Pattern Recognition Letters 000 (2014) 1–9

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Emotion recognition from speech signals via a probabilistic echo-state network ☆

Edmondo Trentin [a,*], Stefan Scherer [b], Friedhelm Schwenker [c]

[a] DIISM, Università di Siena, V. Roma 56, I-53100 Siena, Italy
[b] USC Institute for Creative Technologies, 12015 Waterfront Drive, 90094-2536 Playa Vista, CA, USA
[c] Institute of Neural Information Processing, Ulm University, Oberer Eselsberg, D-89069 Ulm, Germany

ARTICLE INFO

Article history:
Received 26 February 2014
Available online xxx

Keywords:
Emotion recognition
Echo state network
Sequence clustering
Semi-supervised learning

ABSTRACT

The paper presents a probabilistic echo-state network (π-ESN) for density estimation over variable-length sequences of multivariate random vectors. The π-ESN stems from the combination of the reservoir of an ESN and a parametric density model based on radial basis functions. A constrained maximum likelihood training algorithm is introduced, suitable for sequence classification. Extensions of the algorithm to unsupervised clustering and semi-supervised learning (SSL) of sequences are proposed. Experiments in emotion recognition from speech signals are conducted on the WaSeP© dataset. Compared with established techniques, the π-ESN yields the highest recognition accuracies, and shows interesting clustering and SSL capabilities.

1. Introduction

Emotional communication in human-to-human interaction is indefeasible and often crucial. It may convey information that would be missed otherwise (e.g., on the speaker's unspoken thoughts, or on the underlying context in which a dialogue is taking place), information that could grant its real meaning to what is factually spoken. Therefore, in order to render human-computer interaction (HCI) more natural and efficient, machines are sought that can recognize, understand, and express emotional states. In fact, although HCI has been taking a more and more relevant place in our everyday lives, the science of emotion modeling and recognition from audio or video signals is still in its infancy.

On the other hand, several pioneering approaches to emotion recognition from speech signals can be found in the literature. Vlasenko et al. [1] apply Gaussian mixture models (GMM) and hidden Markov models (HMM) defined at both the frame- and turn-level representations of the audio signals, while Wagner et al. [2] thoroughly analyze the behavior of HMMs and support vector machines (SVM) using Mel-cepstra [3] and energy-based features. Schwenker et al. [4] investigate the use of the SVM-GMM Supervector approach relying on PLP and ModSpec features [5]. Dellaert et al. [6] classify speech signals into 4 broad classes of emotions by applying a mixture of k-nearest neighbor [7] experts (with $k = 11$) estimated on different subsets of acoustic features. Depending on the method and on the dataset, these studies observed recognition accuracies ranging mostly between 60% and 85%.

This paper introduces and investigates a novel approach to emotion recognition and clustering from speech signals, the probabilistic echo state network (π-ESN). This article is the journal version of a workshop communication [8], and introduces new algorithms, faces new setups (unsupervised, semi-supervised), reports on a much wider and deeper experimental investigation, and offers an in-depth discussion of the key findings.

The π-ESN realizes a parametric model of the probability density function (pdf) underlying the distribution of a sample of variable-length sequences of real-valued, multivariate random vectors. The model relies on the hybridization between an echo state network (ESN) [9] and a constrained radial basis function (RBF)-like network [10].

While RBFs are feed-forward networks known to realize linear combinations of Gaussian kernels evaluated on a fixed-dimensional, real-valued vector space, ESNs are a particular subclass of the broad family of recurrent neural networks (RNN). A schematic representation of an ESN is shown in Fig. 1. The most important part of the network is its recurrent reservoir. It is a large collection of units that are loosely and randomly connected to each other. The probability of a connection holding between a generic pair of units a_i, a_j is a decreasing function of the reservoir size, and lies typically in the $(0.02, 0.1)$ range [9]. The connection weights W of the reservoir are drawn at random, as well. Additive-sigmoid transfer functions $\psi(.)$ are associated with the units. The input layer is fully connected with the reservoir via connections having weight matrix W^{in}. The reservoir, in turn, is

☆ This paper has been recommended for acceptance by Sanniti di Baja.
* Corresponding author. Tel.: +39 0577 233601; fax: +39 0577 233602.
E-mail address: trentin@dii.unisi.it (E. Trentin).

