10.12871/clicit2014221

A Preliminary Application of Echo State Networks to Emotion Recognition

Claudio Gallicchio
Department of Computer Science
University of Pisa
Largo B. Pontecorvo 3
56127 Pisa, Italy
gallicch@di.unipi.it

Alessio Micheli
Department of Computer Science
University of Pisa
Largo B. Pontecorvo 3
56127 Pisa, Italy
micheli@di.unipi.it

Abstract

English. This report investigates a preliminary application of Echo State Networks (ESNs) to the problem of automatic emotion recognition from speech. In the proposed approach, speech waveform signals are directly used as input time series for the ESN models, trained on a multi-classification task over a discrete set of emotions. Within the scopes of the Emotion Recognition Task of the Evalita 2014 competition, the performance of the proposed model is assessed by considering two emotional Italian speech corpora, namely the E-Carini corpus and the €motion corpus. Promising results show that the proposed system is able to achieve a very good performance in recognizing emotions from speech uttered by a speaker on which it has already been trained, whereas generalization of the predictions to speech uttered by unseen subjects is still challenging.

Italiano. *Questo documento esamina l'applicazione preliminare delle Echo Stato Networks (ESN) per il problema del riconoscimento automatico delle emozioni dal parlato. Nell'approccio proposto, i segnali che rappresentano la forma d'onda del parlato sono usati direttamente come serie temporali di ingresso per i modelli ESN, addestrati su un compito di multi-classificazione su un insieme discreto di emozioni. Entro gli ambiti della Emotion Recognition Task della competizione Evalita 2014, la performance del modello proposto viene valutata considerando due corpora di dati emotivi in lingua Italiana, ovvero il corpus E-Carini e il corpus €motion. I risultati ottenuti sono promettenti e mostrano che il sistema proposto è in grado di raggiungere una buona prestazione nel riconoscimento di emozioni a partire dalle parole pronunciate da un utente sul quale il sistema è stato già addestrato, mentre la generalizzazione delle predizioni per le frasi pronunciate da soggetti mai visti in fase di addestramento rappresenta ancora un aspetto ambizioso.*

1 Introduction

The possibility of recognizing human emotions from uttered speech is a recent interesting area of research, with a wide range of potential applications in the field of human-machine interactions. One of the most prominent aspects of recent systems for emotion recognition from speech relates to the choice of proper features that should be extracted from the waveform signals. Popular choices for such features are continuous features (Lee and Narayanan, 2005), such as pitch-related features or energy-related features, or spectral based features, such as linear predictor coefficients (Rabiner and Schafer, 1978) or Mel-frequency cepstrum coefficients (Bou-Ghazale and Hansen, 2000).

Within the scopes of the Evalita 2014 competition, this report describes a preliminary investigation of the application of Echo State Networks (ESNs) (Jaeger and Haas, 2004) to the problem of identifying speakers' emotions from a discrete set, namely anger, disgust, fear, joy, sadness and surprise. We adopt the paradigm of Reservoir Computation (Lukosevicius and Jaeger, 2009), which represents a state-of-the-art approach for efficient learning in time-series domains, within the class of Recurrent Neural Networks, naturally suitable for treating sequential/temporal information. As such, in our proposed approach, the waveform sig-

nals representing speech are directly used as input for the emotion recognition system, allowing to avoid the need for domain-specific feature extraction from waveform signals. In order to assess the generalization performance of the proposed emotion recognition system, we take into consideration a homogeneous experimental setting and a heterogeneous experimental setting. In the homogeneous setting, the performance of the recognition system is assessed on sentences uttered by the same speaker on which the system has been trained, while in the heterogeneous setting, the performance is assessed on sentences pronounced by unseen subjects during the training process.

2 Description of the System

We took into consideration data coming from two emotional Italian speech corpora, namely the E-Carini corpus (Tesser et al., 2005; Avesani et al., 2004) and the €motion corpus (Galatà, 2010). Each corpus contains waveform signals representing sentences spoken by a single user, see the task report (this volume) for further details. Such data was then organized into two datasets, one for each corpus, segmenting sentences into words, based on the available information. Our emotion recognition system directly uses the sounds waveform of spoken words as input time-series for the neural network model, avoiding the use of feature extraction for speech representation. The only preprocessing step consists in normalizing the input signals to zero mean and unitary standard deviation, using the data pertaining to the extra neutral emotion class for computing the normalization constants, independently for each speaker.

