International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 6, June - 2014

Speaker-Dependent and Independent Factors for Human Voice Visualization

Howard Lei
Dept. of Engineering
CSU – East Bay, Hayward, CA

Gopikrishnan Pallipatta
Dept. of Engineering
CSU – East Bay, Hayward, CA

Farnaz Ganjeizadeh
Dept. of Engineering,
CSU – East Bay, Hayward, CA

***Abstract* - This work aims to visualize the variability of recorded human speech based on differences in speaker ethnicity, gender, age and recording environment. Analysis of this mismatch can be done using Gaussian Mixture Models (GMMs), which have historically been used in the field of speaker recognition, trained on Mel-Frequency Cepstral Coefficient (MFCC) feature vectors extracted from the speech waveforms. Certain parameters in the GMM models are extracted and mapped to two dimensions via the Principal Component Analysis (PCA) technique. The two-dimensional values are visualized and compared based on their ethnicity, age and gender. The database used consists of 42 speakers of different ethnicity, gender and age. The analysis showed that speaker voices recorded in the same noisy environment appear more closely on the two-dimensional plot, confirming that the noise level in every recording plays an important role which differentiates speakers.**

Keywords - Speaker Recognition; GMM-UBM; Principal Component Analysis

I. INTRODUCTION

In speaker recognition, having a proper understanding of the causes of speaker variability is important to designing techniques to compensate for such variability. This work attempts to gain an understanding of the causes of speaker variability arising from factors such as speaker age, gender, ethnicity, and recording environment. Each speaker is modeled using a Gaussian Mixture Model (GMM) [1], which has been popular for speaker recognition up until around 2006. Certain parameters of each speaker's GMM are extracted, and reduced to two-dimensions using Principal Component Analysis (PCA) [2] for visualization. The GMMs are trained in the typical fashion using Mel-Frequency Cepstral Coefficient (MFCC) feature vectors [3]. Once two-dimensional data is obtained for each speaker, it can be visualized on a two-dimensional scatter plot.

Given limited knowledge of the data, GMMs can model feature vector distributions that are difficult to precisely characterize, such as feature vectors resulting from speech waveforms. GMM models are trained using the Expectation Maximization (EM) algorithm [4], an iterative algorithm that finds a maximum-likelihood estimate of the model parameters given the feature vectors. The EM algorithm is similar to the K-means clustering algorithm, except that it uses soft clustering assignments. In soft clustering, each feature vector is assigned a likelihood of belonging to each GMM mixture. The mixture means, covariance, and weights are updated based on the likelihoods of its MFCC vectors.

In every speaker recognition system, a Universal Background Model (UBM) is needed to represent the distribution of a general population of speakers. The UBM is a speaker-independent GMM model that can be used for speaker-dependent GMM training, and is itself trained using the EM algorithm given feature vectors from a large number of speakers. We note that although the classical GMM-based approach to speaker recognition is now obsolete, GMM models themselves are still used in the more state-of-the-art i-vector approach [5][6] for purposes of computing sufficient statistics.

The article is structured as followed: Section 2 describes the data used; Section 3 describes the methodology; Section 4 describes the experiments, results, and provides a discussion, and Section 5 provides a conclusion and future work.

II. DATA

The data consists of recorded speech from California State University, East Bay students and faculty. All the speakers were asked to read a paragraph from an engineering textbook, and the same paragraph was read by all speakers. The recorded subjects include 9 females and 33 males, and roughly two minutes of speech were recorded per speaker. The first 21 speakers were recorded in the first data collection phase (Phase I), while the next 21 were recorded in the second phase (Phase II). The two phases used different recording software with different audio recording qualities. The recordings were taken using a Blue Snowball USB microphone with the Omni-directional microphone setting. Phase I recordings were performed using the Audio Recorder software [7], while Phase II recordings were performed using Audacity [8]. The voice data were recorded in different environments such as conference halls, laboratories, classrooms. These environments had different noise levels and were recorded at different timings. Our aim is to understand the factors that impact the GMM speaker verification models, which are suitable for smaller datasets with fewer speakers. Table I below shows the speaker tag, gender, age, ethnicity and location of the recordings.

