BULGARIAN ACADEMY of SCIENCES 1869

IICT Institute of Information and Communication Technologies

Speech Processing Experiences

Ayyoob Jafari

May 2013

Speech Processing Experiences

- Speech Enhancement (MSC Thesis)

- Text to Speech Synthesis (Gooya System)

Research Center of Intelligent Signal Processing(RCISP) Until 2007,

- Speech Recognition (PHD Thesis)

Speech Enhancement

- Analyzing Different Speech Enhancement Approaches

- Adaptive Wavelet Based Speech Enhancement algorithm

- Proposing a chaotic silence detection system

TTS System

- A Persian Rule-based TTS System (Gooya)
- A diphon and phoneme speech database
- A database for grammar rules
- Join Gooya to a screen reader system (Shiva)

Research Center of Intelligent Signal Processing(RCISP) from 2004 to 2007

Contents of Speech Recognition Part

1-Achievements

2- Methods

3- Database and Speech Recognition System

4-Results

Achievements

- Automatic speech recognition is a speech-to-text process which is done on speech data captured by microphones. Considering recent advances in artificial intelligent researches and Man-Machine interactions, speech recognition has showed very important rule in resent researches and different academic and commercial recognition systems were developed. In such systems, recognition is done with limited success.

Achievements

- In this research, with emphasis on feature extraction methods, considering dimension reduction approaches and speech reconstructed phase space, the improvement of the accuracy of speech recognition systems has been studied. Dimension reduction algorithms studied in this research includes two models of continuous hidden variables and manifold learning algorithms. In usage of chaos theory in speech recognition, nonlinear modeling of speech reconstructed is considered.

Achievements

- The main novel technical contributions of this thesis are as follows. As our first contribution, theoretical foundation and structure of a model is introduced based on non-linear principle component analysis (NLPCA). In this model, introducing an effective algorithm, usual frequency domain features have been transformed to a new subspace. This method improves the accuracy of speech classification about 3.7% for clean speech data and isolated phoneme recognition tests in TIMIT database.

Achievements

- The second contribution of this research is based on a new dimension reduction approach based on Laplacian Eigenmaps latent variable model for speech recognition. This feature extraction approach has showed very interesting improvement in speech recognition accuracy with about 6% improvement in isolated phoneme recognition tests for clean data from TIMIT database.

Achievements

- The third contribution of this research is based on introducing a combinational model for frequency domain features and features obtained from non-linear modeling of speech reconstructed phase space. This method improves isolated phoneme recognition accuracy about 3.4% for clean data from TIMIT database. Next main contribution of this research is based on non-linear modeling of speech reconstructed phase space Poincare sections in combination with frequency domain features. Combination of features was done using fisher discrimination analysis. This method improves isolated phoneme recognition accuracy about 5.7% for clean data from TIMIT database.

Achievements

- The final contribution on this research is based on using phase space theory and Laplacian Eigenmaps. In this proposed method, Poincare sections of speech reconstructed phase space are calculated and then are transformed to a new subspace using Laplacian Eigenmaps method. Modeling is done in this final subspace and obtained features then will be combined with frequency features. This method has showed very interesting performance in robust speech recognition tests. This method improves isolated phoneme recognition accuracy about 5.7% for clean data from TIMIT database.

