International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Robust Speaker Recognition using Enhanced Spectrogram

[1] **Sukhvinder Kaur , [2] J. S. Sohal**

[1]Ph.D Reaearch Scholar , I.K. Gujral PTU, Jalandhar, Kapurthala, India
[2]Director, LCET, Ludhiana, Punjab, India

ABSTRACT

The aim of this paper is to present an efficient, fast and optimized system that identify the speaker in automatic speaker recognition system (ASR). It can be used in voice biometrics. In this proposed technique, the daubechies wavelet transform is used to compress the audio stream in the ratio of 1:4 with 99% of energy; their features are extracted by enhanced spectrogram with non-linear energy operator. Finally, three different distance matrices: T-test, deltaBIC and KL2 were used for feature matching of different speakers. The proposed technique using enhanced spectrogram with t-test distance metric gives fast and better results as compared to delta BIC and KL2.

Keywords : Bayesian Information Criteria, Kullback Leibler Distance Metric, Enhanced Spectrogram, Non-Linear Energy Operator, T-Test Wavelet Transform

I. INTRODUCTION

Speaker recognition is the identification of a person from characteristics of voices (voice biometrics). It is also called voice recognition. There is a difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said). These two terms are frequently confused, and "voice recognition" can be used for both[1]. One of the demanding areas of Automatic Speaker recognition application is in forensics. Usually the case where a crime has been committed and the voice of the criminal needs to be verified from a recorded message[2]. Traditionally this was done by training a specialist who can able to identify the speaker's voice by comparing the visual speech features (spectrograms voice prints) of the speakers. But the accuracy in these methods was found not reliable and not effective. To prove that the suspect is the criminal, it needs to be verified beyond reasonable doubt that the voice of the criminal and the voice of the suspect are the same. So to overcome this problem a Automatic and reliable Speaker Verification system is desired[3].

This paper proposes a method for automatic speaker recognition by using wavelet transform and enhanced spectrogram algorithm as feature extraction and traditional Bayesian information criteria, kullback leibler (KL2) distance metric and T-test algorithm for feature matching. In this study, a discrete wavelet transform (DWT) based compression and de-noising approach is presented to improve the speech quality of speaker and then enhanced spectrogram algorithm is applied. After feature matching, the results are compared using different distance metrics. The following section will introduce the principles of wavelet transform, enhanced spectrogram and feature matching algorithms. The experimental results shown in section 3 prove that there is an improvement in the proposed speaker recognition system. Finally conclusion and future scope is presented in last section.

II. Feature Extraction

Discrete Wavelet Transform

Wavelet Transform is emerged in the 1980s; however it only started being used to solve engineering problems in the 1990s[4]. Discrete wavelet transform (DWT) uses the fact that it is possible to resolve high frequency components within a small time window, and only low frequency components need large time windows[5]. This is because a low frequency component completes a cycle in a large time interval whereas a high frequency component completes a cycle in a much shorter interval. Therefore, slow varying components can only be identified over long time intervals but fast varying components can be identified over short time intervals. The wavelet transform is defined as the inner product of a signal *x(t)* with the mother wavelet *ψ(t)* is as follows:

$$W_{\psi}x(a,b)=\frac{1}{\sqrt{a}}\int_{\infty}^{-\infty} x(t)\psi^{*}_{a,b}(t)\,dt, (1)$$

Where,

$$\psi_{a,b}(t)=\psi\left(\frac{t-b}{a}\right) \quad (2)$$

CSEIT1724135 | Received : 10 August 2017 | Accepted : 21 August 2017 | July-August-2017 [(2)4: 637-640]

Where *a* and *b* are scale and shift parameters respectively. Mother wavelet can be dilated or translated by changing *a* and *b*. The DWT functions at level *m* and time location t_m can be expressed as:

$$d_m(t_m)= x(t)\, \psi_m\left(\frac{t-t_m}{2^m}\right) \quad (3)$$

Where, ψ_m is the decomposition filter at frequency level *m*. The effect of the decomposition filter is scaled by the factor 2^m at stage m, but otherwise the shape is the same at all stages. DWT is used in speaker recognition to decompose the speech signal into two halves, lower frequency components known as approximations and high frequency component represented as details. About 98% of speech information is present in approximation. This algorithm is used to compress and de-noise the speech signal as shown in figure 1

