A Text-Independent Speaker Recognition System

Catie Schwartz
Ph.D. Student, Applied Mathematics and Scientific Computing
Department of Mathematics
University of Maryland, College Park
schwa2cs AT math.umd.edu

Dr. Ramani Duraiswami
Associate Professor, Department of Computer Science and
Department of the University of Maryland Institute of Advanced Computer Studies (UMIACS)
University of Maryland, College Park
ramani AT umiacs.umd.edu

Abstract

Speaker recognition is the computational task of validating a person's identity based on their voice. The two phases of a speaker recognition system are the enrollment phase where speech samples from the different speakers are turned into models and the verification phase where a sample of speech is tested to determine if it matches a proposed speaker. In a text-independent system, there are no constraints in the words or phrases used during verification. Numerous approaches have been studied to make text-independent speaker recognition systems accurate with very short speech samples and robust against both channel variability (differences due to the medium used to record the speech) and speaker dependent variability (such as health or mood of the speaker). A text-independent speaker recognition system using Gaussian mixture models and factor analysis techniques will be implemented in Matlab and tested against the NIST SRE databases for validation.

1 Project Background/Introduction

Humans have the innate ability to recognize familiar voices within seconds of hearing a person speak. How do we teach a machine to do the same? Research in speaker recognition/verification, the computational task of validating a person's identity based on their voice, began in 1960 with a model based on the analysis of x-rays of individuals making specific phonemic sounds [1]. With the advancements in technology over the past 50 years, robust and highly accurate systems have been developed with applications in automatic password reset capabilities, forensics and home healthcare verification.

There are two phases in a speaker recognition system: an enrollment phase where speech samples from different speakers are turned into models and the verification phase where a sample of speech is tested to determine if it matches a proposed speaker, as displayed in Figure 1. It is assumed that each speech sample pertains to one speaker. A robust system would need to account for differences in the speech signals between the enrollment phase and the verification phase that are due to the channels used to record the speech (landline, mobile phone, handset recorder) and inconsistencies within a speaker (health, mood, effects of aging) which are referred to as channel variability and speaker dependent variability respectively. In text-dependent systems, the words or phrases used for verification are known beforehand and are fixed. In a text-independent system, there are no constraints on the words or phrases used during verification. This project will focus on text-independent speaker verification systems.

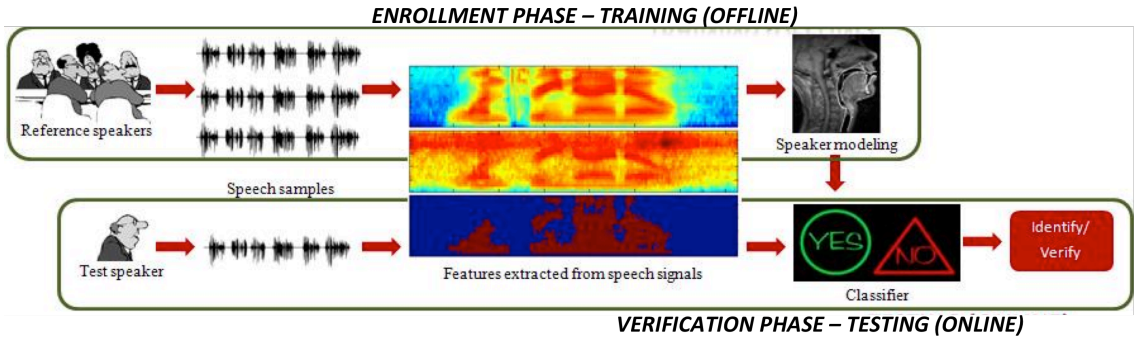

Figure 1: Speaker Recognition System (Courtesy of Balaji Srinivasan)

A variety of different features can be extracted from the speech samples, or utterances. Low level features relate to physiological aspects of a speaker such as the size of the vocal folds or length of vocal tract, prosodic and spectro-temporal features correspond to pitch, energy or rhythm of the speech, and high level features are behavioral characteristics such as accents or pronunciation. When extracting features, voice activity detectors (VADs) can be used to remove segments in an utterance where there is no speech. VADs can be energy based or based on periodicity. Many advanced systems account for multiple features and fusion is used to find the best overall match [2].

For text-independent speaker verification, the most popular modeling approaches are vector quantization (VQ), Gaussian mixture models (GMMs) and support vector machines (SVM). VQ is a technique that divides the features into clusters using a method such as K-means. GMM is an expansion of the VQ model, allowing each feature to have a nonzero probability of originating for each cluster [2]. A universal background model (UBM) representing an average speaker is often used in a GMM-based model. Adaptations of the UBM are used to characterize each of the individual speakers making the models robust even when the full phonemic space is not covered by the training data. SVM take labeled training data and seeks to find an optimized decision boundary between two classes which can be used to discriminate the different speakers.

Various techniques have been researched to assist in compensating for channel variability and speaker dependent variability, including speaker model synthesis (SMS) and feature mapping (FM). Most approaches require the speaker models to be organized into a high- and fixed-dimensional single vector called a supervector so that utterances with varying numbers of features can be represented in a general and compatible form. Popular methods that focus on compensating SVM supervectors include generalized-linear discriminant sequence (GLDS) kernel and maximum likelihood linear regression (MLLR) [2]. Factor analysis (FA) is a common generative modeling technique that is used on supervectors from GMMs to account for variability by learning low-dimensional subspaces. FA methods used in speaker verification include joint factor analysis (JFA) which model channel variability and speaker dependent variability separately, and total variability which model channel variability and speaker dependent variability in the same space. Normalization methods such as nuisance attribute projection (NAP), within-class covariance normalization (WCCN) and linear discriminant analysis (LDA) are also used for intersession variability compensation [2].

2 Approach

In this project, three separate speaker recognition systems will be developed, each of which will build upon the previous. All cases will be text-independent and will be implemented using mel-frequency cepstral coefficients (MFCCs) as the features with a voice activity detector (VAD) used to remove silent frames. The first speaker models will be UBM-adapted GMMs as shown in Figure 2. Once speaker models are created, a log likelihood ratio will be used to classify whether a test speaker is the same as a hypothesized speaker.

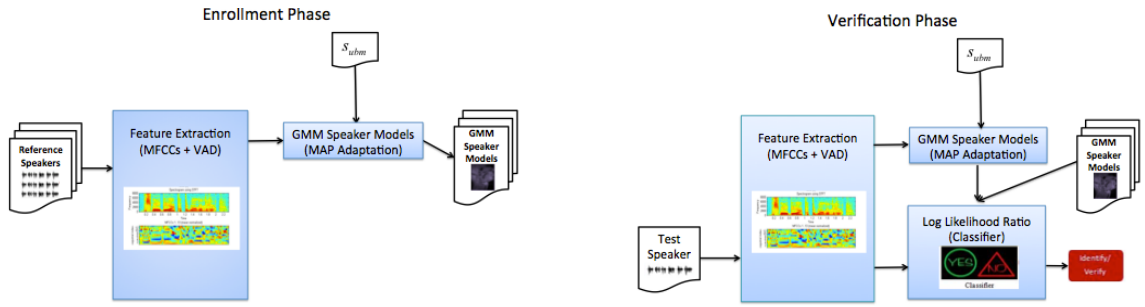

Figure 2: The first implementation for speaker models will be Gaussian mixure models based on the adaptation of the UBM. In the enrollment phase, a gallery of speaker models is generated from reference speakers. In the verification phase, a probe or test speaker is compared to a hypothesized speaker from the gallery using log likelihood ratio

Displayed as an input to the MAP Adaptation algorithm, the GMM Universal Background Model, s_{ubm}, is trained beforehand and only needs to be completed once. Other variables, such as the total variability space, T, and the reduced subspace, A, needed for the other speaker models are also trained prior to the enrollment phases of the different speaker recognition systems. The data used to create these variables are a large set of background speakers that do not overlap with the reference speakers in the speaker verification system. That said the background speakers should be a good representation of the

diversity found within the reference speakers. For example, if the speakers used as reference speaker are all male citizens of the United States, the background-training speakers should also be representative of all male citizens of the United States. Figure 3 illustrates all the training variables that need to be computed prior to speaker models being created.

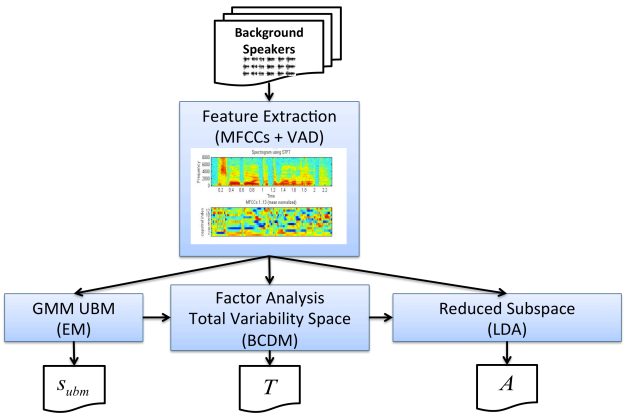

Figure 3: Algorithm flow chart for Background training

The second speaker models will be based on the idea that mean components in the GMMs concatenated into supervectors can create a new space. Factor Analysis (FA) techniques will also be used on the GMM supervectors to learn the low-dimensional total variability space. i-vectors will be extracted from the low-dimensional total variability space which uniquely represent the same information contained in the GMM supervectors. The i-vector speaker model flow chart is displayed in Figure 4. The i-vector speaker models build off of the GMM speaker models and use the total variability space trained using background speaker models. The classifier for i-vectors will be the discrete cosine score (DSC).

