Middle-East Journal of Scientific Research 24 (11): 3527-3541, 2016
ISSN 1990-9233
© IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.3527.3541

A Perfect Balance of Sparsity and Acoustic Hole in Speech Signal and its Application in Speaker Recognition System

Satyanand Sing and Mansour H. Assaf

Department of ECE, CMRIT, School of Engineering and Physics,
University of the South Pacific, Suva, Fiji, Hyderabad, India

Abstract: This paper addresses the issue of balancing the acoustic hole and sparsing the speech signal enrollment for training and testing in Automatic Speaker Recognition (ASR) system. Sparsing techniques involve the representation of a small number of coefficients that hold a large amount of the energy. Sparsity can play a major role in resolving the issue of dealing with big data in ASR by applying speech compression techniques and information storage in databases. Spectral domain compression of the speech signal using novel sparsing algorithms that balance the sparsity of speech signal with the acoustic hole is proposed. The speech signal is converted to a spectral domain using the Discrete Rajan Transform (DRT) and only first and mid-spectrum component in each block of size 8x1 retained forcing the remaining component to zero. The speech signal spectrum can be maximally compressed at 8:1 ratio to the unique one with balancing acoustic hole and synthesized speech signal, which can be used in ASR systems. A balanced spectrally compressed speech signal can be stored in database as a speaker representative and during training and testing time it can be synthesized using the Inverse Discrete Rajan Transform (IDRT). Simulation results, shows acceptable speech signal spectral compression that balances sparsity and the generation of the acoustic hole is 75% with 94.8% efficiency without sparsification and 99.1% efficiency with TIMIT database respectively.

Key words: Discrete Rajan Transform · Inverse Discrete Rajan Transform · Compressive sensing · Cumulative Point Index · Error Dynamic Range · Gaussian mixture model · Universal background model · Linear discriminate analysis · Probabilistic linear discriminant analysis

INTRODUCTION

The sparse speech data training and testing in ASR result from a big challenge because of an absence of acoustic phonemes scope in the speaker space compared with more conversational speech data. In this way, it is profoundly likely that phoneme bungle exists between the restricted prepared acoustic space and info test grouping. We called this marvel "acoustic holes" in the acoustic model space. Speaker recognition robustness in adverse condition has been investigated widely in recent years [1]. There are quite a number of factors affecting the automatic speaker recognition performance including channel/session variability and noise/reverberation. In real-world applications dealing with the mismatched condition is inevitable and any type of mismatch between training and test session will potentially result in degraded performance. Based on the type of the data in national institute of standards and technology (NIST) speaker recognition evaluations, the researchers in speaker recognition field have successfully developed techniques to deal with session/channel variability [2].

Although the state-of-the-art algorithms sensitivity to unseen channel or session variability is partially mitigated, they are highly vulnerable to additive noise and reverberant environment [3]. It has also been shown that even the performance of the state-of-the-art speaker recognition systems degrades substantially when limited speech is available in testing phase [4]. Although there are recent studies to handle reverberation and additive noise in feature and model domain for speaker recognition systems, the compensation techniques with respect to noise and reverberation for speaker recognition systems are still an open question.

Corresponding Author: Satyanand Sing, Department of ECE, CMRIT, School of Engineering and Physics, University of the South Pacific, Suva, Fiji, Hyderabad, India.

Since our civilization, the speech is pure and natural means of human communication. Let us take an example for speaker recognition, a human recognizes a speaker regardless of the text spoken without of any effort for him/her to understand what exactly text spoken by different speakers. Human speech signal carries linguistic information as a major component as well as non-verbal information as a minor component. Based on speaker-specific features of acoustic speech signal a listener can identify his/her gender of the speaker, approximate age and emotional state. In the human being, there is the effective way to automatically extract speaker-specific information from speech signals; the same concept has to be used in automatic speaker recognition by machine. The interference of redundancy in speech signal components hampers a speech signal or speaker recognition system performance [5].

The speech signal characteristically generated vocal tract which is a resonant system otherwise by physical impacts or occasionally together. Speech signal generated by vocal tract as resonant systems contain the number of redundant frequency components, therefore, if the speech signal which is going to be used in ASR is transformed in the spectral domain, a comparatively high degree of sparsity can be obtained. Taking into consideration that speech signal generated by physical impacts then it can be experimentally observed that the largest part of the speech signal is concentrated on time. This observation of speech characteristics permits superior sparse demonstration of the speech signal in time domain. In such kind of situation of the speech signal, Wavlets transform is most suitable for sparsification of the speech signal. Hence, the perception of sparse representation and sparsity in speech processing and ASR is very effective [6].

The compressive sensing (CS) concept in sparsing can be utilized in a number of applications, particularly in speech signal processing that is speech pre-conditioning, Signal to Noise Ration (SNR) improvement and speech coding [7]. Though sparsing is the latest technology, very little research has been done on the application of sparsing on speech signal and its utilization in ASR. In these manner, significant difficulties to apply sparsification in speech signal processing to balance the acoustic hole begins with, finding a good sparse basic and development most efficient measurement matrices [8].

Sparse Representation with Discrete Rajan Transform (DRT) and Inverse Discrete Rajan Transform (IDRT): Rajan Transform (RT) demonstrates a function ϕ: $G \rightarrow H$ is a homomorphism nature if for all g_1, g_2 in G, $(g_1, g_2)\phi = (g_1)\ \phi\ (g_2)\ \phi$ and it is transformation invariant in speech signal. Due to homomorphism nature, it has many applications in image processing like detection of the curve, detection of lines, detection of contour and detection of edge and image point isolation. If signal sequences are highly correlated then error in reconstructed signal is less and vice versa with the application of DRT. Due to the highly correlated non-stationary nature of the speech signal, the DRT plays a very important task in terms of spectral sparsification, compression and original speech signal reconstruction. A U dimensional speech signal vector ‘‘d’’ can be represented as $U = 2^u$ with u being a nonnegative integer. Consider a speech signal d(u), apply DRT on signal then spectrum $D(r)$ can be obtained after u steps. The time domain speech signal can be converted into the spectral domain with a unique operating matrix of dimension $\left(\frac{U}{2^{r-1}} \times \frac{U}{2^{r-1}}\right)$ denoted as Y_r. This unique operation matrix construction is defined as;

$$Y_r = \begin{bmatrix} I_w & I_w \\ -e_r^1, I_w & e_r^1, I_w \end{bmatrix} \tag{1}$$

I Indicates the *with* order identity matrix. For example at r steps the order of identity matrix is $w_r = \frac{U}{2^r}; r \in \{1,2,...,n\}$ and e_r^1 is the “supplementary information” which indicates the equilibrium state condition of the signal during spectrum generation.

There will be a certain inherent phasor relation with ‘supplementary information’ e_r between the sample points 1st and 5th.

$$e_r^i = \begin{cases} -1, & d_r^i(w_r + 1) < d_r^i(1) \\ 1, & otherwise \end{cases} \tag{2}$$

where I = $\{1,2,...,2^{r-1}\}$. At every step r, let F_r denoted as output sequence and it is obtained as:

$$F_r = Y_r D_r = [f_r^1 \quad f_r^2 ... \quad f_r^i] \tag{3}$$

In eqn. (3) at every steps F_r has got $2^r p_r$ elements When r = 1, D_1 = d, at every steps the equilibrium segments are considered for r > 1, 2^{r-1}.

Y_r Is the operator matrix can be constructed at a stage r using supplementary information e_r? Additionally if r > 1 then the output can be restructure in equilibrium segments and it can be defined as:

$$D_{r+1} = [\overline{d}^{1}_{r+1} \quad \Upsilon^{1}_{r}.\overline{d}^{2}_{r+1}\ldots \quad d^{i-1}_{r}.\overline{d}^{i}_{r+1}] \tag{4}$$

where $\Upsilon_r = [\Upsilon^1_r \quad \Upsilon^2_r \ldots \quad \Upsilon^{i-1}_r]$ and

$$\Upsilon^{i-1}_{k} = e^{1}_{r} \times e^{i}_{r} \text{ for } r > 1 \tag{5}$$

Also,

$$D_{r+1} = \begin{bmatrix} f^i_r(1) & f^i_r(2)\ldots & f^i_r(p_r) \\ f^i_r(w_r+1 & f(w_r+2) & f^i_r(2w_r) \\ \vdots & \vdots & \vdots \\ f^i_r(2^{r-1}w_r+1) & \cdots & f^i_r(2^r w_r) \end{bmatrix}$$

$$= [\overline{d}^{i}_{r+1} \quad \overline{d}^{2}_{r+1}\ldots \quad \overline{d}^{i}_{r+1}] \tag{6}$$

D_{R+1} express that signal spectrum into equilibrium segments. Steps will be continuing till the final DRT spectrum is obtained after u steps. As explained already, DRT is a homomorphic function and it also exhibits the isomorphism property when the complementary phasor information is preserved. Since DRT is also viewed as an isomorphic function, one should be able to retrieve the original signal data from its DRT spectrum by means of its inverse transform. Indeed, the IDRT is used to retrieve the input data with the help of e_k^1 and μ_k. Now the DRT operator R_k is obtained using the values of e_k^1 and μ_k. The general expression used to retrieve intermediate signal data at every stage is [9]:

$$\tilde{D}_t = \frac{1}{2}[Y_t F_t] = [d^1_t \quad d^2_t \ldots \quad d^i_t]^T \tag{7}$$

where $t = \{r, r-1, \ldots 1\}$.

As on account of forward DRT calculation wherein the succession is part into balance portions, on account of IDRT calculation, the sections are recombined and data arrangement recovered iteratively.

When $a = r$, $F_t = F_r$ then we can obtained final stage spectral domain signal and for $t > r$,

$$F_{t-1} = [\overline{F}_t(1) \quad \Upsilon^1_r.\overline{F}_t(2)\ldots \quad \Upsilon^{i-1}_r.\overline{F}_t(i)] \tag{8}$$

$$\overline{F}_{a-1} = \begin{bmatrix} d^i_a(1) & d^i_a(2w_r+1)\cdots & d^i_a(2^{r-1}w_r+1) \\ d^i_a(2) & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ d^i_a(2w_r) & d^i_a(2^2 w_r) & d^i_a(2^r w_r) \end{bmatrix} \tag{9}$$

During the IDRT computation the original input speech signal can be obtained.

DRT Sparsification, Compression and Decompression Application on Speech Signal: The experiments are performed on a speech signal is taken from TIMIT and NIST database. Speech signal of male and female taken for 3 sec with the sampling frequency 16 KHz. This experiment is conducted on MATLAB with i3 Intel Core Processor Clock Frequency at 2.53 GHz. The entire speech signal is divided into a number of blocks. Every block contains 8 samples. Here, DRT will be applied to speech signal, it will be Sparsified, compressed, stored and whenever the speech signal required IDRT will be applied to reconstruct the original speech signal.

Fig. 1. shows the block diagram representation of speech production process [10]. The speech production process begins when a speaker formulates a message in his/her mind to transmit to the listener via speech communication. The next steps in the process are the conversion of the message into language code. This corresponds to converting the message into a set of phoneme sequences corresponding to the sounds that make up the words, along with prosody (syntax) markers denoting duration of the sounds, loudness of sounds and pitch associated with the sounds. The acoustic speech is produced at 64kbit/s but in order to understand the information we need only 50bit/s. Once speech signal is produced the recognition parts start at 64kbit/s continuous signal then spectrum analyzer and feature extraction came to 2kbit/s by doing this we have removed a lot of redundancy. Once the feature extracted we need to do the language translation which is discrete in nature. In order to understand the information, only 50bit/s information is required. To understand the information from acoustic wave form we need 50 bit/sec information. But the acoustic wave form generated from vocal tract system used to be 30-64 kbit/s. DRT is a very powerful tool to remove redundancy from acoustic waveform.

The DRT algorithms based sparsification, compression, decompression steps are as given below:

- Read wave files.
- Select beginning 48128X1 sizes of speech data.
- Convert speech data into 8X6016 sizes blocks.
- Apply DRT on all 6016 blocks.
- Keep Cumulative Point Index (CPI) and mid frequency component (the 5th component of each block) and force all other components to zero.
- Preserve CPI and mid frequency component and discard remaining components.

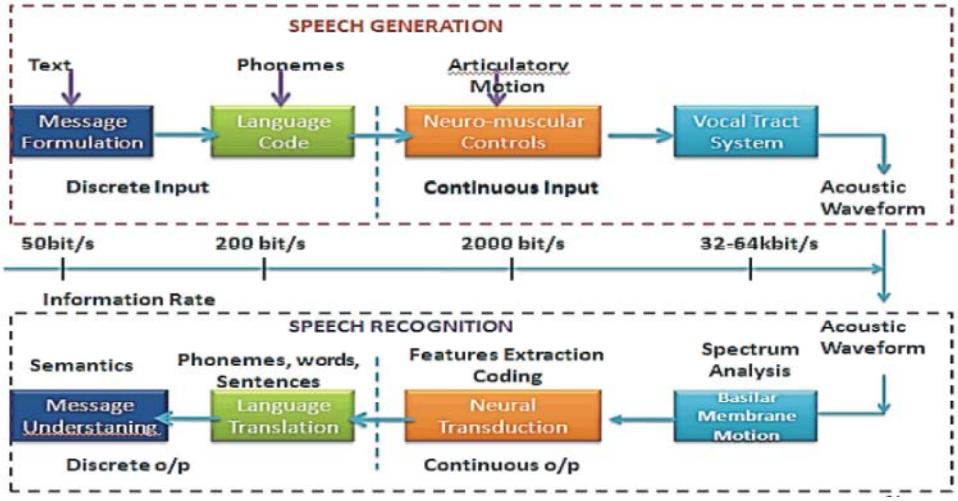

Fig. 1: The Speech production chain

Table 1: Real time speech signal in discrete sequence d(u) of length 64

	S1	S2	S3	S4	S5	S6	S7	S8
B1	-0.01379	0.00461	0.02484	0.03485	0.03903	0.04227	0.04269	0.04135
B2	0.03622	0.02921	0.02109	0.01840	0.02112	0.02051	0.01794	0.00565
B3	-0.01263	-0.02914	-0.03036	-0.02661	-0.02188	-0.01910	-0.01825	-0.02017
B4	-0.01913	-0.01575	-0.01544	-0.01343	-0.00946	-0.00201	0.00699	0.01447
B5	0.01547	0.01031	0.00449	0.00131	0.00049	0.00177	0.00504	0.00406
B6	-0.00180	-0.00815	-0.01065	-0.01447	-0.02026	-0.02347	-0.02332	-0.02151
B7	-0.02188	-0.02130	-0.02695	-0.03314	-0.04178	-0.04562	-0.04996	-0.05539
B8	-0.06128	-0.05380	-0.03781	-0.01907	0.01581	0.11163	0.19968	0.16220

Table 2: D(r) The spectrum of d(u)

