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

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.

Key words:
Phoneme classification, genetic algorithms, wrappers, wavelet packets

*Corresponding author.
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, TE: +54(342)4575233 ext 191, FAX: +54(342)4575224, Santa Fe (3000), Argentina.

Email address: `ldvignolo@fich.unl.edu.ar` (Leandro D. Vignolo)

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

Preprint submitted to Expert Systems with Applications *September 18, 2012*

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 [1]. This task was first addressed by signal processing techniques like filter-banks, linear prediction coding and cepstrum analysis [2]. The most popular feature representation currently used for speech recognition is built from the mel-frequency cepstral coefficients (MFCC) [3], 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 [4] and relative spectra [5] 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 [6]. Besides, the utilization of wavelet based analysis for speech feature extraction has recently been studied [7, 8, 9, 10].

The multi-resolution analysis associated with discrete wavelet transform (DWT) can be implemented as a filter bank decomposition (or filter bank schemes) [11]. 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 [12]. 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 [13]. 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 [14] an automatic extraction of high quality features from continuous wavelet coefficients according to signal classification criteria was presented. In [15], 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 [16]. Similarly, the use of wavelet based decompositions has also been proposed as a tool for the development of robust features for speaker recognition [17, 18]. Another interesting approach was proposed in [19], 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 [20]. Furthermore, the use of a genetic algorithm to optimize WPT based features for pathology detection from speech was proposed in [21], 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 [22, 23] and image compression [24]. Besides, different approaches have been proposed for the optimization of wavelet based representations using swarm intelligence [25, 26]. Many other studies also rely on evolutionary algorithms for feature selection [27, 28, 29] and the optimization of speech representations [30, 31, 32]. 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 [33]. 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* [34, 35], is widely used as it allows to obtain the good solutions in comparison with other techniques [36].

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 $t = 0$ [37]:

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

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

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

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 [11], 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 [38]

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

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

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

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

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

where j is the depth of the node and p indexes the nodes in the same depth, every c_j^p with p even is associated to approximations and every c_j^p 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 (Figure 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* [39]. Another possibility, closer to the classification problem, is to use the *local discriminant basis* algorithm, which provides an appropriate orthogonal basis for signal classification [40]. 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 [31]. 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 [41]. A classic GA consists of three kinds of operators: selection, variation and replacement [42]. Selection mimics the natural advantage of the fittest individuals, giving them more chance to

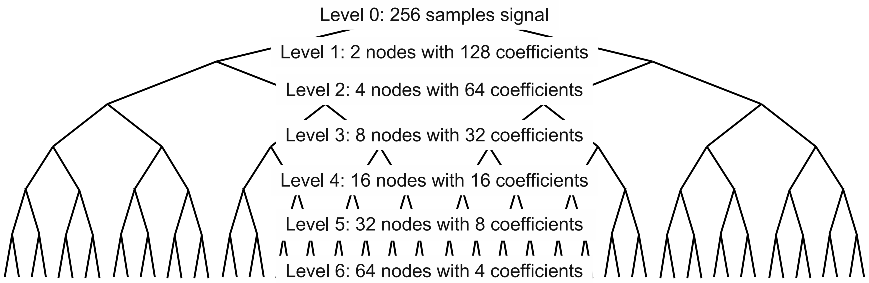
Level 0: 256 samples signal
Level 1: 2 nodes with 128 coefficients
Level 2: 4 nodes with 64 coefficients
Level 3: 8 nodes with 32 coefficients
Level 4: 16 nodes with 16 coefficients
Level 5: 32 nodes with 8 coefficients
Level 6: 64 nodes with 4 coefficients

Figure 1: Wavelet packets tree with six decomposition levels for a 256 samples signals.