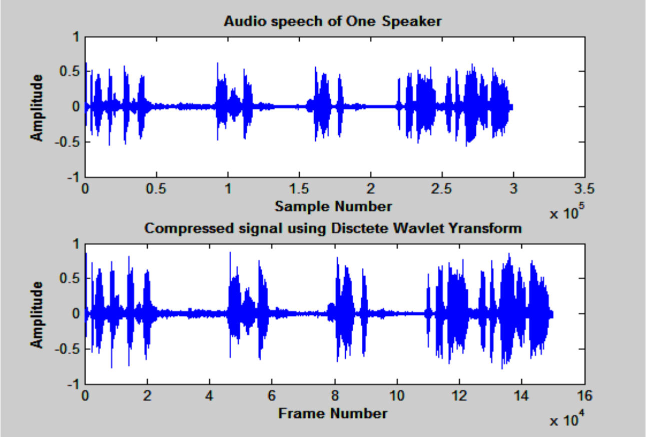

Figure 1. Waveform of Original Audio Signal and its Compressed Form

Enhanced Spectrogram

The enhanced spectrogram, called pyknogram, were first introduced in [6] to facilitate formant tracking and are calculated by applying multiband demodulation in the framework of the AM-FM modulation model[7]. Overlaps in speech data can be detected by using pyknogram [7]. In pyknogram, the resonances (formants) and harmonic structure of speech are enhanced by decomposing the spectral sub-band into amplitude and frequency components.The frequency and amplitude components of a given subband, x(n), is as follows:

$$f = \frac{1}{2\pi}\arccos\left(1-\left(\frac{\psi|x(n)-x(n-1|}{2\psi|x(n)|}\right)\right) \quad (5)$$

where ψ[x(n)] is non linear energy operator (NEO).

$$\psi[x(n)] = x^2(n) - x(n-1)x(n+1) \quad (6)$$

$$|a| = \sqrt{\frac{\psi|x(n)|}{\sin^2(2\pi f)}} \quad (7)$$

The weighted average of the instantaneous frequency components are used to derive a short-time estimate value for the dominant frequency in each subband over a fixed period of time, in this case the duration of a time-frame (typically 12 msec).

$$F_w(t) = \frac{\sum_t^{n+T} f(n)a^2(n)}{\sum_t^{n+T} a^2(n)} \quad (8)$$

where f(n) and a(n) are the instantaneous frequency and amplitude functions calculated for each sample in the t^{th} frame over the frame length (T samples per frame). Resonances and harmonic peaks are located in each frame by comparing the average frequency estimates with filter bank center frequencies [6]. The motivation behind using an energy operator based approach [8] is to avoid assumptions on the number of speakers in the signal. The AM-FM decomposition method relies on signal resonances and does not restrict the signal to a specific structure. The final time-frequency representation is called a pyknogram and is denoted Spyk(t, f) as a function of time (t) and frequency (f) as shown in figure 2.

III. Distance metrics for Feature Matching

Bayesian Information Criteria

Bayesian Information Criterion (BIC) is one of the most popular technique for detecting speaker change point in an audio recording presented in [9]. It's the statistical measure used in statistical hypothesis testing.

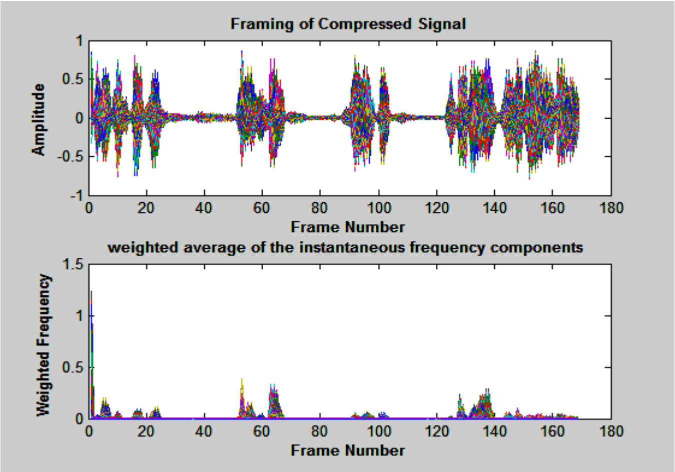

Figure 2. Frames of Compressed Signal and weighted average of instantaneous frequency Component

Let's say the model trained on segment X_1 and X_2 is M_1 and M_2 respectively. Then BIC for each segments are,

$$BIC(X_1, M_1) = \log (p(X_1|M_1)) - \lambda d_l \log N_1 \quad (9)$$

$$BIC(X_2, M_2) = \log (p(X_2|M_2)) – \lambda d_2 \log N_2 \qquad (10)$$

The first term is likelihood term while second term checks for complexity and therefore controls over-fitting. Similarly BIC of segments concatenating X_1 and X_2, let's say X, with respect to model M is calculated. Finally following BIC measure is calculated.

