TEXT INDEPENDENT HUMAN VOICE RANKING SYSTEM FOR AUDIO SEARCH ENGINES USING WAVELET FEATURES

A. Jose Albin[1] and N. M. Nandhitha[2]
[1]Faculty of Computer Science, Sathyabama University, Chennai, Tamil Nadu, India
[2]Faculty of Electrical & Electronics, Sathyabama University, Chennai, Tamil Nadu, India
Email: roemi_mich@yahoo.co.in

ABSTRACT

Performance of conventional text based audio search engines can be improved with feature based search engines. In this paper, text independent audio ranking system for audio engines with audio signal as query is proposed. Discrete Wavelet Transform (DWT) is used for feature extraction. Ranking is obtained using three different distance metrics namely Euclidean distance, Manhattan distance and Maximum distance. An efficient ranking system is identified based on the performance of the proposed technique in terms of accuracy of detection.

Keywords: text independent speaker recognition system, discrete wavelet transform, ranking, Euclidean distance, Manhattan distance, maximum distance, accuracy.

1. INTRODUCTION

In general, audio engines should search and display the audio signals which are close to the key audio signals. In conventional search engines like Yahoo and Google, even today text based query is used as the key. It searches the text i.e. the filename rather than the features and displays the results. Performance of the search engines can be improved if the text based query is replaced with the feature based query. In other words, the features of the key audio signal must be compared with the features of the database signals instead of the texts. In recent years, considerable research work is carried out in this area. However an efficient ranking system is yet to be developed. Hence the proposed research work aims at developing a ranking system that performs feature based comparison. Also the proposed system should be text and language independent. Steps involved in this work are as follows: Acquire the audio signals and create a database; preprocess the audio signals and retain a certain number of samples; convert the domain of the audio signal by choosing an appropriate transform; aggregate the features; determine the dissimilarity using distance metric and rank the audio signals based on the minimum distance.

In this paper, real time audio signals are acquired and statistical features on wavelet co-efficient are determined. Three different distance metrics are used for determining the similarity and the impact of the choice metric is determined in terms of accuracy of identification. This paper is organized as follows; Section 2 gives the related works on speaker recognition system. The proposed methodology is described in section 3. Section 4 gives the results and discussions of the proposed work. This paper is concluded in section 5.

2. OVERVIEW OF RELATED WORKS

Greenberg *et al.* (2014) described an i-vector approach for speaker recognition system. Once the speech activity is detected, mel frequency cepstra and derivatives at 100 feature vectors/second were extracted. GMM is used for speaker modelling. Based on the cosine distance between the i-vector of the test signal and i-vector of the speaker model, the speaker was identified. Srinivas and Rani (2014), proposed a new technique in which wavelet transform is used for feature extraction and neural network model is used for speaker classification. Wrapper based feature selection was included to select the best features and thereby to reduce the run time of the system. The reduced best set of features was ranked and the top ranked features were considered for training the neural network based classifier. Abdallah and Hajaiej (2014) performed a novel feature set extraction with the Gamma chirp filter bank from the voice signal that is based on the characteristics of human auditory system. Gamma chirp auditory filter bank is a non uniform band pass filter used to replicate the frequency resolution of human hearing. Compared to Mel-Frequency Cepstrum Coefficients (MFCC) the proposed auditory based technique provided better results. Prabhakar and Sahu (2014), presented a hybrid technique using MFCC and Linear Prediction Cepstral (LPC) for feature extraction. GMM and Vector Quantization (VQ) with Linde, Buzo and Gray (LBG) algorithm is used for speaker classification. McLaren *et al.* (2013) developed a robust speaker recognition system over channel degraded speech signal. Perceptual linear prediction (PLP), Medium duration modulation cepstrum (MDMC), Power-normalized cepstral coefficient (PNCC), Mean Hilbert envelope coefficient (MHEC) and Sub-band Hidden Markov model (HMM) and Gaussian Mixture Model (GMM) is independently used over the extracted five features. Sekar (2012) introduced a novel text independent speaker identification system based on image processing techniques. In this, the speech signal is converted to spectrogram for analyzing the spectral characteristics. Feature extraction is performed over the spectrogram using Radon Transform and Discrete Cosine Transform. K Nearest Neighbor is the classifier used to identify the speaker. Inter class and intra class variance was used for matching. Farah and Shamim (2013) implemented speaker recognition system for biometrics verification. For feature extraction procedure MFCC and

Linear Prediction Coding (LPC) was used. And for speaker classification, Vector Quantization (VQ) was used. The results obtained using MFCC and LPC were compared with respect to the system efficiency. The authors also investigated the effect of noise and pitch over the system accuracy.

