Noise Robust Continuous Digit Recognition with Reservoir-based Acoustic Models

Azarakhsh Jalalvand, Kris Demuynck, Jean-Pierre Martens
Multimedia Lab, ELIS, Ghent University/iMinds, Sint-Pietersnieuwstraat 41, B-9000, Ghent, Belgium
Email: Azarakhsh.Jalalvand@ugent.be

***Abstract*—Notwithstanding the many years of research, more work is needed to create automatic speech recognition (ASR) systems with a close-to-human robustness against confounding factors such as ambient noise, channel distortion, etc. Whilst most work thus far focused on the improvement of ASR systems embedding Gaussian Mixture Models (GMM)s to compute the acoustic likelihoods in the states of a Hidden Markov Model (HMM), the present work focuses on the noise robustness of systems employing Reservoir Computing (RC) as an alternative acoustic modeling technique. Previous work already demonstrated good noise robustness for continuous digit recognition (CDR). The present paper investigates whether further progress can be achieved by driving reservoirs with noise-robust inputs that have been shown to raise the robustness of GMM-based systems, by introducing bi-directional reservoirs and by combining reservoirs with GMMs in a single system. Experiments on Aurora-2 demonstrate that it is indeed possible to raise the noise robustness without significantly increasing the system complexity.**

I. Introduction

Enhancing the noise robustness of automatic speech recognition (ASR) systems is still an active area of research. In this work we focus on robust continuous digit recognition (CDR). CDR is essential for the recognition of spoken numerical data (e.g. PIN-codes) in many applications which are often operated in a noisy environment and utilized by accented non-native speakers (e.g. tourists) as well as native speakers. The absence of a language model also makes CDR an attractive setup to evaluate the robustness of acoustic modeling techniques.

A modern ASR system treats CDR as a statistical pattern recognition problem which aims at finding the most likely interpretation of a stream of acoustic feature vectors generated by an acoustic front-end. The recognition is achieved by a back-end comprising one left-to-right multi-state Hidden Markov Model (HMM) per digit. In most systems, the likelihood to observe a given feature in a certain state is estimated by means of a Gaussian Mixture Model (GMM). Although state-of-the-art systems can reach high accuracy on low-noise test utterances, they are still susceptible to severe degradations in noisy conditions.

In recent years, many strategies for improving the noise robustness have been proposed [1]. Most of them deal with the speech signal preprocessing in the acoustic front-end and aim to retrieve acoustic features that represent the clean speech component of a noisy signal [2]–[4]. Other methods take the impact of the noise into account in a consistent way during the likelihood computation in a GMM-based back-end [5]. Finally, there have also been several attempts to improve robustness by means of alternative acoustic modeling techniques such as neural networks [6] or Support Vector Machines (SVM) [7].

In recent work, we investigated the potential of Reservoir Computing (RC) [8] as an alternative approach. Reservoir Computing employs reservoir networks – a particular kind of Recurrent Neural Networks (RNN) [9] – as complex dynamical systems that can analyze an incoming input vector stream. The hypothesis is that such a dynamical system can be designed to focus on the speech dynamics, and thus, to be less sensitive to the dynamics of the noise. We were already able to devise an RC-based CDR system that attains competitive recognition accuracies in clean conditions and that outperforms most other systems in noisy conditions [10], [11].

In this paper we extend this previous work. In particular, we investigate whether RC-based systems can profit from front-end techniques that were shown to work well in combination with traditional GMM acoustic models. Since reservoirs only build up some memory of the recent past, we also test bi-directional reservoir systems combining reservoirs that process the speech frames from left to right and from right to left respectively. Such bi-directional systems were already demonstrated to improve phone recognition in continuous speech [12], but they were not yet applied to CDR. Finally, we also study ways of combining reservoir networks and GMMs in a single system.

The rest of the paper is organized as follows: Section II provides a concise outline of the basic principles of RC, Section III describes ways of integrating reservoir networks in an RC-HMM hybrid speech recognizer, Section IV reviews three approaches that were tested for combining reservoir networks with GMMs in a single system and Sections V and VI summarize the experimental setup and the results obtained with it on the Aurora-2 benchmark for CDR. The paper ends with conclusion and future work.