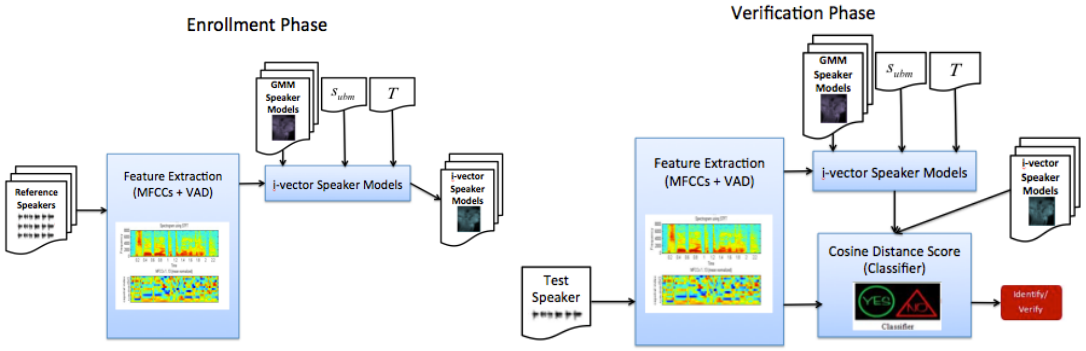

Figure 4: i-vector speaker verification

Linear discriminant analysis (LDA) techniques will then be applied to the i-vectors of the reference speakers to create the third and final speaker recognition system. Using LDA, a lower dimensional subspace will be found that will on maximize the inter-speaker variability and minimize speaker-

dependent variability. The algorithm flow chart for the LDA reduced i-vectors will be very similar to Figure 4, except i-vector speaker models will be used as an input, along with the total variability matrix and the reduced subspace matrix found in the background training section. Discrete cosine scoring (DCS) will also be used for verifying if a test utterance matches a proposed speaker.

2.1 Feature Extraction

The features used for this project will be mel-frequency cepstral coefficients (MFCCs). In order to make the MFCCs more robust, a voice activity detector (VAD) algorithm is implemented to find segments in the utterance in which no speech is detected.

2.1.1 MFCCs

Low-level features called mel-frequency cepstral coefficients (MFCCs) will be extracted from the speech samples and used for creating the speaker models. The mel-frequency scale maps lower frequencies linearly and higher frequencies on a logarithmic scale, as displayed in Figure 5, in order to account for the widely supported result that humans' can differentiate sound best at lower frequencies. Cepstral coefficients are created by taking a discrete cosine transform on the logarithm of the magnitude of the original spectrum. This step removes any relative timing, or phase, information between different frequencies and significantly alters the balance between intense and weak components [3]. MFCCs relate to the physiological aspects of a person such as the size of their vocal folds or length of their vocal tract and were first used starting in the 1980s [2]. They have been found to be fairly successful in speaker discrimination.

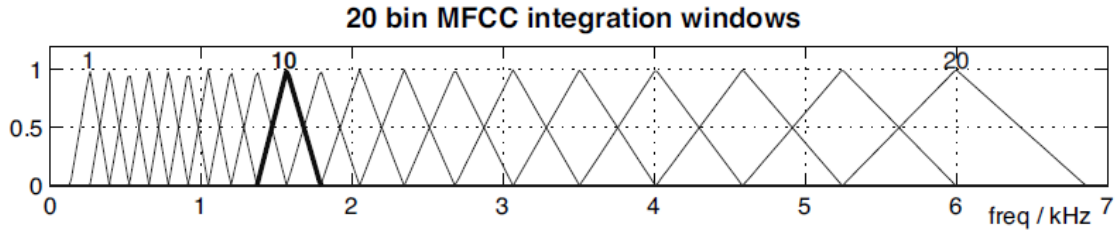

Figure 5: Illustration of 20 bins used for creating a 20-channel mel-frequency filterbank [3]. The highlighted bin 10 shows that most of the emphasis in this channel is on the center frequency with some weight placed on frequencies down to the center of the previous channel and up to the center of the next channel.

Given an utterance, it is first segmented using a 20 ms windowing process at a 10 ms frame rate. If the waveform is sampled at 16kHz, the 20 ms segment will contain 320 samples. The Fast Fourier Transform (FFT) algorithm with 512 points is applied to the speech sample. Then a mel-frequency filter bank is used to obtain an M-channel filterbank denoted as $Y(m), m = 1, \dots, M$. For this project, a 40-channel filterbank will be used. To convert the FFT power spectrum into a 40-channel filterbank, first the mel-frequency centers for the 40 different channels are determined. For each channel, weights for different frequencies will be computed with the most emphasis placed on the center frequency for the current

channel and some emphasis placed on frequencies down to the center of the previous channel and up to the center of the next channel as displayed in Figure 5. The channels are then scaled so that each channel has approximately a constant energy per channel [3]. Once the mel-frequency filterbank is computed, the MFCCs are found using the following formula:

$$c_n = \sum_{m=1}^{M}[\log Y(m)]cos\left[\frac{\pi n}{M}(m-\frac{1}{2})\right] \quad (1)$$

where n is the index of the cepstral coefficient. The 19 lowest DCT coefficients will be used for purposes of this project along plus 1 energy value (c_0). For initial verification, the 12 lowest DCT coefficients will be used along with the energy term. The algorithm to obtain the MFCCs is outlined in Figure 6.

Input: *utterance*; *sample rate*
Output: *matrix of MFCCs by frame*
Parameters: *window size* = 20 ms; *step size* = 10 ms
nBins = 40; *d* = 13 (*nCeps*)

Step 1: Compute FFT power spectrum
Step II : Compute mel-frequency m-channel filterbank
Step III: Convert to ceptra via DCT

$$c_n = \sum_{m=1}^{M}[\log Y(m)]\cos\left[\frac{\pi n}{m}\left(m-\frac{1}{2}\right)\right]$$

(0th Cepstral Coefficient represents "Energy")

Figure 6: MFCC algorithm

2.1.2 VAD

Since it is natural for people to pause while speaking, many of the 20 ms frames will contain no useful information. A simple energy based voice activity detector (VAD) will be applied to the speech signals in order to locate the specific frames that include speech segments and specific frames that are silent [2]. The algorithm for the VAD implemented is displayed in Figure 7. Note that the input and parameters to the VAD must match the input and parameters to the MFCC algorithm. For each 20 ms frame found to be silent, the entire frame is removed.

Input: *utterance*, *sample rate*
Output: *Indicator of silent frames*
Parameters: *window size* = 20 ms; *step size* = 10 ms

Step 1 : Segment utterance into frames
Step II : Find energies of each frame
Step III : Determine maximum energy
Step IV: Remove any frame with either:
a) less than 30dB of maximum energy
b) less than -55 dB overall

Figure 7: VAD algorithm

2.2 Gaussian Mixture Models using a Universal Background Model

Gaussian mixture models (GMMs) were first introduced as a method for speaker recognition in the early 1990s and have since become the *de facto* reference method [2, 4]. GMMs represent each speaker, s_i, by a finite mixture of multivariate Gaussians based on the *d*-dimensional feature vector $\boldsymbol{x}$:

$$p(\boldsymbol{x}|s_i) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \sum_k) \tag{2}$$

where K is the number of components, $\pi_k \geq 0$ represent the mixture weights that are constrained by $\sum_{k=1}^{K} \pi_k = 1$ and

$$N(\boldsymbol{x}|\mu_k, \sum_k) = (2\pi)^{-\frac{d}{2}} |\sum_k|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \sum_k{}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\} \tag{3}$$

where μ_k of dimension $1 \times d$ represents the mean value of mixture component *k* and $\sum_k$ of dimension $d \times d$ represents the covariance of mixture component *k* [4]. Given the sequence of T training vectors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ the GMM likelihood can be rewritten as

$$p(X|s) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\, s). \tag{4}$$

The values of $\pi, \mu, \sum$ representing each speaker, s, will be learned using maximum likelihood (ML) estimation techniques, which seek to find model parameters which maximize the likelihood of the GMM given the input training data, $\boldsymbol{x}$. Using full-covariance GMM normally requires a significant amount of training data and is very computationally intensive, therefore diagonal covariance matrices will be used.

A universal background model (UBM) or speaker-independent model is first created using speech samples from a large number of background speakers. Based on the newly found GMM-UBM components, $\pi^{UBM}, \mu^{UBM}, \sum^{UBM}$, a Bayesian adaptation technique is used to determine the components of the GMM for each individual speaker using an algorithm called Maximum a posteriori (MAP) adaptation.