$$\Delta BIC = BIC(M) - BIC(M_1) – BIC (M_2) \qquad (11)$$

For multivariate Gaussian distributions $M_1 = N(\mu_1,\sum_1)$, $M_2 = N(\mu_2, \sum_2)$ and $M = N(\mu , \sum)$ with model size N_1, N_2 and N_1+N_2 respectively, delta BIC is

$$\Delta BIC = (N_1 + N_2) \log(\sum) - N_1\log(\sum_1) - N_2\log(\sum_2)$$

$$- \lambda(0.5 *(d+0.5*(d+1))) \log N \qquad (12)$$

Where λ is a penalty weight, d is a dimension of the feature space and $\sum_1$, $\sum_2$ and $\sum$ are determinants of covariance matrices for the segments X_1 , X_2 and X respectively. If ΔBIC >0, a local maximum of ΔBIC is found and time ti is considered to be a speaker change point. If ΔBIC < 0, there is no speaker change point at time t_i.

Kullback Leibler (KL2)Dstance Metric

The Kullback-Leibler distance (KL2) is a popular distance metric in speech recognition. If two audio segments are modeled by multivariate Gaussian distribution $N(\mu_1,\sum_1)$ and $N(\mu_2,\sum_2)$, then the KL2 distance between the segments is given as :

$$KL2_{1,2} = \frac{1}{2}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) + \frac{1}{2}tr(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I) \quad (13)$$

This metric more popular in speech processing when used to characterize the similarity of two audio segments.

Student t-test

As discussed in [10] student t-test is an efficient distance metric which is defined as

$$T_d = d(S_a(X), S_b(X)) = \frac{|m_1 - m_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (14)$$

Where m1, σ1, n1, m2, σ2, n2 are respectively the mean, stan- dard deviation and size of two populations S_a (X) and S_b(X). Applying the above formulas in the context of measuring the distance between two speakers $S_1 = \{x_1, x_2, . . . ,x_N\}$ and $S_2 = \{y_1, y_2, . . , y_M\}$, with the following proposed distribution function:

$$a(x)= \log(p(x_i|M_1)) – \log(p(X|M_{UBM})) \qquad (15)$$

$$b(x) = \log(p(y_i|M_2)) − \log(p(X|M_{UBM})) \qquad (16)$$

where, $X = \{ x_1,x_2, . . . , x_N; y_1, y_2, . . . , y_M \}$, x_i; y_i are the feature vectors, M_1 is the model estimated using feature vectors of speaker S_1, M2 is the model of speaker S_2, UBM is the universal background model. The distance between speaker S1 and S2 is then computed using (13); a smaller value of T_d indicates that two speakers are more similar to each other.

IV. Speaker Recognition System

The system follows the standard automatic speaker recognition system framework shown in figure 3. The system includes speaker database, feature extraction and feature matching algorithms as explained in previous section.

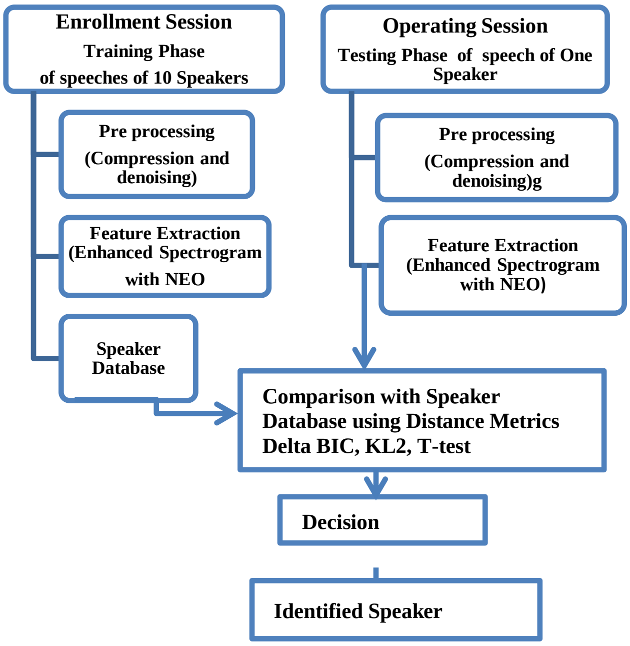
Enrollment Session
Training Phase
of speeches of 10 Speakers
Operating Session
Testing Phase of speech of One Speaker
Pre processing
(Compression and denoising)
Pre processing
(Compression and denoising)g
Feature Extraction
(Enhanced Spectrogram with NEO
Feature Extraction
(Enhanced Spectrogram with NEO)
Speaker Database
Comparison with Speaker Database using Distance Metrics Delta BIC, KL2, T-test
Decision
Identified Speaker

