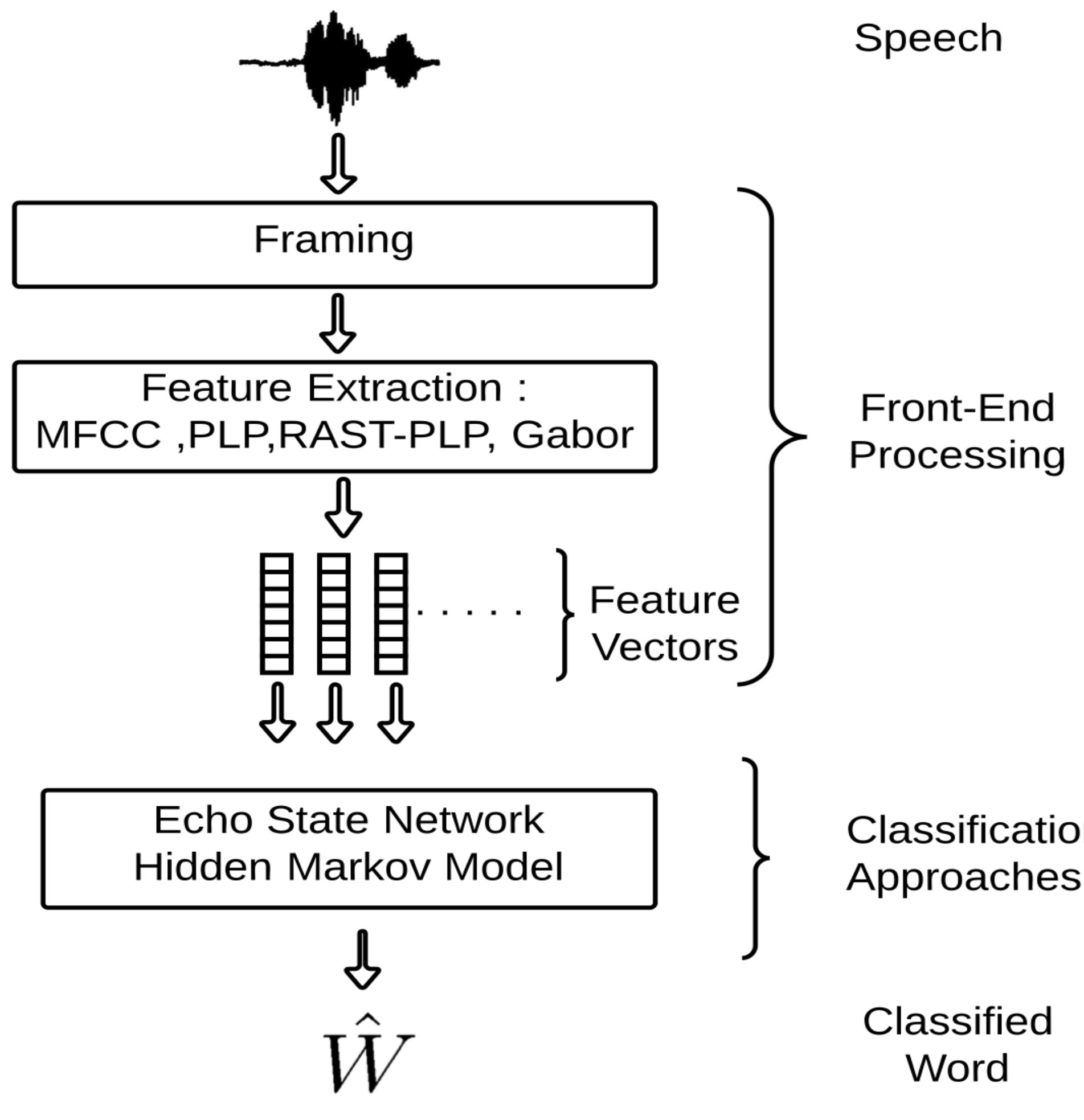
A NOISE ROBUST ARABIC ISOLATED WORD RECOGNITION SYSTEM BASED ON THE ECHO STATE NETWORK

