Robust Real-Time Vowel Classification with an Echo State Network

Ted P. Schmidt[1]
ted@almende.org

Marco A. Wiering[2]
mwiering@ai.rug.nl

Anne C. van Rossum[1]
anne@almende.org

Ronald A.J. van Elburg[2]
r.van.elburg@ai.rug.nl

Tjeerd C. Andringa[2]
t.c.andringa@rug.nl

Bea Valkenier[2]
bea@ai.rug.nl

[1] Almende B.V., The Netherlands
[2] Artificial Intelligence Department, University of Groningen, The Netherlands

Abstract

In the field of reservoir computing echo state networks (ESNs) and liquid state machines (LSMs) are the most commonly used networks. Comparative studies on these reservoirs identify the LSM as the network that yields the highest performance for speech recognition. But LSMs are not always usable in a real-time setting due to the computational costs of a large reservoir with spiking neurons. In this paper a vowel classification system is presented which consists of an ESN which processes cochlear filtered audio. The performance of the system is tested on a vowel classification task using different signal-to-noise ratios (SNRs). The usefulness of this method is measured by comparing it to formant based vowel classification systems. Results show that this ESN based system can get a performance similar to formant based vowel classification systems on the clean dataset with only a small reservoir and even outperforms these methods on the noisy dataset.

1 Introduction

The Echo State Network (ESN) [5] is one of the well known recurrent neural networks (RNN) used in reservoir computing. RNNs have the ability to model highly non-linear systems, and are capable of processing temporal information. The hard part of using RNNs is training the network. Three different types of RNNs have been described to overcome this problem, Echo State Networks [5], Liquid State Machines (LSMs) [8], and Back Propagation Decorrelation (BPDC) [12]. With reservoir computing a randomly connected RNN is used as a reservoir that is not trained but read out by a simple classification layer. The reservoir has the function of a kernel, it projects the input to a higher-dimensional space in which it is better separable. The advantage of a reservoir in comparison to kernel-based methods (e.g. SVM) is the ability to incorporate temporal information.

Reservoir computing has been successfully implemented in several application domains. It has for example been used in dynamic pattern classification, tone generation, object tracking and prediction, reinforcement learning, and also Digital Signal Processing (DSP). For an overview read [10]. Reservoir computing also has successful implementations in speech recognition [9, 11, 16].

The primary goal for developing an LSM was to provide a biologically plausible paradigm for computations in generic cortical micro-circuits, while ESNs have been designed for high performance engineering tasks. LSMs therefore consist of biologically inspired spiking neurons with a small world interconnectivity pattern. Descriptions of ESNs can be found with analogue neurons and several different interconnection structures. Verstraeten et al. [15] compared reservoirs using different node types, for a broad range of parameter settings and tasks. They concluded that the computational cost of a spiking reservoir is higher but the performance was better on a speech recognition task of isolated digit recognition. They also showed that the memory capacity of both, spiking and analogue, reservoirs increases monotonically with the size, and found a strong dependence on the spectral radius for analogue neurons.

Because of the computational constraints of a real-time speech recognition system, experiments will be conducted with an ESN to test whether it is capable of obtaining a performance which is comparable to formant based vowel classification systems. In this paper a comparison will be made with formant based vowel classification systems [13, 2]. For this an experiment will be conducted where vowels from the Hillenbrand American English Vowels (AEV) dataset [3] need to be recognized under signal-noise-ratios (SNR) ranging from 30dB to -14dB with pink noise.

Contributions of this paper. (1) In this paper ESNs are compared to formant based vowel classification systems on the Hillenbrand vowel dataset. (2) The influence of different amounts of noise on these methods is experimentally analysed, and it is shown that ESNs are much more robust to the influence of noise than current formant based vowel classification systems. (3) It is shown that larger ESN topologies lead to significant performance increases.

In the next section the implementation of the ESN will be described, followed by the vowel classification system that consists of a cochlear filter, an ESN, and a linear classifier.

