Computer Speech & Language 65 (2021) 101139

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Speaker clustering quality estimation with logistic regression

Yishai Cohen, Itshak Lapidot*

Afeka Tel-Aviv College of Engineering, ACLP, Israel

ARTICLE INFO

Article History:
Received 2 May 2019
Revised 26 May 2020
Accepted 3 August 2020
Available online 7 August 2020

Keywords:
Clustering quality estimation
Speaker clustering
Mean-shift
Stochastic vector quantization (VQ)
Probabilistic linear
discriminant analysis (PLDA)
Cosine distance
Logistic regression

ABSTRACT

This paper focuses on estimating the quality of a clustering process. The task is to cluster short speech segments that belong to different speakers. A variety of statistical parameters are estimated from the output of the clustering process. These parameters are used to train a logistic regression to serve as a clustering quality estimation system. In this paper, mean-shift clustering with either a cosine distance or *probabilistic linear discriminant analysis* (PLDA) score as the similarity measure, as well as stochastic *vector quantization* (VQ) with cosine distance, are applied in order to cluster the short speaker segments, which are represented by i-vectors. The quality of the clustering is measured using the *average cluster purity* (*ACP*), *average speaker purity* (*ASP*) and *K*, which is the geometric mean of *ASP* and *ACP*. We show that these measures can be estimated fairly well by applying logistic regression. Moreover, clustering quality may be well estimated even if the logistic regression was trained using parameters derived from a different clustering algorithm. This is very important, as it allows the use of a single quality estimation system, without the need for retraining when the clustering method is changed.

Additionally, we showed how the clustering quality estimator could be served as an estimator of the number of clusters. For VQ-based clustering the number of clusters has to be predefined. We perform the clustering with different number of clusters. The best number of clusters is estimated as the clustering that achieved the higher estimation of the *K* value. We will show that this approach estimate the best number of clusters accurately.

1. Introduction

Short-segments speaker clustering has considerable importance both for diarization (Ben-Harush et al., 2012; Lapidot et al., 2017; Senoussaoui et al., 2014; Anguera Miro et al., 2012) and applications such as short *push-to-talk* (PTT) segments clustering (Shapiro et al., 2015; Salmun et al., 2017). The output of clustering is frequently passed as an input to the next task, for instance, speaker verification. The success of the subsequent task is highly dependent on the quality of the clustering. For this reason, measuring the clustering quality is of high importance. In the case of poor clustering, it might be better not to take action at all. Unlike supervised tasks, where the output can be presented in terms of a likelihood ratio (which effectively reflects the level of confidence in the decision), quality estimation for clustering is less straightforward. The current study presents a simple approach in which a supervised logistic regression is trained to carry out clustering quality estimation. In previous work (Cohen and Lapidot, 2017), we showed that logistic regression can be trained to estimate *average cluster purity* (*ACP*), *average speaker purity* (*ASP*) and $K = \sqrt{ASP \cdot ACP}$, Ajmera et al. (2002). The results of the quality estimation are highly dependent on the clustering algorithm (Cohen and Lapidot, 2018a). In this study, we extend this research and show that it can be trained to be clustering-algorithm independent. This means that while the same features (or speaker models) are

*Corresponding author.
E-mail addresses: Yishaic@afeka.ac.il (Y. Cohen), itshakl@afeka.ac.il (I. Lapidot).

https://doi.org/10.1016/j.csl.2020.101139
0885-2308/

used for clustering, e.g., normalized i-vectors, the clustering quality can be well estimated even when different clustering algorithms are applied. For example, clustering might be achieved using the mean-shift with cosine distance as the similarity measure or the mean-shift with *probabilistic linear discriminant analysis* (PLDA). To emphasize the difference to our previous works, the mean-shift algorithms which are presented in Section 2, are based on Shapiro et al. (2015); Salmun et al. (2017) without modifications. The novelty is in the clustering quality estimation and in its analysis. Like in previous works, in the current research, all the experiments are performed in a simulated way using NIST SRE 2008 databases (LDC Catalog https://catalog.ldc.upenn.edu), as there are no public databases which fit the real world push-to-talk data.

In this paper, we investigated the clustering quality estimation on both mean-shift clustering algorithms and a stochastic VQ (vector quantization) with cosine metric-based technique (Cohen and Lapidot, 2018b). The mean-shift was chosen as it showed promising results in previous works (Senoussaoui et al., 2014; Shapiro et al., 2015; Salmun et al., 2017). We present different versions of the mean-shift in the past (Shapiro et al., 2015; Salmun et al., 2017) and now we apply these algorithms as a test cases for clustering quality estimation. For VQ-based clustering, the number of clusters either has to be known in advance or it has to be estimated (Jain, 2010; Bezdek et al., 1984; Kohonen, 1990; Pal et al., 2005). Many objective criteria exist for estimating the number of clusters (Chen and Gopalakrishnan, 1998; Tibshirani et al., 2001; Hansen and Yu, 2001; Figueiredo and Jain, 2002; Bolshakova and Azuaje, 2003). We applied the clustering quality estimator as a subjective, supervised, task dependent criterion to estimate the number of clusters. The VQ was performed on codebooks of different sizes, and the best clustering result was estimated using the logistic regression. The algorithm was tested on a large range of speakers, from 2 to 60. The results were compared to those of the mean-shift clustering method, which has already been tested for this task several times on a number of occasions. The VQ results were somewhat inferior to those of the cosine similarity measure-based mean-shift clustering.

The rest of the paper is organized as follows: Section 2 presents the different variants of the mean-shift clustering algorithm investigated in this study. Section 3 describes the stochastic VQ with cosine distance algorithm. The description of the clustering quality estimator is given in Section 4. The experimental methods and results are presented in Section 5. Finally, Section 6 concludes the paper.

2. Mean-shift algorithm

This section describes all the variants of the mean-shift algorithm applied in this study. They are all variants of the mean-shift algorithm that is based on Euclidean distance (Comaniciu and Meer, 2002). This algorithm has been used for speaker diarization in Senoussaoui et al. (2014), where the authors used a cosine distance instead of the Euclidean distance. Variants of cosine-based mean-shift clustering for short *push-to-talk* segments were presented in Shapiro et al. (2015). In Salmun et al. (2017) a mean-shift algorithm with PLDA score as the similarity measure was trained on short segments. Since one of the goals of the present study was to examine the robustness of the clustering quality estimator to different clustering algorithms, three variants of the mean-shift algorithm were investigated.

2.1. Standard mean-shift algorithm

Mean shift algorithm is a non-parametric iterative algorithm. It estimates the *probability density function* (*pdf*) of a random variable (Fukunaga and Hostetler, 1975). The algorithm is inspired by the Parzen window approach to non-parametric density estimation. The algorithm does not require any prior knowledge regarding the number of clusters, and has no assumptions regarding the shape of the clusters. Dense regions in the feature space correspond to local maxima or to the *pdf* modes. As such, for each data point, in order to reach the local maximum of the *pdf*, a gradient ascent on the estimated local density is performed until convergence is reached. Each stationary point represents a mode of the density function. Data points that are associated with the same stationary point are assigned to the same cluster.

The gradient of the density function is required in order to find the above mentioned modes. Following the mathematical formulation in Senoussaoui et al. (2014), Comaniciu and Meer (2002), Fukunaga and Hostetler (1975), the mean shift vector $m_h(x)$ expression is derived according to (1).

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

where x is the current position of the d dimensional i-vectors; h is the neighborhood (or the bandwidth) from which the gradient is estimated, and $S_h(x)$ is the set of i-vectors that are the neighbors of i-vector x, $S_h(x) \equiv \{x_i : \| x - x_i \| \leq h\}$.

Let $\mathcal{X} = \{x_j\}_{j=1}^{J}$ be d dimensional i-vectors to be clustered, then the mean-shift algorithm is as described in Algorithm 1. For each x_j out of $\mathcal{X}$ an iterative procedure is applied: find the neighbors of x_j; calculate the shift according to Eq. (1); shift x_j; repeat the procedure with the shifted i-vector until convergence. Convergence in this case is defined as a shift's norm which is smaller than a threshold Th_1. The last step is merging the shifted i-vectors which have an Euclidean distance smaller than Th_2 to the same cluster.

2.2. Modifications of the mean-shift algorithm

The modifications to the mean-shift algorithm were based on the following findings in the literature: (a) the PLDA score is preferable to both the Euclidean and the cosine distances, and (b) i-vector length normalization, $\phi = x / \| x \|$, prior to

Algorithm 1. Mean-shift clustering algorithm.

