International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

A Discrete Wavelet Based Approach to Speaker Identification for Different Speaking Styles.

Shanthini Pandiaraj, T. Anita Jones Mary

Abstract*: This paper presents a speaker identification system using speech signals of different speaking styles. Normal and fast speaking styles have been chosen for the analysis. The speech data provided by the CHAINS corpus has been used for the experiment. The speech utterance used for testing is compared with the speech sample in the database. Two set ups, each consisting of three groups of speakers have been used for the experiment. From each of the setup, the information contained in the frequency ranges 300-4 KHz, 300-6 KHz and300-8KHz are extracted using the Daubechies db8 wavelet and the speaker identification accuracy is compared.

***Index Terms*: speaker identification, discrete wavelet transform**

I. INTRODUCTION

Speech is a form of communication. Speaker recognition is the process of identifying the person based on the features extracted from the speech. It consists of two stages. In the first stage, features are extracted from speech. The extracted features are used for identification.

In the first stage, features that are extracted must be able to separate the speakers from one another.[1]. The block diagram of a speaker identification system is shown in Fig 1.

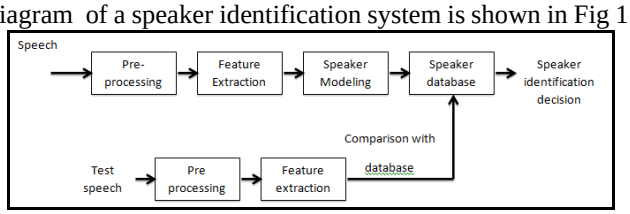

Fig. 1 Block Diagram of Speaker Identification system

The features which are normally gotten by means of Fourier Transforms (FTs), Short Time Fourier Transforms (STFTs) or Linear Predictive coding are utilized for Expert Speaker Identification (ESI). These techniques work on the assumption that a signal is stationary. They do not have the ability to analyse localized events. The LPC method assumes the speech production mechanism to be a all –pole model However, these systems are not suitable for speaker identification, since the speech signal is a non-stationary one, in other words, it is in transient state [2].Wavelet hypothesis was proposed in 1984; Goupillaud et al. presented another transform for the analysis and it is known as wavelet transform.(WT)

II. WAVELET THEORY

Wavelets convey another tool to the speech signal recognition. It tends to be said that the advantages of utilizing wavelets are local; for example the occasion is associated with when it occurs [3]. Wavelets are the shifted and scaled rendition of the first or mother wavelets. The wavelet families are ordinarily symmetrical to each other. This fact is vital since this circumstance gives computational effectiveness and simplicity of numerical usage. Different components affecting the choice of DWT over traditional strategies incorporate their capacity to decide confined highlights. The Fourier change reveals to us that a component happens some place in the sample, however where it occurs is not known. The fundamental preferred standpoint of WT is that the band of investigation can be finely balanced and the coefficients acquired from WT are appeared both the time and frequency domain.[4]

STFT utilizes a constant window length; along these lines, in light of uniform time and recurrence goals, it can't identify sudden changes and transient pieces of signs properly. In most of applications, the energies of the wavelet coefficients are utilized as features.[6]). It was suggested using the logarithm of the filer -bank energies as representation parameters.During discrete wavelet (DWT) disintegration procedurea speech is split into two frequency bands such as lower frequency band (approximation coefficients) and higher frequency band (detail coefficients) Low frequency band is used for further ddecomposition. Consequently DWT gives a left recursive binary tree structure[7]. The fundamental preferred standpoint of WT is that the band of examination can be finely balanced and the outcomes got from WT appear in both the time and recurrence domain[5]

As a result of the way that WT breaks down just the approximations of the signal, it might cause problems while applying WT in specific applications where the vital data is situated in higher frequency components [4]. For speech signals, low-frequency content is the most important part, which gives the signal identity. The high frequency content confers flavor or subtlety. In the event that the high-frequency component so of speech are removed, the voice will sound different but the speech can in any case be understood, but these information are not suitable for classifier due to a lot of information length.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Shanthini Pandiaraj*, Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamilnadu, India
T Anita Jones Mary Pushpa, Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamilnadu, India

Consequently, we have to look for a superior portrayal for the speech features. An vitality file of these sub-band signals was utilized and will be examined in the following section[2].

