Global Journal of Enterprise Information System

DOI: 10.18311/gjeis/2018/16126

Feature Matching Techniques for Speaker Recognition

Pardeep Sangwan*

Department of ECE, Maharaja Surajmal Institute of Technology, New Delhi, India; sangwanpardeep@gmail.com

Abstract

Speaker recognition is a stream of biometric authorization which deals with the automatic identification of individual person using some inherent characteristics of that individual. The last stage of this system is the classification of feature templates generated during the previous stage i.e. feature extraction. This classification stage, also known as feature matching, provides the final decision about the speaker under observation. Hence, it is most important to use appropriate feature matching technique to get the accurate result. There are numerous feature matching techniques which can be used for the purpose. The present work provides an analysis of the various feature matching techniques used in the final step of a speaker recognition system. These techniques can be categorized in Statistical techniques, Soft-computing techniques and hybrid techniques. Statistical techniques include: "Vector Quantization (VQ), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) etc.", while Soft-computing techniques are "Artificial Neural Network (ANN), Support Vector Machine (SVM) and Fuzzy logic etc." Hybrid techniques make use of both the above said techniques.

Keywords: Artificial Neural Network (ANN), Feature Matching, Speaker Recognition, Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Vector Quantization (VQ)

Paper Code (DOI): 16126; **Originality Test Ratio:** 08%; **Submission Online:** 21-May-2017; **Manuscript Accepted:** 29-May-2017; **Originality Check:** 30-May-2017; **Peer Reviewers Comment:** 19-July-2017; **Double Blind Reviewers Comment:** 25-Sep-2017; **Author Revert:** 21-Dec-2017; **Camera-Ready-Copy:** 09-Jan-2018; **Editorial Board Excerpt:** 23-Jan-2018.

Editorial Board Excerpt: *At the Time of Submission (ToS) submitted term paper had a 08% plagiarism which is a terrific suggestion as far as originality report is concerned and falls under an accepted percentage for publication. The editorial board is of an assessment that paper had a consequential seal watch by the blind reviewer's which at a while stages had been rectified and make changes by an author in a range of phases as and when required to do so. The reviewer's had in an initial stages comment with minor revision with a following remark which at a short span reorganized by an author (pardeep sangwan). The rationale related to references, conceptual and body text is perceptible both subject-wise and research wise by the reviewers during inference and further at blind review progression too. All the comments had been communal at a diversity of dates by the authors' in due course of time and same had been integrated by the author in accumulation. By and large all the perspective and reviewer's comments had been integrated in a script at the end and further the paper had been earmarked and decided under* ***"Research Thought"*** *kind as its highlights and draw concentration to the work in relation attribute Matching Techniques for Speaker Recognition*

1. Introduction

Speech is the most important mean of exchanging information among humans. Hence, speech may also be primary mean of exchanging information between humans and machines. Speech recognition is related to the extraction of the linguistic message in the uttered speech while speaker recognition is identifying a person who is speaking. Speaker Recognition (SR) is a stream of biometric authorization which deals with the automatic identification of individual person using some inherent characteristics of that individual. The final step of a SR system is the classification of unknown and known speaker models generated from the information gathered during feature extraction. This classification stage, also known as feature matching, provides the final decision about the speaker under observation. Hence, it is most important to use appropriate feature matching technique to get the accurate result. There are numerous feature matching techniques (Figure 1) which can be used for the purpose. In the present paper some important feature matching techniques are detailed. Generally used speaker recognition models are codebook model, artificial neural network model, statistic model, and template model[1].

Speaker can be recognized with the help of various techniques. Major categories of these techniques are:

- Statistical techniques.
- Soft Computing techniques.
- Hybrid techniques.

*Author for correspondence

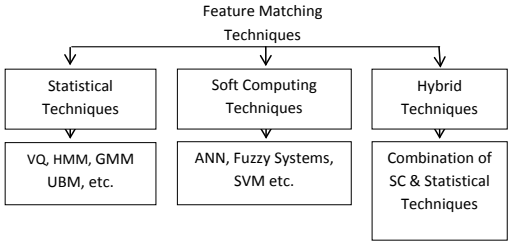

Figure 1. Feature matching techniques.

In section 2, statistical feature matching techniques are discussed and section 3 presents soft-computing techniques for the task. Section 4 gives hybrid techniques and the paper is concluded in the next section.

2. Statistical Techniques

These techniques include 'HMM', 'GMM', 'UBM', 'VQ' and many more. Some are explained in detail here.