Require:

A set of vectors, $\mathcal{X} = \{x_j\}_{j=1}^{J}$ ▷ $x \in \mathbb{R}^{d\times 1}$

Neighborhood size h ▷ $h \in \mathbb{R}^{+}$

A cluster shift threshold Th_1 ▷ $Th_1 \in \mathbb{R}^{+}$

A cluster merging threshold Th_2 ▷ $Th_2 \in \mathbb{R}^{+}$

for $j := 1$ **to** J **step** 1 **do**

Set $\hat{x}_j = x_j$.

repeat

Find the subset $S_h(\hat{x}_j)$.

Calculate $m_h(\hat{x}_j)$, the shift of the vector $\hat{x}_j$, using (1).

$\hat{x}_j \leftarrow \hat{x}_j + m_h(\hat{x}_j)$

until $|m_h(x)| > Th_1$

Cluster the shifted vectors $\hat{\mathcal{X}} = \{\hat{x}_j\}_{j=1}^{J}$ such that the distance between 2 shifted vectors will be less then Th_2.

Return: Cluster index of each vector.

dimensionality reduction makes the algorithm more stable than without normalization Salmun et al. (2017). A further modification is to replace the bandwidth h with maximal number of neighbors k, such that only i-vectors with positive scores are used as neighbors, even if the final number of neighbors is less than the pre-defined value k (replacing $S_h(x)$ by $S_k(x)$). This meant that $S_k(x)$ consisted of at most k i-vectors that are closest to x (or ϕ in the normalized version), all of which had a non-negative PLDA score with respect to x.

Before calculating the PLDA, a dimensionality reduction is performed by applying *principal component analysis* (PCA) to the i-vectors. Whitening (matrix C) and length normalization are applied to the low dimensional i-vectors ($q < d$ dimensional) is according to (2).

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

In previous study we examined the use of PLDA for convergence stopping criterion (Th_1) and for the merging process (Th_2), however, it did not lead to any conclusive improvement and is much more computationally demanded, (Salmun et al., 2017). As such, the Euclidean distance is used as in the original mean-shift algorithm.

The PLDA-based mean-shift algorithm is described in Algorithm 2, and the short-segments clustering process is presented in Algorithm 3. The short-segments clustering apply the mean-shift clustering as the main block after a pre-processing.

Since the goal of this study is to test the robustness of the clustering quality estimator to different clustering algorithms, a modification to the baseline mean-shift algorithm were also examined:

Algorithm 2. PLDA-based mean-shift algorithm.

Require:

A set of vectors, $\Phi = \{\phi_j\}_{j=1}^{J}$ ▷ $\phi \in \mathbb{R}^{d\times 1}$

A set of vectors, $\Phi = \{\varphi_j\}_{j=1}^{J}$ ▷ $\varphi \in \mathbb{R}^{q\times 1}$

Maximal number of neighbors k ▷ $k \in \mathbb{N}$

A cluster shift threshold Th_1 ▷ $Th_1 \in \mathbb{R}^{+}$

A cluster merging threshold Th_2 ▷ $Th_2 \in \mathbb{R}^{+}$

for $j := 1$ **to** J **do**

Set $\hat{\phi}_j = \phi_j$.

repeat

Calculate $\hat{\varphi}_j$ according to eq. 2.

Find k i-vectors with the highest score of $\hat{\varphi}_j$ within Φ.

Find $S_k\left(\hat{\phi}_j\right)$.

Calculate $m_h\left(\hat{\phi}_j\right)$, the shift of the vector $\hat{\phi}_j$, using (1).

$\hat{\phi}_j \leftarrow \hat{\phi}_j + m_h\left(\hat{\phi}_j\right)$

until $|m_h(x)| > Th_1$

Form clusters from the shifted vectors $\hat{\Phi} = \left\{\hat{\phi}_j\right\}_{j=1}^{J}$ such that all pairs of vectors closer than Th_2 are defined as belonging to the same cluster.

Return: Cluster index of each vector.

- A cosine metric is applied to find the best *k* neighbors, instead of the PLDA score. Such a metric has the advantage that it can be applied when there are insufficient labeled and reliable data to train the PLDA. This way, the clustering process is fully unsupervised, as the cosine metric is calculated directly from the normalized i-vectors ϕ_j.

Thus, in the modification we used exactly the same features (ϕ), normalized i-vectors, but with a different optimization criterion (cosine distance instead of PLDA).

3. Stochastic VQ algorithm

In VQ based-clustering methods, the goal is to cluster a set of i-vectors into *Q* clusters, (Cohen and Lapidot, 2018b). This number should be defined *a-priori*. In this algorithm, we used a cosine metric as the optimization criterion instead of the commonly

Algorithm 3. Short speaker-segments clustering algorithm.

Require:

A set of speech segments $\{Seg_j\}_{j=1}^{J}$

Maximal number of neighbors k $\triangleright\ k \in \mathbb{N}$

A cluster shift threshold Th_1 $\triangleright\ Th_1 \in \mathbb{R}^+$

A cluster merging threshold Th_2 $\triangleright\ Th_2 \in \mathbb{R}^+$

1. For each speech segment Seg_j extract an i-vector x_j using the *universal background model* (UBM) and the *total variability* (TV) matrix.
2. Normalize each i-vector x_j to get a normalized i-vector ϕ_j.
3. Using T and C perform dimensionality reduction and spherical normalization according to eq. 2 to obtain the low rank i-vectors $\Phi = \{\varphi_j\}_{j=1}^{J}$.
4. Apply the modified mean-shift as described in Algorithm 1.

Return: Cluster index for each speech segment.

used Euclidean distance, as the latter has been found to be a poor similarity measure for speaker recognition technologies. We applied a stochastic VQ algorithm. Given a training dataset of i-vectors, $\mathcal{X}=\{x_j\}_{j=1}^{J}$, the goal of VQ is to cluster those i-vectors into Q groups that are represented by Q centroids $\mathcal{B}=\{b_q\}_{q=1}^{Q}$. The stochastic training of the VQ-based clustering algorithm is described in Algorithm 4.

It is important to mention that in this clustering approach, the number of clusters, Q, is not estimated and has to be set in advance. The estimation of the best value of Q was performed by clustering according to several *codebook* (CB) sizes and choosing the one with the best estimated clustering quality value, K (described in Section 4). The simplest way to estimate the Q value is shown in Fig. 1.

The clustering algorithm proceeded as follows. Firstly, all the i-vectors were normalized. The new set of normalized i-vectors denoted by Φ. Q i-vectors were then chosen randomly from Φ; these i-vectors constituted the initial CB $\mathcal{B}$. Next, at each iteration, one i-vector $\phi_j(t)$ was randomly chosen with replacement. The closest code-word from $\mathcal{B}$ was found, according to the cosine distance, and was adapted. Finally, the adapted code-word was normalized to have norm 1. This process continued until the termination condition was met. The adaptation factor, α, and the number of iterations, τ, were chosen empirically. We set $\alpha = 0.005$ and $\tau = 10^6$.

4. Clustering quality estimator

In this section we describe the estimator of the clustering quality. Firstly, we present a set of parameters (features) that are calculated from the clustering results. These parameters were used as an input vector to the logistic regression. In the next subsection, 4.1, we define a set of criteria for evaluating clustering quality. Finally, the logistic regression technique is briefly described. This clustering validation system has already been presented in Cohen and Lapidot (2017). However, in the current work we perform a deeper analysis of the generalization capabilities of the system.

Algorithm 4. Stochastic VQ.

Require:

A set of vectors, $\mathcal{X} = \{x_j\}_{j=1}^{J}$ ▷ $x \in \mathbb{R}^{d\times 1}$

Q ▷ Number of centroids in the CodeBook (CB)

α ▷ Adaptation factor $0 \leq \alpha \leq 1$

τ ▷ Number of iterations

Calculate the normalized dataset $\Phi = \{\phi_j\}_{j=1}^{J}$, such that $\phi_j = {x_j}/{\|x_j\|}$ ▷ $\phi \in \mathbb{R}^{d\times 1}$

Create a CB, $\mathcal{B} = \{b_1, \ldots, b_Q\}$, randomly choosing vectors from Φ ▷ $b \in \mathbb{R}^{d\times 1}$

for $t := 1$ **to** τ **do**

Randomly choose $j\,(t) \in \{1, \ldots, J\}$.

$q^* = \underset{q\in\{1,\ldots,Q\}}{\arg\min} \left\{ d_{\cos}\left(\phi_{j(t)}, b_q\right)\right\}$ ▷ $d_{\cos}(\bullet, \bullet)$ is a cosine distance.

