Computers and Electrical Engineering 000 (2016) 1–12

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals

Shabnam Gholamdokht Firooz*, Farshad Almasganj, Yasser Shekofteh

Biomedical Engineering Department, Amirkabir University of Technology, Hafez Ave., P.O. Box 15875-4413, Tehran, Iran

ARTICLE INFO

Article history:
Received 9 September 2015
Revised 6 July 2016
Accepted 6 July 2016
Available online xxx

Keywords:
Automatic speech recognition
Mel-frequency cepstral coefficients
Reconstructed phase space
Recurrence plot
Two-dimensional wavelet transform

ABSTRACT

The spectral-based features, typically used in Automatic Speech Recognition (ASR) systems, reject the phase information of speech signals. Thus, employing extra features, in which the phase of the signal is not rejected, may fill this gap. Embedding the speech signal in the Reconstructed Phase Space (RPS) and then extracting some useful features from it, is a recently considered approach in this field. In this paper, we will follow this approach by evaluating some useful features from the Recurrence Plot (RP) of the embedded speech signals in the RPS; the proposed features are evaluated via applying a two-dimensional wavelet transform to the resulted RP diagrams. The proposed features are examined in an ASR task alone and in combination with the traditional Mel-Frequency Cepstral Coefficients (MFCC). For the second case, using English TIMIT corpus, 3.94% absolute classification accuracy improvement in the phoneme recognition accuracy rate, against using only the MFCC features is gained.

1. Introduction

In recent decades, a variety of linear models for speech coding, synthesis, and recognition with acceptable performances have been introduced. In this way, many types of research achieved improvement in the field of speech recognition by employing novel methods [1,2] or the detection of mispronunciation using Hidden Markov model [3]; however, there are nonlinear aerodynamic phenomena in the human speech production system which generally could not be included in linear models [4]. Therefore, nonlinear methods could potentially provide effective computational models to extract acoustic features which are useful for the nonlinear phenomena detection [4]. Furthermore, some recent studies have shown that utilizing nonlinear characteristics may improve the performances of the ASR systems [5].

Usual ASR systems exploit frequency domain features like Mel frequency cepstral coefficients [6]. The traditional frequency domain methods typically extract only the first and second order properties from the spectral patterns of speech signals [7]. However, there are many signals produced via nonlinear differential equations that have wide spectral characteristics [8]. In such cases, the frequency domain techniques are deficient, because it is impossible to dissociate the information of such a signal only in the frequency domain [5].

In addition, the speech signal shows some chaotic behaviors due to the existence of nonlinear phenomena such as the turbulence [9]. Studies about dynamical systems and chaos theory resulted in a kind of signal representation, a multi-dimensional trajectory embedded in the reconstructed phase space [4–6]. Following this approach, some useful features

* Corresponding author. Fax: +982166495655.
E-mail address: sh.firooz@aut.ac.ir (S.G. Firooz).

http://dx.doi.org/10.1016/j.compeleceng.2016.07.006
0045-7906/

were introduced, such as Lyapunov exponents (LE) and the Fractal Dimensions (FD) which investigate chaotic and nonlinear dynamical properties of the signals. Lyapunov exponents of the signals mapped to the phase space are evaluated to characterize a multi-dimensional chaotic time series [10]. In [11], the Fractal dimension of the speech signal and the MFCC features are simultaneously utilized in the ASR system. By adding these nonlinear features to the entire process, the experiment conducted on the Broadcast News ASR system in Spanish has shown 1.36% correct word rate (CWR) improvement against the baseline system which uses the MFCC features alone [11].

Recently, some non-classic feature extraction algorithms are directly developed over the multi-dimensional RPS transformation of the speech signal. The embedding process is performed using a set of input-output pairs reconstructed via the time delayed based approach. By selecting a sufficient dimension for the reconstructed space, the underlying theory guarantees that the dynamics of signals are fully accommodated in the attractors constructed by embedding them into the RPS. To extract beneficial parameters from the transformed signals, many studies in this field concentrated on the assessment of dynamical invariants [7,10] which do not rely on the initial conditions [4]. Some recent research have utilized statistical distributions such as the Gaussian Mixture Model (GMM) and applied it to signal trajectories appeared in the RPS [5-8]. In [8], the Poincaré section is employed as an effective tool to analyze the trajectories generated in the RPS, and a statistical modeling approach based on the GMM is applied to the Poincaré sections of the speech attractors to extract parameters which could help the ASR performance. Moreover, Ref. [9] applied a multi-dimensional linear method to model the speech trajectory reconstructed inside the RPS, using the Multivariate Autoregressive (MVAR) method, to improve the performance of the involved ASR system. In [12], a set of the Gaussian Mixture Models (GMMs) were trained over the phoneme attractors in the RPS, via which a proper feature vector could be evaluated; the posterior probabilities of different phonemes are then estimated by an MLP-based classifier. By applying this approach, 1.89% absolute accuracy rate improvement (for FARSDAT corpus) was gained, compared to the baseline system which used only the MFCC features.

As aforementioned, the reconstructed trajectories in the phase space could carry nonlinear dynamics of the involved systems; however, these high-dimensional trajectories could not be directly visualized. As the recurrence property is an essential feature of dynamical systems, it could be employed to investigate the system's behavior in the RPS domain. To follow this approach, the Recurrence Plot (RP) as a useful tool may be exploited to analyze the speech trajectories in the RPS [13]. In fact, the RP computes a binary square matrix to represent some main specifications of the phase space trajectories [13].

In this paper, a two-step approach is introduced: first, the speech signal is embedded in the phase space; next, it is parameterized using the recurrence property analysis of dynamical systems. In the final step, the RPS-based evaluated features resulted from the first step are combined with the common MFCC features to improve the extended continuous speech recognition (CSR) task.

