Feature Extraction Based on Speech Attractors in the Reconstructed Phase Space for Automatic Speech Recognition Systems

Yasser Shekofteh and Farshad Almasganj

In this paper, a feature extraction (FE) method is proposed that is comparable to the traditional FE methods used in automatic speech recognition systems. Unlike the conventional spectral-based FE methods, the proposed method evaluates the similarities between an embedded speech signal and a set of predefined speech attractor models in the reconstructed phase space (RPS) domain. In the first step, a set of Gaussian mixture models is trained to represent the speech attractors in the RPS. Next, for a new input speech frame, a posterior-probability-based feature vector is evaluated, which represents the similarity between the embedded frame and the learned speech attractors. We conduct experiments for a speech recognition task utilizing a toolkit based on hidden Markov models, over FARSDAT, a well-known Persian speech corpus. Through the proposed FE method, we gain 3.11% absolute phoneme error rate improvement in comparison to the baseline system, which exploits the mel-frequency cepstral coefficient FE method.

Keywords: Reconstructed phase space, phoneme attractor, feature extraction, speech recognition.

I. Introduction

Traditional automatic speech recognition (ASR) systems, which convert human speech signals into corresponding text, are based on hidden Markov models (HMMs) to model the time varying nature of speech signals. Normally, speech signals are not used directly to model by HMM, so they are partitioned into a series of short quasi-stationary frames. Next, a set of feature vectors are extracted from the segments. The segmented frames are typically overlapped, for example, with a window size of 25 ms and a frame shift of 10 ms [1]. The extracted feature vector of each frame should be compact (small in dimension) and discriminatory. This means that the extracted feature vectors should contain all the information needed to distinguish speech units, for example, phonemes or subword units, and suppress the irrelevant information of speech signals [1], [2].

There are some popular feature extraction (FE) schemes, such as linear prediction coding (LPC), logarithm of filter bank energy (LFBE), mel-frequency cepstral coefficient (MFCC), and perceptual linear prediction (PLP) [3]. The LPC method is based on the AR modeling of speech signals. The second approach utilizes the energy of filter banks applied to the short-term spectrum of a speech signal, and the other two schemes are based on cepstral analysis. In all of the mentioned methods, the nonlinear property of a speech signal is not considered; moreover, in the spectrum-based methods (LFBE, MFCC, and PLP), the phase information of a speech signal is removed [4].

On the other hand, there is experiment evidence that proves the existence of nonlinear characteristics in speech production

Manuscript received Jan. 31, 2012; revised June 29, 2012; accepted July 9, 2012.

Yasser Shekofteh (phone: +98 21 6454 2372, y_shekofteh@aut.ac.ir) and Farshad Almasganj (almas@aut.ac.ir) are with the Biomedical Engineering Department, Amirkabir University of Technology, Tehran, Iran.

http://dx.doi.org/10.4218/etrij.13.0112.0074

systems (for example, turbulence of speech) not considered in the conventional and mentioned FE methods [5], [6]. One of the best domains to represent nonlinear and chaotic properties of a speech signal is the phase space domain. Whitney and Takens introduced the delay coordinate embedding theorem to embed a time series (for example, speech signal) in the phase space domain [7], [8]. This theorem shows that a one-dimensional signal can be embedded in a reconstructed phase space (RPS), a high-dimensional space. Because a recorded speech signal is a one-dimensional signal captured from a nonlinear and dynamical human speech production system, its true dynamic can be reconstructed in the RPS using embedding theory. The RPS can be topologically equivalent to the original system space if its parameters are chosen properly [9]. Therefore, the embedding theory is introduced to represent the actual dynamic and geometric structure of a one-variable time series.

In [10]-[13], some nonlinear dynamic features, such as fractal dimensions, Kolmogorov entropy, correlation dimension, Lyapunov exponents, and radial and scalar distances, were extracted from embedded speech signals to improve the performance of an ASR system. Moreover, a parametric modeling technique and a nonparametric modeling technique based on binning and occurrence counts were introduced to capture the attractor structures of the isolated speech phonemes that appeared in the RPS domain [14], [15]. In [16], a feature vector was suggested whose elements are a combination of the popular features of MFCC and some RPS-based features attained through parametric modeling of the Poincaré section in the RPS.

Most of the RPS-based feature vectors are used in some limited tasks, such as the isolated phoneme recognition; but, in this work, we capture an RPS-based feature vector that could be used in an ASR system, an FE method that benefits from a predefined set of phoneme attractors in the RPS. The proposed method could be considered as a development of Povinelli's method [15].

The rest of this paper is organized as follows. Section II introduces the embedding theorem and the RPS. Sections III and IV detail the requirement of phoneme attractors and the proposed FE method, respectively. In section V, the experimental setup is described. In section VI, our experiment results are introduced and discussed. Finally, we consider the results and present the conclusions in section VII.

