International Journal of Computational and Applied Mathematics.
ISSN 1819-4966 Volume 11, Number 1 (2016), pp. 89-104
© Research India Publications
http://www.ripublication.com

A Novel Algorithm of Sparse Representations for Speech Compression/Enhancement and Its Application in Speaker Recognition System

Dr. Satyanand Singh
Postdoctoral Research Associate, School of Engineering and Physics
Laucala Campus, Suva, Fiji.
E-mail: yogitechno@gmail.com

Mansour H. Assaf
Associate Professor, School of Engineering and Physics
Laucala Campus, Suva, Fiji.
E-mail: mansour.assaf@usp.ac.fj

Abhay Kumar
Ph.D Research Scholar, Sri Satya Sai University of Technology & Medical Sciences (SSSUTMS), Sehore, Madhya Pradesh, India.
E-mail: abhay1880@gmail.com

Abstract

This paper proposes sparse and redundancy representation spectral domain compression of the speech signal using novel sparsing algorithms to the problem of speech compression (SC)/enhancement (SE). In Automatic Speaker Recognition (ASR) sparsification can play a major role to resolve big data issues in speech compression and its storage in the database, where the speech signal can be uncompressed before applying to ASR system. The speech signal is converted to a spectral domain using Discrete Rajan Transform (DRT) and only first and mid spectrum component is retained forcing the remaining component to zero. The speech signal spectrum can be maximally compressed 8:1 ratio to the unique one. Spectrally compressed speech signal can be stored in the database and during training and testing time it can be synthesized using Inverse Discrete Rajan Transform (IDRT) in ASR. Sparsification and spectral compression up to 75% with Equal Error Rate (EER) of ASR is 3%. Percentage of Identification Accuracy (PIA) of ASR with sparsification and speech enhancement is 99.1% and without sparsification 98.8% for TIMIT database respectively.

Keywords: Speech Enhancement, Sparse Representation, Automatic Speaker Recognition, Cumulative Point Index, Support Vector Machine, Gaussian Mixture Model, Universal Background Model, Equal Error Rate.

INTRODUCTION

With advancements in mathematics, linear representation methods (LRBM) have been well studied and have recently received considerable attention [1, 2]. The sparse representation method is the most representative methodology of the LRBM and has also been proven to be an extraordinary powerful solution to a wide range of application fields, especially in signal processing, speech processing, image processing, machine learning, and computer vision, such as image denoising, debluring, in painting, image restoration, super-resolution, visual tracking, image classification and image segmentation [3,4]. Sparse representation has shown huge potential capabilities in handling these problems.

Sparse representation, from the viewpoint of its origin, is directly related to compressed sensing (CS) [5,6], which is one of the most popular topics in recent years. Donoho [4] first proposed the original concept of compressed sensing. CS theory suggests that if a signal is sparse or compressive, the original signal can be reconstructed by exploiting a few measured values, which are much less than the ones suggested by previously used theories such as Shannon's sampling theorem (SST). Candès et al. [7], from the mathematical perspective, demonstrated the rationale of CS theory, i.e. the original signal could be precisely reconstructed by utilizing a small portion of Fourier transformation coefficients. Baraniuk [8] provided a concrete analysis of compressed sensing and presented a specific interpretation on some solutions of different signal reconstruction algorithms. All these literature [9,10] laid the foundation of CS theory and provided the theoretical basis for future research. Thus, a large number of algorithms based on CS theory have been proposed to address different problems in various fields. Moreover, CS theory always includes the three basic components: sparse representation, encoding measuring, and reconstructing algorithm. As an indispensable prerequisite of CS theory, the sparse representation theory [10, 11, 12, 13] is the most outstanding technique used to conquer difficulties that appear in many fields. With regards to CS, it is first expected that the speech signal are approximately sparsed [11, 12, 14]. Contrasted with the original speech signal space occupied, the reconstructed speech signal space decreased considerably with sparsing algorithm. The sparsing algorithm is very useful to tackle the computational complexity in ASR. The sparsification of speech signal representation system can spare a significant improve the computation time and test/train storage in ASR application.

Speaker recognition robustness in adverse condition has been investigated widely in recent years [15]. There are quite a number of factors affecting the ASR 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 [16]. In this research our main focus in on speech enhancement and spectral domain speech compression and its storage in database.

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 [17]. 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 [18]. 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.

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 an 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 [19].

PROBLEM DESCRIPTION OF SPEECH SIGNAL IN ASR

Let us consider a noisy speech signal model applied to ASR in the form of $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{n}$, where x represent noisy speech, y represents clean speech and n represents the noise. Our research goal is to estimate the speech component from available speech signal sparse it, compress it and reconstruct the clean speech to improve the ASR efficiency. A speech signal can be represented as $\mathbf{y} \in \mathbb{R}^{\mathbf{N}}$ which can be sparsed over a fixed length in a block transformed $\mathfrak{J} \in \mathbb{R}^{\mathbf{NXN}}$, then the given speech signal can be expressed as, $\mathbf{y} = \mathfrak{J}\mathbf{L}$, $\|\boldsymbol{\theta}\|\ \mathbf{0} \leq \boldsymbol{S} \ll \mathbf{N}$. L is the length of N coefficient vectors with S none zero elements. Noisy speech signal "x" can be transformed into spectral domain with the forward DRT, sparsed, compressed and clean speech signal "y" can be reconstructed.