1.1. Contribution

Investigating new efficient feature extraction methods to be applied to speech signals is an essential topic in the field of ASR tasks; the ultimate viewpoint of these efforts is basically to improve the performance of the designed and implemented systems. In this paper, we intend to follow and verify that the reconstructed phase space is a proper tool for capturing the signal nonlinear dynamics and compensating the phase information lack which occurs for common popular spectral features like the MFCC; however, the high computational demand of these features and the low accuracy condition that appears while evaluating some of them, makes this approach difficult or somewhere impractical for many real-time applications. To overcome these limitations existed for some previous proposed RPS-based techniques [12], this paper proposes an effective algorithm to extract proper auxiliary features from speech trajectories reconstructed in the phase space. The proposed method benefits from the feature extraction methods introduced in the speech and image fields, simultaneously. In this approach, the one-dimensional speech signal is converted into a two-dimensional image, as is dubbed the RP; the accuracy improvement of the ASR system is obtained by combining the features extracted from the RP-based images of the phone acoustic signals and the traditional MFCC spectral features promised for speech recognition purposes. Moreover, by conducting proper experiments, it is shown that the proposed features keep their performance in noisy conditions too; this is not the case for the features directly evaluated from the RPS domain, for they are typically sensitive to the initial conditions and the environmental noise [8].

The rest of this paper is organized as follows. Section 2 describes the embedding procedure that enables us to reconstruct the trajectory of a speech signal in the phase space. This section also gives a brief synopsis of the recurrence plot and the way it demonstrates some aspects of the dynamics of speech signals [13]. In the following, the proposed feature extraction methods applied to the RP images will be described in detail. Issues of dimensionality reduction based on the Linear Discriminant Analysis (LDA) and the Forward feature selection are also discussed in Section 2. The experimental methodology and data description are described in Section 3. The experimental results and discussion are presented in Section 4. The paper is concluded in Section 5.

2. Material and methods

As mentioned earlier, based on the RPS and RP theories, speech frames are converted into the RP-based images. Image processing methods are then used to extract proper features from the RP images. Since the wavelet transform has been used

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

extensively in the field of image processing, the discrete wavelet transform (DWT) approach is applied to the RP images. These features are then combined with Mel-Frequency Cepstral Coefficients (MFCCs) to improve speech recognition process. Given that the dimension of the wavelet-based features is high, it must be reduced to proper values to be suitable to be fed to the final classifier. Hence, two-dimension reduction methods, the forward feature selection and the linear discriminant analysis (LDA) are examined for this purpose.

2.1. The reconstructed phase space

The theoretical basis of the signal embedding in the phase space could be found in Whitny, Takens and Sauer works [12]. According to embedding theory, a dynamical system could be fully described by its output signals transformed as trajectories in the RPS [10]. By embedding a signal into a phase space with a sufficient dimension, the generated attractor topologically keeps the dynamics of the original signal [4–6]; via this approach, the nonlinearities of a signal could be considered in its related trajectory transformed in the RPS [7]. Suppose a time series of $X = \{x_1, x_2, \ldots, x_n\}$; its transformation matrix with dimension " d" and time lag "τ" is given by

$$X = \begin{bmatrix} x_{1+(d-1)\tau} \\ x_{2+(d-1)\tau} \\ . \\ . \\ x_N \end{bmatrix} = \begin{bmatrix} x_{1+(d-1)\tau} & . & x_{1+\tau} & x_1 \\ x_{2+(d-1)\tau} & . & x_{2+\tau} & x_2 \\ . & . & . & . \\ x_N & . & x_{N-(d-2)\tau} & x_{N-(d-1)\tau} \end{bmatrix} \tag{1}$$

Narayanan and Alwan [14] suggested the mutual information measure for estimation of the proper time lag between speech samples. It is shown that the value of 6 is a proper selection for this parameter in the speech signals with 6 kHz sampling rate [8]. Next, the minimum possible embedding dimension is calculated via an investigative method, like the false neighbors [15]. Kokkinos and Maragos [4] used false neighbors measure to determine the embedding dimension. An appropriate value for the embedding dimension of speech signals is 8 [8].

2.2. The recurrence plot

The recurrence property is a basic attribute of dynamical systems utilized to describe the system's behavior in the RPS [11]. In 1987, Eckmann [16] introduced the recurrence plot (RP) approach to picture the recurrences of dynamical systems. Recurrence plots are also beneficial for the analysis of the short and non-stationary data [13].

The method permits the identification of system features which cannot be seen using other known techniques. If $\{x_i\}_{i=1}^{N}$ is the trajectory of a system's output embedded in the RPS, where N is the number of the observed states [13], the corresponding RP could be evaluated through the following recurrence matrix

$$R_{i,j} = \begin{cases} 1, & x_i \cong x_j \\ 0, & x_i \neq x_j \end{cases} \quad i, j = 1, 2, \ldots, N \tag{2}$$

where $R_{i,j}$ compares the states of the system at times i and j [13]. If the states are similar, this is indicated by scheduling a "1" in the matrix. If the states are rather different, the corresponding value in the matrix must be zero. The RP matrix can be defined as

$$R_{i,j} = \theta\left(\varepsilon - x_i - x_j\right) \quad i, j = 1, 2, \ldots, N \tag{3}$$

where ε is the threshold of the distance, $\theta(.)$ denotes the Heaviside function, and . is the employed vector norm. If ε is put too small, we cannot gain useful information about the system. Vice versa, when ε is put too large, the number of objects increases and causes thicker and longer diagonal structures than they actually are [13]. In this paper, a symmetric RP with a fixed ε and the Euclidean norm [13] is used.

In addition to the recurrence plot approach, the recurrence quantification analysis (RQA) is also utilized as a method to detect dynamical changes occurs in a signal. The RQA is based on quantifying the certain aspects of the nearest neighbors which appear for a signal embedded in the phase space [17]; The RQA involves estimation of some parameters (called recurrence parameters) that describe the structures in the RP [18]. These structures are single dots, diagonal and vertical (or horizontal) lines [19]. The structures within a RP are related to different dynamics within the system. It is shown that this approach is computationally expensive [20]. So, in this work, we preferred to exploit the RP directly to analyze the involved speech frames.