II. Reconstructed Phase Space

One of the interesting topics in the dynamical system theory is the RPS method introduced and utilized by Takens and Sauer [9]. An RPS is a multidimensional space in which its coordinates are produced by shift-delay samples of a one-dimensional signal as a time series. The chaotic behavior of such a signal could be exhibited in the RPS.

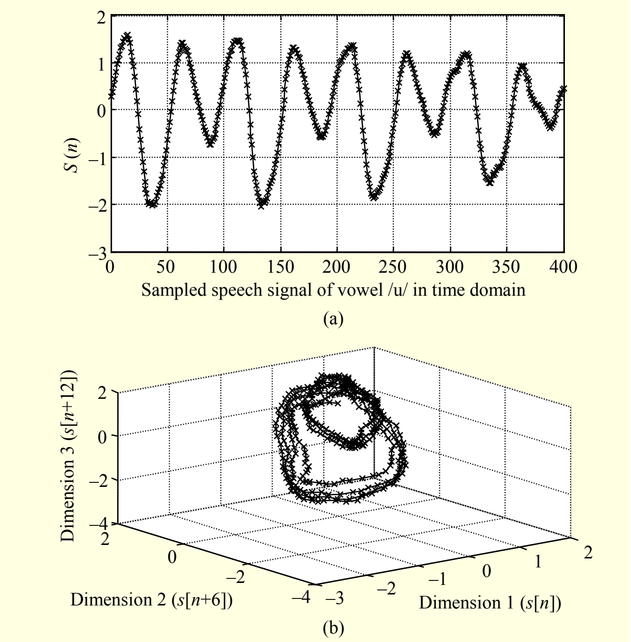

Fig. 1. Phase space reconstruction of speech signal frame (vowel phoneme /u/): (a) speech time series; (b) geometric structure of reconstructed trajectories in three-dimensional RPS (d = 3 and τ = 6).

The sequence of embedded points of a signal in the RPS is commonly referred to as signal trajectory. To construct a signal trajectory, its samples must be embedded in the RPS. If a single point of the embedded signal in the RPS is given by

$$S_l = [s_l \; s_{l+\tau} \; s_{l+2\tau} \; \cdots \; s_{l+(d-1)\tau}], \qquad (1)$$
$$\text{where} \quad s = \{s_1, s_2, s_3, \ldots, s_N\},$$

in which s_l is the l-th sample of an N-point segment of the original one-dimensional signal s, then d is the embedding dimension, and τ is the time lag. The concept of embedding dimension and time lag plays an important role in both the practical and theoretical aspects of the RPS [9], [14]. The minimum possible embedding dimension can be identified by some heuristic procedures, such as false nearest neighbor. Common techniques, including the first minimum of the auto-mutual information function or the first zero crossing of the autocorrelation function, are used to identify the preferred time lag of the RPS [9], [17], [18].

Figure 1 shows a speech frame of voiced phoneme /u/ and its embedded trajectory in a three-dimensional RPS (d=3). As seen in this figure, the behavior of a speech signal trajectory in the RPS, for all the voiced phonemes, creates attractors that are very similar to the process of squeezing in chaotic signals. These chaotic attractors are built up in the RPS by the endless repetition of the stretching and squeezing processes [9]. On the other hand, another chaotic behavior like the stretching behavior of chaotic signals is observed for the fricative and plosive phonemes (such as /b/ and /t/).

In this paper, we consider the consecutive samples of speech frames as a set of time series. So, the RPS representation S of a speech frame s in (1) can be represented by a trajectory matrix defined as

$$S=\begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_L \end{bmatrix}=\begin{bmatrix} s_1 & s_{1+\tau} & \cdots & s_{1+(d-1)\tau} \\ s_2 & s_{2+\tau} & \cdots & s_{2+(d-1)\tau} \\ \cdots & \cdots & \ddots & \cdots \\ s_L & s_{L+\tau} & \cdots & s_{L+(d-1)\tau} \end{bmatrix}, \tag{2}$$

where $L=N-(d-1)\tau$ is the number of embedded points in the RPS. The trajectory matrix is a mathematical representation of the RPS formed by compiling its row vectors from the vectors that are created by (1).

Since the elements of the trajectory matrix of (2) indicate the absolute positions of the embedded points in the RPS, to obtain dynamic information of signal trajectory [18], a flow matrix can be defined as

$$\Delta S=\begin{bmatrix} S_2-S_1 \\ S_3-S_2 \\ \vdots \\ S_L-S_{L-1} \end{bmatrix}, \tag{3}$$

where S_i is the i-th row of the trajectory matrix S. Several experiments have shown that the flow matrix can include useful information to discriminate different attractors from each other in the RPS [18]-[20].