II. Reservoir Computing (RC)

The basic principle of RC is that information can be retrieved from sequential inputs by means of a two-layer RNN with the following characteristics (see Fig. 1). The first layer is a sparsely connected hidden layer, composed of non-linear neurons which, at time t, are driven by inputs U_t and by delayed hidden layer outputs R_{t-1}. The output layer consists of linear neurons which are driven by the hidden layer outputs R_t. Important is that the weights of the hidden neurons are fixed, and only the weights of the output layer are optimized according to a least squares linear regression.

The hidden layer can be envisioned as a reservoir of recurrently interconnected computational neurons, driven by inputs. Together with the output layer it forms a *reservoir*

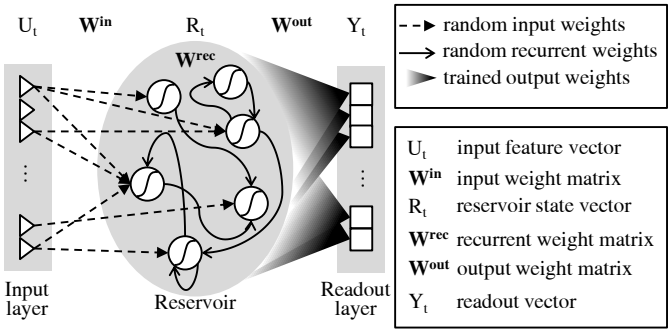


Fig. 1. A basic RC system consists of a reservoir and a readout layer. The reservoir is composed of interconnected **non-linear** neurons with randomly **fixed** weights. The readout layer consists of **linear** neurons with **trained** weights.

network. The network outputs Y_t are usually called *readouts* [8] so as to differentiate them unambiguously from the reservoir outputs R_t. In order to become less sensitive to random inter-frame changes in the inputs (e.g. changes due to the spectral analysis or the ambient noise), one can introduce leaky integration in the reservoir neurons (so-called Leaky Integrator Neurons [13]). The resulting reservoir network is governed by the following equations:

$$R_t = (1-\lambda)R_{t-1} + \lambda\, f_{res}(\mathbf{W}^{in}U_t + \mathbf{W}^{rec}R_{t-1}) \quad (1)$$
$$Y_t = \mathbf{W}^{out}R_t \quad (2)$$

with a leak rate λ between 0 and 1, with $\mathbf{W}^{in}$ and $\mathbf{W}^{rec}$ containing the input and recurrent weights to the reservoir neurons, and with $\mathbf{W}^{out}$ containing the weights of the output neurons.

As mentioned before, the weights of the hidden neurons are fixed. This is achieved by means of a random process characterized by four control parameters (see [14] for more details). These parameters are: (1) α_U, the maximal absolute eigenvalue of the input weight matrix $\mathbf{W}^{in}$, (2) ρ, the maximal absolute eigenvalue of the recurrent weight matrix $\mathbf{W}^{rec}$, (3) K^{in}, the number of inputs driving each reservoir neuron and (4) K^{rec}, the number of delayed reservoir outputs driving each reservoir neuron. The first two parameters control the strengths of the input and the recurrent stimulations of a reservoir neuron, the latter two control the sparsity of the input and recurrent weight matrices. Together with λ they constitute the reservoir control parameters which have to be properly set in order to assure the reservoir is well behaved. Note that any effective reservoir should at least have the so-called echo state property. It states that, with time, the reservoir should forget about the initial state it was in. That is also why a reservoir network was originally called an Echo State Network [8]. It was shown in [8] that the echo state property holds if ρ – called the spectral radius – is smaller than 1.

The reservoir can be envisioned as a predefined but complex non-linear dynamical system that performs a temporal analysis of the input stream. We claim that such a system can extract features that are not so easily corrupted by the presence of noise whose dynamics differ from the speech dynamics.

The output weights are determined so that they minimize the mean squared error between the readouts Y_t and the desired readouts D_t over the training examples [11]. As a consequence, they follow from a set of linear equations. The desired output D_t is a unit vector with a non-zero entry at the position corresponding to the desired HMM-state at time t.

III. Speech recognition with reservoirs

In this section we introduce the architectures that were conceived to perform CDR by means of a reservoir network.