SPARSE REPRESENTATION WITH DISCRETE RAJAN TRANSFORM (DRT) AND INVERSE DISCRETE RAJAN TRANSFORM (IDRT).

Due to homomorphism nature, it has many applications in image processing like detection of the curve, detection of lines, detection of contour, 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 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 spectral domain with a unique operating matrix of dimension $(U/2^{r-1} X\ U/2^{r-1})$denoted as Y_r. This unique operation matrix construction is defined as;

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

$\boldsymbol{I_w}$ indicate the $\boldsymbol{w^{th}}$ order identity matrix. For example at ***r*** steps the order of identity matrix is Wr=U/2r: r∈{1, 2, .., n} and e1ϒ 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’ eϒ 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 steps ***r***, let $\boldsymbol{F_r}$ denoted as output sequence and it is obtained as:

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

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

$\boldsymbol{Y_r}$ is the operator matrix can be constructed at a ***r*** stage using supplementary information $\boldsymbol{e_r}$. Additionally if ***r* > *1*** then the output can be restructured in equilibrium segments and it can be defined as:

$$D_{r+1} = [\bar{d}_{r+1}^1 \quad \Upsilon_r^1 \cdot \bar{d}_{r+1}^2 \cdots \quad d_r^{i-1} \cdot \bar{d}_{r+1}^i] \tag{4}$$

where $\Upsilon_r = [\Upsilon_r^1 \quad \Upsilon_r^2 \cdots \quad \Upsilon_r^{i-1}]$ and also,

$$\Upsilon_k^{i-1} = e_r^1 \times e_r^i for r > 1 \tag{5a}$$

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

$$= [\bar{d}_{r+1}^1 \quad \bar{d}_{r+1}^2 \cdots \quad \bar{d}_{r+1}^i] \tag{5b}$$

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 retry 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^1_k*** and ***μ_k***. Now the DRT operator ***R_k*** is obtained using the values of ***e^1_k*** and ***μ_k***. The general expression used to retrieve intermediate signal data at every stage is [20]:

$$\widetilde{D}_t = \frac{1}{2}[Y_t F_t] = [d_t^1 \quad d_t^2 \ldots \quad d_t^i]^T \qquad (6)$$

Where **t = {r, r-1, .., 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***,

$$F_{t-1} = \left[\bar{F}_t(1) \quad \Upsilon_r^1 \cdot \bar{F}_t(2) \cdots \quad \Upsilon_r^{i-1} \cdot \bar{F}_t(i)\right] \qquad (7)$$

$$\bar{F}_{a-1} = \begin{bmatrix} d_a^i(1) & d_a^i(2w_r+1) \cdots & d_a^i(2^{r-1}w_r+1) \\ d_a^i(2) & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ d_a^i(2w_r) & d_a^i(2^2 w_r) & d_a^i(2^r w_r) \end{bmatrix} \qquad (8)$$

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

DRT SPARSIFICATION, COMPRESSION AND DECOMPRESSION APPLICATION ON SPEECH SIGNAL

The experiments are performed on a speech signal taken from TIMIT 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 enhanced speech signal.

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

- Read wave files.
- Select beginning 1X48128 size 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.
- 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.