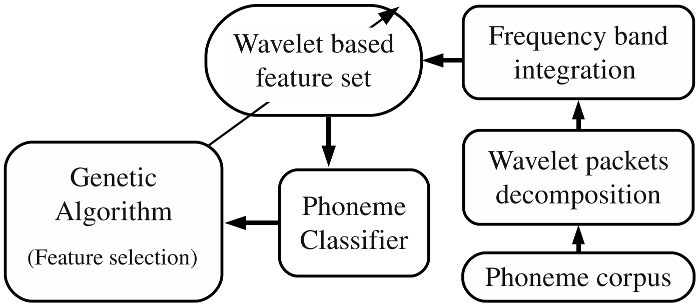
Wavelet based feature set
Frequency band integration
Genetic Algorithm
(Feature selection)
Phoneme Classifier
Wavelet packets decomposition
Phoneme corpus

Figure 2: General scheme of the proposed wrapper method.

reproduce. The purpose of the variation operators is to combine information from different individuals and also to maintain population diversity, by randomly modifying chromosomes. The number of individuals in the current population that are substituted by the offspring is determined by the replacement strategy. The information of a possible solution is coded in a chromosome and its fitness is measured by an objective function, which is specific to a given problem. Parents, selected from the population, are mated to generate the offspring by means of the variation operators. The population is then replaced and the cycle is repeated until a desired termination criterion is reached. Once the evolution is finished, the best individual in the population is taken as the solution for the problem [43]. Genetic algorithms are inherently parallel, and one can easily benefit from this to increase the computational speed [33].

In this case, the objective function needs to evaluate the signal representation suggested by a given chromosome, providing a measure of the class separability. Therefore, the fitness function was defined as a phoneme classifier, based on the optimized learning vector quantization (O-LVQ) tech-

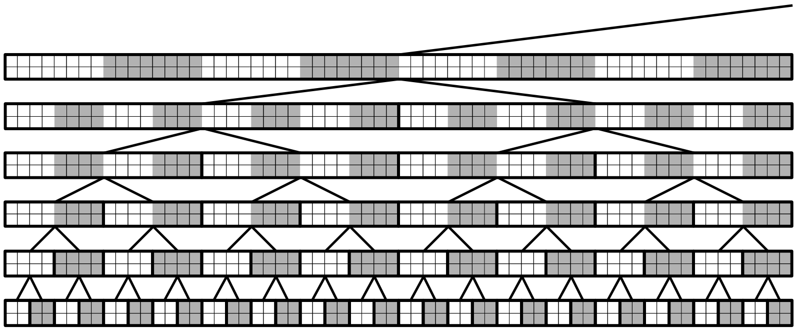

Figure 3: Frequency band integration scheme (half tree).

nique [44]. This classifier uses a set of reference vectors (*codebook*) which are adapted using a set of training patterns in order to represent the distribution of classes. O-LVQ was chosen because it requires much less processing time than those used in state-of-the-art speech recognizers, based on hidden Markov models (HMM). Although, after the optimization, for the validation of the evolved representation an HMM based classifier was also used.

For the evaluation of every individual in the population, the classifier is trained and tested on a phoneme corpus, and the recognition rate is used as the fitness value. The scheme of the proposed wrapper method is shown in Figure 2. The GA uses roulette wheel selection method, the classic mutation and one-point crossover. Also, an elitist replacement strategy was applied, which maintains the best individual to the next generation.

The feature extraction scheme was designed for signals of 256 samples length, this is 32 ms frames at 8 kHz sampling frequency. And the iterative process of filtering and decimation was performed to obtain six decomposition levels, obtaining a full wavelet packet tree consisting of 1792 coefficients. Then, in order to reduce the dimensionality of the search space, the coefficients inside each frequency band were “integrated” by groups. This means that each band was subdivided into groups, and an energy coefficient for each group was obtained by

$$e_j^s = \sum_{\forall c_k \in G_j^s} |c_k|^2, \tag{8}$$

where e_j^s is the energy coefficient for integration group j in scale s, G_j^s, and c_k is the k-th coefficient belonging to this group. Figure 3 illustrates the integration scheme for the first half of the WPT decomposition tree, while the other half is integrated in the same manner. Each small square represents

Table 1: Integration scheme applied to the WPT tree for a 256 sample signal, which reduces from 1792 wavelet coefficients to 208 integration coefficients.

