Comparison of DTW Score and Warping path for Text Dependent Speaker Verification System

Tushar K. Das
ECE Dept.
NIT Silchar
Assam, India
tusharkdnits@gmail.com

Songhita Misra
ECE Dept.
NIT Silchar
Assam, India
msonghita@gmail.com

Suman P. Choudhury
ECE Dept.
NIT Silchar
Assam, India
sumanpc2008@gmail.com

Dinesh K. Sah
CSE Dept.
NIT Silchar
Assam, India
dinesh12159@gmail.com

Ujwala Baruah
CSE Dept.
NIT Silchar
Assam, India
b.ujwala@gmail.com

Rabul H. Laskar
ECE Dept.
NIT Silchar
Assam, India
rabul18@yahoo.com

***Abstract*-- Dynamic Time Warping has always been a popular technique for pattern matching of two speech samples for Automatic Speaker Recognition. DTW score evaluated from the minimum distance matrix is generally used to identify the similarity between two speech samples and thereby giving a decision for the Text Dependent Speaker Verification system. This paper discusses a system based on DTW trace back path for the alignment of two speech samples for the decisions are thereby compared with the performance of the system with the system based on DTW Score. The result shows a significant improvement in the performance of the system suggesting that the trace back path provides a better similarity measure as compared to that of the DTW Score.**

***Keywords*—DTW Score, Warping path, Text Dependent Speaker Verification, MFCC, VAD.**

I. Introduction

The rate of speaking of a sentence by a speaker at different time instants may vary widely and it is essential to align the speech samples for comparing the similarity between two speech samples. In the present scenario DTW finds a wide application in the field of pattern matching of two temporal sequences. Text dependent speaker verification [1] [2] [3] [4]uses the similarity measure to give a binary decision of either accepting the speaker as genuine or rejecting it as an imposter. DTW score is the minimum cost for traversing through the distance matrix of two speech sample matrices and tracing back the minimum distance path which gives an optimal warping path for aligning the two speech samples [8]. The DTW only aligns the speech sample if the two samples vary in the range of 0.5 to 2 times (0.7 to 1.4 times practically), but there may be cases where thespeaking rate variation is very high, falling in the limitation of DTW.

The system implemented uses VoiceActivityDetection (VAD)[5] for truncation of non-speech part of the speech signal and the speaker specific features of the speech were extracted using Mel scale based Cepstral coefficients. Mel-Frequency Cepstral Coefficient (MFCC)

[6] is a well-known feature extraction technique in the field of speech processing. A 13 dimensional MFCC with 40 filter banks, 20 ms frame size with a 10 ms overlap with a 256 DFT was used to extract the speaker specific feature.

The dynamic time warping[10] [11] algorithm developed by Dan Ellis [10]evaluates a total warping cost along the minimum path of the distance matrix on which the first system was developed and it also returns a trace back path depending on which the two aligning matrix is modified and their Euclidean distance gives the decision of accepting or rejecting the speaker.

The decision made in this paper is a cohort decision i.e. the distance between the utterances of the same speaker must be less than the distance between the utterances of two different speaker for the same text, based on which the speaker has to be accepted as a genuine speaker or been rejected otherwise.

A database of 30 speakers is used to develop the systems. Each speakers have 3 training utterances and 20 testing utterances recorded at a lab environment at a sampling frequency of 8KHz. The uttered speech duration is of 2-3 seconds with session variability, and the sentences are selected according to the richness of the phoneme compared to TIMIT sentence.

The rest of the paper is organised as follows. Section II deals with text dependent speaker verification system. DTW Score and warping path is discussed in Section III. Section IV highlights the experimental observations in this study. Section V summarizes the study and also provides with the future scope of work.

978-1-4799-7075-9/15/$31.00 ©2015 IEEE

II. Text Dependent Speaker Verification System

In the present scenario, speaker recognition [1] has become one of the prominent approach for authentication of a person remotely and it is believed to be one most important state of art recognition techniques, for its vast applications.

Text dependent speaker verification [2] [3] [8] is a sub branch of speaker recognition that specifically gives a binary decision on whether to accept or reject a speaker based on the text prompted at the training as well as at the testing sessions.