2 Method

A generic reservoir computing architecture is shown in figure 1. During reservoir simulation the reservoir states and output states with teacher forcing are computed with the following equations:

$$\mathbf{x}(t+1) = f(W_{res}^{res}\mathbf{x}(t) + W_{res}^{inp}\mathbf{u}(t) + W_{res}^{out}\mathbf{y}(t) + W_{res}^{bias}) \tag{1}$$

$$\hat{y}(t+1) = W_{out}^{res}\mathbf{x}(t+1) + W_{out}^{inp}\mathbf{u}(t) + W_{out}^{out}\mathbf{y}(t) + W_{out}^{bias}, \tag{2}$$

where $\mathbf{x}(t)$ is the reservoir neuron state vector for time step t, $\mathbf{u}(t)$ the input vector, f the neuron activation function, $\mathbf{y}(t)$ the teacher input vector, and $\hat{\mathbf{y}}(t)$ the state of the output neurons.

All the connection weights are randomly generated using some kind of distribution of connectivity and connection type, except for W_{out}^{res} which are obtained through learning.

2.1 ESN Implementation

The main architecture of the ESN follows the generic reservoir design. The parameters used for the reservoir are based on work of Jaeger [6] and Venayagamoorthy [14] for the general ESN working, Holzmann [4] for the combination of an ESN with audio processing, and Verstraeten et al. [16] for the overall architecture for speech recognition with a reservoir. The main difference with this approach and that of Verstraeten et al. [16] is the use of an ESN instead of an LSM to obtain a system which can be used in real-time. Because an ESN consists of analog neurons this approach does not need a filter encoding scheme to transform the analog outputs to spike trains as in [16].

The connection weights are generated using a connectivity parameter. With a certain connectivity the connections between the neurons are generated with weight values between -1 and 1. The reservoir and output neurons can be built from different types of neurons, they can have sigmoid or linear activation functions and can optionally be leaky integrator neurons. The state of a reservoir neuron is calculated using the following equation:

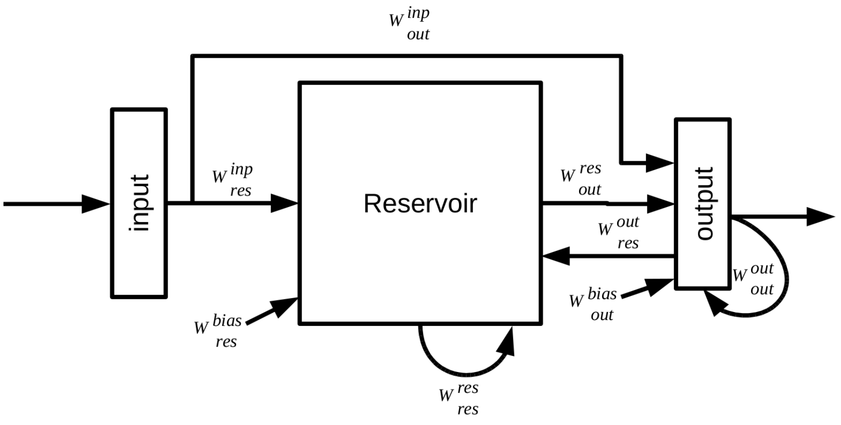

Figure 1: General reservoir computing architecture [10]. The following connection weight vectors are labelled in the figure W^{inp}_{res}: input to reservoir, W^{inp}_{out}: input to output, W^{res}_{res}: reservoir to reservoir, W^{bias}_{res}: a bias value to the reservoir, W^{res}_{out}: reservoir to output, W^{out}_{res}: output to reservoir, W^{out}_{out}: output to output, W^{bias}_{out}: bias value to the output.

$$\mathbf{x}(t+1) = f(W^{res}_{res}\mathbf{x}(t) + W^{inp}_{res}\mathbf{u}(t) + W^{out}_{res}\mathbf{y}(t)) \tag{3}$$

The implemented state update equation is different from the generic update equation, because no teacher forcing and feedback connections are used when simulating the reservoir.
An important property of the reservoir is the spectral radius. The spectral radius is the largest absolute value of the eigenvalues of the reservoir weight matrix. To obtain the echo state property, the network must be on the edge of stability, without constraints on the input this is obtained with a spectral radius between 0 and 1. A spectral radius of 0.8 seems to give a good performance for a variety of tasks [14]. The reservoir weights are first normalized using the largest eigenvalue of the reservoir, and subsequently scaled using the desired spectral radius.