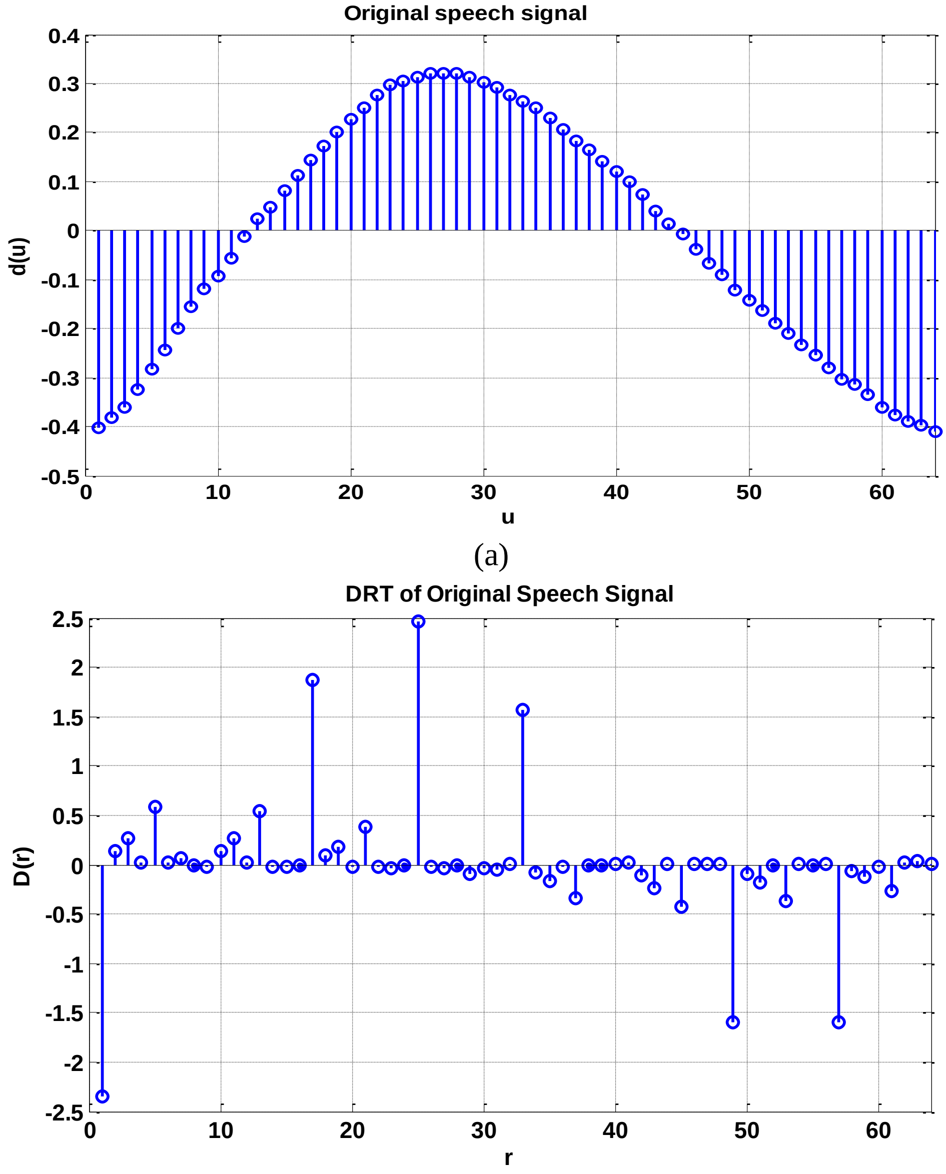
The DRT algorithms have been compared with similar algorithms like DCT, DFT, and DWT in the following section.

Application of DRT on Speech Signal of 64 Sample Size

A 3 sec speech signal of female 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 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 discrete sequenced ***d(u)*** of length 64. Let each block represented by ***B*** and Sample ***S*** respectively.

DRT is applied to ***d(u)*** in block-wise fashion and equivalent spectral blocks obtained as ***D(u).***$\boldsymbol{D}(\boldsymbol{r})$ is the spectral component of the original speech signal $\boldsymbol{d}(\boldsymbol{u})$ and the 1st component of each block have 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. 1 (a). shows the plot 64 points of the original speech signal $\boldsymbol{d}(\boldsymbol{u})$and Fig. 1. (b) shows the plot of $\boldsymbol{D}(\boldsymbol{r})$**.**



(a)

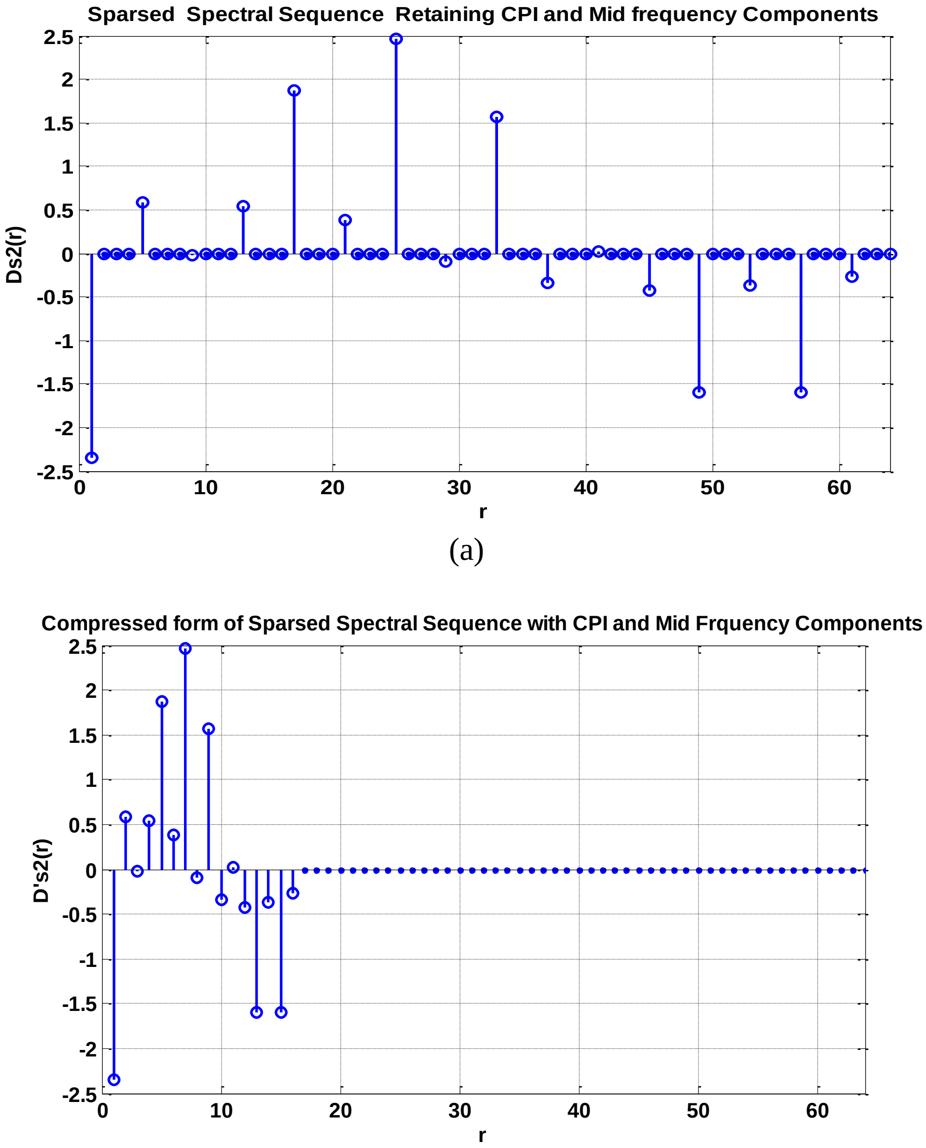
(b)