The basic approach of the text-dependent speaker verification is shown in blocks, in the figure below.

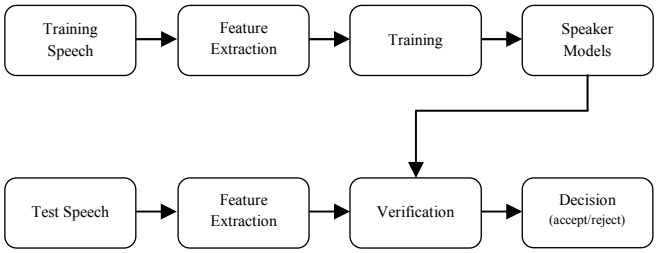

Fig. 1 Basic block diagram of Text Dependent Speaker Verification

A. *Training Phase:*

Three utterances of each speaker is taken for the training of the system. The speech templates are passed through energy based non speech truncation using Voice Activity Detection (VAD) [5] [7]. The Fig. 2 shows the speech input, the non-speech part detected by VAD and the reduced speech extract for a particular input speech.Fig.2 Speech file at various stages.

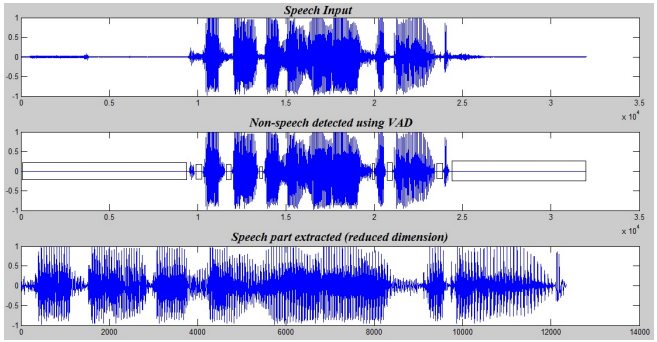

Fig.2 From top, speech input; non-speech part detected using VAD; speech part extracted discarding non-speech part

The extracted part of the speech are further processed for feature extraction using Mel filter based 13 dimensional Cepstral coefficients [6] to evaluate the 13 dimensions of the speaker specific characteristics of the speech.

Three corresponding templates are formed for each individual speaker. One of the training utterances is used for MFCC extraction and is taken as the 1st MFCC reference. Similarly extracted 2nd, 3rd MFCC utterances were time aligned with the first and mean of the three served as the 1st template. Similarly 2nd and 3rd template are formed using the other two utterances of the training phase as the main reference.

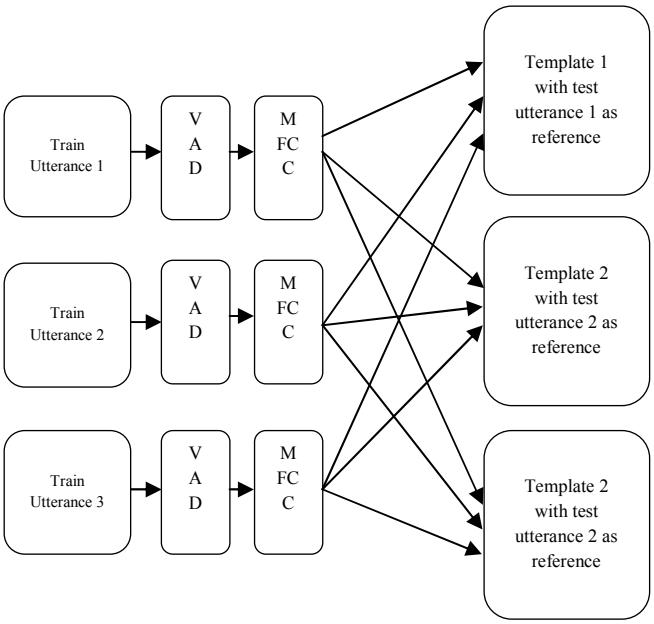

Fig. 3 Training Phase of Text Dependent Speaker Verification System

A. *Testing Phase:*

The test utterance gives a claim of his identity and the corresponding claimed identity template is collected from the saved templates during the training phase. The test utterance is passed through a non-speech part removal block using the energy based Voice Activity Detection (VAD) [5] [7]. The 13 MFCC features [6] are extracted from the test utterance. For the aligning of the templates and the decision making, the test speech is passed through the two DTW based models as discussed in the next section.

