INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Estimating Speaker Clustering Quality Using Logistic Regression

Yishai Cohen, Itshak Lapidot

Afeka Tel-Aviv College of Engineering, ACLP, Israel

`yishaic@afeka.ac.il, Itshakl@afeka.ac.il`

Abstract

This paper focuses on estimating clustering validity by using logistic regression. For many applications it might be important to estimate the quality of the clustering, e.g. in case of speech segments' clustering, make a decision whether to use the clustered data for speaker verification. In the case of short segments speakers clustering, the common criteria for cluster validity are *average cluster purity* (ACP), *average speaker purity* (ASP) and K - the geometric mean between the two measures. As in practice, true labels are not available for evaluation, hence they have to be estimated from the clustering itself. In this paper, mean-shift clustering with PLDA score is applied in order to cluster short speaker segments represented as i-vectors. Different statistical parameters are then estimated on the clustered data and are used to train logistic regression to estimate ACP, ASP and K. It was found that logistic regression can be a good predictor of the actual ACP, ASP and K, and yields reasonable information regarding the clustering quality.

Index Terms: Cluster validity, Logistic Regression, I-vectors, Mean-shift, PLDA.

1. Introduction

Speaker recognition suffers from large degradation in performance when performed over short segments. In some cases many short segments can be supplied, but these segments may belong to many different speakers. So, before concatenating them to have sufficient amounts of speakers' speech, clustering has to be performed. Such a scenario is typical in systems using *push-to-talk* (PTT) equipment, such as taxi stations, airports and homeland security systems.

One of the biggest challenges in clustering is to estimate the cluster validity. One of the issues in cluster validity is answering the question regarding the quality of clustering the data, or how well the data separation works. In many cases the clusters themselves are not well defined, since different clusters can be confirmed based on the same data, depending, for example, on the optimization criterion. In speaker clustering *average cluster purity* (ACP), *average speaker purity* (ASP) and K, which is the geometric mean between the first two measures, are commonly used to validate the clustering quality [1]. In this paper we focus on short segments clustering of many different speakers (where segments' duration is about 2.5 seconds on the average). We applied a mean-shift clustering algorithm [2, 3] to cluster the i-vectors that were extracted from each segment [4, 5]. One of the possible applications of this clustering process, is to concatenate all short segments that were clustered together to a single cluster, and extract an i-vector from a relatively long speech segment, that is then used for speaker verification. However, if the clustering does not work well, we cannot rely on it, and cannot concatenate those short segments together to create single speakers' i-vectors. In this paper we deal with the challenge of training a system that yields good predictions regarding the quality of the clustering.

In this paper we rely on our previous work using mean-shift with *probabilistic linear discriminant analysis* (PLDA) score [6]. In the current work we did not focus on improving the clustering algorithm, rather we focused on constructing a system that estimates the clustering quality. For the purpose of cluster quality estimation, after the clustering we also saved the shifted i-vectors and the attributions of the i-vectors.

In the literature there are many criteria that measure clustering quality [7–9]. However, these measures are mostly task independent, hence not necessarily optimal for our speaker verification application. We use a subset of these criteria as inputs to logistic regression. The criteria we used are: *within single linkage* (WSL), *within complete linkage* (WCL), *within average linkage* (WAL), *between single linkage* (BSL), *between complete linkage* (BCL), *between average linkage* (BAL), *between centroid linkage* (BcenL), *Davies-Bouldin index* (DB), DUNN index, *Hartigan index* (Han), *Krzanowski-Lai index* (KL) and *Separation indexes* (Sep). The outputs of the logistic regression were either ACP, ASP or K. We trained the regression using a training set we created from many clustering runs, and then tested it using a different set of clustering runs, versus the true ACP, ASP and K values. For both training and test sets, we intentionally choose a large variety of speakers ($2-60$) and with different mean-shift setups in order to have all kinds of results from almost perfect clustering to a very poor clustering. This way, the logistic regression could learn all possible scenarios and all the scenarios could be examined on the test set.

The paper is organized as follows: In section 2 the mean-shift algorithm is described; Section 3 presents the cluster validity system. Section 4 contains experimental settings and results, and section 5 concludes the paper.