The two resulting datasets were used to organize two multi-classification task for emotion recognition: a homogeneous task and a heterogeneous task. The homogeneous task includes only the E-Carini corpus dataset, and is designed for assessing the ability of the emotion recognition system to detect human emotions pertaining to a single speaker. Indeed, training and test set for the homogeneous task contain sequences pertaining to the same speaker (test set represents $\approx 30\%$ of the available data). The heterogeneous task includes both the E-Carini corpus and the €motion corpus, and is designed to evaluate the generalization ability of the emotion recognition system when trained on data pertaining to one speaker and tested on data pertaining to a different speaker. In the case of the heterogeneous task, the training set contains data from the E-Carini corpus, while the test set contains data from the €motion corpus. For both the homogeneous and the heterogeneous tasks, the training set was balanced over the class of possible emotions.

Emotion classification is performed by using ESN, which implement discrete-time non-linear dynamical systems. From an architectural perspective, an ESN is made up of a recurrent *reservoir* component, and a feed-forward *readout* component. In particular, the reservoir part updates a state vector which provides the network with a non-linear dynamic memory of the past input history. This allows the state dynamics to be influenced by a portion of the input history which is not restricted to a fixed-size temporal window, enabling to capture longer term input-output relationships. In the context of the specific application under consideration, it is worth noticing that the role of the reservoir consists in directly encoding the temporal sequences of the waveform signals into a fixed-size state (feature) vector, allowing to avoid the need for the extraction of specific features from the uttered sentences. The basic architecture of an ESN includes an input layer with N_U units, a non-linear, recurrent and sparsely connected reservoir layer with N_R units, and a linear, feed-forward readout layer with N_Y units. In particular, for our application we use $N_U = 1$ and $N_Y = 6$, where each one of the output dimensions corresponds to one of the emotional classes considered. In this paper we take into consideration the leaky integrator ESN (LI-ESN) (Jaeger et al., 2007), which is a variant of the standard ESN model, with state dynamics particularly suited for representing the history of slowly changing input signals.

State dynamics of the ESNs follow the word by word segmentation organization considered in the datasets. Accordingly, for each word w, at each time step t, the reservoir computes a state $\mathbf{x}_w(t) \in \mathbb{R}^{N_R}$ according to the equation:

$$\begin{aligned} \mathbf{x}_w(t) = \quad & (1-a)\mathbf{x}_w(t-1) + \\ & af(\mathbf{W}_{in}\mathbf{u}_w(t) + \hat{\mathbf{W}}\mathbf{x}_w(t-1)) \end{aligned} \tag{1}$$

where $\mathbf{u}_w(t)$ is the input at time-step t, $\mathbf{W}_{in}$ is the input-to-reservoir weight matrix, $\mathbf{W}$ is the recurrent reservoir weight matrix, $a \in [0, 1]$ is a leaking rate parameter, f is an element-wise applied activation function (we use $tanh$), and a zero vector

is used for state initialization. After the last time step for word w has been considered, a mean state mapping function is applied, according to:

$$\mathcal{X}(w) = \frac{1}{length(w)} \sum_{t=1}^{length(w)} \mathbf{x}_w(t) \qquad (2)$$

where $length(w)$ is the number of time steps covered by the sentence w. For further information about state mapping functions in general, and mean state mapping in particular, the reader is referred to (Gallicchio and Micheli, 2013).

The classification output is computed by the readout component of the ESN, which linearly combines the output of the state mapping function, according to the equation:

$$\mathbf{y}(w) = \mathbf{W}_{out}\mathcal{X}(w) \qquad (3)$$

where $\mathbf{W}_{out}$ is a reservoir-to-readout weight matrix. The emotional class for each word is set to the class corresponding to the element with the highest activation in the output vector. The final classification of a sentence is computed by a voting process, according to which each sentence is classified as belonging to the emotional class which is more represented among the words that compose that sentence.

Training in ESNs is restricted to only the readout component, i.e. only the weight values in matrix $\mathbf{W}_{out}$ are adapted, while elements in $\mathbf{W}_{in}$ and $\mathbf{W}$ are initialized in order to satisfy the conditions of the *echo state property* (Jaeger and Haas, 2004) and then are left untrained. In practical applications, such initialization process typically consists in a random initialization (from a uniform distribution) of weight values in matrices $\mathbf{W}_{in}$ and $\mathbf{W}$, after which matrix $\mathbf{W}$ is scaled such that its spectral radius $\rho(\mathbf{W})$ is less than 1, see (Jaeger, 2001) and (Gallicchio and Micheli, 2011) for details.