2.2.1 GMM UBM

The parameters of the UBM are found using an expectation-maximization (EM) algorithm which iteratively refines a random initialization of GMM parameters to monotonically increase the likelihood of the estimated model based on the given feature vectors. This algorithm is part of the background training shown in Figure 3. For this algorithm, the MFCCs created from feature extraction will all be concatenated into a large matrix. It is important to have a large amount of data that equally represents the reference speakers that will be part of the speaker verification system. In a case where there is too much data to run on a given system, the data can be scaled down to 2*8*nCeps*K [reference], where K corresponds to the number of Gaussian centers, and still yield good results. In the case where the number of centers and the number of features are both low, the algorithm will choose a maximum between this number and 10000. It important for the data to be scaled down in a uniform and random manner to ensure that there is a good representation of the entire domain. For the initial

implementation, 512 Gaussian centers will be used. In most state of the art speaker recognition systems, 2048 Gaussian centers are currently used.

In the estimation step, the Bayesian statistics are used to determine the probability of mixture component c:

$$\gamma_t(c) = p(c|\boldsymbol{x}_t, s) = \frac{\pi_c N(\boldsymbol{x}|\boldsymbol{\mu}_c, \sum_c)}{\sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \sum_k)} \tag{5}$$

In the maximization step, the following are formulas are used which guarantee a monotonic increase in the model's likelihood value [5]:

Mixture weights:

$$\pi_c = \frac{1}{T} \sum_{t=1}^{T} \gamma_t(c) \tag{6}$$

Means:

$$\boldsymbol{\mu}_c = \frac{\sum_{t=1}^{T} \gamma_t(c) \boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma_t(c)} \tag{7}$$

Variances:

$$\sigma_c = \frac{\sum_{t=1}^{T} \gamma_t(c) \boldsymbol{x}_t^2}{\sum_{t=1}^{T} \gamma_t(c)} - \boldsymbol{\mu}_c^2 \tag{8}$$

The expectation step and the maximization step are iterative. The algorithm will cease to iterate when $\frac{abs(maxlikelihood_{old} - maxlikelihood_{new})}{abs(maxlikelihood_{old})} < 0.01$. There is also a maximum number of iteration that is set such that the algorithm should converge before the designated value. The algorithm is detailed in Figure 8. Remember that diagonal covariances will be used and a tolerance will be set so that the covariances will not grow too small.

Input: *Concatenation of the MFCCs of all background utterances* ($\mathbf{x} = \mathbf{x}_{background}$)

Output: $s_{ubm} = \{\pi^{ubm}, \mu^{ubm}, \Sigma^{ubm}\} = \{\pi, \mu, \Sigma\}$

Parameters: K = 512 (*nComponents*); *nReps* = 10

Step I : Initialize $\{\pi, \mu, \Sigma\}$ randomly

Step II: (Expectation Step)

Obtain conditional distribution of component c

$$\gamma_t(c) = p(c \mid \mathbf{x}_t, s) = \frac{\pi_c N(\mathbf{x}_t \mid \boldsymbol{\mu}_c, \Sigma_c)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \Sigma_k)}$$

Step III: (Maximization Step)

Mixture Weight: $\pi_c = \frac{1}{T}\sum_{t=1}^{T} \gamma_t(c)$

Mean: $\boldsymbol{\mu}_c = \frac{\sum_{t=1}^{T} \gamma_t(c)\mathbf{x}_t}{\sum_{t=1}^{T} \gamma_t(c)}$

Covariance: $\sigma_c = \frac{\sum_{t=1}^{T} \gamma_t(c)\mathbf{x}_t^2}{\sum_{t=1}^{T} \gamma_t(c)} - \boldsymbol{\mu}_c^2$

Step IV: Repeat Steps II and III until

$$\frac{abs(maxlikelihood_{old} - maxlikelihood_{new})}{abs(maxlikelihood_{old})} < 0.01$$

Figure 8: EM Algorithm for GMM

2.2.2 MAP Adaptation for Speaker Models

Given the GMM UBM, the GMM Speaker models can be determined as shown the algorithm flow chart in Figure 2. The speaker models will be made up of the parameters $s = \{\pi^{ubm}, \mu, \Sigma^{ubm}\}$, where π^{ubm}, Σ^{ubm} are values from the UBM and μ will be computed using the MAP algorithm and will be unique for each speaker. The algorithm could be modified to include adaptations of the weights and the covariances, but it has been found to degrade performance compared to only adapting the means.

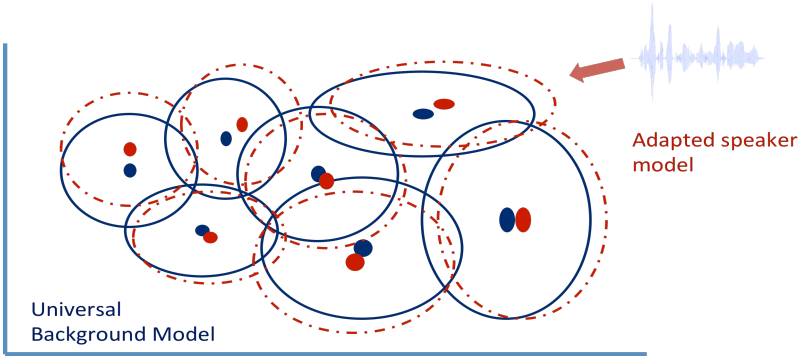

Figure 9: Maximum a posteriori (MAP) algorithm used to adapt the UBM (Courtesy of Balaji Srinivasan)

The first step of the MAP algorithm is the same as steps II and III in EM algorithm, that is, the values of μ_c are found using Bayesian statistics and ML estimations as in (6) and (7) on the reference database corresponding to a specific speaker using the UBM, s_{ubm}. Once these parameters are found, they are used to update the old UBM parameters $\boldsymbol{\mu}_c^{UBM}$ for mixture component c to create the adapted parameters for mixture component c with the equation:

$$\hat{\boldsymbol{\mu}}_c = \alpha_c^m \boldsymbol{\mu}_c + (1 - \alpha_c^m)\boldsymbol{\mu}_c^{UBM} \tag{9}$$

where α_c^m represent the adaptation coefficients controlling the balance between the old and the new estimates for the means [4]. The values of α_i^m are defined as

$$\alpha_c^m = \frac{\sum_{t=1}^{T} \gamma_t(c)}{\sum_{t=1}^{T} \gamma_t(c) + r^m} \tag{10}$$

where r^m is a fixed relevance factor. For this project, $r^m = 16$ will be used since experimental results have found performance to be rather insensitive to values in the range of 8-20 [4]. Using data-dependent adaptation coefficients enables de-emphasis on new parameters when a mixture component has a low probabilistic count with more emphasis on the old parameters. If a mixture component has high probabilistic counts, more emphasis can be placed on the new parameters. Having the ability to adjust the adaptation coefficients based on the data leads to robustness against limited training data. The MAP adaptation concept is illustrated in Figure 9 and the algorithm is outlined in Figure 10. Note that unlike the EM algorithm to determine the UBM, the MAP adaptation is not iterative.

Input: *MFCCs of utterance for speaker* ($\mathbf{x} = \mathbf{x}_{utterence}$);
$s_{ubm} = \{\pi^{ubm}, \boldsymbol{\mu}^{ubm}, \Sigma^{ubm}\}$

Output: $s_i = \{\pi^{ubm}, \boldsymbol{\mu}^{i}, \Sigma^{ubm}\}$

Parameters: K = 512 (*nComponents*); *r*=16

Step I : Obtain $\boldsymbol{\mu}_c$ via Steps II and III in the EM for GMM algorithm (using s_{ubm})

Step II: Calculate $\boldsymbol{\mu}_c^i = \alpha_c^m \boldsymbol{\mu}_c + (1 - \alpha_c^m)\boldsymbol{\mu}_c^{ubm}$

where $\alpha_c^m = \frac{\sum_{t=1}^{T} \gamma_t(c)}{\sum_{t=1}^{T} \gamma_t(c) + r}$

Figure 10: MAP Adaptation Algorithm

The mean components of the GMMs for each speaker can be concatenated into a high- and fixed-dimensional single vector of dimension $Kd \times 1$ where K is the number of Gaussian centers and d is the number of features. The vector is called a supervector and is illustrated in Figure 11. Supervectors are useful because utterances with varying numbers of features can be represented in a general and compatible form [2]. FA techniques will use the GMM supervectors as described in the next section.

Figure 11: Creation of supervectors from GMMs (Courtesy of Balaji Srinivasan)

2.3 Factor Analysis

Factor analysis is a statistical method used to describe variability among observed variables in terms of potentially lower number of unobserved variables called factors [6]. This method can be used to separate variability due to differences in channels or other nuisances from variability inherently within speakers as illustrated in Figure 12.