TABLE I. TABLE OF RECORDED SPEAKERS, INCLUDING AGE, GENDER, ETHNICITY, AND RECORDING LOCATION. "AB" STANDS FOR "AMERICAN BORN".

Speaker Number	Age	Gender	Ethnicity	Location
1	24	Female	Nigerian (Yoruba)	Lab 2
2	22	Male	US English	Lab 1
3	25	Male	Turkish	Lab 2
4	26	Male	Nigerian (Edo)	Lab 2
5	25	Male	Indian	Lab 1
6	26	Male	Chinese	Lab 2
7	57	Male	Caucasian	Lab 2
8	25	Male	Philippines	Lab 2
9	22	Male	Philippines	Lab 2
10	48	Male	Chinese	Lab 1
11	23	Male	German	Lab 1
12	19	Male	AB - Chinese	Lab 1
13	18	Male	AB - mixed	Lab 1
14	52	Male	California	Lab 1
15	30	Male	AB - Chinese	Lab 1
16	29	Female	Taiwanese	Home
17	25	Female	Indian	Lab 2
18	20	Female	Pacific Islander	Lab 1
19	53	Male	Persian (Farsi)	Lab 1
20	24	Male	Sri Lanka + British	Lab 1
21	23	Male	Indian	Lab 1
22	25	Female	Indian	Home
23	44	Male	American	Hotel
24	23	Male	Egyptian	Hotel
25	29	Male	American	Hotel
26	32	Male	Persian	Lab 1
27	30	male	Persian	Lab 1
28	21	Male	Indian	Lab 1
29	22	Male	Indian	Lab 2
30	25	Male	Indian	Lab 2
31	24	Male	Indian	Lab 2
32	27	Male	Indian	Lab 2
33	24	Male	Indian	Lab 1
34	47	Male	American	Conf Room
35	42	Female	American	Conf Room
36	25	Male	Indian	Lab 2
37	53	Female	Chinese	Room
38	53	Male	Middle East	Room
39	50	Female	American	Room
40	21	Female	Indian	Lab 1
41	22	Male	Indian	Lab 1
42	21	Male	Indian	Lab 1

III. METHODOLOGY

To obtain a low-dimensional representation of each speaker for purposes of visualization, GMM models are first trained to model the distribution of the MFCC feature vectors extracted from speech waveforms. MFCC coefficients C0-C12 (a total of 13 dimensions) are used. In addition, the first and second time derivatives of the coefficients of each feature vector dimension are appended to generate vectors of 39 dimensions. The typical feature extraction approach extracts one MFCC feature vector for every 10ms of speech using 25ms windows, such that an entire speech waveform is represented by a sequence of feature vectors. Every minute of speech should hence contain 100 vectors. Each MFCC feature vector dimension is mean- and variance-normalized across the duration of each waveform. The approach involves first training a UBM via the EM algorithm on a set of speech data across a set of 10 speakers (Speaker numbers 1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 42). In our particular implementation of the system, speaker-dependent GMM models are trained using the EM algorithm from each speaker's data, and the UBM is used to initialize the algorithm.

The following equation describes the probably density function (pdf) of a GMM model:

$$p(x;\mu,\Sigma,\omega)=\sum_{m=1}^{M}\omega_m N(x;\mu_m,\Sigma_m) \quad (1)$$

where **x** is a vector, $N(x; \mu_m, \Sigma_m)$ is a pdf of a Gaussian distribution with mean μ_m and covariance matrix Σ_m, and ω_m are the mixture weights. M is the number of Gaussian mixtures.

In our experiments, we used eight mixtures for each GMM (M=8), with full covariance matrices. The number of mixtures is small compared to those used in a typical GMM-UBM system, with 512 to 2,048 mixtures. However, the dataset we are using (1 hour of total speech) is also significantly smaller compared to the typical datasets, and hence fewer mixtures are needed.

