This paper introduces a combinational feature extraction approach to improve speech recognition systems. The main idea is to simultaneously benefit from some features obtained from Poincaré section applied to speech reconstructed phase space (RPS) and typical Mel frequency cepstral coefficients (MFCCs) which have a proved role in speech recognition field. With an appropriate dimension, the reconstructed phase space of speech signal is assured to be topologically equivalent to the dynamics of the speech production system, and could therefore include information that may be absent in linear analysis approaches. Moreover, complicated systems such as speech production system can present cyclic and oscillatory patterns and Poincaré sections could be used as an effective tool in analysis of such trajectories. In this research, a statistical modeling approach based on Gaussian mixture models (GMMs) is applied to Poincaré sections of speech RPS. A final pruned feature set is obtained by applying an efficient feature selection approach to the combination of the parameters of the GMM model and MFCC-based features. A hidden Markov model-based speech recognition system and TIMIT speech database are used to evaluate the performance of the proposed feature set by conducting isolated and continuous speech recognition experiments. By the proposed feature set, 5.7% absolute isolated phoneme recognition improvement is obtained against only MFCC-based features. © 2010 American Institute of Physics. [doi:10.1063/1.3463722]

I. INTRODUCTION

For many years, features extracted from frequency domain analysis, such as MFCCs obtained from perceptual interpretation of speech power spectra, are used in traditional speech recognition systems with an acceptable performance. MFCC features are derived from the fast Fourier transform (FFT) magnitude spectrum by applying a filter bank which has filters evenly spaced on a warped frequency scale (Ashkenazy et al., 2001). The logarithm of the energy in each filter is calculated and accumulated before a discrete cosine transform is applied to produce the MFCC feature vector. The frequency warping scale used for filter spacing in MFCC is the Mel (Melody) scale. The Mel scale is a perceptually motivated scale that was first suggested by Stevens and Volkman in 1937 (Moore, 1995). The first and second derivations of MFCC features are called delta and delta-delta features, respectively.

There are many signals such as signals generated through nonlinear differential or difference equations which typically show wide power spectral specifications. In such cases, the frequency-domain representation does not contain all the information of a signal, because it is impossible to differentiate between such signals that have the same power spectra and different phase space reconstruction (Povinelli et al., 2006). Figure 1 illustrates the time series, frequency spectrum, and reconstructed phase space obtained from logistic map \( x_{n+1} = k \times x_n \times (1-x_n) \) with \( k=4 \) (Theiler et al., 1992). As could be seen in Fig. 1, the reconstructed phase space representation of the system has a more suitable presentation in comparison to frequency domain representation, considering the stationary constraint.

Also, there are some experimental and theoretical evidences introduced in recent researches (McGowan, 1988; Teager and Teager, 1989; Kaiser, 1983; Maragos et al., 2002;
Kokkinos and Maragos, 2005) that proof the existence of strong nonlinear phenomena in speech production system not considered in linear assumptions. One of the main evidences of nonlinearity in speech is turbulence. The turbulence of speech, especially in fricative phonemes, is one of the main problems in speech recognition task which leads to a nonrobust recognition system (Kokkinos and Maragos, 2005).

Considering speech signal turbulence can lead us to application of chaos analysis to speech recognition field. Chaos theory typically leads to extraction of chaotic features such as Lyapunov exponents, fractal dimension, and correlation dimension in signal processing tasks. In these approaches, feature extraction is done on the multidimensional phase attractor reconstructed from one-dimensional speech signal using a time-delay based embedding approach (Pitsikalis et al., 2003; Abarbanel, 1996; Kantz and Schreiber, 1997).

Most researches in application of reconstructed phase space (RPS) focus on estimation of dynamical invariants such as metric (Lyapunov exponents, dimension), density (Ott, 1993), and topology invariants. Correlation dimension (Grassberger and Procaccia, 1983) and box counting (Ott, 1993) are some famous methods in estimation of dimension. However, there is really little literature on applying RPS models or RPS derived models to the tasks of signal classification.

The application of typical chaotic features in signal processing is well-studied in literature (Maragos et al., 2002; Kokkinos and Maragos, 2005; Pitsikalis et al., 2003; Povinelli et al., 2004). As mentioned in literature, such features could not help us to improve speech recognition performance significantly.

Speech signal phase space attractor can be achieved using a time delay based embedding process. If we consider the obvious oscillatory patterns in voiced segments of speech signal and oscillatory model of human vocal tract, we could expect the repetitions in reconstructed phase space attractor (Bohez and Senevirathne, 2001). Usually, Poincaré sections are the cross-sections through these trajectories in phase space attractor that let us to estimate speech deviation (Ku, 1995). Usually, Poincaré sections are triggered by crossing a threshold in multidimensional space, rather than by a time-domain sampling process. The application of Poincaré section in speech requires a phase space reconstruction for the given waveform $x(n)$ using a suitable dimension $d$ and a time delay $\tau$ (Maragos et al., 2002; Kokkinos and Maragos, 2005).