2. Mean-shift algorithm

Mean-shift algorithm for clustering is well known [2, 3], where the standard algorithm is based on Euclidean distance. As Euclidean distance is not fit to work well with i-vectors, it was first replaced by cosine distance [4, 5] and later with PLDA score [6]. Another change to the standard algorithm is replacing the threshold h that determines the neighboring i-vectors, by *k-nearest neighbors* (kNN). It was found in that kNN is much less sensitive to the k value then the h parameter [6]. Let $\mathcal{X} = \{x_j\}_{j=1}^{J}$ be a set of i-vectors from several speakers, and let $S_h(x)$ be the set of the k nearest i-vectors, then the mean shift is given in eq. (1), and the mean-shift algorithm is described in Algorithm 1.

$$m_h(x) = \frac{\sum\limits_{x_i \in S_h(x)} x_i}{k} - x \tag{1}$$

As for the proposed mean-shift, a PLDA score is required, so before performing clustering we train the *universal background model* (UBM) and the *total variability* (TV) matrix for i-vectors extraction. We also train the PCA matrix T, and the

http://dx.doi.org/10.21437/Interspeech.2017-492

Algorithm 1 Mean-shift clustering algorithm

```
Require:
  A set of vectors, $\mathcal{X} = \{x_j\}_{j=1}^{J}$            ▷ $x \in \mathbb{R}^{Q \times 1}$
  A number of neighbors $k$                                     ▷ $k \in \mathbb{N}$
  A cluster merging threshold $Th$                              ▷ $Th \in \mathbb{R}^{+}$
  for $j := 1$ to $J$ step 1 do
      Set $\hat{x}_j = x_j$.
      repeat
          Find $k$ i-vectors with the highest score with $\hat{x}_j$.
          Calculate $m_h(\hat{x}_j)$, the shift of the vector $\hat{x}_j$, using eq. (1).
          $\hat{x}_j \leftarrow \hat{x}_j + m_h(\hat{x}_j)$
      until Convergence
  Cluster the the shifted vectors $\hat{\mathcal{X}} = \{\hat{x}_j\}_{j=1}^{J}$ such that the
  distance between 2 shifted vectors will be less than $Th$.
  Return: Cluster index of each i-vector and the shifted i-vectors $\hat{\mathcal{X}}$.
```

whitening transformation matrix C. We end with low rank normalized i-vectors, as given in eq. (2). As the main goal of this paper is to qualify the clustering and not to find the best clustering algorithm, we use the system described in [6]. It was found that PCA works better than LDA for dimensionality reduction before the PLDA scoring.

$$\varphi = \frac{C\mathrm{T}x}{\|C\mathrm{T}x\|} \tag{2}$$

Having the set of low rank normalized i-vectors $\{\varphi\}$, we then estimate the within covariance matrix W^{-1}, the between covariance matrix B^{-1} and the i-vectors expectation vector μ for the two-covariance model parameters.

3. Cluster validity

In this section we describe the proposed system for estimating the validity of a given clustering. First we describe the parameters (features) that were calculated over the clustering result which serve as inputs to the logistic regression. Then the logistic regression will be shortly presented.

3.1. Cluster validity parameters

The input features that were used for the logistic regression are termed as cluster validity parameters, and are presented below. These features help us understand the degree of separation in our data, and were used as input features for our logistic regression. The distances between our data points are calculated in the feature-space spanned by these validity parameters.
In our work we use the following indexes:
x_{in} - n^{th} vector of the i^{th} cluster. In our case x is an i-vector.
C_i - Cluster i.
c - Number of clusters.
μ_i - The mean of cluster i.
μ - The mean of the whole dataset.
N_i - Number of vectors in cluster i.
N - Total number of vectors in the dataset.
$d(\alpha,\beta)$ - Euclidean distance between two vectors.
$R_{\alpha\beta}$ - Pearson correlation coefficient.
Pearson correlation coefficient between two Q dimensional vectors as defined in 3.

$$\begin{gathered} R_{\alpha\beta} = \frac{\bar{\alpha}^{\top}\cdot\bar{\beta}}{\sqrt{\bar{\alpha}^{\top}\cdot\bar{\alpha}}\sqrt{\bar{\beta}^{\top}\cdot\bar{\beta}}} \\ \bar{\alpha} = \alpha - \frac{1}{Q}\sum_{q=1}^{Q}\alpha_q \qquad \bar{\beta} = \beta - \frac{1}{Q}\sum_{q=1}^{Q}\beta_q \end{gathered} \tag{3}$$

where $\top$ is the transpose operator.