2.3. Feature extraction

Processing of a two-dimensional image is a branch of the signal processing field. Many of the concepts employed to analyze one-dimensional signals are also applicable to images; the main difference between these two appears in the formulations, where two indices are needed to point at an input pixel. The Feature extraction from images can be done in two domains: the spatial domain and the transformed domain. In the first case, processing of a digitized image is carried out directly by point processing in the original space. In the second approach, the original image is first transformed, by a method

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

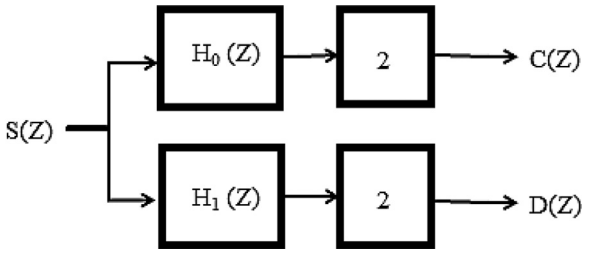

Fig. 1. Two-channel filter bank for implementing the discrete wavelet decomposition.

like the wavelet transform, into a new space, and the proper features are then extracted from it. Applying a professional feature extraction method could improve the following classification phase where needed. Here, two approaches by which the features are extracted from the RP-based images, including the dynamical invariants and the wavelet-based features, are introduced in Sections 2.3.1 and 2.3.2 respectively.

2.3.1. Dynamical invariants extracted from the RPs

Three beneficial and important features extracted from the RPs introduced in the literature are detailed in this section:

1. The distributions of diagonal lines in the RP; these parameters encode some principal properties of the system like the complexity scales and the predictability. The reverse of the longest diagonal in the RP is equivalent to the largest Lyapunov exponent of the system [13]. The number of the found diagonal lines and their mean are evaluated as two distinct structural features [13].
2. Auto mutual information: The mutual information is an effective method to quantify the amount of dependencies within or between time series [13]. The mutual information of a system could be given by

$$I_q^x(\tau) = 2H_q - H_q(\tau) \tag{4}$$

where H_q is the qth-order Renyi entropy of x_i, $H_q(\tau)$ is the qth-order joint Reyni entropy of x_i and $x_{i+\tau}$. Auto mutual information is extracted from the RPs as the third feature [21]. Estimation of the Reyni entropy is carried out using the means of the RP.
3. Entropy: in this part, the Reyni entropy [21] is calculated as the fourth feature

2.3.2. The feature extraction using the wavelet transform

The wavelet transform (WT) is regarded as a revolution in the modern signal processing field. The WT could be employed to present a time-frequency description of a signal. Many researches utilized the wavelet transform as a powerful tool in their feature extraction stages. Wavelets are used in two main structures: The Continuous Wavelet Transforms (CWT) [22] and the Discrete Wavelet Transforms (DWT) [22]. The latter one is typically implemented via a filter bank description. One example of the DWT application, based on a two-channel filter bank, is shown in Fig. 1. The filter bank is consisted of a low pass filter h_0 [n] and a high pass filter h_1[n]. The block of 2↓ represents the down-sampler operator, which acts by a factor of 2. The outputs of the filter bank are two series of coefficients that their z-transforms are shown by C(z) and D(z) so-called approximation and details outputs, respectively [22]. In other words, various information levels can be extracted by decomposing the original signal using a wavelet system with an orthonormal basis [22].

The wavelet transform could be implemented in different decomposition levels; however, high decomposition levels significantly increase the computational cost and time needed for the entire process. So, in this work, we preferred to use the discrete wavelet transform with only two decomposition levels.

A time signal $g(t)$ can be decomposed into approximation and details coefficients as given by

$$g(t) = \sum_{k=-\infty}^{+\infty} c(k)\varphi_k(t) + \sum_{j=0}^{+\infty}\sum_{k=-\infty}^{+\infty} d(j,k)_{j,k}(t) \tag{5}$$

where $c(k)$ and $d(j, k)$, are the approximations and details coefficients, respectively.

Since the wavelet basis functions are short waves ($\varphi_k(t)$) generated by scaling the original mother wavelet ($\Psi_{j,k}(t)$), they are well-localized in time and scale domains [22]. In a classification task, in which the employed features are extracted via a wavelet transform approach, the choice of the mother wavelet is important and needs some careful considerations, because it has deep influences on the output features and classification results [22]. In this work, a 2-Dimensional wavelet packet decomposition scheme, with different decomposition levels, is used as the main transformation method applied to the RP images.

In this study, a simple phone recognition task is used to search for a proper mother wavelet that to be applied to the RP images; we exploit the forward feature selection algorithm and a multiclass support vector machine toolbox (LibSVM) [23],

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

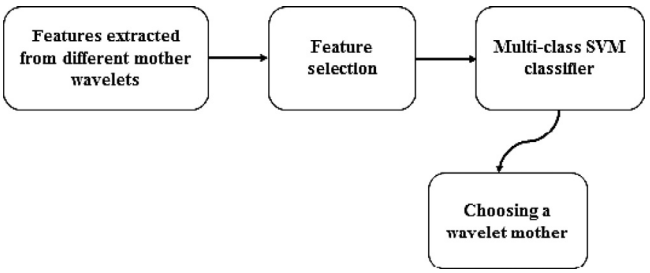

Fig. 2. The block diagram of the search algorithm to find the best mother wavelet among some known ones.

Table 1
Classification results (accuracy rates in %) obtained via 4 different mother wavelets.

Mother wavelet	The selected feature number	Accuracy rate%
Db4	7	64.3
Sym4	3	13.62
Coif2	5	26.81
Bior6.8	6	57.35

as the employed classifier, to find the best mother wavelet among some predefined ones. Fig. 2 shows the block diagram of the employed search algorithm.