- $b^{Temp} = b_{q^*} + \alpha \cdot \left[b_{j(t)} - b_{q^*}\right]$
- $b_{q^*} \leftarrow {b^{Temp}}/{\left\|b^{Temp}\right\|}$

Cluster each normalized i-vector ϕ_j by finding the minimum cosine distance between the data points and all the centroids in $\mathcal{B}$.

Return: Cluster index for each speech segment, the CB $\mathcal{B}$, and the normalized i-vectors Φ.

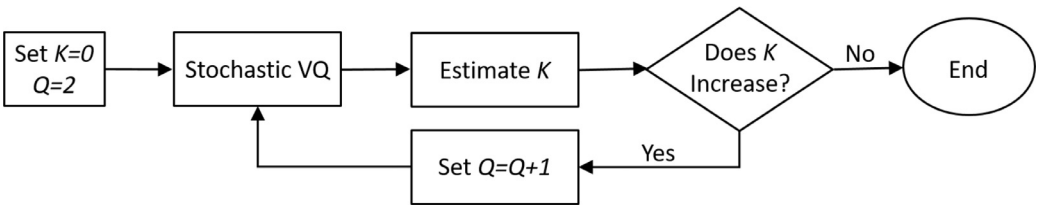

Fig. 1. Clustering system based on stochastic VQ and a clustering quality estimator.

4.1. Clustering validity parameters

The input features for the logistic regression are presented below. These features were used to measure the degree of separation in the data. Firstly we present a set of internal cluster quality measures (features), which do not need the class labels:

ϕ_{ij} - j^{th} normalized i-vector of the ith cluster
C_i - Cluster i
Q - Number of clusters
μ_i - Mean of cluster i
μ - Mean of the whole dataset
J_i - Number of vectors in cluster i
J - 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 d dimensional vectors is defined as in (3).

$$R_{\alpha\beta} = \frac{\overline{\alpha}^{\top} \cdot \overline{\beta}}{\sqrt{\overline{\alpha}^{\top} \cdot \overline{\alpha}}\sqrt{\overline{\beta}^{\top} \cdot \overline{\beta}}}$$
$$\overline{\alpha} = \alpha - 1_d \cdot \frac{1}{d}\sum_{l=1}^{d}\alpha_l \quad \overline{\beta} = \beta - 1_d \cdot \frac{1}{d}\sum_{l=1}^{d}\beta_l \tag{3}$$

where $\top$ is the transpose operator and 1_d id a d dimensional all-ones vector.

When calculating the following parameters, we used a normalized Pearson correlation coefficient, calculated according to [0, 1] as $0.5(1-R_{\alpha\beta}) \rightarrow R_{\alpha\beta}$. In this conversion we convert the correlation coefficient to a "distance". Additionally, we define:

R_{w_i} - The within-cluster dispersion is given in (4).

$$R_{w_i} = \sum_{j=1}^{J_i} R_{\phi_{ij}\mu_i} \tag{4}$$

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

WSL - within single linkage (Bolshakova and Azuaje, 2003):

The minimal Euclidean distance between two data points from the same cluster, defined in (5):

$$WSL = \min_i\{\min_{n\neq m}\{d(\phi_{in}, \phi_{im})\}\} \tag{5}$$

WCL - within complete linkage (Bolshakova and Azuaje, 2003):

The maximal Euclidean distance between two data points from the same cluster, defined in (6):

$$WCL = \max_{1\leq i\leq c}\{\max_{n\neq m}\{d(\phi_{in}, \phi_{im})\}\} \tag{6}$$

WAL - within average linkage (Bolshakova and Azuaje, 2003):

The average Euclidean distance between all pairs of data points from the same cluster, defined in (7):

$$WAL = \underset{i}{\text{mean}}\left\{\underset{n\neq m}{\text{mean}}\{d(\phi_{in}, \phi_{im})\}\right\} \tag{7}$$

BSL - between single linkage (Bolshakova and Azuaje, 2003):

The minimal Euclidean distance between two data points from different clusters, defined in (8):

$$BSL = \underset{i\neq j}{min}\left\{\min_{\phi_{in}\in C_i, \phi_{jm}\in C_j}\left\{d\left(\phi_{in}, \phi_{jm}\right)\right\}\right\} \tag{8}$$

BCL - between complete linkage (Bolshakova and Azuaje, 2003):

The maximal Euclidean distance between two data points from different clusters, defined in (9):

$$BCL = \underset{i\neq j}{max}\left\{\max_{\phi_{in}\in C_i, \phi_{jm}\in C_j}\left\{d\left(\phi_{in}, \phi_{jm}\right)\right\}\right\} \tag{9}$$

BAL - between average linkage (Bolshakova and Azuaje, 2003):

The average Euclidean distance between all pairs of data points from different clusters, defined in (10):

$$BAL = \underset{i \neq j}{\text{mean}} \left\{ \underset{\phi_{in} \in C_i, \phi_{jm} \in C_j}{\text{mean}} \left\{ d\left(\phi_{in}, \phi_{jm}\right) \right\} \right\} \tag{10}$$

BcenL - between centres linkage (Bolshakova and Azuaje, 2003):
The maximal Euclidean distance between the means of all pairs of clusters, defined in (11):

$$BcenL = \underset{i \neq j}{max} \left\{ d\left(\mu_i, \mu_j\right) \right\} \tag{11}$$

DB - Davies-Bouldin index (Bolshakova and Azuaje, 2003):

This index aims at identifying sets of clusters that are compact and well separate, with smaller values indicating a "better" clustering solution. The Davies-Bouldin index is defined in (12):

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

.

DUNN index (Bolshakova and Azuaje, 2003):

The Dunn index is defined in (13):

$$DUNN = \underset{1 \leq i,j \leq Q}{min} \left\{ \frac{R_{b_{ij}}}{max_{1 \leq k \leq Q} R_{w_k}} \right\} \tag{13}$$

Large values of the Dunn index correspond to a good clustering solution.

Han - Hartigan index (Tibshirani et al., 2001):

The Hartigan index is defined in (14):

$$Han(Q) = \left\{ \frac{W_Q}{W_{Q+1}} - 1 \right\} / (J - Q - 1) \tag{14}$$

where

$$W_Q = \frac{1}{2} \sum_{i=1}^{Q} R_{w_i}$$

The Hartigan index was originally defined to estimate the number of clusters, but in this study, we did not have different clustering results corresponding to an increasing number of clusters. Since we could not compare different clusterings, we calculated only W_Q as an input feature to the logistic regression.

KL - Krzanowski-Lai (Tibshirani et al., 2001):

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

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

where

$$DIFF(Q) = (Q-1)^{2/L} W_{Q-1} - Q^{2/L} W_Q$$

As in the case of the Hartigan index, a comparison between different clustering results was not possible. Therefore, we just used the coefficient $Q^{2/L} W_Q$ as an input feature.

Sep - Separation index (Chen et al., 2002):

The separation index is calculated as the weighted average between-cluster dispersion, defined in (16):

$$Sep = \frac{1}{\sum_{i \neq j} J_i J_j} \sum_{i \neq j} J_i J_j R_{b_{ij}} \tag{16}$$

This index reflects the overall dispersion between clusters, with higher values indicating superior clustering results.

4.2. Clustering evaluation criteria

The set of features presented in Section 4.1 can be used as an input vector to the logistic regression, trained with an objective function related to an external clustering quality measure. In this work, clustering quality was evaluated using the same criteria as defined in Ajmera et al. (2002). The concept involved calculating both the *average cluster purity* (*ACP*) and the *average speaker purity* (*ASP*). *ACP* measures the degree to which a cluster is limited to only one speaker, while *ASP* measures the degree to which a speaker is limited to only one cluster. In the ideal case, both *ACP* and *ASP* are equal to 1.0. The geometrical mean of *ACP* and *ASP*, *K*, is applied as an evaluation criterion to compare clustering systems. The formulation of the evaluation criteria is given in (17), where the notation is as follows:

- R - Number of speakers,
- Q - Number of clusters,
- n_{qr} - Total number of i-vectors in cluster q that are associated with speaker r,
- $n_{q.}$ - Total number of i-vectors in cluster q,
- $n_{.r}$ - Total number of i-vectors that are associated with speaker r.

$$\begin{aligned} ACP &= \frac{1}{Q}\sum_{q=1}^{Q} p_{q.} \quad ; \quad p_{q.} = \sum_{r=1}^{R} \frac{n_{qr}^2}{n_{q.}^2} \\ ASP &= \frac{1}{R}\sum_{r=1}^{R} p_{.r} \quad ; \quad p_{.r} = \sum_{q=1}^{Q} \frac{n_{qr}^2}{n_{.r}^2} \\ K &= \sqrt{ACP \cdot ASP} \end{aligned} \tag{17}$$

It is important to note that *ACP* is based on the cluster purities $\{p_{q.}\}_{q=1}^{Q}$, while *ASP* is based on the speaker purities $\{p_{.r}\}_{r=1}^{R}$. These values are not probabilities, as they do not sum to one.

