GENERATIVE MODEL-BASED SPEAKER CLUSTERING VIA MIXTURE OF VON MISES-FISHER DISTRIBUTIONS

Hao Tang[1], *Stephen M. Chu*[2], *Thomas S. Huang*[1]

[1]Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, I.L. 61801, USA
[2]IBM T. J. Watson Research Center, Yorktown Heights, N.Y. 10598, USA

ABSTRACT

This paper proposes a generative model-based speaker clustering algorithm in the *maximum a posteriori* adapted Gaussian mixture model (GMM) mean supervector space. The algorithm can be viewed as an extension of the standard expectation maximization algorithm for fitting a mixture model to the data, which iterates between two steps - a sample re-assignment step (E-step) and a model re-estimation step (M-step) - until it converges. The directional scattering patterns of GMM mean supervectors suggest that we employ a mixture of von Mises-Fisher distributions in the model re-estimation step. In the sample re-assignment step, four sample-to-mixture assignment strategies, namely soft, hard, stochastic, and deterministic annealing assignments, are used. Our experiments on the GALE Mandarin dataset show that the use of a mixture of von Mises-Fisher distributions as the underlying model yields significantly higher speaker clustering accuracies than the use of a mixture of Gaussian distributions. It is further shown that deterministic annealing assignment outperforms soft assignment, that soft assignment is comparable to stochastic assignment, and that both soft and stochastic assignments outperform hard assignment.

Index Terms— Model-based clustering, GMM mean supervectors, mixture of von Mises-Fisher distributions, EM algorithm.

1. INTRODUCTION

Speaker clustering is the process of assigning every utterance in a dataset to its corresponding speakers and is useful in a variety of areas [1, 2, 3, 4, 5, 6, 7]. First, speaker clustering is a key component of speaker diarization (speaker segmentation and clustering) [8, 9, 10]. Second, speaker clustering is a prerequisite for unsupervised speaker adaptation in automatic speech recognition whose purpose is to improve the speech recognition accuracies [11]. Third, the central ideas and techniques for speaker clustering, for instance, those on speaker representations and modeling, can be borrowed or used by other tasks such as speaker recognition (identification and verification) [12].

Current state-of-the-art methods for speaker recognition, diarization, and clustering are primarily based on low-level acoustic features such as mel-frequency cepstral coefficients (MFCCs) [13] or perceptual linear prediction (PLP) coefficients [14]. These acoustic features are often modeled by a Gaussian distribution or a mixture of Gaussian distributions, and a statistical distance metric such as the log likelihood ratio is used for decision making [15]. Recently, the vector space method [16] has begun to gain popularity in speaker clustering. In this method, the acoustic features of an utterance is first fitted by a Gaussian mixture model (GMM) [15], and the mixture means of the fitted GMM are stacked to form a long column vector called a GMM mean supervector [16]. An utterance is thus represented as a single data point in the GMM mean supervector space, and speaker clustering can be deemed as data point clustering in this vector space in a very general sense.

Traditional clustering techniques such as k-means and hierarchical clustering do not assume any underlying probabilistic model of the data and thus are popular for their simplicity. However, when the underlying probabilistic model that governs the data generation process is known, or the data distribution can be described by some probabilistic model reasonably well, these non-model-based clustering techniques would not be able to take advantage of such benefits. In this paper, we propose a generative model-based speaker clustering algorithm in the *maximum a posteriori* (MAP) adapted GMM mean supervector space. Specifically, the acoustic features of all utterances from all speakers in a dataset are first used to training a single GMM, known as the universal background model (UBM) [17]. Then, the UBM is adapted to every utterance in the dataset via MAP adaptation [17] to yield for each utterance a GMM. Next, the mixture means of every GMM are stacked to form a GMM mean supervector [16]. These GMM mean supervectors form the GMM mean supervector space. Finally, in this space, the speaker clustering algorithm iterates between two steps - a sample re-assignment step and a model re-estimation step - until it converges. This algorithm can be viewed as an extension of the standard expectation maximization (EM) algorithm [18] where the E-step corresponds to the sample re-assignment step and the M-step the model re-estimation step. The directional scattering patterns of GMM mean supervectors suggest that we employ a mixture of von Mises-Fisher distributions [19] (analogous to a mixture of Gaussian distributions on the unit hypersphere) in the model re-estimation step. In the sample re-assignment step, four sample-to-mixture assignment strategies, namely soft, hard, stochastic, and deterministic annealing assignments, are used.