2.2 Sound processing

An ESN can be used to classify complex sound patterns because after transfer of the sound patterns to the higher dimensional space constituted by the reservoir, the sound patterns are more likely to be linearly separable. Verstraeten et al. [16] proved that cochlear filtered sound is a good representation to use for sound recognition with an LSM. As in [16] a cochlear filter based on the widely used Lyon Passive Ear model [7] is used to filter the audio signal. This ear model describes the propagation of sound in the inner ear and the conversion of the acoustical energy into neural representations. The cochlea has a strong compressive non-linearity over a wide range of sound intensities. This model unlike many other cochlear models takes the non-linearity into account and explicitly recognizes the purpose of the strong non-linearity as an automatic gain control (AGC) that serves to map a huge dynamic range of physical stimuli into the limited dynamic range of nerve firings [7].

The cochlear data is fed into the reservoir by aligning the channels (frequencies) of the cochlear data to the input of the ESN. Through random connections from the input neurons to the reservoir certain channels are chosen as input for the reservoir. Each time step one frame of the cochlear data is entered into the ESN, which changes the states of the reservoir neurons that are connected to the input neurons. The state of other reservoir neurons that are not connected to the input neurons are changed due to the recurrent connections, no teacher input or feedback connections are used. Based on the states of all the neurons, patterns can be recognized with a readout function.

2.3 Classification

The readout function used to classify the states is the simple linear classification method, ridge regression (Tikhonov regularization). The weights of the readout neurons are calculated as followed:

$$W^{out} = (R + \alpha^2 I)^{-1} P, \tag{4}$$

where $R = S'S$ is the correlation matrix of the extended reservoir states $S = (X + U)$, α^2 is the smoothing factor, I the identity matrix, R^{-1} denotes the inverse of the matrix R, and $P = S'D$ is the cross-correlation matrix of the states S and the desired output D, which is obtained using Fisher labelling [1].

The actual classification is done using a linear projection of the input $\mathbf{u}(t)$ and reservoir states $\mathbf{x}(t)$ to the output $\mathbf{y}(t)$ using weights $\mathbf{w}$:

$$\mathbf{y}(t) = \mathbf{w} \cdot \mathbf{s}(t) \tag{5}$$

The winning class is determined using winner-take-all (WTA) selection, by taking the maximum value of the output **y**(t) over time. The classification performance was also tested using a winning class selection that computed the winner over a predefined amount of echo states, and a final class selection using WTA. But the performance of this method when using a set of echo states was not better than when computing the winning class over all the echo states of one input sample.

2.4 Experiment

The performance and robustness of the vowel classification system is tested on the AEV database [3] with different SNRs of pink noise. This dataset consists of 12 vowels with h-V-d syllables which are uttered by 45 men, 48 women, and 46 children. In the experiment the system is tested with several conditions. The robustness of the system is tested using pink noise with the SNRs 30dB, 20dB, 10dB, 0dB, -2dB, -4dB, -6dB, -14dB. Each of these datasets are tested on a reservoir of 12, 175 and 800 neurons. Pink noise was chosen because is masks speech evenly, and it occurs widely in natural physical systems. The performance of the system is also tested on the clean dataset with these reservoir sizes. The audio samples are first filtered using the cochlear model with a decimation factor of 40, and step factor of 0.28. The filtered audio is used as input for the reservoir using 74 input neurons and a fixed frame size of 115 frames of which the first 10 are dismissed due to the washout-time. To be able to detect the small differences in the values of the cochlear filtered input a large input scale of 10000 is used. The parameter values used for this experiment are optimized by hand, they are shown in table 1.

Table 1: System parameters

Cochlear parameters	
step factor	0.28
decimation factor	40
channel size	74
sample size (fixed)	115 frames

ESN parameters	
input connectivity	0.6
reservoir connectivity	0.5
feedback connectivity	0.0
reservoir activation function	tanh
output activation function	linear
spectral radius	0.8
input shift	0.0
reservoir shift	0.0
feedback shift	0.0
input scale	10000
reservoir scale	1.0
feedback scale	1.0
wash-out time	10 frames

A training and a testing phase exist for each experiment condition. Using 10-fold cross-validation a train and test set is created for each dataset which will then be used to train and test the system on for that particular condition. The conditions vary in speaker type (men, women, children), SNR level, and reservoir size.