4.3. Logistic regression

Logistic regression is a well known algorithm which producing a rating of ordinal data in the range [0,1], (Bishop, 2006). By applying the inverse logit function, a *K* value estimator was trained on the training set. The logistic expression (the inverse logit function) is given in (18):

$$f = \frac{1}{1+e^{-z}} \tag{18}$$

where $z = w^T\psi + b$, and ψ and w are the following column vectors:

ψ - input feature vector (in our case, the different statistical features extracted from the clustering process, described in Section 4.1), after mean substitution and variance normalization of each feature dimension.

w - the weights vector of the linear combination that is estimated on the training set.

f - the output of the logistic expression and is in the range [0,1]; it estimates the *K* value.

b - the bias.

The weights vector w was trained using N clustering trials, and computing for each trial n the corresponding pair $\{\psi_n, K_n\}_{n=1}^{N}$ including the internal cluster statistics vector ψ_n and external quality measure K_n. The applied optimization criterion was the *minimum mean squared error* (MMSE) between the clustering quality value K_n and the regression output f_n. The prediction quality of the estimator was tested on a separate evaluation dataset of size M. For each set of i-vectors, clustering was performed and ψ_m was calculated. The output f_m of the logistic regression was obtained and compared to the true clustering performance K_m. The error of the system was calculated over all M test trials using (19).

$$E = 100\sqrt{\frac{1}{M}\sum_{m=1}^{M}(K_m - f_m)^2} \tag{19}$$

5. Experiment and results

In Cohen and Lapidot (2017) it was shown that *ASP*, *ACP* and *K* can be reasonably estimated when the same clustering algorithm is used for training and evaluation. However, this is not always possible. The clustering algorithm may be changed for a variety of reasons, e.g., a better algorithm is developed or an algorithm that has poorer performance but is much faster, and performs adequately for some cases.

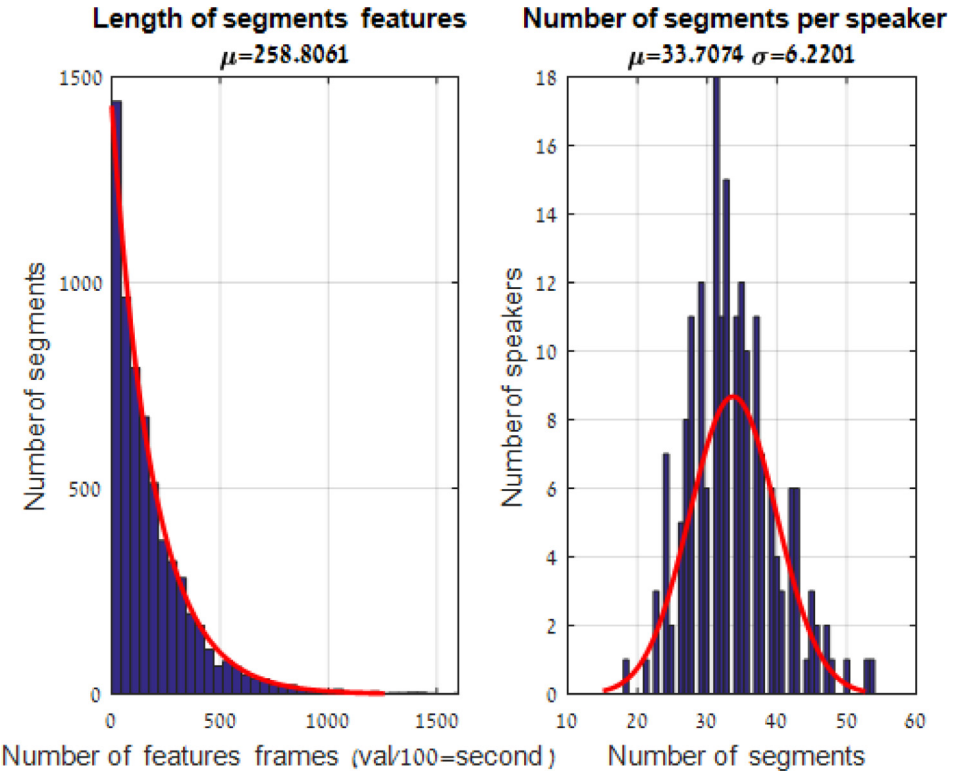

Fig. 2. The distribution of segment lengths (left) and the distribution of the number of segments per speaker (right) (Salmun et al., 2017).

In this work only the K parameter is under examination. It will be shown that with a proper training strategy, logistic regression can accurately estimate the K value using different clustering algorithms. It will be also shown that clustering quality estimator can be used as an estimator for the number of clusters in clustering algorithms that need this parameter to be predefined (such as VQ based-clustering).

5.1. Design of the datasets

All the experiments were carried out using the NIST 2008 Speaker Recognition Database (LDC Catalog https://catalog.ldc.upenn.edu). The test corpus short2-short3-Test7 was employed. Only male speakers were used for speaker clustering and clustering quality estimation. 98 speakers were assigned to the dataset used to train the logistic regression, while the remaining 90 speakers were used as the test dataset. Thus, the two datasets were statistically independent.

For each speech segment, the *Mel frequency cepstral coefficients* (MFCC) were extracted using a 25 ms Hamming window. 19 MFCC features together with log energy were calculated every 10 ms, following by cepstral mean subtraction and variance normalization. The features were augmented by delta and delta delta features, resulting in 60-dimensional feature vectors in total.

A male-only UBM of 2048 Gaussian mixture components was derived by training on the following protocols: Fisher Part ; Switchboard II, Phase 2 ; Switchboard Cellular, Parts 1 and 2 ; and NIST 2004–2006 SREs (LDC Catalog https://catalog.ldc.upenn.edu). A total variability matrix with a low rank (of 400) was also trained, using labeled data from the same databases as for the UBM.

The speech files consisted of 5 min of English telephone speech that we segmented into small segments. The minimum segment length L was $L_{\min}=0.7$ s, and the average length was $L_{av}=2.5$ s. The distribution of the segment length can be approximated using an exponential distribution:

$$L \sim L_{\min} + \exp(\lambda); \lambda = \frac{1}{L_{av} - L_{\min}}$$

The average number of segments per speaker was $\eta=34$ and the standard deviation was $\sigma=6.0$. The distribution of the number of segments per speaker S was approximately Gaussian, $S \sim \mathcal{N}(\eta, \sigma^2)$. The distributions are presented in Fig. 2.

For both the training and the test database, each clustering trial was designed by randomly choosing the number of speakers in the range 2 to 60, while the number of segments per speaker was chosen according to the distribution in Fig. 2. For each dataset 8000 trials were designed.

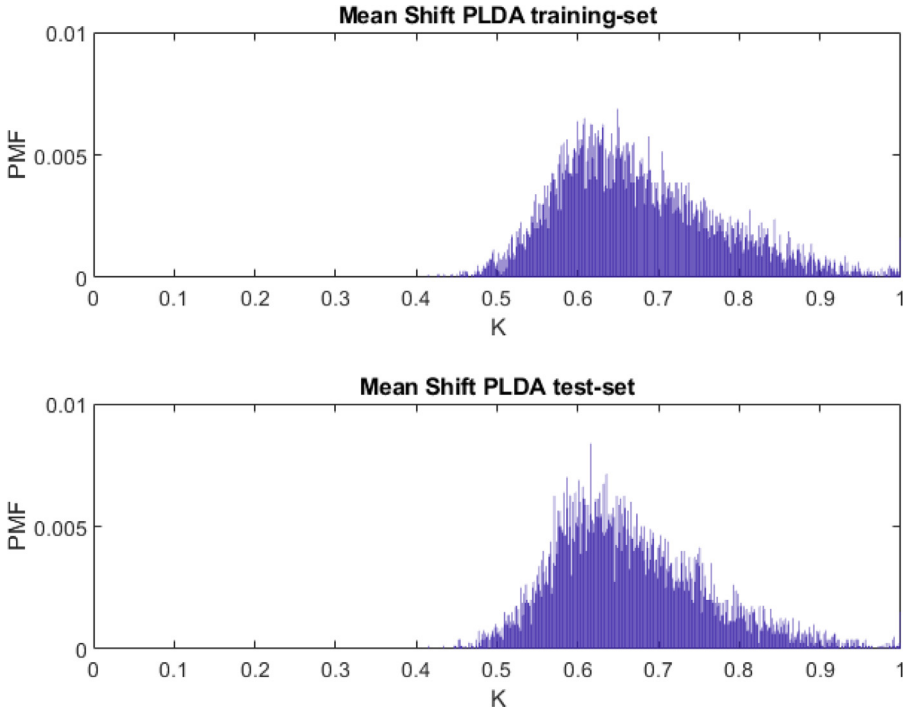

Fig. 3. The PMF of K for each clustering of the PLDA mean-shift training (upper sub-plot) and test (lower sub-plot) datasets.