Povinelli et al. (2006) introduced a statistical modeling approach on speech RPS directly and showed that the extracted features from speech RPS could gain good performances in signal classification tasks. Using some isolated phoneme recognition tests, they compared the power of phase space features and MFCC features (Povinelli et al., 2002). However, their approach did not consider the application of typical MFCC features which one could not simply neglect them in speech recognition field. So, the performance of their method was degraded significantly in continuous speech recognition tests. So, in this research, we focus on two main points. First, phase space features will be extracted from Poincaré sections instead of direct RPS. Second, we combine these RPS-based features with ordinary MFCC features to achieve a suitable and powerful front end for speech recognition purposes (Banbrook et al., 1999).

The organization of this paper is as follows. Section II summarizes the overall structure of the proposed feature extraction method. Section III describes the Poincaré sections applied to speech signal RPSs. This section also includes the embedding procedure used in deriving RPS trajectories from speech signals. Section IV introduces the GMM based statistical modeling approach applied to the Poincaré section. Section V describes the proposed method of pruning the combination of Poincaré-based and MFCC-based features based on Fischer discriminant analysis. Section VI discusses the experimental setup and results. Section VII consists of the discussion about advantages of the proposed approach for speech recognition, followed by the overall conclusions in Sec. VIII.

II. THE OVERALL STRUCTURE OF THE FEATURE EXTRACTION METHOD

As aforementioned, the proposed approach is based on the direct statistical modeling of Poincaré sections of speech RPSs and the combination of the resulted features with usual MFCC features. Figure 2 shows the block diagram of this approach. As shown in Fig. 2, RPSs are reconstructed from speech signal frames using time delay based approach that
will be described in Sec. III. Next, a GMM based statistical modeling is applied to Poincaré sections of the reconstructed phase space. Finally, a feature selection algorithm is applied to the combination of the frequency domain features (MFCC) and phase space features (GMM features). Of course, the selection pattern is derived one time by an offline process (Brown, 1993).

Feature selection is done over a 39-dimensional MFCC feature set that consists of MFCC, delta, and delta-delta MFCC coefficients and a 60-dimensional feature set obtained from the parameters of a four mixture GMM model developed over the Poincaré section of speech frame (Johnson et al., 2005). Finally a pruned 47-dimensional final feature set is obtained as the final feature set. The time delay based approach for speech signal embedding to phase space and Poincaré section estimation will be described in Sec. III. Details of GMM based statistical modeling approach will be described in Sec. IV. Section V discusses the feature selection approach (Cohen, 1986).

III. PHASE SPACE RECONSTRUCTION AND POINCARÉ SECTIONS

A. Phase space reconstruction

In this section, we have a brief review on the speech phase space attractor reconstruction theory. The basis of embedding approach is that we should achieve a mapping that gives us the access to the dynamic structure of a generating system. Whitney (1936) showed that if we have a \( d \)-dimensional topological space, map \( f: \mathbb{R}^d \rightarrow \mathbb{R}^{2d+1} \) can be an embedding from original system and such embedding is a homeomorphic map. Takens (1980) showed that such mapping is continuously differentiable diffeomorphism that describes the dynamics of the original system.

The main step in embedding approach is obtained from Sauer et al. (1991). Work using Takens work and similar theories showed that every time-delay map with sufficient dimension is an embedding and guarantees the topological equivalence with original generating system dynamic. In addition, they found that if \( d_0 \) will be the dimension of the attractor obtained from the system using box-counting method, choosing \( d > 2d_0 \) guarantees that the reconstructed phase space of such a map is topologically the same as the true dynamics of the original system.

Based on such a strong theory, we could use a time-delay based mapping from one-dimensional signal to reconstruct the phase space attractor. Given \( x = \{ x_1, \ldots, x_N \} \) as one-dimensional speech signal time series points, trajectory matrix \( X \) of dimension \( d \) and time lag \( \tau_d \) is defined as

\[
X = \begin{bmatrix}
x_{1+(d-1)\tau_d} & \cdots & x_{1+\tau_d} & x_1 \\
x_{2+(d-1)\tau_d} & \cdots & x_{2+\tau_d} & x_2 \\
\vdots & \ddots & \vdots & \vdots \\
x_N & \cdots & x_{N-(d-2)\tau_d} & x_{N-(d-1)\tau_d}
\end{bmatrix}
\]

We could estimate time lag \( \tau_d \) from the correlations among speech samples. Narayanan and Alwan (1995) suggested the mutual information measure for estimation of the time lags between speech samples. Mutual information measure is defined as

\[
I(T) = \sum_{n=1}^{N-T} P(x(n), x(n+T)) \log_2 \left( \frac{P(x(n), x(n+T))}{P(x(n))P(x(n+T))} \right),
\]

where \( x(n) \) denotes speech samples, \( P \) means probability function, and \( I \) is mutual information measure. The first minimum of \( I(T) \) is the optimum time lag value. Figure 3 shows the \( I(T) \) calculated across the various time delay values for /a/ phoneme from Texas Instruments (TI) and Massachusetts Institute of Technology (MIT) speech corpus.