We compare the use of a mixture of von Mises-Fisher distributions and a mixture of Gaussian distributions as the underlying models of the algorithm in the model re-estimation step. Our experiments on the GALE Mandarin dataset [20] clearly indicate that the use of a mixture of von Mises-Fisher distributions yields significantly higher speaker clustering accuracies than the use of a mixture of Gaussian distributions. This result is consistent with our observation that the GMM mean supervectors show strong directional scattering patterns in a high dimensional space. That is, in the GMM mean supervector space, the directions of the vectors are more informative and indicative than the magnitudes of the vectors. Our experiment results also indicate that for the same underlying model used, the deterministic annealing assignment strategy outperforms the the soft assign-

This work was funded in part by DARPA contract HR0011-06-2-0001.

ment strategy, that the soft assignment strategy is comparable to the stochastic assignment strategy, and that both the soft and stochastic assignment strategies outperform the hard assignment strategy.

2. GMM MEAN SUPERVECTOR SPACE

The GMM mean supervector is a very effective speaker or utterance representation, and has been widely applied to speaker recognition [16]. The acoustic features of an utterance are first used to adapt a UBM, finely trained over all utterances from all speakers in a dataset, to produce an utterance-specific GMM. The mixture means of the GMM are then stacked to form a long column vector - the GMM mean supervector. It is empirically found that the GMM mean supervectors which belong to the same speaker tend to gather closer than those belonging to different speakers in a high dimensional space. This observation is the rudimental motivation of the use of GMM mean supervectors as the speaker or utterance representation for speaker recognition, diarization, and clustering.

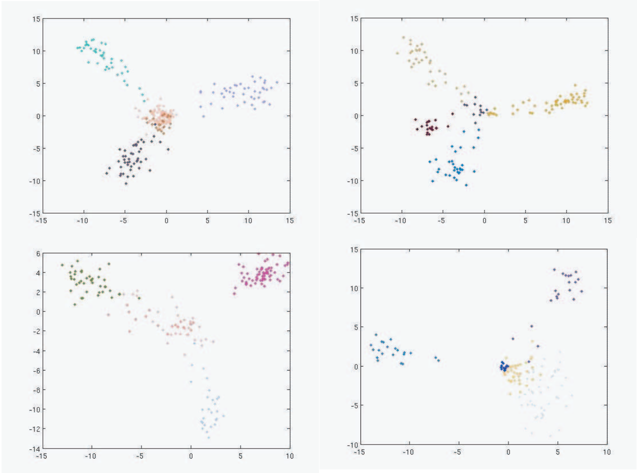

Fig. 1. Scatter plots of GMM mean supervectors projected onto first two principal components.

Fig. 1 shows several scatter plots of the GMM mean supervectors projected onto the first two principal components by principal component analysis (PCA) [21]. In each plot, there are 5 speakers, indicated by different colors, and about 150 utterances, shown as small dots. As we can see, the data points belonging to the same speaker tend to cluster together. Thus, the Euclidean distance metric is a reasonable choice for clustering in the GMM mean supervector space. However, we can also observe that the data points show strong directional scattering patterns. The directions of the data points seem to be more informative and indicative than their magnitudes. This observation motivated us to favor the cosine distance metric over the Euclidean distance metric for clustering the GMM mean supervector space. In fact, our separate experiments prove that the cosine metric consistently outperforms the Euclidean metric in the GMM mean supervector space while using a non-model-based clustering technique such as k-means and hierarchical clustering.

3. GENERATIVE MODEL-BASED CLUSTERING

Generative model-based clustering assumes that there is an underlying probabilistic model which governs the process that generated the data. Usually, a mixture model, such as a mixture of Gaussians, is assumed. This is because that a mixture model can not only approximate arbitrarily complex probability density functions but also is a natural representation of many kinds of observations - those observations generated by a randomly selected source from a set of alternative sources, which are very common in practice. Generative model-based clustering has been applied to document clustering [22]

Generative model-based clustering attempts to fit a mixture model of a specified type to the data to be clustered. The number of mixtures in the model is made equal to the number of clusters. Suppose the mixture model is given by

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^{K} \alpha_k p_k(\mathbf{x}|\lambda_k) \tag{1}$$

where $p_k(\mathbf{x}|\lambda_k)$ is the probability density function of the k^{th} mixture and the λ's denote the model parameters. To determine which mixture (or cluster) among the K mixtures (or clusters) a data point $\mathbf{x}$ belongs to, we compute the posterior probability

$$p(k|\mathbf{x}, \lambda) = \frac{\alpha_k p_k(\mathbf{x}|\lambda_k)}{\sum_{j=1}^{K} \alpha_j p_j(\mathbf{x}|\lambda_j)} \tag{2}$$

where α_k's are the prior probabilities that $\mathbf{x}$ belongs to the k^{th} mixture. The cluster k that yields the highest value of $p(k|\mathbf{x}, \lambda)$ is chosen to be the cluster to which the data point $\mathbf{x}$ belongs.

Given a set of data points X, the standard EM algorithm [18] is used to perform a maximum likelihood estimation of the model parameters. The EM algorithm aims to find a local maximum of the data log likelihood function

$$\log p(X|\lambda) = \sum_{\mathbf{x}\in X} \log(\sum_{k=1}^{K} \alpha_k p_k(\mathbf{x}|\lambda_k)) \tag{3}$$

It is known that the EM algorithm amounts to iterating between an E-step and an M-step until convergence is achieved. In the E-step, based on a prior model λ, the posterior probability that a data point $\mathbf{x}$ belongs to a mixture k (Equation 2) is computed for every $\mathbf{x} \in X$ and every $k = 1, 2, ..., K$. In the M-step, a set of new model parameters λ' are updated by

$$\lambda'_k = arg \max_{\lambda} \sum_{\mathbf{x}\in X} p(k|\mathbf{x}, \lambda) \log p(\mathbf{x}|\lambda) \tag{4}$$

and

$$\alpha'_k = \frac{1}{N} \sum_{\mathbf{x}\in X} p(k|\mathbf{x}, \lambda) \tag{5}$$

where N is the number of data points.

The E-step of the EM algorithm can be interpreted as a sample re-assignment process. That is, each data point is "soft" assigned to every mixture, with a percentage to indicate the degree of ownership of the mixture to the data point. The degree of ownership is determined by the posterior probability given by Equation 2. In this paper, for the purpose of speaker clustering, we extend the standard EM algorithm by adopting, in addition to the soft re-assignment strategy, three other sample re-assignment strategies. These three sample re-assignment strategies are hard assignment, stochastic assignment, and deterministic annealing assignment, and will be explained in detail in Section 5.

4. MIXTURE OF VON MISES-FISHER DISTRIBUTIONS

The von Mises-Fisher distribution (vMF) [19] is a close analogue to the Gaussian distribution for directional data - data that spread on the unit hypersphere. The von Mises-Fisher distribution of a d-dimensional random vector $\mathbf{x}$ can be written as

$$p(\mathbf{x}|\lambda_k) = c_d(\kappa_k)e^{\kappa_k \mu_k^T \mathbf{x}} \tag{6}$$

where $\mathbf{x}$ and μ_k are unit vectors, $\kappa_k \geq 0$ and $d \geq 2$. The distribution is characterized by the mean μ_k and the concentration parameter κ_k. The normalizing constant $c_d(\kappa)$ is given by

$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \tag{7}$$

where $I_r(\cdot)$ is the r^{th} order modified Bessel function of the first kind.

The von Mises-Fisher distribution is one of the simplest parametric distribution for directional data. In this paper, we consider the use of a mixture of von Mises-Fisher distributions to serve as the underlying generative model for speaker clustering in the MAP adapted GMM mean supervector space, where we have observed that the data points in this space are strongly directional. Banerjee et. al. [19] have derive the update equations for estimating the mixture of von Mises-Fisher distributions from a dataset using the EM framework, which are given as follows.

$$\alpha_k = \frac{1}{N}\sum_{\mathbf{x}\in X} p(k|\mathbf{x}, \lambda) \tag{8}$$

$$\mathbf{r}_k = \sum_{\mathbf{x}\in X} \mathbf{x} p(k|\mathbf{x}, \lambda) \tag{9}$$

$$\mu_k = \frac{\mathbf{r}_k}{||\mathbf{r}_k||} \tag{10}$$

$$\frac{I_{d/2}(\kappa_k)}{I_{d/2-1}(\kappa_k)} = \frac{||\mathbf{r}_k||}{\sum_{\mathbf{x}\in X} p(k|\mathbf{x}, \lambda)} \tag{11}$$

5. SAMPLE RE-ASSIGNMENT STRATEGIES

5.1. Hard assignment

The hard assignment strategy is a winner-takes-all heuristic. That is, a data point is assigned to one and only one mixture which has the highest posterior probability. It is implemented by simply replacing the posterior probability of Equation 2 with

$$p(k|\mathbf{x}, \lambda) = \left\{ \begin{array}{ll} 1, & \text{if } k = arg\max_{k'} p(k'|\mathbf{x}, \lambda); \\ 0, & \text{otherwise.} \end{array} \right. \tag{12}$$

where $p(k'|\mathbf{x}, \lambda)$ is given by Equation 2.

5.2. Stochastic assignment

The stochastic assignment strategy can be deemed as a Monte Carlo version of the soft assignment strategy. According to the posterior probability of Equation 2, the data points are stochastically distributed into the K mixtures. More precisely, for a specific data point $\mathbf{x}$, let $\gamma = \{p(k|\mathbf{x}, \lambda)\}_{k=1}^K$ be the parameters of a multinomial distribution $Multi(\gamma)$, and a random draw of k from $Multi(\gamma)$ determines the cluster to which $\mathbf{x}$ is assigned.

5.3. Deterministic annealing assignment

The objective of the standard EM algorithm is to estimate the model parameters so as to maximize the data log likelihood function given by Equation 3. Note Equation 3 is not convex and one of the most significant problems that the standard EM algorithm faces is the local maximum problem [21, 23]. To overcome this problem, the objective function of the EM algorithm is reformulated to form its deterministic annealing version, which aims to find a local maximum of the expected data log likelihood with entropy constraints [24]

$$L = \sum_{\mathbf{x}} p(\mathbf{x}) \sum_{k=1}^K p(k|\mathbf{x}) \log p(\mathbf{x}|\lambda_k) - T \cdot I(X;Y) \tag{13}$$

where $I(X;Y) = H(Y|X) - H(Y)$ is the mutual information of X (data points) and Y (cluster indices), and T is a temperature parameter which gradually decreases during the E-M iterations. During the E-step, the posterior probability is optimized for each T, and it can be shown that the posterior probability becomes

$$p(k|\mathbf{x}, \lambda) = \frac{\alpha_k^{\frac{1}{T}} p_k(\mathbf{x}|\lambda_k)^{\frac{1}{T}}}{\sum_{j=1}^K \alpha_j^{\frac{1}{T}} p_j(\mathbf{x}|\lambda_j)^{\frac{1}{T}}} \tag{14}$$

In the deterministic annealing EM algorithm, the M-step is the same as that of the standard EM algorithm.

6. EXPERIMENTS

We implement the above generative model-based speaker clustering algorithms and conduct experiments on the GALE Mandarin dataset [20]. The GALE Mandarin dataset contains 1900 hours of broadcasting news speech data collected from various TV programs. The waveforms are sampled at 16K Hz and quantized at 16 bits per sample. We select a subset of the GALE Mandarin dataset that contains 603 speakers and 19023 utterances to form our test set. On average, each speaker in the test set contains about 30 utterances, and each utterance is about 3-4 seconds long.

The basic acoustic features are the 13 dimensional PLP coefficients [14], which are extracted using a hamming window of 25ms at a rate of 10ms per frame. In stead of computing the delta and delta-delta features we augment, for each frame, the basic PLP features with the PLP features of the neighboring frames. That is, the PLP features of the current frame and those of the two frames to the left and those of the two frames to the right are concatenated to form a $13 \times 9 = 117$ dimensional long PLP feature vector.

In addition to the test set, an independent data set is selected from the GALE Mandarin dataset in a way that there are no overlapping speakers between the test set and this independent data set. The selected data set contains 498 speakers and 18324 utterances. On average, each speaker contains about 30-40 utterances, and each utterance is about 3-4 seconds long. This independent data set is used to learn a speaker-discriminative feature transformation. Specifically, in the augmented PLP feature vector space, linear discriminant analysis (LDA) [21, 23] is performed on this data set based on the speaker labels to obtain a speaker-discriminative feature transformation. All the augmented feature vectors are then mapped by this feature transformation to a low-dimensional speaker-discriminative feature space. The dimensionality of the low-dimensional feature space is 40 in our experiments.

The independent data set is also used to train a UBM via the EM algorithm, and a GMM mean supervector is obtained for every utterance in the test set via MAP adaptation. The rest of the experiments exclusively deal with the GMM mean supervectors of all utterances in the test set. We conduct four sets of experiments, each involving the use of a different sample re-assignment strategy. For comparison, we conduct another four sets of experiments, using a mixture of Gaussian distributions as the underlying model instead of a mixture of von Mises-Fisher distributions. We divide each set of experiments into 4 cases each associated with a different number of test speakers, namely, 2, 5, 10, and 20, respectively. For each case, this specific number of speakers are drawn randomly from the test set, and all the utterances from the selected speakers are used in the experiments. For each case, 100 trials are run, each of which involves a random draw of the test speakers (the speakers in different trails are different!), and the average of the clustering accuracies (the number of correctly clustered utterances over the total number of utterances) are reported. The experiment results are reported in Table 1.

Table 1. *Average of speaker clustering accuracies over 100 trials for various cases with different numbers of test speakers (%).*

# test speakers	2	5	10	20
vMF (soft)	94.4	87.5	81.2	76.7
vMF (hard)	80.2	77.9	71.3	68.1
vMF (stoch)	93.7	86.6	82.4	77.1
vMF (da)	**96.2**	**88.7**	**85.3**	**79.6**

# test speakers	2	5	10	20
Gaussian (soft)	82.7	76.5	70.4	67.0
Gaussian (hard)	73.3	65.4	58.9	55.4
Gaussian (stoch)	84.0	77.1	69.9	67.4
Gaussian (da)	84.3	78.1	73.6	68.8

Our experiment results clearly show that the use of a mixture of von Mises-Fisher distributions as the underlying model yields significantly higher speaker clustering accuracies than the use of a mixture of Gaussian distributions. This is true for all cases with different numbers of test speakers and different sample-to-mixture re-assignment strategies. Our experiment results also tell that for a particular underlying model (mixture of either vMFs or Gaussians), the deterministic annealing assignment strategy outperforms the soft assignment strategy, the soft assignment strategy is comparable to the stochastic assignment strategy, and both the soft and stochastic assignment strategies outperform the hard assignment strategies.

7. CONCLUSION

This paper proposes a generative model-based speaker clustering algorithm in the MAP adapted GMM mean supervector space. The algorithm iterates between two steps, namely a sample re-assignment step and a model re-estimation step, until convergence. It can be viewed as an extension of the standard EM algorithm with four different types of sample-to-mixture re-assignment strategies in the E-step. The directional scattering patterns of GMM mean supervectors suggest that we employ a mixture of von Mises-Fisher distributions in the model re-estimation step. Our speak clustering experiments on the GALE Mandarin dataset show that the use of a mixture of von Mises-Fisher distributions as the underlying model yields significant better results than the use of a mixture of Gaussian distributions. In addition, it is also shown that the deterministic annealing assignment strategy outperforms the the soft assignment strategy, that the soft assignment strategy is comparable to the stochastic assignment strategy, and that both the soft and stochastic assignment strategies outperform the hard assignment strategy.

8. REFERENCES

[1] H. Jin, F. Kubala, and R. Schwartz, “Automatic speaker clustering,” Proc. DARPA Speech Recognition Workshop’97.

[2] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, “Clustering speakers by their voices,” Proc. ICASSP’98.

[3] D. Reynolds, E. Singer, B. Carson, G. O’Leary, J. McLaughlin, and M. Zissman, “Blind clustering of speech utterances based on speaker and language characteristics,” Proc. ICSLP’98.

[4] S. Chen, and P. Gopalakrishnan, “Clustering via the Bayesian information criterion with applications in speech recognition,” Proc. ICASSP’98.

[5] W. Tsai, S. Cheng, and H. Wang, “Speaker clustering of speech utterances using a voice characteristic reference space,” Proc. ICSLP’04.

[6] R. Faltlhauser, and G. Ruske, “Robust speaker clustering in eigenspace,” Proc. ASRU’01.

[7] I. Lapidot, H. Guterman, and A. Cohen, “Unsupervised speaker recognition based on competition between self organizing maps,” IEEE TNN, 13(4):877-887, 2002.

[8] S. Tranter and D. Reynolds, “An Overview of Automatic Speaker Diarization Systems,” TASLP, 14(5):1557-565, 2006.

[9] Barras, C., Zhu, X., Meignier, S. and Gauvain, J.L., “Multistage speaker diarization of broadcast news,” IEEE Trans. ASLP. vol. 14 no. 5, pp. 1505-1512, Sept. 2006.

[10] C. Wooters and M. Huijbregts, “The ICSI RT07s Speaker Diarization System,” Lecture Notes in Computer Science, 2007.

[11] S. Yamade, A. Baba, S. Yoshikawa, A. Lee, H. Saruwatari, K. Shikano, “Unsupervised speaker adaptation for robust speech recognition in real environments,” Electronics and Communications in Japan, Part 2, Vol. 88, No. 8, 2005.

[12] J. P. Campbell, “Speaker Recognition: A Tutorial,” Proceedings of the IEEE, 85(9), September 1997, 1437 - 1462.

[13] L. R. Rabiner, B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[14] H. Hermansky, “Perceptual linear predictive PLP analysis for speech,” JASA, 87(4):1738–1752, 1990.

[15] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models”, Speech Communication, Vol. 17, Issues 1-2, pp. 91-108, August 1995.

[16] Campbell, W.M., Sturim, D.E., Reynolds, D.A., “Support vector machines using GMM supervectors for speaker verification,” Signal Processing Letters 13(5), 308-311, 2006.

[17] 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, 2000.

[18] Arthur Dempster, Nan Laird, and Donald Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. of Royal Stat. Society, B, 39(1):1C38, 1977.

[19] Banerjee, A., I. Dhillon, J. Ghosh and S. Sra, “Clustering on the Unit Hypersphere using von Mises-Fisher Distributions,” J. of Machine Learning Research 6, 1345-1382, Sept. 2005.

[20] Stephen M Chu, Hong-Kwang Kuo, Lidia Mangu, Yi Liu, Yong Qin, and Qin Shi, “Recent advances in the IBM GALE mandarin transcription system,” Proc. ICASSP’08.

[21] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. JohnWiley & Sons, Inc., 2nd edition, 2001.

[22] Zhong, S. and Ghosh, J., “Generative model-based document clustering: a comparative study,” KIS, 8(3): 374-384, 2005.

[23] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

[24] Rose, K, Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, roceedings of the IEEE, 86, 2210-2239, 1998.