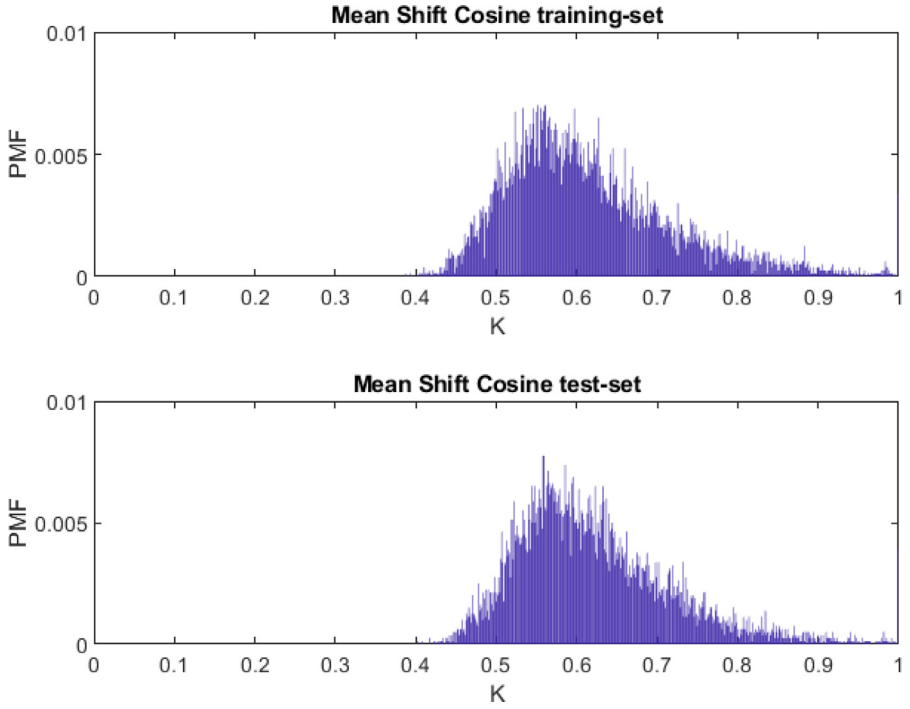

Fig. 4. The PMF of K for each clustering of the Cosine mean-shift training (upper sub-plot) and test (lower sub-plot) datasets.

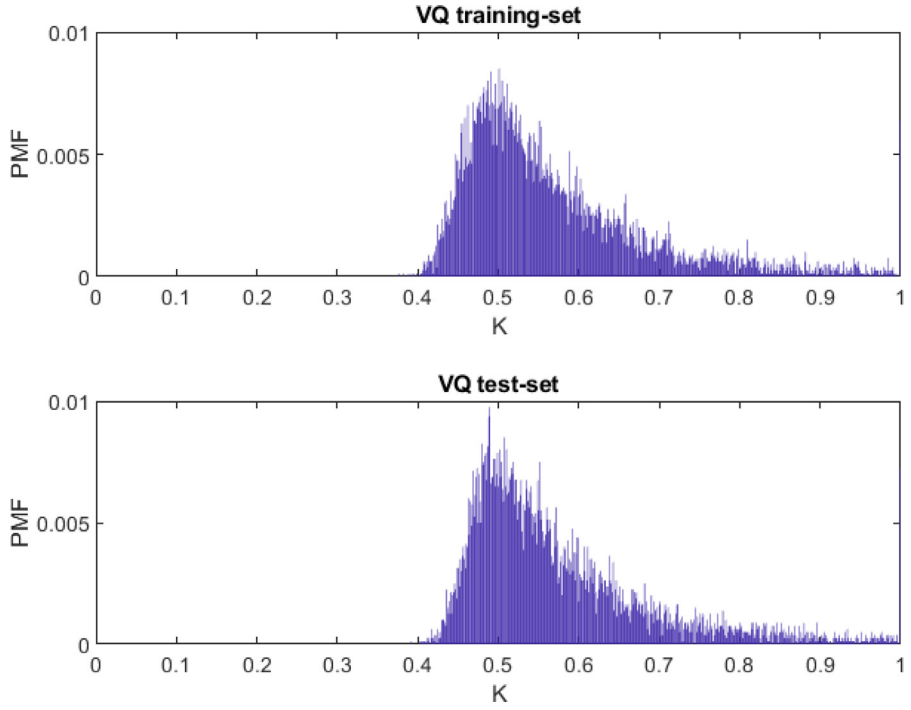

Fig. 5. The PMF of K for each clustering of the stochastic VQ training (upper sub-plot) and test (lower sub-plot) datasets.

Table 1
Speaker clustering quality estimation (MSE [%]) when the logistic regression is trained (rows) and tested (columns) on parameters obtained from the three clustering methods.

	Test		
Train	PLDA	Cosine	VQ
PLDA	5.71	38.94	43.83
Cosine	7.22	5.00	6.80
VQ	34.29	13.56	5.36

5.2. Baseline results

As was described in Sections 2 and 3, four clustering algorithms were used. We notate them as follows:

PLDA - PLDA-based mean-shift with normalized i-vectors,
Cosine - Cosine-based mean-shift with normalized i-vectors,
VQ - Stochastic VQ with normalized i-vectors.

Although the training and test databases contained different speakers, it was important to verify that the two datasets had similar distributions of K values. The comparisons between the *probability mass functions* (PMFs) for the three clustering algorithms are presented in Figs. 3–5. It can be seen that in all cases, there is close similarity between the training and the test K values. As VQ training required knowing the size of the codebook in advance, both for the train and the test phases, and as it was done on the same trials as the mean-shift, it was assumed that the codebook size is the same as has been estimated by the cosine-based mean-shift. In Section 5.4 we will show how to estimate the codebook size using the trained clustering quality estimator.

First we tested the clustering quality estimation for all combinations of training and test clustering methods; see Table 1. The results are calculated according to Eq. (19). As expected, when the logistic regression was trained and tested using the same clustering algorithm, the lowest MSE was obtained (the diagonal of the table). When estimating the clustering quality of **VQ**, with training carried out using the **Cosine** algorithm, the degradation in quality estimation was acceptable. Training on the cosine-based mean-shift data also resulted in acceptable estimation of the quality of **PLDA**. The results presented in Table 1 are the

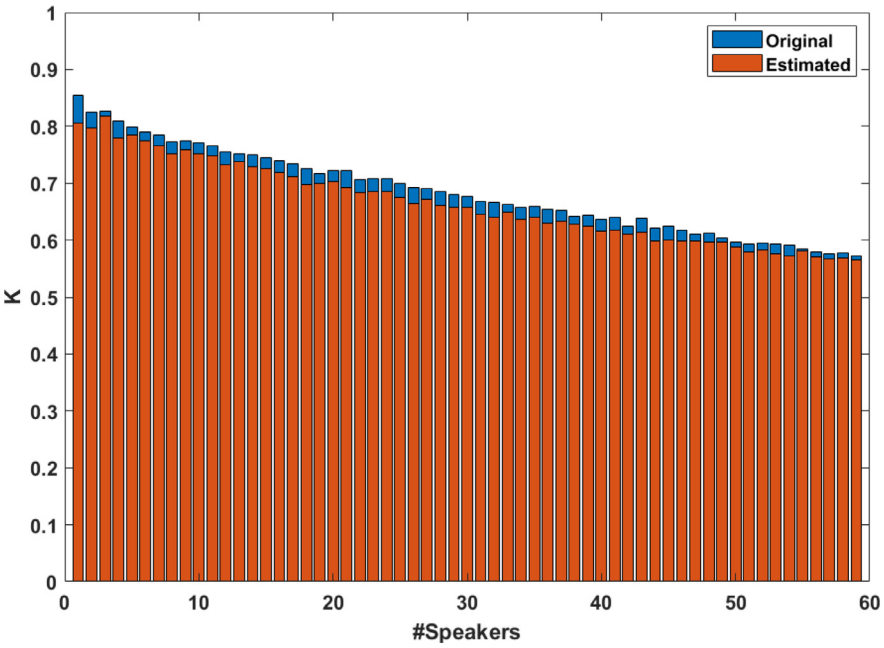

Fig. 6. True K value versus estimated K is presented as a function of the number of speakers in the **PLDA** case.