	S1	S2	S3	S4	S5	S6	S7	S8
B1	0.21585	0.03030	0.07162	-0.01297	0.11484	-0.02652	-0.06613	0.00381
B2	0.17014	-0.02261	-0.04398	-0.00735	-0.03970	-0.00320	0.00790	-0.01602
B3	-0.17816	-0.01190	-0.01263	0.01556	0.01935	0.01361	0.01776	-0.02496
B4	-0.05377	0.02032	0.03894	-0.00134	0.07373	0.00952	0.02692	0.00140
B5	0.04294	-0.00803	-0.01315	-0.00027	-0.02023	0.00864	0.02682	-0.00424
B6	-0.12363	-0.01157	-0.01627	0.00754	-0.05350	0.00876	0.01407	0.00247
B7	-0.29602	-0.01489	-0.03485	-0.00836	-0.08948	-0.00366	-0.00104	0.00519
B8	0.31735	0.08456	0.29263	-0.12204	0.66129	0.03214	0.17624	-0.14456

- Store CPI and mid frequency components as a representative of speaker for ASR application.
- Stored CPI and mid frequency components sequence are the sparsified spectral sequence
- Apply IDRT to reconstruct the time domain speech signal for ASR application
- Compute Mean Square Error, Signal to Noise Ration and PESQ for the reconstructed speech signal with reference to an original speech signal.

The DRT algorithms have been compared with similar algorithms like DCT, DFT and DWT in the following section.

Application of DRT on Speech Signal of 64 Sample Size: A 3 sec speech signal of the male from TIMIT database has 62634 samples. Before applying DRT we need to take sample size which is divisible by 8. Let us take a sample of 48128 and divide it into 8X1 blocks. Now we have total 6016 number of blocks of size 8X1 and DRT applied on every block. A real- time speech signal $d(u)$ of sample size 64 was taken and DRT is applied in the block wise fashion, the corresponding spectrum of the blocks is obtained as $D(r)$.

For instance, let us consider a specimen real-time speech signal in the discrete sequence $d(u)$ of length 64 shown in Table 1. Let each block represented by B and Sample S respectively.

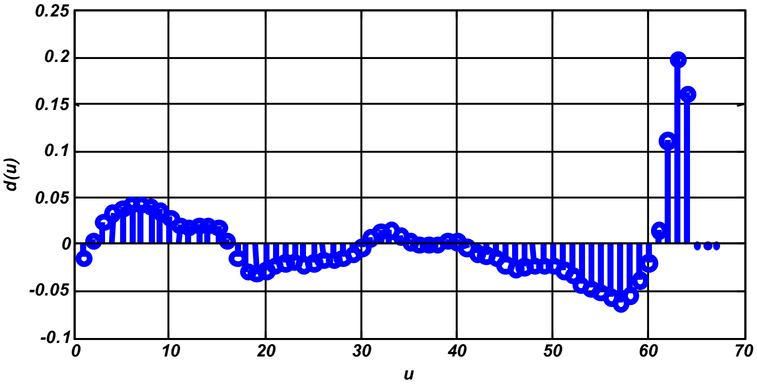

Fig. 2: Plot of 64 points original speech signal d(u)

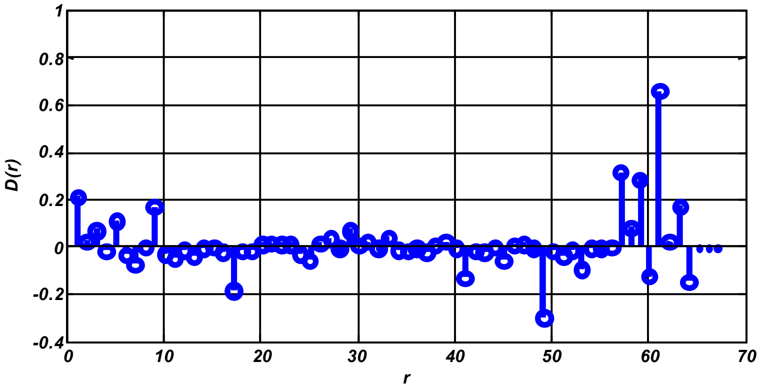

Fig. 3: Plot of D(r) the spectrum of d(u)

DRT is applied to $d(u)$ in block-wise fashion and equivalent spectral blocks obtained as $D(r)$ is shown in Table 2.

$D(r)$ are the spectral component of the original speech signal $d(u)$ and the 1st component of each block having high magnitude compared to remaining? These components are also called CPI and carrying all speech intelligence. The first component of each block can be retained and remaining components can be simply discarded. Fig. 2. Shows the plot 64 points of the original speech signal $d(u)$ and Fig. 3. Shows the plot of $D(r)$.

Sparsing of Speech Data by Retaining CPI Alone: $D(r)$ Is the spectral domain signal and can be sparsed by keeping the CPI alone in each block of length 8 and compelling remaining components to 0.

At that point, the sparse spectrum can be represented by D_{s1} (r) and shown in Table 3.

Compressing Speech Data with CPI Alone: D_{s1} (r) Is the spectral domain sparsed speech data having only 8 non-zero elements. D'_{s1} (r) Is the compressed version of D_{s1} (r) after ignoring all 56 samples of spectral components of zero values? Table 4. Represents the compressed spectrum speech signal of size 8X1 instead of 64X1.

The compressed speech signal D'_{s1} can be stored in a database as a representative biometric vector of a speaker. The scale of compression and hence sparsity acquired by keeping the first component of the spectral is 12.5%.

Fig. 4. Shows the plot of $D_{s1}(r)$ the sparsed spectral sequence and Fig. 5 shows the compressed form of sparsed spectral sequence D'_{s1} (r).

Sparsing of Speech Data by Retaining Cpi and mid Frequencies Alone: In this case $D(r)$ the spectrum of original speech signal is sparsed by keeping the CPI and the mid frequency segment in each block of length 8 and driving remaining components to 0. Mid frequency component = [(total number of samples in a block/2) +1] = [(8/2) +1] =5th sample of each block will be treated as mid frequency component and it will be retained along with CPI. At that point, the sparse speech data sequence is D_{s2} (r) represented in Table 5.

