Improved Speaker Verification through Probabilistic Subspace Adaptation

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

In this paper we propose a new adaptation technique for improved text-independent speaker verification with limited amounts of training data using Gaussian mixture models (GMMs). The technique, referred to as *probabilistic subspace adaptation* (PSA), employs a probabilistic subspace description of how a client's parametric representation (i.e. GMM) is allowed to vary. Our technique is compared to traditional maximum a posteriori (MAP) adaptation, or *relevance adaptation* (RA), and maximum likelihood eigen-decomposition (MLED), or *subspace adaptation* (SA) techniques. Results are given on a subset of the XM2VTS databases for the task of text-independent speaker verification.

1. Introduction

Maximum a posteriori (MAP) adaptation offers certain advantages [1] over approaches, such as maximum likelihood (ML) training, which ignore the fact that the parameter (i.e. client's speaker model) is itself a random variable. However, as is often the case with MAP type methods, the nature and creation of the prior distribution governing how one's parametric representation varies is unclear.

The *rapid* estimation 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. Unlike mobile speech recognition applications, feasible mobile speaker verification applications, due to security and computational 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. Often this form of estimation is referred to as *adaptation* where one takes a pre-existing parametric representation for a known class (i.e. all speakers), where the representation is known to be well trained, and adapts it using a small amount of training observations to a less known class (i.e. single speaker); the less known class is usually a subset or a variant on the well known class. The resulting parametric model is often more accurate and robust than models trained purely from the less known observations alone.

We outline a technique that is able produce robust and generalizable client model's by employing probabilistic a priori knowledge of how a speaker's parametric representation can vary within a subspace that preserves the principal modes of parametric variation across all speakers. This technique is able to make estimates of "unseen" phonemic events[1] by learning many of the dependencies that exist between these events a priori from a developments set of well trained parametric speaker models. The adaptation technique in this paper, which we refer to as probabilistic subspace adaptation (PSA), is an extension to the maximum likelihood eigen-decomposition (MLED) adaptation technique proposed by Kuhn et. al [2] initially for the task of speech recognition; we refer to this adaptation technique simply as subspace adaptation (SA).

PSA is able to employ a Bayesian perspective to SA, by using a MAP instead of ML criterion, which results in more robust and stable client models, especially in the presence of scarce amounts of training observations. PSA is able to use a well defined subspace-prior distribution for MAP estimation, as the reduced dimensionality gained from the subspace representation allows the calculation of stable statistics from a development set of speakers. Previously, without the subspace representation, gaining such accurate statistics from the full parametric space was untenable.

We compare PSA's performance to the well known MAP technique first presented by Gauvain and Lee [3] and used with much success by Reynolds et. al [4] for the task text-independent speaker verification, we refer to this adaptation technique as relevance adaptation (RA). Throughout this paper we shall only be concerned with adapting the means of the mixture components, as the majority of class distinction between speakers can be attributed to the mixture component mean positions.

2. MAP and ML estimation using the EM-algorithm

Given that we have a set of training observations $\mathcal{S}_{trn}$ i.i.d. from an unknown distribution $f(\mathbf{o})$, but has an approximately known parametric form $\boldsymbol{\lambda}$, our task in MAP estimation is to find,

$$\boldsymbol{\lambda}_{MAP} = \arg \max_{\boldsymbol{\lambda}} f(\mathcal{S}_{trn}|\boldsymbol{\lambda}) g(\boldsymbol{\lambda}|\boldsymbol{\phi}) \qquad (1)$$

where $g(\boldsymbol{\lambda}|\boldsymbol{\phi})$ is the prior distribution of parametric form $\boldsymbol{\phi}$ governing how $\boldsymbol{\lambda}$ varies in parametric space.