III. Phoneme Attractor Models

The proposed FE method is based on the posterior probability evaluation of the embedded samples of a new given frame of speech in the RPS, considering the statistical models developed previously for the selected speech units embedded into the RPS. Therefore, first, we must define a set of reasonable speech units (such as phoneme, diphone, biphone, and so on). In this paper, we use phonemes as the speech units to be probabilistically modeled in the RPS. Similar to Povinelli's experiments, we utilize parametric distribution models based on Gaussian mixture model (GMM) distributions [14], [15]. Povinelli and others utilized the RPS embedding and GMM approaches to model phoneme units for an isolated phoneme recognition application using a Bayes maximum likelihood (ML) classifier. Meanwhile, based on the proposed method, the extracted features are the obtained posterior probability of the phoneme attractors embedded in the RPS, and the following continuous phoneme recognition task is done using a standard HMM.

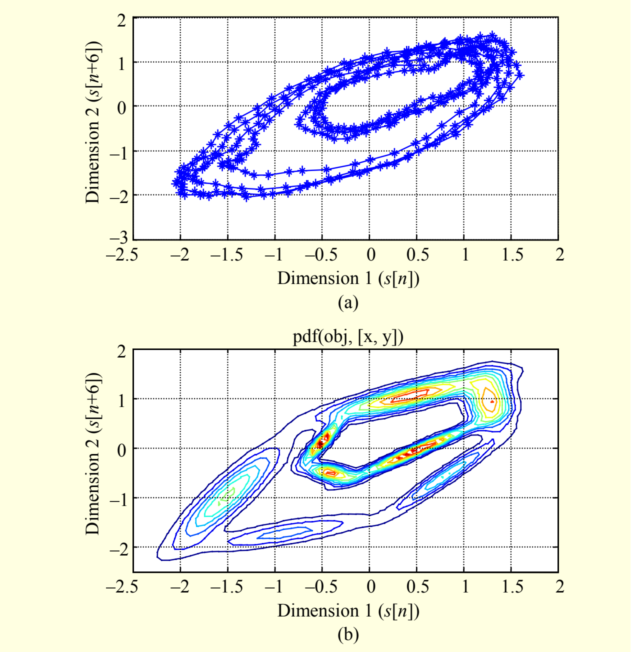

Fig. 2. Distribution modeling of embedded speech frame of Fig. 1 (vowel phoneme /u/): (a) geometric structure of its reconstructed trajectories in two-dimensional RPS; (b) GMM-based distribution modeling in two-dimensional RPS, using M=8 GMM components.

Figure 2 shows a phoneme distribution of a vowel phoneme /u/ in a 2-dimensional RPS. A visualization of its GMM modeling, using eight Gaussian components, is also depicted in this figure, in which every centered ellipse corresponds to one of the Gaussian components of the GMM model. Such RPS embedded trajectories can be considered the phoneme attractors, because they exhibit the behavior of speech trajectories in the RPS [21]. In this paper, to identify these attractors, the following procedure is proposed:

1) The isolated phoneme waveforms, taken from the training set, are normalized using mean subtraction and variance normalization methods to address amplitude variation across the phoneme instances.

2) Using embedding theory discussed in section II, the trajectory and flow matrices of each isolated phoneme waveform are computed with the commonly used parameters embedding dimension 8 and time lag 6 [16]. These matrices are then appended and used as the training RPS-based feature sets needed for the GMM modeling. Therefore, in this manner, a raw (2*d)-dimensional feature set is defined as

$$X=[S,\Delta S], \tag{4}$$

where S is the trajectory matrix, and ΔS is the flow matrix, defined in (2) and (3), respectively.

3) Using the prepared training data for each phoneme (to find its attractor), the parameters of the GMM specialized to model its attractor is estimated using the well-known iterative method of the expectation maximization algorithm. By iterating this phase for other phonemes one by one, we get a set of GMM probability distributions; each GMM defines a probability distribution for its assigned phoneme in the RPS, C_i, which could be interpreted as a phoneme attractor. Using phoneme attractor C_i, the probability of sample x inside the RPS is evaluated by

$$p(x \mid C_i) = \sum_{m=1}^{M} w_m N_m(x; \mu_m, \Sigma_m), \tag{5}$$

where M is the number of mixtures (Gaussian components), $N_m(x; \mu_m, \Sigma_m)$ is a normal distribution with mean vector μ_m and covariance matrix Σ_m, and w_m is the mixture weight that results in $\sum_{m=1}^{M} w_m = 1$. The optimum number of mixtures M is determined empirically in this task and considered in section VI.

