
HETEROGENEOUS RESERVOIR COMPUTING MODELS FOR PERSIAN SPEECH RECOGNITION

Zohreh Ansari

Biomedical Engineering Department
Meybod University
Meybod, Iran
z_ansari@meybod.ac.ir

Farzin Pourhoseini

Biomedical Engineering Department
Meybod University
Meybod, Iran
pourhoseinifarzin@gmail.com

Fatemeh Hadaeghi

Institute of Computational Neuroscience
University Medical Center Hamburg-Eppendorf (UKE)
Hamburg, Germany
f.hadaeghi@uke.de

ABSTRACT

Over the last decade, deep-learning methods have been gradually incorporated into conventional automatic speech recognition (ASR) frameworks to create acoustic, pronunciation, and language models. Although it led to significant improvements in ASRs' recognition accuracy, due to their hard constraints related to hardware requirements (e.g., computing power and memory usage), it is unclear if such approaches are the most computationally- and energy-efficient options for embedded ASR applications. Reservoir computing (RC) models (e.g., echo state networks (ESNs) and liquid state machines (LSMs)), on the other hand, have been proven inexpensive to train, have vastly fewer parameters, and are compatible with emergent hardware technologies. However, their performance in speech processing tasks is relatively inferior to that of the deep-learning-based models. To enhance the accuracy of the RC in ASR applications, we propose heterogeneous single and multi-layer ESNs to create non-linear transformations of the inputs that capture temporal context at different scales. To test our models, we performed a speech recognition task on the Farsdat Persian dataset. Since, to the best of our knowledge, standard RC has not yet been employed to conduct any Persian ASR tasks, we also trained conventional single-layer and deep ESNs to provide baselines for comparison. Besides, we compared the RC performance with a standard long-short-term memory (LSTM) model. Heterogeneous RC models (1) show improved performance to the standard RC models; (2) perform on par in terms of recognition accuracy with the LSTM, and (3) reduce the training time considerably.

Keywords Automatic speech recognition · Deep echo state networks · Heterogeneity · Farsdat · Persian (Farsi) language · Recurrent neural networks · Reservoir computing (RC)

1 Introduction

1.1 Practical relevance and motivation

Over decades of research, hidden Markov models and different variations of neural networks have been extensively exploited to perform speech recognition, speaker identification, text to speech conversion, and other tasks that are relevant for speech processing applications [1, 2, 3, 4]. During the past decade, and thanks to remarkable advancements in graphics processing unit (GPU) and cloud computing technologies, a large variety of deep-learning-based methods have been designed and tested to accomplish challenging speech processing tasks on large datasets [5, 6, 7, 8]. In conventional ASR systems, these models are primarily employed to extract features, to uncover the relationships between the audio signal and the phonemes (or other linguistic units), to learn pronunciation lexicons, and to recognize

(and use) sequences of words. These developments, however, have been mainly focused on improving recognition accuracy by training models on a multitude of collected data. Imprecision and false interpretation, although momentous, are not the only challenges faced by practical ASR systems. Since algorithms have to be computationally efficient, less data-hungry, and compatible with emerging micro-device technologies, further investigations are needed to develop inexpensive yet accurate processing methods. Besides, it is of crucial importance to make speech recognition available for more languages and dialects.

In contrast to deep-learning-based methods, reservoir computing (RC) models (e.g., echo state networks (ESNs) [9] and liquid state machines (LSMs) [10]) have been proven inexpensive to train, have vastly fewer parameters, and reported to perform well in processing complex temporal and spatio-temporal data in real-world applications [11, 12, 13]. It has also been suggested that RC models can provide accurate subject-specific classifiers that are adaptable to the unique characteristic features of temporal data recorded from a target person and do not rely on a vast amount of data collected from other individuals [14]. Besides, the essential ingredient of a reservoir computing model is a random excitable medium that non-linearly projects an input signal into a higher-dimensional signal space. Therefore, researchers from computing theory and microchip technologies have considered RC as a computational scheme compatible with “unconventional” physical or computational platforms such as analog electrical circuits [15, 16], optical media [17], and chemical (molecular) substrates [18]. It is, therefore, promising to design and implement functional sensors, processors, and controllers based on this computational framework.

However, up to this point, the performance of RC in speech processing tasks has been relatively inferior to that of the deep-learning-based models [19] and needs further improvements to fit conventional or end-to-end ASR systems with practical exploitation. In this regard, a major upgrade to shallow RC systems was introducing deep echo state networks (deep ESNs) that are able to capture temporal context of the input signal at different time-scales through several successively stacked RC layers [20, 21]. It enhanced the performance of RC in time-series prediction [20], short-term memory capacity (MC) task, and classification of experimental data in the field of computational biology [22]. However, it has remained to be further assessed if this alteration could also strengthen RC-based speech recognition models. In this study, therefore, we chose to explore applicability of RC models in speech recognition. Besides, since to the best of our knowledge, neither standard RC nor deep ESN have been employed to conduct any Persian ASR tasks, we trained conventional single-layer and deep ESNs to perform a speech recognition task on the Farsdat Persian speech dataset.

