ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/90003, November 2016* ISSN (Online) : 0974-5645

Speaker Identification using a Novel Prosody with Fuzzy based Hierarchical Decision Tree Approach

K. Manikandan[1*] and E. Chandra[2]

[1]Department of Computer Science, PSG College of Arts and Science, Coimbatore – 641015, Tamil Nadu, India; Prof.k.manikandan@gmail.com
[2]Department of Computer Science, Bharathiar University, Coimbatore – 641046, crcspeech@gmail.com

Abstract

Objectives: The proposed speaker identification using a novel prosody with fuzzy based hierarchical decision tree approach and is used to modifying the limitations of existing traditional methods. It improves the performance of speaker identification in given population under noisy environments. **Methods/Statistics:** The key idea of this approach is to achieve an enhanced efficiency speaker cluster group using prosody features with fuzzy clustering at each level to construct the hierarchical decision tree. At each level, a speaker at belong to same groups are constructed. The proposed method has novelty of prosody as pitch and loudness with fuzzy clustering are used. **Findings:** An experimental result shows that the proposed model using prosody features outperforms the efficiency of speaker accuracy rate of 93.75 when compare to vocal source accuracy rate of 81.25 under noisy environments. **Applications:** Gender and age identification, banking and smart voice based technology operation.

Keywords: Fuzzy Clustering, Large Population Speaker Identification, Prosody Feature Extraction, Prosody with Fuzzy based Hierarchical Decision Tree

1. Introduction

A speaker-recognition system recognizes the speaker by his/her voice. It can be either identification or verification[1-15]. In this proposed method speaker identification is implemented using text-dependent data. The spoken utterance constraint includes digits from 0-9 or some fixed sentences[16-20]. The contextual information present in stored data and the input new or existing audio set of data is compared by 1: n ratio. In text independent problem is more difficult as there is no constraint on utterances[21-23]. In both cases, the idea is to identify the voice metrics (speaking habits) is same, i.e., variation in the articulator's organs.

Speech recognition other hand is a task of understanding what is being said rather than who is speaking. Speaker and speech recognition are subsets of a more general area known as pattern recognition[24-26]. The major technique for speaker identification is based on MFCC and GMM[27]. Some important GMM based approaches include the Universal Background Model. Another emerging technique which implement in very large population is i-vector. The first drawback of MFCC GMM UBM approach is difficult to match between the training and testing phase[28-31]. The second drawback is noisy environment[32].

2. Hierarchical Decision Tree

This approach aims at using a hierarchical decision tree to partition the given population data set in to group of registered speakers and sub groups of very small speaker group at leaf node. In this proposed model are extracting the prosody features such as syllable level non uniform extraction region features. It in includes the features such as pitch and loudness which are extracted from the speech data as an individual unique set of cluster group. The accuracy rate of the speaker identification is

* *Author for correspondence*

enhanced and it is improved the reliability of the speaker identification system.

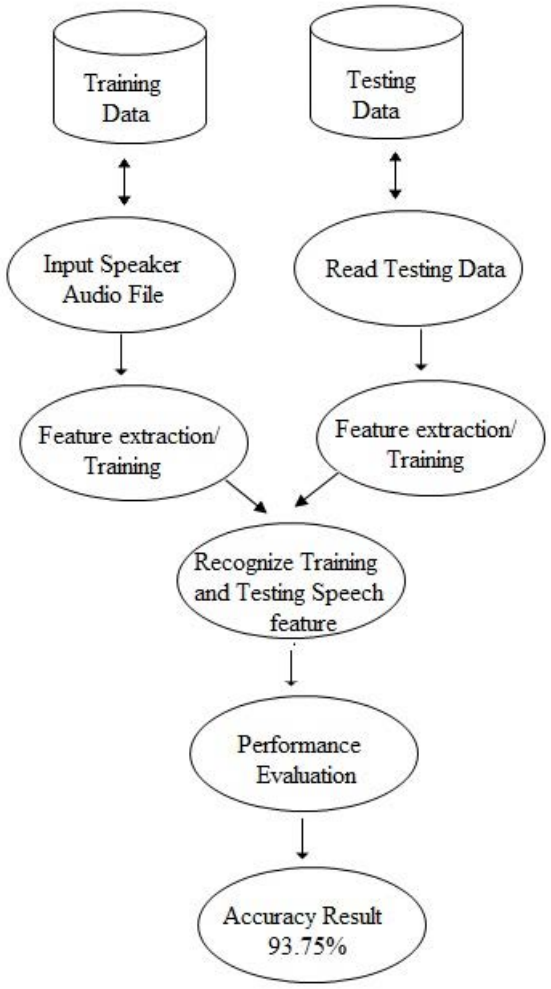

Figure 1. Speaker identification in voice metric system.