A. A hybrid RC-HMM

Like any other neural network based hybrid system [15], a hybrid RC-HMM assumes that every network output corresponds to an HMM state. It transforms these outputs to state likelihoods and performs a standard Viterbi search for the best path through a looped HMM. So, the readouts $y_{t,i}$ (with i indexing the network outputs) are transformed to new outputs $z_{t,i} \approx P(y_{t,i}|q_t = i)/P(i)$. Using these outputs, one can determine the best state sequence as

$$\hat{q} = \arg\max_q P(q|y) = \arg\max_q \prod_{t=1}^{T} z_{t,q_t}\, P(q_t|q_{t-1}),$$

and derive the digit sequence thereof. The admissible state sequences can represent an arbitrary sequence of digits (possibly interleaved with silences). A transition probability P_0 is introduced on the transition from the final to the initial state that controls the balance between deletions and insertions.

The reservoir network can be a simple network with one reservoir, but it can as well be a hierarchical network, obtained by stacking multiple reservoir networks (called layers) on top of each other (see Fig. 2). The argument for cascading layers is that new layers can correct some of the mistakes made by the preceding layers because they offer additional temporal modeling capacity and a new inner space to model the state distributions. This argumentation is supported by experiments showing enhanced digit and phone recognition in continuous speech [11], [12]. The layers are trained one after the other and per layer good settings of the reservoir control parameters emerge from an efficient user-controlled search procedure (see [12]).

The readouts of each layer represent the same set of HMM states and their weights are trained to minimize the mean squared differences between the computed readouts Y_t and the desired readouts D_t. Under these circumstances, the readouts are assumed to adhere to posterior probabilities, meaning that $P(q_t = i|y_{t,i}) \approx y_{t,i}$ and $z_{t,i} \approx y_{t,i}/P(q_t = i)$. However, since $y_{t,i}$ is not confined to [0,1] and since likelihoods must remain positive, one has to map $y_{t,i}$ to a variable that is confined to [0,1] and that can be embedded in the formula for $z_{t,i}$. This mapping can be accomplished by e.g. a sigmoid function or a non-parametric function. These functions can be optimized to approximate the true posterior $P(q_t = i|y_{t,i})$ as it emerges from two histograms of $y_{t,i}$: one over the frames assigned to state i and one over all frames [11]. However, it was experimentally verified that

$$z_{t,i} = \frac{\max(y_{t,i}, y_o)}{\max_j(y_{t,j})}\,\frac{1}{P(q_t = i)}, \quad y_o \ll 1 \quad (3)$$

leads to basically the same results, and therefore, we opted for this simple clip-and-scale approach here.

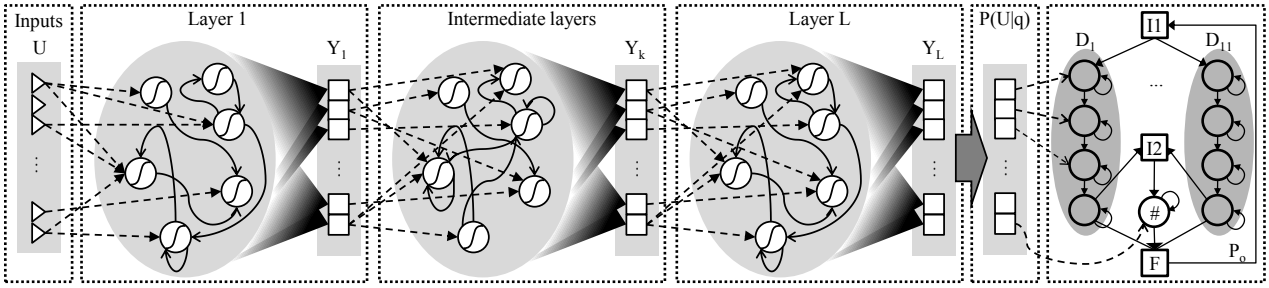


Fig. 2. Architecture of an RC-HMM hybrid comprising a multi-layer reservoir network for CDR. The HMM has two initial states (I1 and I2), one final state (F) and it comprises 11 multi-state digit models (D1 ... D11) and a single state silence model (#)

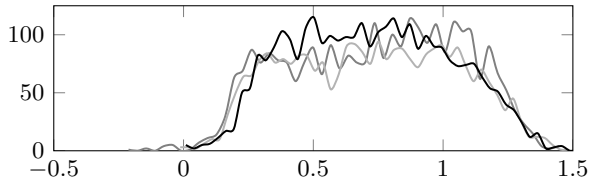


Fig. 3. Histograms of three randomly selected readouts on the frames that were assigned to the their corresponding states by the Viterbi search.