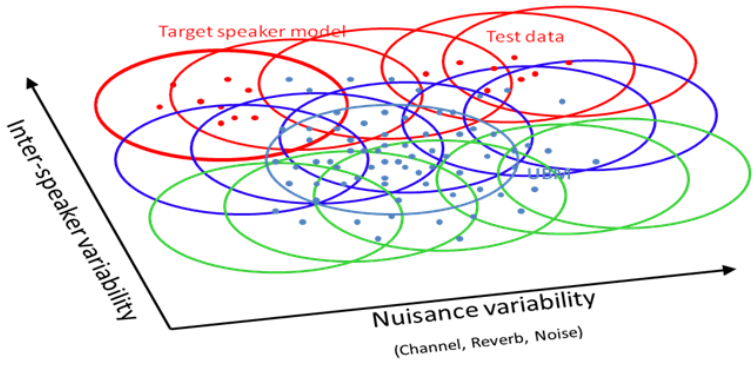

Figure 12: Inter-speaker variability versus nuiance variability. (Courtesy of Balaji Srinivasan)

The initial paradigm that incorporated factor analysis techniques looked into modeling channel-dependent variability explicitly different than speaker-dependent variability. This technique is called Joint Factor Analysis (JFA) and it separated the supervector of a speaker model μ_i into a speaker supervector s and a channel supervector c:

$$\mu_i = s + c \tag{11}$$

where s and c are normally distributed. The idea is to get both s and c in low-dimensional spaces, which is completed for the speaker supervector s by decomposing the supervector into speaker factors and residual factors:

$$s = m + Vy_i + Dz_i \tag{12}$$

In this equation, m is the speaker- and channel-independent supervector (UBM), V is a rectangular matrix of low rank, D is a diagonal matrix, and y and z are independent random vectors with standard normal distributions. The channel supervector c can be rewritten as

$$c = Ux_i \tag{13}$$

where U is a rectangular matrix of low rank and x has a standard normal distribution. Combining the two equations shows that the speaker model μ_i can be decomposed into low dimension spaces.

$$\mu_i = m + Vy_i + Ux_i + Dz_i \tag{14}$$

Dehak et al. [7] found that the subspaces U and V are not completely independent; therefore proposed a combined "total variability" space that will be used in this project. The speaker model supervector, μ_i will be decomposed as shown in the following equation

$$\mu_i = m + Tw_i \tag{15}$$

where T is rectangular matrix of low-rank representing the total variability space and w_i has a standard normal distribution. w_i represents the total variability factors and are often called intermediate/identity

vectors or i-vectors. Equation (15) implies that μ_i is normally distributed with mean vector m and covariance matrix TT^*.

The rank of T is set prior to training. The value 400 is normally used but a smaller number can be used given limited data. To train T, an algorithm using concepts of estimation-maximization (EM) is used. The method is very similar to a Probabilistic Principal Component Analysis (PPCA) approach [8], except PPCA is more restricting as it assumes isotropic Gaussian noise and FA does not. The algorithm used to train T is the same algorithm used to train the V matrix in JFA with only difference between training the V matrix in JFA is that in JFA, all recordings of a given speaker are considered to belong to the same person. In the total variability space, all utterances produced by a given speaker are regarded as having been produced by different speakers [7].

First, the Baum-Welch statistics [8] are calculated for a given speaker s and acoustic features $x_1, x_2, \dots, x_T$ for each mixture component c using equation (5):

$$N_c(s) = \sum_{t=1}^{T} \gamma_t(c) \tag{16}$$

$$F_c(s) = \sum_{t=1}^{T} \gamma_t(c) x_t \tag{17}$$

$$S_c(s) = diag(\sum_{t=1}^{T} \gamma_t(c)\, x_t x_t^*) \tag{18}$$

where $N_c(s)$, $F_c(s)$ and $S_c(s)$ represent the 0th, 1st and 2nd order statistics respectively. The 1st and 2nd order Baum-Welsh statistics are then centralized:

$$\tilde{F}_c(s) = F_c(s) - N_c(s) m_c \tag{19}$$

$$\tilde{S}_c(s) = S_c(s)(diag(F_c(s) m_c^* + m_c F_c(s)^* - N_c(s) m_c m_c^*) \tag{20}$$

Several matrices and vectors are defined based on the Baum-Welsh statistics. Let $\mathrm{NN(s)}$ be the CFxCF diagonal matrix whose diagonal block are $N_c(s)I$ $(c = 1, \dots, C)$. Let $\mathrm{FF(s)}$ be the $\mathrm{CFx1}$ supervector obtained by concatenating $\tilde{F}_c(s)$ $(c = 1, \dots, C)$. Let $\mathrm{SS(s)}$ be the CFxCF diagonal matrix whose diagonal blocks are $\tilde{S}_c(s)$ $(c = 1, \dots, C)$.

An iterative method is now used to determine the matrix T as described in [8, 9]. The first step is to determine the posterior distribution of the variables $w(s)$ given T. For the first iteration, a random initialization can be used for T. For each speaker, the following equation is defined

$$l_T(s) = I + T^* \sum^{-1} NN(s) T. \tag{21}$$

This will result in the posterior distribution of $w(s)$ conditioned on the acoustic observations of the speaker to be Gaussian distributed with mean $l_T^{-1}(s) T^* \sum^{-1} \tilde{F}(s)$ and covariance matrix $l_T^{-1}(s)$ [10]. The maximum-likelihood re-estimation step requires accumulating statistics over all the training speaker:

$$N_c = \sum_s N_c(s) \quad (c = 1, \dots, C) \tag{22}$$

$$A_c = \sum_s N_c(s) l_T^{-1}(s) \quad (c = 1, \dots, C) \tag{23}$$

$$\mathbb{c} = \sum_s FF(s)(l_T^{-1}(s) T^* \sum^{-1} FF(s))^* \quad (c = 1, \dots, C) \tag{24}$$

Given these values, a new estimate of the total variability space can be computed

$$T = \begin{bmatrix} T_1 \\ \vdots \\ T_C \end{bmatrix} = \begin{bmatrix} A_1^{-1} \mathbb{c}_{\mathbf{1}} \\ \vdots \\ A_1^{-1} \mathbb{c}_{\boldsymbol{c}} \end{bmatrix} \tag{25}$$

where

$$\mathbb{c} = \begin{bmatrix} \mathbb{c}_{\mathbf{1}} \\ \vdots \\ \mathbb{c}_{\boldsymbol{c}} \end{bmatrix}. \tag{26}$$

Several iterations (approximately 20) will be competed to obtain the trained total variability space. Once the space is defined, i-vectors are extracted are extracted using the knowledge that from (21), the expected value of an acoustic feature $w(s)$ is $l_T^{-1}(s) T^* \sum^{-1} \tilde{F}(s)$.

2.4 Linear Discriminant Analysis

Another dimensionality reduction technique called linear discriminant analysis (LDA) will be used. Once the total variability space, T, and the i-vectors, w from equation (15) are learned, LDA can be used to project the i-vectors into a lower-dimensional space, ω using the following equation:

$$\omega = Aw \tag{27}$$

The matrix A is chosen such that within-speaker, or speaker-dependent, variability (via the within-speaker covariance) is minimized and inter-speaker variability (via the inter-speaker covariance) is maximized within the space. The within-speaker covariance can be found for K speakers as:

$$S_W = \sum_{k=1}^{K} S_k \tag{28}$$

where

$$S_k = \sum_{n \in C_k} (x_n - m_k)\,(x_n - m_k)^T \tag{29}$$

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x_n \tag{30}$$

and N_k is the number of utterances for Speaker C_k. In order to determine the inter-speaker covariance matrix, we first consider the total covariance matrix

$$S_T = \sum_{n=1}^{N} (x_n - m)(x_n - m)^T \tag{31}$$

where m is the mean of the total data set

$$m = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\sum_{k=1}^{K} N_k m_k \quad (32)$$

and $N = \sum_k N_k$is the total number of utterances. The total covariance matrix can be decomposed as the sum of the within-speaker covariance matrix plus the inter-speaker or between speaker covariance matrix S_B.

$$S_T = S_W + S_B \quad (33)$$

In order for (33) to be true, S_B must be defined as

$$S_B = \sum_{k=1}^{K} N_k (m_k - m)\,(m_k - m)^T \quad (34)$$

Similar matrices can be construction that represent the covariances, s_W, s_B of the projected subspace. In order to construct a matrix in which the inter-speaker covariances is large and the within-speaker covariance is small, one could maximize the matrix

$$J(A) = Tr\{s_w^{-1} s_B\} \quad (35)$$

Solving this problem is the same as finding the eigenvectors of $S_W^{-1}S_B$. The dimension of the new subspace depends on the number of speakers used for training. For initial implementation, the dimension of the new subspace will be 200.

2.4 Classifiers

Two classifiers will be used for the accept/reject decision. A log-likelihood ratio test will be used based on the GMMs models and cosine distance scoring will be used on both the i-vectors and the intersession-compensated LDA reduced i-vectors.

2.4.1 Log-likelihood ratio test

Given a GMM speaker model s_{hyp} and the GMM-UBM s_{UBM}, a log-likelihood ratio test can be applied on the extracted features of a test utterance, X using the following formula [4]:

$$\Lambda(X) = \log p\big(X|s_{hyp}\big) - \log p(X|s_{UBM}) \quad (29)$$

where $\Lambda(X) \geq \theta$ will lead to verification of the hypothesized speaker, s_{hyp} , and $\Lambda(X) < \theta$ will lead to rejection.

2.4.2 Discrete cosine score

The discrete cosine score (DCS) can be applied to both the i-vectors, w, and the intersession-compensated i-vectors using LDA, ω using the following equation [9]:

$$score(\omega_1, \omega_2) = \frac{\omega_1^* \omega_2}{\|\omega_1\|\|\omega_2\|} = \cos\left(\theta_{\omega_1,\omega_2}\right) \quad (30)$$

where $score(\omega_1, \omega_2) \geq \varphi$ will lead to verification of the hypothesized speaker and $score(\omega_1, \omega_2) < \varphi$ will lead to rejection.

