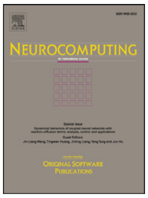
Neurocomputing 000 (2017) 1–13

Contents lists available at ScienceDirect

Neurocomputing

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



Online phoneme recognition using multi-layer perceptron networks combined with recurrent non-linear autoregressive neural networks with exogenous inputs

Diana A. Bonilla Cardona[a,b,*], Nadia Nedjah[a,b], Luiza M. Mourelle[a,b]

[a] Department of Electronics Engineering and Telecommunication, Engineering Faculty, State University of Rio de Janeiro, Brazil
[b] Department of System Engineering and Computation, Engineering Faculty, State University of Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 28 March 2016
Revised 2 September 2016
Accepted 21 September 2016
Available online xxx

Keywords:
Pattern recognition
Phoneme identification
Multi-layer perceptron
NARX neural networks
Dynamic neural networks

ABSTRACT

Off-line pattern recognition in speech signals is a complex task. Yet, this task becomes harder when the recognition result is required online or in real-time. The present work proposes an online identification of the Portuguese language phonemes using a non-linear autoregressive model with exogenous inputs, commonly called NARX. The process first conditions the input speech signal, and extracts its frequency characteristics. Then it pre-classifies the extracted features into one of the ten possible groups of phonemes, as available in the Portuguese language. This pre-classification is done using a multilayer perceptron network (MLP) with a supervised learning. Subsequently, the MLP output vector, together with the vector that carries the input frequencies, feeds a NARX neural network by means of a temporal delay of four times and feed-backward recurrent links that encompass the results of all hidden layers of the network. As a result of this process, the proposed phoneme recognition process improves the accuracy of an online identification of the Portuguese spoken phonemes during a natural conversation. When the phoneme input signal is well conditioned and continuous over time, the proposed recognition process can provide the correct classification in real-time, with an acceptable accuracy rate.

1. Introduction

For humans, one of the most important forms of communication, simple and most used since antiquity is speech. The information that is relevant both to the person transmitting the message and to the recipient is transmitted through the human voice. However, the written language has been of great importance in the development of mankind, since the information is preserved over time and available to more people. In addition, many people with physical or hearing handicap need technological aids to obtain or transmit a message. A Speech To Text system (STT) is a device that receives as input a speech signal, pre-processes it and then applies an automated method to identify each of the spoken words included in the pre-processed input signal, resulting in a text that is displayed to the user via graphical playback devices.

Automatic speech recognition (ASR) converts a given speech signal into an equivalent word sequence [2]. The ASR is an area of research that seeks to develop human-machine interfaces controlled by voice, and thus replacing traditional interfaces based on keyboards, touchscreen and similar devices. Despite the fact that speech recognition is one of the most intuitive tasks for humans, the process is not that evident as far as machines are concerned. For a long time the implementation of ASR was affected by several aspects that compromise the speech signal the system has to process. The external noise, the speaker voice characteristics and phonetics of each language, in addition to the semantic rules has influenced a great deal the development of automatic speech recognition process.

Nowadays, technological research work regarding this area is increasing and evolving thanks to the availability of more computationally powerful yet affordable computer systems along with the development of more accurate models and algorithms that can be applied to automatic speech recognition. It is noteworthy to advocate that transcribed speech information does facilitate interactivity between users and smart devices such as computers, smartphones and SmartTV. However, ASR systems are far from providing universal solutions that can be used in any situation. Generally, these require the tuning of many particular characteristics that depend on the operating conditions as well as on the

[*] Corresponding author at: Department of Electronics Engineering and Telecommunication, Engineering Faculty, State University of Rio de Janeiro, Brazil.

E-mail addresses: dabonillac@gmail.com (D.A. Bonilla Cardona), nadia@eng.uerj.br (N. Nedjah), ldmm@eng.uerj.br (L.M. Mourelle).

http://dx.doi.org/10.1016/j.neucom.2016.09.140
0925-2312/© 2017 Elsevier B.V. All rights reserved.

application characteristics too. The techniques used to implement an automatic speech recognition are categorized as follows:

- *Topological*: These methods are mainly based on Dynamic Time Wrapping (DTW), which is based on computation and comparison of distances [15].
- *Probabilistic*: The main tool of these methods consists of the Hidden Markov Models (HMM), which are based on generative models of the vocabulary words [19].
- *Artificially intelligent*: These techniques are diversified and include Artificial Neural Networks (ANN), knowledge-based systems, Expert systems, among others [13].