1.2 Related works

The FARSDAT dataset [23] was collected in 1996 and since then has been used as the standard benchmark for developing Persian ASR systems such as Shenava [24] and Nevisa [25]. Similar to other ASR models, both systems comprise three modules for feature extraction, and acoustic and language modelling. Dimension reduction techniques such as principle component analysis, wavelet transforms, filter banks, Cepstral analysis, Mel frequency cepstrum, and kernel based methods are commonly employed to extract relevant features from each 20-30 ms frames (segments) of the speech signal. While standard Mel frequency Cepstral coefficient (MFCC) analysis was dominantly utilized in the feature extraction modules of the first editions of Persian ASRs, it has been recently suggested that LHCB (Logarithm of Hamming Critical Band filter banks) [26], convolutional neural networks (CNNs), and deep-belief-networks (DBNs) may return features that further increase the accuracy of Persian speech recognition in neural network models [27].

In the acoustic and language modelling modules, combinations of Gaussian mixture models (GMMs) and deep neural networks with hidden Markov models have been extensively investigated to convert the extracted features to a probability over characters in the Persian alphabet, and to turn these probabilities into words [28, 29, 30, 31]. In addition to conventional Persian ASR platforms, end-to-end systems based on modular deep neural architectures, and uni- and bi-directional long-short-term-memory (LSTM) framework were recently developed and tested on Farsdat dataset [32, 33, 27]. The objective of current study is to investigate if reservoir computing models can be further integrated into either conventional or end-to-end Persian ASR systems.

1.3 Contributions

We present, to the best of our knowledge, the first study that explores capabilities of reservoir computing in the context of Persian speech recognition. Specifically, we employed RC algorithms that are suitable for application to speech recognition. Moreover, we propose heterogeneous single and multi-layer ESNs to create non-linear transformations of the inputs that capture temporal context at multiple scales. The RC-based algorithms are also compared to a de-facto standard LSTM as problem-tailored state-of-the-art deep learning RNN solution.

2 Model Architectures

2.1 Baseline architectures

2.1.1 LSTM

Long short-term memory RNN architectures are widely used for sequence labeling and prediction tasks, including speech processing applications such as spoken language translation [34, 35] and speech recognition [36]. In LSTM networks, the standard hidden layer of recurrent neural networks (RNNs) has been replaced with purpose-built gates and memory cells to filter and store the information. This modification has been proven particularly effective in tackling the vanishing gradient problem and fruitful in finding, memorizing, and exploiting long range dependencies in sequential data [37, 36]. In this study, we took the standard deep-learning LSTM architecture as de-facto standard in our speech recognition task.

2.1.2 Shallow Reservoir Computing

Reservoir computing provides a computationally efficient framework for RNN design and training and has been successfully used in a large range of practical signal processing applications across different fields [9, 13, 12]. The core to a typical reservoir computing model is a random, large, fixed recurrent neural network comprising a set of sparsely connected non-linear nodes (see Fig. 1.A). Through the internal variables of this dynamical system (i.e., reservoir states), the input signal presented to the RNN is non-linearly mapped into a higher dimensional signal space. These states are used to train a feed-forward readout module that is the only trained part of the network. The time-dependent output is computed as a linear combination of these random representations. Depending on the task, randomly generated output to reservoir (all-to-all) feedback connections may also be included in the architecture.

Unlike traditional RNN training methods, the RC technique proposes that the values of input-to-reservoir and reservoir connection weights are not critical and can be selected at random within some pre-defined intervals to obtain the best performance. Training only takes place in the readout layer where the signals from the individual nodes are fitted to a training signal, usually by a linear fit. Since only the output connections are trained and the optimization of the output layer only consists in a linear regression, reservoir computer can be faster and computationally more efficient than training a conventional recurrent neural network.

The dynamics of a reservoir computer with real-time continuous value units is typically described by the following equations:

$$\mathbf{x}(t + \Delta t) = (1 - \eta a)\mathbf{x}(t) + \eta f(\mathbf{W}^{\text{in}}\mathbf{u}(t + \Delta t) + \mathbf{W}\mathbf{x}(t) + \mathbf{W}^{\text{fb}}\mathbf{y}(t)), \quad (1)$$