First, 20 percentages of FARSDAT corpus is randomly selected and the related RP images are extracted from these selected speech signals. A two-level wavelet packet decomposition scheme, using distinctively 4 different mother wavelets, Daubechies4 (Db4), Symlet4 (sym4), Coiflet2 (coif2) and Biorsplines6.8 (bior6.8), are then applied to the RPs extracted from the experimental data. Next, Shanon entropy [21] and energy values are evaluated via the wavelet transform of the RP-based images. This leads to a 32-dimensional feature vector for each of the employed mother wavelets. Next, the forward feature selection method is utilized for dimensionality reduction of the features which led to different numbers of features remained for different mother wavelets. Finally, the selected features are separately fed to a multiclass SVM classifier with a linear kernel [23], employed for a simple isolated phoneme recognition task; in the following, the phoneme classification accuracy rates are distinctly calculated for different mother wavelets. The results of this experiment are shown in Table 1. According to the results, for the current task, the Db4 is the best mother wavelet among the examined mother wavelets. According to the first row of Table 1, for Db4 mother wavelet, one of the 7 remained features is from the set of 16 energy coefficients and the other 6 ones are from the entropy coefficients; these two kinds of features are distinctly evaluated over the wavelet decomposed images. The decomposed parts of a sample phone /a/ is shown in Fig. 3.

2.4. Dimensionality reduction methods

Two dimensionality reduction methods, the forward feature selection algorithm and the linear discriminant analysis (LDA), needed to reduce the dimension of the crude evaluated RPS-based features, are briefly explained in Subsections 2.4.1 and 2.4.2, respectively.

2.4.1. Feature selection

The feature selection is an important topic in the statistics and machine learning areas. Feature selection algorithms (FSAs) select subsets of the features which contain useful information from the data to improve the performance of the subsequent classifier [24]. The best subset is consisted of the least number of features that via which significantly growth in the accuracy is gained; so, by applying this approach, to implement more efficient classification phase, unimportant attributes are formerly being eliminated. In the literature, there are many feature selection methods introduced for dimensionality reduction in the feature space. In many works, this is a necessary preprocessing phase to remove irrelevant features, because they may cause greater computational costs and lead to some over fitting situations. As a basic idea, it is logical to neglect those input features which have a little impact on the final classifier output. It is shown that the feature selection process could enhance accuracy in many classification applications [24].

In this field, two common techniques, the forward and backward feature selection [24] are widely used. In this paper, the forward selection algorithm is used to reduce the dimensionality of the feature space. This approach begins by choosing an empty initial subset and iteratively, one by one, adding the selected features to it. The feature addition procedure continues until it does not significantly decrease the predefined error criterion. Let assume a set of input features as $\{x_1, x_2, \ldots, x_M\}$, where M is the original dimension of the feature space. First, a Leave-one-out cross validation (LOOCV) error is measured for the existed one-component subsets, $\{x_1\}, \{x_2\}, \ldots, \{x_M\}$, to find the best individual feature. Next, the algorithm searches for the best subset including two components, which one of them is the first selected feature; in the following, the input

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

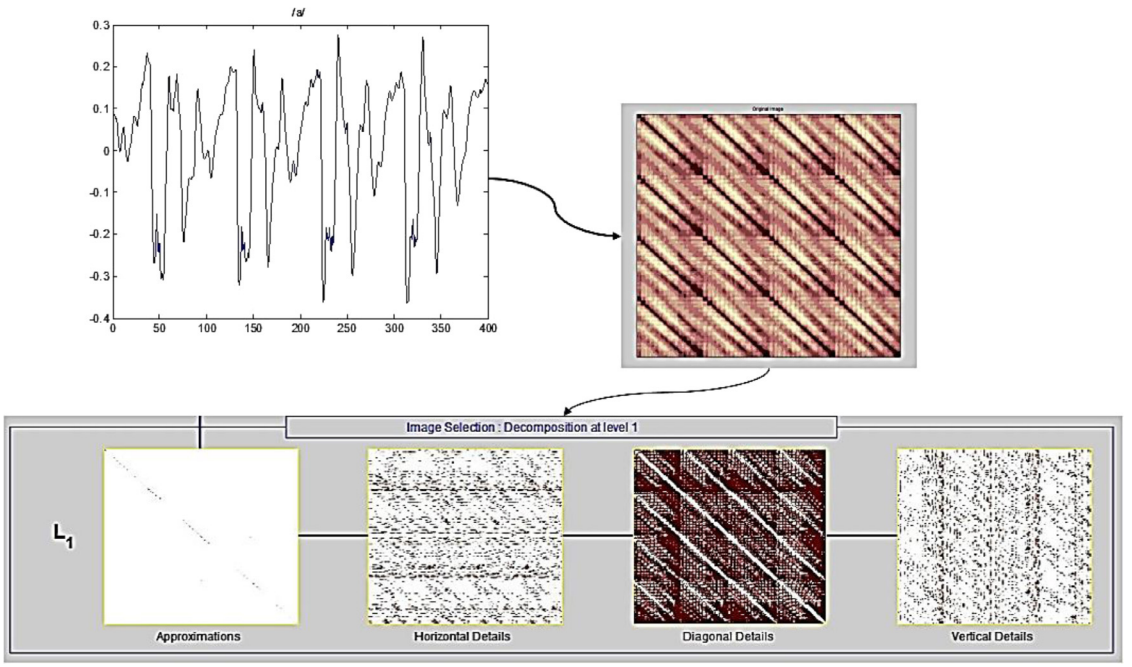

Fig 3. The one-level decomposition of a phoneme /a/ sample.

subsets with three, four, and more features are appraised. Finally, the best subset of the input features is selected that gains the most reduction in the error criterion [24].

2.4.2. Feature transformation

The dimensionality reduction in a feature space could also be done using feature transformation methods. Linear Discriminant Analysis (LDA) is an effective method to find a linear transformation of features for dimensionality reduction purpose. The LDA considers the differences among classes of data [25] and simultaneously decorrelates the transformed features. Although, in the case of the sophisticated systems with non-Gaussian distributions, the LDA may face difficulties to preserve essential data for the classification purpose, this approach has been a successful approach in the major of its applications. In this work, the LDA is employed as a transformation algorithm to reduce the dimensionality of the input space. For the purpose of dimensionality reduction, we can project original features only into the several first bases; in this manner, the useful information is almost kept in the transformed features to be employed in the final classification process [25]. The LDA transformation is defined as

$$F_{LDA} = W_{LDA}^{T}.F \tag{6}$$

where F denotes the feature vector of the underlying system, W_{LDA} is appraised from Eigen decomposition method and the superscript T is the transpose operator.