Fig. 1 illustrates the GMM training process.

Once the speaker-dependent GMMs are trained, the first 13 dimensions of the GMM mixture means (corresponding to MFCC C0-C12) are extracted for all speakers. The eight mean vectors corresponding to each speaker-dependent GMM are averaged together to form a single 13-dimensional vector that represents each speaker. PCA is then trained on and applied to each of the 13-dimensional vectors.

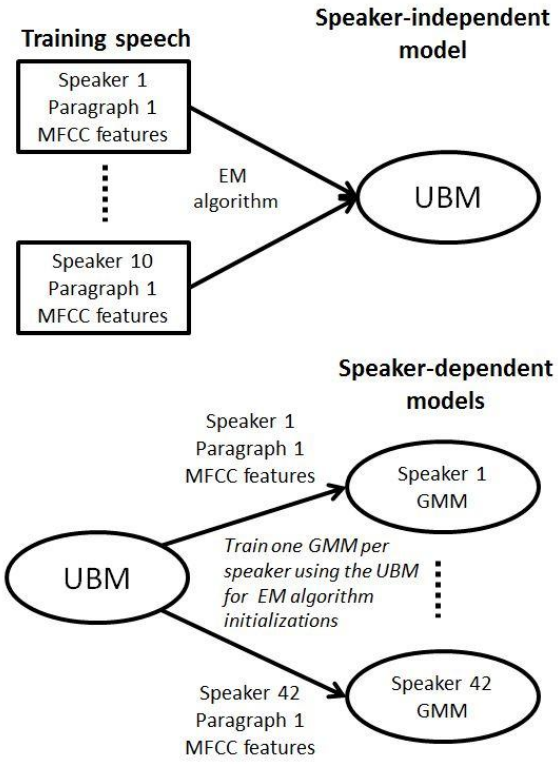

Fig. 1. Training of Speaker-Independent UBM and speaker-dependent GMM models

The process for training and applying the PCA is described as follows: The 13-dimensional vectors are first loaded into a 13x42 matrix (13 dimensions and 42 speakers) **X**. The covariance of the 13-dimensional vectors in the matrix are obtained, forming a 13x13 covariance matrix C. Singular value decomposition (SVD) is applied to the covariance matrix, and the matrices **U**, **S**, and **V*** are found. We note that because the covariance matrix is symmetrical, **U**=**V**. The first two dimensions of the unitary matrix (corresponding to the first two eigenvectors of the matrix decomposition) are used. The product of the two eigenvectors with the original 13x42 matrix is then plotted on a two-dimensional scatter plot for visualization and analysis.

Note that our implementation of the GMM-UBM speaker recognition system along with the data analysis and visualization was done using publically available MATLAB scripts under the BSD license.

In summary, our work consists of following steps:

1. Extract MFCC feature vectors from speaker audio recordings.
2. Train a UBM from the MFCC vectors of a set of 10 speakers.
3. Train speaker-dependent GMMs for each individual speaker using the UBM, along with the MFCC feature vectors for the speaker.
4. Extract the GMM mean vectors for each speaker, and obtain a single averaged mean vector for each speaker.
5. Use PCA to reduce the dimensionality of the averaged mean vectors down to two dimensions.
6. Plot the two-dimensional vectors
7. Analyze the plot and obtain conclusions.

IV. EXPERIMENTS, RESULTS AND DISCUSSIONS

Table II shows the speaker clusters used for our analysis. The clusters are formed based on the categories of data collection phase, gender, age, and ethnicity. Note that the total number of speakers in the clusters of a single category is 42.

Fig. 2 and

Fig. 3 show the two-dimensional scatter plots of each speaker for data collection Phases I and II, respectively. Table III shows the mean and variances of the two-dimensional data points for speakers in each cluster.

