Expert Systems with Applications 40 (2013) 2350–2359

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Genetic wavelet packets for speech recognition

Leandro D. Vignolo *, Diego H. Milone, Hugo L. Rufiner

Research Center for Signals, Systems and Computational Intelligence, Departamento de Informática, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, CONICET, Argentina

ARTICLE INFO

Keywords:
Phoneme classification
Genetic algorithms
Wrappers
Wavelet packets

ABSTRACT

The most widely used speech representation is based on the mel-frequency cepstral coefficients, which incorporates biologically inspired characteristics into artificial recognizers. However, the recognition performance with these features can still be enhanced, specially in adverse conditions. Recent advances have been made with the introduction of wavelet based representations for different kinds of signals, which have shown to improve the classification performance. However, the problem of finding an adequate wavelet based representation for a particular problem is still an important challenge. In this work we propose a genetic algorithm to evolve a speech representation, based on a non-orthogonal wavelet decomposition, for phoneme classification. The results, obtained for a set of spanish phonemes, show that the proposed genetic algorithm is able to find a representation that improves speech recognition results. Moreover, the optimized representation was evaluated in noise conditions.

1. Introduction

Automatic speech recognition systems need a pre-processing stage to make phoneme key-features more evident, in order to obtain significant improvements in the classification results (Rabiner & Juang, 1993). This task was first addressed by signal processing techniques like filter-banks, linear prediction coding and cepstrum analysis (Rabiner & Schafer, 1978). The most popular feature representation currently used for speech recognition is built from the mel-frequency cepstral coefficients (MFCC) (Davis & Mermelstein, 1980), which are based on a linear model of voice production together with the codification on a psycho-acoustic scale. However, due to the degradation of recognition performance in the presence of additive noise, many advances have been conducted in the development of alternative feature extraction approaches. In particular, techniques like perceptual linear prediction (Hermansky, 1990) and relative spectra (Hermansky & Morgan, 1994) incorporate features based on the human auditory system and provides some robustness in ASR. Also, significant progress has been made with the application of different artificial intelligence techniques in the field of speech processing (Hassanien, Schaefer, & Darwish, 2010). Besides, the utilization of wavelet based analysis for speech feature extraction has recently been studied (Avci & Akpolat, 2006; Nehe & Holambe, 2012; Patil & Dixit, 2012; Wu & Lin, 2009).

The multi-resolution analysis associated with discrete wavelet transform (DWT) can be implemented as a filter bank decomposition (or filter bank schemes) (Vetterli & Herley, 1992). Wavelet packet transform (WPT) is a generalization of the DWT decomposition which offers a wider range of possibilities for signal representation in the time-scale plane (Hess-Nielsen & Wickerhouser, 1996). To obtain a representation based on this transform, usually, a particular orthogonal basis is selected among all the available basis. Nevertheless, in phoneme classification applications there is not evidential benefit on working with orthogonal basis. Moreover, it is known that the analysis performed at the level of the auditory cortex is highly redundant and, therefore, non-orthogonal (Munkong & Juang, 2008). Without this restriction the result of the full WPT decomposition is a highly redundant set of coefficients, from which a convenient representation for the problem in hand can be selected.

Many approaches addressing the optimization of wavelet decompositions for feature extraction have been proposed. For instance, in Ray and Chan (2001) an automatic extraction of high quality features from continuous wavelet coefficients according to signal classification criteria was presented. In Wang, Miao, and Xie (2011), an approach based on the best basis wavelet packet entropy method was proposed for electroencephalogram classification. Also, a method for the selection of wavelet family and parameters was proposed for the phoneme recognition task (Rufiner & Goddard, 1997). Similarly, the use of wavelet based decompositions has also been proposed as a tool for the development of robust features for speaker recognition (Kumar & Chandra, 2011;

* Corresponding author. Address: Research Center for Signals, Systems and Computational Intelligence, Departamento de Informática, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Ciudad Universitaria CC 217, Ruta Nacional No. 168 Km 472.4, Santa Fe 3000, Argentina. Tel.: +54 (342) 4575233x191; fax: +54 (342) 4575224.

E-mail address: ldvignolo@fich.unl.edu.ar (L.D. Vignolo).

URL: http://fich.unl.edu.ar/sinc (L.D. Vignolo).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.050

Tiwari & Singhai, 2011). Another interesting approach was proposed in Daamouche et al. (in press), in which a novel approach for generating the wavelet that maximizes the discrimination capability of ECG beats using particle swarm optimization. Also, the use of evolutionary computation techniques in order to optimize over-complete decompositions for signal approximation was proposed in Ferreira da Silva (2003). Furthermore, the use of a genetic algorithm to optimize WPT based features for pathology detection from speech was proposed in Behroozmand and Almasganj (2007), where an entropy criterion was minimized for the selection of the wavelet tree nodes. Similar approaches propose the optimization of wavelet decomposition schemes using evolutionary computation for denoising (Ferreira da Silva, 2005; El-Dahshan, 2011) and image compression (Salvador, Moreno, Riesgo, & Sekanina, 2011). Besides, different approaches have been proposed for the optimization of wavelet based representations using swarm intelligence (Daamouche & Melgani, 2009; Zhao & Davis, 2009). Many other studies also rely on evolutionary algorithms for feature selection (Chatterjee & Bhattacherjee, 2011; Li, Kwong, He, He, & Yang, 2010; Pedrycz & Ahmad, 2012) and the optimization of speech representations (Vignolo, Rufiner, Milone, & Goddard, 2009, 2011a, 2011b). However, the flexibility provided by the full WPT decomposition has not yet been fully exploited in the search for a set of features to improve speech recognition results. When this search is not restricted to non-redundant representations, there is a large number of non-orthogonal dictionaries to be explored, leading to a hard combinatorial problem.