where Δt (here, $\Delta t = 1$) is the real time sampling period, $\mathbf{u}(t) \in \mathbb{R}^{N_u}$ is the input sequence, $\mathbf{x}(t) \in \mathbb{R}^{N_x}$ is the N_x -dimensional reservoir state, and f is a nonlinear function. $\mathbf{W} \in \mathbb{R}^{N_x \times N_x}$, $\mathbf{W}^{\text{in}} \in \mathbb{R}^{N_x \times N_u}$, and $\mathbf{W}^{\text{fb}} \in \mathbb{R}^{N_x \times N_y}$ are the input-to-reservoir, recurrent, and output-to-reservoir feedback weight matrices, respectively. In this equation, $a > 0$ and $\eta = 1$ are reservoir neurons' leakage rate and time constant. The output, $\mathbf{y}(t) \in \mathbb{R}^{N_y}$ is then obtained from the extended system state (i.e., $\mathbf{z}(t) = [\mathbf{x}(t); \mathbf{u}(t)]$ where $[\cdot; \cdot]$ stands for a vertical vector concatenation):

$$\mathbf{y}(t) = g(\mathbf{W}^{\text{out}}\mathbf{z}(t)), \quad (2)$$

where g is an output activation function and \mathbf{W}^{out} is the readout weight matrix. Training the readouts is done by computing the linear regression weights of the target outputs on the harvested states of reservoir units and the inputs via pseudoinverse method:

$$\mathbf{W}^{\text{out}} = \mathbf{Y}_{\text{target}}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \gamma^2 I)^{-1} \quad (3)$$

where \mathbf{Z} is the matrix of extended system states, $\mathbf{Y}_{\text{target}}$ denotes the target sequence, I is the identity matrix and $\gamma \geq 0$ is a regularization factor.

2.1.3 Deep Reservoir Computing

The study of hierarchically organized recurrent neural networks suggests that deep RNNs are able to progressively develop multi-scale representations of the temporal information in their internal states, making them attractive models for complex sequential learning tasks such as text, speech and language processing [38]. Such multi-layer hierarchical architecture comprises an input layer fed by the external stimuli, several intermediate recurrent network layers which receive the hidden state of the previous layer as input, and an output layer that only has access to the hidden state of the

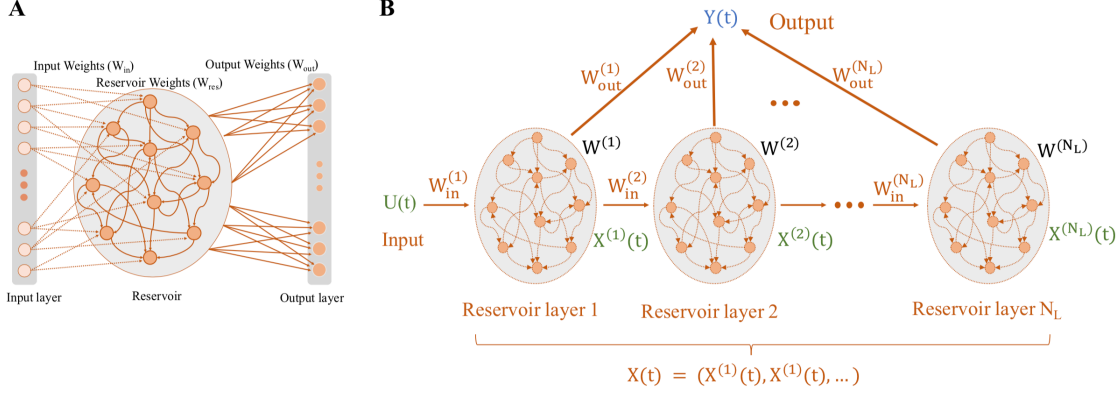


Figure 1: Schematic illustration of **A**: a shallow and **B**: a deep reservoir computing model.

final hidden layer. The training process in these models entails computationally expensive techniques such as temporal back-propagation [38, 21], therefore, investigations in the field of deep RNN are predominantly about the learning process. Deep reservoir computing, in contrast, provides a computationally efficient alternative approach for training. As depicted in Fig. 1.B, core components of this model are sequentially and unidirectionally stacked hidden layers where connections from higher to lower layers as well as connections from the input to layers other than the first level are avoided. Hidden states of these components are further exploited by a readout mechanism to learn the desired temporal task. Analogous to training a shallow ESN, in the case of a deep ESN, the state of the reservoirs at all the layers are concatenated in a predefined order to train a feed-forward readout module that is the only trained part of the network. Following the mathematical notation explained in [21], internal state update equation of the i -th layer of a deep ESN is as follows:

$$\mathbf{x}^{(i)}(t + \Delta t) = (1 - a^{(i)})\mathbf{x}^{(i)}(t) + a^{(i)} \tanh(\mathbf{W}_{in}^{(i)}\mathbf{x}^{(i-1)}(t) + \boldsymbol{\theta}^{(i)} + \mathbf{W}^{(i)}\mathbf{x}^{(i)}(t)). \quad (4)$$