2.1 Vector Quantization

VQ model was proposed for speaker recognition in 1980s[2] and it was originated basically from data compression. VQ is a very simple text-independent speaker model. Primarily VQ is utilized for ensuring the computation at fast rate and for light-weight practical implementation[3]. But competitive accuracy can also be achieved by combining VQ with background-model adaptation. The average quantization distortion can be given as:

$$D_Q(X,\dot{R}) = \frac{1}{T}\sum_{t=1}^{T} \min_{1\le n\le N} d(x_t, r_n) \quad (1)$$

Where X = {x_1, x_2, ..., x_T} and $\dot{R}$ = {r_1, r_2, ..., r_N} are test utterance feature vector and reference vectors respectively. D (x,r) is the Euclidian distance $\| x_t - r_n \|$. Smaller is the value of D_Q higher is the likelihood that X and $\dot{R}$ originates from same speaker. Hypothetically, all the training vectors can be used as reference vector , but to reduce computational complexity some clustering method is used to reduce number of vectors. For instance, K-means method can be used to get a reduced set of vectors (codebook). Moreover, optimization of codebook size is important than clustering method. Vector quantization is basically reduces the size of feature vector by mapping vectors of a larger distribution to smaller number of areas in the space. Every mapped area is known as a "cluster" and may be classified by the center of the area known as "*a code-word*". These "code-words" for all the clusters is known as "*Codebook*".

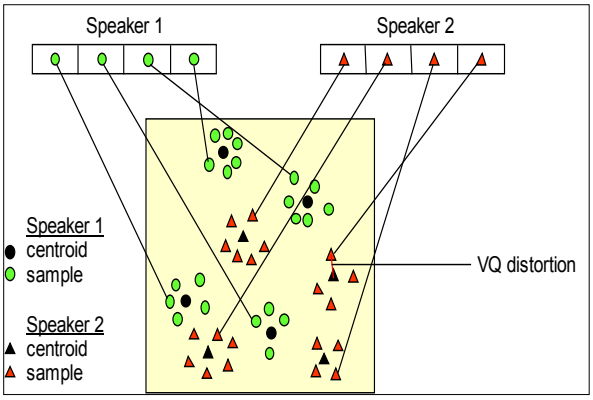

Figure 2. Vector quantization codebook formation.

Figure 2 illustrates a typical speaker identification task. It shows merely a two dimensional space representing the acoustical space of single speaker each. Small 'circles' of green colour are representing "Speaker 1" and red 'triangles' are for "Speaker 2". A vector quantized code book is obtained for all the speakers known to the system by utilizing this "clustering algorithm" during the "training phase" of the SR system. These generated "code-words" known as "centroids" are represented in Figure 2 by "black circles and black triangles" for respective speaker 1 and speaker 2. The discrimination of both the speakers can be done on the basis of centroid's locations. The difference in positions of a vector and the nearest code-word of "codebook" is known as "VQ-distortion". Then, in the next stage i.e. "recognition phase", speech sample of an unknown speaker is "vector-quantized" utilizing previously trained "codebook" and resulting VQ-distortion is calculated. The speaker having lowest distortion is declared as recognized speaker of input speech sample.

2.2 Gaussian Mixture Model (GMM)

This method is an extension of vector quantization model having overlapped clusters. It means that a feature vector have a non-zero probability of evolving from individual clusters. GMM has become a base method for robust speaker recognition[4]. A GMM is represented by "λ" and it's "Probability Density Function (PDF)" as:

$$p(x \mid \lambda) = \sum_{i=1}^{K} P_i G\left(x \mid \mu_i, \sum_i \square\right) \quad (2)$$

where K is no. of "Gaussian components", P_i is "prior probability" of i^{th} component and,

$$G\left(x \mid \mu_i, \sum_i \square\right) = (2\prod)^{-\frac{d}{2}} \mid \sum_i \square \mid^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T \sum_i \square^{-1} (x-\mu_i)\right\} \quad (3)$$

is a "d-variate Gaussian density function" having "mean vector μ_i", and "covariance matrix $\sum_i$". Also $P_i \geq 0$ are constrained as $\sum_{i=1}^{K} P_i = 1$. A huge amount of training data is the requirement for estimating the parameter of a full covariance GMM and this process is also computationally expensive. Thus, the diagonal covariance matrices of GMM are generally used to align the principle axes of the Gaussian ellipse with the coordinate axes as it reduces the computational complexity. To train a GMM, the parameters $\lambda = \left\{ P_i, \mu_i \sum_i \right\}_i^K = 1$ are estimated from a sample $X = \{x_1, x_2, \ldots\ldots\ldots, x_T\}$. Generally ML (Maximum Likelihood) estimation is used. The "average log likelihood of X with respect to λ" can be given as:

$$\text{LLavg}(X, \lambda) = \frac{1}{T} \sum_{i=1}^{T} \log \sum_{i=1}^{K} P_i G\left(x_t \mid \mu_{\cdot}, \sum_i \right) \quad (4)$$