B. Bi-directional RC-HMM

Reservoirs only provide a fading memory of the past; they make no use of the future. On the other hand, the theory of co-articulation states that a phone is also influenced by the forthcoming phone. In order to account for such anticipation as well, we introduce bi-directional reservoir networks. Such a network is composed of two identical reservoirs, one that processes the data stream from left-to-right, the other that processes them from right-to-left. However, there is a single output layer: the readouts at time t are computed as a linear combination of the outputs of both reservoirs [12].

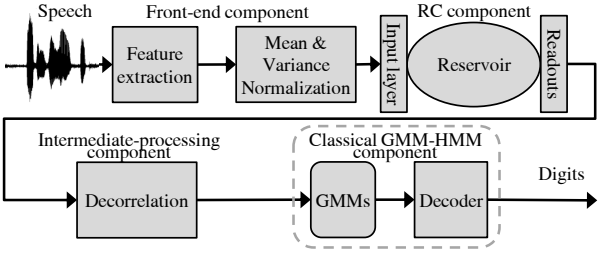
IV. Model combination

In this section we advocate two approaches for combining RC-based likelihoods with GMM-based likelihoods. The former are computed in a large and randomly fixed high-dimensional feature space that is affected by long-term dynamics, the latter are computed in a well conditioned low-dimensional feature space that solely describes local dynamics.

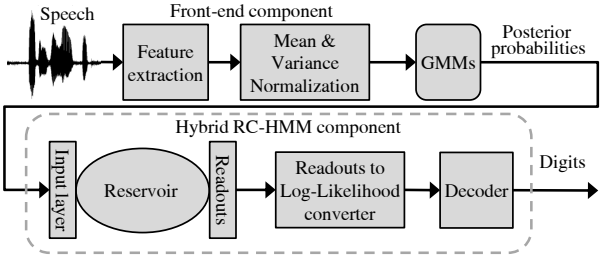
A. An RC-GMM tandem (T-RC-GMM)

In an RC-GMM tandem, the readouts Y_t are supplied as acoustic features to a traditional GMM-HMM system. In an MLP-GMM tandem [16] (MLP stands for Multi-Layer Perceptron), one usually considers a non-linear transformation of the Y_t followed by a dimensionality reduction and decorrelation to obtain GMM inputs that qualify better for being modeled by means of mixture of gaussian distributions with diagonal covariance matrices.

Typical non-linear transformations employed on MLP-outputs are a logarithm or an inverse sigmoid. Their aim is to reduce the skewness of the MLP output distributions. However, reservoir network outputs are linear combinations of zero mean reservoir state variables. Hence they are not hard-limited to the range of [0,1] and therefore, they do not exhibit skewed



(a) An RC-GMM tandem architecture. The reservoir network outputs are decorrelated before they are supplied to the GMM-HMM component.



(b) A GMM-RC tandem architecture. The GMM outputs can be supplied as such to the RC-HMM component.

Fig. 4. An RC-HMM tandem architecture: the front-end, the Reservoir Computing component, the intermediate-processing of the readouts and the GMM-based decoder.

distributions, as demonstrated by the histograms depicted in Fig. 3. Nonetheless, we did experiment with non-linear transformations, but they added nothing. This means that the T-RC-GMM architecture can be reduced to the scheme depicted in Fig. 4(a). The dimensionality reduction and decorrelation is achieved by means of a Mutual Information Discriminative Analysis (MIDA) [17], a technique that can be regarded as a special form of Linear Discriminant Analysis (LDA).

A variation on the proposed tandem is one in which the GMM-HMM component is supplied with a combination of the original acoustic vectors and the reservoir readouts. Like [18], we conjecture that in such a case, dimensionality reduction is inevitable to control the number of free parameters of the GMMs.