http://dx.doi.org/10.1016/j.patrec.2014.10.015
0167-8655/

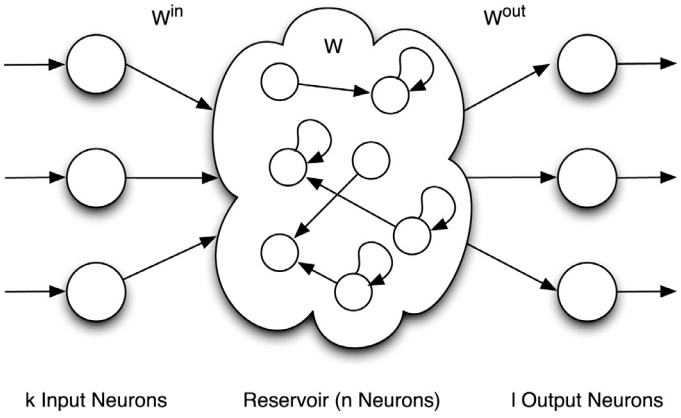

Fig. 1. Schematics of an ESN. Inputs and outputs are fully connected to the reservoir via W^{in} and W^{out}, respectively. Connections of the reservoir and their weights W are random.

fully connected (with weight matrix W^{out}) to the linear output layer. The loose connectivity of the reservoir leads to the formation of small cycles of units that are recursively connected to each other. These cycles are sensitive to certain dynamic phenomena in the signal received through the input units and from other adjacent neurons in the reservoir. Since there are feed-backward and recursive connections within the reservoir, the output at any given time t is a function of the current input pattern and of the *state* (i.e., the value yielded by the corresponding transfer function) of each of the other units in the reservoir at time t (which can be thought of as a non-linearly filtered history of all the inputs up to time $t-1$).

ESNs found application to such different tasks as classification, pattern generation, and control [9,11–13]. They offer advantages with respect to traditional RNNs, e.g. their stability toward noisy inputs [11] and the efficient weight adaptation method [13]. Moreover, ESNs possess the universal computation property, i.e. they can approximate arbitrarily well any non linear filter having bounded memory [14]. Since there is no backpropagation of partial derivatives through the reservoir (albeit there are bounds on the ESN memory [15] in the general case), ESNs do not suffer from major learning problems that affect classic RNNs [16]. This is utterly relevant to our purposes, making the ESN a viable candidate for encoding multivariate time sequences, including the modeling of typical dynamics found in the speech signals such as the prosody of emotional expressions. The expectation is corroborated by the empirical evidence reported in [8,12].

The basic idea pursued in the paper is that the recurrent reservoir of the ESN realizes an encoding of an input sequence by means of the pattern of activation of its state units. The trainable state-to-output weights and the linear output layer of the ESN are replaced by an RBF architecture. The RBF is trained in order to estimate the pdf underlying the distribution of these patterns of activation within the encoding space. Training is realized according to a constrained gradient-ascent algorithm, presented in Section 2, aimed at the maximization of the likelihood of the parameters of the model given the input sequence. Constraints are required to ensure that the estimated model satisfies the axioms of probability. The training scheme is inherently unsupervised and non-discriminative, along the line of statistical parametric pdf estimation techniques that rely on the maximum-likelihood (ML) criterion [7]. Nonetheless, it can be applied in classification tasks by using a separate π-ESN to estimate the class-conditional pdf [7] for each of the classes $\omega_1, \ldots, \omega_c$ involved in the problem, and by applying Bayes decision rule.

The algorithm has been introduced in the framework of sequence classification, assuming that class-labels of specific emotions are associated with all the acoustic observation sequences in the training set. Unfortunately, there are two major issues with this assumption: (1) the categorization of emotions into classes is intrinsically ill-defined, possibly overlapping, and even subjective; (2) emotion processing in real-world scenarios (i.e., not relying on pseudo-emotions simulated by actors) would require huge amounts of mostly unlabeled spontaneous speech data. These issues are faced in the paper by extending the π-ESN training algorithm to fit the unsupervised clustering and the semi-supervised learning (SSL) setups [17] with adaptive number of clusters/classes. This is achieved in Section 3 by exploiting the probabilistic nature of the π-ESN within a (quasi)cross-validated likelihood model selection strategy [18].