Figure 1. (a) Plot of 64 points original speech signal $d(u)$.
(b) Plot of $D(r)$ the spectrum of (u) .

Sparsing of Speech Signal Data by Retaining CPI and mid Frequencies alone

In this case $\boldsymbol{D}(\boldsymbol{r})$ 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. Compressing Speech Signal Data with CPI and mid Frequency alone. $\boldsymbol{D}_{s2}(\boldsymbol{r})$is the spectral domain sparsed speech data having only 16 non-zero elements. $\boldsymbol{D'_{s2}(r)}$is the compressed version of $\boldsymbol{D_{s2}(r)}$ after ignoring all 48 samples of spectral components of zero values. $\boldsymbol{D'_{s2}(r)}$ 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.2. (a) shows the plot of $\boldsymbol{D_{s2}(r)}$ the sparsed spectral sequence with CPI and mid frequency component and Fig.2. (b) compressed form of sparsed spectral sequence $\boldsymbol{D'_{s2}(r)}$.



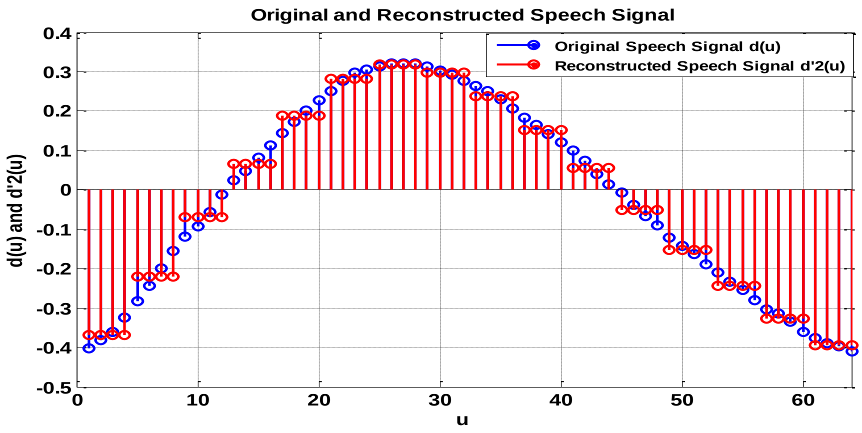
(a)

(b)

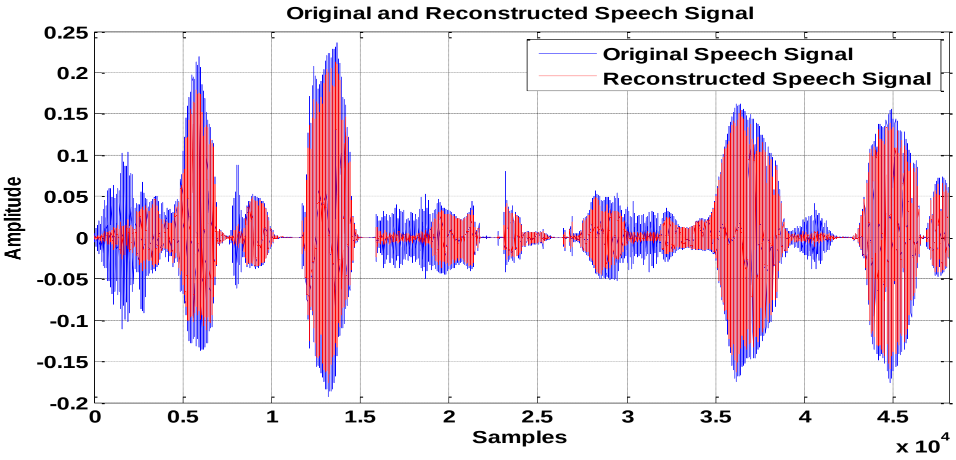
Figure 2.(a) Plot CPI and mid frequency components $\boldsymbol{D_{s2}(r)}$.
(b) Plot compressed CPI and mid frequency components $\boldsymbol{D'_{s2}(r)}$.