Calculating the following parameters, we used normalized Pearson correlation coefficient in the range $[0,1]$ as $0.5\,(1-R_{\alpha\beta}) \rightarrow R_{\alpha\beta}$. Additionally, we also define:
R_{w_i} - The within cluster dispersion.

$$R_{w_i} = \sum_{n=1}^{N_i} R_{x_{in}\mu_i} \tag{4}$$

$R_{b_{ij}} = R_{\mu_i\mu_j}$ - Dispersion between clusters i and j.

WSL – within single linkage [7]:

The minimal Euclidean distance between two data points from the same cluster as in eq. (5):

$$WSL = \min_{1\le i\le c}\left\{\min_{n\ne m}\{d(x_{in},x_{im})\}\right\} \tag{5}$$

WCL - within complete linkage [7]:

The maximal Euclidean distance between two data points from the same cluster as in eq. (6):

$$WCL = \max_{1\le i\le c}\left\{\max_{n\ne m}\{d(x_{in},x_{im})\}\right\} \tag{6}$$

WAL - within average linkage [7]:

The average Euclidean distance between all pairs of data points from the same cluster as in eq. (7):

$$WAL = \underset{i}{\mathrm{mean}}\left\{\underset{n\ne m}{\mathrm{mean}}\{d(x_{in},x_{im})\}\right\} \tag{7}$$

BSL – between single linkages [7]:

The minimal Euclidean distance between two data points from different clusters as in eq. (8):

$$BSL = \min_{i\ne j}\left\{\min_{i\in C_i, j\in C_j}\{d(x_{in},x_{jm})\}\right\} \tag{8}$$

BCL - between complete linkages [7]:

The maximal Euclidean distance between two data points from different clusters as in eq. (9):

$$BCL = \max_{i\ne j}\left\{\max_{i\in C_i, j\in C_j}\{d(x_{in},x_{jm})\}\right\} \tag{9}$$

BAL – between average linkages [7]:

The average Euclidean distance between all pairs of data points from different clusters as in eq. (10):

$$BAL = \underset{i\ne j}{\mathrm{mean}}\left\{\underset{i\in C_i, j\in C_j}{\mathrm{mean}}\{d(x_{in},x_{jm})\}\right\} \tag{10}$$

BcenL – between centers linkages [7]:

The maximal Euclidean distance between all pairs of clusters centroids as in eq. (11):

$$BcenL = \max_{i \neq j} \{d(\mu_i, \mu_j)\} \tag{11}$$

DB - Davies-Bouldin index [7]:

This index aims to identify sets of clusters that are compact and well separated. The Davies-Bouldin index is defined as in eq. (12):

$$DB = \frac{1}{c} \sum_{i=1}^{c} \max_{j \neq i} \left\{ \frac{R_{w_i} + R_{w_j}}{R_{b_{ij}}} \right\} \tag{12}$$

Smaller value indicates a "better" clustering solution.

DUNN index [7]:

Dunn index is defined as in eq. (13):

$$DUNN = \min_{1 \leq i,j \leq c} \left\{ \frac{R_{b_{ij}}}{\max\limits_{1 \leq k \leq c} R_{w_k}} \right\} \tag{13}$$

Large values of Dunn index correspond to good clustering solution.

Han - Hartigan index [8]:

Hartigan equation is as in eq. (14):

$$Han(c) = \left\{ \frac{W_c}{W_{c+1}} - 1 \right\} / (N - c - 1) \tag{14}$$

where,

$$W_c = \frac{1}{2} \sum_{i=1}^{c} R_{w_i}$$

Hartigan index was initially defined to estimate the number of clusters, but here we do not have several clusterings. As we cannot compare different clusterings, we calculate only W_c as a feature to the logistic regression.

KL - Krzanowski-Lai [8]:

The Krzanowski and Lai index is defined as in eq. (15):

$$KL(c) = \left| \frac{DIFF(c)}{DIFF(c+1)} \right| \tag{15}$$

where,

$$DIFF(c) = (c-1)^{2/Q} W_{c-1} - c^{2/Q} W_c$$

As in Hartigan index case, we cannot compare different clusterings, so we just used the coefficient $c^{2/Q} W_c$.

Sep - Separation index [9]:

Separation index is calculated as the weighted average between cluster dispersion as in eq. (16):

$$Sep = \frac{1}{\sum\limits_{i \neq j} N_i N_j} \sum_{i \neq j} N_i N_j R_{b_{ij}} \tag{16}$$

Separation index reflects the overall dispersion between clusters. Increasing separation index suggests an improvement in the clustering results.