3 Implementation

In Phases II-IV (described in Section 7), a simple yet complete speaker recognition system will be implemented in Matlab on a modern Dell desktop computer. A software package that extracts the MFCCs from the utterances will be used [12], but all other code will be developed. Two classifier tests will be included to validate code at different the phases.

The implemented code will not be able to processes large amounts of data which is typically necessary for a robust speaker recognition system, especially in obtaining the GMM-UBM and the total variability matrix, T. Therefore, lower dimensional features and modest sized training sets will be used for initial test and validation. To test on larger data sets, numerical complexities and high memory requirements are expected and techniques will have to be implemented make the code work satisfactorily.

The results of Phase II-IV will impact the decision of what to implement in Phase V. If reasonable results are obtained using the implemented code from Phase II-IV, more features may be added to the system. If the code written to obtain the GMM-UBM or the total variability matrix is found to be inefficient, Phase V may be to parallelize/optimize the inefficient code. If this is the case, the task will be too computationally intensive to complete the task on a single computer and will therefore be completed on a cluster. The code will most likely be implemented in Matlab, using C and MPI if necessary. Another option of Phase V is to complete more extensive testing using different inputs into the vetted speaker recognition system.

4 Databases

The National Institute of Standards and Technology (NIST) has coordinated Speaker Recognition Evaluations (SRE) approximately every two years since 1996. In support of the SRE, datasets consisting of *.wav and *.sph formatted files with a sampling rate of 8kHz are provided for use of the participants. The databases that will be used for this project is the NIST 2008 SRE database and the NIST 2010 SRE database, both of which contain speech data in several different conditions including data from interviews, microphones, telephone conversations. The NIST 2008 SRE database will only be used if the amount of data from the NIST 2010 SRE is too much to process for phases II-IV. The NIST 2010 SRE database contains utterances from approximately 12000 different speakers.

The TIMIT dataset designed by SRI International, Texas Instruments, Inc (TI) and Massachusetts Institute of Technology (MIT) in 1990 will be used for initial testing in order to obtain results on a smaller set of speakers. This database contains recording of 630 speakers from eight different American dialect regions, each reading ten sentences that have been chosen to encapsulate the phonetic space. The waveforms are recorded at 16kHz. Performance on the TIMIT dataset is expected to be good as the

TIMIT files are relative clean and have little variations due to channel variability and speaker-dependent variability.

5 Validation

5.1 Feature Extraction

5.1.1 MFCCs

The code to create the MFCCs was modified from a tool set created by Dr. Dan Ellis from Columbia University. In order to validate this code, the results of the modified code were compared to the results of the original code given the same inputs.

5.1.2 VAD

As shown in Figure 13, a tool was created to display an input waveform to the VAD algorithm along with the VAD results marking the detected speech segments. By visual inspection, it seems as though the detected speech segments looks reasonable. The tool also enables the user to listen to the original waveform, the concatenation of the silent frames along with the concatenation of the frames in which speech was detected. There is never any loss of speech data in the new speech segment that can be heard and silent frames sound silent. When the silent frames are normalized, the result sounds like noise.

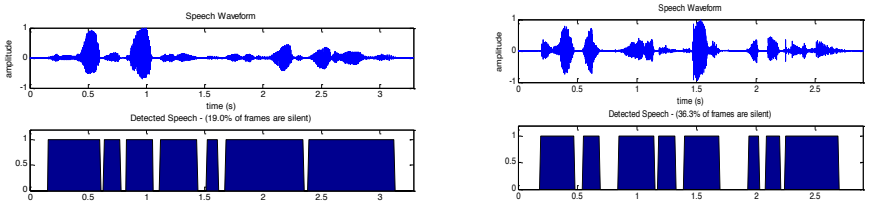

Figure 13: Speech Detection from VAD on two different waveforms

5.2 GMM

In order to test the GMM speaker models, a tool was created in order to make toy set of examples in two dimensional features space. The tool enables the user to choose the number of Gaussian centers for the data. Besides the number of Gaussian centers, the number of test speakers and the number of training speakers must be designated. Figure 14 shows examples in which the number of Gaussian centers were chosen to be 3 for (a) and 128 for (b). Each time the function is run, a new set of centers is created, even if the number of Gaussian centers is chosen to be the same. Numerous examples were used for validation but only one per number of components will be used in this presentation. The number 3 was chosen in order to let the reader visually understand the algorithm's performance. The number 128 was chosen to enable to reader to see the difficulties in understanding the results in 2-

dimensional space when only 128 components are used. The smallest set of true data used in the speaker recognition project will be 512 Gaussian centers in 13-dimensional feature space.

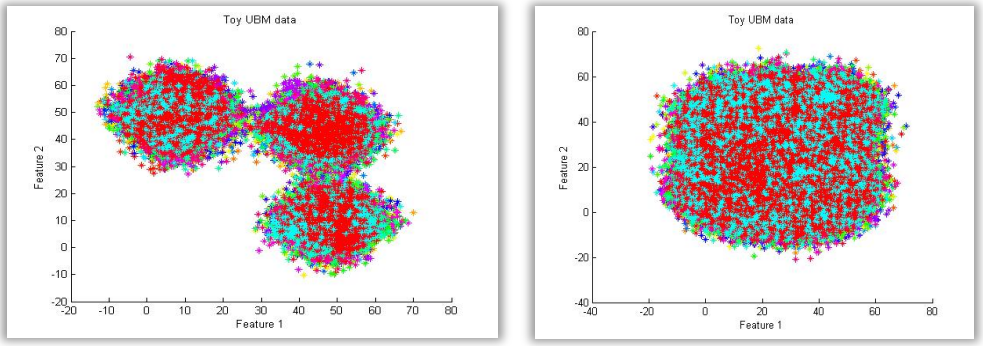

Figure 14: Result of tool to create toy datasets using different numbers of Gaussian Centers.
a) User designated 3 Gaussian centers b) User designated 128 Gaussian centers.

First, the true Gaussian centers are determined randomly on a uniform space. Then, for each speaker, a mean offset and covariance is chosen for the speaker for each true Gaussian component. The speaker can have a random number of data points for a given component, and different components have a different probability that speakers will have data related to the component. Once the mean offsets, the covariance and the number of data points is determined for a given speaker and a given component, random features are generated and added to the speaker's feature matrix that represents the features in an utterance. Two separate utterances are generated for each speaker that use the same mean offsets and covariances, but will have a different number of data points per component and different random data points. In Figure 14, each speaker is represented in a different color.

The same method of creating features for the speaker's utterances is used for both the training speakers and the test speakers, but all have different mean offsets and covariances.

5.2.1 EM For GMM UBM

The results of both utterances for all of the speakers in the training set are used to create the GMM UBM. Figure 15 displays all the iterations of a random initialization of the EM algorithm for GMM. Each iteration comes closer to the true answer until the values converge. Figure 16 displays examples of other random initializations and the number of iterations it took for the EM algorithm to converge. The black data points represent all of the data points used to train the GMM UBM. The green x denotes the mean of a given Gaussian component and the green ellipses in which the x's are the centers represent the covariances of the given Gaussian components. Note that not every random initialization will yield a desirable result. Therefore it is advised to complete multiple repetitions and keep only the parameters that yielded the largest maximum likelihood.

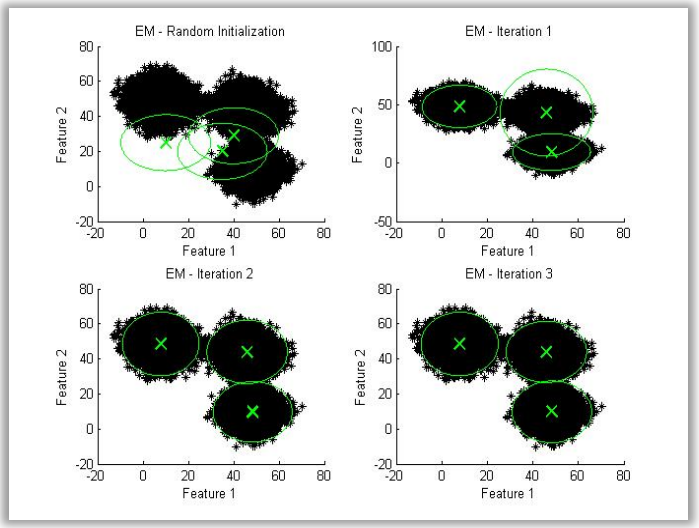

Figure 15: Iteration in the EM algorithm showing convergence

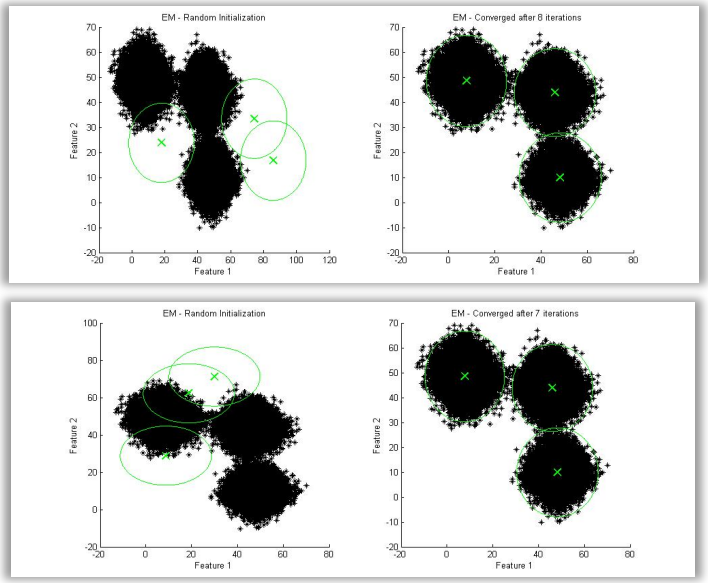

Figure 16: Convergence of the EM algorithm with different initializations.
a) converged after 8 iterations. b) converged after 7 iterations.