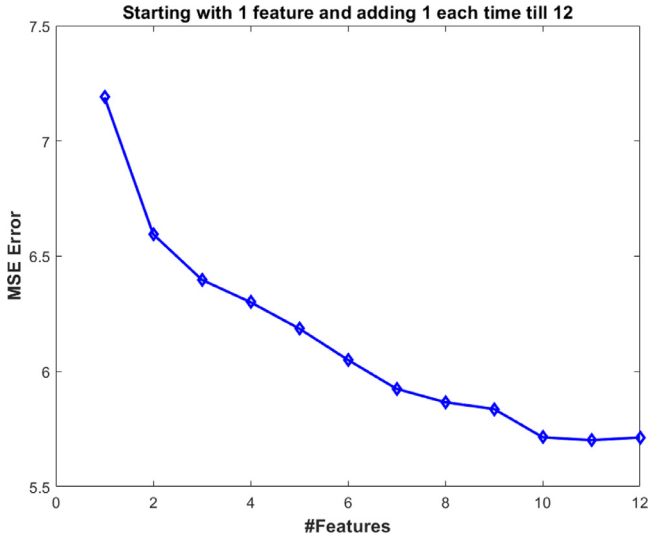

Fig. 7. K value as a number of features in the **PLDA** case.

average over the speaker numbers from 2 to 60. In Fig. 6, an example of true K value versus estimated K is presented as a function of the number of speakers in the **PLDA** case. The tendencies of the true and the estimated K values are similar; however, logistic regression under-estimate the true value between $4\% - 11\%$.

Another issue is the relevance of the features that are used for estimation. The results are presented in Fig. 7. We apply a greedy search, starting with the best feature and adding one feature each time. According to the results, it seems that the optimal number of features is between 10 and 12. In the rest of the experiments we will use 12 features.

The obtained results raise several questions: Does the quality depend on whether training is carried out using PLDA or cosine? Does the quality depend on whether training is carried out using mean-shift or VQ? To have a better understanding of the results we compare the distribution of K for each of the clustering methods. The PMFs are presented in Fig. 8. It can be seen that the K values for the **Cosine** system lie in between the values of the **PLDA** and the **VQ** systems, and hence the **Cosine** system generalizes these systems well. The results indicate that the optimization criterion and the precise clustering method are not the key factors

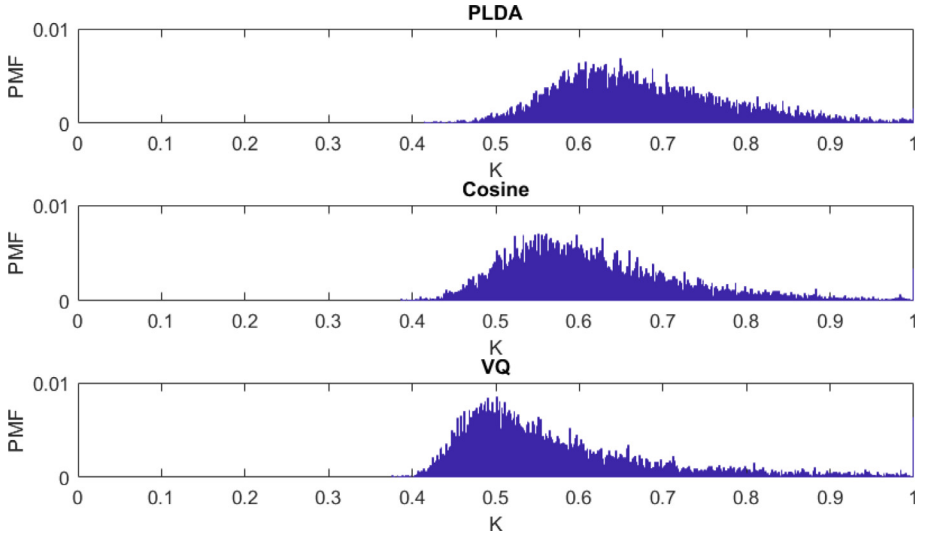

Fig. 8. The PMF of K values for each clustering method: Upper plot - **PLDA**; Middle plot - **Cosine**; Lower plot - **VQ** .

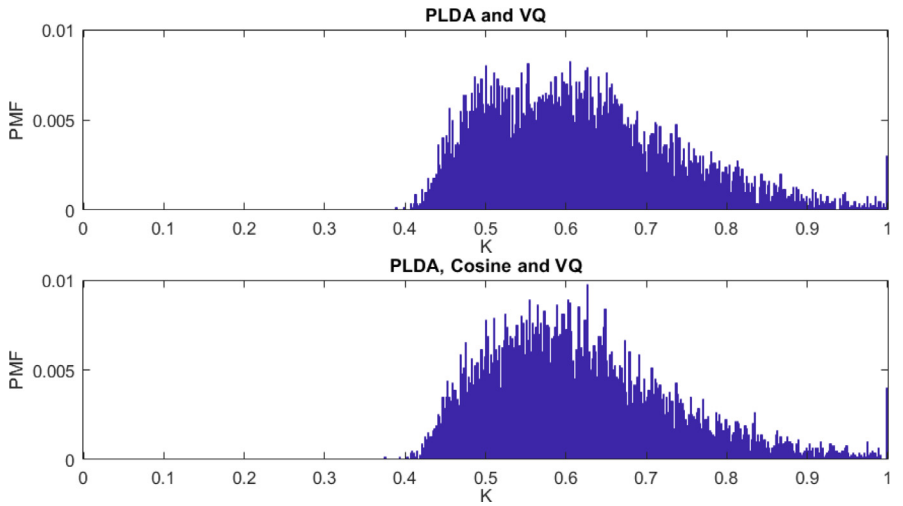

Fig. 9. PMFs of K from combined training sets: Upper plot - **PLDA** and **VQ**; Lower plot - **PLDA, Cosine** and **VQ**.

Table 2
Prediction error (MSE [%]) using multi-class training of the logistic regression.

Test set	PLDA and VQ	All	Best
PLDA	6.01	6.20	5.71
Cosine	7.51	5.49	5.00
VQ	6.36	6.19	5.36

for ensuring successful generalization. Rather, the important criterion seems to be the span of the K values: if they span the entire range, then logistic regression can learn to produce a reasonable estimate of the clustering quality.

5.3. Further generalization abilities

Following the experimental results of the previous subsection and, in particular, the PMFs in Fig. 8, a new experiment was conducted. We combined several training sets from different speaker clustering algorithms and tested whether the resulting logistic regression has improved generalization abilities relative to those of single-algorithm training. The PMFs are presented in Fig. 9. The upper plot shows the PMF corresponding to training sets from **PLDA** and **VQ**, the lower PMF corresponding to training

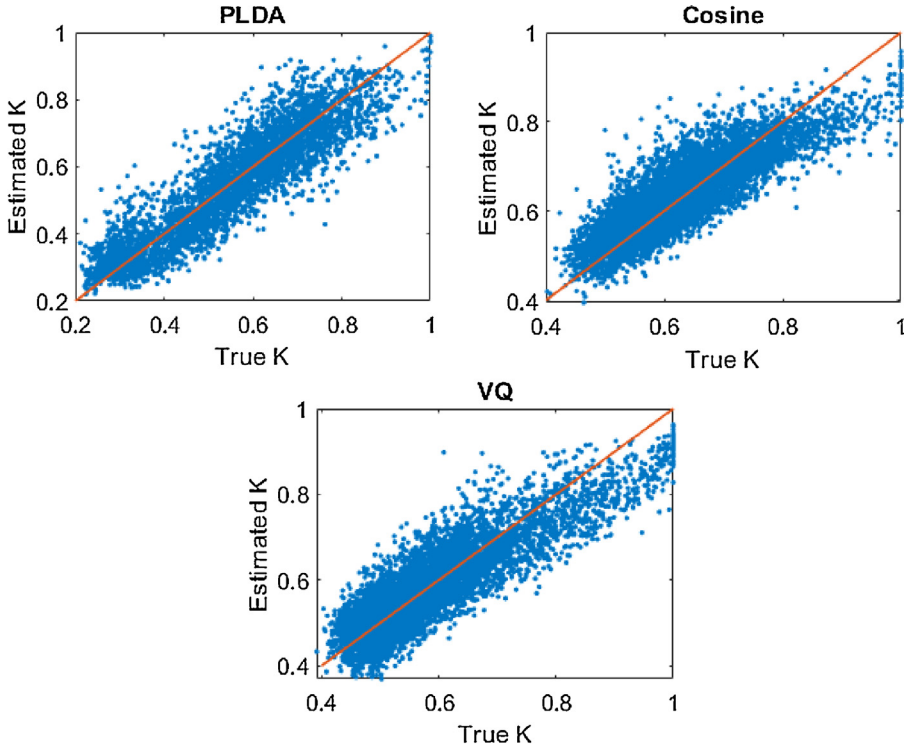

Fig. 10. True K vs estimated K values (in red the line $x=y$) for matched conditions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
PCC values of different systems vs training data type of the logistic regression.