Decompression and speech enhancement from D’s1(r) and D’s2(r)

Amid the speaker recognition testing stage $D'_{s2}(r)$ is uncompressed to acquire $D_{s2}(r)$. In the of case 25% of sparsity the algorithm will inject three zeros after the CPI and mid frequency to acquire$D_{s2}(r)$. Presently, IDRT algorithm applied to $D_{s2}(r)$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 $d'_2(u)$. Fig.3.(a) shows original 64 samples speech signal $\boldsymbol{d(u)}$ and IDRT reconstructed enhanced speech signal $\boldsymbol{d'_2(u)}$represented in the same plot. Fig. 3. (b). shows original speech signal $\boldsymbol{d(u)}$and IDRT reconstructed speech signal $\boldsymbol{d'_2(u)}$exhibited in the same plot of 48128 sample size of speech data. The enhanced speech signal $\boldsymbol{d'_2(u)}$is the discourse signal reproduced from 25% of spectral data. Reconstructed speech signal $d'_2(u)$ of a 48128 samples size of the speech signal of a speaker from the sparsified information keeping the CPI and the mid frequency from each of block of length 8 and constraining different components to 0, that is $\boldsymbol{D'_{s2}(r)}$**,** is particularly closer to the original speech signal $\boldsymbol{d(u)}$**.**



(a)



(b)

Figure 3.(a) Plot of $\boldsymbol{d(u)}$and $\boldsymbol{d'_2(u)}$.(b) Original speech signal $\boldsymbol{d(u)}$and IDRT reconstructed speech signal $\boldsymbol{d'_2(u)}$**.**

PERFORMANCE OF SPEECH QUALITY MEASUREMENT MATRIX

Following three different performance parameters are used to measure the quality of reconstructed enhanced speech signal. Here we have measured the performance of ***d'₂(u)*** with reference signal ***d(u)***.

The Mean Squared Error (MSE) is apparently the essential paradigm used to assess the quality of the reconstructed signal. The 3 sec original of speech signal and IDRT synthesized speech signal with "u" range time index covering the measurement intervals, then the MSE is defined as:

$$MSE = \sum_u \frac{\left(d(u) - \tilde{d}_1(u)\right)^2}{u} \tag{9}$$

In digital speech processing MSE represents the quantity by which IDRT reconstructed speech signal fluctuates from the original speech. Signal-to-noise ratio is defined as the ratio of the power of an original speech signal and the power of the error signal and mathematically defined as:

$$SNR = 10log_{10}\left(\frac{\sum_u (d(u))^2}{\sum_u \left(d(u) - \tilde{d}_1(u)\right)^2}\right) \tag{10}$$

Table 1. and Table 2. show comparison of mean square Error (MSE) and signal to noise Ratio (SNR) of DRT, DFT, DCT, DWT for 100 different speeches randomly selected from TIMIT database.

Table 1. 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 2. 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

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 75% of sparsification the PESQ of the reconstructed speech signal does not deviate so much from the standard value that is 3.2331.

FEATURE EXTRACTION BY MEL-FREQUENCY CEPSTRAL COEFFICIENTS

The speaker specific feature extracted from the short duration speech segment/frame of 25-30 milliseconds. Mel-Frequency Cepstral Coefficients (MFCC) is the most popular acoustic features used in ASR. Following steps, we adapt to extract MFCC.

i. Divide test/train speech signal into short overlapping segments (25 milliseconds)
ii. Windowing (Hamming and Hanning)
iii. Take logarithm
iv. Nonlinear scale Mel scale filtering analysis (24 channel filter bank energy coefficients) [21].
v. Apply DCT on the filter bank energy parameters (Retaining the12 coefficients after DCT and discarding the remaining)

SPEAKER MODELING

Once the speaker specific feature has been extracted and converted to feature vectors then next step in ASR is modeling. In general, how we are going to describe the feature vector of a speaker is called modeling. The model must be capable of comparing the unknown utterance of the speaker and identifying the exact one. If characterization process of speaker-specific features is not much more affected by unwanted distortions then the modeling is called as robust. State-of-the-art ASR modeling techniques make a lot of mathematical assumption on speaker-specific feature vectors like Gaussian distribution. If the assumed mathematical properties are not met by the data then ASR modeling itself is introducing certain error.

