Retraction

Retraction: Improving the performance of speech clustering methods using I-vectors (IOP Conf. Ser.: Mater. Sci. Eng. 1070 012066)

Published 17 January 2023

This article has been retracted by IOP Publishing following an allegation that this article may contain verbatim overlap with other published work and tortured phrases [2].

IOP Publishing has investigated in line with the COPE guidelines and have found the article contains significant similarities to the source without citation and agree this article should be retracted.

The authors neither agree/disagree to this retraction.

[1] R Suneetha Rani et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1070 012066
[2] https://pubpeer.com/publications/F6FACCBB45D35BBB0F30171AFBE89#

Retraction published: 17 January 2023
Improving the performance of speech clustering methods using I-vectors

R Suneetha Rani, G LV Prasad, A Prakash and A Ramesh
QIS College of Engineering and Technology, Ongole, AP-523272, India

Abstract. Identification of speaker for the speech segment is a challenging task in speech clustering. Similar segments (i.e., those segments are spoken by the same speaker) are grouped into a cluster. Thus, segment labelling is very important in speaker detection. Existing systems use the Gaussian mixture model (GMM) for deriving the GMM mean supervector of speech segments in segment labelling process. i-vectors are tougher than GMM mean supervectors and i-vectors are between GMM mean supervectors high-dimensional space and low-dimensional manifolds. Existing GMM based speech clustering method is modified by i-vectors for achieving an effective speech clustering method. In order to show the feasibility of the proposed work, Experiments are performed on real time speech datasets.

Keywords: Speech clustering, GMM, GMM mean supervectors, i-vectors.

1. Introduction

Speech clustering [1] is an unlabeled process and follows a two-step process, which is a) finding voice segment labels b) finding related segment classes. Improvement of segment labelling problem is attempted in this paper for obtaining of efficient speech clustering. Based on similarity characteristics of a predefined set of speech segments, this segment labelling [2] is carried out.

These characteristics of similarity are calculated using distance metrics. The distance metrics therefore play a key role in computing similarity characteristics. Acoustic modeling involves segment labelling of speech segments. Speech data modeling determines the features of voice data i.e. the mean and variance of the information is extracted. This modeling is carried out using a well-known statistical technique called the Gaussian Mixture Model. The GMM[3] is used for the derivation of mean GMM supervectors[4] and the variation of segments of expression. For each speech segment, the GMM mean supervectors are extracted and comparability between GMM mean speech segment supervectors is determined using distance metrics such as Euclidean, cosine, etc. Segment labelling is done using characteristics of similarity in the speech parts Later based on the labels of speech segments, clusters are formed.

This clustering process is carried out using traditional clustering methods such as k-mean [5], graph-based clustering [6]. The mean supervectors of the GMM are in high-dimensional form and are translated into low-dimensional manifolds. The i-vectors are between high-dimensional and low-dimensional intermediate vectors. This paper also addressed another issue of cluster tendency of set of
speech segments. The process of detecting the number of speakers of speech dataset is called as cluster tendency of speech dataset. The efficiency of speech clustering techniques also depends on the trend of clusters. The contributions in this document to the proposed work are summarized as follows:

1. To find the GMM model parameters of each and individual speech segments.
2. To extract the i-vectors of GMM mean supervectors.
3. To extract and find the similarity characteristics of i-vectors of speech segments.
4. To generate the clusters by proposed methods.
5. To conduct an empirical study to show the efficacy of the methods proposed.

Section 2 addressed the context of the work, Section 3 presented the proposed work, Section 4 displayed the experimental results and its discussion, and Section 5 presented the conclusion along with the possible scope of the paper. The remainder of the paper is structured as follows.

2. Related Work

In speech clustering, extraction of mel-frequency-cepstral coefficients (MFCC) [8] of speech datatinitiated and model parameters of GMM are derived for the processing of speaker identification. Currently, most of research in speech clustering is focused on solving the problem of cluster tendency. Bottom-up and top-down approaches [9], k-means, and other traditional clustering methods [10] are used for obtaining the clustering results of speech data. In the past 6 years, unsupervised acoustic model training has been developed using Gaussian mixture model (GMM), hidden Markov model (HMM)[11], discriminative clustering methods [12], Bayesian estimation of GMM or HMM [13]. In the existing system, most of the methods are use the Gaussian mixture model for defining models of speech segments. The GMM consists of m-number of components, so it is abbreviated as m-GMM. 64-GMM, 128-GMM, 256-GMM, 512-GMM, 1024-GMM etc. are the most used formats for GMM models.