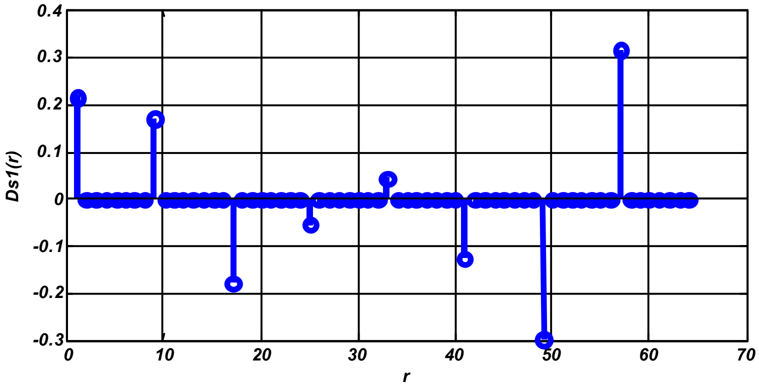

Fig. 4: Plot of $D_{s1}(r)$ sequence

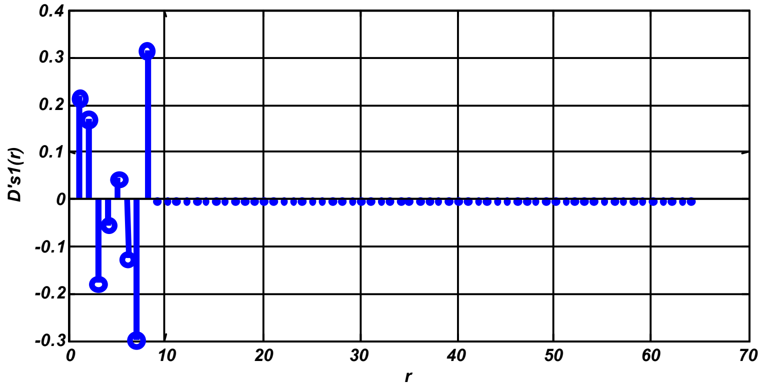

Fig. 5: Plot of $D'_{s1}(r)$ sequence

Table 3: The sparsed spectrum $D_{s1}(r)$

	S1	S2	S3	S4	S5	S6	S7	S8
B1	0.21585	0	0	0	0	0	0	0
B2	0.17014	0	0	0	0	0	0	0
B3	-0.17816	0	0	0	0	0	0	0
B4	-0.05377	0	0	0	0	0	0	0
B5	0.04294	0	0	0	0	0	0	0
B6	-0.12363	0	0	0	0	0	0	0
B7	-0.29602	0	0	0	0	0	0	0
B8	0.31735	0	0	0	0	0	0	0

Table 4: Compressed spectrum $D'_{s1}(r)$

	S1	S2	S3	S4	S5	S6	S7	S8
B1	0.21585	0.17014	-0.17816	-0.05377	0.04294	-0.12363	-0.29602	0.31735

Table 5: The sparsed spectrum $D_{s2}(r)$

	S1	S2	S3	S4	S5	S6	S7	S8
B1	0.21585	0	0	0	0.11484	0	0	0
B2	0.17014	0	0	0	-0.03970	0	0	0
B3	-0.17816	0	0	0	0.01935	0	0	0
B4	-0.05377	0	0	0	0.07373	0	0	0
B5	0.04294	0	0	0	-0.02023	0	0	0
B6	-0.12363	0	0	0	-0.05350	0	0	0
B7	-0.29602	0	0	0	-0.08948	0	0	0
B8	0.31735	0	0	0	0.66129	0	0	0

Table 6: Compressed spectrum $D'_{s2}(r)$

	S1	S2	S3	S4	S5	S6	S7	S8
B1	0.21585	0.11484	0.17014	-0.03970	-0.17816	0.01935	-0.05377	0.07373
B2	0.04294	-0.02023	-0.12363	-0.05350	-0.29602	-0.08948	0.31735	0.66129

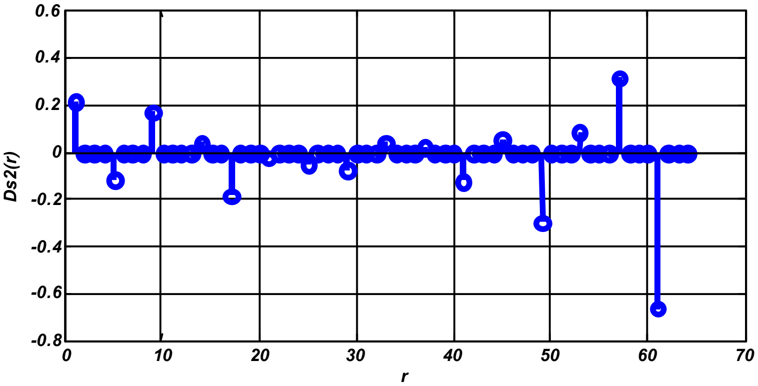

Fig. 6: Plot CPI and mid frequency components $D_{s2}(r)$

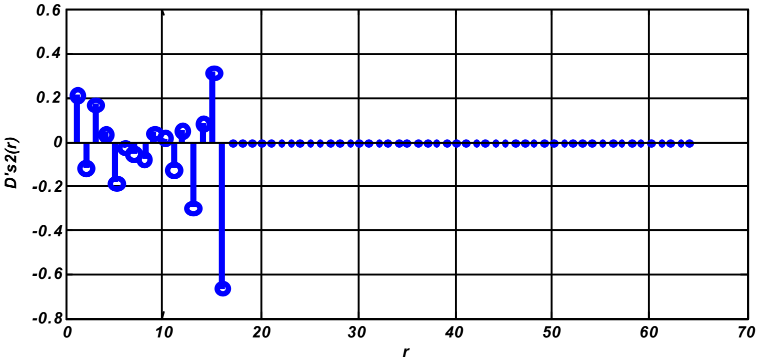

Fig. 7: Plot compressed CPI and mid frequency components $D'_{s2}(r)$

Compressing Speech Data with CPI and mid Frequency Alone: $D_{s2}(r)$ Is the spectral domain sparsed speech data having only 16 non-zero elements. D'_{s2} (r) Is the compressed version of $D_{s2}(r)$ after ignoring all 48 samples of spectral components of zero values? The D'_{s2} (r) is presented in Table 6.

D'_{s2} (r) Can be stored in a database as a representative biometric vector of a speaker of size 16X1 instead of 64X1. The scale of compression and hence sparsity acquired by keeping the first and mid frequency component of the spectral is 25%.

Fig. 6 shows the plot of D_{s2} (r) the sparsed spectral sequence with CPI and mid frequency component and Fig. 7. Compressed form of sparsed spectral sequence D'_{s2} (r).

Decompressing the Speech Signal from $D'_{s1}(r)$ and $D'_{s2}(r)$ with IDRT: Amid the speaker recognition testing stage, D'_{s1} (r) and D'_{s2} (r) is uncompressed to acquire D_{s1} (r), D_{s2} (r) in case of 12.5% and 25% of sparsity respectively. Presently, IDRT algorithm applied to D_{s1} (r) and D_{s2} (r) to reconstruct the speech signal which can be used during testing and training phase of ASR. Here we can rename the reconstructed speech signal as d'_1 (u) and d'_2 (u) respectively. Table 7 Represents the time domain reconstructed speech signal d'_1 (u) from D'_{s1}.

Table 8 Represents the time domain reconstructed speech signal d'_2 (u) from D'_{s2} (r)

Fig. 8 shows original 64 samples speech signal $d(u)$ and IDRT reconstructed speech signal d'_1 (u) represented in the same plot. Fig. 9 shows original 64 samples speech signal $d(u)$ and IDRT reconstructed speech signal $d'_2(u)$ represented in the same plot.

Fig. 10 Shows original speech signal $d(u)$ and IDRT reconstructed speech signal $d'_1(u)$ exhibited in the same plot of 48128 sample size of speech data. Likewise, $D(r)$ is sparsed by holding CPI values alone recurrence parts of all the 6016 blocks. Because of this sparsing, 12032 unearthly values would involve 12.5% of the real memory dispensed to oblige 48128 examples.