III. Dynamic Time Warping

In time series analysis, dynamic time warping (DTW) [10] [11] is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. DTW finds its application in many of the audio, video processing and data mining. In general, DTW is a method that calculates an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" non-linearly in the time dimension

to determine a measure of their similarity independent of certain non-linear variations in the time dimension.

In this paper, speaker verification is addressed by the DTW in two different ways:

1. *Model 1:*

In this model, the DTW [10] is used for aligning the MFCC of the test utterance with that of the claimed template in usual traditional manner. The whole speech matrix is aligned using the warping path provided by DTW, followed by the cohort based decisions process where the Euclidean distance between the test utterance and the training templates of that claimed speaker should be minimum when compared with some randomly chosen training templates of other speakers.

2. *Model 2:*

In this model, the total DTW cost incurred between two speech matrices, is used as the speaker defining parameter for making a decision on the claim. The MFCC [6] of the test utterance is aligned with that of the claimed speaker's template and the total cost is calculated for the speech matrices. On the basis of the total cost, the cohort based decision is taken to recognise the speaker as genuine or as an imposter.

For two speech matrices, the distances matrix is calculated as discussed above. Let the distance matrix D be of $n \times m$ dimension then the cost matrix C is calculated by

$$C = \begin{cases} D_{ij}, & i = 1 \; or \; j = 1 \\ \min\left(D_{(i-1)j}, D_{i(j-1)}, D_{(i-1)(j-1)}\right) + D_{ij}, & i \neq 1 \; and \; j \neq 1 \end{cases} \quad \ldots(1)$$

The total cost between two speech matrices is given $C_{n\times m}$ which is taken as the parameter for making the cohort decision in Model 2.

The cohort based decision as shown in Fig. 4 depends on the ground that intra-speaker variability should be less as compared to that of the inter speaker variability. The system discussed in this paper compares the distance calculated from 3 claimed template with that of the 9 randomly selected non-claimed templates. If the claimed identity's distances are found to be less than that of the randomly selected not-claimed identity distances, the speaker is accepted as a genuine speaker.

IV. Experimental Observations

The experimental observations for the text-dependent speaker verification system, using both the DTW approaches for genuine speakers are shown in Table I and that for imposter speakers are shown in Table II. The comparing parameter, Equal Error Rate (EER) was calculated for both the Models based on the success rate of the genuine and imposter speakers as shown in Table III.

The observations in Table I depicts that out of 600 genuine tests, the speakers recognised correctly by the machine is 510 for Model 1 giving a success rate of 85% whereas for Model 2, 522 tests utterances are correctly recognised , giving a success rate of 87% were accepted for 522 tests for Model 2 giving a success rate of 87%.

In Table II, the system is analysed for the imposter speakers with the same number as carried out for the genuine.

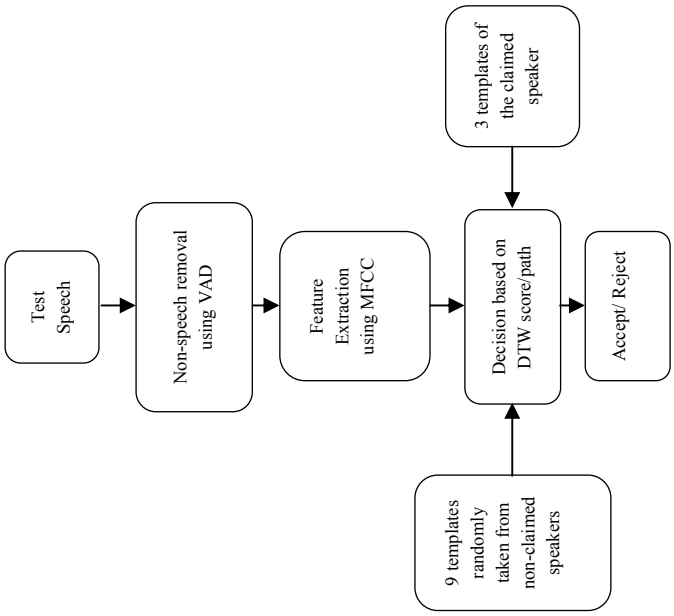

Fig.4 Block wise approach for cohort based decision

Here, 468 test utterances are rejected correctly (132 test utterance of the imposters are accepted)for Model 1, giving a success rate of 78% and for Model 2, the test utterances rejected were 492 (108 test utterance of the imposters are accepted) giving a success rate of 82%.