Here we propose a new approach to optimize over-complete decompositions from a WPT dictionary, which consists in the use of a genetic algorithm (GA) for the selection of wavelet based features. In order to evaluate the solutions during the search, the GA uses a learning vector quantization (LVQ) classifier. Some preliminary results with this strategy were presented in Vignolo et al. (2006). The methodology, referred to as *genetic wavelet packets* (GWP), relies on the benefits provided by evolutionary computation to find a better signal representation. This feature selection scheme, known as *wrapper* (Durbha, King, & Younan, 2010; Hsu, Hsieh, & Lu, 2011), is widely used as it allows to obtain the good solutions in comparison with other techniques (Kohavi & John, 1997).

The organization of this paper is as follows. In Section 2, brief descriptions of the properties of WPT and GA are presented. Next, our wrapper method for the selection of the WPT components is described. The following section discusses the obtained recognition results for a set of spanish phonemes. Finally, the general conclusions and future work are presented.

2. Materials and methods

2.1. Wavelet and wavelet packet transforms

In contrast with sine and cosine bases, wavelet bases are simultaneously localized in time and frequency. This feature is particularly interesting in the case of signals which present both stationary and transient behaviors. Wavelets can be defined, in a simplified manner, as a function of zero mean, unitary norm and centered in the neighborhood of $t = 0$ (Mallat, 2008):

$$\psi(t) \in L^2(\mathbb{R}); \quad \int_{-\infty}^{\infty} \psi(t)dt = 0; \quad \|\psi(t)\| = 1. \tag{1}$$

A family of time–frequency atoms is obtained by scaling and translating the wavelet function:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right), \tag{2}$$

with $u,\ s \in \mathbb{R}$. This way, the continuous wavelet transform of the signal $x(t)$ is defined as the inner product with this family of atoms

$$W_x(u,s) = \langle x(t), \psi_{u,s}(t)\rangle = \int_{-\infty}^{+\infty} x(t)\,\psi^*_{u,s}(t)\,dt. \tag{3}$$

The *discrete dyadic wavelet transform* (DWT) of $x[n] \in \mathbb{R}^N$ is obtained by discretizing translation and scaling parameters in (3), as $u = m$ and $s = 2^j$. A fast implementation of the DWT based multiresolution analysis exists (Vetterli & Herley, 1992), which uses low-pass and high-pass filters to decompose a signal to detail ($d_j[n]$) and approximation ($a_j[n]$) coefficients. Since the filter outputs contain half the frequency components of the original signal, both approximation and detail can be sub-sampled by two, maintaining the number of samples. This process is iteratively repeated for the approximation coefficients, increasing the frequency resolution on each decomposition step. As result a binary decomposition tree is obtained, where each level corresponds to a different scale j (Mallat, 1989)

$$d_{j+1}[m] = \sqrt{2}\sum_{n=-\infty}^{\infty} g[n-2m]a_j[n], \tag{4}$$

$$a_{j+1}[m] = \sqrt{2}\sum_{n=-\infty}^{\infty} h[n-2m]a_j[n], \tag{5}$$

here $g[n]$ and $h[n]$ are the impulse responses of the high-pass and low-pass filters associated with the wavelets and scaling functions, respectively.

The WPT could be considered as an extension of the DWT which provides more flexibility on frequency band selection. With the same reasoning above, details (high frequency components) can be decomposed as well as approximations (low frequency components). In a similar way to the DWT, the full wavelet packets decomposition tree is obtained by

$$c^{2p}_{j+1}[m] = \sqrt{2}\sum_{n=-\infty}^{\infty} g[n-2m]c^p_j[n], \tag{6}$$

$$c^{2p+1}_{j+1}[m] = \sqrt{2}\sum_{n=-\infty}^{\infty} h[n-2m]c^p_j[n], \tag{7}$$

where j is the depth of the node and p indexes the nodes in the same depth, every c^p_j with p even is associated to approximations and every c^p_j with p odd is associated to details.

The wavelet packet analysis allows to represent the information contained in a signal in a more flexible time-scale plane, by selecting different sub-trees from the full decomposition (Fig. 1). For the selection of the best tree it is possible to make use of the knowledge about the characteristics of the signal and to obtain an efficient representation in the transform domain. For the case of signal compression the criteria is based on "entropy" measures, method named as *best orthogonal basis* (Coifman & Wickerhauser, 1992). Another possibility, closer to the classification problem, is to use the *local discriminant basis* algorithm, which provides an appropriate orthogonal basis for signal classification (Saito, 1994). These criteria are based on the assumption that an orthogonal basis is convenient. Nevertheless, for the case in study there is not evidence on the convenience of a non-redundant representation. Moreover, the redundancy often provides additional robustness for classification tasks in adverse conditions (Vignolo, Rufiner, Milone, & Goddard, 2011b). Because of this, a method which explores a wider range of possibilities should be studied.

2.2. Genetic wavelet packets

Genetic algorithms are meta-heuristic optimization methods motivated by the process of natural evolution (Sivanandam & Deepa, 2008). A classic GA consists of three kinds of operators: