Speaker Clustering Based on Non-negative Matrix Factorization

Masafumi Nishida, Seiichi Yamamoto

Department of Information System Design, Doshisha University, Kyoto 610-0321, Japan

{mnishida,seyamamo}@mail.doshisha.ac.jp

Abstract

This paper addresses unsupervised speaker clustering for multi-party conversations. Hierarchical clustering methods were mainly used in previous studies. However, these methods require many processes, such as distance calculation and cluster merging, when there are many utterances in conversation data. We propose a clustering method based on non-negative matrix factorization. The proposed method can perform fast and robust clustering by decomposing a matrix consisting of distances between models. We conducted speaker clustering experiments using a Bayesian information criterion based method, a method based on the likelihood ratio between Gaussian mixture models, and the proposed method. Experimental results showed that the proposed method achieves higher clustering accuracy than these conventional methods.

Index Terms: unsupervised speaker clustering, non-negative matrix factorization, agglomerative hierarchical clustering, multi-party conversation

1. Introduction

Speaker clustering is a technique for clustering utterances from the same speaker, and is useful for retrieving the utterances of a specific speaker and for improving automatic speech recognition performance based on speaker adaptation of the acoustic model. Speaker clustering has been studied mainly for broadcast news audio, recorded meetings, and telephone conversations [1].

Typically, agglomerative hierarchical clustering (AHC) has been used for speaker clustering. The Bayesian information criterion (BIC) is used as a stopping criterion of clustering in which clusters are represented by parametric probability densities such as a single Gaussian with full covariance [2]. The incremental Gaussian mixture cluster modeling method has been proposed as a BIC-based inter-cluster distance measure within the framework of AHC [3]. This method gradually increments the complexity of cluster models from single Gaussian distributions to Gaussian mixture models (GMMs) with multiple mixture components as more data become available for better and more dynamic representation of clusters during AHC. The phonetic subspace mixture (PSM) model has been proposed to replace the single Gaussian model in the BIC distance measure [4]. The acoustic feature space is divided into a set of phonetic subspaces according to the corresponding phonetic content with this model.

The generalized likelihood ratio (GLR) has been widely adopted as an inter-cluster distance measure. However, it tends to be affected by the size of the clusters considered, which could result in erroneous selection of the cluster pair to be merged during AHC. To solve this problem, a combination method of GLR and information change rate (ICR) has been proposed as the inter-cluster distance measure [5]. Nishida and Kawahara have proposed a framework in which an optimal speaker model (GMM or Vector Quantization (VQ)) is selected based on the BIC, which reflects the amount of speech data [6]. This framework makes it possible to use a discrete model when the training data is sparse and to seamlessly switch to a continuous model after sufficient data is obtained. Bozonnet et al. have proposed an integrated speaker diarization system, which harnesses the benefits of both top-down and bottom-up approaches through their fusion at the heart of the clustering and segmentation stage [7].

In AHC, it is necessary to compute distances between clusters and incrementally merge the closest pair of clusters. Such methods must perform many processes when there is a large amount of utterances in speech data. We propose a speaker clustering method based on non-negative matrix factorization (NMF). Non-negativity is a useful constraint for matrix factorization that can learn a part of the representation of the data [8]. The non-negative basis vectors that are learned are used in distributed, yet still sparse combinations to generate expressiveness in the reconstructions. The proposed method can achieve robust speaker clustering only by decomposing a matrix consisting of distances between utterances. We conducted speaker clustering experiments using the Corpus of Spontaneous Japanese (CSJ) [9] to prepare a variety of test sets that have a different number of speakers and utterances to demonstrate the robustness of the proposed method and compare it to conventional methods such as BIC-based and GMM-based methods. Iso has also used the CSJ database in speaker clustering experiments [10].

The paper is organized as follows. Section 2 introduces non-negative matrix factorization, Section 3 describes the proposed speaker clustering method based on non-negative matrix factorization, Section 4 presents the specifications of the evaluation data and the experimental results of speaker clustering, and Section 5 concludes the paper.

2. Non-negative matrix factorization

Non-negative matrix factorization can be applied to the statistical analysis of multivariate data. Given a set of multivariate n-dimensional data vectors, the vectors are placed in the columns of an $n \times m$ matrix V where m is the number of samples in the data set. The standard NMF problem is to find two new reduced-dimensional matrices W and H to approximate the original matrix V by the product of $W * H$ in terms of a metric. Given a non-negative V, find non-negative matrix factors W and H such that:

$$V \approx WH \tag{1}$$

This matrix is approximately factorized into the $n \times r$ matrix W and the $r \times m$ matrix H. Each column of W contains a basis vector while each column of H contains the weights

needed to approximate the corresponding column in V using the basis from W. Usually r is chosen to be smaller than n or m, so that W and H are smaller than the original matrix V. The choice of r is generally application dependent.

By setting vector w_i as the rth column of W and vector h_r as the rth row of H, that is, $W = \{w_1, w_2, \cdots, w_r\}$, $H = \{h_1, h_2, \cdots, h_r\}^t$, Eq. (1) can be re-written as:

$$V \approx \sum_{i=1}^{r} w_i h_i \tag{2}$$

Regarding each vector w_r as a basis feature in V, the corresponding h_r is the vector of coefficients or encoding of this feature.

To find an approximate factorization, we need to define a cost function that quantifies the quality of the approximation. We used the Kullback-Leibler divergence as the cost function.

$$D(V||WH) = \Sigma_{ij}(V_{ij} log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}) \tag{3}$$

The divergence $D(V||WH)$ is minimized with respect to W and H, subject to the constraints W, $H \geq 0$. Although the function $D(V||WH)$ are convex in W or H, they are not convex in both combined variables. Therefore, it is unrealistic to expect an algorithm to solve a problem of minimization in the sense of finding global minima.

Matrices W and H are estimated by following multiplicative update rules.

$$W_{ij} \leftarrow W_{ij} \frac{\Sigma_k H_{jk} V_{ik} / (WH)_{ik}}{\Sigma_a H_{ja}} \tag{4}$$

$$H_{jk} \leftarrow H_{jk} \frac{\Sigma_i W_{ij} V_{ik} / (WH)_{ik}}{\Sigma_b W_{bj}} \tag{5}$$

These processes iterate until convergence or after p iterations. Initialization is performed using positive random initial conditions for W and H. Clearly, the approximations of W and H remain non-negative during the updates. It is generally best to update W and H simultaneously, instead of updating each matrix fully before the other. In this case, after updating a row of H, we update the corresponding column of W.

3. Speaker clustering based on non-negative matrix factorization

Next, we present the speaker clustering procedure based on the proposed method. Speaker clustering is performed through an iterative process of computing distances between speaker models for utterances and then merging them. In the first step, GMMs are trained as the speaker models for utterances. We use the cross likelihood ratio (CLR) [11] which is based on likelihoods for corresponding utterances, as the distance measure between models.

We used the CLR between utterances as elements in V of NMF. The dimensions n and m of V are the number of utterances. We let the elements of V be a reciprocal of the CLR because the CLR is the distance and where the smaller value is similar. The matrices W and H are obtained by decomposing V. Matrix W contains basis vectors obtained from the CLR between utterances. Matrix H contains weights to the basis vectors based on the CLR between utterances in each utterance. The proposed method performs speaker clustering by merging the corresponding utterances when the basis vectors of which the weights are maximum, are the same.

The procedure of the proposed method is described in detail as follows.

1. Training: For each cluster, GMMs are trained using an utterance of the cluster. Each utterance forms one cluster.

2. Distance calculation: The distance between utterances is computed based on the CLR. The CLR d_{ij} for utterances is given by

$$d_{ij} = \log \frac{P(X_i|\lambda_i)}{P(X_i|\lambda_j)} + \log \frac{P(X_j|\lambda_j)}{P(X_j|\lambda_i)} \tag{6}$$

$$\log P(X_i|\lambda_j) = \frac{1}{T_i} \sum_{t=1}^{T_i} \log P(x_{it}|\lambda_j)$$

where X_i is an utterance i, x_{it} is a feature vector of the tth frame of utterance i, T_i is the number of frames of utterance i, λ_i is the parameters of a GMM for utterance i, and $\log P(X_i|\lambda_j)$ is the average log likelihood of utterance i given by the GMM λ_j.

3. Decomposition of V: Matrix V is constructed using the CLR between utterances. The matrices W and H are estimated using the multiplicative update rules in Eqs.(4) and (5). We need to set a reduced-dimension number r and the number of iterations in Eqs.(4) and (5).

4. Clustering of utterances: The utterances are merged if the basis vectors with a maximum weight are the same in H obtained in Step 3. An example of V, W, and H when the number of utterances is 9 and the reduced-dimension number is 3 is shown as follows.

$$\begin{bmatrix} v_{11} & v_{12} & \dots & v_{19} \\ v_{21} & v_{22} & \dots & v_{29} \\ \vdots & \vdots & \vdots & \vdots \\ v_{91} & v_{92} & \dots & v_{99} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ \vdots & \vdots & \vdots \\ w_{91} & w_{92} & w_{93} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & \dots & h_{19} \\ h_{21} & h_{22} & \dots & h_{29} \\ h_{31} & h_{32} & \dots & h_{39} \end{bmatrix} \tag{7}$$

where v_{ij} is the CLR between utterances i and j, w_{ij} is an element of the jth basis vector, and h_{ij} is a weight to the ith basis vector by utterance j.

If the value of h_{21} is maximum in utterance 1 and the value of h_{29} is maximum in utterance 9, utterances 1 and 9 are merged because it shows that they have high similarity. Therefore, the number of dimensions r corresponds to the number of clusters.

In AHC, it is necessary to recompute the distance between clusters if utterances are merged and the distance between utterance pairs is only used as a similarity between utterances. However, it is not necessary to calculate distances in the clustering process by first calculating the distance between utterances in the proposed method. Moreover, the proposed method can take into account the similarity between all the utterances by decomposing the matrix that uses distance between utterances as matrix element. Therefore, the proposed method is able to perform fast speaker clustering by only decomposing the matrix.

Table 1: *Details of each test set*

Test set	Number of speakers	Number of utterances	Min & Max utterances	Total time (min.)
A1	6	116	15-28	25.8
A2	6	125	9-30	28.4
A3	6	138	12-41	33.3
A4	6	117	15-27	28.0
A5	6	158	14-34	37.5
B1	8	162	12-27	37.9
B2	8	163	10-26	39.5
B3	8	175	14-33	41.0
B4	8	172	15-30	40.6
B5	8	181	12-32	43.0

4. Experiments

4.1. Database and procedure

We used the CSJ as the material for the speaker clustering experiments. The CSJ consists of 3302 talks (662 hours, 1417 unique speakers) from academic conference presentations (on the natural and social sciences and engineering) and extemporaneous speeches on everyday topics. The talks are segmented into utterances automatically at every points where pauses were longer than 300 ms.

We conducted speaker clustering experiments for ten test sets. Table 1 lists details of each test set. The $Number\ of\ utterances$ is the total number of utterances, $Min\ \&\ Max\ utterances$ is the minimum and maximum number of utterances for speakers, and $Total\ time$ is the total duration (minutes) of utterances in each test set. The duration of an utterance ranges from 10 to 20 seconds. There is a variation in the number of utterances for every speaker to make the test sets as close to actual discussions. Each test set consisted of utterances of multiple speakers randomly chosen from the CSJ. Different speakers were used for each test set.

4.2. Experimental conditions and evaluation measures

The speech data were divided into utterance units based on energy parameters. The speech data were sampled at 16 kHz and the acoustic features consisted of 12 Mel-frequency cepstral coefficients.

We compared the proposed method with the conventional methods, AHC with the BIC-based method and the GMM-based method using CLR. In the GMM-based method, the merged cluster keeps the original component speaker models and the clustering process finishes if all mean distances for utterances in the merged clusters are larger than the threshold. The threshold in the GMM-based method and the penalty factor in the BIC-based method were set after preliminary experiments. There were four mixtures in the GMM-based and proposed methods.