IV. Overall Structure of Proposed Feature Extraction Method

In this section, we briefly introduce the proposed FE method based upon the dedicated GMM modeling of the subword attractors in the RPS. The subword attractors are considered the phoneme units of speech segments. In the proposed method, for an embedded speech signal X corresponding to an input speech frame (observed as an RPS-based vector), its posterior probability (conditional probability of phoneme attractor model given observed RPS-based vector X) is calculated for all of the previously modeled phoneme attractors. The computed posterior probability of some of the selected phoneme attractors in the RPS is then used as the final feature vector, that is, PPRPS, given by

$$F(s[n]) = [p(C_1 \mid X^n), \ldots, p(C_K \mid X^n)], \tag{6}$$

where K is the number of selected phoneme attractors, $s[n]$ is the n-th frame of the seech signal, and $F(s[n])$ is the proposed feature vector, PPRPS. In addition, X^n is the embedded data of the n-th frame in the RPS introduced in the previous section. The elements of PPRPS are evaluated via the following equation, using the likelihood function introduced in (5):

$$\begin{aligned} p(C_i \mid X^n) &= \frac{1}{L}\sum_{l=1}^{L} p(C_i \mid X_l^n), \\ p(C_i \mid X_l^n) &= \frac{p(X_l^n \mid C_i)\, p(C_i)}{p(X_l^n)}, \end{aligned} \tag{7}$$

where L is the number of the embedded points X^n of the n-th speech frame in the RPS and X_l^n is the l-th point of the embedded frame X^n. Here, we assume that the prior probability $p(C_i)$ of all of the considered phoneme attractors is equal.

One of the most important parameters of the proposed feature vector is the number of primarily selected attractors K. Finding attractors through which discriminative features can be properly extracted is an open problem in implementing the proposed method. The feature selection methods could be employed to solve this problem. In this work, first of all, each element of the obtained feature vector corresponds to one of the phoneme attractors. Next, to find the best final feature set, we utilize an isolated phoneme recognition-based method (similar to that in Povinelli's experiments [14]) to rank and select phoneme attractors that corresponded to phonemes with higher recognition accuracy rates. In this work, the parameter K is selected as 13, 26, or 30 to make possible an equitable comparison between the proposed feature vector and the extracted traditional speech feature vector.

Moreover, we employ the linear discriminative analysis (LDA) method to increase the discriminative ability of the extracted features. LDA is a supervised method used in pattern recognition and machine learning to find a linear combination of features that separates two or more classes of data [1], [2]. This linear transformation method attempts to maximize the linear separation between the classes of data. The new resulting combination is commonly used as the final features, which are decorrelated and are ready to feed the final classifier. The LDA transformation is made, as given by

$$F_{\text{LDA}}(s[n]) = W_{\text{LDA}}^{T} \cdot F(s[n]), \tag{8}$$

where W_{LDA} is estimated from an eigendecomposition approach [1], [2] and the transpose operator is denoted by a superscript T.

V. Database and ASR Setup

In this paper, a well-known Persian speech corpus, FARSDAT, is used to conduct the experiments [22]. FARSDAT includes a variety of Persian speech data collected from 304 adult speakers differing in age, gender, dialect, and educational level. Each participant spoke 20 sentences in two sessions. These speech samples were manually segmented and labelled for each phoneme and word.

Speech data of the first 250 speakers was assigned to a training stage, and the rest of the data was assigned to a test stage. A set of 44 phonetic labels was employed in the labelling stage of the corpus. In this work, the labels are reduced to 30 classes by merging some of them.

A state-of-the-art continuous phoneme speech recognition (CPSR) system, utilizing HTK functions, is used to conduct the main experiments [23]. The following points are considered to implement the context-dependent HMM recognizer using the HTK toolkit.

- Bigram phoneme language model
- Well-trained HMM triphone models initially produced from 30 initial monophone models, similar to previous work [18], [24]
- Decision tree used to tie single Gaussians of states to overcome lack of training data for some triphone models
- Number of Gaussian mixtures of HMMs finally increased to eight

The training procedure leads to a collection of continuous density HMM triphone models. The results are obtained using gender independent models in a single-pass decoding framework, without speaker adaptation. The recognition results of CPSR are introduced by the phoneme error rate (PER) measure, as given by

$$PER = \frac{I + D + S}{N} \times 100\,(\%), \qquad (9)$$

where N is the total number of phones in the reference phonetic transcription, and I, D, and S are the number of insertions, deletions, and substitutions of decoded phones, respectively.

In this paper, feature extraction real-time factor (FE-RTF) is a time factor used as a speed measuring metric; we define this parameter as the ratio of the FE time to duration of the processed speech file. Therefore, lower values of FE-RTF indicate better runtimes for the FE phase of the whole speech recognition process.