Often times in statistics, it is not easy to select an appropriate prior distribution [1]. It is instead convenient to use an improper distribution (non-informative prior) that is represented by a nonnegative density function whose integral over the whole parameter space is infinite. We refer to this special

---

[1]The term phonemic event in this context refers to the mixture components found in a client's GMM, estimated from a data-driven clustering criteria as opposed to pre-ordained psychoacoustic labels.

case of MAP estimation as ML estimation where all knowledge about $\boldsymbol{\lambda}$ stems from the observations.

$$\boldsymbol{\lambda}_{ML} = \arg\max_{\boldsymbol{\lambda}} f(\mathcal{S}_{trn}|\boldsymbol{\lambda}) \qquad (2)$$

Dempster et. al [5] were able to develop an iterative algorithm referred to as the EM-algorithm, made up of an expectation (E) step and maximization (M) step, that is able to obtain a unique solution to $\boldsymbol{\lambda}$ provided their parametric form stems from the exponential family so that their well-known convexity property [5] can be taken advantage of. For ML estimation of $\boldsymbol{\lambda}$ whose parametric form is a mixture of M Gaussians (i.e. GMM) the EM-algorithm requires the maximization of the auxiliary function,

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(n)}) = E\{\log f(\mathcal{S}_{trn}, \mathbf{q}|\boldsymbol{\lambda})|\boldsymbol{\lambda}^{(n)}\} \qquad (3)$$

where $\mathbf{q} = [q_1, \ldots, q_R]$ is the hidden mixture component sequence where $q_r \in [1, \ldots, M]$ and $\boldsymbol{\lambda}^{(n)}$ is the previous iteration's estimate of $\boldsymbol{\lambda}$. The EM-algorithm can be applied equally well to MAP estimation [5] as long as the parametric form of ϕ belongs to the conjugate family of the complete-data density (i.e. the exponential family). The auxiliary function to be maximized according to the MAP criterion is defined by,

$$R(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(n)}) = E\{\log f(\mathcal{S}_{trn}, \mathbf{q}|\boldsymbol{\lambda})|\boldsymbol{\lambda}^{(n)}\} + \log g(\boldsymbol{\lambda}|\phi) \quad (4)$$

for both MAP and ML estimation using the EM-algorithm iterations are continued until a stable result is obtained.

Gaussian mixture models (GMMs) have been shown [4] 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,

$$f(\mathbf{o}|\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \qquad (5)$$

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 [5] based on either a maximum likelihood (ML) or maximum a posteriori (MAP). K-means clustering was used to provide initial estimates of these parameters. Using $M = 80$ mixture components received good results in our experiments.

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 $p(\boldsymbol{\lambda})$. An explicit form of MAP adaptation, which we refer to as relevance adaptation (RA), has been shown [3, 4] to greatly improve automatic text-independent speaker verification performance over traditional ML training.

There are a variety of ways to gain a priori information about the distribution of λ. In speaker verification, the employment of a *world*, or *universal background model* as it is sometimes referred to, in conjunction with a *relevance factor* has been shown [4] 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 a ML criterion and thus usually requires large amounts of training data to be trained satisfactorily.

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 the world model, to maximize the parametric representation of an GMM. This results in the following update equation,

$$\boldsymbol{\mu}_{c_m} = (1 - \alpha_m)\boldsymbol{\mu}_{w_m} + \alpha_m \frac{\sum_{r=1}^{R} \gamma_m(\mathbf{o}_r)\mathbf{o}_r}{\sum_{r=1}^{R} \gamma_m(\mathbf{o}_r)} \qquad (6)$$

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,

$$\alpha_m = \frac{\sum_{r=1}^{R} \gamma_m(\mathbf{o}_r)}{\tau + \sum_{r=1}^{R} \gamma_m(\mathbf{o}_r)} \qquad (7)$$

for our experiments an $\tau = 16$ received good results, it must be reemphasized that the work in this paper is concerned *only* with adapting the means. The total number of parameters per client is $M \times d$.

4. Subspace adaptation

Gauvain et. al [3], in their development of RA, made an assumption of independence between mixture components, such that a mixture component can *only* move if it has observations "seen" in it. This assumption, although empirically valid in the presence of reasonable amounts of training data, severely limits the ability to train robust speaker models with small amounts of data as there is often a number of "unseen" mixture components. Reasonable performance is still obtained using RA [4], as the relevance factor ensures that "unseen" mixture components remain close to their world model (i.e. average speaker-independent) positions.

An obvious approach to lessen the effect of these "unseen" mixture components is to try and learn the dependencies that exist between mixture components, such that mixture components can still move in an appropriate direction even if their are minimal to no observations "seen" in them. However, due to the parametric size of a speaker's GMM model (e.g. 80 mixture components with $d = 26$ results in 2080 free parameters for the mean representations) it is infeasible to gain accurate statistics of *all* these dependencies (e.g. it would require at *least* 2080 linear independent speaker models to obtain a fully ranked sample covariance matrix).