This value indicates whether the unknown vectors are evolved from λ or not. For a given data likelihood may be maximized utilizing "Expectation Maximization (EM) algorithm"[5]. EM algorithm may be initialized using K-means and only few EM iterations are required. Estimation of optimal number of EM iterations is very important for a given task. Research has shown that separate model for male and female speaker have better performance than a single model for both. The adaptation of the new speaker model during the enrolment is carried out with the background model of the respective gender.

There are several methods used for adaptation out of which most important are "Maximum A Posteriori (MAP)" and "Maximum Likelihood Linear Regression (MLLR)". The adaptation method is selected on the basis of available amount of training data. MAP is applied where hundred hours of "training data" is used while MLLR is more effective for short utterances of few seconds. The matching each frame with others in GMM requires intensive computations. In GMM-UBM model, the score (13) is calculated fast by determining the top-C (generally C=5) scoring Gaussian from UBM for individual test utterance[6].

Additional techniques for fast computation includes Gaussian component evaluations, reduced number of vectors, and speaker models. In Hidden Markov Model (HMM) phonetic information is used for speech recognition but in GMM no such information is used explicitly and all "spectral features" of separate phonetic classes are combined to form training set for GMM. Due to this reason, test feature is phonetically misaligned with Gaussian component and it could bias the match score.

The problem of mismatching of phoneme is elaborated with phonetically motivated tree structure and an independent GMM for different phonetic classes. For example, P-GMM (phonetic-GMM) utilizing a "Neural-Network Classifier" for broad phone classes from 11 different languages is described in[7].

3. Soft Computing Techniques

These include "Artificial Neural Network (ANN)", "Genetic Algorithm (GA)", "Support Vector Machine (SVM)" and "Fuzzy Systems" etc.

3.1 Artificial Neural Network

It is used in several pattern classifying applications. It has the advantage that feature extraction and pattern matching can be performed with one artificial network which enables the simultaneous optimization of both. It is also very easy to combine different subsystems using ANN[8]. Artificial neural network is network of computing "neurons", and represents "parallel-distributed processing" structure. ANN is inspired by the bio-logical structure of human brain which is made up of the neurons. An important property of NN is its capability of approximating an arbitrary non-linear function. As "Artificial Intelligence (AI)" requires typically higher capability of taking non-linear decisions, NN could be a better choice in AI. ANN contains nodes, commonly arranged in different layers, and connections are made with the help of "weight elements", known as synapse. At all nodes, "weighted inputs" are "aggregated", "thresholded", and applied to "activation function" for generating output of a particular node. This process is illustrated in the Figure 3.

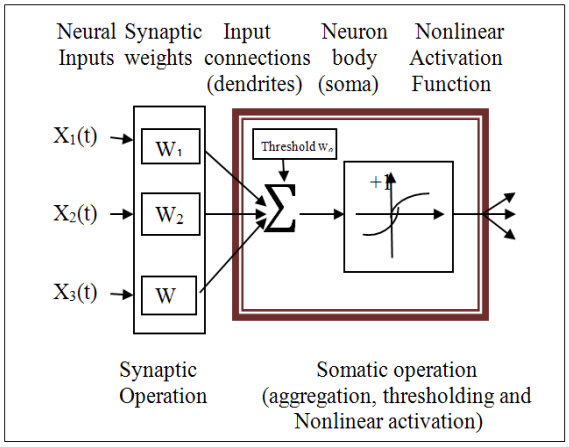

Figure 3. Operation at a node of NN.

3.2 Support Vector Machine

SVM is probably the most powerful classifier for speaker recognition systems. SVM can increase the accuracy when combined with GMM. SVM is a robust discriminative classifier which is equally applicable on "spectral", "prosodic", and "high-level" features. Furthermore, this technique is advantageous due to its ability to classify unseen data. It is a *binary* classifier modelling a decision boundary (*Separating Hyperplane*) between two classes as shown in Figure 4.