As shown in Fig. 3, the best value for the time lag is 6. After fixing \( \tau_d = 6 \), we had to determine the value of the embedding dimension. We used false neighbors measure to determine the embedding dimension. False neighbors are points on reconstructed phase space that are initially far apart to come near to each other. We should select embedding dimension in a manner that the false neighbor’s percentage becomes minimum (Kokkinos and Maragos, 2005). Using this measure for 1000 speech phonemes taken randomly from TIMIT corpus, we found \( d = 8 \) as the best value for the embedding dimension of speech signal. Figure 4 shows a sample of three-dimensional reconstructed phase space attractor for /a/ phoneme taken from TIMIT database.

B. Poincaré sections

According to cyclic properties of speech signal in its voice segments, the related reconstructed phase space trajectories repeat by every pitch period. As Poincaré sections have been proved as an appropriate tool for cyclic trajectories analysis and dimension reduction purposes in reconstructed phase space of signals, this could be a reasonable motivation for its application in the field of speech processing.

Poincaré sections are like the snapshots of trajectories taken at regular intervals. Let us consider the reconstructed
phase space of a three-dimensional dynamical system \((X_1, X_2, X_3)\); choosing a section plane in this space and marking points through which the trajectory crosses this section gives the sense of a Poincaré section. In higher order embedding dimensions, like those involved in speech processing tasks, we had to select hyperplanes instead of two-dimensional sections. The idea of Poincaré sections for three-dimensional systems is generalized for trajectories with higher dimensions. In general, for an \(n\)-dimensional trajectory, we had to use a \((n-1)\)-dimensional hyperplane that is transverse to the processed trajectory. Considering such a defined structure for Poincaré sections, we can reduce the feature set dimension by one. Figure 4 presents a three-dimensional RPS and its two-dimensional Poincaré section for a sample phone of /a/ taken from TIMIT corpus. In this instance, \(\tau = 6\).

The main problem in Poincaré section evaluations for speech frames raises for the proper selection of initial points and flow directions to form the transverse hypersurfaces which cross the reconstructed phase space trajectories. In this research, to calculate Poincaré sections, the following steps are followed:

1. For a selected speech frame, the reconstructed phase space is constructed using time delay approach.
2. For an initial point \(X(n_0)\), the neighborhood \(N(n_0)\) is searched for the \(k\) closest points, according to Euclidian distance measure.
3. The mean flow direction \(f(n_0)\) of the trajectories in neighborhood \(N(n_0)\) is computed by

\[
f(n) = \text{mean}(x[n+1] - x[n]) \quad \forall n \in N(n_0).
\]

To form the direction, we use the condition \(f[n]f[n_0] > 0.9\). This condition lets us to only select trajectories which are roughly in the same direction as the initial flow vector and prevents them from scattering around the mean flow direction, which may be occurring because of selecting trajectories which are placed in different directions.
4. Poincaré hyperplane is defined as a plane which is perpendicular to \(f(n_0)\) and simultaneously passes through \(X(n_0)\).
5. The intersection points of a reconstructed phase trajectory with its Poincaré hyperplane could be calculated...
through an interpolation process. This involves matched points of the trajectory samples which are in the neighborhood of the Poincaré section in different sides.

IV. STATISTICAL MODELING APPROACH OF RECONSTRUCTED PHASE SPACE

The nonlinear modeling of phase space attractor can be interpreted as an effective approach to overcome the unreliability of extracted features from embedded phase space due to sensitivity to initial conditions, noisy conditions, and small data sets. Kokkinos and Maragos (2005) evaluated some modeling approaches such as local polynomials, radial basic function (RBF) neural networks, fuzzy Takagi–Sugeno–Kang (TSK) models, and support vector machines in reconstructed phase space. Povinelli et al. (2006) used some statistical modeling methods such as bin-based and GMM method for modeling of speech attractor for isolated phoneme recognition.

Based on literature and our experiments, for automatic speech recognition (ASR) systems, we choose Gaussian mixture model for the modeling of Poincaré sections of reconstructed phase space of speech frames (Mitchell, 1997). Gaussian mixtures are used widely in speech processing tasks. They play a key role in hidden Markov model (HMM) as speech classifier. Here, we do not use GMM as classifier, but we use them to model data distribution in phase space attractor. GMM has two main advantages for us in modeling data distribution. First, their decision regions have soft boundaries, and second, the number of required mixtures does not require to increase exponentially with respect to embedding dimension.

In GMM, the probability of \( x_n \) is defined as

\[
p(x_n) = \sum_{m=1}^{M} w_m p_m(x_n) = \sum_{m=1}^{M} w_m N(x_n; \mu_m, \sigma^2_m),
\]

where \( P \) denotes the probability function. \( M \) is the number of selected mixtures and \( N \) denotes a normal Gaussian distribution with mean \( \mu_m \) and covariance \( \sigma^2_m \). \( w_m \) is the weight factor of \( m \)th mixture.