B. A GMM-RC tandem (T-GMM-RC)

Another type of tandem is a GMM-RC tandem in which the likelihoods computed by the GMMs are supplied to an RC-

HMM back-end. Since RC does not make any assumptions regarding the distributions of the individual inputs nor about the correlations between these inputs, the GMM outputs can be supplied to the RC-HMM component without any transformation or decorrelation, as shown in Fig. 4(b).

C. Likelihood fusion (F-GMM-RC)

Since reservoir-based likelihoods are computed in a high-dimensional feature space which was randomly fixed and designed to expose long-term memory effects, they may differ considerably from GMM likelihoods that are computed in a low-dimensional space of well established local features. Consequently, it seems sensible to fuse those likelihoods in the Viterbi search.

If the two likelihood sets apply to the same HMM states (digit states + silence state) the fusion is straightforward to achieve by considering a weighted mean of the two log-likelihoods as the state log-likelihood to control the Viterbi search. Obviously, such a state-based combination scheme assumes that the composing acoustic models are kind of time synchronous, meaning that they support state transitions at the same time instances. This may not be entirely true but it drastically simplifies the decoding. In our experiments, we pursued time synchrony by imposing the state-level segmentation provided by the reservoir during GMM training.

There are two popular ways of computing a mean likelihood. One is to compute a (weighted) linear combination of likelihoods, another is to compute a (weighted) log-linear combination of the likelihoods. The former strategy is believed to be ideal for reducing the effects of noise on the likelihoods, the latter is believed to be preferable for combining complementary information streams as it complies better with the log-linear combination of likelihoods across frames. As we contemplate that the two likelihoods attribute complementary information, we opt for the log-linear combination approach. For simplicity, we consider just one weight, irrespective of the state. The value of this so-called stream weight is determined from recognition experiments on the development data.

V. Experimental setup

In this section we present the experimental framework that was adopted to test the proposed approaches.

A. Speech corpus

All experiments are conducted on the Aurora-2 database [19]. This database contains clean and noisy utterances, sampled at 8 kHz and filtered with either a G712 or a MIRS filter. There are 8440 clean training samples, each counting 1 to 7 digits. The corpus also includes various noise corrupted copies of each clean utterance. The noisy utterances were created by artificially adding different noise types in different degrees, leading to Signal-to-Noise Ratios (SNR) between 20 and -5dB. The vocabulary consists of the digits 0 to 9 and the letter 'oh' (a substitute for 'zero'). We adhere to the training and test sets that were defined as part of the Aurora-2 benchmark. We have trained systems on clean speech (= clean speech training) and on clean + noisy speech (= multi-style training) and tested them on the test sets A - C.

B. Front-end setups

We investigate three acoustic feature sets: MFCCs (log energy and $c_1 \dots c_{12}$), 24 log Mel filterbank features (MelFB), and the ETSI Advanced Front-End features (AFE) (denoised $c_0 \dots c_{12}$ without dropping non-speech frames) [2]. In all cases, the analysis is performed on 30 ms Hamming-windowed frames and the hop size between frames is $\tau_{fr} = 10$ ms. In order to provide some context information, each feature set is supplemented with Δ and $\Delta\Delta$ features. The frame-wise feature extraction is followed by an utterance-wise normalization that creates zero-mean and unit-variance features.

C. Reservoir component setup

Based on previous work, the reservoir control parameters were determined in the same way for each layer. Defining $\tau_\lambda \doteq -\tau_{fr}/\ln(1-\lambda)$ as the leaky integration time constant and T as the expected state duration, we select $(\rho, \tau_\lambda, K^{in}, K^{rec}) = (0.8, T, 10, 10)$. The parameter α_U is chosen so that the average variance of the resrevoir outputs reach a certain level [14]. Since layers 2 and 3 see basically the same inputs, α_U is taken the same for both layers.

The size of the reservoir – defined as the number of neurons it contains (N^{res}) – is considered to be an independent variable. Note that since K^{in} and K^{rec} are kept fixed to 10, the CPU-time needed for calculating the readouts scales linearly with the size of the reservoir.

The reservoir networks are trained by means of a Tikhonov regression [20]. When comparing bi-directional to uni-directional systems we maintain the number of trainable parameters. This implies that a bi-directional system contains two reservoirs of half the size of the reservoir embedded in the uni-directional system it is compared to.

Each digit is modeled by a 7-state left-to-right HMM whilst the silence is modeled by a single state.