Both Figures 15 and 16 displayed results in which the number of Gaussian centers entered in as an input to the EM algorithm matched the number of true Gaussian centers. It is an interesting study to understand what happens when these numbers do not agree. As shown in Figure 17, if the value used in the EM algorithm is smaller than the number of true Gaussian components, the result is poor because the algorithm is forced to merge at least two true Gaussian components together. This will result in parameters that do not match any true Gaussian.

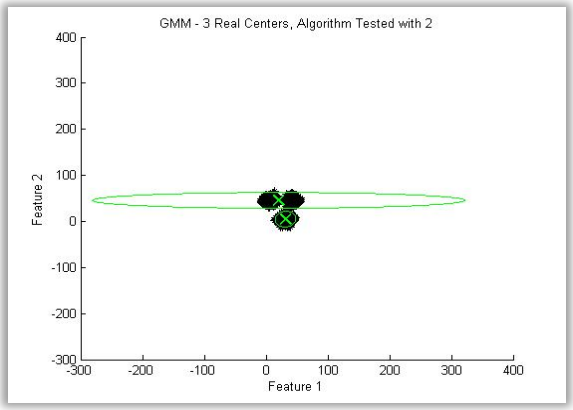

Figure 17: K=2 is less than the number of true Gaussian components (3)

If the value entered into the EM algorithm is larger than the number of true Gaussian components and there is a sufficient amount of data, the true Gaussian components are estimated well and any additional parameters represent made up Gaussian centers that are created by less than 5% of the data per center and often much less. This is true even when the number of proposed Gaussian centers for the EM algorithm is over twice the number of true centers, as displayed in Figure 18 b.

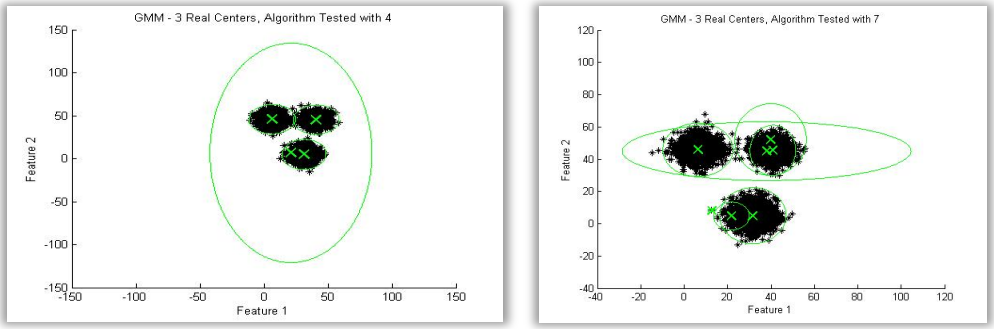

Figure 18: a) K=4 and b) K = 7 is more than the number of true Gaussian components (3)

All of the above examples were using a data set in which there were 3 true Gaussian components. The lowest number of Gaussian components analyzed in the speaker recognition project will be 512 and will be in 13 dimensional space. To illustrate the difficulties in visualizing the data set, Figure 19 shows an example in 2 dimensional space with 128 components. The covariance for only three components are drawn, otherwise the entire figure would be covered with green ellipses. In this case, the algorithm converged after 3 iterations.

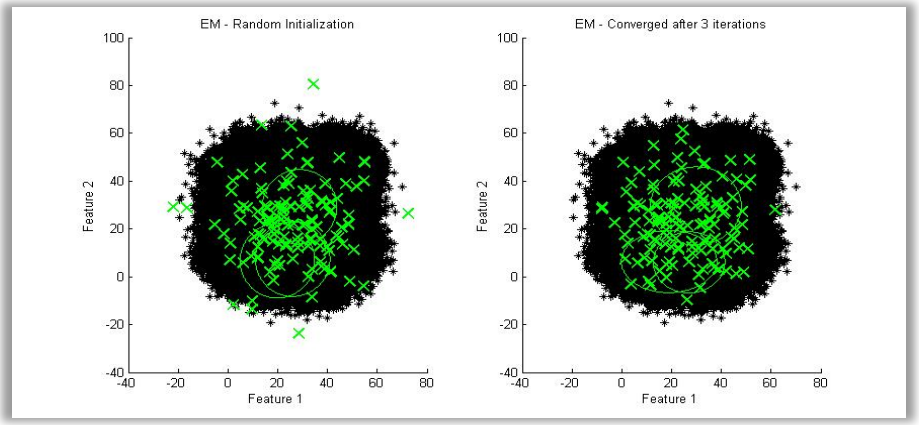

Figure 19: EM algorithm results with 128 true components and K=128

5.2.2 MAP Adaptation For GMM Speaker Models

The MAP adaptation algorithm should set the means for a speaker model to be the means of the UBM if there is very little or no data representing a given component for a given speaker. If there is sufficient data, the means for the speaker model should be representative of the mean for the speaker increasingly as the amount of data for the component per speaker is increased. Given these expectations, a visual inspection can be used to validate the algorithm when the number of components can be observed. Figure 20 shows results for 4 different speakers when the number of true components is equal to 3. The black data points represent the data points in the Universal Background Model whereas the red data points represents the data points for the specified speaker. As expected, when the number of data points for a given component for the speaker is increased, the mean for the component lays on top of the mean for the speaker model. When there is no data for the speaker for a given component, as observed in one of the components for speaker 2, the mean is the same for the speaker as it is for the UBM. In cases where there are some data points corresponding to a component, but not a significant amount, it should be observed that the mean for the speaker is in between the mean of the UBM and the estimated mean of the speaker for the given component.

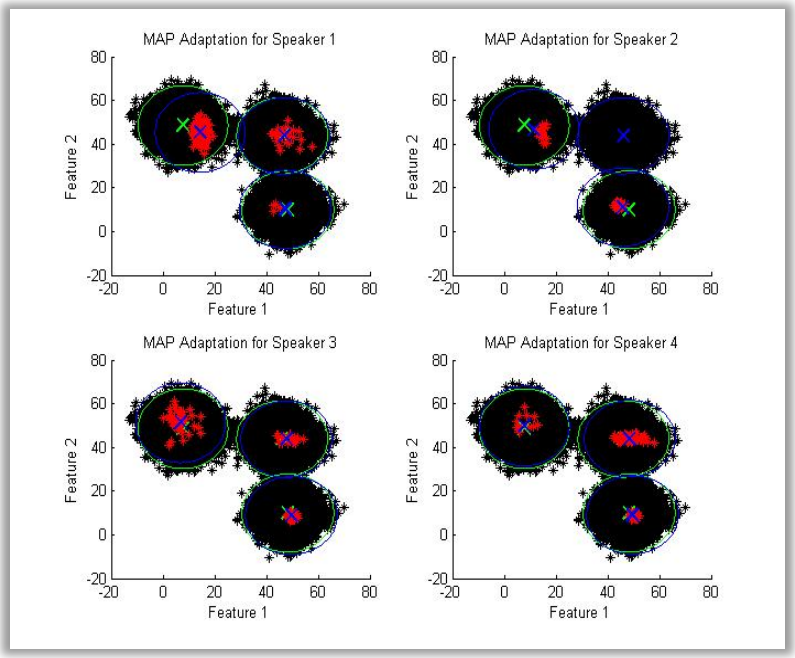

Figure 20: The results of the MAP algorithm of four separate speakers

Figure 21 illustrates the MAP adaptation results in the case in which there were 128 Gaussian components. There are many more cases in which a component for a given speaker is not represented by the data in a given utterance and therefore the means of the components are often represented by the mean of the UBM.

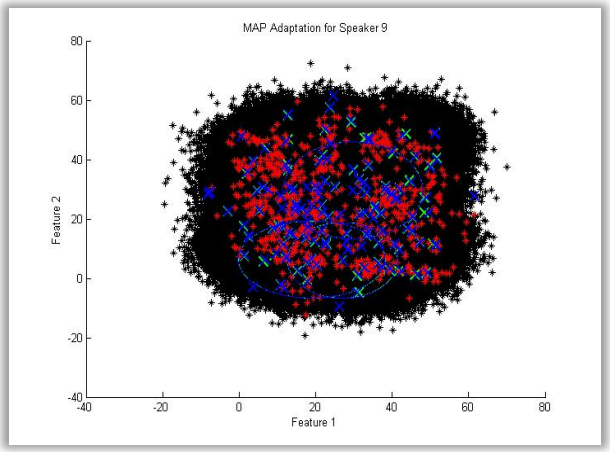

Figure 21: Example MAP Adaptation for speaker with the number of Gaussian Components being 128

5.3 Factor Analysis

Similar to GMM validation, a toy example will be created in order to ensure that the factor analysis algorithm is implemented properly by visual validation. As of now, the algorithm is implemented and some methods have been implemented for validation but the validation process is not complete.