In this work, the LDA algorithm is performed over the extracted features obtained via the Db4 wavelet transform (32-dimensional feature vector) of the RP images created from the speech signals. The employed speech signals are taken from FARSDAT corpus. After sorting the Eigen values evaluated by implementing the LDA method, the first 10 highest ones and their assigned Eigen vectors were kept to construct the transformation matrix. In this manner, a 32×10-dimensional transformation matrix is obtained. Using this transformation matrix leads to 10-dimensional transformed feature vectors. The diagram of the sorted Eigen values is shown in Fig. 4. We see that the first 6 Eigen values have considerable values and the 4 rest are very small; so, it is reasonable to make the transformation matrix by using only the corresponding Eigen vectors of the first 6 selected Eigen values; this leads to a 32×6 transformation matrix. Hence, the dimensionality of the final transformed feature vector reduces to 6 and the compressed wavelet-based features will appear as 6-dimensional vectors for each frame.

3. Experimental methodology

3.1. System overview

In this work, the proposed features are evaluated from the recurrence plot representation of speech frames transformed to the RPS; it is shown that the combination of the proposed features with the MFCC features causes considerable improvement in the performance of the conventional used ASR system. Fig. 5 shows the schematic diagram of the proposed feature

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

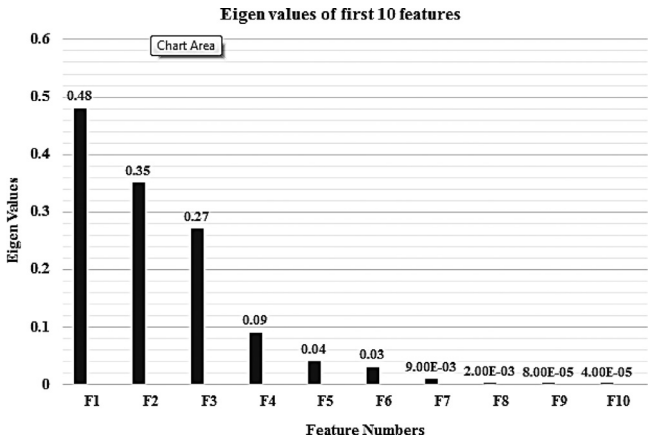

Fig. 4. The Diagram of the first ten Eigen values obtained via the LDA algorithm.

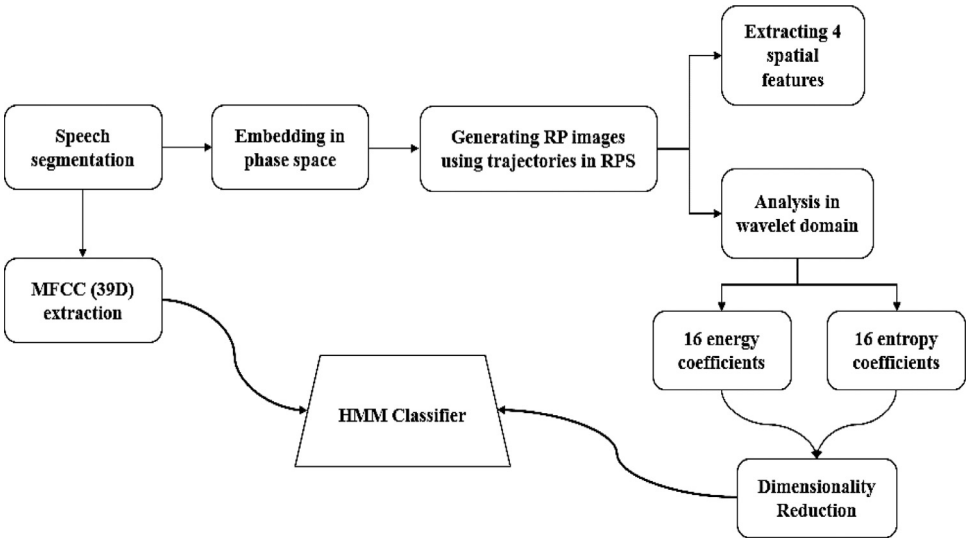

Fig. 5. The overall scheme of the proposed feature extraction and the following recognition approach.

extraction method and the following experimental setup employed for the validations of the methods. As shown in this figure, first of all, segmentation of the speech signals into 25 ms frames with 60% overlap is done. The resulted frames are reconstructed in the RPS to create the related trajectories, as shown in Fig. 5. Next, the recurrence plots of these trajectories are evaluated to probe the signal behavior. A variety of methods could be used to extract proper features from the generated RPs. We utilize dynamical invariants of the RP and 2D wavelet transform to extract useful features from the RPs. By applying one level discrete wavelet transform to the RP images, some energy and entropy based features are evaluated from them. Because, the extracted RP-based features typically appear in high dimensional vectors, a proper feature selection or reduction method is needed to be next applied to them. The combination of the conventional frequency domain features, here the MFCC, and the proposed features are then fed to a phoneme recognition system which basically is reconstructed from the Hidden Markov Models. To examine the benefits of the proposed features, the phoneme recognition accuracy obtained via employing different feature combinations is considered.

3.2. Database

The experiments are performed exploiting two TIMIT and FARSDAT corpora, which include English and Persian speech files, respectively. FARSDAT [26] is a small continuous speech corpus with a limited vocabulary of about 1200 words. The sampling rate of the wave signals is 16 kHz. FARSDAT includes 6080 utterances from 304 speakers, 20 sentences spoken in two sessions by each speaker. The speakers are from 10 dialect regions of Iran. FARSDAT speech signals are phonetically tagged by precise hand-labeled boundaries which could be used to extract desired isolated phoneme sets.