The clustering results were aligned with the ground truth speaker labels to measure their accuracy based on the speaker diarization error rate (DER) [12]:

$$DER = \frac{U_{miss} + U_{wrong}}{U_{ref}} \tag{8}$$

where U_{miss} is the total length of utterances not aligned with the speaker labels, U_{wrong} is the total length of utterances aligned with the wrong speaker labels, and U_{ref} is the total length of all utterances in a test set. We also calculated the purity metric [13]:

$$Purity = \frac{U_{pure}}{U_{ref}} \tag{9}$$

where U_{pure} is the total length of the speaker label with the longest utterance duration for each cluster.

Table 2: *Diarization error rate (%) for each test set*

Test set	BIC	GMM	NMF
A1	19.3 (5)	21.7 (15)	6.9 (6)
A2	31.1 (6)	13.0 (13)	17.6 (6)
A3	30.9 (6)	30.2 (24)	9.2 (6)
A4	16.8 (6)	27.9 (12)	16.7 (6)
A5	28.0 (8)	24.9 (17)	14.1 (6)
B1	29.4 (9)	20.1 (24)	9.4 (8)
B2	23.0 (8)	32.2 (22)	20.6 (8)
B3	28.9 (8)	13.5 (23)	4.8 (8)
B4	27.7 (8)	20.0 (18)	16.0 (8)
B5	23.3 (6)	19.5 (22)	9.6 (8)

Table 3: *Cluster purity (%) for each test set*

Test set	BIC	GMM	NMF
A1	80.1	77.6	92.2
A2	68.0	86.4	83.2
A3	69.6	68.1	90.6
A4	81.2	70.9	82.1
A5	70.3	73.4	85.4
B1	69.8	77.2	90.1
B2	74.8	64.4	76.1
B3	69.7	84.6	94.9
B4	71.5	77.9	83.7
B5	75.7	79.0	90.1

4.3. Experimental results

The results with the DER as a speaker clustering accuracy measure are listed in Table 2. The results with cluster purity as a speaker clustering accuracy measure are listed in Table 3. In these tables, BIC are the results from using the BIC-based method, GMM are results from using the GMM-based method using CLR, NMF are the results from using the proposed method when the reduced -dimension for test sets A and B varied. The numbers in parenthesis in Table 2 indicates the number

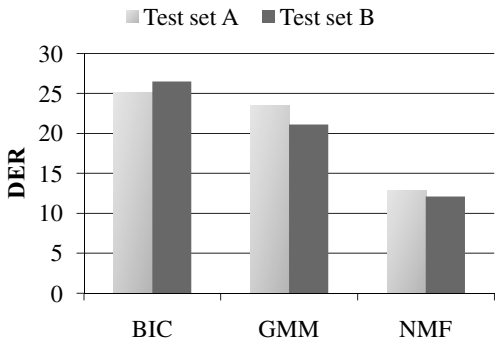

Figure 1: *Clustering accuracy (DER) by each method.*

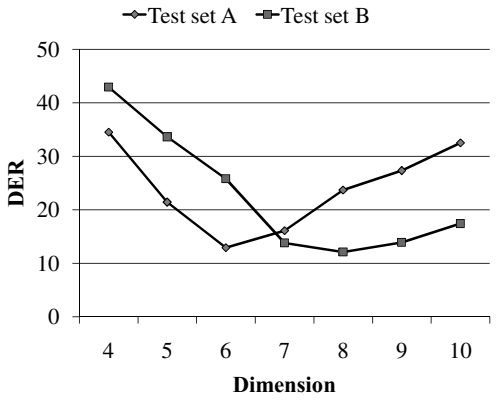

Figure 2: *Clustering accuracy (DER) versus reduced-dimension in NMF.*

of obtained clusters. The number of clusters is not included in Table 3 because it is the same as that in Table 2.