Automatic speech recognizers can process the input text in three distinct ways. The first way refers to batch recognition, also known as off-line recognition, wherein the signal corresponding to a speech is first collected, and then processed, so all the collected speech is recognized in one go. This could happen at any time after the speech signal capture. The second way refers to online recognition, wherein the speech signal is processed immediately while it has been collected. In this case, the user usually only has to wait a short time to receive the response of the recognition process. Online phoneme recognition enables the user to input the speech and get the result of the processing of that data immediately, *i.e.* after a short time: one phoneme after the other one. The third way, which is actually a subset of the online processing, refers to real-time recognition. In this case, phonemes are continuously and automatically acquired from microphones, which is processed immediately in order to respond to the input in as short time as possible. After the system is finished responding, it obtains the next phoneme immediately to process it. This system does not need a user to control it as it works automatically. Mainly real-time processing can act without the need of a user intervention nor does it impose a long processing time beforehand. Most existing recognition works deal with words instead of phonemes, which makes the recognition process complex and time consuming, especially in the case of those that use topological and probabilistic speech recognition. This kind of ASR base their final decision on some pre-defined dictionaries, which require long search times. Note that this analysis concerns the test, validation and operation phases of the speech recognition system. Nonetheless, the training phase of the system is always done off-line.

At the base of efficient of automatic speech recognition is the phoneme recognition. In linguistics, phoneme is the smallest sound unit in any word. Among phonemes, there are the vowels, semivowels and consonants, which compose a large group of classes. The phonemes depend heavily on the language. For instance the groups of phonemes available in the English language are different from that available in Mandarin or Arabic. In this work, we are interested in dealing with automatic recognition of the phonemes of the Portuguese language. The sounds available in the Portuguese phonetic can be classified as in Fig. 1. For instance, note that rhotic consonants do not exist in the English language.

Neural networks configurations as used in phonemes recognition are complex and represent a major challenge during classification. This is mainly due to the fact that within distinct class groups of phonemes, there are phonetic elements with similar characteristics. This can negatively impact the accuracy of the classification process [16]. Given the available dataset for training, test and validation, in this work, we consider only a subset of these classes, as it will be explained when appropriate.

In an original approach, the work presented in this paper proposes the exploration of the architectures of Multi-Layer Perceptron (MLP) together with a Dynamic Neural Network (DNN) for the recognition of the Portuguese language phonemes. In order to cope with the dynamic nature of speech signal as to achieve speech recognition, we use the Time Delay Neural Network (TDNN) architecture, but for performance purposes, we restrict the recurrence between the output and input layers. So, we use a Nonlinear AutoRegressive eXogenous model (NARX) to implement the time delay dynamic neural network. In order to evaluate the performance of the proposed network, the comparison results of accuracy of this work and those of a simple NARX are presented and analyzed. Besides improving the recognition accuracy rate, the phoneme recognition, as proposed in this work can be used to provide classification results online. Furthermore, whenever the speech input is acquired in a controlled environment, so that no signal conditioning required, the proposed recognition process is able to provide the phoneme recognition result in a real-time pace. It is noteworthy to emphasize that even though the idea behind this work seems simple, it improves a great deal the accuracy rate of phoneme recognition and yet it can provide the classification result in real time. These characteristics are of colossal importance to any embedded system that may be needed to take care of speech within final electronic products.

The remaining of this paper is organized in seven sections. Initially, in Section 2, some of the related work is described briefly. Then, in Section 3, a global view of the architectures of the artificial neural networks explored in this work, which are the multilayer perceptron neural network and the dynamic neural networks, is provided. Subsequently, in Section 4, the proposed architecture for the recurrent network learning, regarding phoneme recognition is presented and commented. After that, in Section 5, the main steps of the implementation of the proposed network for Portuguese phonemes recognition, including signals pre-processing and extraction of the phoneme main characteristics, is given. Later on, in Section 6, the results as obtained during the performed experiments, are analyzed and discussed. Finally, in Section 7, conclusions are drawn and some directions for future work are pointed out.

2. Related works

In this section, we present the main research work related to automatic speech recognition using dynamic autoregressive neural network with exogenous inputs.

In [20], several ANN-based applications for speech processing, including speech attribute extraction, phoneme estimation and classification are presented. It was proven that ANNs play a key role in several important speech applications, wherein there is large vocabulary continuous speech recognition (LVCSR) and automatic language recognition.

In [6], recurrent neural networks (RNN) [8] are used to recognize phonemes. In this work, three recurrent neural networks are used for phoneme recognition. Each network allows three inputs and three outputs, representing the possible vowels. The work was done aiming at the classification of vowels, initial consonants, which are those that appear at the beginning of a word and final consonants, which are those that appear at the end of a word. The reported results on hit rate was satisfactory for a set of networks with one hidden layer with a number of neurons dependent on the difficulty of task related to phoneme recognition. It is noteworthy to point out that the RNN for initial consonants required a larger number of neurons than the other two RNNs, *i.e.* those dedicated to recognizing vowels and final consonants.

In [12], a NARX architecture with one hidden layer is used to improve processes that have high dependence on past events. This architecture has been tested on a grammatical inference problem. In this case, the use of NARX improved the learning process, increasing the performance of a recurrent network. Note that the network was trained using a gradient descent learning algorithm and converged to a minimum error faster than conventional ANNs.