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

Table 2
The evaluated features are indexed by numbers 1 to 56 as scheduled below.

Feature Index	Description
1–39	(1) MFCC coefficients
40–41	(2) Properties of the RP Diagonal lines (Section 2.3.1)
42	(3) Auto mutual information (Section 2.3.1)
43	(4) Reyni entropy (Section 2.3.1)
44	(5) One energy coefficient selected from the wavelet-based features (Section 2.4.1)
45–50	(6) Entropy coefficients selected from the wavelet-based features (Section 2.4.1)
51–56	(7) wavelet-based features transformed by the LDA (Section 2.4.2)

TIMIT [9, 27] is a well-known English database which contains 6300 utterances (about 322 minutes) from 630 speakers, from eight greater dialect regions of the United States. The sampling rate of the speech signals is 16 kHz. In this research, we have utilized the training and test parts of TIMIT to conduct the needed experiments. The training subdivision includes 10 sentences from each of 462 speakers, but only the SI and SX sentences are eliminated, resulting in 3696 sentences. The SI and SX sentences (eight sentences at all) are pronounced by all the speakers in TIMIT corpus. We omitted these sentences from the training dataset, to avoid the problem of over fitting during training phase of the proposed system. The used test subdivision provides 1344 sentences from 168 speakers without the SA sentences. The original 64 phoneme TIMIT corpus is decreased to 39 phonemes set (CMU/MIT Standards) to be used in constructing the acoustic models.

In this work, speech signals are segmented into 25 ms frames with 60% overlap and are then processed and applied to the involved HMMs.

4. Experimental results

In this section, we evaluate the efficiency of the proposed RP-based features inside a typical CSR task.

The Hidden Markov Model Toolkit (HTK) [28] is especially designed for development of Hidden Markov Models (HMMs). Moreover, it is supported by some auxiliary functions which make it a well-known toolkit for speech recognition purposes [28]. As mentioned earlier, this free and open source software is employed to conduct the experiments.

To show the benefits of the proposed features for the continuous speech recognition task, we conduct some valuable experiments in this field. The recognition system is composed of six-state monophonic HMMs with eight Gaussian mixtures in each state [8]. Moreover, the silent is modeled by a three-state HMM. The performance of the CSR system is evaluated by the accuracy (ACC %) of the recognized phone sequences as given by

$$ACC\% = \left(1 - \frac{S + I + D}{N}\right) \times 100 \tag{7}$$

where S is the number of substitutions, I is the number of insertions, D is the number of deletions, and N denotes the original number of phones.

In Section 2.3, a number of different RP-based feature extraction methods were introduced. Some of these approaches are relatively old and are introduced in the literature; the remained ones are newly proposed in the current study. We combine these features in different ways with the MFCC to examine the benefits of adding the RP-based features to the typical MFCC features in the CSR task. Previously, some CSR experiments are conducted which employ the proposed features in different groups to investigate their benefits when the MFCC features are not presented. Table 2 shows how the different groups of the evaluated features are numbered. MFCC features (row 1 of the table) are consisted of 13 MFCC features plus its delta and delta-delta values that leads to a 39-dimensional vector. Rows 2, 3 and 4 of the table, show the RP-based features introduced in Section 2.3.1. Features indexed in rows 5 and 6 together are the best 7 wavelet-based features selected via the forward selection algorithm, aforementioned in Section 2.4.1; they were formerly denoted in row 1 of Table 1. The one scheduled in the row 5 of Table 2 is selected from the energy coefficients evaluated from the wavelet decomposition of the RP-based image, and the other 6 ones, selected from the entropy of the wavelet coefficients, are put in the row 6 of the table. The last row of this table is assigned to wavelet features which the LDA is applied to them to reduce their dimension from 32 to 6. These features were introduced in Section 2.4.2.

The CSR results using the proposed features in individual groups for FARSDAT and TIMIT corpora are shown in Table 3. Fig. 6 shows the results for different combination of the proposed RP-based features with MFCC features.

Typically, speech recognition systems work reasonably well in clean conditions but work poorly under noisy conditions. In this paper, the preliminary experiments were conducted over TIMIT database without adding noise; following these experiments, to move toward more realistic conditions, different noises with various power levels were added to the clean speech signals. Two different white and babble noises at different SNRs were added to the TIMIT corpus and the experiments were conducted over this constructed noisy database to evaluate the robustness of the proposed method against the noisy conditions. Again, the obtained results show the superiority of the proposed features combined with the MFCC features against the MFCC features alone.

The results are shown in the Figs. 7 and 8 for “Babble” and “White” noises, distinctively.

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

Table 3
The CSR accuracy results (in %) using individual groups of the RP-based features for the FARSDAT and TIMIT test sets.

Features Indexes	FARSDAT(Persian speech data)	TIMIT(Enlish speech data)
40–43	14.18	10.54
42–43	34.39	28.43
44–50	45.48	38.31
42–50	**51.56**	**45.48**
51–56	44.96	38.73
42–43 & 51–56	**51.64**	**41.85**

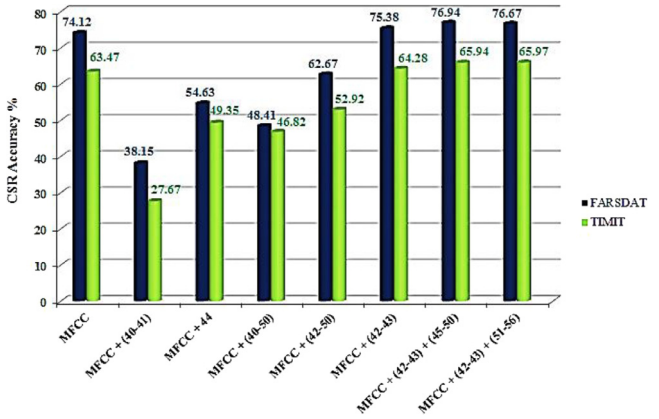

Fig. 6. The CSR accuracy results (in %) using the combination of the MFCC and the RP-based features for FARSDAT and TIMIT test sets.