Where $a^{(i)} \in [0, 1]$ is the leaking rate parameter of the i -th layer. With a simple assumption that all the reservoir layers comprise N_R neurons, $\mathbf{W}_{in}^{(i)} \in \mathbb{R}^{N_R \times N_R}$ and $\mathbf{W}^{(i)} \in \mathbb{R}^{N_R \times N_R}$ denote the input and recurrent weight matrices, respectively, and $\boldsymbol{\theta}^{(i)} \in \mathbb{R}^{N_R}$ is the bias-to-reservoir weight vector for this layer. Since the first layer (i.e., $i = 1$) receives the external inputs (i.e., $\mathbf{x}^{(0)}(t) = \mathbf{u}(t)$), $\mathbf{W}_{in}^{(1)} \in \mathbb{R}^{N_R \times N_u}$ denotes a random input-to-reservoir weight matrix. Considering l reservoir layers, one possible strategy to compute the output of a deep ESN at each time step, t , would be connecting the reservoir neurons in all the layers to the readout unit(s) and creating the extended system state as:

$$\mathbf{z}(t) = [\mathbf{u}(t); \mathbf{x}^{(1)}(t); \dots; \mathbf{x}^{(l)}(t)]. \quad (5)$$

Similar to the standard RC, the output at each time step, and the readout weights are further obtained by applying Eq. 2 and Eq. 3. It should be noted that in this study, randomly generated output-to-reservoir feedback connections were not included in the architecture.

2.2 Heterogeneous Reservoir Computing

One technique to enhance the accuracy of the RC in ASR applications could be enforcing multi-scale spatiotemporal reservoir dynamics through increasing variability in the hidden layer. Therefore, different gradient-based learning methods have been previously employed to optimize global hyper-parameters as well as learnable neuron-specific variables to adjust timescales of this dynamical system such that the required memory to solve a given task is created [39, 40, 41, 42]. In this study, we propose a laminar-specific architecture to develop heterogeneous single and multi-layer ESNs which diversifies dynamics of individual units.

In shallow ESN, we introduced unit variability by organizing the hidden layer into three sub-groups (i.e., laminar layers) of interconnected units with slightly different state update equations. As denoted in Eqs. 6 and 7, the future state of

reservoir nodes in each ensemble, $\mathbf{x}_{(i)}(t + \Delta t)$, is given in terms of the values of the other state variables at previous times:

$$\mathbf{x}(t + \Delta t) = (1 - a)\tilde{\mathbf{x}}(t) + f(\mathbf{W}^{\text{in}}\mathbf{u}(t + \Delta t) + \mathbf{W}\tilde{\mathbf{x}}(t)), \quad (6)$$

where

$$\tilde{\mathbf{x}}(t) = [\mathbf{u}(t); \mathbf{x}_{(1)}(t - \tau_1); \dots; \mathbf{x}_{(l)}(t - \tau_l)]. \quad (7)$$

In this notation, $\mathbf{x}_{(i)}$ stands for i -th pre-defined sub-group of internal neurons. It should be noted that the recurrent weight, $\mathbf{W} \in \mathbb{R}^{N_x \times N_x}$, and the input-to-reservoir weights, $\mathbf{W}^{\text{in}} \in \mathbb{R}^{N_x \times N_u}$, are initiated randomly before assigning multiple time delays to different sub-groups.

For deep ESN architecture, we applied the same multiple timescale framework by assigning gradually increasing time delays to successive layers in a way that the state variables in deeper layers depend on a longer history of their own states. To implement that, we simply modified the state equation and the extended system introduced in Eqs. 4 and 5 as follows:

$$\mathbf{x}^{(i)}(t + \Delta t) = (1 - a^{(i)})\mathbf{x}^{(i)}(t - \tau_i) + a^{(i)} \tanh(\mathbf{W}_{\text{in}}^{(i)}\mathbf{x}^{(i-1)}(t) + \boldsymbol{\theta}^{(i)} + \mathbf{W}^{(i)}\mathbf{x}^{(i)}(t - \tau_i)). \quad (8)$$

$$\mathbf{z}(t) = [\mathbf{u}(t); \mathbf{x}^{(1)}(t - \tau_1); \dots; \mathbf{x}^{(l)}(t - \tau_l)], \quad (9)$$

where τ_i , the internal delay in i -th ensemble/layer, can be either fine-tuned as a hyperparameter or trained via gradient based methods.

3 Materials and methods

3.1 Data-set and data characteristics

All the experiments in this manuscript were conducted on the FARsi Speech DATAbase (FARSDAT)[23] which has been created for Persian speech and speaker recognition purposes. This collection comprises 608 wave files, consisting 20 seconds long of 9 – 12 sentences spoken by 304 native Persian speakers with non-identical accents from different age groups. FARSDAT audio signals were recorded with sampling rate of 22.5 KHz, and the signal-to-noise ratio is 34 dB. Manual annotations at phoneme as well as word-level have been provided for all utterances.