Based on the means of each cluster, the Phase I and Phase II clusters, along with the male and female clusters, appear to have the most significant separation. Phase I and Phase II cluster data points are distributed with means of [**0.07**, -0.03] and [**0.21**, -0.06] respectively, and variances of [0.02, 0.01] and [0.01, 0.01] respectively. Male and female clusters have means of [**0.13**, -0.05] and [**0.18**, -0.04] respectively, and variances of [0.02, 0.01] and [0.02, 0.01] respectively. We note that in neither case, the variances are high, which would render the differences in the means to be less significant. The separation is less clear for clusters of different age groups and ethnicities.

Based on the 2D scatter plots, it's apparent that data points that are closely clustered may contains speakers of different genders, ages, and ethnicities. For instance, speakers 10, 14, 18 are closely clustered, but have different ethnicity, ages and gender according to Table I. Based on these observations, clusters that are observed on the scatter plots may not result from the effects of the factors of gender, age, and ethnicity.

TABLE II. SPEAKER CLUSTERS USED, AND THE NUMBER OF SPEAKERS PER CLUSTER.

Speaker Cluster	No. Of Speakers
Based on data collection phase	
Phase I	21
Phase II	21
Based on gender	
Male	33
Female	9
Based on age	
18-22	10
23-39	22
40 and above	10
Based on ethnicity	
American	9
Afro & Euro	3
Asians w/o Indians	16
Indians	14