Methods

Continuous Latent Variable Models

Manifold Learning Algorithms

Dimension Reduction Methods

Usual Chaotic Features

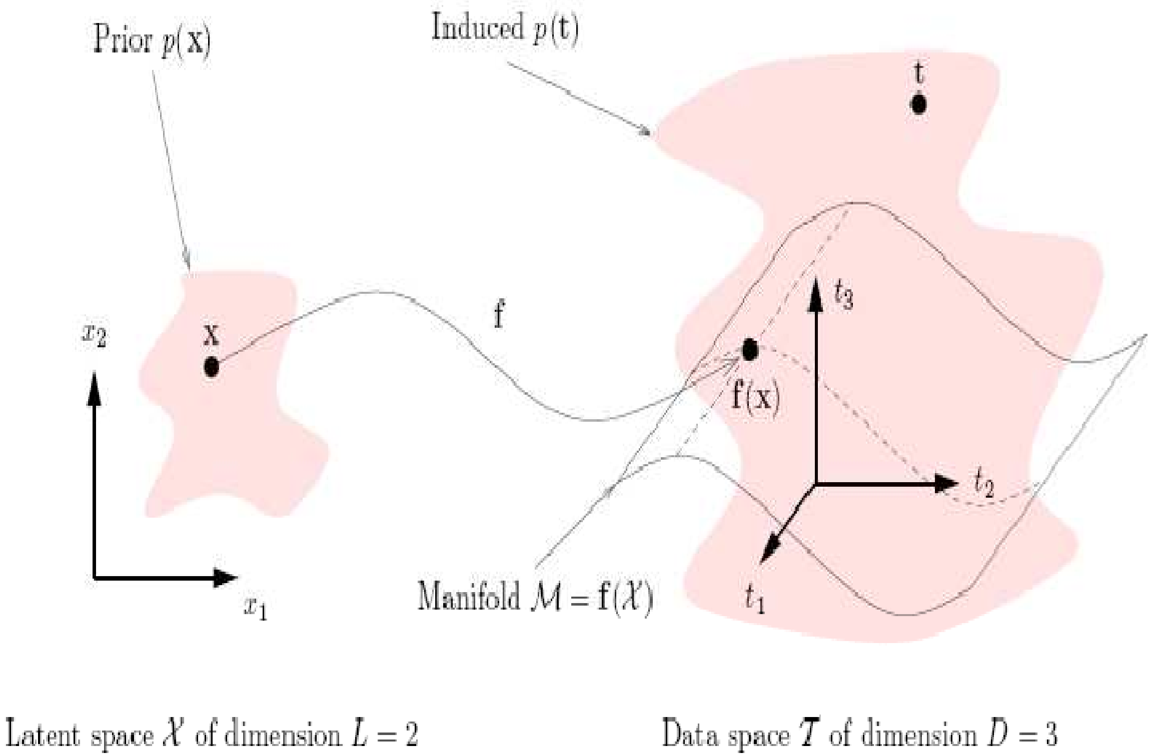
Reconstructed Phase Space Modelling

Chaotic Analysis

Methods

Continuous Latent Variable Models

Dimension Reduction Methods



Parameter Estimation Using EM algorithm

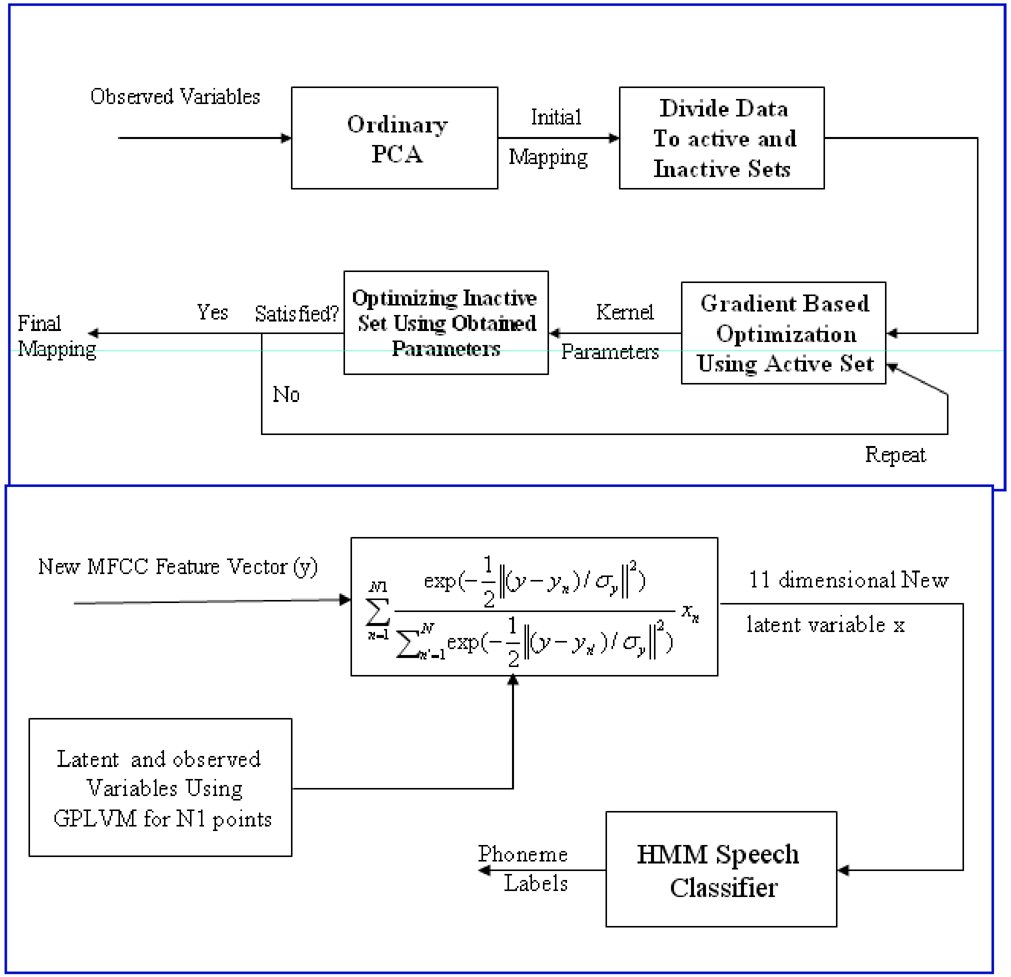
Methods

Continuous Latent Variable Models

Model	*Prior in latent space* $p(\mathbf{x})$	*Mapping* $\mathbf{f}$ $\mathbf{x} \to \mathbf{t}$	*Noise model* $p(\mathbf{t}\|\mathbf{x})$	*Density in observed space* $p(\mathbf{t})$
Factor analysis (FA)	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	linear	diagonal normal	constrained Gaussian
Principal component analysis (PCA)	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	linear	spherical normal	constrained Gaussian
Independent component analysis (ICA)	unknown but factorised	linear	Dirac delta	depends
Independent factor analysis (IFA)	product of 1D Gaussian mixtures	linear	normal	constrained Gaussian mixture
Generative topographic mapping (GTM)	discrete uniform	generalised linear model	spherical normal	constrained Gaussian mixture

Methods

Proposed Approach Using CLVM



Methods

Manifold Learning Approaches

- Locally Linear Embedding
- Laplacian Eigenmaps
- Isomaps

Methods

Laplacian Eigenmaps Manifold Learning Approach

$$X_s = (x_1, \ldots, x_N)$$

$$Y_s = (y_1, \ldots, y_N)$$

$$Y_u = (y_{N+1}, \ldots, y_{N+M})$$

$$Ly = \lambda D y$$

$$\min_x tr(XLX^T) \ s.t. \ XDX^T = I, XD1 = 0$$

$$\min_{X_u} tr\left((X_s, X_u) \begin{pmatrix} L_{ss} & L_{su} \\ L_{us} & L_{uu} \end{pmatrix} \begin{pmatrix} X_s^T \\ X_u^T \end{pmatrix} \right)$$

$$L_N = \begin{bmatrix} L_s & L_{su} \\ L_{us} & L_{uu} \end{bmatrix}$$

$$\min_{X_u} tr(X_s L_s X_s^T + X_s L_{su} X_u^T + X_u L_{us} X_s^T + X_u L_{uu} X_u^T)$$