Figure 3. Proposed Speaker Recognition System

V. Experiments and Results

Database Used

In this research work the database consists of 10 speakers in which 8 of them are females. The recordings are taken using regular phone under normal environmental conditions. There are three segments for each speaker with different lengths: 10, 15, and 20 seconds. We measured the distances between every pair of segments and based on the distance values, each pair of segments was judged to be from a same speaker or not.

Experimental Results

We compare our proposed distance metric-test with deltaBIC and KL2 measure using Enhanced spectrogram of speech signals of various speakers as shown in figure 4. The graph is plotted for speeches of five speakers. Among five speakers, second speaker is compared with all and found that when it is compared with itself, the distance is zero using T-test as shown in third subplot. KL2 and delta BIC also shows comparable results.

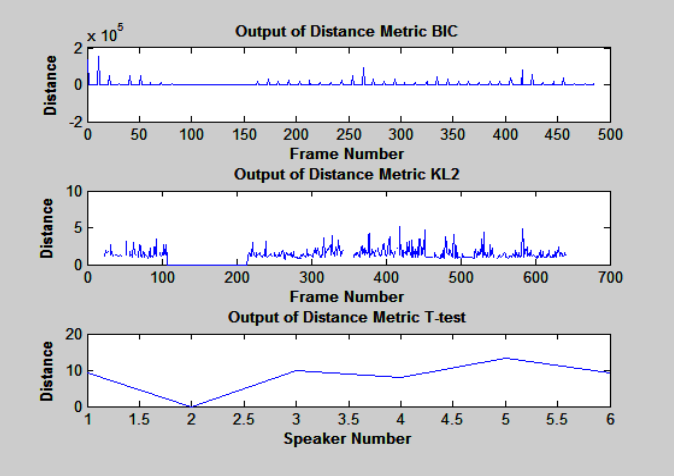

Figure 4. Comparison of Three Distance Metrics deltaBIC, KL2 and T-Test

VI. Conclusion and Future Scope

In this research work, enhanced spectrogram is proposed for feature extraction and uses it with different distance matrices for similarity measure between two speakers. From figure 4 it is concluded that T-test and delta BIC with enhanced spectrogram results better as compared to KL2. Evaluation of results using Detection Error Trade-off (DET) curves and Receiver Operating Characteristics (ROC) are the future scope of this research.

VII. REFERENCES

[1]. A. Alexander and A. Drygajlo, "Speaker Recognition : A Simple Demonstration Using".

[2]. M. W. Mak and H. B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," Comput. Speech Lang., vol. 28, no. 1, pp. 295-313, 2014.

[3]. T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition : from Features to Supervectors," 2009.

[4]. J. I. Agbinya and N. S. Wales, "Processing," pp. 1-6, 1996.

[5]. J.-D. Wu and B.-F. Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition," Expert Syst. Appl., vol. 36, no. 2, pp. 3136-3143, 2009.

[6]. A. Potamianos and P. Maragos, "multiband energy demodulation," vol. 99, no. 6, pp. 3795-3806, 1996.

[7]. N. Shokouhi, A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Robust Overlapped Speech Detection And Its Application In Word-Count Estimation For Prof-Life-Log Data Navid Shokouhi , Ali Ziaei , Abhijeet Sangwan , John H . L . Hansen Center for Robust Speech Systems (CRSS) The University of Texas at Dallas , Richar," no. 978, pp. 4724-4728, 2015.

[8]. P. Maragos, S. Member, J. F. Kaiser, T. F. Quatieri, and S. Member, "Application to Speech Analysis S :, s :," vol. 41, no. 10, pp. 3024-3051, 1993.

[9]. P. S. Gopalakrishnan, "Clustering Via The Bayesian Information Criterion With," pp. 645-648, 1998.

[10]. T. H. Nguyen, E. S. Chng, and H. Li, "T-Test Distance and Clustering Criterion for Speaker Diarization," no. 4, pp. 2-5.