Support Vector Machine (SVM)

The objective of an SVM in speaker recognition is to locate the ideal isolating hyperplane which expands the edge of the training speech data and constructed from the sum of its kernel function K(.) [22]

$$f(x) = \sum_{i=1}^{N} a_i t_i \, K(x.x_i) + d \tag{11}$$

t_i= output, $\sum_{i=1}^{N} a_i t_i = 0$, $a_i > 0$, The vectors x_i are support vector and acquired from the training speech data. The perfect yields are either 1 or - 1, subject on whether the comparing support vector is in class 0 or class 1. For collection, a class choice is based upon whether the value of classification falls which side of threshold. Based on the Mercer condition the kernel properties is defined as follows,

$$K(p, q) = b(p)^t b(q) \qquad (12)$$

Input space mapping is represented by $b(p)$. The Mercer condition guarantees that the edge idea is proper, and the streamlining of the SVM is all well defined.

The Gaussian Mixture Model (GMM) Supervector

In speaker recognition system the training and testing speech data are always a different durations. ASR system needs to compare the training and testing utterances of different time duration. The concept single utterance fixed dimensional representation has played a major role in effective speaker recognition. The mathematical representation of GMM and Universal Background Model (UBM) is represented by the following equation:

$$g(x) = \sum_{i=1}^{N} \lambda_i \, \mathrm{N}(x; m_i; \textstyle\sum_i) \qquad (13)$$

Where λ_i is the mixture weight,$\mathrm{m_i}$ is the mean and $\sum_\mathrm{i}$is the covariance of Gaussian N(). During the training GMM UBM is performed of the given utterance by MAP adaptation of the mean. The GMM supervector has been formed by this adapted model. The formation of adaptation based GMM supervector can be considered as a mapping between the testing utterances and a high-dimensional vector. This design fits well with the possibility of a SVM sequence kernel [23]. Basically SVM sequence kernel compares two speech utterances $\mathrm{speech_a}$ and $\mathrm{speech_b}$with $\mathrm{K(speech_a, speech_b)}$. As per the Mercer condition the Kernel can be written as $\mathrm{K(speech_a, speech_b) = b(speech_a)^t b(speech_b)}$. It will be very easy to map $\mathrm{speech_a}$ to $\mathrm{b(speech_a)}$ with the help of GMM supervector.

Linear Kernel of GMM Supervectors

Let us consider two speech utterances $\boldsymbol{speech_a}$ and $\boldsymbol{speech_b}$ and train the ASR model with the help of GMMs, $\mathrm{g_a}$, $\mathrm{g_b}$ by MAP adaptation technique. The Kullback-Leibler (KL) divergence is the method to find out the natural distance between two speech utterances,

$$D(g_a \parallel g_b) = \int_{R^n}^{n} g_a(x) \, log\left(\frac{g_a(x)}{g_b(x)}\right) dx \qquad (14)$$

Since kernel matrix separations specifically in view of symmetric KL divergence does not fulfill the Mercer conditions, i.e., it is not a positive definite matrix we require a further stride to produce a legitimate kernel. Therefore, we can’t use SVM directly. With the help of log-sum inequality we can bound the divergence as follows [24],

$$D(g_a \parallel g_b) \leq \sum_{i=1}^{N} \lambda_i D\,(\mathrm{N}(.\,;\mathrm{m}_\mathrm{i}^\mathrm{a}, \textstyle\sum_i) \parallel \mathrm{N}(.\,;\mathrm{m}_\mathrm{i}^\mathrm{b}, \textstyle\sum_i)) \tag{15}$$

Where m^a and m^b is represents adapted supervector. Considering diagonal covariance, the closed form approximation can be computed from (15) and written as,

$$d(m^a, m^b) = \frac{1}{2}\sum_{i=1}^{N} \lambda_i \left(m_i^a - m_i^b\right) \sum_{i}^{-1} \left(m_i^a - m_i^b\right) \tag{16}$$

The final step inequality can be written as,

$$0 \leq \mathrm{D}(\mathrm{g_a} \parallel \mathrm{g_b}) \leq \mathrm{d}\left(\mathrm{m^a}, \mathrm{m^b}\right) \tag{17}$$

From (17) the divergence is small hence the distance between m^a and m^b is small. The kernel function can be computed from (16) as follows,

$$\begin{aligned} K(speech_a, speech_b) &= \sum_{i=1}^{N} \lambda_i m_a^i \sum_{i=1}^{-1} m_i^b \\ &= \sum_{i=1}^{N} \left(\sqrt{\lambda_i} \sum_{i=1}^{-\frac{1}{2}} m_i^b\right)^t \left(\sqrt{\lambda_i} \sum_{i=1}^{-\frac{1}{2}} m_i^b\right) \end{aligned} \tag{18}$$