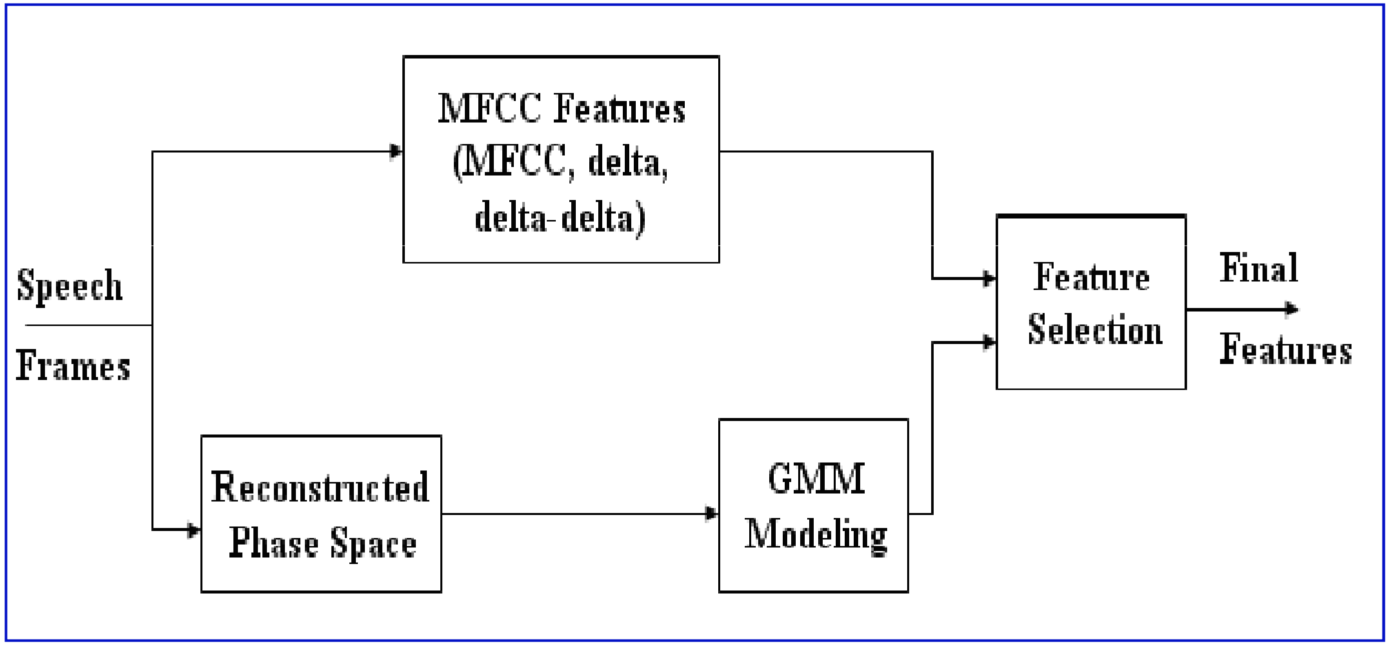
$$X_u = -X_s L_{su} L_{uu}^{-1}$$

$$x = F(y) = -\frac{1}{l_{uu}} X_s L_{su} = \frac{X_s K(y)}{1^T K(y)}$$

$$= \sum_{n=1}^{N} \frac{K(y, y_n)}{\sum_{n'=1}^{N} K(y, y_{n'})} x_n$$

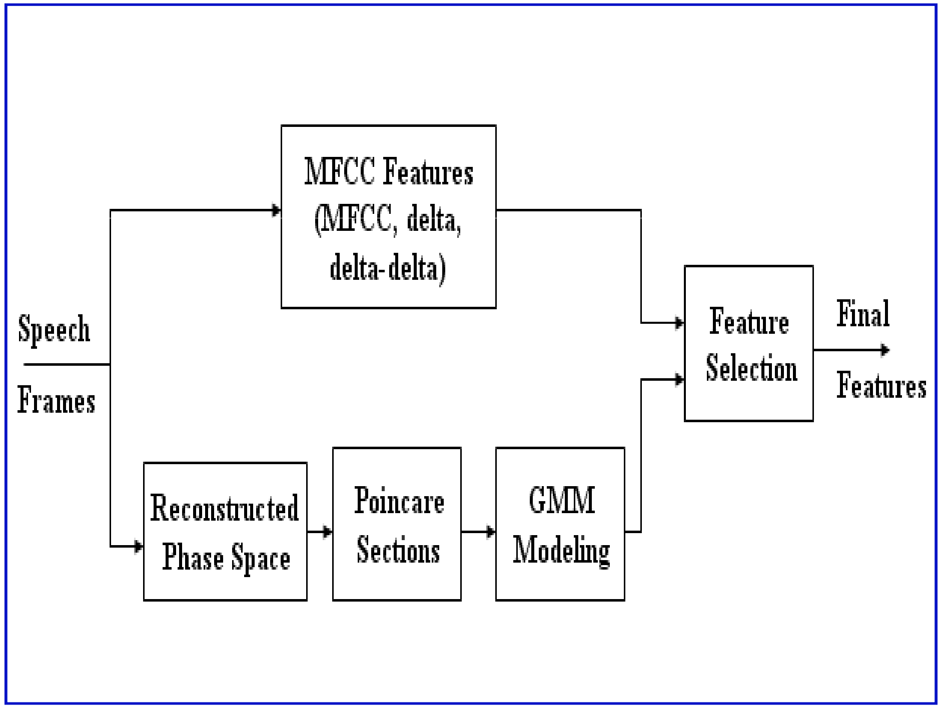
Methods

Chaotic Features Application



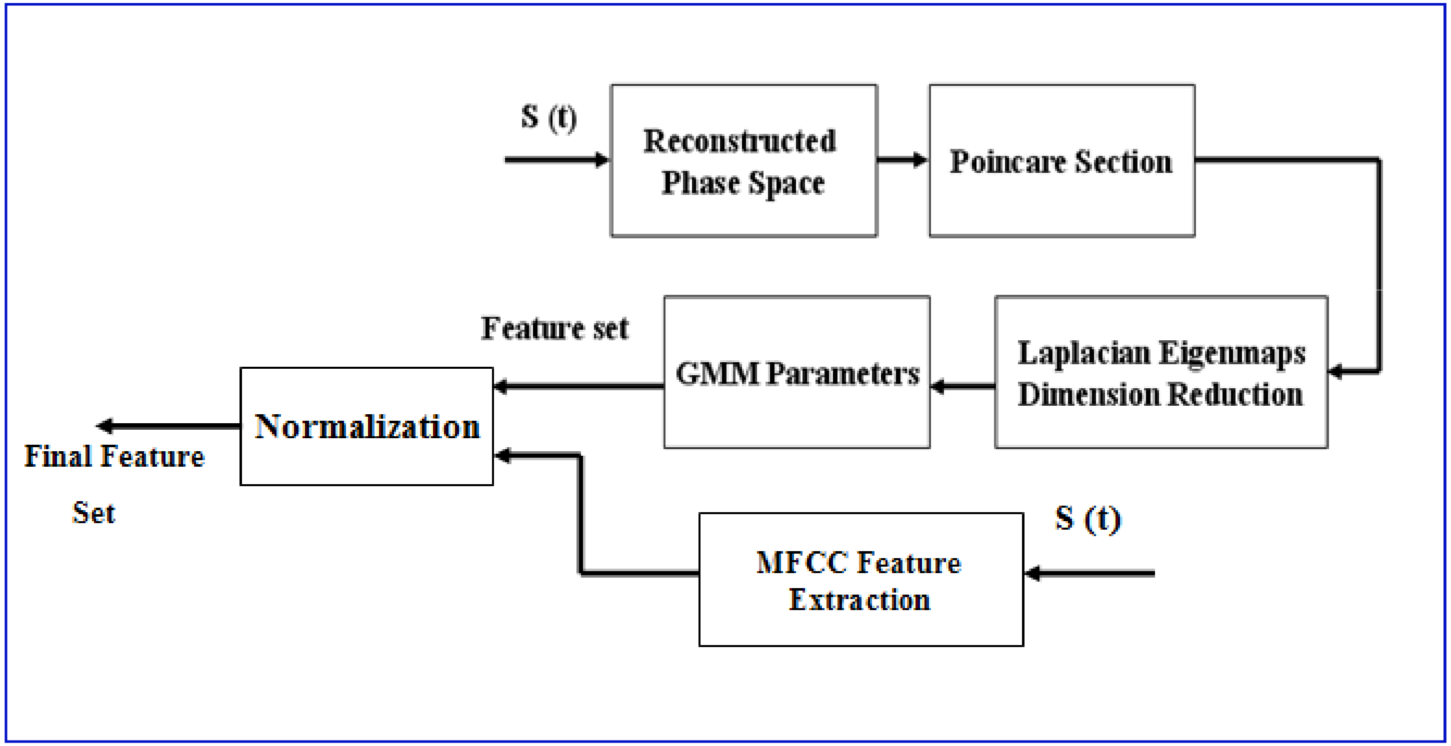
Methods

Chaotic Features Application



Methods

Proposed LE and Chaotic Subspace



Database and Speech Recognition System

- An English Speech Database (TIMIT) and a Persian Speech Database (Farsdat) are used in experiminets
- TIMIT Database consists of 6300 sentence from 10 speaker with complete test set with 1344 sentence (27% of all dataset) and core test set with 192 sentence. All phonemes categoried to 39 classes.

- Farsdat database consists of 2000 sentences from persian speakers provided by research center of intelligent signal processing (RCISP) in Iran.