3.2. Logistic regression

Logistic regression is a well known algorithm to estimate a grade of an ordinal data in the range $[0, 1]$ [10]. By applying the inverse logit function it is possible to design a model to estimate information based on a training data set. The logistic expression (the inverse logit function) is as in eq. (17):

$$f = \frac{1}{1 + e^{-z}} \tag{17}$$

where $z = w^T \phi$, ϕ and w are column vectors:
ϕ - input feature vector (in our case, different statistics we extract after the clustering), after mean substitution and variance normalization of each feature dimension.
w - the weights vector of the linear combination we wish to estimate.
f - is the output of the logistic expression and its range is $[0, 1]$. In our case, f will be the estimate of the clustering quality, either ACP, ASP or K.

Having a training set, that includes pairs of clustering statistics vectors ϕ_n (the parameters of this vector are described in the sub-section, 3.1) and calculated clustering quality parameter O_n (O_n can be either ACP, ASP or K), $\{\phi_n, O_n\}_{n=1}^{N}$, the weights vector w is trained. The optimization criterion was *minimum mean squared error* (MMSE) between the clustering quality value O_n and the regression output f_n. The prediction quality is tested on a separate set. For each set of i-vectors, clustering is performed, and the statistics are calculated and denoted as ϕ_m. The output f_m is obtained and compared to the true output O_m. The systems's error is calculated over all M test trials using eq. 18.

$$E = 100 \sqrt{\frac{1}{M} \sum_{m=1}^{M} (O_m - f_m)^2} \tag{18}$$

4. Experiments and results

4.1. I-vectors training data

We used *Mel frequency cepstral coefficients* (MFCC), that were extracted using a 25 ms Hamming window with a frame rate of 100 frames per second. 19 MFCC features together with log energy. Cepstral mean subtraction and variance normalization were applied to the MFCCs. These vectors were augmented by the delta and delta delta to produce all together 60-dimensional feature vectors. Male only UBM of 2048 Gaussians mixture components was trained using Fisher Part 1; Switchboard II, Phase 2; switchboard Cellular, Parts 1 and 2; and NIST 2004-2006 SREs. Then, the total variability matrix with a low rank of 400 was trained using labeled data from same databases as for the UBM. In total, 975 unique male speakers with 10705 sessions were used. Part of this data was also used to estimate the PCA and the whitening transformations and the PLDA score parameters. PLDA score was calculated on low dimension i-vectors, of dimension 250, after PCA and whitening.

4.2. Experiment setup

The i-vectors database is based on NIST2008 which has 188 different speakers. We used 98 speakers for our training set, and the remaining 90 speakers for testing in order to have a wide and statistically independent data sets. The database contains male speakers only. For each of the two sets, training and test ones, we forced our mean-shift algorithm to run over several different numbers of speakers $(2, 3, 5, 7, 12, 15, 17, 22, 24, 30, 48, 60)$, different k for the kNN $(3, 5, 7, 9, 13, 17)$, and different amounts of segments per speaker. The large variety of parameters ensure that some of the clustering trials will be good, others bad, and some should be ”in between”. We did it so the logistic regression could learn all scenarios of clustering performances. The segments length distribution and the number of segments per speaker are shown in figure 1.

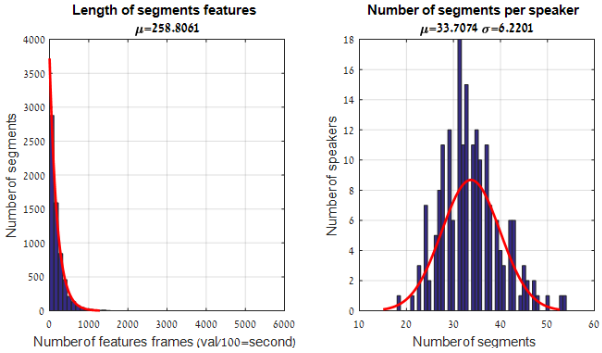

Figure 1: *PTT duration distribution (on the left), and the distribution of the number of segments per speaker (on the right)*

We randomly choose the clustering data for $16,000$ times, $8,000$ for training and $8,000$ for testing. For all the $16,000$ sets of segments, mean-shift clustering is performed. The shifted i-vectors and their attribution to the clusters were saved. The resulting ACP, ASP and K parameters were calculated. Using the information about the i-vectors positions pre- and post-mean-shift, and their clusters attribution, we calculate cluster validity parameters (according to the description in section 3.1). We calculated these parameters twice. First, using the original i-vectors (named **Before** in the results table); Then, using the shifted i-vectors (named **After** in the results table). Then we concatenated the **Before** and the **After** features, and obtained a double size length vector (named **Both** in the results table). For each clustering trial, three sets of parameters are extracted (12 for **Before** and **After** cases and 24 for the **Both** case) and used as feature vectors to the logistic regression.