III. METHODOLOGY

Three diverse scope of frequencies were chosen for the analysis. The frequency ranges chosen are 300-4800 Hz,300-6000 Hz and 300-8000 Hz.

Pre-Processing

1. A 6th order Butterworth filter was utilized to bandlimit the signals.
2. The signal was first pre-emphasized applying the filter.

$$x'(t) = x(t) - \alpha x(t-1) \quad \text{---------- (1)}$$

where, α=0.97.

The objective of the filter is to improve the high frequencies of the range, which is decreased amid the speech generation process.

3.Each speech signal was isolated to casings of length 1024 samples with 512 overlapping samples. A Hamming window was connected to each frame before processing.

Feature Extraction

DWT was applied to the speech sample utilizing the Daubechies-8 wavelet decomposition filters. The speech signal was decomposed to 8 levels. Every one of the coefficients were considered for the analysis. The log energy of the coefficients was utilized as the component.

Speaker Modelling

The features were fed to a Gaussian Mixture Model (GMM). GMM is widely utilized for classification in speaker identification experiments. This is on the grounds that GMM can easily estimate the density of the data clusters. The model was trained with speech signals in the Normal style. Its exactness and execution are known to be exceptional.

Testing Phase

In this phase speech signals in Normal style were used. The normal of 10 readings was taken as the speaker recognizable proof rate.

IV. DATABASE

Speech gives a rich flag which is broadly accepted to contain adequate data to exceptionally distinguish an individual. This data might be generally assembled into two separate classes.

1. Static data emerging from the subtleties of vocal tract life structures
2. Dynamic data which is accessible through the demonstration of talking.

It is critical to see how change in talking style impacts speech. It is fundamental to comprehend which attributes of a people voice stays invariant. The CHAINS corpus is a discourse database explicitly intended to help portray speakers as people. The corpus contains chronicles of 36 speakers acquired in two unique sessions with a period partition of two months (F. Cummins, 2006).

The NORM condition which has a place with the primary session and FAST style which has a place with the second session were utilized in the examination. From the discourse material, CSLU's phonetically rich expressions sentences 1-9 were utilized for testing. The TIMIT sentences were utilized for preparing.

Every one of the accounts are available as 16 bit PCM encoded WAV records with a testing rate of 44.1kHz.

Two sets of information were utilized for the examination. Each set comprises of three groups of 8,12 and16 speakers individually. In both the sets, equal number of male and female speakers were used. In the main set, voice tests of speakers having a place with various geographic areas were utilized so as to test the capacity of the speaker distinguishing proof framework to adjust to various accents.

The separation subtleties of speakers picked for the main set is as per the following:

8 speakers - 2 male and 2female from eastern part of Ireland.
2 male and 2 female from UK and California USA

12 speakers- 3 male and 3 female from eastern part of Ireland
3 female from California 2 male from UK and 1 from USA

16 speakers - 4 male and 4 female from eastern part of Ireland.
3 female from California , 1from UK , 2 male from UK
what's more, 2 from USA

The second set utilized voices of speakers having a place with the equivalent geographic area yet talking at various paces.

The separation subtleties of speakers picked for the second set is as per the following:

8 speakers - 4 male and 4 female speakers from eastern part of Ireland

12 speakers - 6 male and 6 female from eastern part of Ireland and

16 speakers - 8 male and 8 female from eastern part of Ireland.

From the speech data given in CHAINS corpus, sentences s1-s9 were utilized for testing. Sentences s10 to s33 were utilized for preparing the GMM. [8].

V. RESULTS AND DISCUSSIONS

The accuracy of GMM classifier was assessed utilizing speech samples recorded in NORM style and FAST style. Results were acquired changing the quantity of speakers, the measure of training material utilized and the quantity of Gaussian components in the GMM. Features were extracted using Daubechies wavelet db 8.The features were extracted in the frequency range 300-4800Hz.