In case of verification, training vector of reference speaker is kept in one class and may be labelled as +1 while the training vectors of background (imposter) population are contained in second class labelled as "1". These labelled vectors are used to find a hyperplane which maximize the "margin of separation" of both classes. Discriminative function of SVM can be defined as[9]:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + d \quad (5)$$

Here $t_i \in \{+1, -1\}$ is the ideal output value, $\sum_{i=1}^{N} \alpha_i t_i = 0$ and $\alpha_i > 0$. x_i (Support vector), α_i (weight of x_i) and d (bias term) are obtained from training data by some optimization process. K (kernel function) can be given as $K(x, x_i) = \varnothing(x)^T \varnothing(x_i)$, where ø is mapping of the input space to high dimensional kernel feature space.

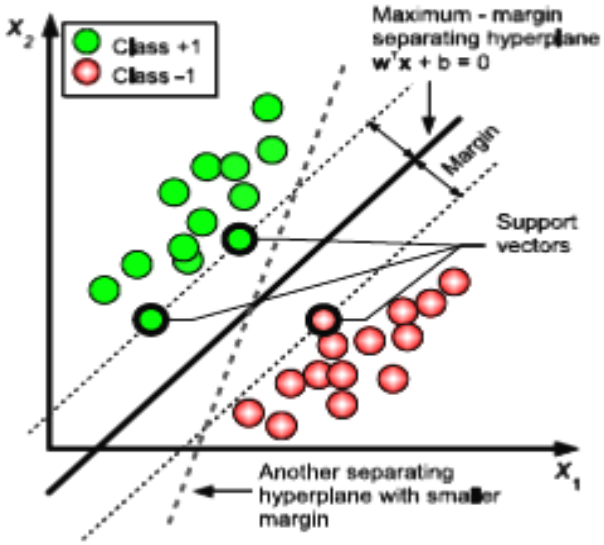

Figure 4. Principle of support vector machine[9].

3.3 Fuzzy Systems

"Fuzzy Logic (FL)" technique is utilized in image processing for "edge detection", "feature extraction", "classification", and "clustering". FL is capable enough to imitate human brain in effective manner based on logical reasoning. Reasoning explores decision making in respect of precision and certainty that involves processing costs. The extent to which "imprecision and uncertainty" can be tolerated, is explored by considering basic human capability for understanding "distorted speech", "decipher sloppy handwriting", "comprehend nuances of natural language", "summarize text", and "recognize and classify images". Fuzzy system may be used to design an inference system mapping fuzzy if-then-else rules. Fuzzy system makes use of linguistic variables matching human thought process. Fuzzy system has the capability for modelling of arbitrarily complex non-linear function to certain accuracy. Fuzzy systems can model a multi-input, multi-output system.

Zadeh was the first to introduce this term "Fuzzy Logic" in the work "Fuzzy sets," which explained mathematical aspects of "Fuzzy set theory". But Lukasiewicz was the pioneer in proposing a systemized replacement to the "bi-value logic of Aristotle" having choice of only "True or False". Lukasiewicz's proposal was to add one more value to these two as "possible," and a numerical value is assigned along with the two possible values. Then, four and five valued logics are also proposed by him. After that, it was proposed by him that, in fact, an infinite valued logic can possibly be derived. Fuzzy logic can also model the inherently imprecise conditions.

The crisp sets are extended to fuzzy sets. Crisp set allows either complete or no memberships, while fuzzy set allows partial memberships too. In crisp sets, whether an element x is a member or non-member of set A, it is represented by a membership function μA(x). If μA(x)=1 then $x \in A$ and μA(x)=0 then $x \notin A$. Fuzzy sets introduced "partial membership" in addition. A "fuzzy set A" on a universe of discourse U is defined by a characteristic function μA(x) that can take values in between [0,1]. Fuzzy set represents common sense linguistic labels like slow, fast, small, large, heavy, low, medium, high, tall, etc. A membership function is essentially a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1[10].

4. Hybrid Techniques

Next generation techniques are hybrid techniques which make use of both statistical techniques as well as soft computing techniques to achieve the higher efficiency in speaker recognition. These techniques are very much useful in Forensic Speaker Recognition because forensic data is generally not a clean data and hence robust techniques are required for this kind of systems.

5. Conclusion and Future Scope

The importance of feature matching techniques in a SR system is discussed in the present work. Three categories of these techniques namely, statistical, soft-computing and hybrid techniques are detailed in the paper giving an insight of various feature matching techniques from each category. Out of the various available techniques, "VQ, GMM, SVM, ANN and Fuzzy logic" are discussed in detail. Each of the above techniques has their respective advantages and disadvantages, but researches have shown that ANN and GMM are the best techniques for speaker recognition. For future research, the combination of two or more of the above mentioned techniques can be utilized to evolve a new hybrid technique for increasing the efficiency of the speaker recognition system.