3 Results

In our experiments we considered ESNs with reservoir dimension $N_R \in \{100, 200\}$, 10% of reservoir units connectivity, spectral radius $\rho = 0.999$ and leaky parameter $\alpha = 0.01$. For every reservoir hyper-parametrization, results were averaged over a number of 10 reservoir guesses. The readout part of the ESNs was trained using pseudo-inversion and ridge regression with regularization parameter $\lambda \in \{10^j | j = -5, -4, -3, -2, -1, 0, 1, 2, 3\}$. Reservoir dimension and readout regularization were chosen on a validation set (with size of $\approx 30\%$ of the training set size), according to a hold out cross validation scheme for model selection.

The performance of the emotion recognition is assessed by measuring the accuracy for the multi-classification task, i.e. the ratio between the number of correctly classified sentences and the total number of sequences. Average training and test accuracy obtained on both the homogeneous and heterogeneous tasks are reported in Table 3.

Task	**Training**	**Test**
homogeneous	$0.86(\pm 0.01)$	$0.82(\pm 0.01)$
heterogeneous	$0.91(\pm 0.02)$	$0.27(\pm 0.03)$

Table 1: Average training and test performance accuracy achieved by ESNs on the homogeneous task and on the heterogeneous task.

For the sake of performance comparison, notice that the accuracy achieved by a chance-null model is 0.17 on both the tasks. The averaged accuracy achieved on the test set of the homogeneous task is 0.82, which is comparable with literature results on emotion recognition from speech in homogeneous training-test condition (Ayadi et al., 2011). The averaged accuracy achieved on the test set of the heterogeneous task is 0.27. Note that, although such performance is far from the one achieved on the homogeneous task, it is still definitely beyond the performance of the null model. The result achieved by the system trained on the heterogeneous case on the full test set of the Evalita 2014 competition, comprising data from 5 different unseen speakers, is 0.24.

4 Discussion

In this report we have described a preliminary application of ESNs to the problem of recognizing human emotions from speech. The proposed emotion recognition system directly uses as input the time series of the waveform signals corresponding to the uttered sentences, avoiding the need for a specific feature extraction process. Two experimental settings have been considered, with training and test data pertaining to sequences pronounced by the same speaker (homogeneous setting) or not (heterogeneous setting). Performance results achieved by ESNs are promising. In particular, a very good predictive performance is ob-

tained when the system is assessed considering unseen sentences pronounced by a speaker on which the system has already been trained. On the other hand, the generalization of the emotion predictions to speech uttered by speakers on which the system has not been trained still remains a challenging aspect. Overall, given the characteristics of efficiency and simplicity of the proposed approach, and in view of a possible integration with domain-specific techniques for the multi-speaker case, we believe that the proposed system can represent an interesting contribution for the design of tools in the area emotional speech processing.

References

Cinzia Avesani, Piero Cosi, Elisabetta Fauri, Roberto Gretter, Nadia Mana, Silvia Rocchi, Franca Rossi, and Fabio Tesser. 2004. Definizione ed annotazione prosodica di un database di parlato-letto usando il formalismo tobi. In *Il parlato Italiano*, pages 1–14.

Moataz El Ayadi, Mohamed Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Sahar E. Bou-Ghazale and John Hansen. 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. *Speech and Audio Processing, IEEE Transactions on*, 8(4):429–442.

Vincenzo Galatà. 2010. Production and perception of vocal emotions: a cross-linguistic and cross-cultural study. PhD Thesis, University of Calabria, Italy, (unpublished).

Claudio Gallicchio and Alessio Micheli. 2011. Architectural and markovian factors of echo state networks. *Neural Networks*, 24(5):440 – 456.

Claudio Gallicchio and Alessio Micheli. 2013. Tree echo state networks. *Neurocomputing*, 101:319–337.

Herbert Jaeger and Harald Haas. 2004. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80.

Herbert Jaeger, Mantas Lukosevicius, Dan Popovici, and Udo Siewert. 2007. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Networks*, 20(3):335–352.

Herbert Jaeger. 2001. The ”echo state” approach to analysing and training recurrent neural networks. Technical report, GMD.

Chul Min Lee and Shrikanth Narayanan. 2005. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303.

Mantas Lukosevicius and Herbert Jaeger. 2009. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Lawrence Rabiner and Ronald Schafer. 1978. *Digital Processing of Speech Signals*. Pearson Education.

Fabio Tesser, Piero Cosi, Carlo Drioli, and Graziano Tisato. 2005. Emotional festival-mbrola tts synthesis. In *INTERSPEECH*, pages 505–508.