There is no analytical solution to determine the optimum number of mixtures actually needed for a GMM developed for a specific problem. This is dependent upon the complexity of the involved data set. In addition, as in this work, we are going to exploit the parameters of the developed GMM as the feature set which will be fed to the ASR system, we need to limit the dimension of the final feature set to have a reasonable complexity for the resulted recognition process. So, to reduce the total number of the parameters, we forced all elements of the covariance matrix to zero except of diagonal elements. In this research, we heuristically selected \( M = 4 \) in order to nearly have an acceptable dimension for the final obtained feature set. By increasing this number, the feature set dimension may dramatically increase and this causes some troubles for the used classifier, for example, may degrade its performance in addition to increasing the complexity of the classification process (Leung, 1998).

In training stage, the parameters for the GMM can be selected using the expectation-maximization algorithm (Dempster et al., 1977). As the dimension of extracted Poincaré sections of speech RPS is 7 (dimension of RPS minus 1), every Gaussian mixture has the following parameters:

(a) one-dimensional weight factor \( w_m \);
(b) seven-dimensional mean vector \( \mu_m \); and
(c) seven-dimensional covariance vector \( \sigma^2_m \).

So, for every Gaussian mixture, we have a 15-dimensional feature vector. As the used GMM has only four mixtures, we have finally a 60-dimensional feature vector that resulted from statistical modeling of Poincaré sections.

V. FEATURE SELECTION

The main goal of this research is to show the benefits of the features extracted from the speech Poincaré sections in combination with frequency domain MFCC features in speech recognition task. So, in the first step we conduct some experiments using MFCC features and the parameters of fitted model in phase space attractor separately to analyze their performance in speech classification for isolated phoneme recognition and ASR applications. We then combine these features in a unit vector after applying a normalization stage. In combination of features, we have to combine 39-dimensional MFCC features vector with 60-dimensional GMM features. So, the size of final feature set, without any extra modifications, will be 99. This size of feature vector is unacceptable in real time applications for three reasons:

1. The complexity of the recognition process will increase dramatically by the size of the input features.
2. The memory needed to save classifier parameters will increase.
3. The recognition performance may decrease because of the large dimension of the input vectors.

To overcome these limitations, Fischer discrimination analysis is used to reduce the dimension of the final feature set.

A. Fischer discrimination analysis

As cleared in Sec. IV, we finally have a 99-dimensional feature vector which must be fed to the speech classifier. As we discussed, it is reasonable to restudy this set of features and search for a pruned version, because many features may contain overlapped information and could be pruned without any loss of information. This must be noticed that the classifiers in the hand are mostly suffering from different limitations. An increase in the number of input features generally not only increases the complexity of the classification process, but also may decrease its performance. To overcome this problem, we could search for a reduced version of the original feature set, which its discrimination ability is kept (Darbyshire and Broomhead, 1996). To follow this approach, we need an assessment criterion for the features. We use Fisher discriminator analysis in order to have a discrimination assessment factor for the transformed feature set.
Linear discriminant analysis and the related Fisher’s linear discriminant are methods used in statistics and machine learning to find a linear combination of features which characterize or separate two or more classes of objects or events. The resulting combination may be used by a linear classifier or more commonly for dimensionality reduction of features before applying to a more complex classifier (Mika et al., 2001).

Fisher discriminator analysis for a two-class problem leads to Eq. (5), which evaluates Fisher’s measure or $F$-ratio,

$$ F_{\text{ratio}}(W) = \frac{W^T S_B W}{W^T S_W W}. \tag{5} $$

$S_B$, the “between class scatter matrix,” is given by

$$ S_B = (\hat{\mu}_1 - \hat{\mu}_2)(\hat{\mu}_1 - \hat{\mu}_2)^T, \tag{6} $$

where $\hat{\mu}_i$ is the mean of the $N_i$ samples of class $w_i$ in the $n$-dimensional space. The $S_W$ matrix is the sum of scatter matrices for two classes and is given by

$$ S_W = S_{W1} + S_{W2}. \tag{7} $$

$W$ is consisted of the largest eigenvectors of $S_W^{-1}S_B$. For the two-class case, we have

$$ W \approx S_W^{-1}(\mu_0 - \mu_1). \tag{8} $$

Fisher’s measure or “$F$-ratio” defines the ratio between classes scattering to within class scattering of data points (Mika et al., 1999). Higher value of $F$-ratio for a group of samples from different classes implies more discrimination power among the involved classes. So $F$-ratio value could be a measure to find a properly pruned version of the original features. Since, for a $K$-class problem, $S_B$ is of rank $K-1$, the nonzero eigenvectors identify a vector subspace containing the variability between features and Fisher discriminant analysis involves $(K-1)$ discriminant functions. Considering $W_d \times (K-1)$, we have to update two notations we defined for two-class problem, $S_B$ and $S_W$. $S_W$ for multiclass problem is given by