3. PROPOSED METHODOLOGY

A research database is created with real time signals. Ten speech signals were collected each from ten different speakers. Mean, Standard Deviation, Skewness, Kurtosis, Energy, Zero Crossing Rate, Second Order Moment and Third Order Moment were extracted from the acquired audio signals. Approximation and detailed coefficients using Discrete Wavelet Transform (DWT) with Discrete Meyer wavelet (Dmey) is used for feature extraction (Albin *et al.* 2014, Nandhitha *et al.* 2007, Selvarasu *et al.* 2010). When a speech signal of unknown speaker is given, the various distance metrics namely Euclidean distance, Manhanttan distance and Maximum distance between the Dmey features of the unknown speaker and the Dmey features stored in the database are calculated. The obtained distances are arranged in ascending order and ranking is done. The best match gives a minimum distance and is given the highest rank. The speaker assigned with the highest ranks is identified as the target speaker.

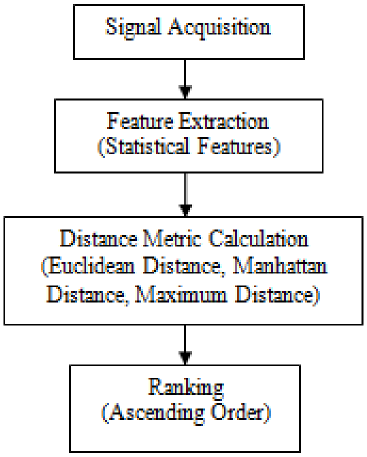

Signal Acquisition
Feature Extraction
(Statistical Features)
Distance Metric Calculation
(Euclidean Distance, Manhattan
Distance, Maximum Distance)
Ranking
(Ascending Order)

Figure-1. Flow diagram of the proposed work.

The various distances between the unknown speaker and the database speakers is calculated using the following formula [9][10][11][15].

Euclidean Distance

$$d(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_n-q_n)^2} \qquad (1)$$

Manhattan Distance

$$d(p,q)=\left|(p_1-q_1)\right|+\left|(p_2-q_2)\right|+\ldots.+\left|(p_n-q_n)\right| \qquad (2)$$

Maximum Distance

$$d(p,q)=Max[\left|(p_1-q_1)\right|+\left|(p_2-q_2)\right|+\ldots.+\left|(p_n-q_n)\right|] \qquad (3)$$

Performance of the proposed technique is measured in terms of accuracy [12].

$$Accuracy=\frac{No.ofTP+TN}{No.ofTP+FP+FN+TN}*100 \qquad (4)$$

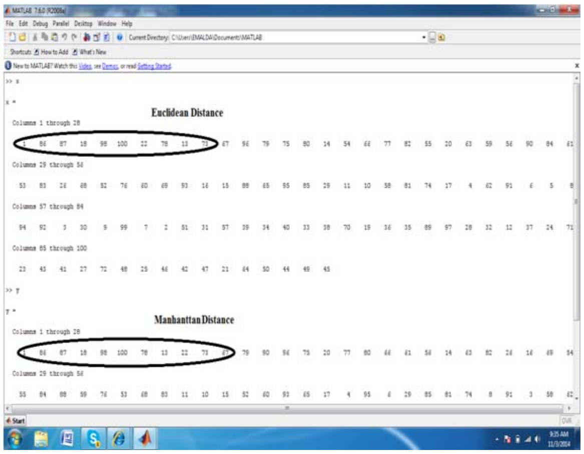
Where TP is True Positive, identifying speaker 1as speaker 1and so on. FN is False Negative, identifying speaker 1 as some other speaker and so on. TN is True Negative, correctly rejecting the speakers other than the key speaker. Speaker 1 is identified as other speaker is False Positive FP.

4. RESULTS AND DISCUSSIONS

Performance of the proposed work is measured in terms of accuracy and the impact of the choice of the distance metrics is also shown in Table 1. Here audio signals are numbered as 1-10 for speaker 1, 11-20 for speaker 2, 21-30 for speaker 3 and 91-100 for speaker 10. Every signal is given as the input query signal and the ranking is determined. The first 10 signals with minimum distance are considered and the accuracy is determined out of these 10 signals. For example from the first row of Table 1, it is interpreted that with speaker 1's first signal as input, irrespective of the distance metric, only one signal corresponding to speaker 1 was ranked correctly and displayed as the first 10 signals. On the other hand, from the second row it is found that Euclidean and Maximum distance metrics provide 9 out of 10 signals correctly (When the second signal of speaker 2 is given as the query signal). It is also shown in Figure-2 and Figure-3.