Section 4 reports (and discusses in depth) experiments based on a corpus containing pseudo-words spoken in six different emotional prosodies, WaSeP© [19]. The behavior of the π-ESN in supervised, unsupervised, and semi-supervised tasks is analyzed and (favorably) compared w.r.t. established techniques. Final remarks are drawn in Section 5.

2. The probabilistic echo-state network

As we stated in the previous section, a separate, class-specific, and independent π-ESN is used for each emotion involved in the task. Thence, in the following we will focus on a generic π-ESN, trained over the corresponding emotion-specific training sample, with the understanding that the algorithm has to be subsequently applied to as many π-ESNs as the number of classes at hand. Albeit intrinsically unsupervised, the algorithm is foremost oriented to supervised classification tasks.

Let $\mathcal{T} = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_N\}$ be a random sample of N acoustic observation sequences, identically and independently drawn (iid) from the unknown pdf $p(\mathcal{Y})$. The π-ESN is devised as a plausible parametric model of $p(\mathcal{Y})$, namely $p(\mathcal{Y}|\theta)$, determined uniquely by the value of its parameters $\theta = (\theta_1, \ldots, \theta_k)$. A parameter learning algorithm is pursued that maximizes the likelihood $p(\mathcal{T}|\theta)$ of θ given $\mathcal{T}$. Relying on the iid assumption, we can write $p(\mathcal{T}|\theta) = \prod_{i=1}^{N} p(\mathcal{Y}_i|\theta)$. Before proceeding, it is necessary to specify a well-defined form for $p(\mathcal{Y}|\theta)$, as follows. Let us assume the existence of an integer d and of two functions, $\phi : \{\mathcal{Y}\} \to \Re^d$ (where $\{\mathcal{Y}\}$ is the universe of all possible observation sequences) and $\hat{p} : \Re^d \to \Re$, s.t. $p(\mathcal{Y})$ can be decomposed as:

$$p(\mathcal{Y}) = \hat{p}(\phi(\mathcal{Y})). \tag{1}$$

It is seen that there exist (infinite) functions $\phi(.)$ and $\hat{p}(.)$ that satisfy Equation (1), the most trivial being $\phi(\mathcal{Y}) = p(\mathcal{Y}), \hat{p}(x) = x$. We call $\phi(.)$ the *encoding*, while $\hat{p}(.)$ is simply referred to as the "likelihood". Again, we assume parametric models $\phi(\mathcal{Y}|\theta_\phi)$ and $\hat{p}(\mathbf{x}|\theta_{\hat{p}})$ for the encoding and for the likelihood, respectively, and we set $\theta = (\theta_\phi, \theta_{\hat{p}})$ and $p(\mathcal{Y}|\theta) = \hat{p}(\phi(\mathcal{Y}|\theta_\phi)|\theta_{\hat{p}})$.

A hybrid two-block connectionist/statistical model is proposed for $p(\mathcal{Y}|\theta)$ as follows. The function $\phi(\mathcal{Y}|\theta_\phi)$ is realized via an ESN, suitable to map sequences $\mathcal{Y}$ into real vectors $\mathbf{x}$. Let θ_ϕ be the set of the ESN weights. A RBF-like network is then used to model $\hat{p}(\mathbf{x}|\theta_{\hat{p}})$, where $\theta_{\hat{p}}$ is the parameter vector of the RBF. In order to ensure that a pdf is obtained, constraints have to be placed on the hidden-to-output connection weights of the RBF (assuming that normalized Gaussian kernels are used).

First, let us focus on the ESN-based model for $\phi(\mathcal{Y}|\theta_\phi)$. The topology and the weight matrix W of the reservoir are generated at random. W is normalized s.t. its spectral radius is $\alpha \leq 1$ [13]. This scaling of the weight matrix is accomplished so that the maximal eigenvalue $\lambda_{\max}$ of W satisfies $|\lambda_{\max}| \leq 1$. The encoding of $\mathcal{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_T$ (where T is not fixed, but sequence-specific) is accomplished as follows:

1. initialize the state $\hat{\mathbf{x}}$ of the reservoir at random
2. feed the ESN with the first L acoustic feature vectors[1] $\mathbf{y}_1, \ldots, \mathbf{y}_L$
3. save the resulting state $\mathbf{x}_0$ of the ESN at time L as the starting state

[1] This is done to minimize the influence of the random initial conditions [13].