$$ S_W = \sum_{i=1}^{K} S_{W_i}. \tag{9} $$

If we define $S_T$ as the sum within class and between class matrices as

$$ S_T = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T, \tag{10} $$

then $S_B$ could be obtained from

$$ S_B = S_T - S_W. \tag{11} $$

$F$-ratio is a good measure to compare extracted features with each other in an involved task. It is expected that a feature vector with a higher $F$-ratio acts better in an involved classification process. So, we compute $F$-ratio for the overall feature vector and use this measure for dimension reduction of this vector (Kadke, 1995). We have done a complementary test to confirm the results obtained by this measure. By deleting features which have the smaller $F$-ratio, we could reduce the size of the feature vector. We will employ this approach to reduce the size of the final feature set and verify its performance by conducting proper phoneme recognition experiments in Sec. VI B.

Using this approach, we selected a final 47-dimensional feature set used in the train and test stages of the speech classifier (Munkherjee et al., 1997).

VI. EXPERIMENTAL SETUP AND RESULTS

In this section, we evaluate the application of the RPS-based feature extraction method to speech classification. Experiments are done in two main categories: isolated phoneme recognition and continuous speech recognition.

In the remainder part of this section, we first introduce the used speech database and HMM-based speech classifier. The results of the isolated phoneme recognition experiments over TIMIT database are brought in the third part of this section (Packard et al., 1980). Next, the suggested combination feature set consisted of the parameters of the nonlinear modeling approach of speech signal and MFCC features are evaluated over a continuous speech recognition task using TIMIT corpus and hidden Markov model toolkit (HTK) toolbox. In this research, for both the isolated and continuous speech recognition experiments, the recognizer output is the recognized phone stream and the recognition rate is evaluated by measuring the related phoneme recognition rate (PRR). Moreover, in both isolated and continuous speech recognition tests, the sensitivity of different systems to noisy conditions will be evaluated using four types of noises taken from Noisex.92 database (Petry et al., 2002).

As speech recognition systems are typically degraded dramatically in noisy conditions, this is important to know the degradation degree of different approaches in such conditions. A good feature set should keep its performance in noisy conditions, too. So, in addition to clean speech experiments, we arrange extra experiments for noisy conditions using four different types of additive noises.

A. Speech database

TIMIT corpus (Garofolo et al., 1993) is the used speech database in the experiments. The sampling rate of this corpus signal is 16 kHz. TIMIT contains a total of 6300 utterances, ten sentences spoken by each of 630 speakers from eight major dialect regions of the United States. TIMIT speech signals have expertly labeled time-stamped phoneme boundaries, which can be used to extract desired isolated phoneme sets. TIMIT corpus is consisted of two distinct parts: a main core and a complete test set.

The complete test set consists of a total of 168 speakers and 1344 uttered sentences, accounting for about 27% of the total speech materials of the whole database. Inside the complete test set, a smaller core test set is assigned which consists of 24 speakers with 192 related sentences. In this work, the core test set is used as the evaluating test set in isolated phoneme recognition experiments. In continuous speech recognition experiments, the complete test set plays this role.
Phonemes in TIMIT test set have 48 different phonemic labels. By combining some of these labels, 39 remained labels are usually used in phoneme recognition tests as is typical with the literature.

To perform noisy condition tests, we exploited Noisex.92 database, which is a known noise database. Four types of noise used in our experiments are white, pink, bubble, and factory noises in different values of signal to noise ratio (SNR).

HTK (Cambridge Research Laboratory) toolbox is the baseline speech recognizer used in the experiments. HTK is an open source and famous speech classifier which uses HMMs. HMM models with six-state and eight Gaussian mixtures are used in all the experiments. Silent parts of speech signals are also modeled by a three-state HMM model. Speech signals are divided into 25.6 ms frames with 50% overlap and then are fed to HTK system.

### B. Feature selection experiments

As explained in Sec. IV, using GMM with four mixtures leads to a 60-dimensional RPS-based feature set. GMM modeling is done using the code provided in GMM toolbox for MATLAB by Kämäräinen and Paalanen. We selected four for GMM mixtures and seven-dimensional Poincaré section points are clustered via this approach.

To add the benefits of these features to typical MFCC features in speech recognition tasks, we combined them together. Adding 39 MFCC-based features to 60-dimensional GMM-based features leads to a total of 99-dimensional feature set. Table I shows the labels that we assigned to these features.