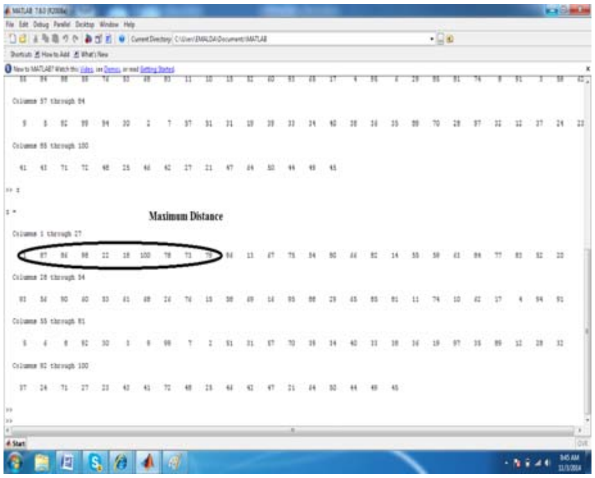
Table-1. Accuracy using various distance for speaker 1.

Sample	**Accuracy** (%)		
	Euclidean distance	**Manhattan distance**	**Maximum distance**
1	10	10	10
2	90	80	90
3	90	80	90
4	70	70	70
5	70	70	70
6	70	70	70
7	80	80	90
8	80	70	80
9	90	80	90
10	40	40	30


MATLAB 7.6.0 (R2008a)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\EMALDA\Documents\MATLAB
Shortcuts How to Add What's New
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Euclidean Distance
Manhanttan Distance
Start

```
>> x

x =

  Columns 1 through 28

     1    86    87    18    98   100    22    78    13    73    67    96    79    75    80    14    54    66    77    82    55    20    63    59    56    90    84    61

  Columns 29 through 56

    53    83    26    68    52    76    60    69    93    16    15    88    65    95    85    29    11    10    58    81    74    17     4    62    91     6     5     8

  Columns 57 through 84

    94    92     3    30     9    99     7     2    51    31    57    39    34    40    33    38    70    19    36    35    89    97    28    32    12    37    24    71

  Columns 85 through 100

    23    43    41    27    72    48    25    46    42    47    21    64    50    44    49    45

>> y

y =

  Columns 1 through 28

     1    86    87    18    98   100    78    13    22    73    67    79    90    96    75    20    77    80    66    61    56    14    63    82    26    16    69    54

  Columns 29 through 56

    55    84    88    59    76    53    68    83    11    10    15    52    60    93    65    17     4    95     6    29    85    81    74     8    91     3    58    62
```

Figure-2. Euclidean distance and Manhattan distance for sample 1 of speaker 1.


MATLAB 7.6.0 (R2008a)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\EMALDA\Documents\MATLAB
Shortcuts How to Add What's New
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Maximum Distance
Start

```
    55    84    88    59    74    53    68    83    11    10    15    52    60    93    65    17     4    95     6    29    85    81    74     8    91     3    58    62

  Columns 57 through 84

     9     5    92    99    94    30     2     7    57    51    31    19    39    33    34    40    38    36    35    89    70    28    97    32    12    37    24    23

  Columns 85 through 100

    41    43    71    72    48    25    46    42    27    21    47    64    50    44    49    45

>> z

z =

  Columns 1 through 27

     1    87    86    98    22    18   100    78    73    79    64    13    67    75    54    90    66    82    14    55    58    61    84    77    83    92    23

  Columns 28 through 54

    93    56    90    60    53    61    68    26    76    15    38    69    16    95    88    29    65    85    81    11    74    10    62    17     4    94    91

  Columns 55 through 81

     5     6     8    92    30     3     9    99     7     2    51    31    57    70    39    34    40    33    19    36    19    97    35    89    12    28    32

  Columns 82 through 100

    37    24    71    27    23    43    41    72    48    25    46    42    47    21    64    50    44    49    45

>>
>>
```

Figure-3. Maximum distance for sample 1 of speaker 1.

This procedure is repeated for all the samples of 10 speakers and is listed in Table-2. Figure-2 depicts the average accuracy for the considered 10 speakers using various distance metrics.

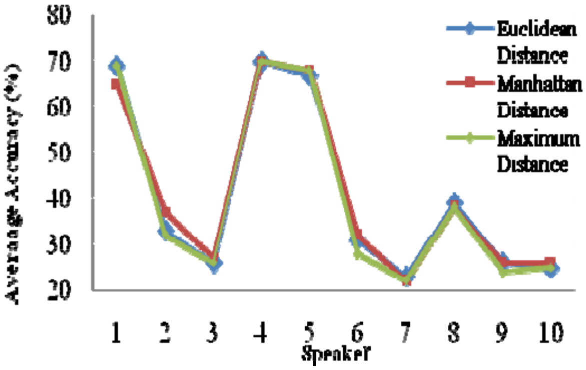


Figure-4. Average accuracy for speakers using various distances.

Table-2. Average accuracy for three different distance metrics for 10 speakers.