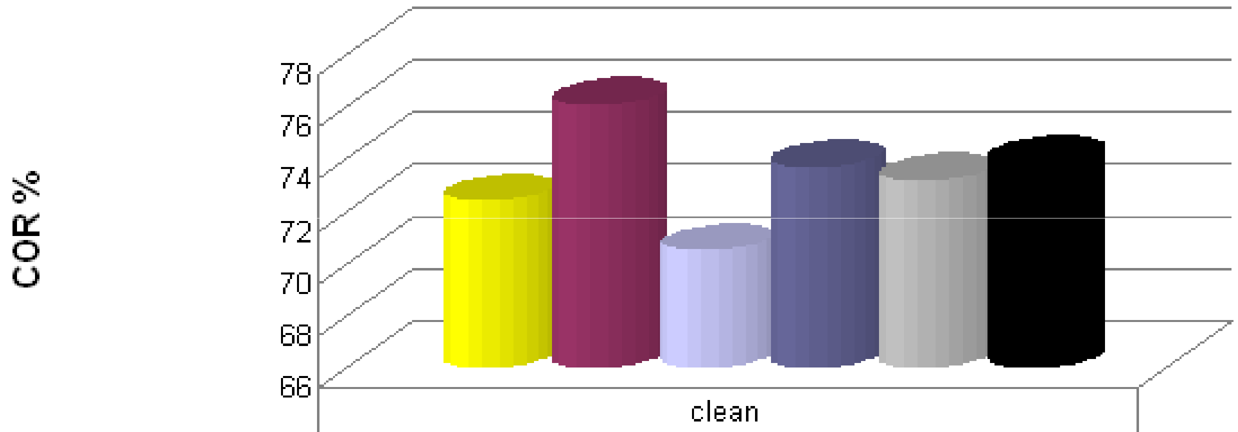
Database and Speech Recognition System

- Noisex.92 noise database used for additive noise signals.
- We used HMM toolbox for matlab and HTK toolkit (Cambridge Univeristy) for speech recognition engine in our experiments.
- 6 state with 8 Gaussian mixtures used in HMM model and frames

With 25.6 ms are used.

Results

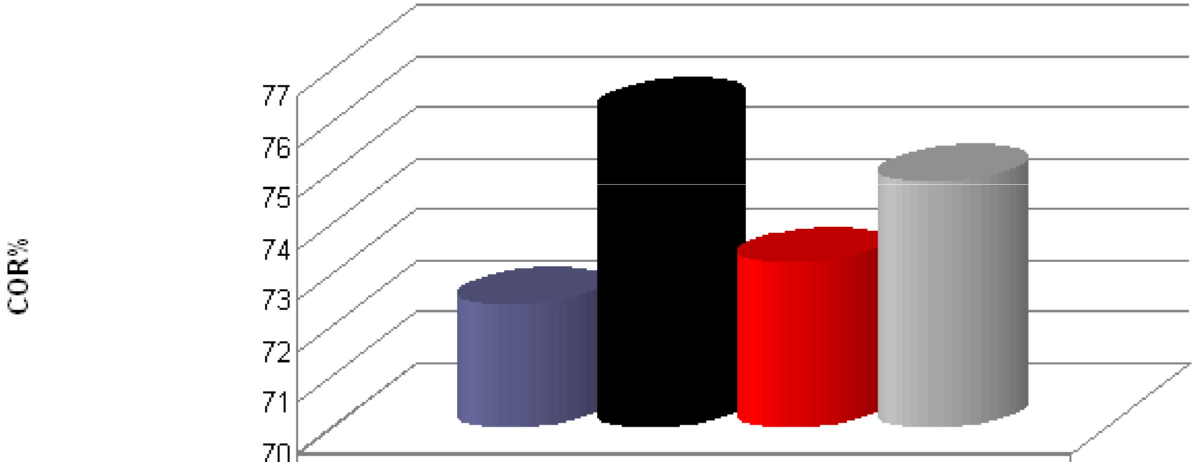
Basic MFCC,Proposed NLPCA, PCA Features COR% For TIMIT Database



	clean
Basic MFCC	72.4
Proposed NLPCA	76.1
PCA(Dim=20)	70.5
PCA (Dim=22)	73.7
PCA(Dim=24)	73.2
PCA (Dim=28)	73.9