VI. Experiment Results and Discussion

This section details the experiment results obtained from the proposed RPS-based FE method. As mentioned in section IV, to implement the proposed method, the most effective features (assigned to the posterior probability of some phoneme attractors) must be selected in the first step. First, to find the best features, we implement an isolated phoneme recognition test, a simpler task compared to the final experiments with an ASR system, similar to Povinelli's work [14], to select the proper phoneme attractors.

1. Isolated Phoneme Recognition Experiments

Using the learned and evaluated GMM models for all the phoneme attractors, we construct the preliminary PPRPS (K=30) needed for the isolated phoneme recognition, according to (6). The implemented evaluation of isolated phoneme recognition is done utilizing a multiclass support vector machine (SVM) to consider the discriminative ability of different compositions of the parameters stated in the proposed feature vector. To train SVM classifiers, the LIBSVM (Library for SVM) toolbox is utilized with different kernel functions, such as linear, polynomials, and radial basis function (RBF) kernels [25]. The main idea of the used kernel function is to map the original feature space into a high-dimensional space, in which the features may be linearly separated. Using LIBSVM, the appropriate parameter C, the regularization term in the Lagrange formulation, and the RBF (Gaussian) kernel parameter are found by searching over a grid space, along with fivefold cross validation. For multiclass SVM classification, LIBSVM implements the "one-against-one" approach [26]. If K is the number of classes, $K(K–1)/2$ classifiers are constructed and each one is trained with the data of two different classes (supervised). In the classification stage, a voting strategy is then utilized.

Table 1. ACC of isolated phoneme recognition, using proposed feature vector with K=30 (29 Persian phoneme attractors + silence attractor).

Feature vector	SVM-kernel	*K*	ACC (%)
Povinelli's method [14]	-	-	50.12
PPRPS	Linear	30	60,24
PPRPS	**Pol. *D*=2**	**30**	**63.59**
PPRPS	Pol. *D*=3	30	61.48
PPRPS	RBF	30	62.76

Table 1 shows the obtained results of the preliminary experiments (isolated phoneme recognition). In this table, phoneme classification accuracy rate (ACC) via the proposed feature vector is shown using different types of kernel functions in the SVM classifier. The ACC is commonly used as a classification accuracy measurement in isolated phoneme recognition tasks. It can be computed by the ratio of the number of correctly recognized phones to the total number of isolated reference phones.

In addition, a classification accuracy rate based on the isolated phoneme recognition of Povinelli's method [14] (classification accomplished through the Bayes ML method) is scheduled in the first row of Table 1. In the first experiment, we utilize M=128 and K=30. Assuming there are 30 phoneme classes (number of phoneme labels, including silence), K=30 is equal to utilizing all the possible Persian phoneme attractors in the RPS.

As shown in Table 1, the results obtained via the proposed PPRPS are noticeable. The best classification accuracy rate of the proposed method is 63.59%, which shows superiority over Povinelli's method by a 13.47% absolute rate. This superiority is gained using PPRPS and the SVM classifier with a second-order polynomial kernel. This shows that a nonlinear transformation over the features extracted via the proposed RPS-based method could help the classification process.

In Povinelli's method, the classification is done based on a Bayes ML schema, directly using the involved GMMs. However, in the proposed method, some features are first evaluated via the posterior probability of each phoneme attractor in (6), and the SVM classifier is then applied to them. In the proposed method, we improve feature observation utilizing the obtained information of phoneme attractors in the RPS. This is not the case in Povinelli's method. So, using the PPRPS method, RPS-based information is extracted as a proper feature vector for speech frames, to be properly exploited in an ASR task.

2. Selection of Phoneme Attractors

In the next step, some of the phoneme attractors must be selected to be employed in the further continuous phoneme recognition application. This is done according to the ACC of different phonemes. Heuristically, we choose those attractors belonging to the well-recognized phonemes for this purpose. Therefore, the features that correspond to them are kept in the final feature vector (PPRPS) and the others are removed. In Table 2, the classification accuracy of 29 Persian phonemes, for the isolated phoneme recognition task obtained via Povinelli's method and the proposed FE method (PPRPS), are scheduled. The phonemes are ranked through the ACC of the PPRPS experiment. Table 2 shows the superiority of PPRPS features over Povinelli's framework in more detail. In the proposed method, the members of the final feature vector (for example, in (6)) are selected from the phonemes that have higher accuracy, as shown in Table 2.