As discussed in Sec. V, we use Fisher discrimination analysis to reduce the size of the combined feature set. The algorithm of feature selection method was described in Sec. V. Using Fisher discrimination analysis, we first evaluate $F$-ratio for each of the features scheduled in the total 99-dimensional combined feature set. Figure 5 shows the individual Fisher values calculated for 30 features, in a left to right descending order. The features are sorted by their pruning effects in $F$-ratio values. In order to reduce the dimension of the final feature set, we start to delete features with lower $F$-ratio values from the feature set, one by one. In each step, after pruning one feature, we calculate phone recognition results using the reduced feature sets over TIMIT core test set. At last, features that had the minimum negative effects on the recognition results are permanently deleted from the total feature set. In this manner, 47 features are remained as the final feature set. The finalized feature set consists of the following:

1. MFCC features (1, 2, 3, 4, 5, 6, 7, 8, 9, 10),
2. Delta MFCC features (1, 2, 3, 4, 7, 9, 11, 12, 13),
3. Delta-delta MFCC features (1, 2, 3, 4, 5, 6, 8), and
4. GMM features (1, 2, 3, 4, 5, 6, 7, 11, 12, 15, 16, 17, 18, 20, 21, 27, 29, 30, 33, 34, 40),

which are 26 selected features from the frequency-based features and 21 selected features from the RPS-based GMM modeling parameters.

There is an interesting point about the features selected automatically from the RPS-based features by the proposed feature selection approach. The remained feature set consists of 16 mean, 2 weight, and 3 variance parameters of the developed GMM model. This is a very interesting point. As

<table>
<thead>
<tr>
<th>Description</th>
<th>Feature index</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC coefficients</td>
<td>1–13</td>
<td>MFCC1 to MFCC13</td>
</tr>
<tr>
<td>Delta coefficients</td>
<td>14–26</td>
<td>Delta1 to Delta13</td>
</tr>
<tr>
<td>Delta-delta coefficients</td>
<td>27–39</td>
<td>Delta-delta1 to Delta-delta13</td>
</tr>
<tr>
<td>Mean values of GMM mixtures</td>
<td>40–67</td>
<td>GMM1 to GMM28</td>
</tr>
<tr>
<td>Weight value of GMM mixtures</td>
<td>68–71</td>
<td>GMM29 to GMM32</td>
</tr>
<tr>
<td>Variances of GMM mixtures</td>
<td>72–99</td>
<td>GMM33 to GMM60</td>
</tr>
</tbody>
</table>

To add the benefits of these features to typical MFCC features in speech recognition tasks, we combined them together. Adding 39 MFCC-based features to 60-dimensional GMM-based features leads to a total of 99-dimensional feature set. Table I shows the labels that we assigned to these features.

As discussed in Sec. V, we use Fisher discrimination analysis to reduce the size of the combined feature set. The algorithm of feature selection method was described in Sec. V. Using Fisher discrimination analysis, we first evaluate $F$-ratio for each of the features scheduled in the total 99-dimensional combined feature set. Figure 5 shows the individual Fisher values calculated for 30 features, in a left to right descending order. The features are sorted by their pruning effects in $F$-ratio values. In order to reduce the dimension of the final feature set, we start to delete features with lower $F$-ratio values from the feature set, one by one. In each step, after pruning one feature, we calculate phone recognition results using the reduced feature sets over TIMIT core test set. At last, features that had the minimum negative effects on the recognition results are permanently deleted from the total feature set. In this manner, 47 features are remained as the final feature set. The finalized feature set consists of the following:

1. MFCC features (1, 2, 3, 4, 5, 6, 7, 8, 9, 10),
2. Delta MFCC features (1, 2, 3, 4, 7, 9, 11, 12, 13),
3. Delta-delta MFCC features (1, 2, 3, 4, 5, 6, 8), and
4. GMM features (1, 2, 3, 4, 5, 6, 7, 11, 12, 15, 16, 17, 18, 20, 21, 27, 29, 30, 33, 34, 40),

which are 26 selected features from the frequency-based features and 21 selected features from the RPS-based GMM modeling parameters.

There is an interesting point about the features selected automatically from the RPS-based features by the proposed feature selection approach. The remained feature set consists of 16 mean, 2 weight, and 3 variance parameters of the developed GMM model. This is a very interesting point. As

![FIG. 5. (Color online) Individual F-ratio values for 30 different features (sorted by their pruning effects in F-ratio values).](image-url)
could be seen, 16 mean parameters (the major part of the total 28 parameters) of four-mixture GMM remained after the pruning process. This confirms that the centers of the Gaussian mixtures, corresponding to the centers of the data point groups placed on the Poincaré section, are the most meaningful and important features of the RPS representation of speech signals. This was predicted earlier and is confirmed by the obtained results (Senevirathne et al., 1992).

### C. Isolated phoneme recognition tests

We introduced the proposed nonlinear modeling of Poincaré sections of reconstructed phase space attractor using GMM in Secs. III and IV. In this section, to evaluate the performance of this modeling scheme and select the best structure for ASR application, we perform some isolated phoneme recognition tests using 4200 utterances (isolated uttered phones) taken from TIMIT core test set. Phoneme correction rate (COR) is used as the accuracy measurement in isolated phoneme recognition tests. This measure could be evaluated by simply the ratio of the number of correctly recognized phones to the total number of the isolated phones and means as the accuracy of classification.