Among the uttered sentences, two are common to all speakers and contain all the Persian alphabet letters except for “fe”. In the experiments conducted in this article, recordings of these two sentences uttered by 100 speakers were adopted for reporting the primary results of each RC structure, as well as, hyper-parameter optimization. Therefore, within one-fold cross-validation setting, the training set includes the speech signal of 70 speakers (20 percent of which was randomly selected for validation) and the test set contains audio signals recorded from another 30 speakers. In addition to the performance on the small set of common sentences, we report the recognition results on total FARSDAT speech database, which we refer to as *complete FARSDAT* database where the total sentences uttered by 297 speakers (i.e. 5940 sentences) were selected as the training set and those recorded from the other speakers (140 sentences) were utilized to test the models.

3.2 Pre-processing

During the experiments, each audio signal was segmented to overlapped frames of the same lengths (usually 23 ms with 12.5 ms overlap) to ensure the signal segments remain statistically static. The frames were further windowed by the Hamming window to calculate the spectrum of the frame using the Fast Fourier Transform (FFT). The received spectrum was afterward passed through the 18 filter banks on the Bark scale. Then, the logarithms of the energies under each filter were separately calculated to obtain 18 LHCB (Logarithm of Hamming Critical Band filter banks) representation vectors [26]. Since, LHCB features are sensitive to noise and signal variations, the longitudinal norm-2 (i.e., mean and variance normalization) was used to normalize LHCB components to enhance the accuracy and efficiency of the developed models.

The features extracted from training data were afterward exploited to adjust models' hyper-parameters and to train learnable parameters of the neural networks. Finally, the obtained models were evaluated on two test data, the small set of common sentences and the *complete FARSDAT*.

3.3 Evaluation metrics

In this study, the accuracy of the models on test sets is measured at the frame level. Frame recognition rate is computed based on the frame-wise comparison of the reference labels for each frame with model predictions.

4 Experiments and results

All our RC models are grounded on a reservoir of leaky-integrator neurons [39] with a few control parameters whose values determine the emergent behavior of the system in response to the features extracted from the input audio signals. Number of the input units is determined by the number of extracted features and the extent of consecutive frames that provide the phoneme recognition networks with the acoustic context of each input sequence. In this study, we extracted 18 features from the input frames and considered 14 consecutive frames for each sentence to present the RC models with $18 * 14$ input sequences at a time. That is, the recurrent neural network slides on the sequence of features extracted from each sentence and updates its states. At each time point, a sequence of 14 components from feature vectors are fed to the network that has to recognize the phoneme label of the central vector. All models comprise 35 output units representing 35 Persian phonemes.

We implemented our models using the commercial toolbox MATLAB 2017b (computer hardware: Intel Core i7 (2.7 GHz), 16 GB RAM, NVIDIA GeForce GT 650M).

4.1 Shallow RC

In our experiments on shallow RC, we gradually increased the size of the reservoir to examine the role of this parameter in the performance of the network. The activation function of the reservoir units and the output layer were considered as tangent hyperbolic and identity functions, respectively. We calculated the readout weights through the pseudo inverse method. Both input-to-reservoir and reservoir internal weights were sampled from a uniform distribution. Extensive grid searches were conducted to optimize hyper-parameters of the model. As a result, we set spectral radius, $\rho = 0.3$ and the leakage rate, $a = 0.5$. We also sampled input-to-reservoir weights from a uniform distribution with real values between $[-0.1, 0.1]$. Since the initial values for the input-to-reservoir and reservoir internal weights are random, for each reservoir hyper-parameter, the training and testing of the RC were repeated 5 times and the average results were reported (see Fig. 2).

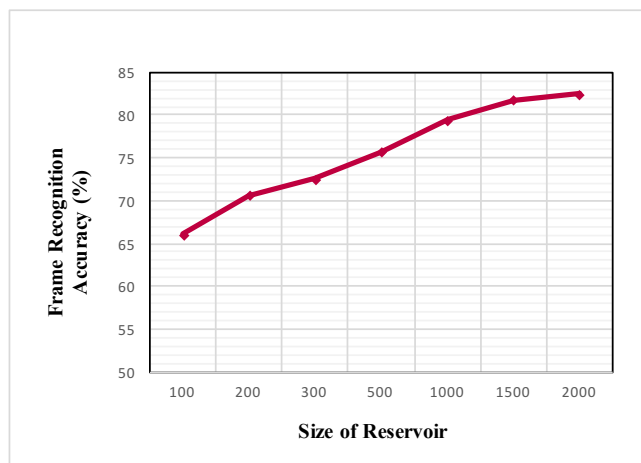


Figure 2: Effect of the reservoir size in the performance of the shallow RC as the speech recognition model. The models were trained on speech signal of 56 speakers and validated on audio signals recorded from 14 speakers.

As expected, increasing the size of the hidden layer enhances the performance such that the largest network with 2000 neurons achieved 82.52% frame recognition accuracy.