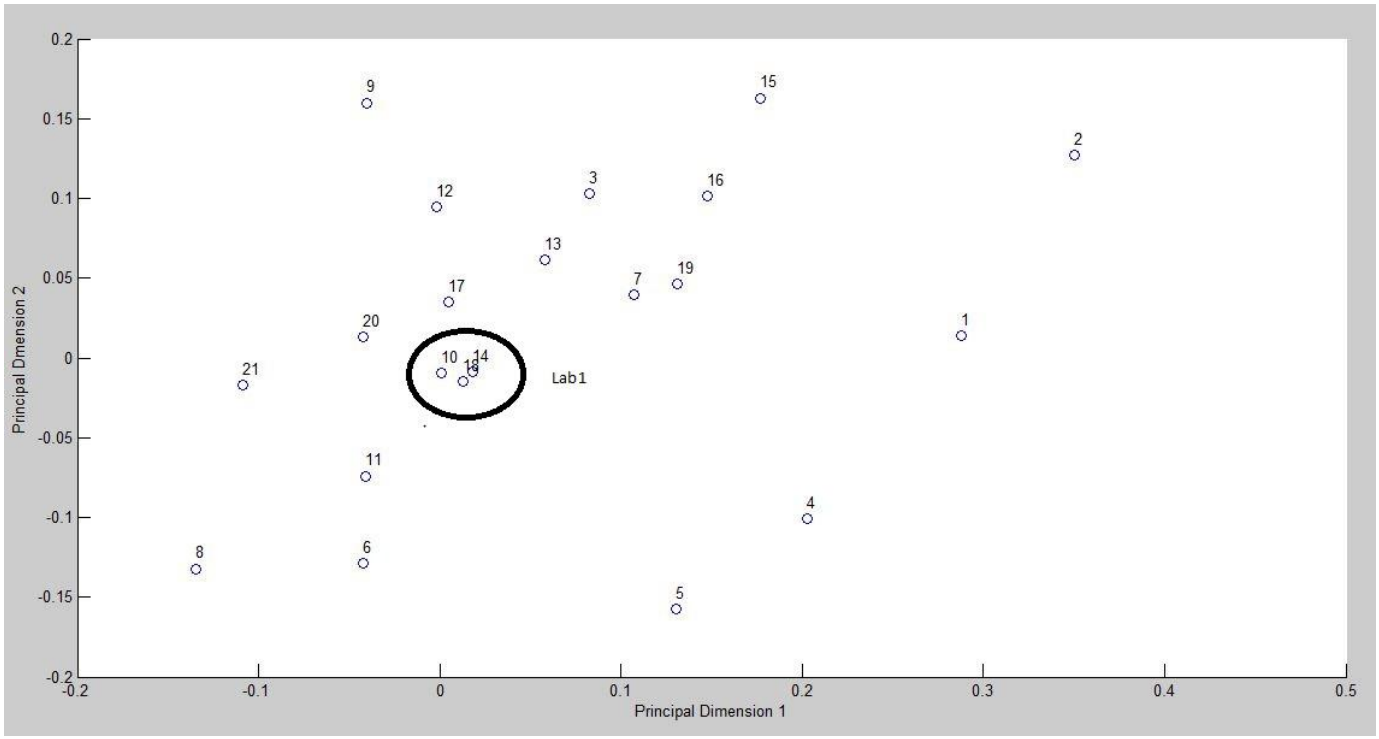

Fig. 2. Plot of speaker numbers 1-21, which were recorded in the first data collection phase.

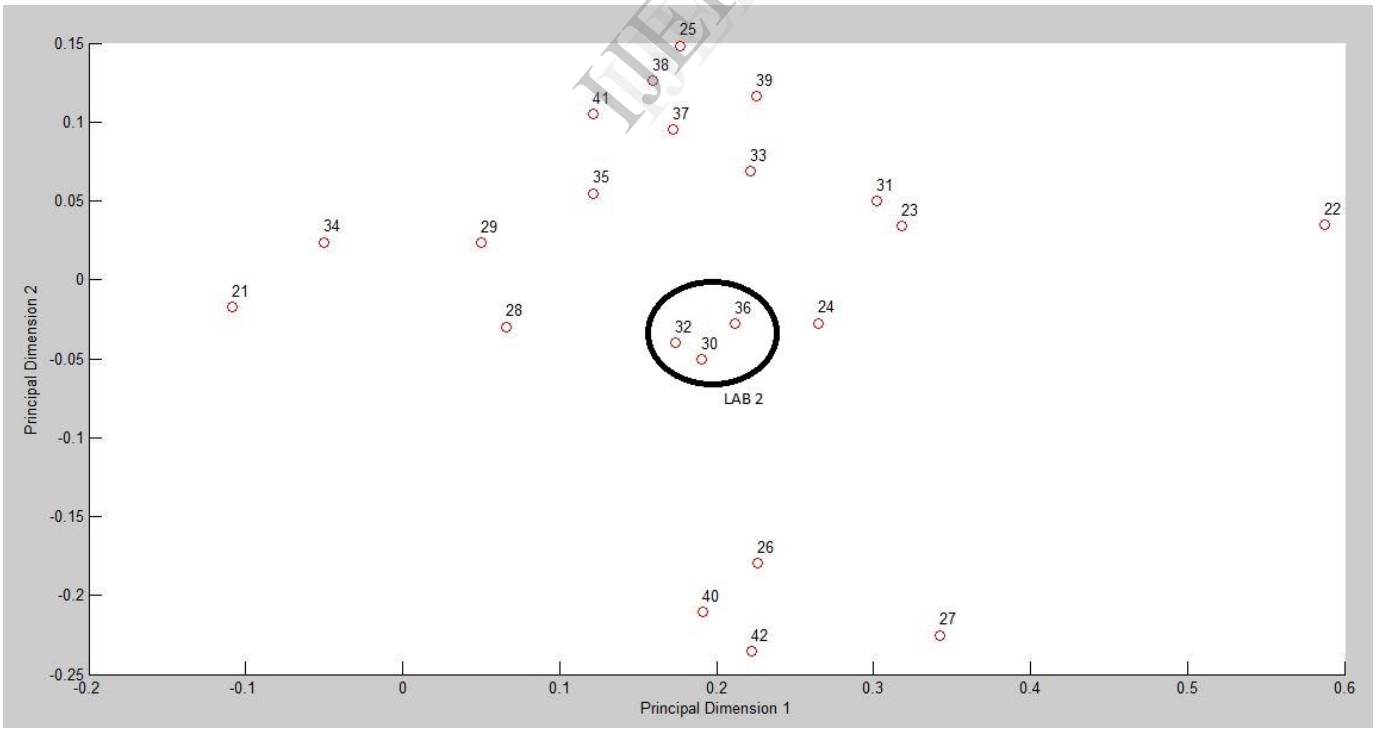

Fig. 3. Plot of speaker numbers 22-42, which were recorded in the second data collection phase.