The $8,000$ feature vectors of the training set were used to train three logistic regressions, for ACP, ASP and K ($K = \sqrt{ACP \cdot ASP}$). In total, 9 regressions were trained (3 for the original i-vectors, 3 for the shifted ones, and 3 for the concatenated ones). All the systems were trained to minimize the *mean squared error* (MSE).

4.3. Experiments

After the 9 systems were trained, we used the 3 sets of features extracted from the clustering output on the test data. We calculated the MSE between the true outputs (ACP, ASP and K) and the same measures as estimated by the logistic regression systems.

4.4. Results

The results are presented in table 1. The prediction error is according to eq. 18. It can be seen that the prediction of the logistic regression is better for all criteria (ACP, ASP, K) when using the original i-vectors (before applying mean-shift) compared to the shifted i-vectors, while the lowest error rate was obtained by the concatenated features. We have no explanation as to why the results of the i-vectors **Before** the mean-shift yielded better results than those of the **After** mean-shift. More in-depth research is required to understand this phenomenon as it was consistent for all the experiments. Moreover, we find it interesting to calculate K out of the prediction of ACP and ASP, according to the formula for K (the K1 column in table 1). It was found that the calculated K is almost as accurate as the predicted one by the logistic regression.

Table 1: *Prediction error for ACP, ASP, K1 - the geometric mean, and the predicted K.* ***Before*** *and* ***After*** *mean-shift and* ***Both*** *concatenated.*

	ACP	**ASP**	**K1 =** $\sqrt{ACP \cdot ASP}$	**K**
Before	8.03	9.60	7.79	7.51
After	9.24	13.36	11.47	10.73
Both	7.62	8.45	6.79	6.57

The best score for all criteria was achieved by the **Both** variant. Moreover, estimating K directly yields the best MSE score, so it predicts a better score than the geometric mean of the predicted ACP and ASP, although they were quite close.

5. Conclusions

Clustering validation is an important issue and may have different consequences on further processes. In this paper, we show that good prediction of clustering quality can be done using logistic regression. It was shown that different measures can be estimated - ACP, ASP or K; They can all be estimated well using the same set of statistical parameters extracted from the clustering process and results.

The clustering was done using the mean-shift algorithm with PLDA score, but prediction of the clustering quality in the **Before** case does not relate to the specific clustering algorithm, while **After** and **Both** are strongly dependent on the mean-shift algorithm. It was shown that the **Before** mode is a good predictor of ACP, ASP and K, while parameters extracted from the shifted i-vectors carry some supplementary information thus improving the prediction.

In the future we intend to replace the logistic regression by an artificial neural network to see whether the prediction may be even more accurate.

6. Acknowledgements

The authors thanks Irit Opher on her great help in all aspects to make the paper clear and understandable.

7. References

[1] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, “Unknown-multiple speaker clustering using HMM,” in *In Proceedings of ICSLP-2002*, 2002, pp. 573–576.

[2] K. Fukunaga and L. Hostetler, “The estimation of the gradient of a density function, with applications in pattern recognition,” *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, Jan 1975.

[3] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[4] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, “A study of the cosine distance-based mean shift for telephone speech

diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–227, Jan 2014.

[5] I. Shapiro, N. Rabin, I. Opher, and I. Lapidot, "Clustering short push-to-talk segments," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 3031–3035.

[6] I. Salmun, I. Opher, and I. Lapidot, "On the use of plda i-vector scoring for clustering short segments," in *ODYSSEY 2016 -The Speaker and Language Recognition Workshop*, 2012.

[7] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," *Signal Processing*, vol. 83, no. 4, pp. 825 – 833, 2003, genomic Signal Processing.

[8] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[9] G. Chen, S. A. Jaradat, N. Banerjee, T. S. Tanaka, M. S. H. Ko, and M. Q. Zhang, "Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data," *Statistica Sinica*, pp. 241–262, 2002.

[10] C. M. Bishop, "Pattern recognition and machine learning," 2006.