From the results in Tables 2 and 3, the GMM-based method obtained higher clustering accuracy than the BIC-based method. However, the estimation accuracy of the number of speakers with the BIC-based method was higher than that with the GMM-based method. The proposed method obtained higher clustering accuracy than both conventional methods for all test sets except test set A2.

The average clustering accuracy with each method is shown in Fig. 1. The DER with the BIC-based method was 25.2% for test set A and 26.5% for test set B, that with the GMM-based method was 23.5% for test set A and 21.1% for test set B, and that with the proposed method was 12.9% for test set A and 12.1% for test set B.

Figure 2 shows the clustering accuracy with each reduced -dimension of NMF in the proposed method. The clustering accuracy was the highest when the number of reduced -dimensions was six in test set A and when was eight in test set B. The correct number of speakers was six in test set A and eight in test set B. The proposed method obtained the highest accuracy when the number of reduced -dimensions was the same as the correct number of speakers.

The proposed method can achieve fast and robust speaker clustering only by decomposing a matrix based on the distances between utterances. Therefore, we demonstrated that speaker clustering based on NMF was effective for utterances of multiple speakers. We will develop a method for selectings an optimal reduced -dimension beforehand.

5. Conclusions

We proposed a speaker clustering method based on NMF using distances between utterances. The proposed method achieves fast and robust speaker clustering by decomposing a matrix based on the distances between utterances. We conducted speaker clustering experiments using academic conference presentations to evaluate the proposed method and compared it with two conventional methods. From the results of the experiments, the DER, as the average accuracy measure for test sets A and B, with the BIC-based method was 25.9%, the DER with the GMM-based method was 22.3%, and the DER with the proposed method was 12.5%. We demonstrated that the proposed method was effective in speaker clustering.

For future work, we will evaluate the proposed method using the National Institute of Standards and Technology (NIST) databases to demonstrate its generality. It is also necessary to study how to select the optimal number of reduced -dimensions in NMF.

6. References

[1] S. E. Tranter and D. A. Reynolds, "An Overview of Automatic Speaker Diarization Systems", IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, No.5, pp.1557-1565, 2006.

[2] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp.127-132, 1998.

[3] K. J. Han, S. S. Narayanan, "Agglomerative Heirarchical Speaker Clustering Using Incremental Gaussian Mixture Cluster Modeling", Proc. INTERSPEECH, pp.20-23, 2008.

[4] I. F. Chen, S. S. Cheng, and H. M. Wang, "Phonetic Subspace Mixture Model for Speaker Diarization", Proc. INTERSPEECH, pp.2298-2301, 2010.

[5] K. J. Han and S. S. Narayanan, "A Novel Inter-cluster Distance Measure Combining GLR and ICR for Improved Agglomerative Hierarchical Speaker Clustering", Proc. ICASSP, pp.4373-4376, 2008.

[6] M. Nishida and T. Kawahara, "Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing", IEEE Transactions on Speech and Audio Processing, Vol.13, No.4, pp.583-592, 2005.

[7] S. Bozonnet, N. Evans, C. Fredouille, D. Wang, and R. Troncy, "An Integrated Top-Down/Bottom-Up Approach To Speaker Diarization", Proc. INTERSPEECH, pp.2646-2649, 2010.

[8] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization", Proc. NIPS, pp.556-562, 2000.

[9] K. Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation", Proc. ISCA & IEEE Workshop on SSPR, pp.7-12, 2003.

[10] K. Iso, "Speaker Clustering Using Vector Quantization and Spectral Clustering", Proc. ICASSP, pp.4986-4989, 2010.

[11] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. J. McLaughlin, and M. A. Zissman, "Blind Clustering of Speech Utterances based on Speaker and Language Characteristics", Proc. ICSLP, pp. 3193-3196, 1998.

[12] S. E. Tranter and D. A. Reynolds, "Speaker Diarisation for Broadcast News, " in Odyssey04 - The Speaker and Language Recognition Workshop, pp.337-344, 2004.

[13] D. Liu and F. Kubala, "Online Speaker Clustering", Proc. ICASSP, vol.I, pp.572-575, 2003.