3. Novel Fuzzy Clustering based Decision Tree

3.1 Fuzzy Clustering Method

Fuzzy Clustering Method (FCM) i.e., fuzzy c-means is an approach that enables data files to fall into various clusters. It depends on the objective function as follows:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, 1 \le m \le \infty$$

where m is real number higher than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i^{th} of measured data, C_j is the numeral of the data objects in the j^{th} cluster and ||*|| is any norm exhibiting the resemblance between features (scores) of the data objects and the center. Fuzzy separating is executed by concluding an optimization iteratively of the objective function as shown above, with the membership update u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}, c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}.x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration stop when $\max_{ij}\left\{\left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right|\right\} < \varepsilon$,

where ε is a termination measured between 0 and 1, whereas k are the repetition steps. This approach converges to a local smallest or a saddle point of J_m.

The subsequent collection of step:

Step 1: Initialize the speakers number
Step 2: count the speaker number $C_{n1,n2,...nl}$
Step 3: Initialize the speaker and feature index value
Step 4: Statistic Analysis
Step 5: Data set for Clustering and mean and standard deviation of speaker i
Step 6: count the number of M cluster
Step 7: partition vector $P_{0,}$ $P1_{...}$ $P_{m-1.}$
Step 8: Initialize Speaker Index
Step 9: Initialize Cluster Index
Step 10: Confidential interval is conducted speaker i.
Step 11: Terminate node.

In the FCM technique rather than the similar assigned datum sometime it will not belong merely to a same group of cluster. In such case, the membership function represents a datum might assigned to other group of clusters with constant.

4. System Implementation

4.1 The Novel Voice Metric Method Implementation Flow

- Feature Extraction
- Feature Evaluation
- Fuzzy Clustering
- FCM Based Hierarchical Decision Tree
- Performance evaluation

4.2 Feature Extraction

4.2.1 Pitch Extraction

In this work, YIN algorithm is used for pitch feature extraction[27]. Figure 5-1 shows the input and the output of the pitch extraction module using YIN algorithm. From the figure, given a continuous speech as the input, the module first decomposed it into frames. The frame length is 25 ms and the frame shift length is 10 ms. For the i^{th} frame (i=1, 2, …N_F), its obtain the pitch estimation and the probability of the frame being voiced. Since the reasonable pitch range of human speech is from 50 Hz to 550 Hz, in this work drop all pitch estimations which are lower than 50 Hz or higher than 550 Hz and also discard all estimations of pitch extracted from frames whose probability of being voiced are below 0.8. By doing so, it can remove all potential outliers and obtain a set of reliable pitch estimations.

4.2.2 Prosody Features Extraction

In this module developed our own algorithm to extract features as pitch extraction. The Prosodic features represent the characteristics of speaker like intonation, timing, and loudness. The first one, the phone-in-word-duration, hears durations of phones are model as an individual word. The secondly, the state-in-phone-duration, hears durations are features as condition based frames. Loudness represents weak and strong syllables. It also represents speaker characteristics like anger, quiet, and tense.

4.3 Feature Evaluation

In this model evaluate proposed model justifies the input data with known set of data.

The pitch features are robust to additive noise which is usually trained and tested in high feature set of data. For example, robust are extracted from voiced speech

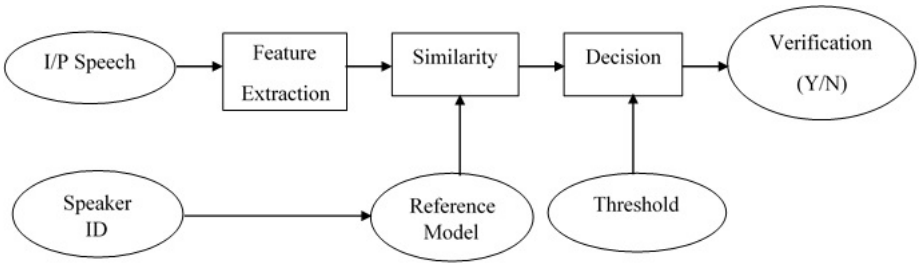

Figure 2. System flow diagrams.