The kernel $K(speech_a, speech_b)$ in (18) is linear and satisfying Mercer condition in the GMMsupervector. Now the SVM model can be mathematically derived as follows,

$$f(x) = \left(\sum_{i=1}^{L} a_i t_i b(x_i)\right)^t b(x) + d = w^t b(x) + d \tag{19}$$

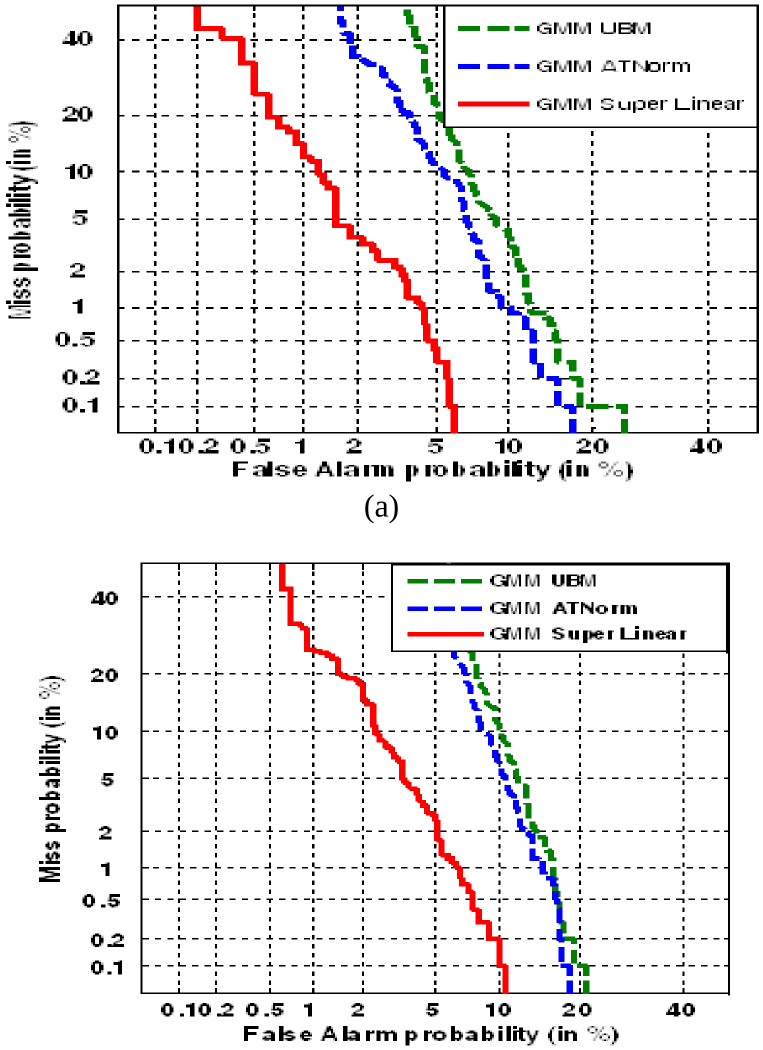
This implies we just need to process a solitary inward product between the objective model and the GMM supervector to acquire a score.

EXPERIMENTS

Experimental assessment of the original and sparsified reconstructed speech signal continued with the 160 speakers from TIMIT corpus. We have taken corpus of 10 male and 10 female from each dialect group. In the experimental evaluation following combination has been taken into consideration.

i. 2048 GMMUBM mixture components
ii. GMM MAP training, mean with relevance factor 16.
iii. GMM super vector generation using MAP adaptation
iv. SVM training utilizing GMM supervector label.
v. An SVM background buildup with the help of GMM supervectors.
vi. For registration of speakers to ASR 8 GMM supervector has been generated from 8 conversations.
vii. SVM model training with the help of target GMM super vector.

We comparer the result of ASR system of TIMIT corpus of sparsified reconstructed/enhanced with GMMUBM, GMMATNorm and GMM Super Linear based modeling techniques in Fig.4.(a). The EER of GMM Super Linear is 3%, GMMATnorm is 6% and GMM UBM is 8% respectively. Performance of linear GMM super vector kernel is best compare to standard GMM configuration. The less computational complexity results the excellent performance in linear GMM super vector kernel based ASR system. In Fig. 4. (b) we comparer the result of ASR system TIMIT corpus with GMMUBM, GMMATNorm and GMM Super Linear based modeling techniques. The EER of GMM Super Linear is 4%, GMMATnorm is 8% and GMM UBM is 10% respectively. In the case of the original TIMIT corpus performance of linear GMM super vector kernel is best compare to standard GMM configuration.



(a)

(b)

Figure 4. **(a)** Comparison of DET curve of different modelling based ASR system for sparsified reconstructed TIMIT corpus. (b) Comparison of DET curve of different modelling based ASR system for TIMIT corpus.

CONCLUSIONS

Recent year an extensive research work has been done on sparse representation. This paper introduces a novel algorithm for speech signal sparsing, compression and enhancement and its application in ASR system. The TIMIT ASR evaluation corpus is used to conduct the experiment. It was shown in the experimental result that the ASR efficiency is better after application of novel sparse/enhancement algorithm. The sparse representation application and scope has been merged with machine learning and computer vision. The ASR efficiency has been compared in two cases and found that the speech enhancement with sparsification improves significantly. Similar ASR efficiency trends have been observed in all different modeling techniques. The EER performances of ASR with application of novel algorithm of sparsification on TIMIT corpus are 3% (GMM Super Linear), 6% (GMMATnorm) and 8% (GMM UBM) correspondingly. EER performances of ASR have been observed without sparsification for TIMIT corpus as 4% (GMM Super Linear), 8% (GMMATnorm) and 10% (GMM UBM) correspondingly. One can observed 1% of EER improvement in case of GMM Super Linear and 2% of EER improvement in case of GMMATnorm, GMMUBM.

As a part of future work to enhance the ASR efficiency the algorithm can be modified in terms of mid frequency component.

REFERENCES

[1] B. K. Natarajan, “Sparse approximate solutions to linear systems,” SIAM J. Comput., Vol. 24, No. 2, pp. 227-234, 1995.

[2] M. Huang et al., “Brain extraction based on locally linear representation based classification,” NeuroImage, Vol. 92, pp. 322-339, May 2014.

[3] X. Lu and X. Li, “Group sparse reconstruction for image segmentation,” Neurocomputing, Vol. 136, pp. 41-48, Jul. 2014.

[4] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, “Sparse representation for computer vision and pattern recognition,” Proc. IEEE, Vol. 98, No. 6, pp. 1031-1044, Jun. 2010.

[5] D. L. Donoho, ``Compressed sensing,” IEEE Trans. Inf. Theory, Vol. 52, No. 4, pp. 1289-1306, Apr. 2006.

[6] E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. Inf. Theory, Vol. 52, No. 2, pp. 489-509, Feb. 2006.

[7] E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. Inf. Theory, Vol. 52, No. 2, pp. 489-509, Feb. 2006.

[8] A. M. Bruckstein, D. L. Donoho, and M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images,” SIAM Rev., Vol. 51, No. 1, pp. 34-81, 2009.

[9] D. L. Donoho, “Compressed sensing,” IEEE Trans. Inf. Theory, Vol. 52, No.

4, pp. 1289-1306, Apr. 2006.

[10] E. Candès and J. Romberg, “Sparsity and incoherence in compressive sampling,” Inverse Problems, Vol. 23, No. 3, p. 969, 2007.

[11] M. Elad, M. A. T. Figueiredo, and Y. Ma, “On the role of sparse and redundant representations in image processing,” Proc. IEEE, Vol. 98, No. 6, pp. 972-982, Jun. 2010.

[12] M. Elad, “Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing,” New York, NY, USA: Springer-Verlag, 2010.

[13] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, “Sparse representation for computer vision and pattern recognition,” Proc. IEEE, Vol. 98, No. 6, pp. 1031-1044, Jun. 2010.

[14] J.L. Starck, F. Murtagh, and J. M. Fadili, “Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity,” Cambridge, U.K.: Cambridge Univ. Press, 2010.

[15] D. Garcia, Romero, X. Zhou, and C. Y. Espy-Wilson, “Multicondition training of gaussianplda models in i-vector space for noise and reverberation robust speaker recognition,” In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2012), 2012.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Trans. Audio, Speech and Language Processing, Vol.19, No. 4, pp. 788 -798, May 2011.

[17] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “The effect of noise on modern automatic speaker recognition systems,” In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2012), 2012.

[18] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, “ Evaluation of i-vector speaker recognition systems for forensic application,” In Proc. Interspeech 2011, pp. 21-24, 2011.

[19] SadaokiFurui, “50 Years of Progress in Speech and Speaker Recognition Research”, ECTI Transactions on Computer and Information Technology, 2005, Vol.1, No.2, pp 64-74.

[20] Prashanthi,Satyanand, Dr. E.G. Rajan and Pat Krishanan,“Sparsification of Voice Data Using Discrete Rajan Transform and its Applications in Speaker Recognition,” IEEE International Conference on Systems, Man, and Cybernetics October 5-8, San Diego, CA, USA, pp. 429-434, 2014.

[21] John H.L. Hansen and Taufiq Hasan, “Speaker Recognition by Machines and Humans,” IEEE Signal Processing Magazine, November, pp. 74-99, 2015

[22] C. Nello and S.-T. John, “Support Vector Machines,” Cambridge, U.K.: Cambridge Univ. Press, 2000.

[23] W. M. Campbell, “Generalized linear discriminant sequence kernels for speaker recognition,” In Proceedings of ICASSP, pp. 161.164, 2002.

[24] Minh N. Do, “Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models,” IEEE Signal Processing Letters, Vol. 10, No. 4, pp. 115-118, 2003.