claimant speaker c and world model w respectively. A threshold Th needs to be found so as to make the decision. Speaker verification performance is evaluated in terms of two types of error being false rejection (FR) error, where a true client speaker is rejected against their own claim, and false acceptance (FA) errors, where an impostor is accepted as the falsely claimed speaker. The FA and FR errors increase or decrease in contrast to each other based on the decision threshold Th set within the system. A simple measure for overall performance of a verification system is found by determining the equal error rate (EER) for the system, where FA = FR.

8. RESULTS

Results were obtained for SA using various types and sizes of subspace. RA obtained results using a relevance factor of $\tau = 0$ and 16. Results for both SA and RA can be seen in Figure 1. Tests were constructed with the emphasis being placed on how well the client models generalize, irrespective of what digit utterance was being said. To this end the training digits used to train each client model were drawn randomly from the pool of 60 digit utterances available for each client.

9. DISCUSSION

From the results in Figure 1 one can see that SA actually performs slightly better than RA when extremely modest amounts of observations are used for adaptation (i.e. 3 random digit utterances) using a PCA-TI based subspace. This result is significant as SA seems to be able to constrain client models more effectively than RA, indicating that SA may have benefits over RA for speaker verification when adaptation data is at a premium. A cautionary note must be made to emphasize that SA performance, in its current form, is not significantly better than RA but the result is of importance concerning the further investigation and development of SA. From previous testing we have found a subspace size of $K = 20$, for PCA based SA, received best results when small amounts of data was used for adaptation. If one starts increasing the subspace size the average performance across all amounts of data improves;

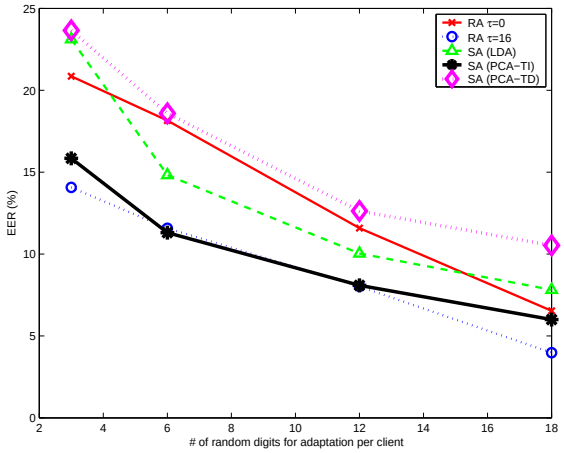

Fig. 1. A comparison between RA and SA across varying amounts of randomly drawn training digits.

but at the cost of performance when the amount of adaptation data is small.

SA using a LDA or PCA-TD subspace does not perform as well as PCA-TI. The LDA subspace does however outperform PCA-TD, as it tries to minimize the affect of

As expected good results are attained for RA using a relevance factor $\tau = 16$. Results for $\tau = 0$ are also shown to depict the ML case when no adaptation is performed. It was found that when there was a large amount of adaptation data (typically greater than 45 digits), as expected, ML and the MAP based RA approach each other, however with modest amounts of adaptation data the performance of ML is quite poor. Conversely, when the number of observations for adaptation increases past a certain point (typically 25 digits) then the performance of *all* SA techniques steadily degrade relative to the RA results.

One can attempt to explain these results if one thinks about SA as a quantized approximation to MAP adaptation. By enforcing a client's parametric representation to vary only within a pre-defined subspace one is essentially forming a binarized informative prior where parameters are able to vary freely within the subspace and are clamped within the residual space (i.e. where the parameters are not allowed to vary at all outside the subspace). This prior distribution is however improper (i.e. integrates to infinity within the subspace) so that it cannot be formulated naturally within the traditional MAP framework due to the prior not satisfying belonging to the conjugate family.

This quantization performs well when K is relatively small as it circumvents the "curse" of dimensionality, as the larger the dimensionality of $\mathbf{x}$ gets the more observations are required to gain adequate statistics. The increased stability stemming from the dimensionality reduction comes at cost, with less parametric variation available to discriminate between speaker classes. However, this harsh quantization tends to perform badly as K is increased with minimal amounts of adaptation data as their are too many degrees of freedom available within the subspace. PCA-TI is obviously a good, if not the best, criteria to employ for designing the parametric subspace as it ensures that the most varying portion of a client's parametric representation lies in the unclamped subspace, and that all variation stems solely from text-independent client variation. Other techniques like PCA-TD or LDA generated subspaces tend to perform worse as they often clamp these highly varying client portions, due to their differing criteria, while allowing variations due to differing text. SA using PCA-TD actually suffers from "catastrophic adaptation" as the number of adaptation observations increase. Catastrophic adaptation refers to the case where the adaptation performance attained is worse than the performance of the classifier with no adaptation (i.e. the ML ($\tau = 0$) case). As Kuhn et. al [1] highlighted this can be fundamentally attributed to the SA technique being inherently limited in the degrees of freedom the parametric representation can take on; such that in the limiting case with vast amounts of adaptation data, unlike MAP based RA, SA will *not* approach ML performance.

10. FUTURE WORK

The catastrophic adaptation problem faced by SA is inherently related to the harsh quantization that occurs from defining a subspace outside of which parameters cannot vary. An obvious shortcoming of SA currently is there is *no* constraint on variation of parameters within the subspace. It is felt by the authors that the incorporation of constraints, such as the eigenvalues of the space, could aid in improving performance. The integration of such a concept within the traditional MAP framework will be a topic of future research. An additional problem still remains however in the limiting case of SA, as catastrophic adaptation will still occur due to the suppression of any variation of parameters outside the subspace.

11. ACKNOWLEDGEMENTS

We would like to thank the Sony Corporation for their continuing support of this research.

12. REFERENCES

[1] R. Kuhn, J. Junqua, P. Ngyuen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on speech and audio processing*, vol. 8, no. 6, pp. 695–707, 2000.

[2] 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.

[3] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press Inc., 24-28 Oval Road, London NW1 7DX, 2nd edition, 1990.

[5] J. Mariethoz and S. Bengio, "A comparative study of adaptation methods for speaker verification," in *International Conference on Spoken Language Processing*, Denver, Colorado, September 2002, pp. 581–584.

[6] O. Thyes, R. Kuhn, P. Nguyen, and J.C. Junqua, "Speaker identification and verification using eigenvoices," in *International Conference on Spoken Language Processing*, Beijing, China, October 16-20 2000.

[7] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Audio- and Video- Based Biometric Person Authentication*, Washington, D.C., March 1999, pp. 72–77.

[8] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, April 1994.

[9] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Royal Statistical Society: B Series*, vol. 61, no. 3, pp. 611–622, 1999.

[10] D. K. Kim and N. S. Kim, "Bayesian speaker adaptation based on probabilistic principal component analysis," in *International Conference on Spoken Language and Processing*, 2000.