Abdulrahman Alalshekmubarak and Leslie S. Smith, Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, Scotland, UK
(aal/l.s.smith @cs.stir.ac.uk)

Introduction

Noise-resilient automatic speech recognition (ASR) is needed for real applications, but remains a major challenge. We present a noise-resilient system based on the Echo State Network (ESN), a type of recurrent neural network. We have developed a publicly available single word Arabic speech corpus (`http://www.cs.stir.ac.uk/~lss/arabic`), consisting of 10,000 examples in total, with 50 speakers uttering 20 isolated Arabic words. We compare feature extraction methods (MFCC, PLP [1], RASTA-PLP[2], and a simple spectrotemporal Gabor filter), and classification techniques (ESN, and a baseline HMM): six models in all. These were trained on clean data, then tested on both (unseen) clean data and noise-corrupted data. ESN models almost always outperformed the HMM models, with best performance obtained using RASTA-PLP with ESN.

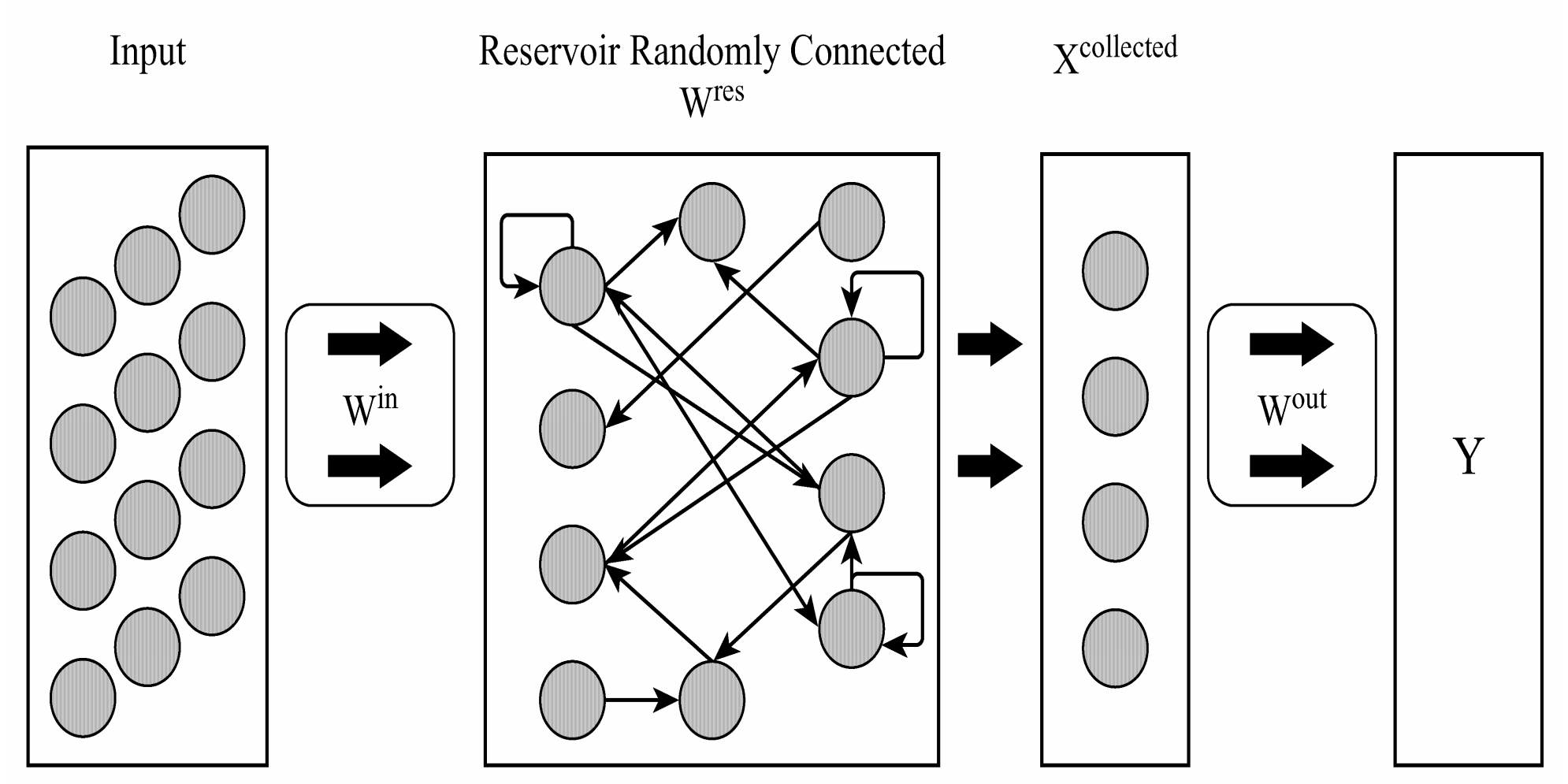
Overview



Reservoir Computing

Reservoir computing (RC) is an emerging field that offers a novel approach to training recurrent neural networks. RC contains several techniques that have been derived from different backgrounds. However, all of them share the main idea of the random initialisation between the weight of the recurrent nodes and only learning weights in the output layer by using simple read-out functions. The two major approaches that lie under the umbrella of RC are the echo state network (ESN) and the liquid state machine (LSM). Here we adopt ESN to develop our model due to its efficient implementation and the ease of interpreting the reservoir response compares to LSM where the response is in the form of spike trains that need to be further processed to obtain real valued representations that are fed to the read-out function.