\[
P(x|\lambda) = \sum_{i=1}^{m} w_i N(x|\mu_i, \Sigma_i)
\]

(1)

\[
P(i \mid X_t) = \frac{w_{oi} N(X_t|\mu_{oi}, \Sigma_{oi})}{\sum_{j=1}^{m} w_{oj} N(X_t|\mu_{oj}, \Sigma_{oj})}
\]

(2)

where \(x\) refers the d-dimensional vector, \(w_i\) can be prior probability with condition of \(\sum_{i=1}^{m} w_i = 1\). Now, the GMM is defined by its parameters \(\lambda = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{m}\). Thus, we focus on GMM to find model parameters that are solved by the technique of maximum likelihood estimation (MLE), namely the algorithm of expectation-maximization (EM). In Eqn, the GMM median supervisors are seen:

\[
S = [\mu_1^T \mu_2^T \mu_3^T \mu_4^T \ldots \mu_m^T]^T
\]

(3)

The supervectors of the equation ‘\(S\)’ are in high-dimensional form. To reduce the time complexity of speech clustering, these vectors need to be converted into low-dimensional shapes. There is a risk of losing the features data in this reduction of dimensionality. This paper is intended to use better alternative representation of i-vectors because of this fact. The time complexity of speech clustering, these vectors need to be converted into low-dimensional forms. In this reduction of dimensionality, there
is a risk of losing knowledge about the functions. Because of this fact, this paper suggests using better representation of alternative i-vectors. In terms of all components, these i-vectors compensate for the channel variations. This paper therefore addresses the issue of reduction of dimensionality and channel compensation in the proposed work by refining the current GMM-based bunching strategies with i-vectors, since i-vector depiction is a kind of minimized representation, and when compared to low-dimensional vector-based bunching results, these vectors support strong grouping results. In proposed bunching strategies, front-end factor analysis of i-vectors is used and it is applied in the following area.

3. Proposed Work

The method for generating i-vectors and the proposed scheme are discussed as follows.

3.1. i-vectors

The contribution of speech knowledge is converted into the organization of acoustic highlights. (i.e., in MFCC structure). For every 10ms, the MFCC is extracted over a window of 25ms, the GMM is utilized to view discourse data. The technique of factor inquiry is extended to GMM supervectors to assess the variability of the speaker and to compensate for channel irregularities [14]. This approach is referred to as a Total Variability Approach (TVA) and appears as follows.

\[ M = m + Tw \]  

Here, 'm' is a free supervector speaker and it is taken from the enormous GMM (called Universal Context Model-UBM), and 'M' meets the subordinate supervector. The UBM is prepared by a gigantic measure of speech data and is used to talk about the autonomous dissemination of acoustic highlights to the speaker [15]. For the complete inconstancy space, the word ‘T’ characterizes the low-position rectangular lattice, 'w' is a low-dimensional irregular vector and it is referred to as an absolute factor vector or an i-vector, representing a middle vector. The UBM is prepared for the free distribution of acoustic highlights through an immense measure of discourse knowledge and is used to talk to the speaker [15].

3.2. Dissimilarity Matrix Computation

The Euclidean and Cosine separation measurements are as often as possible utilized measurements and Cosine based uniqueness highlights are significantly prevailing in speaker segregation process [7]. Henceforth, this paper is proposed to utilize Cosine metric in speech grouping. The equation of Cosine separation metric is appeared in Eqn. (5).

\[
D(x, y) = 1 - \left( \frac{x \cdot y}{\|x\|\|y\|} \right)
\]  

This equation is given by the distance of cosine between two segments of speech x and y.

3.3. Finding Number of Involved Speakers

The quality of clusters is dependent on group inclination estimation. Deciding the amount of speech data included by speakers is known as a community pattern tendency. The Visual Access Tendency [19] is one of the well-non-parametric visualized methods used to identify the number of speakers from different highlights of discourse results.

3.4. Enhanced Clustering Methods

The clustering alludes to the way the extracted comparable i-vectors are collected together, i.e., one category includes i-vectors of a related speaker's speech fragments. The number of speakers in this article is regulated by VAT. In the following equation, the proposed ‘i-vectors VAT dependent k-implies (iVK)’ is represented.
Algorithm: iVK

Input: \( SS = \{S_1, S_2, \ldots, S_n\} \), set of ‘n’ no.of Speech segments
Output: \( k \)-no.of speakers
        \( C = \{C_1, C_2, \ldots, C_k\} \), \( k \)-clusters