Speaker	Accuracy (%)		
	Euclidean distance	Manhattan distance	Maximum distance
1	69	65	69
2	33	37	32
3	26	27	26
4	70	70	70
5	67	68	68
6	31	32	28
7	23	22	22
8	39	38	38
9	26	26	24
10	25	26	25

From the Table-2, it is found that Manhattan distance based classifier provides nearly perfect matching at more instances. It is also emphasized from Table-3 which provides the average accuracy, maximum and minimum accuracy of each speaker. The average accuracy of all the speakers using various distance is listed in Table-3. From the Table-3, it is clear that for the considered speech signals, Manhattan distance provides better accuracy than Euclidean and Maximum distance.

Table-3. Average accuracy using various distance metrics.

Distance metric	Average accuracy	Maximum accuracy	Minimum accuracy
Euclidean distance	40.9	70	23
Manhattan distance	41.1	70	22
Maximum distance	40.2	70	22

5. CONCLUSIONS

An efficient ranking technique is required for a text independent audio mining system. In this paper, DWT is used for feature extraction after which ranking based on various distances is performed. Distance metrics like Manhattan distance, Euclidean distance and Maximum distance were considered. Average accuracy using the various distances over the maintained speech database is calculated. From the results, it is proved that the Manhattan distance is the best ranking technique for the text independent speaker recognition system.

REFERENCES

[1] Craig S. Greenbergm., Désiré Bansé., George R. Doddington., Daniel Garcia-Romero., John J. Godfrey., Tomi Kinnunen., Alvin F. Martin., Alan McCree., Mark Przybocki., Douglas A. Reynolds. 2014. Odyssey 2014: The Speaker and Language Recognition Workshop, 2014, Joensuu, Finland. pp. 224-230.

[2] V. Srinivas., Ch. Santhi rani., T. Madhu 2014. Neural Network based Classification for Speaker Identification. International Journal of Signal Processing, Image Processing and Pattern Recognition, 7(1), 109-120.Amina Ben Abdallah, Zied Hajaiej (2014). Improved Closed Set Text Independent Speaker Identification System using Gammachirp Filterbank in Noisy Environments. IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14), 978-1-4799-3866-7/14/ $31.00 ©2014 IEEE.

[3] Om Prakash Prabhakar., Navneet Kumar Sahu. 2014. Speaker Identification system using Mel Frequency Cepstral Coefficient and GMM technique. International Conference on Advances in Engineering & Technology. pp. 51-56.

[4] Mitchell McLaren., Nicolas Scheffer., Martin Graciarena., Luciana Ferrer., Yun Lei. 2013. Improving Speaker Identification Robustness To Highly Channel-Degraded Speech Through Multiple System Fusion. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6773-6777.

[5] Sekar K. 2012. Performance Analysis of Text-Independent Speaker Identification System. International conference on modeling optimization and computing. Elsevier Procedia Engineering. 38: 1925- 1934.

[6] Shahzadi Farah., Azra Shami. 2013. Speaker Recognition as System Using Mel-Frequency Cepstrum Coefficients, Linear Prediction Coding and Vector Quantization. IEEE International Conference on Computer, Control and Communication. (IC4).

[7] A. Jose Albin., N.M. Nandhitha. and S. Emalda Roslin. 2014.Text Independent Speaker Recognition System using Back Propagation Network with Wavelet Features. IEEE International Conference on Communication and Signal Processing, 978-1-4799-3357-0, 942-946.

[8] http://en.wiktionary.org/wiki/Manhattan_distance

[9] http://en.wikipedia.org/wiki/Euclidean_distance

[10] http://en.wikipedia.org/wiki/Absolute_difference

[11] http://en.wikipedia.org/wiki/True_positive#true_positive

[12] N. M. Nandhitha., N. Manoharan., B.sheela Rani., B. Venkataraman., P. Kalyana Sundaram. and Baldev Raj. 2007. Detection and Quantification of Tungsten Inclusion in Weld Thermographs for on- line weld monitoring by region growing and Morphological Image Processing Algorithm. Proc. Of International Conference on Computational Intelligence and Multimedia Applications. pp. 513-518.

[13] N. Selvarasu., Alamelu Nachiappan., N. M. Nandhitha. 2010. Extraction and Quantification Techniques For Abnormality Detection From Medical Thermographs In Human. International Journal of Technology And Engineering System (IJTES). 1(2): 120-124.

[14] N. Selvarasu., Alamelu Nachiappan., N.M. Nandhitha. 2010. Abnormality Detection from Medical Thermographs in Human Using Euclidean Distance Based Color Image Segmentation. International Conference on Signal Acquisition and Processing. pp. 73-75.

[15] A. Jose Albin., N.M. Nandhitha., S. Emalda Roslin. 2014. ART Network Based Text Independent Speaker Recognition System for Dynamically Growing Speech Database. International Conference on Frontiers of Intell. Comput. (FICTA). Advances in Intelligent Systems and Computing. 1: 473-480.