Table 7: Reconstructed speech signal $d'_1(u)$ from $D'_{s1}(r)$

	S1	S2	S3	S4	S5	S6	S7	S8
B1	0.02698	0.02698	0.02698	0.02698	0.02698	0.02698	0.02698	0.02698
B2	0.02127	0.02127	0.02127	0.02127	0.02127	0.02127	0.02127	0.02127
B3	-0.02227	-0.02227	-0.02227	-0.02227	-0.02227	-0.02227	-0.02227	-0.02227
B4	-0.00672	-0.00672	-0.00672	-0.00672	-0.00672	-0.00672	-0.00672	-0.00672
B5	0.00537	0.00537	0.00537	0.00537	0.00537	0.00537	0.00537	0.00537
B6	-0.01545	-0.01545	-0.01545	-0.01545	-0.01545	-0.01545	-0.01545	-0.01545
B7	-0.03700	-0.03700	-0.03700	-0.03700	-0.03700	-0.03700	-0.03700	-0.03700
B8	0.03967	0.03967	0.03967	0.03967	0.03967	0.03967	0.03967	0.03967

Table 8: Reconstructed speech signal $d'_2(u)$ from $D'_{s2}(r)$

	S1	S2	S3	S4	S5	S6	S7	S8
B1	0.01263	0.01263	0.01263	0.01263	0.04134	0.04134	0.04134	0.04134
B2	0.02623	0.02623	0.02623	0.02623	0.01630	0.01630	0.01630	0.01630
B3	-0.02469	-0.02469	-0.02469	-0.02469	-0.01985	-0.01985	-0.01985	-0.01985
B4	-0.01594	-0.01594	-0.01594	-0.01594	0.00249	0.00249	0.00249	0.00249
B5	0.00790	0.00790	0.00790	0.00790	0.00284	0.00284	0.00284	0.00284
B6	-0.00877	-0.00877	-0.00877	-0.00877	-0.02214	-0.02214	-0.02214	-0.02214
B7	-0.02582	-0.02582	-0.02582	-0.02582	-0.04819	-0.04819	-0.04819	-0.04819
B8	-0.04299	-0.04299	-0.04299	-0.04299	0.12233	0.12233	0.12233	0.12233

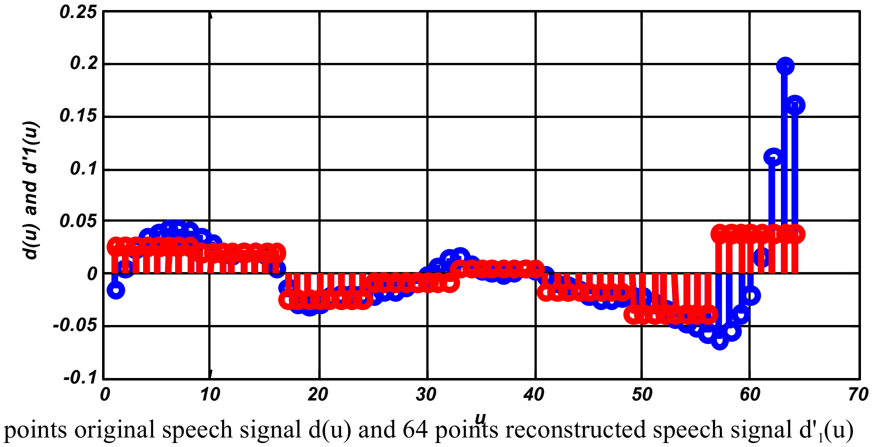

Fig. 8: Plot of 64 points original speech signal d(u) and 64 points reconstructed speech signal $d'_1(u)$

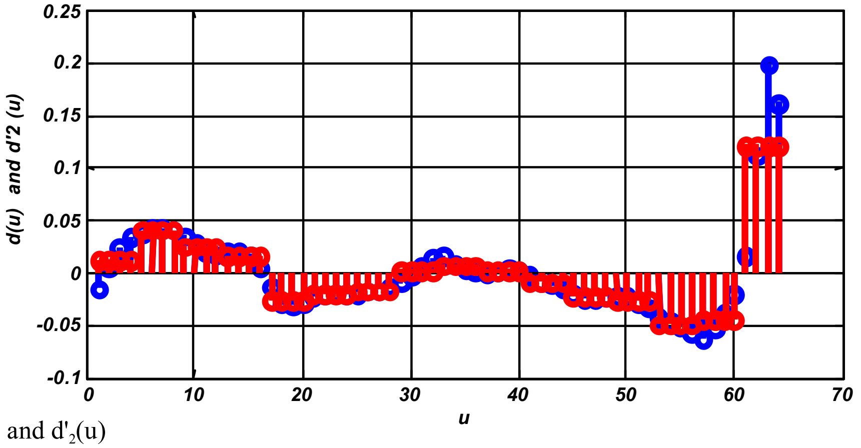

Fig. 9: Plot of d(u) and $d'_2(u)$

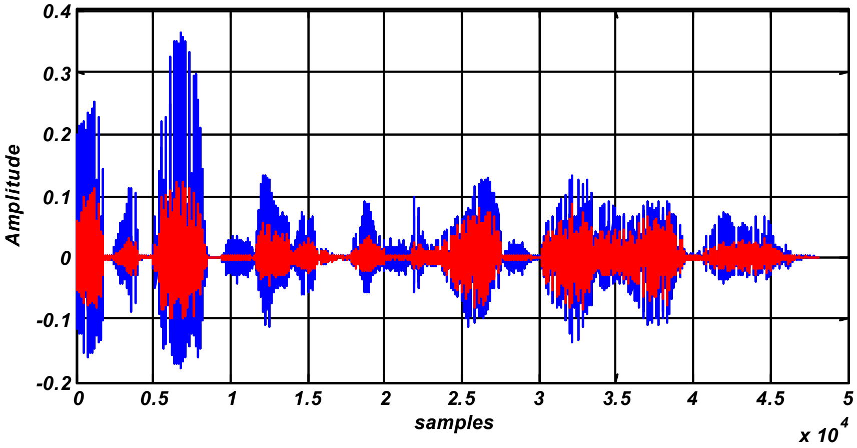

Fig. 10: Original speech signal d(u) and IDRT reconstructed speech signal d'$_1$(u)

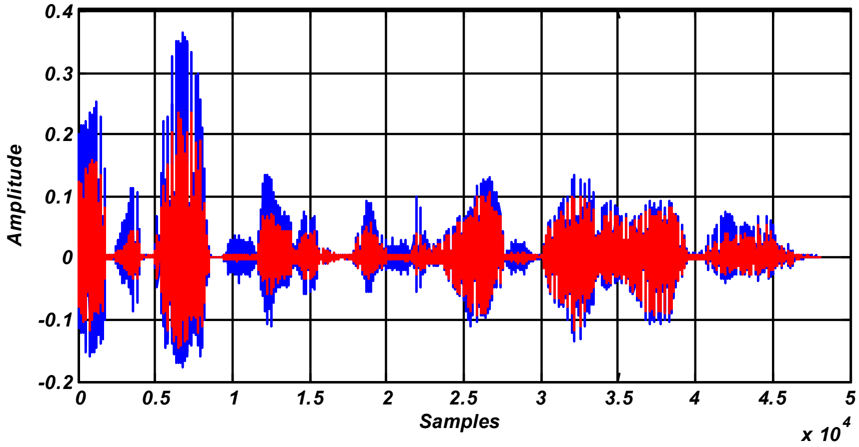

Fig. 11: Original speech signal d(u) and IDRT reconstructed speech signal d'$_2$(u)

Fig. 11. Indicates original speech signal $d(u)$ and IDRT reconstructed speech signal $d'_2(u)$ the discourse signal reproduced from 25% of speech data. Reconstructed speech signal $d'_2(u)$ of a 48128 samples size of the speech signal of a speaker from the scarified information keeping the CPI and the mid frequency from each of block of length 8 and constraining different components to 0, that is D'_{s2} (r), is particularly closer to the original speech signal $d(u)$.

Error Dynamic Range (EDR) because of remaking from 25% of voice information is less when contrasted with the EDR because of reproduction from 12.5% of voice information and henceforth the previous is superior to the last in speaker recognition application.