3 Results

The results of the vowel classification experiments on the clean dataset are presented in table 2. The performance of the system is tested for three different reservoir sizes. The best performing reservoir is the largest with 800 neurons. It is able to obtain an overall vowel classification of 81.7% with the highest classification of 84.2% on the dataset with male speakers. The overall performance difference between the reservoir with 800 neurons and the reservoir with 175 neurons is only a bit smaller than between the reservoir with 175 neurons and the reservoir with 12 neurons, while the reservoir size difference is about 5 times and 15 times smaller respectively.

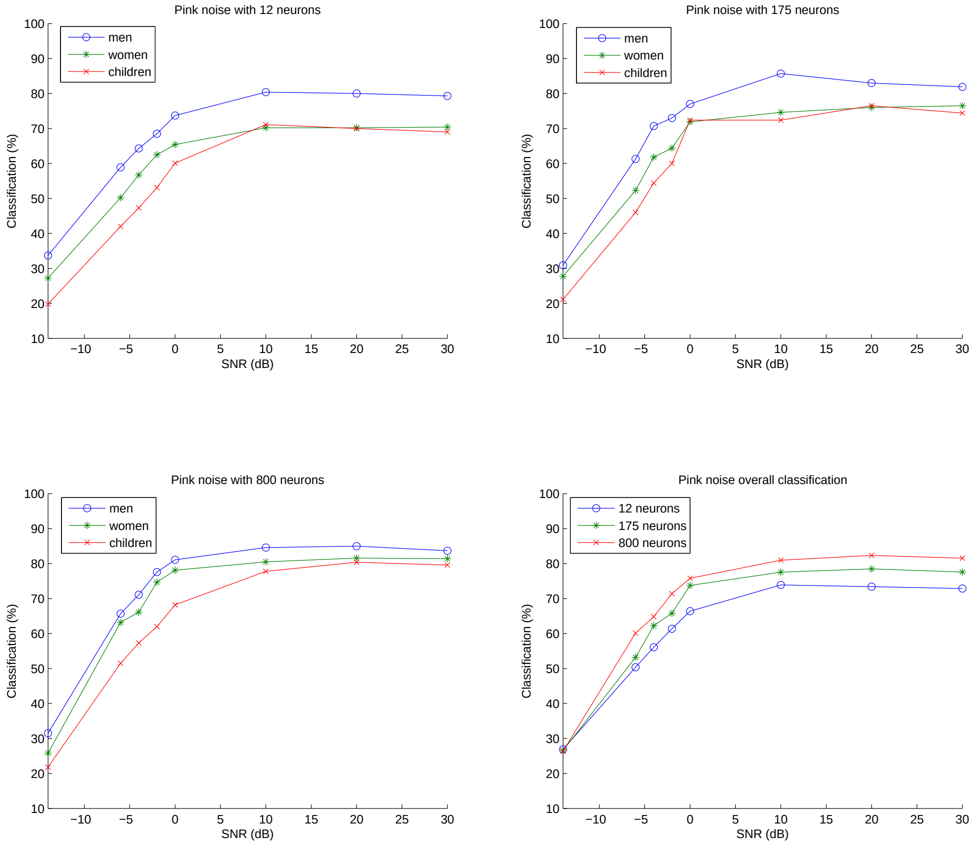

Figure 2: Classification performance on the datasets; Top left: pink noise and 12 neurons; Top right: pink noise and 175 neurons; Bottom left: pink noise and 800 neurons; Bottom right: the average overall performance on pink noise.

The results of the experiments on the dataset with different SNRs is presented in figure 2. The performance on the pink noise dataset with an SNR of -14dB was for all the reservoir sizes a reasonable 27%. The reservoirs with 175 and 800 neurons have nearly the same performance of 75% on this dataset for an SNR of 0dB, and for an SNR of 30dB the reservoir with 800 neurons has a performance of 82% which is 4% higher than the performance of the reservoir with 175 neurons.