Kuhn et. al [2] 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,

$$\boldsymbol{\mu}_c = \mathbf{V}\mathbf{x} + \boldsymbol{\mu}_w \qquad (8)$$

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.

A variant of the EM-algorithm [2, 5] is employed so as to maximize the auxiliary function $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(n)})$ with respect to the subspace representation $\mathbf{x}$, whose update equation can be represented compactly in matrix form as,

$$\mathbf{x}_{ML} = \mathbf{A}^{-1}\mathbf{b} \tag{9}$$

where,

$$a_{k,j} = \sum_{m=1}^{M} \left(\sum_{r=1}^{R} \gamma_m(\mathbf{o}_r) \right) \mathbf{v}'_{k,m} \boldsymbol{\Sigma}_m^{-1} \mathbf{v}_{j,m} \tag{10}$$

$$b_k = \sum_{m=1}^{M} \sum_{r=1}^{R} \gamma_m(\mathbf{o}_r) \mathbf{v}'_{k,m} \boldsymbol{\Sigma}_m^{-1} \left(\mathbf{o}_r - \boldsymbol{\mu}_{w_m} \right) \tag{11}$$

and $\mathbf{v}_{k,m}$ represents the subvector of the eigenvector $\mathbf{v}_k$ corresponding to the mth mean mixture component. In a similar fashion to the calculation of the world model $\boldsymbol{\mu}_w$, the eigenvectors in $\mathbf{V}$ are calculated using principal component analysis (PCA) from a development set of well trained speaker models independent of the clients to be adapted.

5. Probabilistic subspace adaptation

An obvious shortcoming of SA, and PCA in general, is there is *no* constraint on the variation of parameters within the subspace. Tipping and Bishop [6] addressed this problem, for PCA, in the form of probabilistic PCA (PPCA) which models the subspace spanned by the eigenvectors as a Gaussian,

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \tag{12}$$

where $\mathbf{D}$ is the diagonal matrix containing the K largest eigenvalues corresponding to the eigenvectors $\mathbf{v}_k$.

Applying a MAP criterion, resulting in the maximization of $R(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(n)})$ with respect to $\mathbf{x}$ by assuming the distribution of $g(\boldsymbol{\lambda}|\boldsymbol{\phi})$ is described by the Gaussian in Equation 12, results in the update equation,

$$\mathbf{x}_{MAP} = \left(\mathbf{A} + \mathbf{D}^{-1} \right)^{-1} \mathbf{b} \tag{13}$$

We refer to the iterative application of Equation 13 as probabilistic subspace adaptation (PSA).

6. 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 [7], 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.

7. Experiments

Experiments were conducted on the acoustic digit portion of the XM2VTS [8] 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 [8] 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. 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.

To ensure the separation of clients the first 100 speakers in the client set were used as the 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).

8. 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,

$$\log f(\mathbf{O}|\boldsymbol{\lambda}_c) - \log f(\mathbf{O}|\boldsymbol{\lambda}_w) \mathop{\lessgtr}_{\text{accept}}^{\text{reject}} Th \tag{14}$$

where $f(\mathbf{O}|\boldsymbol{\lambda}) = \prod_{t=1}^{T} f(\mathbf{o}_t|\boldsymbol{\lambda})$ is the likelihood score describing how likely the test utterance $\mathbf{O} = [\mathbf{o}_1, \dots, \mathbf{o}_T]$ 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.

9. Results and discussion

Figure 1 contains comparative results between traditional RA and subspace based techniques SA and PSA. SA was evaluated in the form conceived by Kuhn et. al [2]; which we denoted as SA* and essentially set the world means in Equation 11 to zero ($\boldsymbol{\mu}_w = \mathbf{0}$) with all parametric variations being modelled in the subspace representation. We also evaluated SA (denoted without the * superscript) using the same world means used in RA; as a consequence SA and SA* required the generation of different subspaces[2] from the development set. The world means were used during the attainment of results for PSA. For the three subspace techniques (SA*, SA, PSA) in Figure 1 the best EERs were quoted for each training data amount from across 4 subspace sizes (K=20,40,60 and 80).