Demonstration of Speaker Specific Information and Acoustic Hole in IDRT Reconstructed Speech Signal: Sparse and compressed speech signal was reconstructed using IDRT algorithm from 12.5% and 25% compression of a speech waveform from TIMIT database, to demonstrate the appearance of acoustic hole and similarity of other speaker specific information. The transitions of formants are much clearer in both reconstructed speech signal. The appearance of phonemes is compared with original speech signal and reconstructed speech signal. In Fig. 12. the top one is original spectrogram and bottom is CPI alone reconstructed speech signal spectrogram. It has been observed that the phonemes in reconstructed speech signal have been burst and a lot of acoustic holes are generated. It indicates that the up to 12.5% of sparsed and compressed speech signal may not be suitable for ASR application. The part wise comparisons of spectrogram have been shown in different colors of windows. In extreme right side white window there is no similarity of spectrogram with the original and reconstructed signal.

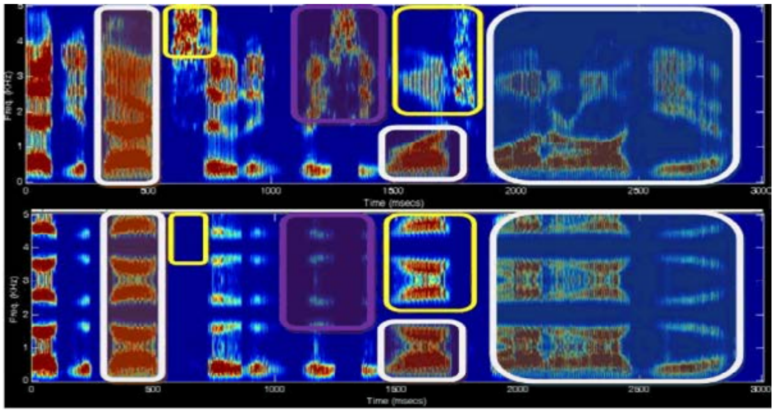

Fig. 12: Spectrogram of original and CPI alone reconstructed speech signal

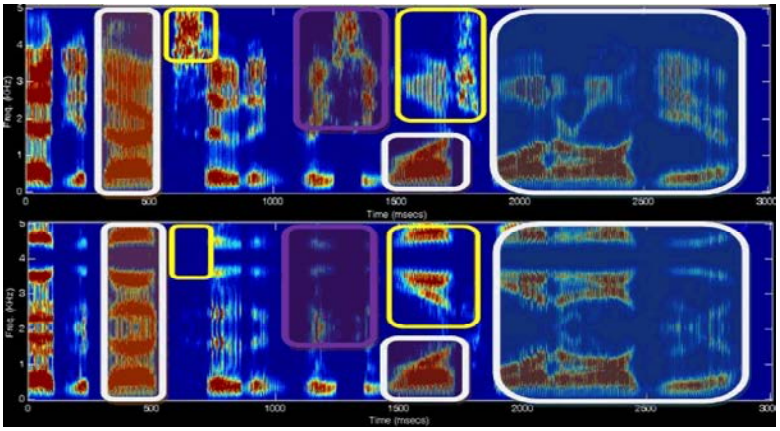

Fig. 13: Spectrogram of original and CPI+ mid frequency reconstructed speech signal

In Fig. 13. the top one is original spectrogram and bottom is CPI+mid frequency alone reconstructed speech signal spectrogram. It has been observed that the phonemes in original and reconstructed speech signal are almost similar. Comparatively, the acoustic hole is less with CPI alone reconstructed signal. It indicates that the up to 25% sparsed and compressed speech signal will be suitable for ASR application. The part wise comparisons of spectrogram have been shown in different colors of windows. In the extreme right side white window there is almost similar spectrogram with the original and reconstructed signal.

CPI along with mid frequency component is balancing the acoustic hole in reconstructed signal. In Fig. 12 it has been demonstrated that if mid frequency component will be forced to zero will generate lots of acoustic hole.

Speech Quality Performance Measurement Matrix: Following three different performance parameters are used to measure the quality of reconstructed speech. Here we have measured the performance of $d'_1(u)$ and $d'_2(u)$ with reference to original speech signal $d(u)$.

The Mean Squared Error (MSE) is apparently the essential paradigm used to assess the quality of the reconstructed signal. The 48128 samples of original of speech signal and IDRT synthesized speech signal with "u" range time index covering the measurement intervals, then the MSE is defined as:

$$MSE = \sum_u \frac{[d(u) - d_1^{'}(u)]^2}{u} \tag{10}$$

In digital speech processing MSE represents the quantity by which IDRT reconstructed speech signal fluctuates from the original speech.

SNR is defined as the ratio of the power of an original speech signal and the power of the error signal and mathematically defined as:

$$SNR = 10\log_{10}\left(\frac{\sum_u (d(u))^2}{\sum_u [d(u) - d_1(u)]^2} \right) \tag{11}$$

Table 9: MSE of speech signal after applying different transform

Sparsification by retaining	MSE DRT	MSE DFT	MSE DCT	MSE DWT
Only CPI	0.001900	0.036900	0.014300	0.015400
CPI +Mid frequency	0.000647	0.034700	0.012200	0.011500

Table 10: SNR of speech signal after applying different transform

Sparsification by retaining	SNR DRT (dB)	SNR DFT (dB)	SNRDCT (dB)	SNR DWT (dB)
Only CPI	25.47	10.9125	15.03	14.7028
CPI +Mid frequency	27.39	11.7032	17.97	15.9656

Table 9 and Table 10 represents MSE and SNR of DRT, DFT, DCT, DWT for 100 different speeches randomly selected from TIMIT database.

MSE of DRT is least and SNR is more, therefore, DRT is suitable for sparsification of the speech signal in ASR application.

PESQ is a generally utilized, upgraded perceptual estimation for voice quality in information transfers. By and large, speech quality appraisal can be categorized as one of two classifications; subjective and objective quality measures. Subjective quality measures depend on the examination of original and reconstructed speech signal by an audience or a board of audience members. The scope of PESQ varies from 0.5 to 4.5, with the lower values interpreted as poor speech quality.

It is observed that for the case of up to 25% data compression the PESQ of the reconstructed speech data does not deviate so much from the standard value that is 3.2331. Indeed, for the case of up to 12.5% data compression, the PESQ of the reconstructed speech data deviates considerably from the standard value that is 2.1543.

Experimental Results: Experimental assessment of the DRT sparsification algorithms with CPI+mid frequency components continued with the 100 speakers from TIMIT database and NIST database. The sparsification with retaining CPI alone and reconstructed speech signal generate acoustic hole, therefore, we are not evaluating ASR with this speech signal. In this paper, we are comparing sparsified reconstructed speech (CPI + Mid frequency component) and original speech signal based ASR.

Speaker Specific Feature Extraction of Original and Sparsified Reconstructed Speech Signal: The speaker specific information of original and reconstructed speech signal is captured in this research by the most prominent features Mel-frequency Cepstral Coefficients (MFCC) [11]. The steps involved in capturing acoustic features are as follows:

- Framing and windowing the speech signal of duration 25-30 milliseconds.
- For each frame compute the periodogram and estimate the power spectrum.
- Apply Mel-filter bank (A typical 29 filter bank is taken into consideration) to the power spectrum [12].
- Sum the energy in each filter.
- Take the logarithm of all filter bank energies.
- Take the DCT of the log filter bank energies.
- Keep 19 DCT coefficients and discard the rest.
- RASTRA Filtering [13].
- Delta and Delta Delta filtering to achieve 60-dimension feature vectors
- Frame dropping
- Feature Warping [14].