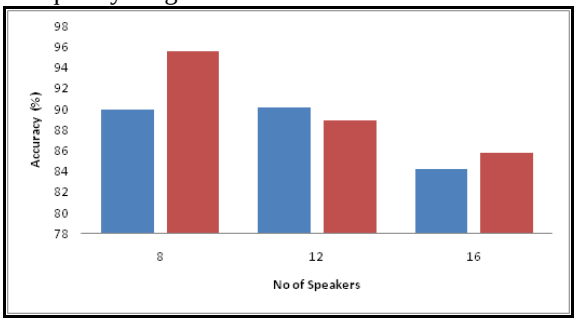

Fig. 2 shows the accuracy of GMM classifier varying the number of speakers

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Fig. 2 demonstrates the exactness of GMM classifier changing the quantity of speakers. The length of preparing material was 30 seconds and the length of test material was 10seconds. The quantity of Gaussian parts was 12.If the quantity of speakers was expanded keeping the measure of preparing material steady the precision of the framework decreased. NORM style was utilized for preparing and testing.

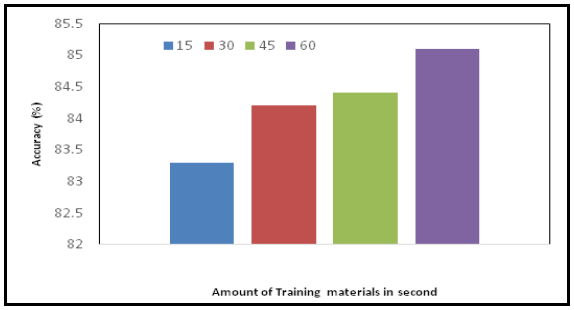

Fig. 3 SOLO: 300-4800 Hz Varying amount of training material

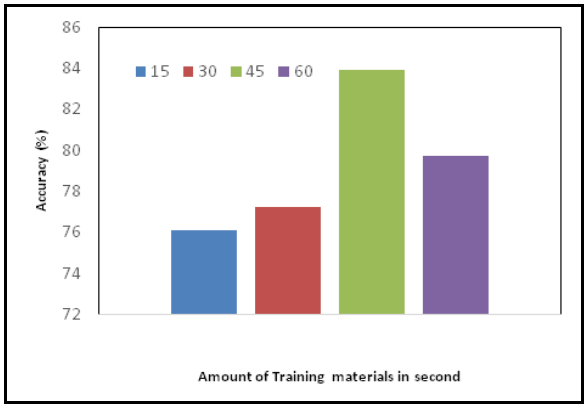

Fig. 4 FAST: 300-4800 Hz Varying amount of training material

Fig. 3 and 4 abridge the precision of the classifier varying the amount of preparing material. The number of speakers was 16. The quantity of Gaussian segments was 12. The speech signals were within 300-4800 Hz. It was observed that, as the amount of training material increased, the accuracy of identification also increased.

In the following part of the investigation, the quantity of Gaussian segments was differed. The quantity of speakers was set to 16.30seconds of training material and 10 seconds of testing material were utilized. The experiment was performed for the three frequency ranges. The identification accuracy using NORM style and FAST style have been tabulated in Table 1 and Table 2 individually. It was seen that the Gaussian parts G8,G12 and G16 gave better execution.

Table 1 Speaker identification accuracy using NORM style of speech.

No of Gaussian	300 – 4800 Hz	300- 6 kHz	300 – 8kHz
G4	59.7	63.9	65.3
G8	81.8056	84.2	85.4
G12	84.2	83.6	83.8
G16	80.3	84.4	84.3
G32	70.4	74.2	75.9
G64	60.1	55.9	63.5

Table 2 Speaker identification accuracy using FAST style of speech.

No of Gaussian	300 – 4800 Hz	300- 6 kHz	300 – 8 kHz
G4	76.7	80.7	80.4
G8	73.9	75.7	77.5
G12	77.2	77.4	78.9
G16	76.7	78.2	81.1
G32	73.5	70.4	76.7
G64	54	61.8	67.5

A. Role of different frequencies

In this segment, highlights were extricated in three diverse recurrence interims specifically 300-4800 Hz, 300-6kHz and 300-8kHz. For testing purposes, the length of speech material utilized for training was 15 seconds and 10 seconds of discourse material for testing. The number of Gaussian components was fixed at 12.The outcomes are shown in Table 3. It was seen that the precision of set1 was more compared to set2 for NORM style of discourse. It was likewise seen that the accuracy of identification was more in set2 for FAST style.

Table 3. Role of different frequencies for 15seconds of training material and 10 seconds of testing material.