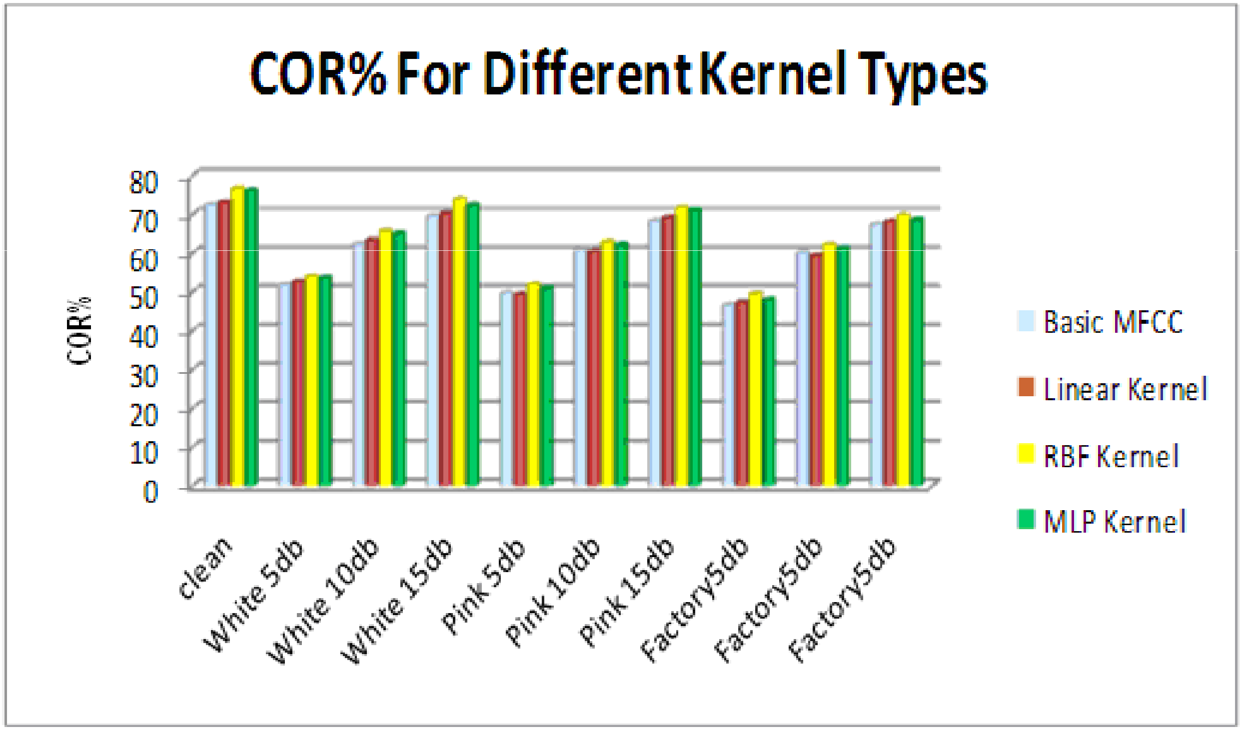
Results

Basic MFCC,Proposed NLPCA, LDA, LDA-MLLT For TIMIT Database



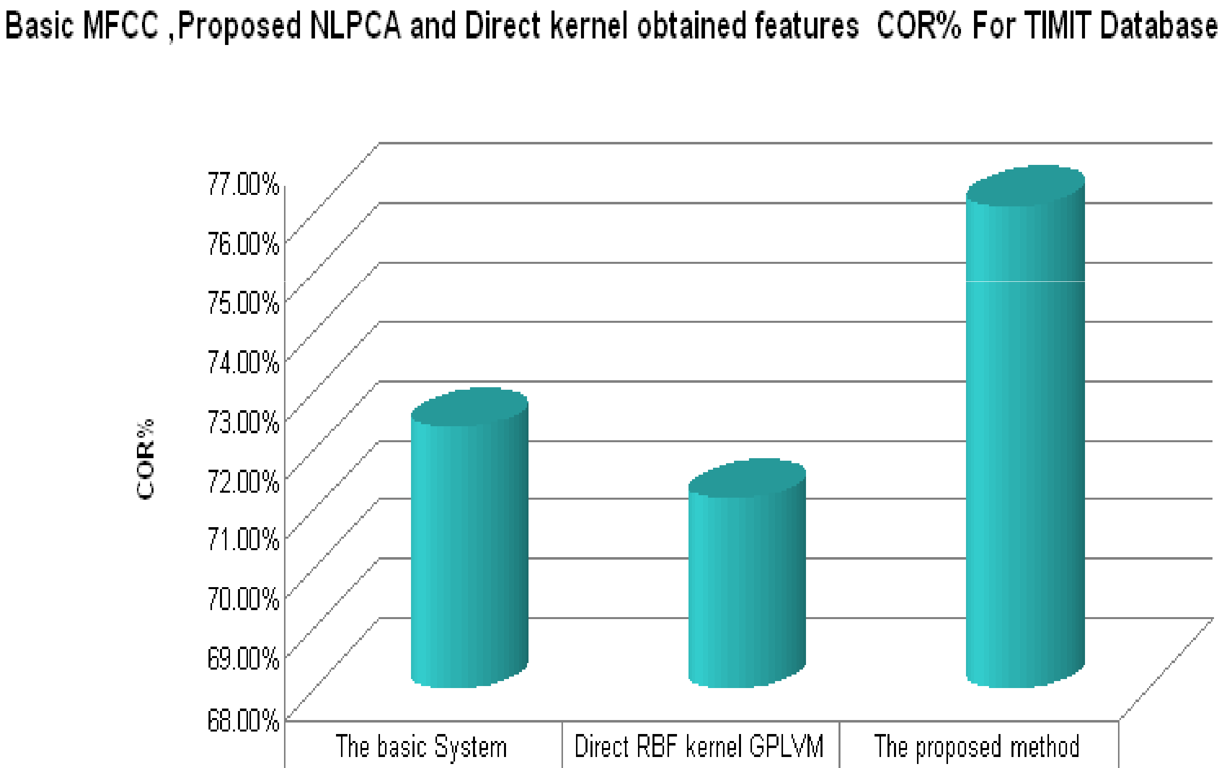
	clean
Basic MFCC	72.4
Proposed System	76.1
LDA	73.2
LDA-MLLT	74.8

Results



COR% For Different Kernel Types
COR%
80
70
60
50
40
30
20
10
0
clean
White 5db
White 10db
White 15db
Pink 5db
Pink 10db
Pink 15db
Factory5db
Factory5db
Factory5db
Basic MFCC
Linear Kernel
RBF Kernel
MLP Kernel

Results

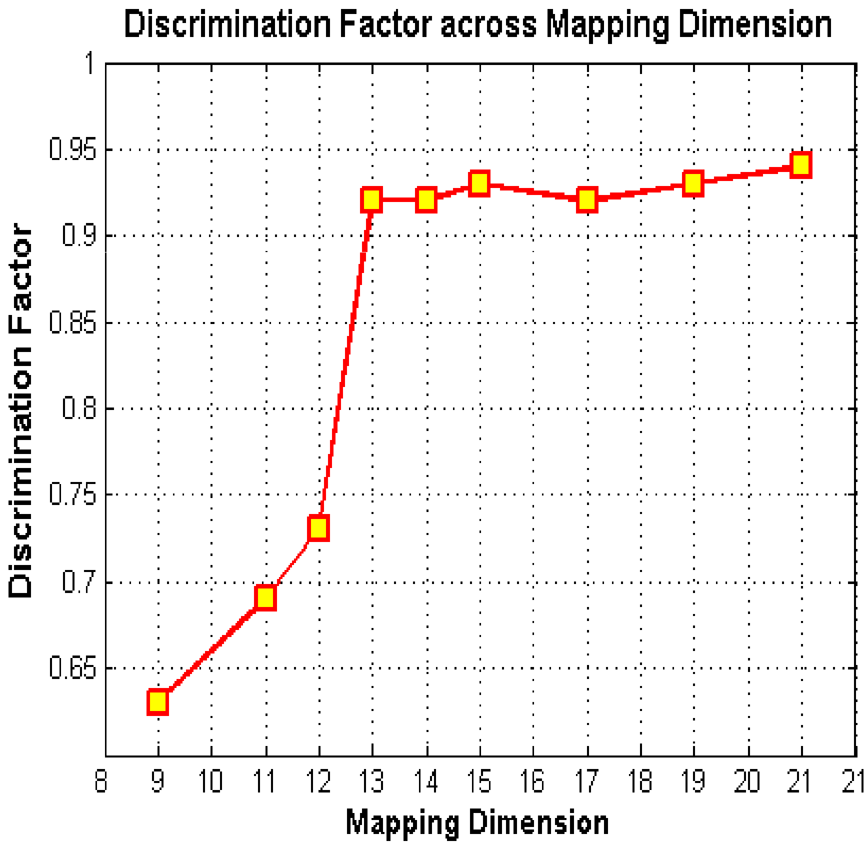


	The basic System	Direct RBF kernel GPLVM	The proposed method
COR%	72.40%	71.20%	76.10%

Results

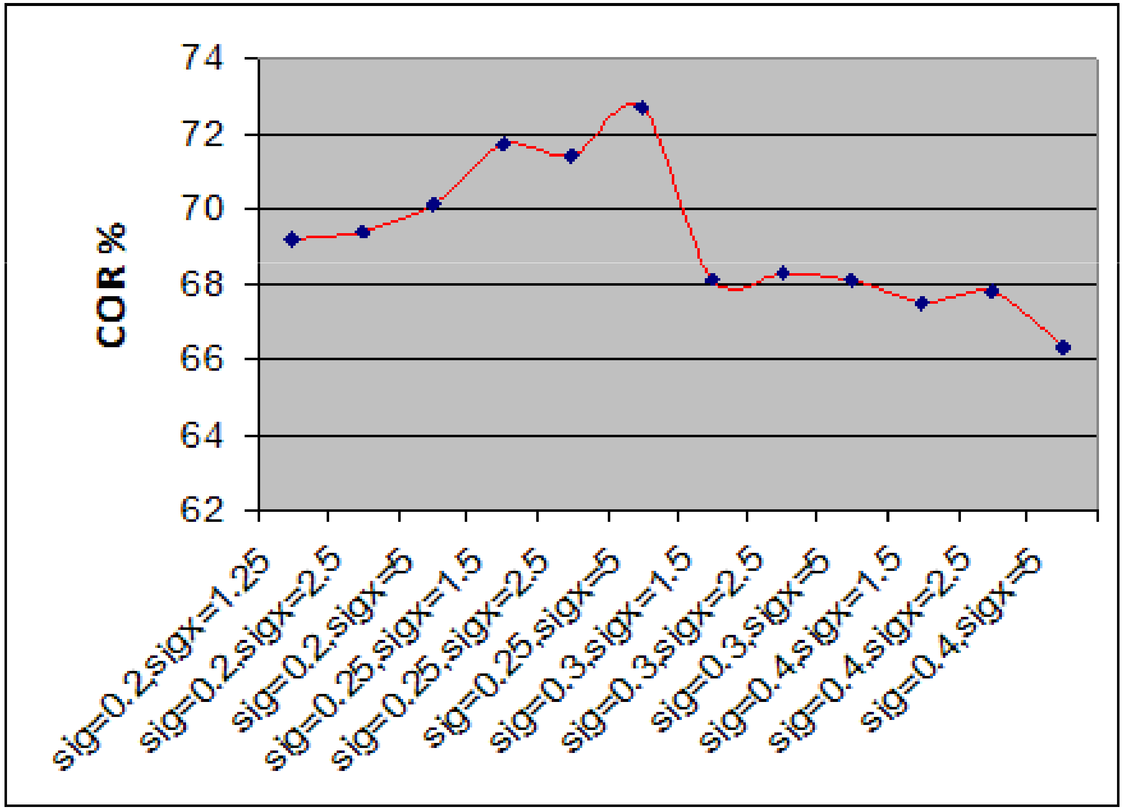
Manifold Learning

Optimum Dimension Detection



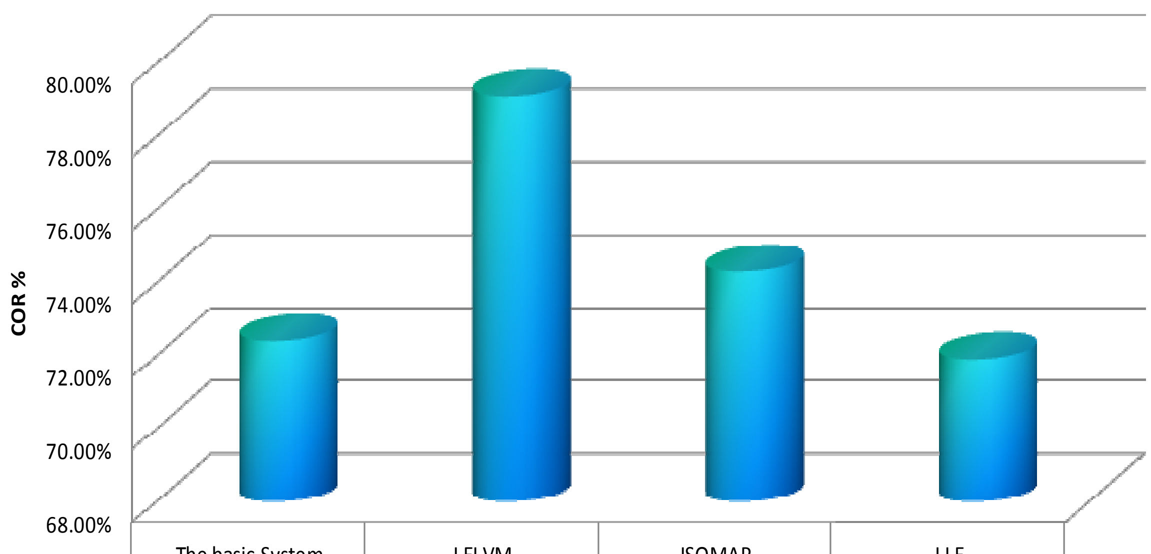
Results

Manifold Learning



Results

Manifold Learning Algorithms

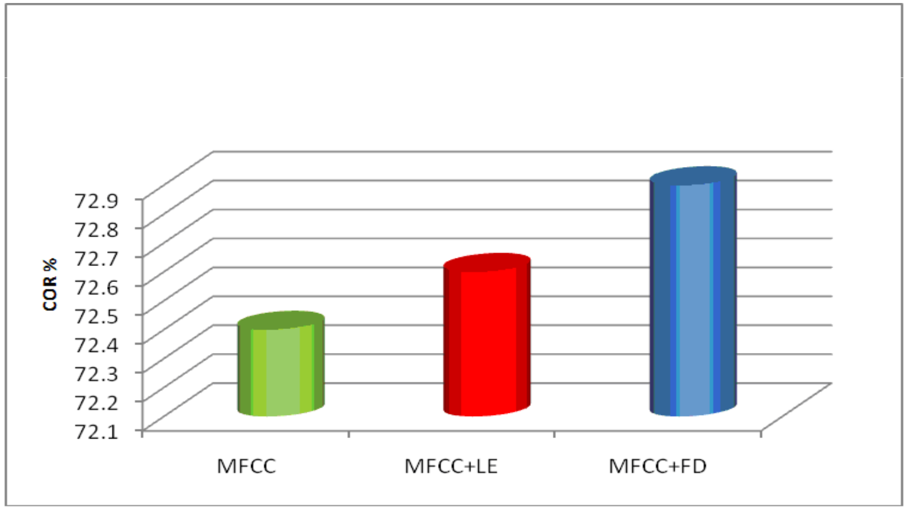


	The basic System	LELVM	ISOMAP	LLE
Series1	72.40%	79.10%	74.30%	71.90%

Results

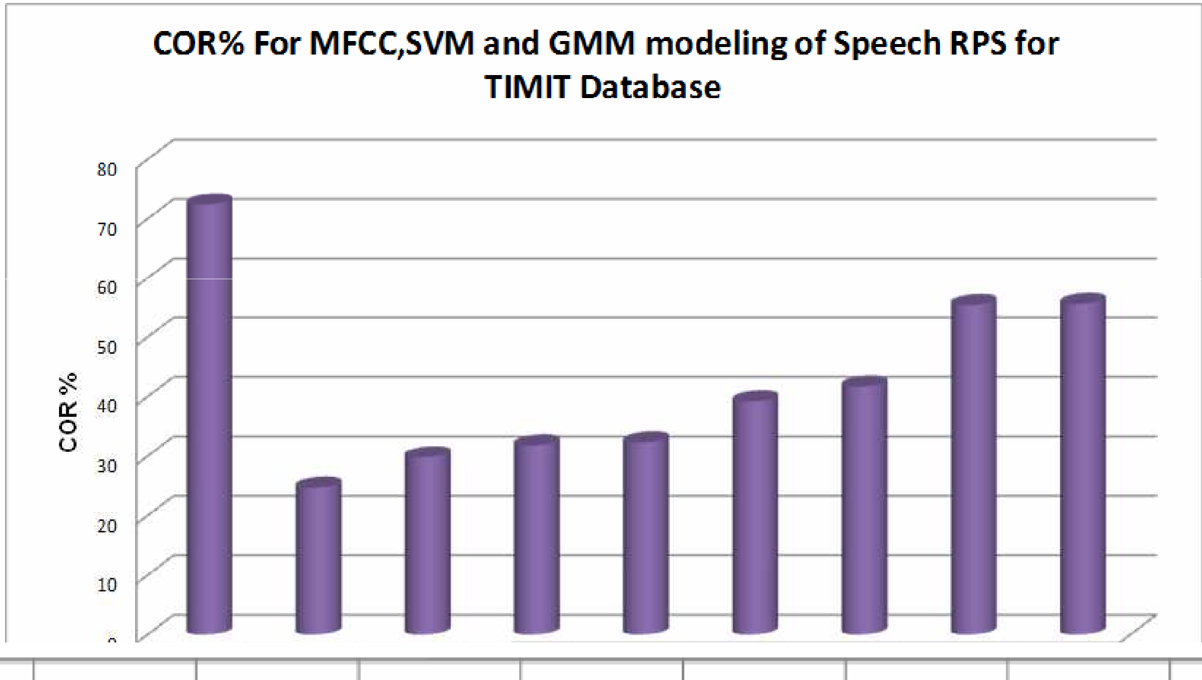
Chaotic Features Application

Phones/LE's	/ aa /(30)	/u/(30)	/s/(30)
λ_1	$0.051 \pm .012$	$.103 \pm .040$	$-.032 \pm .016$
λ_2	$-.003 \pm .006$	$-0.014 \pm .05$	$-0.50 \pm .016$
λ_3	$-0.93 \pm .021$	$-.128 \pm .039$	$-0.76 \pm .021$