Baseline Speaker Modeling Techniques in ASR System: In this paper, two modelling configurations for ASR have been taken into consideration.

- Gaussian mixture model (GMM) [15].
- State-of-the-art i-vector framework [16].

With the GMM model, we have tested the ASR system with original and sparsified reconstructed speech signal for short utterances. The i-vector-based ASR framework was assessed by having right around 10 seconds of original and sparsified reconstructed speech for training and testing. i-vectors are one case of subspace ASR modelling approaches that can be utilized to minimize the dimensionality of the training speech data before applying classifiers to perceive the dialect utilized as a part of utterances.The dimensionality decrease should make training of classifiers less computational costly, which could empower us to train the framework with more data. We meant to exhibit the outcomes for the GMM-based framework as a proof of idea and considered the i-vector-based ASR system for analyzing the comparative efficiency of the system. For the i-vector based ASR system, as it happens in genuine legal applications, we took a cutting edge ASR system [17],

Table 11: ASR efficiency of TIMIT and NIST Corpora

		Train/ Test with Original Speech Signal			Train/ Test with Reconstructed Speech Signal		
Modeling	Corpora	Male	Female	Together	Male	Female	Together
GMM	TIMIT	98.9%	98.4%	98.7%	94.4%	93.7%	94.2%
	NIST	98.6%	98.2%	98.4%	94.2%	93.3%	94.1%
i-vector	TIMIT	99.3%	98.8%	99.1%	94.9%	93.9%	94.8%
	NIST	99.1%	98.4%	98.9%	94.6%	93.3%	94.5%

[18] off-the-rack where ASR parameters can't be adjusted to the test condition as a result of insufficient data. The i-vector-based ASR system requires a modest bunch of highlight vectors for solid extraction of adequate statistics which is uncommon in short utterances and we utilized an expression determination convention not the same as the one in the GMM case.

The GMMs with a corner to corner covariance structure were trained with 32 Gaussians. The i-vector-based ASR system was created in Radboud University Nijmegen as a part of the college submission to NIST ASR assessment in 2012 [17, 18]. A male/female based universal background model (UBM) [19] with 2048 speaker specific information was trained utilizing a subset of NIST SRE 2004–2006, Switchboard cell stage 1 and 2 and the Fisher English corpora. To factorize the GMM mean supervectors, the aggregate variability space [20] was trained with the same information concerning UBM with 400 data. During pre/post-processing of speech signal level i-vectors, we utilized linear discriminate analysis (LDA) projection to upgrade the detachability of classes and diminish the i-vectors' measurement to 200. Prior to probabilistic linear discriminant analysis (PLDA) [21] modeling, we eliminated the mean, performed whitening utilizing within-class covariance normalization (WCCN) [22] and standardized the length of i-vectors [23].

Experimental Results: Experimental assessment of the original and sparsified acoustical balanced reconstructed speech signal continued with the 100 speakers from TIMIT and NIST speaker recognition evaluation corpora. Table 11. presents the over all ASR efficiency for TIMIT/NIST corpora.

It is observed from Table 11. in the TIMIT/NIST corpora the performance of ASR little inferior for the female speaker but both corpora demonstrate the same general trend with GMM and i-vector modeling techniques. The i-vector based ASR gives the highest efficiency up to 99.1%, 94.80 % and 98.90%, 94. 5% respectively for original and Scarified reconstructed speech signal for TIMIT/NIST corpora.

Performance Evaluation of Speaker Recognition Systems: With a specific end goal to check the performance of sparsification and acoustic hole balancing algorithm based ASR, we processed the real match scores with original speech signal and Scarified reconstructed speech signal with the impostor match scores. The Detection Error Trade-off (DET) of every experiment has appeared in the accompanying figures. Fig. 14. compares GMM based ASR system performance with TIMIT corpora in two different modalities: that is original speech and sparsified reconstructed speech signal. GMM based ASR performance with TIMIT corpora an Equal Error rate (EER) value of about 1.6% and with sparsified reconstructed speech about 4.3%.

GMM based ASR performance with NIST corpora an EER value of about 1.9% and with sparsified reconstructed speech the EER of about 4.8%. Fig. 15. compares GMM based ASR system performance with NIST corpora in two different modalities: that is original speech and sparsified reconstructed speech signal. The Same trend can be observed that TIMIT corpora performance is superior that NIST.

Based on the analysis of the DET curves in Fig. 14. and Fig. 15. it is clear that by employing the sparsification based algorithms which are perfectly balancing the acoustic hole, spectrum can be compressed up to 25% without degrading much more ASR system performance.

Fig. 16 compares i-vector based ASR system performance with TIMIT corpora in two different modalities: that is original speech and sparsified reconstructed speech signal. I-vector based ASR performance with TIMIT corpora an EER value of about 1.4% and with sparsified reconstructed speech the EER of about 4.1%.

Fig. 17 compares i-vector based ASR system performance with NIST corpora in two different modalities: that is original speech and sparsified reconstructed speech signal. The same trend can be observed that TIMIT corpora performance is superior that NIST. i-vector based ASR performance with NIST corpora an EER value of about 1.5% and with sparsified reconstructed speech the EER of about 4.2%.