ECHO STATE NETWORK



The structure of the ESN and readout system. On the left, the input signal is fed into the reservoir network through the fixed weights W^{in}. The reservoir network recodes these non-adaptively, and the output from the network is read out using the readout network weights W^{out}, which are learned.

Results

Dataset		Feature Extraction	HMM	ESN
Clean		MFCCs	97.65%	98.97% (0.15)
		PLP	98.45%	99.16%(0.11)
		RASTA-PLP	98.8 %	99.38 %(0.11)
		Gabor	———	**99.78**%(0.08)
White Noise	30 db	MFCCs	96.4 %	98.03%(0.21)
		PLP	91.3 %	90.13%(0.36)
		RASTA-PLP	98.1 %	99.04%(0.11)
		Gabor	———	**99.17**%(0.17)
	20 db	MFCCs	85.29 %	94.914 %(0.37)
		PLP	51.13 %	56.07 %(6.66)
		RASTA-PLP	96.05 %	**97.32** % (0.33)
		Gabor	———	93.39%(0.55)
	10 db	MFCCs	45.67 %	77.19 %(2.12)
		PLP	12.06 %	19.83 %(3.83)
		RASTA-PLP	81.99 %	**87.48** %(1.47)
		Gabor	———	34.07%(3.85)
Babble Noise	30 db	MFCCs	95.85 %	97.23 %(0.29)
		PLP	97.05 %	97.87 %(0.36)
		RASTA-PLP	98.65 %	99.22 %(0.19)
		Gabor	———	**99.46**%(0.14)
	20 db	MFCCs	78.49 %	89.72 %(0.87)
		PLP	86.64 %	89.47 % (2.43)
		RASTA-PLP	96.75 %	97.18 % (0.42)
		Gabor	———	**97.27** %(0.41)
	10 db	MFCCs	31.77 %	64.12 % (2.31)
		PLP	54.23 %	56.23 % (4.82)
		RASTA-PLP	85.14 %	**85.45** % (8.6)
		Gabor	———	66.56 %(1.15)

Table 1

System	Result
TM [3]	93.10%
CHMM[4]	94.09%
LoGID[5]	95.99%
ESN(This work)	**99.06%** (0.23)

Table 2

Table 1 :
Table 1 shows the results for the Arabic Speech Corpus for Isolated Words. It contains about 10000 utterances of 20 words spoken by 50 male native Arabic speakers. The best result on the clean set is achieved when combining the Gabor approach with ESN whereas PLP-RASTA with ESN provides a superior performance on the noisy sets.