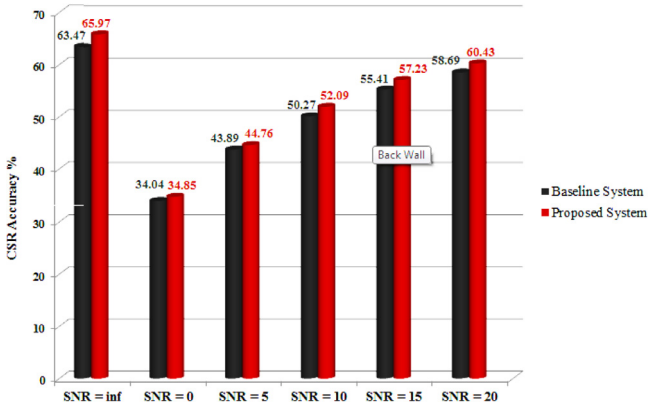

Fig. 7. The CSR accuracy results (in %) obtained over the Noisy TIMIT test set (created by adding "Babble" noise), using the different combinations of the MFCC and the proposed RP-based features.

4.1. Discussion

The current work shows that the information extracted from the recurrence plots of speech trajectories in the phase space is valuable for the following automatic speech recognition task. Of course, the spectral based features like the MFCC have proved their benefits in this task and could not be easily replaced by novel features. In this study, as a start point to evaluate this idea, the proposed features are not seriously targeted to be individually exploited in the ASR task; but, are viewed as some complementary features to be added to the known and proved features like the MFCC. This is worthy to remind that to find the typical spectral features, the phase of the speech signal is neglected; in this point of view, the RPS-based features may be proper features to compensate a part of this lack of information. The recurrence plot representation could effectively transfer the specifications of cyclic trajectories embedded in the phase space. The experimental results verify that by first applying the RP transformation to speech signals and then extracting some proper features from the resulted 2-dimensional signal, we are supported by some features of the underlying speech production system that could not be fully captured by the common spectral based ones.

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

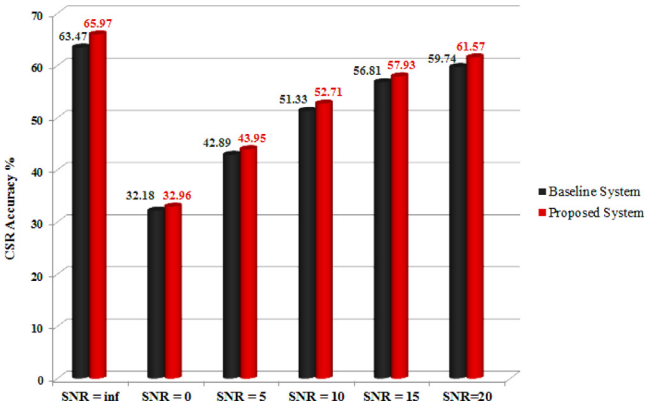

Fig 8. The CSR accuracy results (in %) obtained over the Noisy TIMIT test set (created by adding "White" noise), using the different combinations of the MFCC and the proposed RP-based features.

The RP transformation leads to a 2-dimensional signal, like an image. We examined that the application of wavelet packet decomposition over the resulted image, using the db4 mother wavelet, could be led to some proper features which have considerable benefits for the involved speech recognition task. Table 3 shows that using only the proposed features in an ASR system is not adequate to obtain satisfactory results; moreover, it shows that the wavelet-based features selected via the forward selection algorithm provided better performance in comparison to the other RP-based features.

Of course, we found that the combination of the RP-based features with the spectral-based features could considerably promote the following CSR process. Features no. 40, 41, 44 did not show good performances in the conducted experiments. As shown in Fig. 6, using FARSDAT and TIMIT corpora, the combination of RP-based features (selected by the forward feature selection method) with MFCC features respectively led to 3.80% and 3.94% absolute classification accuracy improvement against the baseline-system which uses only the MFCC as the input feature set. Furthermore, for noisy speech recognition, using the combination of the MFCCs and the RP-based features still yields to better results, compared to using only the MFCC features.

Moreover, the experiments indicate that in low SNRs, the proposed features are less sensitive to additive white noise than the additive babble noise. The difference could be resulted from the nature of the white noise which is coherently different from the speech signal and when they are simultaneously transformed to the RP plane, could be better discriminate from each other. This could be compared with conditions that these two are mixed in the time or frequency domain.

Improvements resulted from the proposed features, for the clean and noisy data (15 dB and 10 dB, white noise), respectively are equal to 3.94%, 3.29% and 3.62%. These results verify that the proposed features keep their benefits for the following CSR task even for nearly powerful noises. In the other words, the combined features have better performance in comparison to only MFCC features whether for clean or noisy data. Of course, the performance of the CSR system degrades at high levels of noise. For example, they became useless for low SNRs, like 0 dB, as could be seen in Figs. 7 and 8.

5. Conclusion

In this work, we searched for a nonlinear feature extraction method, employing the RPS theory, to capture some nonlinear features from the human speech production system. Speech frames were first embedded in the RPS, utilizing Taken's theory. Using the speech trajectories in the phase space and the recurrence property of the dynamical system, the related RP pattern was generated. The known dynamical invariants were next derived from the resulted RPs; moreover, the wavelet decomposition was applied to the RPs and then by evaluating Shanon entropy and the energy of the coefficients, some pure wavelet-based features were resulted. The linear discriminant analysis and the forward feature selection algorithms were then applied to them.

There is an important notice that, in the speech recognition field, the proposed features alone, even with enough embedding dimension and applying appropriate feature selection methods, do not work as well as the proved features obtained via frequency domain analysis. So, we combined them with the traditional MFCC features. The proposed feature set showed good performance in the ASR experiments. In this way, we gained 3.80% and 3.94% absolute classification accuracy growth in the speech recognition accuracy for FARSDAT and TIMIT corpora, respectively. This is very considerable, and verifies that the proposed RP-based features could significantly help to CSR applications which typically are faced to very short time series (frames with about 25 ms lengths). Of course, this study does not reject this probability that by modifying this approach in the future, the RPS-based features may become a complete replacement for the spectral based features like the MFCC.