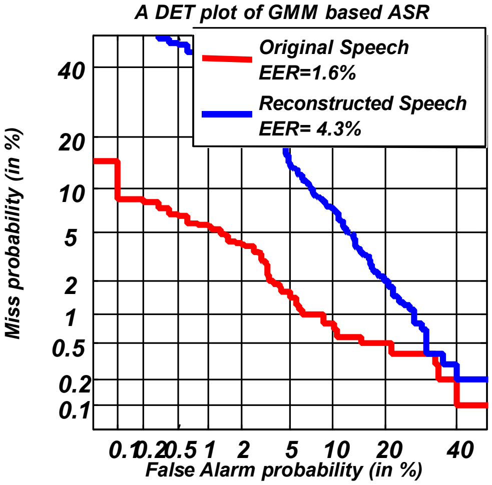

Fig. 14: DET curve of GMM based ASR for TIMIT Corpora

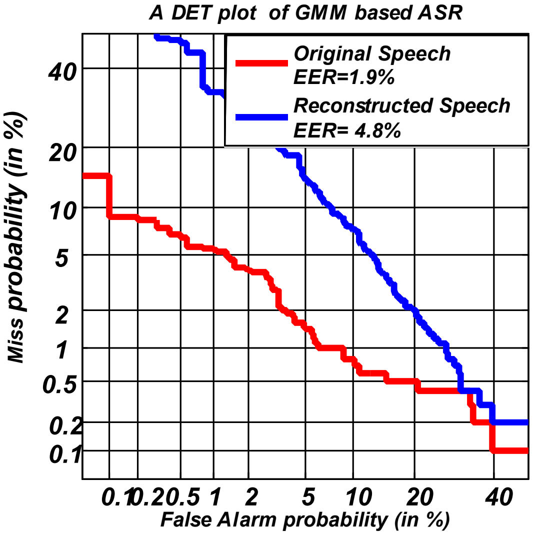

Fig. 15: DET curve of GMM based ASR for NIST Corpora

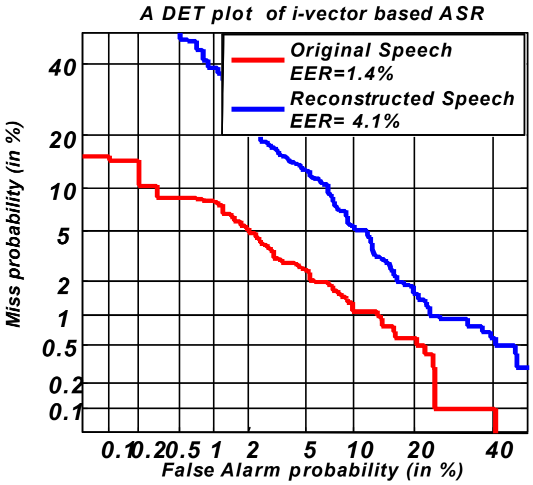

Fig. 16: DET curve of i-vector based ASR for TIMIT Corpora

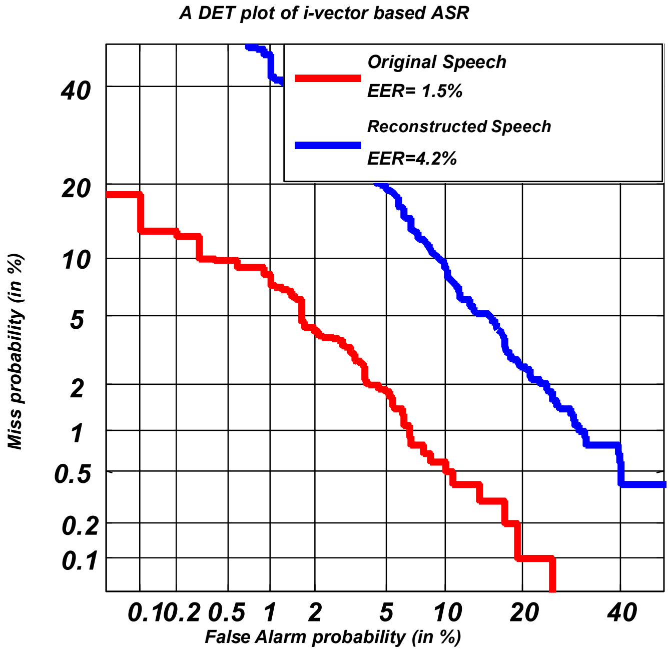

Fig. 17: DET curve of i-vector based ASR for NIST Corpora

CONCLUSION

In this research, we have presented an efficient ASR system based on original speech and acoustically hole balanced sparsified reconstructed speech of TIMIT and NIST corpora. The study capered with GMM and i-vector modeling method. The I-vector modeling technique was adopted in this work, due to its high accuracy. Achievable spectral compression of voice samples was found to be about 75% with ASR efficiency of 94.8% with baseline efficiency of 99.1% in case of i-vector modeling of TIMIT corpora. In the case of NIST corpora, the ASR efficiency is 94.5% with sparsified reconstructed speech and 98.9% the original speech signal respectively.

The EER of TIMIT corpora and i-vector-based modeling is 1.4% for original speech signal and 4.1% for sparsified reconstructed speech signal respectively. We have achieved a reduction of 75% of the speech signal with scarification of EER 2.7% only. Future work would be to consider improving the sparsified reconstructed speech signal efficiency to approach to the baseline efficiency.

REFERENCES

1. Hansen, John H.L. and Taufiq Hasan, 2015. Speaker Recognition by Machines and Humans: A tutorial review. IEEE Signal Processing Magazine, 32(6): 74 -99.
2. Dehak, N., P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech and Language Processing, 19(4): 788-798.
3. Mandasari, M.I., M. McLaren and D.A. VAN Leeuwen, 2012. The effect of noise on modern automatic speaker recognition systems. In Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2012).
4. Mandasari, M.I., M. McLaren and D.A. Van Leeuwen, 2011. Evaluation of i-vector speaker recognition systems for forensic application, In Proc. Interspeech, pp: 21-24.
5. Furui Sadaoki, 2005. 50 Years of Progress in Speech and Speaker Recognition Research, ECTI Transactions on Computer and Information Technology, 1(2): 64-74.
6. Plumbley, M.D., T. Blumensath, L. Daudet, R. Gribonval and M.E. Davies, 2010. Sparse Representations in Audio and Music: from Coding to Source Separation, In Proceedings of IEEE, 98(6): 995-1005.
7. Wang, Y., Z. Xu, G. Li, L. Chang and C. Hong, 2011. Compressive Sensing Framework for Speech Signal Synthesis Using a Hybrid Dictionary, 4th International Congress on Image and Signal Processing (CISP), Shanghai, 5: 2400-2403.

8. Desai Siddhi and Nakrani Naitik, 2014. Improved Performance of Compressive Sensing for Speech Signal with Orthogonal Symmetric Toeplitz Matrix, International Journal of Signal Processing, Image Processing and Pattern Recognition, 7(4): 371-380.
9. Prashanthi Satyanand, Dr. E.G. Rajan and Pat Krishanan, 2014. Sparsification of Voice Data Using Discrete Rajan Transform and it's Applications in Speaker Recognition, IEEE International Conference on Systems, Man and Cybernetics, San Diego, CA, USA, pp: 429-434.
10. Rabiner, L. and B.H. Jung, 1993. Fundamental of Speech Recognition, Prentice- Hall, New Jersey.
11. Davis, S. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Processing, 28(4): 357-366.
12. Yuan Lan, Zongjiang Hu, Yeng Chai Soh and Guang-Bin Huang, 2013. An extreme learning machine approach for speaker recognition, Neural Comput & Applic, 22: 417-425.
13. Hardt, D. and K. Fellbaum, 1997. Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification, in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'97), Munich, Germany, pp: 867-870.
14. Pelecanos, J. and S. Sridharan, 2001. Feature warping for robust speaker verification, in Proc. Speaker Odyssey: Speaker Recogn, Workshop (Odyssey'01), Crete, Greece, pp: 213-218.
15. Saeidi Rahim, Paavo Alku and Tom, 2016. Feature Extraction Using Power-Law Adjusted Linear Prediction with Application to Speaker Recognition under Severe Vocal Effort Mismatch, in IEEE/ACM Transaction on Audio, Speech and Language Processing, 24(1): 42-53.
16. Dehak, N., P.J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, 2011. Frontend factor analysis for speaker verification, IEEE Trans. Audio, Speech, Lang. Process., 19(4): 788-798.
17. Garcia-Romero, D. and C.Y. Espy-Wilson, 2011. Analysis of I-vector length normalization in speaker recognition systems, in Proc. Interspeech, pp: 249-252.
18. Dehak, N., P.J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, 2011. Front-end factor analysis for speaker verification, IEEE Trans. Audio, Speech, Lang. Process., 19(4): 788-798.
19. Reynolds, D., T. Quatieri and R. Dunn, 2000. Speaker verification using adapted Gaussian mixture models, Digital Signal Process., 10(1): 19-41.
20. Dehak, N., P.J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, 2011. Frontend factor analysis for speaker verification, IEEE Trans. Audio, Speech, Lang. Process., 19(4): 788-798.
21. Prince, S.J.D. and J.H. Elder, 2007. Probabilistic linear discriminant analysis for inferences about identity, in Proc. 11th Int. Conf. Comput. Vis, pp: 1-8.
22. Hatch, A.O., S. Kajarekar and A. Stolcke, 2006. Within-class covariance normalization for SVM-based speaker recognition, in Proc, Interspeech (ICSLP'06), Pittsburgh, PA, USA, pp: 1471-1474.
23. Garcia-Romero, D. and C.Y. Espy-Wilson, 2011.Analysis of i-vector length normalization in speaker recognition systems, in Proc. Interspeech, 11: 249-252.