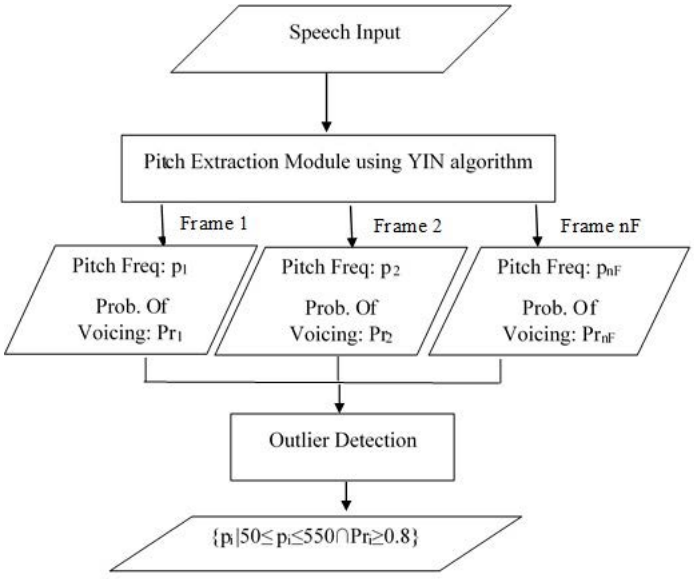

Figure 3. Pitch feature extractions.

frames. SNR is relatively increased when compared to conventional method. AWGN noise are reduced and maintained various speech are recorded are studied.

4.4 Fuzzy Clustering

To achieve an enhanced high accuracy fuzzy clustering at leaf node to root node is applied and decision-tree-based classification is derived. In this model, the prosody with fuzzy clustering technique is employed to differentiate from the conventional hard clustering, is a class of algorithms for cluster analysis that allow the objects to belong to several clusters simultaneously, with different degrees of membership. Hear is proposed to derive unique group of structure into same level of leaf node to root node. Normally the speaker group value lies either to 0 or to 1 as a binary output. But, in proposed model the values lies in-between 0 to 1 are considered. At each level of the tree, the classification is done and the performance accuracy is measured. The error mainly comes from those boundary speakers between different speaker groups or those speakers who have relatively low feature stability are studied and by means of FCM. The repetitions of same set of speaker group are eliminated and gain more accuracy using fuzzy classification. The problem is overcome by solving duplicate determination using cluster boundaries. Secondly, it reduced abnormal feature like inaccurate behavior of clustering output.

4.5 FCM based Hierarchical Decision Tree

The FCM Based Hierarchical Decision Tree approach shows the classification of various node levels from top to bottom. At all levels speakers clusters are grouped and classified in to unique structure. The leaf node to root node, fuzzy based decision tree is applied and speakers with same set of groups is significantly arranged in lower node. The speaker clusters are taken into decision and form a hierarchical tree structure. Thus for a input signal is undergoes two phases i.e., training phase where speaker cluster take place and in testing phase where the speaker group at the leaf node to root node by fuzzy decision.

4.6 Performance Evaluation

4.6.1 Accuracy

Accuracy rate could be calculated from formula given as follows,

$$accuracy = \frac{no.of\ true\ positives + number\ of\ true\ negatives}{no.of\ true\ positives + false\ positives + false\ negatives + true\ negatives}$$

4.6.2 Precision

Precision value with their true positive and false positive can be defined as,

$$precision = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + false\ positives}$$