5.4 Linear Discriminant Analysis

Also similar to GMM validation, a toy example will be created in order to ensure that the Linear Discriminant Analysis algorithm is implemented properly by visual validation. As of now, the algorithm is implemented and some methods have been implemented for validation but the validation process is not complete.

5.5 Validation in general

Three commonly used metrics will be used for validation in this project. Equal error rate (ERR) is a measure that gives the accuracy at decision threshold for which the probabilities of false rejection (miss) and false acceptance (false alarm) are equal. This measure is good at obtaining a first quick understanding of whether there are any bugs because large values are not expected. Detection error trade-off (DET) curves will also be used for visual inspection. Lastly, the MinDCF algorithm used by NIST in the evaluation of the SRE will be examined.

Validation will ensure that the code is working properly in order to complete Phases II, III and V. Phase II marks the completion of using the EM and the MAP algorithm to generate the speaker model supervectors. A likelihood ratio test will be used as the classifier to validate results at this phase. Phase III uses FA techniques and LDA to create a low-dimensional space in which interspeaker variability is maximized and with-speaker variability is minimized. Discrete cosine scoring (DCS) can be used as the classifier and results can be tested after the FA step on the i-vectors and after the LDA step. Results from the DCS of the i-vectors should be an improvement over the likelihood ratio test in Phase II. The results from LDA should be an improvement over both of the previous score.

Validation in this manner mixes in concepts of testing and therefore results of validation in general will be placed in the test section.

6 Testing

Initial testing has been done on the TIMIT database. Using the MFCC algorithm with 12 cepstral coefficients + 1 energy feature (c_0), the VAD, and the GMM model with 512 components, the Detection Error Tradeoff (DET) curve is displayed in Figure 22.

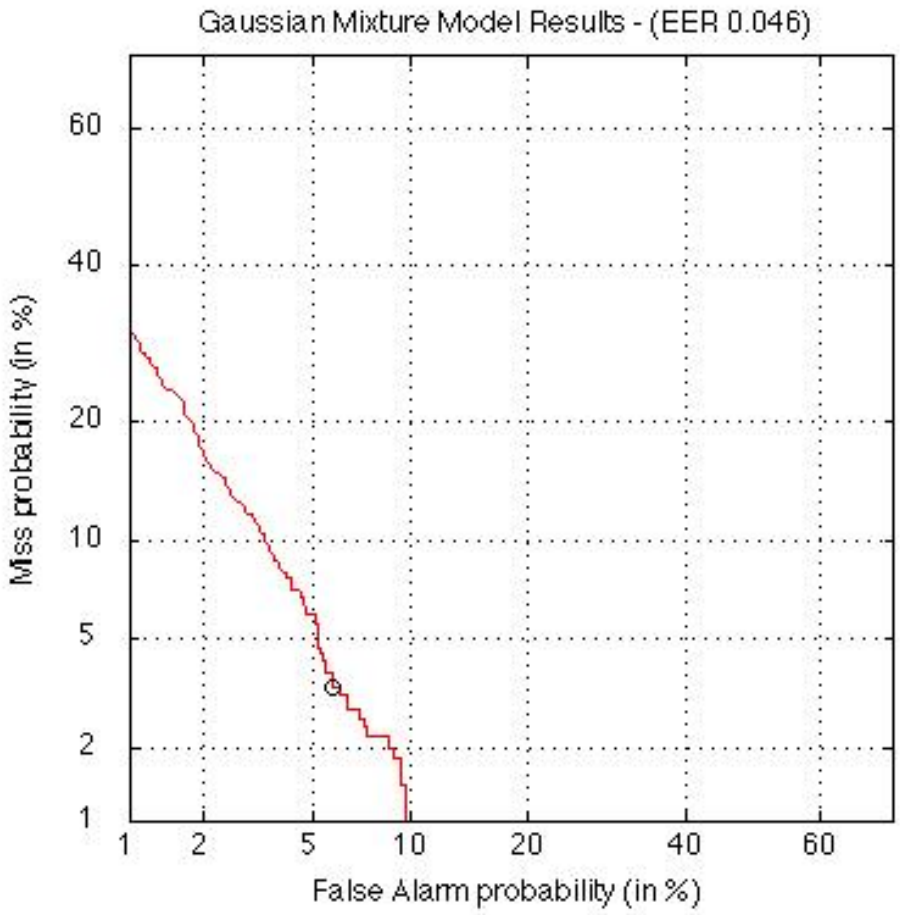

Figure 22: DET of the GMM Speaker Models with 12 cepstral coefficients + 1 energy based

These results shown are comparable to the state-of-the-art results in 2008. A number of small tests have been completed for comparison of different inputs given the algorithms. It was interesting to see that turning the VAD algorithm off lead to better results. This is not expected and will need to be analyzed in depth.

The results using factor analysis techniques are better, as expected, as shown in Figure 23.

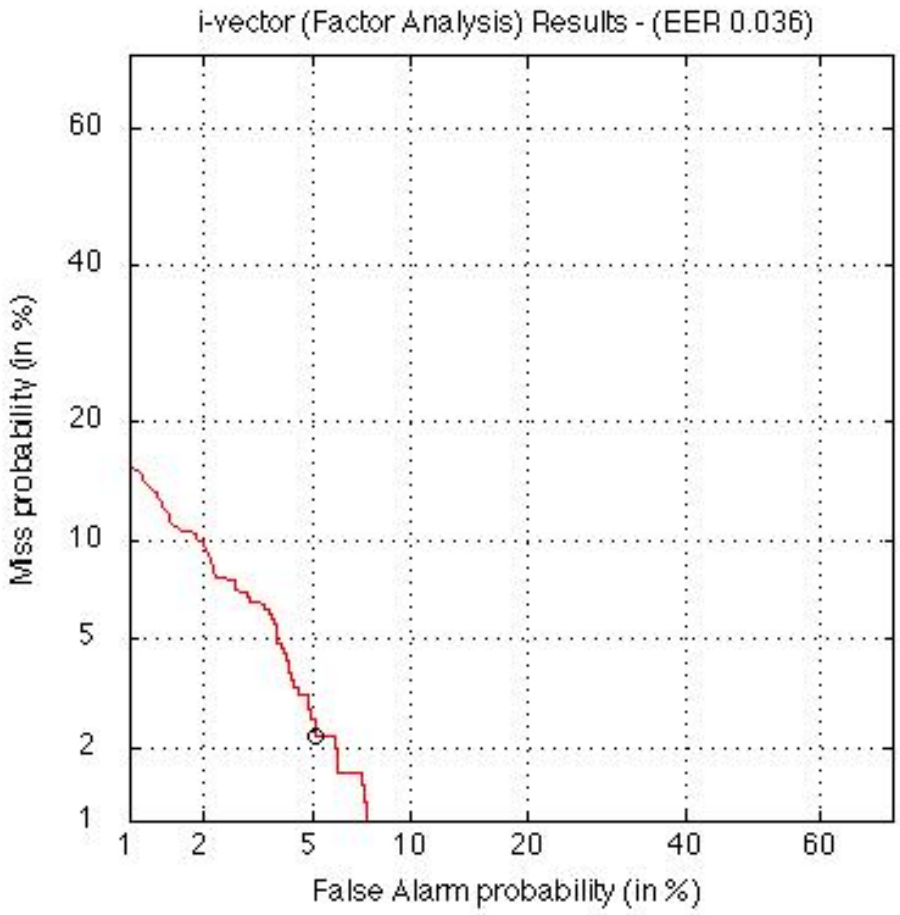

Figure 23: DET of the i-vector (FA) Speaker Models with 12 cepstral coefficients + 1 energy based

With the addition of LDA techniques, the results continue to improve as displayed in Figure 24.