To evaluate the proposed approach in this work, we have to compare its performance with some typical known approaches. In addition to usual MFCC features, there are other feature extraction approaches developed in speech processing tasks. One of the famous ones is RASTA (RelAtive SpectrA), which has shown interesting results in the field of speech recognition (Hermansky and Morgan, 1994). This features were implemented using MATLAB code provided by the electrical engineering department of Columbia University. So, in the following, we will have speech recognition experiments with MFCC, RASTA, and other approaches which are especially discussed in this paper.

Table II shows the phoneme correction rates (COR%) obtained via the different interested approaches. As discussed earlier, we perform the recognition experiments using 39-dimensional frequency domain features (MFCC, delta, and delta-delta), RASTA features, the proposed 60-dimensional feature set extracted only from GMM model parameters, and the pruned 47-dimensional combined feature set that consists of 26 selected frequency domain features (MFCC-based features) and 21 selected RPS-based features.

As shown in Table II, the results obtained via the nonlinear modeling approaches alone are noticeable, but are completely far from the result of MFCC features. This could be the motivation of employing the combination of these features to have the benefits of both feature sets. The finally proposed combined 47-dimensional feature set has shown a COR improvement about 5.7% against MFCC features.

Table II shows that using RASTA features can improve recognition results about 1% against MFCC features. This is not noticeable in comparison with the COR improvement that resulted by the pruned version of the combined feature set.

Moreover, this is important to know the degradation degree of the performance of an ASR system which exploits a defined set of speech features, while some environmental noises are added to its input signals.

### D. Continuous speech recognition experiments

To show that the nonlinear modeling of reconstructed phase space is applicable to continuous speech recognition (CSR), and keeps its benefits in this field too, we conduct a complete set of CSR experiments using three of different feature sets involved in the earlier isolated phoneme recognition experiments (Skijarov and Bortnik, 2005).

In the CSR experiments, the recognition system consists of six-state monophone HMM models with eight mixtures in every state. All the experiments are done over the TIMIT complete core test set, which contains about 20 000 phones at all.

For CSR experiments, PRR is given by

\[
PRR = 1 - \frac{S + I + D}{N},
\]

where PER is the phone error rate, \(S\) is the number of substitutions, \(I\) is the number of insertions, \(D\) is the number of deletions, and \(N\) is the original number of phones. Somewhere in the literature, PRR is referred to as the accuracy rate (ACC%). Table III shows the results of the conducted CSR experiments.

To evaluate the sensitivity of the proposed features to additive noises, in ASR experiments, we compare the performances of MFCC features against the proposed features in the noisy conditions, too. Table IV presents the details of the conducted experiments in this field.

### VII. DISCUSSION

The current research verifies that the information extracted from the RPSs of speech signals is valuable for automatic speech recognition systems. Moreover, it shows that, in the speech recognition field, the features extracted by sta-
tistical modeling of Poincaré sections of speech RPS have more benefits against features obtained via directly modeling of RPS, as done by Povinelli et al. (2006). The CORs scheduled in the rows of 3 and 4 of Table III clearly verify this noticeable point.

This is an interesting result and verifies that Poincaré section is an appropriate tool for analysis of cyclic trajectories like speech signal. Usual applications of reconstructed phase space in signal processing tasks consist of extraction of some global metric features such as Lyapunov exponents and fractal dimension. These metrics are very sensitive to initial conditions and noise. We need enough samples of time series to reconstruct more reliable attractors, in order to extract reliable features from them. In signal processing tasks, we usually assume that a dynamical system which produces time series is stationary and such assumption lets us to use typical statistical analysis techniques. In speech signal, the stationary assumption is nearly valid only for time series with about 20–25 ms periods. In such periods, the time series length is about 400 samples for 16 000 Hz sampling rates. So, to reconstruct the phase space attractor from a 400-sample speech frame with eight embedding dimension, at least \(8+8 \times 8 = 72\) neighboring data points are necessary to build an appropriate approximation around that point, which is about 17% of trajectory points. So, we are mostly faced to very short time series and these kinds of extracted features are not too appropriate choices (Tishby, 1990).

To overcome the short length limitation, one interesting idea is to use nonlinear modeling of phase space attractor. Such an approach lets us to have more suitable representation from phase space information. In this research, we applied statistical modeling approach to Poincaré sections of phase space attractors. Poincaré sections could be useful especially for cyclic trajectories embedded in phase space. For speech signal, there are obvious cyclic patterns for voiced words that the nonlinear modeling could extract a proper representation from phase space trajectory with an appropriate interpolation of trajectory points (Tishby, 1990).