Potentially, one of the most challenging aspects in implementing the proposed method is its computational cost. If we assume $O(g)$ as the required computational cost to calculate the likelihood function of a single Gaussian model, the proposed PPRPS method needs about $L*K*M*O(g)$ computations to extract the features assigned to one frame of a speech signal where L is the number of the embedded points in the RPS, which generally cannot be reduced; therefore, the chosen values of K (number of the selected GMMs) and M (number of mixtures for GMM) must be small whenever possible.

In much of the reported work on ASR, the length of the used feature vector (in many cases, MFCC and its derivations) in different cases was selected as 13, 26, or 39 to make it possible to have an equitable comparison of the proposed method with a typical baseline system. In each of the experiments conducted to evaluate our proposed method, we choose and examine the value 13 or 26 for the parameter K. So, based on the classification rate scheduled in Table 2, the first K phonemes and their assigned attractor models are selected to be used in our next experiments.

Table 2. ACC of 29 phoneme attractors in isolated phoneme recognition task for Povinelli's method (Pov.) and PPRPS-2 FE approach (silence is not considered as attractor in this table).

Rank	Ph.	ACC (%)		Rank	Ph.	ACC (%)	
		Pov.	PPRPS			Pov.	PPRPS
1	a	71.37	85.76	16	t	35.39	53.24
2	s	71.18	82.59	17	u	61.81	51.87
3	sh	64.35	82.18	18	n	25.69	51.59
4	aa	68.07	77.82	19	o	52.82	50.91
5	kh	58.12	70.02	20	h	42.68	49.69
6	e	41.80	68.01	21	zh	43.01	48.00
7	b	56.11	67.33	22	j	46.80	44.33
8	z	54.12	67.18	23	v	56.23	43.49
9	i	44.62	64.33	24	f	43.31	42.04
10	d	31.42	63.99	25	p	39.44	41.98
11	k	23.33	58.53	26	l	26.74	25.97
12	m	42.30	57.40	27	q	28.79	23.84
13	gs	50.16	57.17	28	g	13.57	16.58
14	ch	76.74	56.48	29	y	27.23	12.10
15	r	35.94	55.06	-	**ALL**	**50.12**	**63.59**

3. Continuous Phoneme Recognition Experiments

Table 3 shows the results obtained via the proposed PPRPS FE method by a continuous phoneme speech recognition system, for different values of the number of Gaussian components (M) of the employed GMM model. All the experiments are conducted using 26 phoneme attractors (K=26), selected as previously mentioned. FE-RTF values representing different cases are obtained based on the runtime of the FE phase of the experiments, implemented on a computer equipped with an Intel 3.07 GHz processor and 4 GB of RAM. The programs are developed and run in a MATLAB software domain. Using C-based language programming and hardware equipped with multicore processors could reduce the FE runtime significantly.

By investigating the results scheduled in Table 3, we find that the best result (the lowest PER) is 46.53%, obtained for M=16. Of course, the choice of M=4 is a desirable choice because the complexity of the FE process is reduced about four times,

Table 3. PER and FE-RTF results of continuous recognition experiment, using proposed feature vectors with different number of mixtures (*M*).

Feature vector	*M*	*K*	PER (%)	FE-RTF
PPRPS	128	26	50.93	46.63
PPRPS	**16**	**26**	**46.53**	**5.82**
PPRPS	8	26	47.40	2.94
PPRPS	4	26	48.41	1.46

Table 4. Results obtained for continuous phoneme recognition, using MFCC and proposed feature vector (GMM of phoneme attractors with *M*=4 mixture with/without LDA transformation, for different values of *K*) in different feature vector dimensions (*Dim*).

Feature vector	*K*	Dim.	PER (%)	FE-RTF
MFCC (baseline)	-	13	44.62	0.08
MFCC + LDA13	-	13	43.79	-
PPRPS	13	13	49.55	0.74
PPRPS	26	26	48.41	1.46
PPRPS+LDA26	**26**	**26**	**41.51**	**1.46**
PPRPS+LDA13	26	13	43.31	-
PPRPS+MFCC+LDA26	**26**	**26**	**38.57**	-

while the PER increases only about 1.9%.

As discussed in section IV, we could modify the primarily extracted feature vector, utilizing LDA transformation, as given in (8). In the following, we employ the proposed PPRPS FE method with *M*=4 because of its lower computational cost. To evaluate the performance of the final proposed features, we compare it with the MFCC features, which are the most frequently used in the current ASR systems [1]. The MFCCs are features evaluated through the magnitude-only spectrums calculated from short-time frames of a speech signal. In our experiments, we use the first 12 cepstral coefficients of the MFCC features combined with the zero-order cepstral coefficient, which together lead to a 13-dimensional feature vector. To extract the MFCC feature vector, a filter bank including 20 triangular filters is [0]selected[0]. The obtained results are shown in Table 4.