Level	1	2	3	4	5	6
Nodes	2	4	8	16	32	64
Integration groups per node	8	8	4	2	1	1
Wavelet coefficients per group	16	8	8	8	8	4
Integration coefficients per level	16	32	32	32	32	64

a single component, in the first row (level 0) this is a sample of the temporal signal and for the other rows (levels 1 to 6) each of them correspond to a single wavelet coefficient. White and gray zones delimit the different integration groups and the tree nodes in each decomposition level are indicated with thick line rectangles. Table 1 shows the number of components and coefficients in each integration group. This integration scheme was heuristically designed, considering the most relevant frequency bands in speech and their temporal resolutions.

After band integration, a normalization was applied: if $w_p[k]$ is the k-th energy coefficient corresponding to the pattern p, then the normalized coefficient will be given by

$$\hat{w}_p[k] = \frac{w_p[k]}{\max\limits_i \{w_i[k]\}}. \tag{9}$$

A canonical evolution model with binary chromosomes was used, in which every individual represents a different selection of the WPT band-integrated coefficients. Each gene in a chromosome represents one normalized coefficient, and its value indicates whether that coefficient should be used to parametrize the signals (Figure 4). Once the data base is processed, each feature vector is composed by the normalized and band-integrated WPT coefficients. These labeled patterns are used to train and test the O-LVQ based classifier. When a particular individual is evaluated, each feature vector is reduced to the subset of coefficients indicated by the chromosome.

The selection of individuals should be done considering the set of coefficients represented by each chromosome. The chromosomes which codify the best signal representations, those which allow better classification results, should be assigned high probability. As the codification could be redundant and no restriction is imposed for coefficients combination, the GA initialization consists on a random settling of the genes in the chromosomes. All

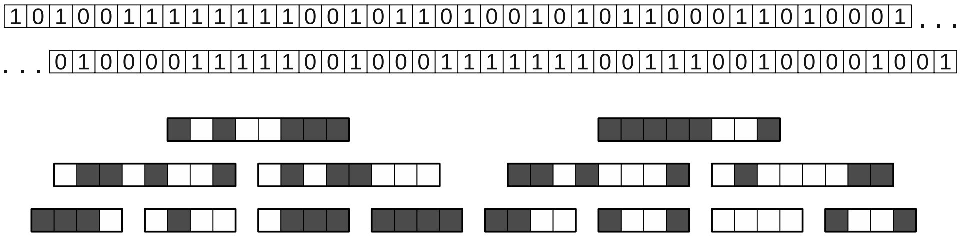

Figure 4: Codification example with a 80 genes chromosome and the corresponding WPT tree. Dark boxes represent the used coefficients and the white boxes represent those that are discarded.

Algorithm 1: Optimization for GWP.

```
Obtain full WPT for each phoneme example in the corpus using (6) and (7)
Apply the band-integration scheme to WPT coefficients using (8)
Normalize the integrated coefficients for each pattern according to (9)
Initialize the GA population
**Evaluate population** (Algorithm 2)
**repeat**
    Select parents (roulette wheel)
    Create a new population from selected parents
    Replace population
    **Evaluate population**
**until** *stopping criteria is met*
```

the steps involved in the GWP feature selection strategy are summarized in Algorithm 1, while the details for the population evaluation are shown in Algorithm 2.

2.3. Phoneme corpus

The speech data used for experimentation is a subset of the Albayzin geographic corpus [45, 46], named Minigeo. This subset consists of 600 utterances, spoken by twelve different speakers (six men and six women) which where between fifteen and fifty five years old. This speech data has been phonetically segmented using a speech recognition system based on hidden Markov models [47]. This way the temporal localization of every phoneme within each utterance was obtained. The extracted phonetic corpus was partitioned in three groups, a training set and a testing set to be used during evolution, and a third data set which was used only for the validation of best

Algorithm 2: Evaluate population.

```
for each individual in the population do
    Re-parameterize training/test patterns according to the chromosome
    Train the LVQ based classifier on the training set
    Test the LVQ based classifier on the test set
    Assign classification rate as the current individual's fitness
```

solutions after the feature selection process.

The experiments included the phonemes /a/, /e/, /i/, /o/, /u/, /b/, /d/, /p/ and /t/ from Spanish. The five vowels were included because of their obvious importance in the language, while the four occlusive phonemes have been included because of their similar characteristics, that make a set particularly difficult to distinguish [48]. Even though it can be suggested that all phonemes should be included, our hypothesis is that this strategy simplifies the task of the GA while still allows to find features useful in continuous ASR. In order to avoid adding additional complexity to the search of the GA, every sample used in the optimization consisted in a single speech frame of 32 ms length, which was extracted from the center of the phoneme utterance.