1. Train UBM with an immense amount of voice data
2. Build the GMM models for speech segments of ‘SS’
3. Adapt the UBM into GMM and generate GMM mean supervectors
   \( S = [\mu_1^T, \mu_2^T, \mu_3^T, \ldots, \mu_m^T]^T \)
4. Find i-vectors using Eqn. (4)
5. Compute the Dissimilarity Matrix ‘D’ of obtained i-vectors (’w’)
6. \( k = \text{VAT}(D) \) // The VAT procedure is in [16]
7. Discover the clustering results of \( k \)-means at obtained ‘k’ value from step 6. The results of \( k \)-means is ‘C’

In iVK, initially UBM is built and it is adapted with GMM of speech segment to estimate the accurate model parameters of speech segments in Step 2. Step 3 shows the generation of GMM-UBM based supervectors. In Step 4, the i-vectors are generated using TVA. Step 5 shows the cosine based dissimilarity matrix computation. Assessment of clusters is performed by VAT in Step 6. The effects of clustering are discovered using k-means in Step 7.

Similarly, using i-vectors and VAT methods, the MST-based clustering technique is extended, and this suggested method is called iVM. These suggested methods work effectively to shape the number of speakers and discover the effects of speech clustering.

4. Experiment Results and Its Discussion

TSP speech datasets are the subject of the experiments and this dataset can be found in [16]. This dataset contains the set of speech utterances of different speakers in .wav files format. The experimental descriptions of the dataset are detailed in Table 1. In the MATLAB setting, the proposed methods are established.

In both methods i.e. iVK and iVM, i-vectors of GMM mean supervectors are derived. This dataset contains for different speakers a list of speech utterances in the .wav file format. The standard k-means and MST-based methods of clustering discover the effects of clustering. using GMM Mean super vectors. However, the proposed method uses the i-vectors for producing of quality of clusters.

| Description                                | Training set | Test set |
|--------------------------------------------|--------------|----------|
| No. of speakers                            | 40           | 5        |
| No. of speech statements (or speech fragments) | 4000         | 200 (max) |
| No. of statements of each speaker          | 100 to 200   | 30 to 40 |
| Statement (.wav file) duration             | 10 to 20 sec | 10 to 20s |
Two clustering metrics, such as clustering accuracy (CA), normalized mutual knowledge (NMI) [17], are used to test the output results of iVK and iVM. The comparative analysis between conventional and suggested clustering approaches is shown in Table 2 and Table 3.

**Table 2**: Outcomes of k-means and proposed methods of clustering

| Speakers data | k-means | i-VK |
|---------------|---------|------|
|               | CA      | NMI  | CA   | NMI  |
| 2-speakers data | 0.71    | 0.65 | 0.85 | 0.76 |
| 3-speakers data | **0.80** | 0.55 | 0.81 | 0.58 |
| 4-speakers data | 0.65    | 0.55 | 0.66 | 0.58 |
| 5-speakers data | 0.66    | 0.60 | 0.70 | 0.61 |
| 6-speakers data | 0.55    | 0.44 | 0.58 | 0.46 |

**Table 3**: Results of MST-based-clustering and proposed clustering method

| Speakers data | MST-clustering | i-VM |
|---------------|----------------|------|
|               | CA      | NMI  | CA   | NMI  |
| 2-speakers data | 0.55    | 0.43 | **0.62** | **0.45** |
| 3-speakers data | **0.77** | 0.56 | 0.76 | **0.58** |
| 4-speakers data | 0.55    | 0.47 | **0.58** | 0.48 |
| 5-speakers data | 0.46    | 0.40 | **0.47** | 0.44 |
| 6-speakers data | 0.52    | 0.43 | **0.55** | 0.45 |

Table 2 presents the discourse bunching results for testing of Two speakers, Three speakers, Four speakers, Five speakers, and Six speakers grouping results utilizing proposed iVK. It likewise presents the similar examination between k-means and iVK techniques. The k-means utilizes the GMM mean supervectors and iVK utilizes the i-vectors. From Table 2, it perceived that critical improvement advanced in iVK than k-means as for two performance measures, clustering accuracy (CA) and normalized mutual information (NMI).

The comparative analysis of MST-based clustering and iVM is shown in Table 3. It was perceived from the CA and NMI estimates of Table 3 that the proposed technique (iVM) achieves more precision than the clustering approach based on MST. Fig. 1 illustrates performance assessment between the two proposed iVK and iVM methods, in which iVK clustering accuracy is greater than iVM. Using Fig 2. Comparative perception is made regarding the measure of NMI efficiency It is henceforth assumed that iVK and iVM are carried out in a way that is better than k-means and MST-based methods of clustering, and iVK is more suitable than iVM.
Fig. 1 as well as Fig. 2 shows an experimental review of current and proposed clustering accuracy strategies and structured mutual knowledge. It is observed from these findings that iVK is better than iVM strategies.