As shown in Table 4, the baseline system has a phoneme error rate of 44.62% and FE-RTF of 0.08. The best PER result is obtained by the simultaneous usage of the PPRPS and LDA, without any dimensionality reduction. In this case, the PER of 41.51% and FE-RTF of 1.46 are yielded. Other experiments including different values of *K* and dimensions of PPRPS features are shown in Table 4. Moreover, the recognition result of a combination of different MFCC-based and RPS-based acoustic features is given in the last row of Table 4. This combination of acoustic features is carried out directly on the level of feature vectors. Thirteen MFCC features are directly concatenated with 26 PPRPS features, and the resulting 39-dimensional feature is reduced to a 26-dimensional vector via the LDA. In this manner, a significant reduction in PER is achieved, 2.94% and 6.05% against individually using the best PPRPS and MFCC feature vectors, respectively.

To analyze the benefits of the proposed features, we could claim that the proposed method is particularly suited in differentiating between signals wherein their phases have important discriminative information. Of course, the phase information can be captured by the RPS. Therefore, the theory of time delay embedding is addressed by transforming a speech signal into the RPS, which has a mathematical correspondence with the true dynamic of the underlying system. The RPS is a time-invariant domain, so it can represent the trajectory of embedded signals or access the state structure of the systems. Based on the existence of geometric structures underlying the transmission of embedded speech points in the RPS (speech trajectory), some machine learning approaches can fit statistical distributions (for example, GMM) to model geometric structures of speech attractors. So, GMM modeling can be used for a continuous parametric model to analytically determine closed-form conditional distributions for each speech phoneme class and derive a specific phoneme attractor. In this manner, the posterior probabilities obtained from different attractors in the RPS domain capture extra information, among which might be phase information.

The obtained results confirm that the extraction of proper features from signals embedded in the RPS gives our proposed method a significant discriminatory ability, enabling it to outperform the conventional MFCC method in continuous phoneme recognition applications. Additionally, the experiment results show the superiority of the proposed PPRPS-based features over Povinelli's RPS-based method in the isolated speech recognition framework and verify the fact that the proposed features can be considered effective features to employ in continuous speech recognition systems.

In this paper, we compare the results of only using static features via the proposed features with the results of only using static features via MFCC-based ones. Based on our extended experiments, the PER of a baseline ASR system including 13 MFCC + 13 delta + 13 delta-delta (39 Dim. feature vector) is 23%, whereas the PER achieved using the proposed feature vector and its delta features is 31%. Therefore, we may conclude that the usual and simple method of evaluating delta features [23], which is suitable for spectral-based or cepstral-based features (for example, LFBE, PLP, and MFCC), is not a suitable approach to extract dynamic information from the posterior-based features, such as the proposed PPRPS method. On the other hand, using a neural network framework, similar to the TANDEM framework [27], on some consecutive speech frames is expected to solve this problem, making dynamic information as valuable as the delta features of MFCC. The verification of this idea could be addressed in future works.

VII. Conclusion

In this work, we proposed an FE method, considering nonlinear characteristics of a speech signal. First, we defined a set of speech units as phoneme attractors in the phase space, constructed by the embedding theory. To identify the proper phoneme attractors, phonemes isolated through training were embedded in the RPS, utilizing Taken's theory. Using the obtained trajectory and flow matrices, some GMMs were learned over the phoneme attractors reconstructed in the RPS. Finally, to extract the feature vector for a test signal, after embedding it in the RPS, the posterior probability of their attractors was evaluated, considering the previously trained GMMs. Next, a linear LDA transformation was applied to the features. These features were fed to an HMM classifier to implement a continuous phoneme recognition task and showed superiority over the MFCC features, by a 3.11% absolute reduction of the PER. Combining the MFCC and PPRPS features increased this reduction to 6.05%.

References

[1] X. Liu, *Discriminative Complexity Control and Linear Projections for Large Vocabulary Speech Recognition*, doctoral dissertation, Cambridge University Engineering Department, Cambridge, England, UK, 2005.

[2] Y. Tang and R. Rose, "A Study of Using Locality Preserving Projections for Feature Extraction in Speech Recognition," *Proc. ICASSP*, 2008, pp. 1569-1572.

[3] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *J. Acoustical Soc. America*, vol. 87, no. 4, 1990, pp. 1738-1752.

[4] A. Errity, J. McKenna, and B. Kirkpatrick, "Dimensionality Reduction Methods Applied to Both Magnitude and Phase Derived Features," *Proc. Interspeech*, 2007, pp. 1957-1960.

[5] I. Kokkinos and P. Maragos, "Nonlinear Speech Analysis Using Models for Chaotic Systems," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, 2005, pp. 1098-1109.