TABLE III: MEANS & VARIANCES OF 2D DATA POINTS OF SPEAKER CLUSTERS

Speaker Cluster	No. Of Speakers	Means	Variances
Based on data collection phase			
Phase I	21	[0.07,-0.03]	[0.02,0.01]
Phase II	21	[0.21,-0.06]	[0.01,0.01]
Based on gender			
Male	33	[0.13,-0.05]	[0.02,0.01]
Female	9	[0.18,-0.04]	[0.02,0.01]
Based on age			
18-22	10	[0.12,-0.04]	[0.01,0.02]
23-39	22	[0.15,-0.07]	[0.03,0.01]
40 and above	10	[0.14,0.01]	[0.01,0.01]
Based on ethnicity			
American	9	[0.15,0.00]	[0.02,0.00]
Afro & Euro	3	[0.11,-0.11]	[0.03,0.00]
Asians w/o Indians	16	[0.11,-0.03]	[0.18,0.01]
Indians	14	[0.16,-0.08]	[0.02,0.12]

The recording quality and environment appeared to contribute more to the clustering of the data. The Phase I and Phase II data, which were recorded under different recording environments and with different recording software, have the most data separation. An observation of the plots shows that for several instances, speakers recorded in the same location lie close to each other. Examples include speakers 20 and 21 (Lab1 Location), speakers 23 and 24 (Hotel Location), speakers 38 and 39 (Room Location), and speakers 30 and 32 (Lab2 Location). Lastly, it's interesting to note that speakers 8, 9, 21, 22, 27, 40, 42 lie at the edges of the plots, and their .wav file playback sound clearer than .wav files from most other speakers.

The analysis suggests that factors such as recording environment and recording quality, which affect the recording noise levels, can have a large impact in determining where the speakers lie on the scatter plots. This impact can overshadow impacts from factors such as speaker gender, age, and ethnicity.

V. SUMMARY AND FUTURE WORK

After analyzing the scatter plots of the two-dimensional data points for every speaker in our dataset, the following points can be made. Ethnicity, accent, age and gender do not seem to play vital role in the differentiation of speakers in our recorded database. The noise level in every recording is different and in a few of them, they are high. The more clear the sound, the more the speaker seems to lay distinguished from other speakers. From the analysis of the plots, we can conclude that the location of the recording and the noise associated with the location are the main factors that determine where the speaker appears on the 2D plot. The noise levels at the locations make the speakers in those locations similar according to the 2D plots.

Future work will attempt to explain in greater detail why gender, ethnicity, and age between the speakers did not affect the speaker scatter as much as the recording environments themselves did. We hope to also expand our dataset to contain more speakers from the same environment so that by controlling the environment, differences in the scatter plot may be more attributable to gender, ethnicity, and age. Different speaker modeling and visualization techniques will be explored as well. Lastly, the distance from the recording microphone to the speaker (i.e. near-field vs. far-field microphone recordings) is another factor we plan to examine.

REFERENCES

[1] D.A. Reynolds, T.F. Quatieri, and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," in *Digital Signal Processing*, Vol. 10 No. 3, 2000, pp. 19–41.
[2] Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4.
[3] S. Davis and P. Mermelstein, "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences", in Proceedings of *ICASSP*, 1980.
[4] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Series B*, Vol. 39(1), 1977, pp. 1-38.
[5] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification", in Proceedings of *Interspeech*, Brighton, UK, 2009, pp. 1559-1562.
[6] L. Burget, P. Oldrich, C. Sandro, G. Oldrej, M. Pavel, and N. Brummer, "Discriminantly Trained Probabilistic Linear Discriminant Analysis for Speaker Verification", in Proceedings of ICASSP, Brno, Czech Republic, 2011.
[7] Audio Recorder, http://mac.softpedia.com/get/Audio/Audio-Recorder.shtml
[8] Audacity software, http://audacity.sourceforge.net/