4.6.3 Recall

The alternatively sensitivity represents Recall, where there are two different incorrect conclusions which could be drawn in a statistical hypothesis test and it can be inappropriate. An analyzed data for a positive test which accurately reflects the tested for activity. Let us consider p is the prediction, TPR represents RECALL or TRUE POSITIVE RATE, which could defined as,

$$True\ positive\ rate = \frac{True\ Positive}{Prediciton}$$

$$Prediciton = True\ Positive + False\ negative$$

The Table 1 shows the enhanced accuracy, precision

Table 1. Speaker recognition for vocal and prosody features

Speaker	Vocal source accuracy	Prosody feature accuracy	Vocal source precision	Prosody feature precision	Vocal source recall	Prosody feature recall
1	4.00	4.00	1.00	1.00	1.00	1.00
2	3.00	4.00	1.00	1.00	0.75	1.00
3	3.00	4.00	0.75	1.00	0.75	1.00
4	3.00	4.00	0.75	1.00	0.75	1.00
5	3.00	4.00	0.75	1.00	0.75	1.00
6	2.00	3.00	0.75	0.75	0.50	0.96
7	2.00	3.00	0.75	0.80	1.00	0.96
8	4.00	4.00	1.00	1.00	1.00	1.00
Accuracy	**81.2500**	**93.7500**	**0.8229**	**0.9438**	**0.8125**	**0.9375**

and recall value is compared with vocal and prosody features as follows.

4.6.4 F-Measure Comparison

F-measure differentiates the classification of labels of correct documents within various classes. In essence, it assesses the algorithm effectiveness on a class, and the higher it is, the better is the clustering. It is defined as below:

$$2((\text{precision}*\text{recall}) / (\text{precision} + \text{recall}))$$

The Table 2 shows the F-measure value is compared with vocal and prosody features.

Table 2. F-measure result for the speaker

Speaker	Vocal source F-measure	Prosody feature F-measure
1 to 8	0.8177	0.9406

4.7 Data Preparation

For this, work take data from the database recorded speech corpus. It is designed to help to characterize speakers in to various clusters. In order to provide good phonetic coverage, the CHAINS CORPUS database 40 individual data's are taken as the input for the speaker characterization and so, those have been converted to .wav format by software and read by MATLAB to get the signals.

In dataset1- I have selected 28 persons voice .Ten speech samples were collected from each of 28 people, in total 28*10=280 samples. The datasets for Cinderella story are processed.

5. Conclusion and Future Work

In this proposed research work a novel prosody with fuzzy based hierarchical decision tree approach is used to improve the accuracy rate of speaker identification under noisy conditions. This approach partition the given data to achieve an enhanced efficiency speaker cluster group using prosody features with fuzzy clustering at each level to construct the hierarchical decision tree. At each level, a speaker at belong to same groups are constructed. In addition, internal speech recognizer features include the Prosodic Features which contains the pitch and loudness features. The accuracy rate of the speaker identification is enhanced in this proposed system which is compared to the existing system. As well as it is improved the reliability of the speaker identification system. Experimental result shows that the proposed research work shows the output accuracy rate is high effective than the conventional method.

5.1 Future Work

To further validate the superiority of this model decision tree approach, more experiments should be conducted to test of this approach on datasets with larger population in different scenarios of additive noise such as interfering speakers' voice, background music, etc. Automatic speech recognition, it is a common practice to automatically determine the order of the features in decision tree.

6. References

1. Reynolds D, Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Speech Audio Process. 1995 Jan; 3(1):72–83.
2. Makhoul J. Linear prediction: A tutorial review. Proceedings of IEEE; 1975 Apr. p. 561–80.
3. Reynolds D. Large population speaker identification using clean and telephone speech. IEEE Signal Process Lett. 1995 Mar; 2(3):46–8.
4. Reynolds D. Speaker identification and verification using Gaussian mixture speaker models. Speech Commun. 1995; 17(1–2):91–108.
5. Pellom B. Hansen J. An efficient scoring algorithm for Gaussian mixture model based speaker identification. IEEE Signal Process Lett. 1998 Nov; 5(11):281–4.
6. Baraldi A, Blonda P. A survey of fuzzy clustering algorithms for pattern recognition. IEEE Trans Syst, Man, Cybern, B: Cybern. 1999 Dec; 29(6):778–5.
7. Reynolds D, Quatieri T, Dunn R. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 2000; 10(1–3):19–41.
8. Wang C. Prosodic modeling for improved speech recognition and understanding [PhD dissertation]. Cambridge, MA: Mass. Inst. of Technol.; 2001.
9. Ezzaidi H, Rouat J, O'Shaughnessy D. Towards combining pitch and MFCC for speaker identification systems. Proceedings of 7th *European Conference* on *Speech Communication* and *Technology*; 2001.
10. Spoken Language Processing. In: X. Huang, et al, Editors. Upper Saddle River, NJ: Prentice-Hall; 2001.
11. Chaudhari U, Navrratil J, Ramaswamy G, Maes S. Very large population text-independent speaker identification using transformation enhanced multi-grained models. Proceedings IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'01); 2001. p. 461–4.