4.2 Deep RC

Afterward, we investigated the effect of stacking hidden layers in frame recognition accuracy. To this end, different deep RC structures with 3, 5, 7 and 10 hidden layers were implemented. For the sake of simplicity, we used the same reservoir size and the hyper parameter values for all hidden layers. Again through extensive grid search, we optimized spectral radius and the leakage rate for each structure. Fig. 3 depicts frame recognition rates of different deep RC structures with varying reservoir sizes. Noticeably, the best accuracy, 83.38%, is obtained from a 3-layer deep ESN with 2000 neurons in each layer, showing the potential overfitting effect after stacking more layers. In the case of 5, 7 and 10 hidden layers, after the reservoir size exceeds 1000 neurons, the recognition rate declines. Therefore, we didn't report the frame recognition rate of the larger reservoir sizes.

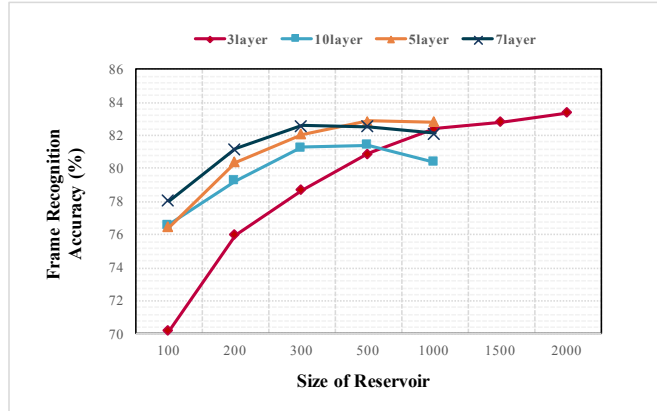


Figure 3: Effect of the reservoir size in the performance of the deep ESN as the speech recognition model. The models were trained on speech signal of 56 speakers and validated on audio signals recorded from 14 speakers.

4.3 Heterogeneous RC

Our experiments on standard shallow and deep RC models suggest that organizing the model in a hierarchical structure with three layers only enhances the frame recognition rate by 0.27%. Therefore, we, firstly, sought to see if introducing heterogeneity to shallow architecture can show comparable impact. Then, we investigated the effect of layer variability in deep RC structures. As reported in Table. 1 presence of multiple time-scales prolongs the internal memory and further improves the performance of the network in speech recognition. In fact, with only 2000 neurons, it outperforms the performance of the deep ESN with 6000 internal units.

In this study, we only considered three sub-groups with $\tau_i \in \{1, 3, 5\}$ and studied the effect of number of nodes in each population on the performance. Our experiments showed that, the best result (reported in Table. 1) is obtained when all ensembles had identical sizes. The underlying rationale for the choice of $\tau_i \in \{1, 3, 5\}$ was that on the one hand, introducing longer time lags enables RNNs to incorporate extra temporal context which leads to better frame recognition scores. On the other hand, however, RNNs seem to only benefit from having access to a limited range of previous contextual information in framewise phoneme classification (e.g., see Fig.3 in [43]). Therefore, we pre-set the lags and optimized the size of each population.

We also explored heterogeneous deep ESN networks where intermediate layers have non-identical temporal scales. We assigned $\tau_i = \{1, 3, 5\}$ to superficial, intermediate and deep layers, respectively. Again, we studied networks with various identical and non-identical layers and noticed the best performance is obtained with a 3 layer network where each layer comprises 2000 internal nodes. As shown in Fig. 4 and reported in Table. 1, the proposed heterogeneity improves recognition accuracy of the deep ESN by approximately 1.25% on small dataset.

Conducting preliminary experiments on the *complete FARSDAT* database, consistency of these findings were confirmed (see Table. 1). However, in this study, we only tested RC models with 500 neurons.

4.4 LSTM

We compared RC speech recognition performance with an LSTM network with three hidden layer of 100 LSTM cells and a Softmax readout layer. Similar to RC models, we presented the models with sequences of feature vectors extracted from consecutive frames representing spoken sentences. This network was trained using the stochastic gradient descent

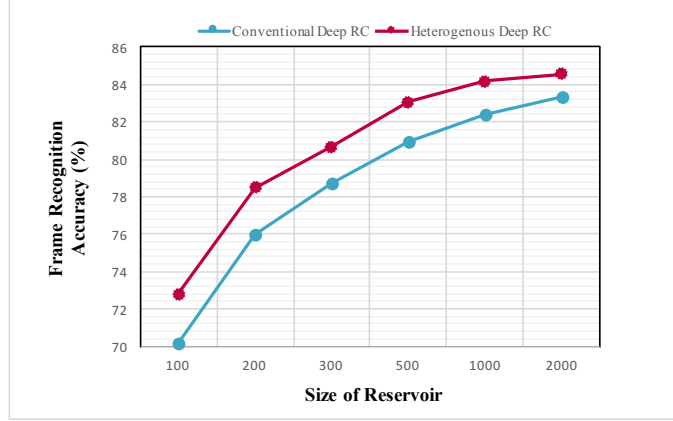


Figure 4: Effect of the proposed heterogeneity in the performance of deep ESN models. The models were trained on speech signal of 56 speakers and validated on audio signals recorded from 14 speakers.