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

References

[1] Varela O, San-Segundo R, Hernandez LA. Combining pulse-based features for rejecting far-field speech in a HMM-based Voice Activity Detector. Comput Electr Eng 2011;37:589–600.
[2] Kopparapu SK, Bhuvanagiri KK. Recognition of subsampled speech using a modified Mel filter bank. Comput Electr Eng 2013;39:655–62.
[3] z Ge, Sharma SR, Smith MJT. Improving mispronunciation detection using adaptive frequency scales. Comput Electr Eng 2013;39:1464–72.
[4] Kokkinos I, Maragos P. Nonlinear speech analysis using models for chaotic systems. IEEE Trans Speech Audio Process 2005;13:1098–109.
[5] Povinelli RJ, Johnson MT, Lindgren AC, Roberts FM, Ye J. Statistical models of reconstructed phase spaces for signal classification. IEEE Trans Signal Process 2006;54:2178–86.
[6] Johnson MT, Povinelli RJ, Lindgren AC, Ye LiuX, Indrebo KM. Time-domain isolated phoneme classification using reconstructed phase spaces. IEEE Trans Speech Audio Process 2005;13:458–66.
[7] Povinelli RJ, Johnson MT, Lindgren AC, Ye J. Time series classification using gaussian mixture models of reconstructed phase spaces. IEEE Trans Knowl Data Eng 2004;16:779–83.
[8] Jafari A, Almasganj F, NabiBidhendi M. Statistical modeling of speech poincaré sections in combination of frequency analysis to improve speech recognition performance. Chaos 2010;20:1–11.
[9] Shekofteh Y, Almasganj F. Autoregressive modeling of speech trajectory transformed to the reconstructed phase space for ASR purposes. Digit Signal Process 2013;23:1923–32.
[10] Zhang J, Man KF, Ke JY. Time series prediction using Lyapunov exponents in embedding phase space. In: IEEE International conference on systems, man, and cybernetics, vol. 2; 1998. p. 1744–9.
[11] Ezeiza A, Ipia KL, Hernndez C, Barroso N. Enhancing the feature extraction process for automatic speech recognition with fractal dimensions. (Springer) Cogn Comput 2013;5:545–50.
[12] Shekofteh Y, Almasganj F, Daliri A. MLP-based isolated phoneme classification using likelihood features extracted from reconstructed phase space. J Eng Appl Artif Intell 2015;44:1–9.
[13] Marwan N, Romano MC, Thiel M, Kurths J. Recurrence plots for the analysis of complex systems. Phys Rep 2007;438:237–329.
[14] Narayanan S, Alwan A. A nonlinear dynamical systems analysis of fricative consonants. J Acoust Soc Am 1995;97:2511–24.
[15] Shekofteh Y, Almasganj F. Feature extraction based on speech attractors in the reconstructed phase space for automatic speech recognition systems. J ETRI 2013;35:100–8.
[16] Eckmann JP, Kamphorst SO, Ruelle D. Recurrence plots of dynamical systems. Europhys Lett 1987;4:973–7.
[17] Fatoorehchi H, Zarghami R, Abolghasemi H, Rach R. Chaos control in the cerium-catalysed Belousov-Zhabotnisky reaction using recurrence quantification analysis measures. ChaosSolitons Fract 2015;76:121–9.
[18] Liop MF, Gascons N, Liauro FX. Recurrence plots to characterize gas-solid fluidization regimes. Int J Multiple Flow 2015;73:43–56.
[19] Tahmasebpoora M, Zarghamib R, Sotudeh-Gharebaghb R, Mostoufi N. Characterization of fluidized beds hydrodynamics by recurrence quantification analysis and wavelet transform. Int J Multiphase Flow 2015;69:31–41.
[20] Yan J, Wang Y, Ouyang G, Yu T, Li X. Using max entropy ratio of recurrence plot to measure electrocardiogram changes in epilepsy patients. Physica A 2016;443:109–16.
[21] Bromiley PA, Thacker NA, Bouhova-Thacker E. Shannon entropy, Renyi entropy, and information. Stat Segment Series 2010.
[22] ErfanianSaeedi N, Almasganj F, Torabinejad F. Support vector wavelet adaptation for pathological voice assessment. J Comput Biol Med 2011:822–8.
[23] Chang CC, Lin CJ. LIBSVM. a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2011;2:1–27.
[24] Ladha L, Deepa T. Feature selection methods and algorithms. Int J Comput Sci Eng (IJCSE) 2011;3.
[25] McLachlan GJ. Discriminant analysis and statistical pattern recognition. Wiley; 2004.
[26] FARSDAT. Persian speech database. Available: http://catalog.elra.info/product_info.php?products_id=18.
[27] Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D, Dahlgren N. DARPA TIMIT acoustic-phonetic continuous speech corpus [CD-ROM]; 1993. Available: https://catalog.ldc.upenn.edu/LDC93S1.
[28] Young S, Evermann G, Gales M, Th Hain, Kershaw D, Liu X, et al. The HTK book; 2006. Version 3.4.

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006

Shabnam Gholamdokkht Firooz received her BS in biomedical engineering from Amirkabir University of Technology, Tehran, Iran in 2011. She received her MS in biomedical engineering from Amirkabir University of Technology in 2013. She is currently a PhD candidate in the Biomedical engineering at Tehran University. Her research is mainly focused on signal processing and speech recognition.

Farshad Alamsganj 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 Biomedical Engineering Department of Amirkabir University of Technology. His research interests include automatic detection of voice disorders, speech recognition, prosody, and language modeling for ASR systems.

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 and PhD in biomedical engineering from Amirkabir University of Technology in 2008 and 2013 respectively. His research is mainly focused on signal processing, speech recognition, and keyword spotting.

Please cite this article as: S.G. Firooz et al., Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers and Electrical Engineering (2016), http://dx.doi.org/10.1016/j.compeleceng.2016.07.006