D. Model combination setup

In order to investigate the proposed model combination strategies, we needed a GMM-based component incorporating states that directly map to the reservoir network outputs. It was created with the SPRAAK toolkit[1], meaning that it works with semi-continuous HMMs, that is, all GMMs select members from the same global pool of Gaussians that emerged from an unsupervised clustering procedure. Consequently, there is an extensive parameter tying which is beneficial for small databases such as Aurora-2. The number of Gaussians and the number of mixture per state are determined automatically from the size and the statistics of the data, so that the risk of over-fitting is low.

E. Evaluation setup

In the development phase, two thirds of the training set are used for training, the held out third is used for control parameter optimization (e.g. the transition probability P_0). In the final evaluation phase the acoustic models are trained on the complete training set but using the control parameters that

[1] SPRAAK: Speech Processing, Recognition and Automatic Annotation Kit [http://www.spraak.org]

were optimal in the development phase. In this paper, we only report the results of the final evaluation experiments.

We report average Word Error Rates (WERs) on tests A-C for all SNRs, and we consider both clean speech training and multi-style training. In multi-style training the training set consists of clean utterances and utterances with SNRs between 20 and 5dB.

VI. Experimental results

In this section we review the results obtained with the proposed approaches and we compare them to reference results published in the literature.

A. Reference systems

First of all, we report some state-of-the-art reference system performances (see Table I). In particular, we consider the ML-based GMM systems using AFE-features proposed in [21], the ML-based and MCE-based GMM systems proposed in [3] and [4], two GMM systems embedding more sophisticated back-ends based on joint uncertainty decoding (JUD) and Vector Tylor Series (VTS) respectively [5] and the tandem system embedding deep belief networks reported in [6]. The figures show that the best reference systems incorporating a standard back-end are the ones employing the AFE features, be it that in multi-style training the impact of the front-end (compare AFE with MVN) is low. The figures further show that advanced back-end techniques (JUD and VTS) lead to a significant gain in noise robustness, but it is not clear how they affect the results for clean speech.

B. Self-developed GMM systems

Since we want to compare the effects of the front-end in GMM-based and RC-based systems and since we aim to combine the two systems in one recognizer we also developed a number of additional GMM systems with the SPRAAK toolkit. The performances of these systems are listed in the second section of Table I. Compared to the reference GMM (AFE) system the self-developed systems exhibit a much better clean speech performance, because the comprehensive parameters tying used by SPRAAK allows it to make detailed models even on small databases like Aurora-2. However, this is at the expense of lower noise robustness in the clean speech training case. This might also be the reason why it completely fails when using Mel-filter bank spectra as its inputs.

C. Impact of the front-end in RC-HMM hybrids

In a first experiment we test three acoustic feature sets in combination with RC-HMM hybrids incorporating a three-layer reservoir network, with each layer embedding a reservoir of 8K neurons. The results in Table I show that for clean speech training the AFE features lead to significant improvements in noise robustness. The differences are significant in moderately mismatched conditions of 0-20dB (from 12.3% to 10.0% WER) and substantial in the strongly mismatched condition of -5dB (from 73.9% to 58.4% WER). In the case of multi-style training, the impact of the front-end is much more modest, as it was for the GMM-based systems. Like in [12], we also tested larger reservoirs (up to 32K nodes) and more layers, but they yield only a small extra gain in performance for a substantial increase of the computational load.

TABLE I. Comparing average WERs (in %) per condition for test sets A-C of Aurora-2 using a 3-layer hybrid RC-HMM for both clean and multi-style training.