3. Results and discussion

3.1. Genetic optimization of wavelet packets

In [49], various wavelet families have been tested in order to find which one is the most convenient for signal classification. In this work the tests included the most widely used families, among which we can mention Meyer, Daubechies, Symmlets, Coiflets y Splines [50]. As result, the 4th order Coiflet family was chosen to be used on the following experiments.

For the first experiment, a codebook of 117 vectors (13 per phoneme) was used within the LVQ classifier and the initial learning rate was set on 0.02. The classifier training was made in 6 epochs with 1449 patterns and 252 patterns were left for testing. For the GA, the population size was set on 100 individuals, while crossover and mutation probabilities were set on 0.9 and 0.05, respectively. The performance of the best solution found was 57.94% of correct classifications.

As the LVQ codebook initialization has a random component, repeated evaluations for the same individual may result on different fitness values.

Table 2: Summary of obtained fitness and validation results.

Strategy	Convergence	Fitness	Validation	
	(# generations)		average	STD
Random LVQ initialization	26	57.94%	53.69%	2.33%
Fixed LVQ initialization	216	57.78%	56.67%	2.9%
Generational LVQ initialization	355	64.07%	59.16%	2.91%

Then, in order to obtain a good estimation of the performance for the best individual, after the evolution, it was evaluated ten times with a validation data set In this process we used 2637 training patterns and 450 test patterns which were not used during the evolution. As result, an average of 53.69% correct classifications with a standard deviation of 2.33% was obtained. Also, in order to analyze the effect of the random LVQ initialization on the evolution, two different alternatives were considered. In the first case, the randomness was eliminated from the codebook initialization. In this case, the GA was able to find an individual with a fitness (classification rate) of 57.78%. With the validation procedure, described earlier, an average classification rate of 56.67% was obtained. Even though an improvement was obtained, it is possible that the evolution was biased by this fixed codebook initialization. Then, another strategy was considered, in which a fixed initialization sequence was used for each generation. In order to allow individuals evaluated with different conditions (initializations) to coexist in the same generation a generational gap of 10 individuals was used, maintaining more information from one generation to the next. This means that, in addition to the best individual which is preserved by the elitist strategy, another 10 individuals are chosen by the selection algorithm to be maintained to the next generation. In this case the best solution achieved 64.07% correct classifications, and an average classification rate of 59.16% was obtained with the validation data. Table 2 summarizes these results, showing that this last strategy allowed to improve the generalization capability.

Table 3 shows a confusion matrix obtained from validation results. As this matrix shows, the /t/ phoneme is mainly (61.51%) classified as /p/. This error turn up because the experimental data was taken from the central part of the samples and the plosive phonemes, like /p/ and /t/, have their most particular attributes at the beginning (the phoneme plosion). A similar problem happens with phoneme /d/, and this might be solved for all plosive phonemes by considering their context (i.e. a number of precedent

Table 3: Confusion matrix obtained from the validation of the best GA solution, giving 59.16% average classification rate .

	/a/	/e/	/i/	/o/	/u/	/b/	/d/	/p/	/t/
/a/	**84.85**	00.30	00.00	11.82	01.21	01.21	00.30	00.30	00.00
/e/	02.73	**76.06**	01.82	05.15	03.64	03.33	01.51	03.64	02.12
/i/	00.00	08.18	**86.97**	00.00	00.30	02.73	00.61	00.30	00.91
/o/	25.15	10.61	00.00	**42.42**	14.24	04.24	02.73	00.61	00.00
/u/	08.79	00.00	01.51	08.48	**59.39**	14.24	00.61	05.15	01.82
/b/	00.30	02.42	00.00	04.54	09.70	**62.12**	06.06	13.03	01.82
/d/	10.30	31.82	09.09	07.57	04.54	04.54	**10.61**	17.27	04.24
/p/	00.00	00.00	00.00	00.00	00.00	03.03	02.42	**78.18**	16.36
/t/	00.00	00.00	00.00	00.30	00.00	04.54	01.82	61.51	**31.82**

and posterior frames).