No.of speakers		300 – 4800 Hz		300 – 6 kHz		300 – 8 kHz	
		SOLO	FAST	SOLO	FAST	SOLO	FAST
Set 1	8	91.1	85.6	87.8	84.7	88.3	83.1
	12	85.4	80	85.2	82.4	85	80.9
	16	79.2	76.1	91.7	75.3	75	79.9
Set 2	8	90	94.7	90.8	95	90.6	94.7
	12	83.9	78.9	82.9	80.9	82.8	77.4
	16	81.7	75.9	82.2	81.4	81.3	81.7

B. Role of different frequencies for 30 seconds of training material.

The duration of training material was fixed at 30 seconds. 10 seconds of speech material was used for testing. The number of Gaussian components used were 12.The results are shown in Table 4.

Table 4. Role of different frequencies for 30seconds of training material and 10 seconds of testing material.

No.of speakers		300 – 4800 Hz		300 – 6 kHz		300 – 8 kHz	
		Solo	Fast	Solo	Fast	Solo	Fast
Set 1	8	90	85.83	94.4	86.94	93.3	85
	12	90.19	78.52	91.3	82.96	88.3	82.4
	16	84.17	77.2	83.6	77.36	83.75	78.9
Set 2	8	95.56	94.7	93.3	95.6	92.2	94.4
	12	88.89	80	88.15	81.5	83.15	84.1
	16	85.83	79.7	83.06	84.17	84.86	84.0

C. Role of Different

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

D. Frequencies for 60 seconds of training material.

The duration of training material was 60 seconds. 10 seconds of speech material was used for testing.The number of Gaussian components were fixed at 12.

Table 5. Role of different frequencies for 60seconds of training material and 10 seconds of testing material.

No.of speakers		300 – 4800 Hz		300 – 6 kHz		300 – 8 kHz	
		SOLO	FAST	SOLO	FAST	SOLO	FAST
Set 1	8	92.2	88.61	93.89	90.56	93.3	90
	12	92.03	81.67	92.59	84.4	91.85	86.48
	16	82.5	79.58	86.1	81.25	86.1	82.08
Set 2	8	94.17	94.4	91.85	96.39	92.77	95.3
	12	89.1	83.3	89.26	83.52	88.3	82.78

VI. CONCLUSION

The experimental evaluation indicates that in all the three frequency ranges, the identification rate increases with increase in training material. The experimental study returned maximum efficiency of speaker identity in the 300-6000 Hz frequency range for solo style. Similarly, the fast style showed optimum identification of speakers in the 300-8000 Hz frequency range. This shows the presence of speaker identification characteristics at high frequencies. It was also observed on experimenting with the number of Gaussian components used, that using 8, 12 or 16 Gaussian components yielded the most optimal results.

REFERENCES

1. Jian-Da Wu, Bing-Fu Lin, Speaker identification using discrete wavelet packet transform technique with irregular decomposition, Expert Systems with Applications 36 (2009) 3136 – 3143.
2. Engin Avci, Zuhtu Hakan Akpolat, Speech recognition using a wavelet packet adaptive network based fuzzy inference system, Expert Systems with Applications 31 (2006) 495 – 503.
3. Derya Avci, An expert system for speaker identification using adaptive wavelet sure entropy, Expert Systems with Applications36(2009) 6295-6300
4. Sami Ekici,selcuk Yildirim, Mustafa Poyraz, Energy and entropy – based feature extraction for locating fault on transmission lines by using neural network and wavelet packet decomposition, Expert Systems with Applications 34 (2008) 937 – 2944
5. Shung-Yung Lung, Wavelet feature selection based neural networks with application to the text independent speaker identification, Expert Systems with Applications 39 (2006) 1518 – 1521.
6. M.Hariharan,C.Y.Fook,R.Sindhu,Abdul Hamid Adom,Sazali Yaacob,Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy,Digital Signal Processing 23(2013)952-959
7. Hamid Reza Tohidypour, Seyyed Ali, Seyyyedsalehi, Hossein Behbood,Hossein Roshandel, A new representation for speech frame recognition based on redundant wavelet filter banks, Speech comunication 54(2012) 256-271
8. F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains corpus: Characterizing individual speakers," in Proc. SPECOM'06, St. Petersburg, Russia, 2006, pp. 431-435.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)