Table 1: Performance evaluation of the reservoir computing models in terms of frame recognition accuracy. Test sets are both small and complete FARSDAT datasets.

	small FARSDAT	complete FARSDAT
Shallow ESN	82.05%	67.66%
Deep ESN	82.32%	71.86%
Heterogeneous shallow ESN	82.72%	68.74%
Heterogeneous deep ESN	83.57%	72.51%

(SGD) with Momentum method and the learning rate of 0.0001. On the small and complete FARSDAR dataset, the mean frame recognition rates are 84.99% and 84.26%, respectively.

5 Discussion and conclusions

The present study explores the potentials of reservoir computing (RC) for Persian speech recognition as a common practice in ASR applications. The underlying rationale was that RC represents a computationally efficient RNN-based approach to learn temporal features. Our results suggest that the RC models perform in terms of frame recognition accuracy at least on par with the de-facto standard in speech processing, the LSTM. It should, however, be noted that the implemented RC models have a vastly fewer trainable parameters (approximately 70000 in the current study) than the LSTM (> 305500). In the current experiment, this led to a reduction of training time from 150 minutes for the LSTM to 90 minutes for the RC system. It should also be mentioned that LSTM training was performed on GPU and was already highly optimized for GPU usage, while the RC training was on CPU and was not optimized for parallel computing.

Moreover, introducing heterogeneity in terms of variable time-scales further improved recognition accuracy of both shallow and deep ESN up to 1.25%. On the complete FARSDAT dataset, however, we preliminary tested our RC models with only 500 neurons. Therefore, the performance is not comparable to LSTM. It remains to be shown that our models can be improved by further increasing the size of hidden layers. As a further extension to this study, the effect of additional number of heterogeneous sub-groups and different values of temporal delays in both shallow and deep ESN will be systematically explored. Besides, in this study, we only reported the performance in terms of frame recognition accuracy. It remains for future works to conduct recognition tasks at phoneme and word levels.

6 Acknowledgments

Fatemeh Hadeaghi's research was supported by the Deutsche Forschungsgemeinschaft, Germany (TRR 169/A2).

References

- [1] A. Ganapathiraju, J. E. Hamaker, and J. Picone, “Applications of support vector machines to speech recognition,” *IEEE transactions on signal processing*, vol. 52, no. 8, pp. 2348–2355, 2004.
- [2] B. J. Shannon and K. K. Paliwal, “A comparative study of filter bank spacing for speech recognition,” in *Microelectronic engineering research conference*, vol. 41. Citeseer, 2003, pp. 310–12.
- [3] M. Gales and S. Young, *The application of hidden Markov models in speech recognition*. Now Publishers Inc, 2008.
- [4] N. Morgan, “Deep and wide: Multiple layers in automatic speech recognition,” *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 7–13, 2011.
- [5] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.
- [6] D. C. de Andrade, S. Leo, M. L. D. S. Viana, and C. Bernkopf, “A neural attention model for speech command recognition,” *arXiv preprint arXiv:1808.08929*, 2018.
- [7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [8] D. Yu, L. Deng, and F. Seide, “The deep tensor neural network with applications to large vocabulary speech recognition,” *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 2, pp. 388–396, 2012.
- [9] H. Jaeger and H. Haas, “Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication,” *science*, 2004.
- [10] W. Maass, T. Natschläger, and H. Markram, “Real-time computing without stable states: A new framework for neural computation based on perturbations,” *Neural computation*, vol. 14, no. 11, pp. 2531–2560, 2002.
- [11] F. Hadaeghi, B.-P. Diercks, D. Schetelig, F. Damicelli, I. Wolf, and R. Werner, “Spatio-temporal feature learning with reservoir computing for T-cell segmentation in live-cell Ca^{2+} fluorescence microscopy,” *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.
- [12] F. Triefenbach, A. Jalalvand, B. Schrauwen, and J.-P. Martens, “Phoneme recognition with large hierarchical reservoirs,” *Advances in neural information processing systems*, vol. 23, 2010.
- [13] M. Lukoševičius, H. Jaeger, and B. Schrauwen, “Reservoir computing trends,” *KI-Künstliche Intelligenz*, vol. 26, no. 4, pp. 365–371, 2012.
- [14] F. Hadaeghi, “Reservoir computing models for patient-adaptable ECG monitoring in wearable devices,” *arXiv preprint arXiv:1907.09504*, 2019.
- [15] X. He, T. Liu, F. Hadaeghi, and H. Jaeger, “Reservoir transfer on analog neuromorphic hardware,” in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 1234–1238.
- [16] F. Hadaeghi, “Neuromorphic electronic systems for reservoir computing,” in *Reservoir Computing*. Springer, 2021, pp. 221–237.
- [17] M. Freiberger, A. Katumba, P. Bienstman, and J. Dambre, “On-chip passive photonic reservoir computing with integrated optical readout,” in *2017 IEEE International Conference on Rebooting Computing (ICRC)*. IEEE, 2017, pp. 1–4.
- [18] A. Goudarzi, M. R. Lakin, and D. Stefanovic, “DNA reservoir computing: a novel molecular computing approach,” in *International Workshop on DNA-Based Computers*. Springer, 2013, pp. 76–89.
- [19] F. Triefenbach, A. Jalalvand, K. Demuynck, and J.-P. Martens, “Acoustic modeling with hierarchical reservoirs,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2439–2450, 2013.
- [20] Q. Ma, L. Shen, and G. W. Cottrell, “Deepresn: A deep projection-encoding echo-state network,” *Information Sciences*, vol. 511, pp. 152–171, 2020.
- [21] C. Gallicchio and A. Micheli, “Deep echo state network (deepesn): A brief survey,” *arXiv preprint arXiv:1712.04323*, 2017.
- [22] C. Gallicchio, A. Micheli, and L. Pedrelli, “Deep reservoir computing: A critical experimental analysis,” *Neuro-computing*, vol. 268, pp. 87–99, 2017.