The tests were performed to analyse the importance of full warping path as compared to that of the minimum cost for genuine and imposter tests which reveals that the full warping path provides a comparable success rate as that of minimum cost.

Table I: Results for genuine trials based on Model 1 and Model 2

	No. of Tests	No. of Tests Passed	Success Rate (%)
DTW Score (Model 1)	600	510	85
Warping Path (Model 2)	600	522	87

Table II: Results for imposter trials based on Model 1 and Model 2

	No. of Tests	No. of Tests Passed	Success Rate (%)
DTW Score (Model 1)	600	132	78
Warping Path (Model 2)	600	108	82

The results observed for genuine and imposter tests were characterised on basis of EER for a comparable analysis in Table III. As the table suggests the system based on Model 1 DTW approach gives a False Acceptance Rate (FAR) of 22% and False Rejection Rate (FRR) of 15% resulting an EER of 18.5% whereas the system based on Model 2 DTW approach gives a FAR of 18% and FRR of 13% with an EER of 15.5%.

Table III: Results based on error for Model 1 and Model 2

	FAR (%)	**FRR (%)**	**EER (%)**
DTW Score (Model 1)	22	15	18.5
Warping Path (Model 2)	18	13	15.5

V. Summary & Conclusion

DTW is considered to be the most efficient aligning technique in time domain analysis particularly in automatic speaker recognition techniques. This paper compares the performance of DTW warping path based speaker verification technique with that of the DTW Score based approach for speaker verification system. The study also suggests that there may be an improvement of EER of about 3%in the EER for the speaker verification system when the decision is based on the DTW Score which may signify that the total cost of the matrix may give a better approximation of the speaker as compared to that of the warping path. The study was carried on a controlled noise environment, on microphonic database. Rigorous study may be further carried out in this path, followed by its use in various applications for betterment of the system performance.

Acknowledgement

The authors highly acknowledge the Department of Electronics & Information Technology (DeitY), Ministry of Communications & IT, Government of India for the resources provided and also their never ending support and motivation for the research work.

References

[1] H. Matthieu, "*Text-Dependent Speaker Recognition*," Springer Handbook of Speech Processing, pp. 743-762, 2008

[2] "*A Tutorial on Text-Independent Speaker Verification*," EURASIP Journal on Applied Signal Processing, pp. 430–451, 2004

[3] K. Tomi, L. Haizhou, "*An overview of text-independent speaker recognition: From features to supervectors*," Speech Communication, Vol. 52, Issue 1, 2010

[4] S. Shukla, S. R. M. Prasanna and S.Dandapat, "*Speech Recognition under Stress Condition*," in 15thNational Conference on Communications, IIT Guwahati, Jan 2009, pp. 299-302.

[5] Boyd, Ivan, and Daniel K. Freeman. "*Voice activity detection*," U.S. Patent No. 5,276,765, Jan 1994

[6] Hasan, MdRashidul, et al. "*Speaker identification using mel frequency cepstral coefficients*," 3rd International Conference on Electrical & Computer Engineering ICECE, Vol. 2004, 2004

[7] Kondoz, A. M. "*Voice Activity Detection,*" Digital Speech: Coding for Low Bit Rate Communication Systems, Second Edition, pp. 357-377, 2004

[8] K.S.R. Murty, B. Yegnanarayana, "*Combining evidence from residual phase and MFCC features for speaker recognition*," IEEE Signal Processing Letters, pp. 52-55, 2006

[9] R.R. Lawrence, E.R. Aaron and E.L. Stephen, "*Considerations in dynamic time warping algorithms for discrete word recognition*," Acoustical Society of America, 2005

[10] http://www.ee.columbia.edu/ln/rosa/matlab/dtw/

[11] http://www.google.com/patents/US8099288