Figure 1. Clustering Accuracy for Proposed Methods

Figure 2. Normalized Mutual Information for Proposed Methods

5. Conclusion and Future Enhancements

This paper discussed two issues of speech clustering by suggested approaches, such as cluster tendency and efficient clustering based on i-vectors. The latest conventional clustering method produces clustering results based on GMM without knowing the previous data about the number of speakers included. However, it is the key issue and progressively solved using VAT method and better speaker’s discrimination is performed using i-vectors in the proposed framework of the system. The proposed structure for the noise-free speech data clustering scheme is further elaborated in this paper.
References

[1] Roberto T and Daniel P 2011 An Overview of Speaker Identification: Accuracy and Robustness Issues *IEEE circuits and systems magazine* pp 23-61.

[2] Haipeng W, Tan L, Cheung-chi L and Bin M 2015 Acoustic Segment Modeling with Spectral Clustering Methods *IEEE Trans. on Audio, speech, and Lang. Proceedings* pp 264-277.

[3] Chu S, Tang H and Haung T 2009 Fisher Voice And Semi-Supervised Speaker Clustering *Proc. IEEE Int. Conf. on Acoustic speech and signal processing* pp 4089-4092.

[4] Douglas A, Reynolds, Thomas F, Quatieri, Robert B and Dunn 2000 Speaker Verification Using Adapted Gaussian Mixture Models *Digital signal processing* pp 19–41.

[5] Han J and Kamber M 2011 Data Mining: Concepts and Techniques, 3rd Edn. (Morgan Kaufmann, San Francisco).

[6] Arun K and Pujari 2001 Data Mining Techniques (Universities Press).

[7] Mohammad S and Patrick K 2014 A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization *IEEE/ACM Trans. on Audio, Speech, and Language Processing* pp 217-227.

[8] Tang H and Chu S M 2012 Partially supervised speaker clustering *IEEE Trans. Pattern Anal. Machine Intell*, pp 959-971.

[9] Senoussaoui M, Patrick K, Themostastafylakis and Pierre D 2013 Efficient Iterative Mean Shift Based Cosine Dissimilarity for Multi-Recording Speaker Clustering *Proc. ICASSP* pp 7712-7715.

[10] Jain A K and Dubes R C 1988 Algorithms for Clustering data, *Prentice Hall*.

[11] Varadarajan B, Khudanpur S and Dupoux E 2008 Unsupervised Learning of Acoustic Subword Units, *Proceedings of ACL-08: HLT* pp 165-168.

[12] Anguera X 2012 Speaker Independent Discriminant Feature Extraction for Acoustic Pattern Matching, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[13] Lee C Y and Glass J 2012 A Nonparametric Bayesian Approach To Acoustic Model Discovery Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp 40-49.

[14] Dehak N, Kenny P, Dehak R, Dumouchel P and Ouellet P 2010 Front-End Factor Analysis For Speaker Verification, *IEEE Trans. Audio, Speech, Lang. Process*, pp 788–798.

[15] Stephen H, Shum, Najim D, Reda D, James. R and Glass 2013 Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach, *IEEE Trans. on Audio, Speech, and Language Processing* pp 2015-2026.

[16] http://www.mmsp.ece.mcgill.ca/Documents/Data/

[17] Cai D, He X and Han J 2015 Document Clustering Using Locality preserving indexing, *IEEE Transactions on Knowledge and Data Engineering*, pp 1624-1637.

[18] Sangeeta B, Johan R and Koichi S 2015 Autonomous Selection of I-Vectors For PLDA Modeling in Speaker Verification, *Elsevier Science*, pp 32-46.

[19] Liang W, Xin G, James B, Christopher L and Ramamohanarao K 2010 Enhanced Visual Analysis for Cluster Tendency Assessment and Data Partitioning *IEEE Trans. on Know. and Data Engg.*, pp 1401 – 1414.

[20] Suneetha Rani T and Krishna Prasad M.H.M 2018 Access the number of speakers through Visual Access Tendency for Effective Speech Clustering, *International Journal of System Assurance Engineering and Management* pp 559–566.

[21] Suneetha Rani T, Anjaiah P and K.Keerthi 2018 Unification of Artificial Intelligence and Big Data using NLP, *Journal of Advanced Research in Dynamical and Control Systems*, pp 1744- 1751.
Acknowledgments

The authors would like to thank the management team of QIS College of Engineering and Technology, Ongole, Andhra Pradesh, India for providing excellent resources for completing this research work.