Table 2 :
The results on the SAD[6], summarised in table 2, show the superior performance of the proposed system compared to the systems found in the literature. This Corpus contains 8800 samples but is available only as MFCC vectors, so that we can not apply different feature extraction methods or add noise to this corpus.

Conclusion & future work

A noise robust system based on ESN was proposed, and evaluated on a newly developed corpus and the well-known spoken Arabic digits (SAD). Different feature extraction methods considered in this study include mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction(PLP), RASTA-perceptual linear prediction, and a simple spectro-temporal Gabor filter. The result was compared with a baseline hidden Markov model (HMM). These models were trained on clean data and then tested on unseen data with different levels and types of noise. ESN models outperformed HMM models under almost all the feature extraction methods, noise levels, and noise types. The best performance was obtained by the model that combined RASTA-PLP with ESN. Future work will include an investigation of the system usability in Arabic continuous speech, the possible use of a language model and further developing the Gabor approach.

Front-End Processing Techniques

Mel-frequency Cepstral Coefficients In ASR systems, MFCCs are by far the most adopted approach. The process of computing MFCCs from the acoustic signal consists of six steps: pre-emphasis, framing & windowing, computation of the discrete Fourier transform (DFT), mel filter bank application, and finally the inverse DFT is calculated.

Perceptual Linear Prediction (PLP) Perceptual Linear Prediction was proposed in[1] as a technique that is more consistent with human hearing. The main limitation of this approach is its sensitivity to noise, which can limit its adoption in real-world applications. The main strength of this technique is the ability to compress speaker-dependent information while maintaining the relevant information needed to identify different linguistic traits even when low order prediction is used.

RASTA- Perceptual Linear Prediction (RASTA-PLP) In order to overcome the limitations of PLP, the RASTA-Perceptual Linear Prediction (RASTA-PLP) approach was introduced in[2]. It provides a low-dimensional representation with robust performance in noisy environments. Unlike short-term spectral analysis, RASTA-PLP makes use of the context information.

The Gabor Technique The Gabor technique starts from a biologically inspired gammatone filter front-end, and provides input based on signal energy, input based on onset[7], and input based on a spectro-temporal Gabor filter designed to be sensitive to amplitude modulation.

References

[1] Hermansky, H.: Perceptual linear predictive (plp) analysis of speech. The Journal of the Acoustical Society of America **87** (1990) 1738

[2] Hermansky, H., Morgan, N.: Rasta processing of speech. Speech and Audio Processing, IEEE Transactions on **2** (1994) 578–589

[3] Hammami, N., Bedda, M.: Improved tree model for arabic speech recognition. In: Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on. Volume 5. (2010) 521–526 ID: 1.

[4] Hammami, N., Bedda, M., Nadir, F.: The second-order derivatives of mfcc for improving spoken arabic digits recognition using tree distributions approximation model and hmms. In: Communications and Information Technology (ICCIT), 2012 International Conference on. (2012) 1–5 ID: 1.

[5] Cavalin, P.R., Sabourin, R., Suen, C.Y.: Logid: An adaptive framework combining local and global incremental learning for dynamic selection of ensembles of hmms. Pattern Recognition **45** (2012) 3544–3556

[6] Bache, K., Lichman, M.: Uci machine learning repository (2013)

[7] Smith, L.S., Fraser, D.S.: Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses. Neural Networks, IEEE Transactions on **15** (2004) 1125–1134