		PLDA	Cosine	VQ
Training	**PLDA**	0.82	0.65	0.53
	Cosine	0.72	0.86	0.72
	VQ	0.74	0.83	0.88

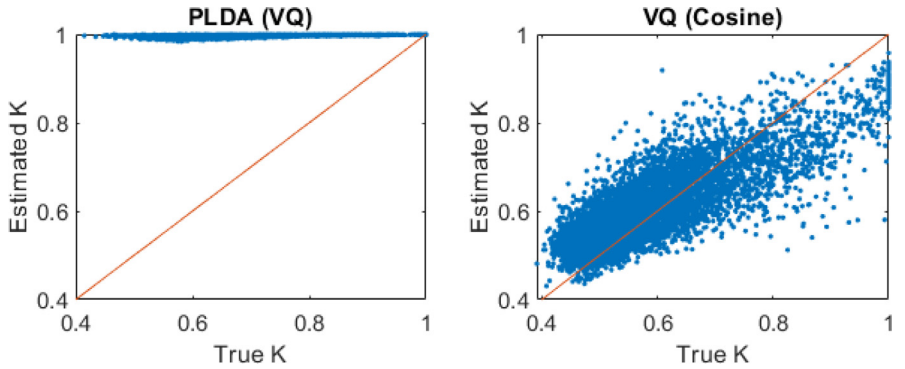

Fig. 11. Examples of true K values vs estimated K values (in red the line $x=y$) for unmatched conditions. The data conditions used for training Logistic Regression is given in parentheses. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sets from all the speaker clustering algorithms (**PLDA, Cosine**, and **VQ**). The combined training datasets were created by randomly choosing 8000 examples from the group of all datasets. This way, the effect on the generalization performances is only due to the change in the distribution of the data and not from enlarging the training dataset size. The data were taken from the original distributions as shown in Fig. 8. The clustering quality estimation results are shown in Table 2. The second column shows the results when the training process uses **PLDA** and **VQ** only. The third column shows the results when the training process uses

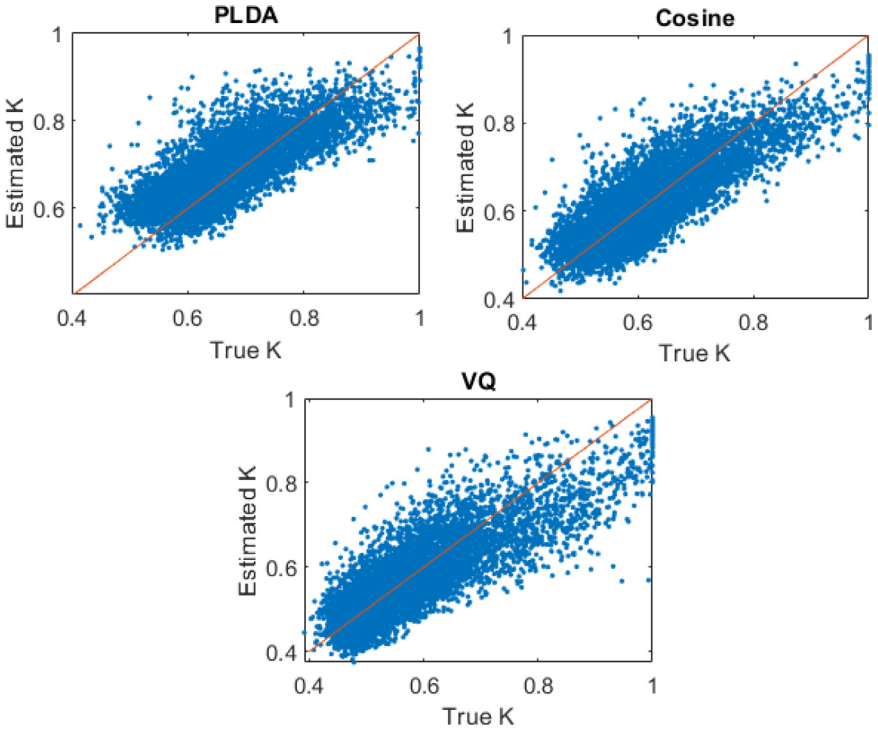

Fig. 12. True K values vs estimated K values (in red the line $x=y$), trained using the combined train data from all the clustering methods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
PCC values as a function of combined training data for the logistic regression.

PLDA	Cosine	VQ
0.78	0.83	0.84

Table 5
Speaker clustering quality estimation (MSE [%]) when the logistic regression is trained versus $\hat{K}$.

	Test			
Train	PLDA	normPLDA	Cosine	VQ
Logistic regression	7.29	5.71	5.00	5.36
$\hat{K}$	16.95	9.37	9.69	11.44

all datasets. The last column shows the optimal results, which correspond to matched training and test conditions. Comparing the second and third shows negligible difference except the **Cosine** results that improved by almost 27%, relative improvement. Comparing the second and the third columns with the last column shows that multi-class training shows promise for generalization.

In order to validate the results, we draw the graphs of true K values versus the estimated K values and calculate the sample *Pearson correlation coefficient* (PCC) between the true K values and the estimated K values. When there is a match between the training conditions, the expectation is that the true K versus estimated K lies on a line with slope equals 1 and PCC close to 1. The results are presented in Fig. 10. All the representations show the same behavior - there is a good match between the true and estimated K values. The same can be seen on the main diagonal of Table 3, which shows the PCC for the matched conditions. For all these cases, the PCC is good, which indicates that the estimated K values are highly correlated with the true K values. Part of the results for the non-matched conditions are presented in Fig. 11. The results can vary significantly. Moderate estimation is observed when VQ based clustering K is estimated with logistic regression, trained using data from Cosine-based mean shift clustering (right graph). On the other hand, a very bad estimation can be observed for estimating the K values of the PLDA-based clustering, while the training is performed based on VQ clustering (left graph). This is consistent with our previous observations that

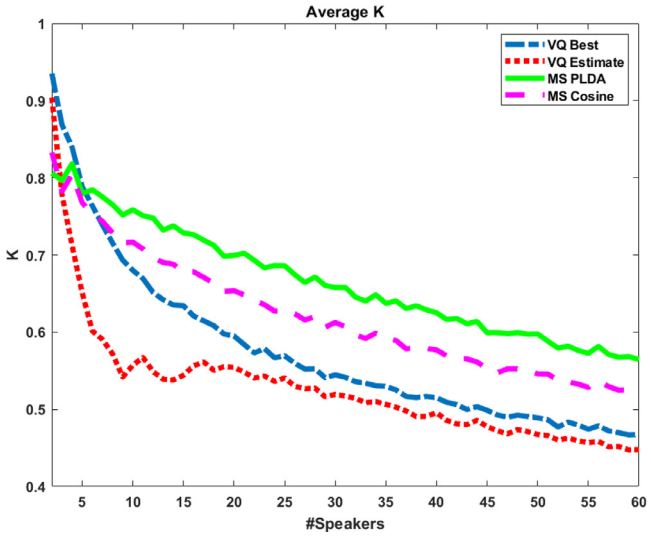

Fig. 13. The average K value as a function of the number of speakers.

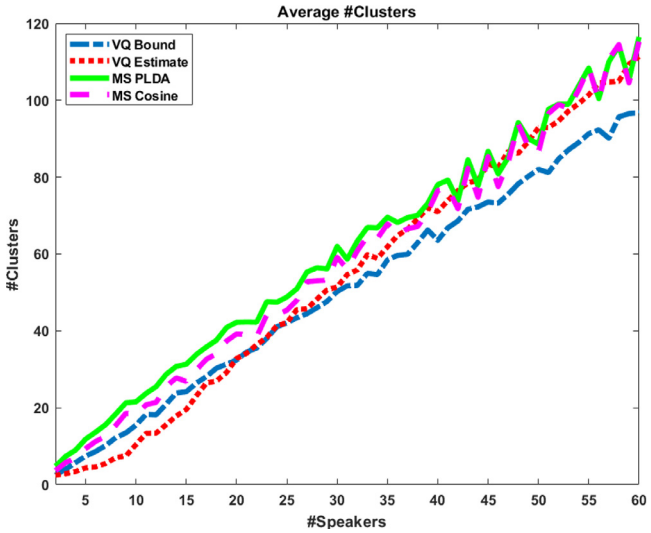

Fig. 14. The average number of estimated clusters as a function of the actual number of speakers.

when the PMFs highly overlap each other, the K estimation is better. The out of-diagonal values of PCC in Table 3 show a decrease in the cross-correlation as the PMFs are less overlapping each other.