12. De Cheveigne A, Kawahara H. Yin, a fundamental frequency estimator for speech and music. J Acoust Soc Amer. 2002; 111:1917.
13. Xiong Z, Zheng T, Song Z, Soong F, Wu W. A tree-based kernel selection approach to efficient Gaussian mixture model-universal background model based speaker identification. Speech Commun. 2006; 48(10):1273–82.
14. Kenny P, Boulianne G, Ouellet P, Dumouchel P. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans Audio, Speech, Lang Process. 2007 May; 15(4):1435–47.
15. Hosseinzadeh D, Krishnan S. Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMS. IEEE 9th Workshop Multimedia Signal Process (MMSP'07); 2007. p. 365–8.
16. Narvaez L, Perez J, Garcia C, Chi V. Designing 802.11 WLANS for VOIP and data. IJCSNS. 2007; 7(7):248.
17. Brummer N, Burget L, Cernocky J, Glembek O, Grezl F, Karafiat M, van Leeuwen D, Matejka P, Schwarz P, Strasheim A. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. IEEE Trans Audio, Speech, Lang. Process. 2007 Sep; 15(7):2072–84.
18. Grimaldi M, Cummins F. Speaker identification using instantaneous frequencies. IEEE Trans Audio, Speech, Lang Process. 2008 Nov; 16(6):1097–111.
19. Apsingekar V, De Leon P. Speaker model clustering for efficient speaker identification in large population applications. IEEE Trans Audio, Speech, Language Process. 2009 May; 17(4):848–53.
20. Sarkar A, Rath S, Umesh S. Fast approach to speaker identification for large population using MLLR and sufficient statistics. Proceedings of National Conference on Communication (NCC); 2010. p. 1–5.
21. Togneri R, Pullella D. An overview of speaker identification: Accuracy and robustness issues. IEEE Circuits Syst Mag. 2011; 11(2):23–61.
22. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. IEEE Trans Audio, Speech, Lang Process. 2011 May; 19(4):788–98.
23. Hu Y, Wu D, Nucci A. Pitch-based gender identification with two-stage classification. Security Commun Netw. 2011.
24. Wang N, Ching P, Zheng N, Lee T. Robust speaker recognition using denoised vocal source and vocal tract features. IEEE Trans Audio, Speech, Lang Process. 2011 Jan; 19(1):196–205.
25. Diez M, Penagarikano M, Varona A, Rodriguez-Fuentes L, Bordel G. On the use of dot scoring for speaker diarization. Pattern Recogn and Image Anal. 2011:612–9.
26. Nakagawa S, Wang L, Ohtsuka S. Speaker characterization and recognition-speaker identification and verification by combining MFCC and phase information. IEEE Trans Audio, Speech, Lang Process. 2012 May; 20(4):1085.
27. Hu Y, Wu D, Nucci A. Fuzzy-clustering-based decision tree approach for large population speaker identification. IEEE Transactions on Audio, Speech and Language Processing. 2013 Apr; 21(4).
28. Chandra E, Manikandan K, Sivasankar M. A proportional study on feature extraction method in automatic speech recognition system. IJIREEICE. 2014 Jan; 2(1).
29. Subhashree R, Rathna GN. Speech emotion recognition: Performance analysis based on fused algorithms and GMM modeling. Indian Journal of Science and Technology. 2016 Mar; 9(11). Doi no:10.17485/ijst/2016/v9i11/88460
30. Chithra PL, Aparna R. Performance analysis of windowing techniques in automatic speech signal segmentation. Indian Journal of Science and Technology. 2015 Nov; 8(29). Doi no:10.17485/ijst/2015/v8i29/83616
31. Sajeer K, Rodrigues P. Novel approach of implementing speech recognition using neural networks for information retrieval. Indian Journal of Science and Technology. 2015 Dec; 8(33). Doi no:10.17485/ijst/2015/v8i33/81115.
32. Kim D-I, Kim B-C. Speech recognition using hidden markov models in embedded platform. Indian Journal of Science and Technology. 2015 Dec; 8(34). Doi no: 10.17485/ijst/2015/v8i34/85039