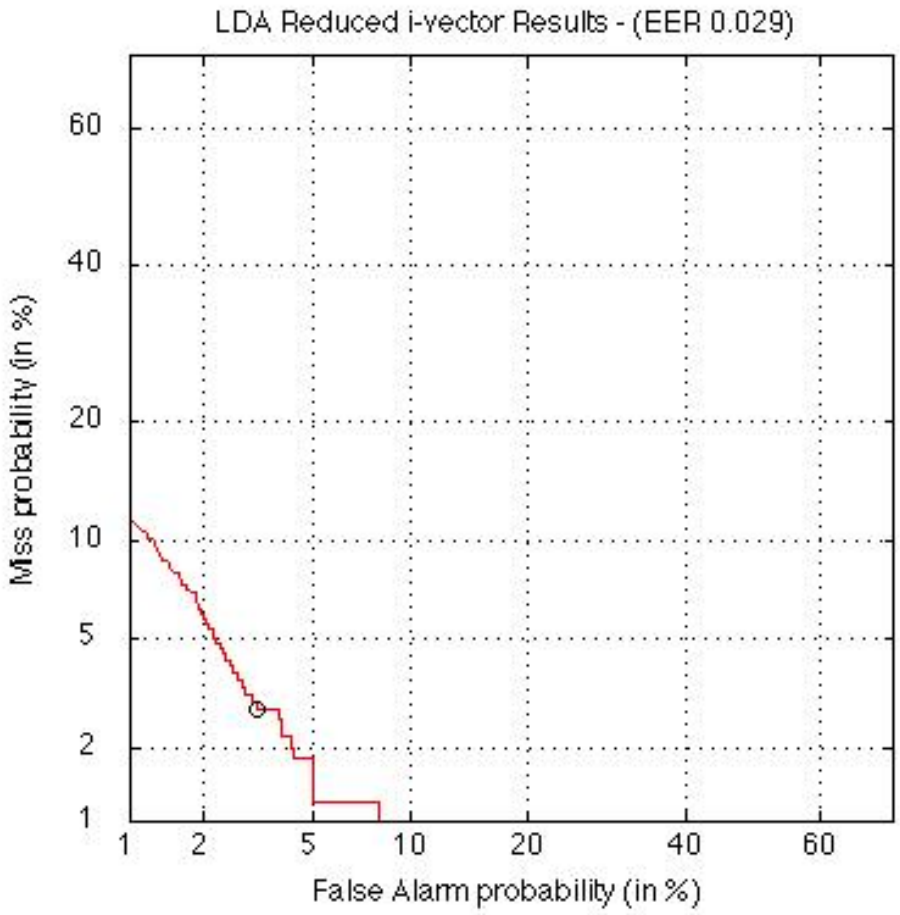

Figure 24: DET of the LDA reduced i-vector Speaker Models with 12 cepstral coefficients + 1 energy based

Several different tests will be conducted during Phase IV. Selection of tests will be chosen from the variety of different conditions made available by the NIST 2010 SRE databases. Smaller scaled testing will be completed in order to minimize the probability of running into difficulties processing the data on a modern desktop computer. If needed, the NIST 2008 SRE database can be used which contains smaller datasets. Larger tests will be used to determine the capabilities of the Matlab tool.

All tests completed on the Matlab code will also be tested on an already vetted speaker recognition system created by researchers at UMD and JHU. Side by side results can be compared and used as a higher level means validation.

7 Project Schedule

Fall 2011

Phase I:	***~(5 weeks)***
Aug. 29 – Sept. 28 ~(4 weeks)	✓ Read a variety of Text-Independent Speaker Identification papers to obtain an understanding of the proposed project
Sept. 28 – Oct. 4 ~(1 week)	✓ Write proposal and prepare for class presentation
Phase II:	***~(4 weeks)***
Oct. 5 – Oct. 21 ~(2 weeks)	✓ Be able to extract MFCCs from speech data and apply simple VAD algorithm ✓ Understand SRE databases
Oct. 22 – Nov. 4 ~(2 weeks)	✓ Develop EM algorithm to trained UBM ✓ Add MAP algorithm to create speaker models ✓ Add likelihood ratio test as a classifier ✓ Validate results using likelihood ratio test as classifier with EER and DET curves, bug fix when necessary
Phase III:	***~(5 weeks)***
Nov. 5 – Dec. 2 ~(3 weeks + Thanksgiving Break)	✓ Create supervectors from GMMs ✓ Write code to train total variability space ✓ Add ability to extract i-vectors from the total variability space ✓ Add cosine distance scoring (CDS) as a classifier ✓ Validate results using the CDS classifier with EER and DET curves, bug fix when necessary
Dec. 3 – Dec. 9 ~(1 week) *overlap*	✓ Prepare Project Progress Report
Dec. 3 – Dec. 19 ~(2 week) *overlap*	✓ Implement LDA on the i-vectors ✓ Validate results using the CDS classifier with EER and DET curves, bug fix when necessary

Spring 2012

Phase IV:	➢ ***~(4 weeks)***
Jan. 25 – Feb. 24 ~(4 weeks)	➢ Obtain familiarity with vetted a speaker recognition system ➢ Test algorithms of Phase II and Phase III on several different conditions and compare against results of vetted system ➢ Bug fix when necessary
Phase V	➢ ***~(7 weeks)***
Feb. 25 – Mar. 2 ~(1 week) *overlap*	➢ Make Decision to either: (1) parallelize/optimize inefficient code, (2) Add more features, or (3) test in various conditions ➢ Read appropriate background material to make decision
Feb. 25 – Mar. 2 ~(1 week) *overlap*	➢ Work on Project Status Presentation
Mar. 3 – Apr. 20 ~(6 weeks + Spring Break)	➢ Update Schedule to reflect decision made in Phase IV ➢ Finish (1) or (2) in a 6 week time period including time for validation and test
Phase VI:	➢ ***~(3 weeks)***

Apr. 21 – May 10 ~(3 weeks)	➢ Create final report and prepare for final presentation

8 Milestones

Fall 2011

October 4	✓ Have a good general understanding on the full project and have proposal completed. Present proposal in class by this date. o *Marks completion of Phase I*
November 4	✓ Validation of system based on supervectors generated by the EM and MAP algorithms o *Marks completion of Phase II*
December 19	✓ Validation of system based on extracted i-vectors ✓ Validation of system based on nuisance-compensated i-vectors from LDA ✓ Mid-Year Project Progress Report completed. Present in class by this date. *Marks completion of Phase III*

Spring 2012

Feb. 25	➢ Testing algorithms from Phase II and Phase III will be completed and compared against results of vetted system. Will be familiar with vetted Speaker Recognition System by this time. o *Marks completion of Phase IV*
March 18	➢ Decision made on next step in project. Schedule updated and present status update in class by this date.
April 20	➢ Completion of all tasks for project. o *Marks completion of Phase V*
May 10	➢ Final Report completed. Present in class by this date. o *Marks completion of Phase VI*

9 Deliverables

A fully validated and complete Matlab implementation of a speaker recognition system will be delivered with at least two classification algorithms.

Both a mid-year progress report and a final report will be delivered which will include validation and test results.

10 Bibliography

[1]*Biometrics.gov - Home*. Web. 02 Oct. 2011. <http://www.biometrics.gov/>.

[2] Kinnunen, Tomi, and Haizhou Li. "An Overview of Text-independent Speaker Recognition: From Features to Supervectors." *Speech Communication* 52.1 (2010): 12-40. Print.

[3] Ellis, Daniel. "An introduction to signal processing for speech." *The Handbook of Phonetic Science,* ed. Hardcastle and Laver, 2nd ed., 2009.

[4] Reynolds, D. "Speaker Verification Using Adapted Gaussian Mixture Models." *Digital Signal Processing* 10.1-3 (2000): 19-41. Print.

[5] Reynolds, Douglas A., and Richard C. Rose. "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models." *IEEE Transations on Speech and Audio Processing* IEEE 3.1 (1995): 72-83. Print.

[6] "Factor Analysis." *Wikipedia, the Free Encyclopedia*. Web. 03 Oct. 2011. <http://en.wikipedia.org/wiki/Factor_analysis>.

[7] Dehak, Najim, and Dehak, Reda. "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification." *Interspeech 2009 Brighton.* 1559-1562.

[8] Kenny, Patrick, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel. "A Study of Interspeaker Variability in Speaker Verification." *IEEE Transactions on Audio, Speech, and Language Processing* 16.5 (2008): 980-88. Print.

[9] Lei, Howard. "Joint Factor Analysis (JFA) and i-vector Tutorial." *ICSI. Web. 02 Oct. 2011.* http://www.icsi.berkeley.edu/Speech/presentations/AFRL_ICSI_visit2_JFA_tutorial_icsitalk.pdf

[10] Kenny, P., G. Boulianne, and P. Dumouchel. "Eigenvoice Modeling with Sparse Training Data." *IEEE Transactions on Speech and Audio Processing* 13.3 (2005): 345-54. Print.

[11] Bishop, Christopher M. "4.1.6 Fisher's Discriminant for Multiple Classes." *Pattern Recognition and Machine Learning*. New York: Springer, 2006. Print.

[12] Ellis, Daniel P. W. *PLP and RASTA (and MFCC, and Inversion) in Matlab*. *PLP and RASTA (and MFCC, and Inversion) in Matlab*. Vers. Ellis05-rastamat. 2005. Web. 1 Oct. 2011. <http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/>.

[13] Tipping, Michael E., and Christopher M. Bishop. "Probabilistic Principal Component Analysis." *Journal of the Royal Statistical Society* B 3 (1999): 611-22. Print.

[14] Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern Classification*. New York: Wiley, 2001. Print.

[15] Shum, Stephen. *Unsupervised Methods for Speaker Diarization*. Thesis. Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 2011. Print.

[16] Garcia-Romero, Daniel., Espy-Wilson, Carol Y. "Joint Factor Analysis for Speaker Recognition Reinterpreted as Signal Coding Using Overcomplete Dictionaries." *Odyssey 2010: The Speaker and Language Recognition Workshop* (2010). Print.

[17] D'Souza, Aaron. "Derivation of Maximum Likelihood Factor Analysis Using EM." *www-clmc.usc.edu/* Web. 21 Oct. 2011. <www-clmc.usc.edu/~cs599_an/factor_analysis.pdf>.