- [23] M. Bijankhan, J. Sheikhzadegan, and M. R. Roohani, "Farsdat-the speech database of farsi spoken language," in *Proceeding Australian Conference On Speech Science and Technology*. Proceeding Australian Conference On Speech Science and Technology, 1994.
- [24] F. Almasganj, S. A. Seyedsalehi, M. Bijankhan, H. Sameti, and J. Sheikhzadegan, "Shenava-1 a farsi spontaneous speech recognition system," in *Iranian Conference on Electrical Engineering (ICEE)*. Iranian Conference on Electrical Engineering (ICEE), 2001.
- [25] H. Sameti, H. Veisi, M. Bahrani, B. Babaali, and K. Hosseinzadeh, "Nevisa, a persian continuous speech recognition system," in *Computer Society of Iran Computer Conference*. Springer, 2008, pp. 485–492.
- [26] I. Nejadgholi and S. A. Seyedsalehi, "Nonlinear normalization of input patterns to speaker variability in speech recognition neural networks," *Neural computing and applications*, vol. 18, no. 1, pp. 45–55, 2009.
- [27] H. Veisi and A. Haji Mani, "Persian speech recognition using deep learning," *International Journal of Speech Technology*, vol. 23, no. 4, pp. 893–905, 2020.
- [28] S. G. Firooz, F. Almasganj, and Y. Shekofteh, "Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals," *Computers & Electrical Engineering*, vol. 58, pp. 215–226, 2017.
- [29] Z. Ansari and S. A. Seyedsalehi, "Toward growing modular deep neural networks for continuous speech recognition," *Neural Computing and Applications*, vol. 28, no. 1, pp. 1177–1196, 2017.
- [30] Z. Ansari, F. Almasganj, and S. J. Kabudian, "Rapid speaker adaptation based on combination of KPCA and latent variable model," *Circuits, Systems, and Signal Processing*, vol. 40, no. 8, pp. 3996–4017, 2021.
- [31] M. Asadolahzade Kermanshahi and M. Homayounpour, "Improving phoneme sequence recognition using phoneme duration information in DNN-HSMM," *Journal of AI and Data Mining*, vol. 7, no. 1, pp. 137–147, 2019.
- [32] S. Alisamir, S. M. Ahadi, and S. Seyedin, "An end-to-end deep learning model to recognize farsi speech from raw input," in *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE, 2018, pp. 1–5.
- [33] M. A. Kermanshahi, A. Akbari, and B. Nasersharif, "Transfer learning for end-to-end ASR to deal with low-resource problem in persian language," in *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*. IEEE, 2021, pp. 1–5.
- [34] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.
- [35] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [36] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.
- [37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [38] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," *Advances in neural information processing systems*, vol. 26, 2013.
- [39] H. Jaeger, M. Lukoševičius, D. Popovici, and U. Siewert, "Optimization and applications of echo state networks with leaky-integrator neurons," *Neural networks*, vol. 20, no. 3, pp. 335–352, 2007.
- [40] L. Manneschi, M. O. Ellis, G. Gigante, A. C. Lin, P. Del Giudice, and E. Vasilaki, "Exploiting multiple timescales in hierarchical echo state networks," *Frontiers in Applied Mathematics and Statistics*, vol. 6, p. 76, 2021.
- [41] L. Manneschi, A. C. Lin, and E. Vasilaki, "Sparce: Improved learning of reservoir computing systems through sparse representations," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [42] G. Tanaka, T. Matsumori, H. Yoshida, and K. Aihara, "Reservoir computing with diverse timescales for prediction of multiscale dynamics," *arXiv preprint arXiv:2108.09446*, 2021.
- [43] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.