[6] J.J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in Voice, from Modeling to Measurement," *J. Voice*, vol. 20, 2006, pp. 2-17.

[7] H. Whitney, "Differentiable Manifolds," *Annals Math.*, 2nd series, vol. 37, 1936, pp. 645-680.

[8] F. Takens, "Detecting Strange Attractors in Turbulence," *Proc. Dynamical Syst. Turbulence*, 1980, pp. 366-381.

[9] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge, England, UK: Cambridge University Press, 1997.

[10] A. Ezeiza et al., "Combining Mel Frequency Cepstral Coefficients and Fractal Dimensions for Automatic Speech Recognition," *Proc. NOLISP*, 2011, pp. 183-189.

[11] V. Pitsikalis, I. Kokkinos, and P. Maragos, "Nonlinear Analysis of Speech Signals: Generalized Dimensions and Lyapunov Exponents," *Proc. Eurospeech*, 2003.

[12] S. Prasad et al., "Nonlinear Dynamical Invariants for Speech Recognition," *Proc. Int. Conf. Spoken Language Process.*, 2006, pp. 2518-2521.

[13] S. Yu, D. Zheng, and X. Feng, "A New Time-Domain Feature Parameter for Phoneme Classification," *Proc. WESPAC IX*, 2006.

[14] M.T. Johnson et al., "Time-Domain Isolated Phoneme Classification Using Reconstructed Phase Spaces," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, 2005, pp. 458-466.

[15] R.J. Povinelli et al., "Statistical Models of Reconstructed Phase Spaces for Signal Classification," *IEEE Trans. Signal Process.*, vol. 54, no. 6, 2006, pp. 2178-2186.

[16] A. Jafari, F. Almasganj, and M. NabiBidhendi, "Statistical Modeling of Speech Poincaré Sections in Combination of Frequency Analysis to Improve Speech Recognition Performance," *Chaos*, vol. 20, 2010, pp. 033106:1-11.

[17] J. Sun, N. Zheng, and X. Wang, "Enhancement of Chinese Speech Based on Nonlinear Dynamics," *Signal Process.*, vol. 87, no. 1, 2007, pp. 2431-2445.

[18] Y. Shekofteh and F. Almasganj, "Using Phase Space Based Processing to Extract Proper Features for ASR Systems," *Proc. 5th Int. Symp. Telecommun.*, 2010, pp. 596-599.

[19] A.C. Lindgren, M.T. Johnson, and R.J. Povinelli, "Speech Recognition Using Reconstructed Phase Space Features," *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2003, pp. 61-63.

[20] A.C. Lindgren, M.T. Johnson, and R.J. Povinelli, "Joint Frequency Domain and Reconstructed Phase Space Features for Speech Recognition," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2004, pp. 533-536.

[21] J. Ye, M.T. Johnson, and R.J. Povinelli, "Phoneme Classification over Reconstructed Phase Space Using Principal Component Analysis," *Proc. NOLISP*, 2003, pp. 11-16.

[22] FARSDAT (Farsi Speech Database). Available: http://catalog.elra.info/product_info.php?products_id=18

[23] S. Young et al., *The HTK Book*, Version 3.4, Cambridge University Engineering Department, Cambridge, England, UK, 2006. Available: http://htk.eng.cam.ac.uk

[24] Y. Shekofteh, F. Almasganj, and M.M. Goodarzi, "Comparison of Linear Based Feature Transformations to Improve Speech Recognition Performance," *Proc. ICEE*, 2011, pp. 1-4.

[25] C.C. Chang and C.J. Lin, "LIBSVM: A Library for Support

Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, article 27.

[26] C.W. Hsu and C.J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, 2002, pp. 415-425.

[27] F. Grezl and M. Karafiat, "Integrating Recent MLP Feature Extraction Techniques into TRAP Architecture," *Proc. Interspeech*, 2011, pp. 1229-1232.

Yasser Shekofteh received his BS in biomedical engineering and electrical engineering from Amirkabir University of Technology, Tehran, Iran, in 2005 and 2006, respectively. He received his MS in biomedical engineering from Amirkabir University of Technology in 2008. He is currently a PhD candidate in the Biomedical Engineering Department at Amirkabir University of Technology. His research is mainly focused on nonlinear and chaotic signal analysis, speech recognition, keyword spotting, and pathological speech signal processing.

Farshad Almasganj received his MS in electrical engineering from Amirkabir University of Technology, Tehran, Iran, in 1987 and his PhD in biomedical engineering from Tarbiat Modares University, Tehran, Iran, in 1998. He is currently an associate professor in the Biomedical Engineering Department of Amirkabir University of Technology. His research interests include automatic detection of voice disorders and signal processing, speech recognition, prosody, and language modeling for ASR systems.