One can see in Figure 1 the benefit of using the world means

---

[2]For SA* a principal subspace was found using PCA on the development set without subtracting the world means $\boldsymbol{\mu}_w$; for SA the principal subspace was found after subtracting $\boldsymbol{\mu}_w$.

in the SA process, with equal to slightly better performance being seen across all amounts of training data. This improvement may be explained from the fixing of the world model in SA as opposed to SA* so that the variations modelled in the subspace concentrate only on the relative differences between speakers.

For PSA, with small amounts of training data, one can clearly see the performance improvement from employing a probabilistic description of how $\mathbf{x}$ is allowed to vary within the subspace. We think the benefit of the technique stems largely from the additional stability gained from incorporating the eigenvalues as well as the eigenvectors into the adaptation process. This point can be reinforced by comparing Equations 9 and 13. In Equation 9 the presence of minimal training observations will cause $\mathbf{A}$ to be poorly ranked making its inversion highly unstable. However, in Equation 13 the inversion of $\left(\mathbf{A} + \mathbf{D}^{-1}\right)$ is well ranked due to the condition of $\mathbf{D}$ being a diagonal matrix of eigenvalues. This point can be further argued from an empirical perspective if one inspects Figure 2.

In Figure 2(a) one can see that verification performance for SA is poorer with a large subspace size (K=80) than with a small subspace size (K=20) in the presence of small amounts of training data. This performance can largely be attributed to 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. For SA a tradeoff must be made between stability and variability. PSA can largely circumvent this tradeoff by ensuring stable performance, even when the dimensionality is reasonably high (ensuring more variability), as one can see from the results in Figure 2(b).

In this paper we have elucidated upon a technique that is able to learn a priori the *principal* modes of dependency between phonemic events (mixture components) for a population of speakers. The probabilistic description provided by PSA largely removes the tenuous balance between stability and variability found in SA. PSA however, still suffers the serious problem faced by SA in the presence of ample training observations as the result does not converge to the normal ML result as traditional RA does. Future work will try and incorporate work by Kim and Kim [9] that try and employ a variant of SA that provides a latent-variable full parametric space based on the work of Tipping and Bishop [6].
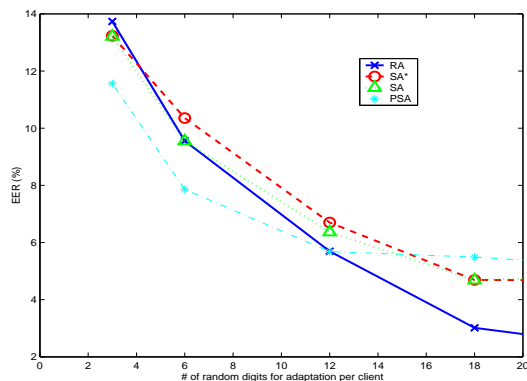

Figure 1: A comparison of EERs for RA ($\tau = 16$) SA, SA* (* denotes no world model used) and PSA across varying amounts of randomly drawn training digits. For subspace techniques the lowest EER was quoted from subspace sizes (K=20,40,60,80).
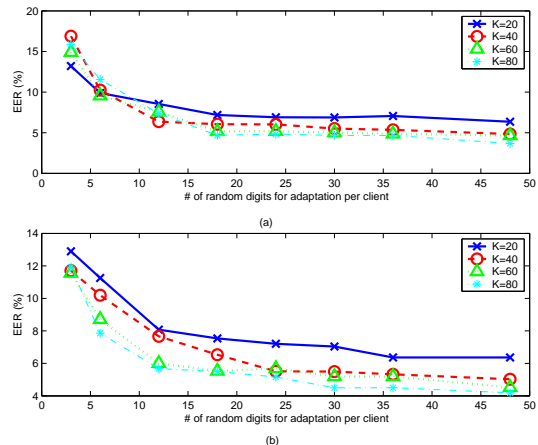

Figure 2: EERs for (a) SA and (b) PSA across different amounts of randomly drawn training digits and subspace sizes (K).

10. Acknowledgements

11. References

[1] M. H. Degroot, *Optimal statistical decisions*, McGraw Hill, 1970.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.