### TABLE IV. CSR results in details for three different feature sets.

<table>
<thead>
<tr>
<th>Signal</th>
<th>SNR</th>
<th>PRR% for MFCC features</th>
<th>PRR% for the proposed features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>inf</td>
<td>66.9</td>
<td>71.1</td>
</tr>
<tr>
<td>White</td>
<td>5</td>
<td>49.4</td>
<td>53.1</td>
</tr>
<tr>
<td>Pink</td>
<td>5</td>
<td>48.6</td>
<td>51.8</td>
</tr>
<tr>
<td>Factory</td>
<td>5</td>
<td>47.1</td>
<td>50.5</td>
</tr>
<tr>
<td>Bubble</td>
<td>5</td>
<td>48.3</td>
<td>51.7</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>57.8</td>
<td>61.6</td>
</tr>
</tbody>
</table>

### TABLE V. COR% for isolated phoneme recognition for different noisy conditions.

<table>
<thead>
<tr>
<th>Signal</th>
<th>SNR</th>
<th>COR% for the basic system</th>
<th>COR% for the proposed system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>inf</td>
<td>72.4</td>
<td>78.1</td>
</tr>
<tr>
<td>White noise</td>
<td>5</td>
<td>52.4</td>
<td>57.5</td>
</tr>
<tr>
<td>Pink noise</td>
<td>5</td>
<td>51.8</td>
<td>55.7</td>
</tr>
<tr>
<td>Factory noise</td>
<td>5</td>
<td>48.7</td>
<td>52.4</td>
</tr>
<tr>
<td>Bubble noise</td>
<td>5</td>
<td>49.4</td>
<td>53.1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>57.1</td>
<td>61.3</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>62.2</td>
<td>65.9</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>64.8</td>
<td>67.6</td>
</tr>
</tbody>
</table>

GMM, which is widely used in different signal processing tasks, could be observed as a statistical clustering tool that its assigned region boundaries are soft. In addition, to cluster a data set sample, the number of its mixtures need not be increased by the space dimension. In this research, the number of GMM mixtures is limited to 4. As described in Sec. IV, this leads to a 60-dimensional feature set. Putting more mixtures in the model increases the number of involved features and may decrease the performance of the used speech classifier, while increasing the classification complexity. This is not suitable for real time speech processing tasks (Rosenstein et al., 1992). However, in the finalized feature set, this limitation is compensated by adding some MFCC-based features. This combination is necessary because MFCC features are proved features for ASR applications, and the extracted features from the phase space alone are not adequate for ASR applications. They seem to be good auxiliary features which may complete some weak points of the strong traditional features, like MFCC. The final combinational feature set, as shown in Table V, improves the recognition rate of the ASR system significantly about 5% in the application of the isolated phone recognition (Vapnik et al., 1996).

Considering experimental results obtained for isolated and continuous speech recognition, it is obvious that the proposed nonlinear modeling approach has overcome the short time limitation of speech time series used in speech phase space reconstruction. We could explain this superiority with these words that the nonlinear modeling could extract a proper representation from phase space trajectory with an appropriate interpolation of trajectory points (Rapp et al., 2002).

Also, as could be found from the results scheduled in Tables V and IV, in the noisy conditions, where some noises are added to the speech signals, the proposed feature set keeps its performance perfectly (Sciamarella and Mindlin, 2001). As described in Sec. VI B, the major of the selected RPS-based features for the final pruned feature set are the mean coefficients of the GMM mixtures. These features have this property that their deviations by noise are minimal, because in noisy conditions the displacements of the centers of the Gaussian mixtures will be minimal. Moreover, the effect of noise in Poincaré section trajectories is less than its effect on direct RPS trajectories. So, it is reasonable that the proposed approach keeps its benefits in noisy conditions, too. As shown in Table IV, the best accuracy improvement is obtained for the clean speech data, and in the noisy conditions, the minimum improvement is obtained for bubble and fac-
tory noises and the best improvement for white noise. This subject can be studied furthermore using a real noisy database such as Aurora2.

VIII. CONCLUSION

In this research, we searched for an efficient combination of usual frequency features and the features obtained from the application of a nonlinear modeling approach to the Poincaré sections of reconstructed phase space attractors of speech signals. In this way, there were some applicable problems, for example, the shortness of the speech frames which must be processed as individual speech frames, and the curse of dimensionality of the resulted feature set. We showed that we could successfully solve these problems and benefit from the information which could be extracted directly from RPS trajectories of speech signals (Yamada and Saiki, 2007).

There is an important notice that RPS-based features alone, even with enough embedding dimension and appropriate modeling approach in reconstructed phase space, will not work as well as frequency domain analysis in speech recognition tasks. So, after applying Poincaré algorithm to speech RPS and modeling of the data distribution in the obtained subspace, we added the parameters of the developed models, as RPS-based features, to the typical MFCC-based features. The combinational feature set, following an extensive pruning process, showed good performances in different isolated and continuous phoneme recognition experiments. In a final CSR experiment, using HMM monophone models, the proposed feature set gained 4.1% absolute growth in phoneme recognition accuracy rate against only MFCC-based features. This is a very noticeable result, and verifies that RPS-based features could also help noticeably to CSR applications which are faced to very short time series (frames with about 25 ms lengths). Some experiments conducted on the speech signals corrupted by some additive noises showed that the proposed features keep their benefits in noisy conditions, too.