System	Clean	Clean 0-20	Clean -5dB	Multi	Multi 0-20	Multi -5dB
GMM (AFE) [21]	0.77	13.2	69.9	0.83	8.4	59.2
GMM (MVN) [3]	0.84	19.7	82.2	1.77	8.5	59.1
GMM (MVN-MCE) [4]	0.41	15.7	77.2	0.92	6.4	55.3
GMM (VTS) [5]	-	9.4	-	-	-	-
GMM (JUD) [5]	-	10.3	-	-	-	-
T-DBN-GMM [6]	1.26	21.0	74.6	-	-	-
GMM (MelFB-SPRK)	0.24	51.2	92.9	0.39	7.1	58.2
GMM (MVN-SPRK)	0.24	20.7	81.1	0.59	8.3	66.3
GMM (AFE-SPRK)	**0.20**	15.5	74.2	**0.39**	6.2	54.2
RC (MelFB)	0.59	12.3	73.9	0.98	6.1	51.1
RC (MVN)	0.78	11.8	63.5	1.28	7.1	52.0
RC (AFE)	0.82	10.0	58.4	1.31	6.2	47.5
biRC (MelFB)	0.75	10.9	64.0	1.07	**5.5**	45.5
biRC (MVN)	0.96	11.1	60.5	1.47	6.3	47.1
biRC (AFE)	0.86	**9.0**	**54.4**	1.43	5.8	**43.3**
T-GMM-biRC (AFE)	0.87	16.2	73.1	1.28	7.3	54.0
T-biRC-GMM (AFE)	0.77	10.6	58.7	1.19	5.7	43.9
F-GMM-biRC (AFE)	0.53	10.8	63.8	0.80	5.4	46.6

It is clear that the noise robustness of RC-HMM hybrids is better than that of GMM-based systems with a traditional back-end, but that they cannot compete with the self-developed GMMs on clean speech utterances. An important finding is that with clean speech training an RC-HMM with a simple back-end can compete with a GMM system incorporating a much more complex VTS-based back-end.

D. Effect of bi-directional processing in RC-based systems

In a second experiment, we test three-layer bi-directional reservoir systems with two 4K-node reservoirs per layer (biRC-HMM). The results in Table I show that bi-directional processing offers extra noise robustness at the expense of a small loss in clean speech performance. The bi-directional system now competes well with a GMM system with a VTS back-end.

Comparing the average WERs for the 0-20dB conditions shows that a bi-directional system yields around 10% and 6% relative improvement (on clean and multi-style training) over a unidirectional system, without much changing the complexity of that system. Table II provides the WER of such a 3-layer bi-directional RC-HMM per test set and per SNR.

E. Effect of model combination

In a third experiment we investigate the effect of the proposed model combination techniques on the system performance and the noise robustness. We employed our best system that utilizes the AFE as the front-end and a bi-directional reservoir as the RC-component (biRC-AFE).

In the case of tandem systems, feeding the GMM likelihoods into a reservoir system is apparently not a good idea. Supplying the reservoir network outputs to a GMM system does not hurt, but it does not help either. Likelihood fusion on the other hand can help to improve the clean speech recognition performance while maintaining most of

TABLE II. WERs (IN %) FOR TEST SETS A - C OBTAINED WITH A 3-LAYER BI-DIRECTIONAL HYBRID RC-HMM AND THE AFE FEATURES.

	Set	Clean	20	15	10	5	0	-5	0-20dB
Clean	A	0.82	1.64	2.26	4.8	10.6	26.1	55.5	9.1
	B	0.82	1.66	2.39	4.2	9.5	24.4	53.3	8.4
	C	0.93	1.75	2.73	5.1	11.0	27.0	54.2	9.5
	Avg.	0.86	1.68	2.46	4.7	10.3	25.8	54.4	9.0
Multi	A	1.37	1.30	1.69	2.8	5.8	15.6	42.6	5.4
	B	1.37	1.53	2.01	3.0	6.2	16.8	43.8	5.9
	C	1.54	1.48	2.13	3.2	6.7	16.7	43.5	6.1
	Avg.	1.43	1.44	1.94	3.0	6.2	16.3	43.3	5.8

the noise robustness. However, the bottom line is that model combination does not lead to improved noise robustness.

VII. CONCLUSION AND FUTURE WORK

In this paper we studied reservoir based acoustic modeling for noise robust continuous digit recognition. A reservoir based acoustic model computes the state likelihoods in an HMM by means of a two-layer recursive neural network. This network is peculiar in the sense that it consists of a hidden layer of recurrently connected non-linear neurons with fixed (= non-trained) coefficients – called a reservoir – and an output layer of linear neurons with coefficients that can be trained using a simple Tichonov regression method. A particular advantage of reservoir networks is that they are not easily over-trained.