At the last example the results are presented for the training data from all the clustering algorithms. The results are presented in Fig. 12. The PCC are summarized in Table 4. All the results are in accordance with Table 2. The training with all the algorithms well estimates the K values of new clusterings.

Until now, the results of the training performance were presented. One can argue that the results are not very good. It is possible to estimate the clustering quality with similar quality by using a very naive estimator $\widehat{K} = \frac{1}{M}\sum_{m=1}^{M} K_m$, which is the MMSE estimator of K using a constant. The comparison between the quality estimation using the logistic regression and the mean value are presented in Table 5. The worst case is when the K distribution is uniform over the range [0, 1] and then the error is 28.87%, however, the range is more narrow and the distribution is not uniform. As such, the square root MSE observed for the $\widehat{K}$ is much lower. Even though, logistic regression estimator improvement is between 39.06% to 56.74%. After saying that, the improvement in the prediction of the K is not the only important issue. As we saw in the previous figures, logistic regression estimator is highly correlated with the true K value. As we will show in the next sub-section, it is helpful to estimate the optimal number of clusters in clustering algorithms like VQ-based clustering. When using the estimated mean value, the estimator and the true value are uncorrelated. This means that it is impossible to learn anything about the true K from the estimator. We also observed that on

average, for every number of speakers, the estimator can be viewed as a pessimistic estimation, Fig. 6. It estimates K value a bit lower than the true one. It may be a good property as it is a kind of a lower bound on the real K value, while the actual K value may be a bit higher, but still close to the estimated value.

5.4. Performance evaluation of the stochastic VQ

For the VQ-based method, the number of clusters has to be defined prior to the VQ. However, this information is frequently unavailable and has to be estimated. Here, the quality estimator was used as a means of estimating the optimal number of clusters.

In this experiment, 8000 trials were performed. The number of speakers, number of segments per speaker and the segments themselves were all chosen randomly. Four systems were compared: Stochastic VQ with logistic regression, as described in Section 3; Stochastic VQ with the best calculated K value (this system was only included to provide a best-case comparison; it is not a practical system); Mean-shift with the cosine similarity measure; Mean-shift with the PLDA score.

The comparison is presented in Fig. 13, as a function of the number of speakers. As expected, the PLDA-score based mean-shift gave the best results, as the PLDA was trained in the same conditions as those under which it was tested. The next best performance was achieved by the cosine-similarity based mean-shift algorithm, which is known to be inferior to PLDA-based clustering (Salmun et al., 2017). The VQ-based clustering system, labeled as "VQ Estimated", could not compete with the mean-shift systems and showed consistently lower performance. When the performance of this system is compared with the best possible result, labeled as "VQ Best", it can be seen that the stochastic VQ closely approximates the upper bound, provided there are at least 20 speakers. As mentioned, the upper bound is a theoretical case, as it uses the number of clusters that yield the highest K value. However, in practice, this number of clusters is not known and has to be estimated. The finding that, for a low number of speakers, the estimates of the best K value are not reliable was unexpected and has not been observed in previous experiments.

It is also interesting to compare the different algorithms in terms of the number of clusters estimated. It is known that the mean-shift method usually over-clusters the data and produces many more clusters than the actual number of speakers. This comparison is presented in Fig. 14. The results are approximately the same for all methods, with the ratio of the number of clusters to the number of speakers averaging about 1.5. The stochastic VQ system, however, tends to estimate fewer clusters than the other systems at low numbers of speakers. When examining the VQ Bound, we see that its estimate of the number of speakers is superior to that of other systems when the number of speakers is high.

6. Conclusions

In this study, we presented a simple way of estimating the speaker clustering quality based on a logistic regression. It was shown that if the logistic regression is trained on the data from a particular clustering algorithm, the estimation results have a small mean square error. However, when the logistic regression is trained with data from one clustering algorithm, but applied to the output of another algorithm, the error increases dramatically. Examination of the K value distributions showed that each clustering algorithm spans a different range of values. When the logistic regression was trained using a combination of datasets from different clustering algorithms, the generalization significantly improved (Table 2). We assume that different clustering algorithms produce different types of error that cannot be captured by using a single clustering algorithm, however, training on several clustering algorithms together, allow logistic regression to make the generalization.

In the second set of experiments, the performance of the stochastic VQ system was investigated. For this system, the quality estimator was used as a means of estimating the optimal number of clusters. This is because the stochastic VQ requires the number of clusters to be stated in advance, and, if this number is not known, it should be estimated. The estimation of the number of clusters using logistic regression resulted in performance measures close to the upper bound of the stochastic VQ-based clustering. This can be seen in Figs. 13 and 14.

Although all the experiments were performed on a simulated data, the practical results on real data confirm the results we presented in this paper.

In future work we intend to investigate the possibility of improving both clustering quality estimation and generalizability using *deep neural networks* (DNNs); check the influence of regularization; and apply this approach to estimate the number of speakers for speaker diarization.

Declaration of Competing Interest

None

References

Ajmera, J., Bourlard, H., Lapidot, I., McCowan, I., 2002. Unknown-multiple speaker clustering using HMM. In Proceedings of ICSLP-2002, pp. 573–576.
Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O., 2012. Speaker diarization: a review of recent research. IEEE Trans. Audio Speech Lang.Process. 20 (2), 356–370.
Ben-Harush, O., Ben-Harush, O., Lapidot, I., Guterman, H., 2012. Initialization of iterative-based speaker diarization systems for telephone conversations. IEEE Trans. Audio Speech Lang.Process. 20 (2), 414–425.
Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 10 (2), 191–203.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.
Bolshakova, N., Azuaje, F., 2003. Cluster validation techniques for genome expression data. Signal Process. 83 (4), 825–833.
Chen, G., Jaradat, S.A., Banerjee, N., Tanaka, T.S., Ko, M.S.H., Zhang, M.Q., 2002. Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. Stat. Sin. 241–262.
Chen, S., Gopalakrishnan, P., 1998. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion.
Cohen, Y., Lapidot, I., 2017. Estimating speaker clustering quality using logistic regression. In: Proc. Interspeech 2017, pp. 3577–3581.
Cohen, Y., Lapidot, I., 2018. Robust speaker clustering quality estimation. 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), pp. 1–5. https://doi.org/10.1109/ICSEE.2018.8646164.
Cohen, Y., Lapidot, I., 2018. Speakers clustering with stochastic VQ and clustering quality estimator. 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), pp. 1–5.
Comaniciu, D., Meer, P., 2002. Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24, 603–619.
Figueiredo, M., Jain, A., 2002. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 24 (03), 381–396. https://doi.org/10.1109/34.990138.
Fukunaga, K., Hostetler, L., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inf. Theory 21 (1), 32–40 .
Hansen, M.H., Yu, B., 2001. Model selection and the principle of minimum description length. J. Am. Stat. Assoc. 96 (454), 746–774.
Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognit. Lett. 31 (8), 651–666.Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)
Kohonen, T., 1990. The self-organizing map. Proc. IEEE 78 (9), 1464–1480.
Lapidot, I., Shoa, A., Furmanov, T., Aminov, L., Moyal, A., Bonastre, J., 2017. Generalized Viterbi-based models for time-series segmentation and clustering applied to speaker diarization. Computer Speech Lang. 45, 1–20.
Pal, N.R., Pal, K., Keller, J.M., Bezdek, J.C., 2005. A possibilistic fuzzy c-means clustering algorithm. IEEE Trans. Fuzzy Syst. 13 (4), 517–530.
Salmun, I., Shapiro, I., Opher, I., Lapidot, I., 2017. Plda-based mean shift speakers' short segments clustering. Comput. Speech Lang. 45, 411–436.
Senoussaoui, M., Kenny, P., Stafylakis, T., Dumouchel, P., 2014. A study of the cosine distance-based mean shift for telephone speech diarization. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (1), 217–227.
Shapiro, I., Rabin, N., Opher, I., Lapidot, I., 2015. Clustering short push-to-talk segments. In: Proceedings of Interspeech 2015.
Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B 63 (2), 411–423.