Results

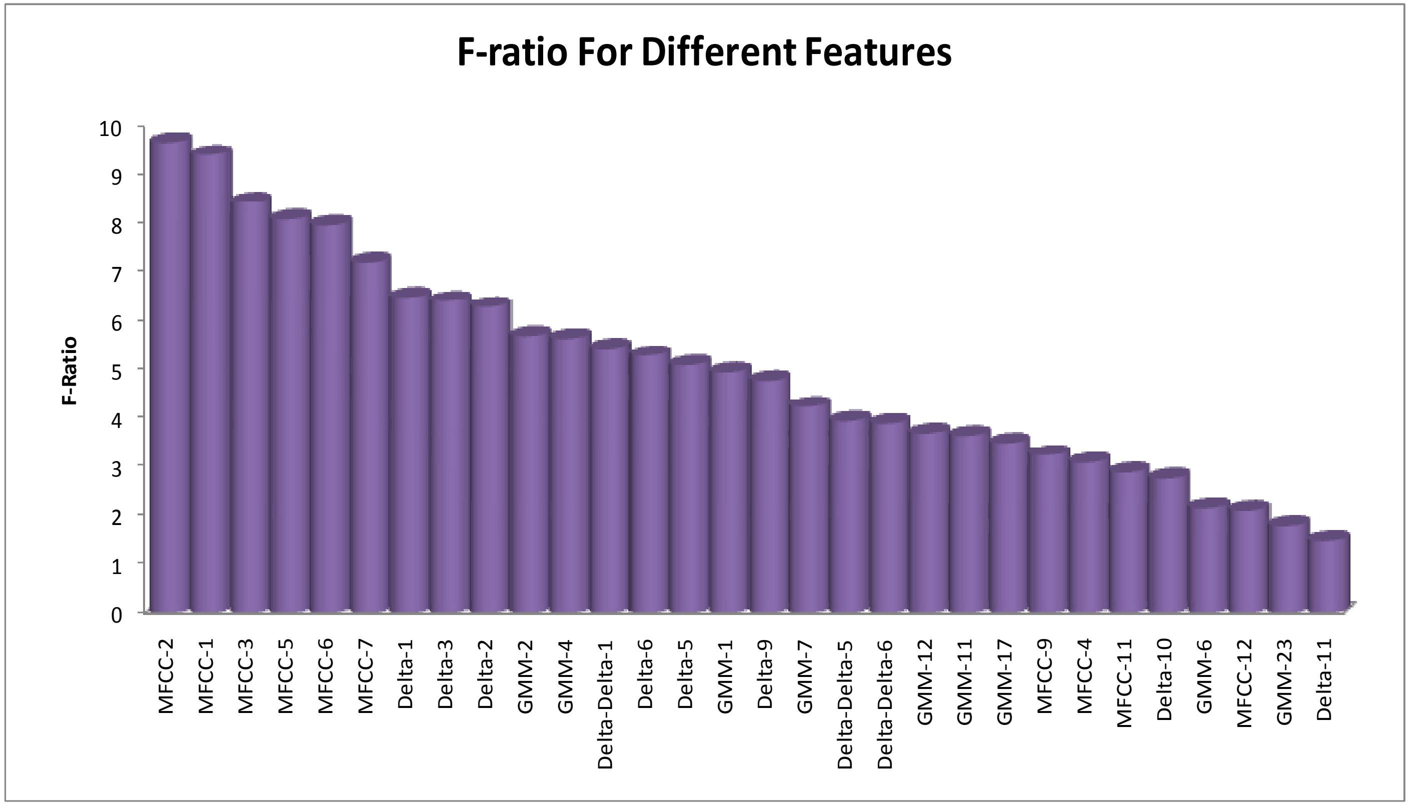
Chaotic Features Application



	MFCC	SVM-4	SVM-8	SVM-16	SVM-32	GMM-4	GMM-8	GMM-16	GMM-32
COR%	72.4	24.7	29.8	31.8	32.3	39.2	41.7	55.4	55.6

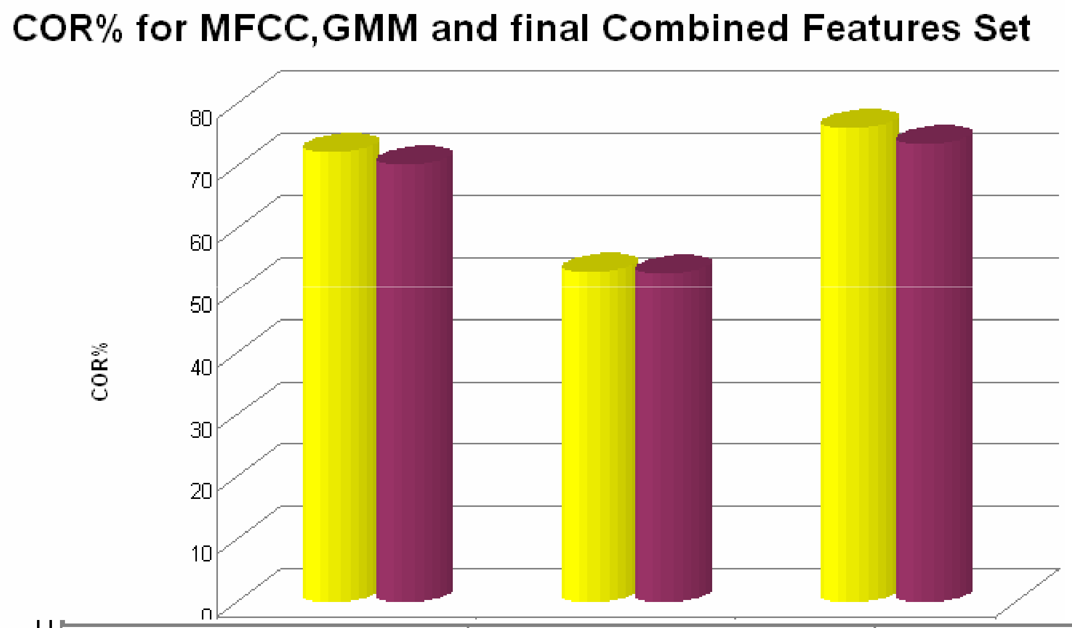
Results

Chaotic Features Application



Results

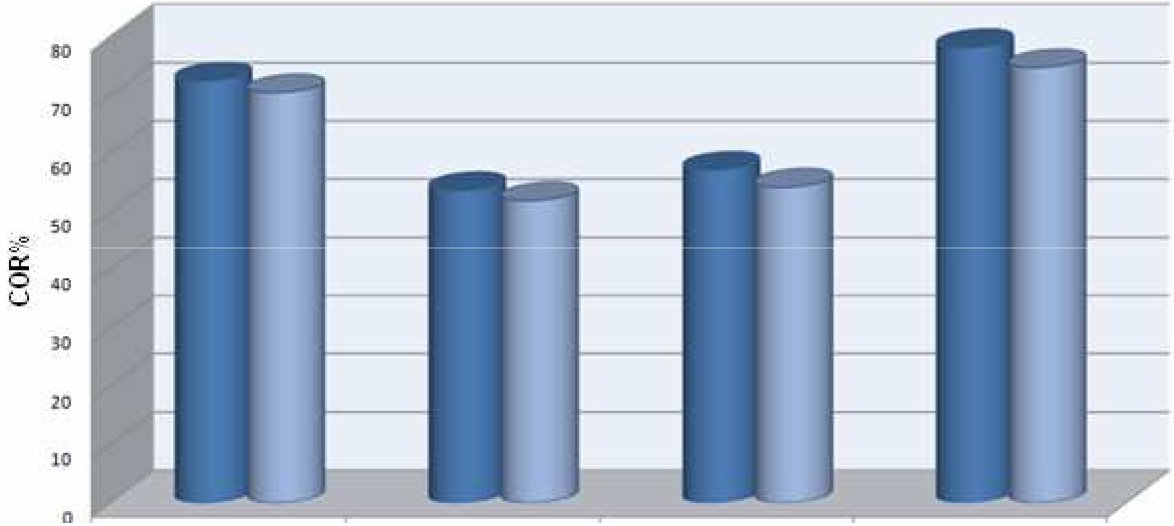
Chaotic Features Application



	MFCC(39 Features)	GMM(68 Features)	Combined Features
TIMIT	72.4	53.1	76.2
FarsDat	70.4	52.9	73.7

Results

Poincare Sections

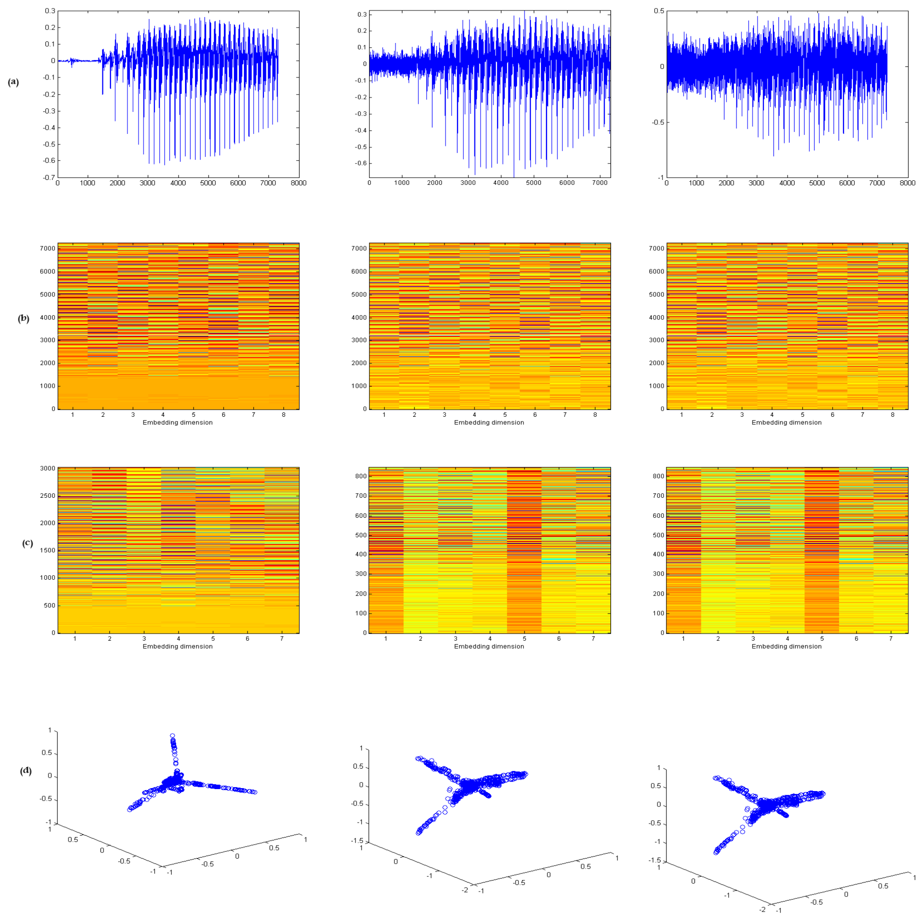


0

	MFCC	GMM (Povinelli et al)	GMM(Poincare)	Proposed Combined Method
TIMIT	72.4	53.6	57.1	78.1
FarsDat	70.4	51.9	54.2	74.7

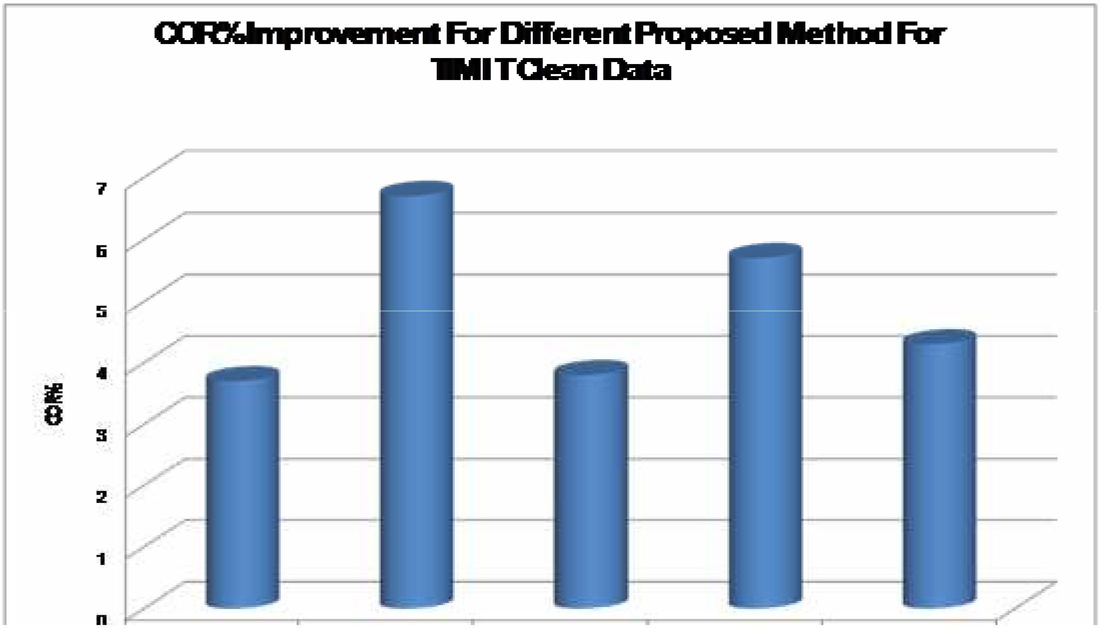
Results

Manifold Learning And Chaotic Analysis



Conclusions

All Methods Results for Clean Speech Signal



	NLPCA	LELVM	RPS-Based	Poincare	Poincare-LELVM
COR%	3.7	6.7	3.8	5.7	4.3