ON IMPROVING THE CLASSIFICATION CAPABILITY OF RESERVOIR COMPUTING FOR ARABIC SPEECH RECOGNITION

UNIVERSITY OF STIRLING

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)

UNIVERSITY OF STIRLING

ICANN 2014 Hamburg September 2014

Introduction

Designing noise-resilient systems is a major challenge in the field of automated speech recognition (ASR). These systems are crucial for real-world applications where high levels of noise tend to be present. We introduce a noise robust system based on Echo State Networks and Extreme Kernel Machines which we call ESNEKM. To evaluate the performance of the proposed system, we used our recently released public Arabic speech dataset and the well-known spoken Arabic digits (SAD) dataset[6]. Different feature extraction methods considered in this study include mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLP) and RASTA- perceptual linear prediction. These extracted features were fed to the ESNEKM and the result compared with a baseline hidden Markov model (HMM), so that nine models were compared in total. ESNEKM models outperformed HMM models under all the feature extraction methods, noise levels, and noise types. The best performance was obtained by the model that combined RASTA-PLP with ESNEKM.

Proposed System

The proposed model aims to improve the classification capability of the ESN by applying the EKM classifier on the output layer instead of the linear classifier used in the conventional approach. We have found few attempts in the literature to overcome the limitation of the linear readout function and replace it with a nonlinear classifier. The added training time (and the binary nature of some classifiers such as SVMs) makes this impractical for many tasks, specifically for multi-label tasks with a relatively large number of classes. However, using EKM yields the benefits of using the nonlinear function while maintaining a single-shot convex solution that can handle multi-label tasks even when the number of labels is large. This is important not only for the efficiency aspect but also for reproducibility: many nonlinear classifiers such as multilayer perceptrons are sensitive to their initial weights.

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). In this paper 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.

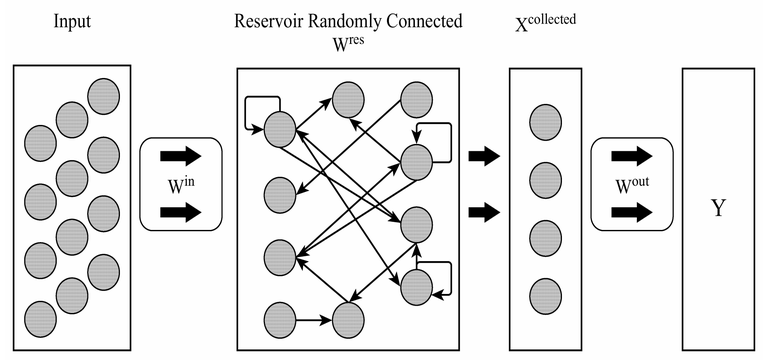
Front-End Processing Techniques

Mel-frequency Cepstral Coefficients
In ASR systems, using MFCCs is by far the most adopted approach. The process of computing MFCCs from the acoustic signal consists of six steps: pre-emphasis, framing & windowing, discrete Fourier transform is computed, mel filter bank and finally the inverse discrete Fourier transform 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 a small number of order 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 context information.

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	ESNEKM
Clean		MFCCs	97.65%	98.97%(0.15)	99.59%(0.05)
		PLP	98.45%	99.16%(0.11)	99.31%(0.09)
		RASTA-PLP	98.8 %	99.38%(0.11)	99.69%(0.06)
White Noise	30 db	MFCCs	96.4%	98.03%(0.21)	99.05%(0.13)
		PLP	91.3%	90.13%(0.36)	97.59%(0.17)
		RASTA-PLP	98.1%	99.04%(0.11)	99.59%(0.06)
	20 db	MFCCs	85.29%	94.91%(0.37)	94.82%(0.30)
		PLP	51.13%	56.07%(6.66)	75.39%(0.97)
		RASTA-PLP	96.05%	97.32% (0.33)	98.41%(0.07)
	10 db	MFCCs	45.67%	77.19%(2.12)	79.50%(0.85)
		PLP	12.06%	19.83%(3.83)	35.35%(1.96)
		RASTA-PLP	81.99%	87.48%(1.47)	90.29%(0.53)
Babble Noise	30 db	MFCCs	95.85%	97.23 %(0.29)	99.35 %(0.18)
		PLP	97.05%	97.87%(0.36)	99.02%(0.06)
		RASTA-PLP	98.65%	99.22%(0.19)	99.65%(0.06)
	20 db	MFCCs	78.49%	89.72%(0.87)	94.41%(0.34)
		PLP	86.64%	89.47% (2.43)	96.64%(0.22)
		RASTA-PLP	96.75%	97.18%(0.42)	98.30%(0.14)
	10 db	MFCCs	31.77%	64.12%(2.31)	65.48%(0.86)
		PLP	54.23%	56.23%(4.82)	81.23%(0.32)
		RASTA-PLP	85.14%	85.45%(8.6)	90.76%(0.44)

Table 1

System	Result
TM [3]	93.10%
CHMM[4]	94.09%
LoGID[5]	95.99%
ESN(This work)	**99.06%** (0.23)
ESNEKM(This work)	**99.16%** (0.12)

Table 2

Table 1 :
Results obtained by the proposed system, ESN and a baseline hidden Markov model (HMM) on the Arabic Speech Corpus for Isolated Words. It contains about 10000 utterances of 20 words spoken by 50 native male Arabic speakers. In ESN and ESNEKM, we report the means and the standard deviations over ten runs.

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 and available only as MFCC vectors, so we can not apply different feature extraction methods or add noise to this corpus.

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

Conclusion & Future Work

A novel speech recognition model based on RC and EKM which we call ESNEKM 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) and RASTA-perceptual linear prediction. The result was compared with a baseline hidden Markov model (HMM), so that nine models were compared in total. These models were trained on clean data and then tested on unseen data with different levels and types of noise. ESNEKM models outperformed HMM models under all the feature extraction methods, noise levels, and noise types. The best performance was obtained by the model that combined RASTA-PLP with ESNEKM. Future work will include an investigation of the system usability in Arabic continuous speech and the possible use of a language model.