Table 2: Performance (%) on the clean dataset

neurons	men	women	children	overall
800	84.2	80	81	81.7
175	83.7	77	75.8	78.8
12	81.1	70.7	70	73.9

In table 3 the confusion matrix of the classification for an ESN with 800 neurons and an SNR of -14dB is shown. The matrix shows the real classes (rows) with the classifications made by the system (columns). The values are the total classifications of the three speaker classes.

Table 3: Confusion matrix of the classification for an ESN with 800 neurons on the dataset with an SNR of -14dB

	ae	ah	aw	eh	ei	er	ih	iy	oa	oo	uh	uw
ae	17	17	18	42	3	2	1	9	14	1	14	0
ah	3	84	33	3	0	2	0	4	1	0	7	2
aw	6	33	77	7	0	0	0	4	2	0	9	0
eh	40	5	24	31	0	4	0	2	6	6	19	2
ei	4	11	3	7	18	5	21	10	15	17	1	27
er	7	5	8	6	19	19	14	7	14	16	7	17
ih	1	5	2	2	25	12	19	16	8	13	0	35
iy	3	9	3	3	5	1	9	74	0	2	1	29
oa	16	11	4	20	10	10	6	4	21	17	11	8
oo	3	9	3	5	18	16	14	11	18	13	5	23
uh	17	15	34	40	0	4	0	6	7	1	13	1
uw	1	6	4	3	8	3	20	36	1	2	2	51

4 Discussion

In this paper a method was described for vowel classification with an ESN. By processing cochlear filtered audio with an ESN, a robust and rather efficient vowel classification system is proposed. The performance of the system is tested on the vowel dataset AEV [3]. With several noise conditions and reservoir sizes the robustness and efficiency of the system is tested.

The results from the experiments on the clean dataset show that the system yields high performance with only a small reservoir of 12 neurons. Using a larger reservoir size of 800 neurons does increase the performance by 8% to 82%, averaged over all the three speaker classes: men, women and children. A performance increase together with reservoir size was expected since Verstraeten et al. [15] showed that the memory capacity of the reservoir monotonically increases with the reservoir size. The steepness of the performance increase due to the reservoir size is rather low. A reason for this could be due to the fact that some audio samples are even hard to classify for most human listeners [3].

The robustness against noise was tested for different SNRs with pink noise. As with the clean dataset the difference in performance between the reservoir with 175 and 800 neurons is rather small. What can be seen is that the system is very robust against noise, it yields good performance with an SNR of 0dB and from there on the performance increases gradually. The performance for the negative SNRs increases rather steep but an overall performance of 27% for an SNR of -14dB for all the reservoir sizes also shows the good robustness of the system.

The confusion matrix (table 3) gives some insight on the type of speech information that is still available for the ESN with an SNR of -14dB. The confusion matrix shows us that the error clusters still show a structure that can be related to the internal structure of vowels. The position of the first formant is called vowel height; a so called high vowel [i,u] has a relatively low first formant and a low vowel [a] has a high first formant. Since the first formant has the highest energy, it can be expected to be most noise robust. Visual inspection shows that also in pink noise the first formant stays visible longer than the other formants. Vowels that are close to each other in vowel height, for example iy [i], ih [I], and uw [u] seem to be confused more often than vowels more different on vowel height, for example ih [I] and ae [æ]. This indicates that this discriminable feature can still be exploited by the system.

4.1 Comparison

Comparing these results to the results of the formant based vowel classification systems [13, 2] on the same datasets shows that this ESN based vowel classification system is more robust against noise (table 4). The ESN was able to obtain the same classification results as in [13] on the clean dataset with a small reservoir. When comparing the noise robustness for pink noise this ESN based method obtained a classification performance that is 23% higher than in [13] for an SNR of 0dB averaged over the three speaker classes with a reservoir of 12 neurons. When comparing these results to the results from [2], this ESN method obtained a classification performance that is 15% higher than in [2] for an SNR of 0db averaged over the male and female speaker class.