The main objective of our work was to demonstrate that an RC-based system comprising a cascade of reservoir networks can outperform GMM-HMM systems in noisy conditions with different front-ends. The introduction of noise robust features (AFE) and bi-directional reservoir networks clearly lead to lower WERs, both in matched and mismatched conditions. Our present systems now outperform all other neural-based approaches we know of that were recently evaluated for continuous digit recognition.

We also investigated different combinations of RC-based and GMM-based systems to find a way of improving the reservoir performance in clean conditions. Particularly, we introduced RC-GMM and GMM-RC tandems as well as a simple fusion approach to combine the reservoir and GMM likelihoods. Our experiments showed that although adding the information from a GMM marginally improves the performance in the matched condition, it degrades the performance in the mismatched environments.

Given the above observations, our future research will investigate more front-end and back-end approaches that can further improve a hybrid RC-HMM system. One direction is to train a reservoir network to denoise the acoustic features, another is to investigate the combination of RC-HMM and the uncertainty decoding framework. Furthermore, we plan to evaluate the potential of reservoir systems in the context of noise robust large vocabulary recognition (e.g. Aurora-4).

ACKNOWLEDGMENT

The research leading to the results presented here has received funding from Flemish Science Foundation (FWO) under grant agreement G.0088.09N (RECAP).

REFERENCES

[1] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, no. 1, p. ID942617, 2009.

[2] ETSI, "Speech processing, transmission and quality aspects STQ; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ES 202 050, Tech. Rep., 2002.

[3] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 257–270, jan. 2007.

[4] X. Xiao, J. Li, E.-S. Chng, H. Li, and C.-H. Lee, "A study on the generalization capability of acoustic models for robust speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1158–1169, 2010.

[5] H. Xu, M. Gales, and K. Chin, "Joint uncertainty decoding with predictive methods for noise robust speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1665–1676, 2011.

[6] O. Vinyals and S. Ravuri, "Comparing multilayer perceptron to deep belief network tandem features for robust ASR," in *Proc. ICASSP*, 2011, pp. 4596–4599.

[7] S.-X. Zhang and M. Gales, "Structured support vector machines for noise robust continuous speech recognition," in *Proc. Interspeech*, 2011, pp. 989–992.

[8] H. Jaeger, "The ''echo state'' approach to analysing and training recurrent neural networks - with an erratum note," GMD Report 148, German National Research Center for Information Technology, Tech. Rep., 2001.

[9] T. Robinson and F. Fallside, "A recurrent error propagation network speech recognition system," *Computer Speech and Language*, vol. 5, no. 3, pp. 259–274, Jul 1991.

[10] A. Jalalvand, F. Triefenbach, D. Verstraeten, and J.-P. Martens, "Connected digit recognition by means of reservoir computing," in *Proc. Interspeech*, 2011, pp. 1725–1728.

[11] A. Jalalvand, F. Triefenbach, and J.-P. Martens, "Continuous digit recognition in noise: Reservoirs can do an excellent job!" in *Proc. Interspeech*, 2012, p. ID:644.

[12] F. Triefenbach, A. Jalalvand, K. Demuynck, and J.-P. Martens, "Acoustic modeling with hierarchical reservoirs," *IEEE Trans. Audio, Speech and Language Processing*, vol. PP, no. 99, 2013.

[13] 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, Apr 2007.

[14] M. Lukoševičius, "A practical guide to applying echo state networks," in *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7700, pp. 659–686.

[15] H. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, ser. The Kluwer International Series in Engineering and Computer Science. Springer US, 1994, vol. 247.

[16] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.

[17] K. Demuynck, J. Duchateau, and D. Van Compernolle, "Optimal feature sub-space selection based on discriminant analysis," in *Proc. Eurospeech*, 1999, pp. 1311–1314.

[18] F. Valente, M. Magimai-Doss, and W. Wang, "Analysis and comparison of recent mlp features for lvcsr systems," in *Proc. Interspeech*, 2011, pp. 1245–1248.

[19] H.-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *Automatic Speech Recognition: Challenges for the Next Millennium*. ISCA ITRW, 2000, pp. 181–188.

[20] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural Computation*, vol. 7, pp. 108–116, 1994.

[21] H. G. Hirsch and D. Pearce, "Applying the advanced ETSI frontend to the Aurora-2 task," version 1.1, Tech. Rep., 2006.