6. References

1. Roberto T, Danie P. An overview of speaker identification: Accuracy and robustness issues, IEEE CSM. 2011; 11:23-61.
2. Soong FK, Rosenberg AE, Juang BH, Rabiner LR. A vector quantization approach to speaker recognition, AT and T Technical J. 1987; 14-26. https://doi.org/10.1002/j.1538-7305.1987.tb00198.x.
3. Saastamoinen J, Karpov E, Hautamaki V, Franti P. Accuracy of MFCC based speaker recognition in series 60 device, EURASIP J. ASP. 2005; 2816-27.
4. Reynolds DA, Quatieri T, Dunn R. Speaker verification using adapted Gaussian mixture models, DSP. 2000; 19-41.
5. Bishop C. Pattern recognition and machine learning. New York: Springer Science+Business Media; 2006.
6. Saeidi R, Mohammadi H, Ganchev T, Rodman R. Particle swarm optimization for sorted adapted Gaussian mixture models, IEEE Trans. ASLP. 2009; 17(2):344-53. https://doi.org/10.1109/TASL.2008.2010278.
7. Castaldo F, Colibro D, Dalmasso E, Laface P, Vair C. Compensation of nuisance factors for speaker and language recognition, IEEE Trans. ASLP. 2007; 15(7):969-78. https://doi.org/10.1109/TASL.2007.901823.
8. Tong R, Ma B, Lee K, You C, Zhu D, Kinnunen T, et al. Fusion of acoustic and tokenization features for speaker recognition, ISCSLP. 2006. https://doi.org/10.1007/11939993_59.
9. Campbell W, Campbell J, Reynolds D, Singer E, Torres-Carrasquillo P. Support vector machines for speaker and language recognition, CSL. 2006; 210-29. PMid: 16338636.
10. Karrey FO, DeSilva C. Soft-Computing and Intelligent System Design, Pearson Education. 2006.

Annexure-I

Feature Matching Techniques for Speaker Recognition

ORIGINALITY REPORT

8%

SIMILARITY INDEX

PRIMARY SOURCES

1	doiserbia.nb.rs Internet	53 words — 2%
2	ethesis.nitrkl.ac.in Internet	31 words — 1%
3	Kulkarni, Arun D., G. B. Giridhar, and Praveen Coca. "<title>Neural-network-based fuzzy logic decision systems</title>", Intelligent Robots and Computer Vision XIII Algorithms and Computer Vision, 1994. Crossref	17 words — 1%
4	cs.uef.fi Internet	17 words — 1%
5	www.wseas.us Internet	13 words — 1%
6	www.uef.fi Internet	13 words — 1%
7	Li, Qiang, and Yan Hong Liu. "SVM-GMM Based Speaker Identification", Advanced Materials Research, 2014. Crossref	10 words — < 1%
8	ijircce.com Internet	10 words — < 1%
9	A. M. Jinturkar. "Determination of water quality index by fuzzy logic approach: a case of ground water in an Indian town", Water Science & Technology, 04/2010 Crossref	9 words — < 1%
10	R. Sant'Ana. "On the Performance of Hurst- Vectors for Speaker Identification Systems", Lecture Notes in Computer Science, 2005 Crossref	8 words — < 1%
11	eprints.utm.my Internet	8 words — < 1%
12	Srinivasan, A.. "Real time speaker recognition of letter 'zha' in Tamil language", 2013 Fourth International Conference on Computing Communications and Networking Technologies (ICCCNT), 2013. Crossref	6 words — < 1%

EXCLUDE QUOTES ON EXCLUDE MATCHES OFF
EXCLUDE BIBLIOGRAPHY ON

Source: http://www.ithenticate.com/

Prevent Plagiarism in Publication

The Editorial Board had used the ithenticate plagiarism [http://www.ithenticate.com] tool to check the originality and further affixed the similarity index which is 8% in this case *(See Annexure-I)*. Thus the reviewers and editors are of view to discover it suitable to publish in this *Volume-10, Issue-1, January-March, 2018*.

Citation:
Pardeep Sangwan
"Feature Matching Techniques for Speaker Recognition",
Global Journal of Enterprise Information System. Volume-10, Issue-1, January-March, 2018. (http://informaticsjournals.com/index.php/gjeis)
DOI: 10.18311/gjeis/2018/16126
Conflict of Interest:
Author of a Paper had no conflict neither financially nor academically.