Table 4: Comparison of performance (%) with the formant based methods of Valkenier et al. [13] and Wet et al. [2], for an SNR of 0dB. The results are averaged over the speaker classes.

classes	[13]	[2]	12 neurons	175 neurons	800 neurons
man, women	38	55	70	74	80
man, women, children	43	-	66	74	76

5 Conclusion

In this paper a vowel classification system based on reservoir computing was presented. In literature of reservoir computing [15] LSMs seem to outperform ESNs on speech recognition tasks but higher computational costs seems to be the toll. To test whether an ESN can be used for robust real-time vowel classification, a vowel classification experiment was conducted and compared to formant based vowel classification methods [13, 2]. This ESN based vowel classification system showed a good performance on the vowel classification task. To be able to use this system in a real-time setting the performance of the reservoir was also tested with a small reservoir size. This system was able to obtain good results on a clean dataset with a small reservoir, and outperformed methods found in the literature [13, 2] on noisy data. With these results it can be concluded that the ESN which is computationally cheaper than an LSM can be successfully used for robust real-time vowel classification.

References

[1] Antonelo, E., Schrauwen, B., Dutoit, X., Stroobandt, D., and Nuttin, M. Event detection and localization in mobile robot navigation using reservoir computing. *Lecture Notes in Computer Science 4669* (2007), 660.

[2] de Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., and Bourlard, H. Evaluation of formant-like features on an automatic vowel classification task. *The Journal of the Acoustical Society of America 116* (2004), 1781.

[3] Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. Acoustic characteristics of american english vowels. *Journal of the Acoustical Society of America 97*, 5 (1995), 3099–3111.

[4] Holzmann, G. Echo state networks in audio processing. *Internet Publication* (2007). Available from: http://grh.mur.at/sites/default/files/ESNinAudioProcessing.pdf.

[5] Jaeger, H. The ”echo state” approach to analysing and training recurrent neural networks. GMD Report 148, GMD - German National Research Institute for Computer Science, 2001. Available from: http://www.faculty.jacobs-university.de/hjaeger/pubs/EchoStatesTechRep.pdf.

[6] JAEGER, H. Tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the ”echo state network” approach. Tech. rep., Fraunhofer Institute AIS, St. Augustin-Germany, 2002.

[7] LYON, R. Automatic gain control in cochlear mechanics. *The Mechanics and Biophysics of Hearing* (1991).

[8] MAASS, W., NATSCHLAGER, T., AND MARKRAM, H. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput 14*, 11 (2002), 2531–60. Institute for Theoretical Computer Science, Technische Universitat Graz, A-8010 Graz, Austria. maass@igi.tu-graz.ac.at.

[9] MAASS, W., NATSCHLAGER, T., AND MARKRAM, H. A Model for Real-Time Computation in Generic Neural Microcircuits. In *NIPS 2002* (2003), Advances in Neural Information Processing Systems, MIT Press, pp. 229–236.

[10] SCHRAUWEN, B., VERSTRAETEN, D., AND CAMPENHOUT, J. V. An overview of reservoir computing: theory, applications and implementations. In *Proceedings of the 15th European Symposium on Artificial Neural Networks* (2007), pp. 471–482.

[11] SKOWRONSKI, M. D., AND HARRIS, J. G. Automatic speech recognition using a predictive echo state network classifier. *Neural Networks 20*, 3 (April 2007), 414–423. Available from: `http://dx.doi.org/10.1016/j.neunet.2007.04.006`.

[12] STEIL, J. J. Backpropagation-decorrelation: online recurrent learning with o(n) complexity. In *Proc. IJCNN* (Jul 2004), vol. 1, pp. 843–848.

[13] VALKENIER, B., KRIJNDERS, J., VAN ELBURG, R., AND ANDRINGA, T. Robust vowel detection. In *Proceedings of NAG/DAGA* (2009), vol. 303, pp. 1306–1309.

[14] VENAYAGAMOORTHY, G. K., AND SHISHIR, B. Effects of spectral radius and settling time in the performance of echo state networks. *Neural Networks 22*, 7 (2009), 861–863.

[15] VERSTRAETEN, D., SCHRAUWEN, B., D’HAENE, M., AND STROOBANDT, D. An experimental unification of reservoir computing methods. *Neural Networks 20*, 3 (2007), 391–403. Available from: `http://dblp.uni-trier.de/db/journals/nn/nn20.html#VerstraetenSDS07`.

[16] VERSTRATEN, D., SCHRAUWEN, B., STROOBANDT, D., AND VAN CAMPENHOUT, J. Isolated word recognition with the liquid state machine: a case study. *Information Processing Letters 95*, 6 (2005), 521–528.