3.2. Comparative analysis

In order to compare this results, the same classifiers were trained with other state-of-the-art speech features: the classic MFCC [3], the alternative cepstral features based on Slaney's filter-bank [51], and a representation based on the standard DWT. It is worth pointing out that, as the speech data used in these experiments was sampled at 8 kHz, the original parameters of the filter-bank proposed by Slaney were modified by following the reasoning in [52]. Table 4 shows the results obtained with the above mentioned representations and using a validation data set consisting of 450 patterns, which were not used for the optimization. As it can be seen, the best average classification rate was obtained with the GWP. This shows that by the genetic optimization of the full WPT decomposition it is possible to improve the classification results, in contrast to the classic cepstrum based representations. On the other hand, it can be noticed that the phoneme /d/ seems to be the most difficult for all representations, except for DWT.

Even though the state-of-the-art speech recognizers use HMM for acoustic modeling, the significant improvements provided by GWP when using an LVQ based classifier should not be disregarded. These results clearly show that a more efficient class separation is provided in the GWP features space. It should be taken into account that this straightforward classifier was used as objective function to guide evolution, and only the central frame was considered for each phoneme pattern during the optimization.

Table 4: Classification results obtained with an LVQ based classifier.

Phoneme	Cepstral coefficients		Wavelets	
	MFCC(13)	Slaney(18)	DWT(256)	**GWP**(104)
/a/	82.80	70.40	69.39	**84.85**
/e/	**77.40**	61.60	54.54	76.06
/i/	**90.00**	61.60	84.54	86.97
/o/	46.20	28.40	**54.54**	42.42
/u/	31.60	21.40	31.51	**59.39**
/b/	45.20	39.60	58.48	**62.12**
/d/	08.80	09.00	**59.69**	10.61
/p/	55.60	45.20	04.85	**78.18**
/t/	48.60	**58.40**	31.21	31.82
Average	54.02	43.96	49.86	**59.16**

3.3. Evaluation with hidden Markov models

In this work we have raised the hypothesis that, by using a simple classifier in the optimization, the class separability would be maximized and this could also be beneficial to a more sophisticated classifier, like HMM [53]. In order to verify this, the performance of an HMM based classifier was evaluated for each of the representations in Table 4. This classifier is based on a continuous HMM, using Gaussian mixtures with diagonal co-variance matrices for the observation densities, as common in ASR [54]. We used a three state model with mixtures of four gaussians, constructed with the tools provided in the HMM Toolkit (HTK) [47]. These tools use the Baum-Welch algorithm [55] to train the HMM parameters, and the Viterbi algorithm [53] to search for the most likely state sequence. This classifier was evaluated in a ten-fold cross-validation process with random partitions, each of which consisted of 2484 and 621 patterns for training and testing, respectively. It is important to point out that, because of the nature of HMM, in the evaluation of this classifier all the successive frames composing a phoneme were used. While, for the LVQ based classifier, only the central frame was used for a particular phoneme.

For the features based on DWT the training of the HMM classifier could not converge, which is mainly because the gaussian mixtures are not able to adequately model the probability densities of these coefficients [56]. Another problem for training the models with DWT coefficients is due to the high dimensionality of this representation. Then, a post-processing based on principal component analysis (PCA) [57] was applied in order to obtain a representation of lower dimensionality, with probability densities more similar to

Table 5: Classification results obtained with an HMM based classifier.

Phoneme	Cepstral coefficients		Wavelets	
	MFCC(13)	Slaney(18)	DWT+PCA(134)	**GWP**(104)
/a/	**59.70**	58.26	49.71	54.21
/e/	**67.55**	64.21	45.50	60.30
/i/	59.00	63.49	62.02	**76.82**
/o/	31.74	27.53	27.38	**34.78**
/u/	37.68	51.02	37.82	**58.99**
/b/	**43.76**	26.08	42.61	30.59
/d/	**30.72**	16.52	22.45	26.81
/p/	36.96	40.00	36.37	**49.44**
/t/	**71.46**	60.00	59.43	53.33
Average	48.74	45.24	42.60	**49.48**

gaussians. The best result for DWT+PCA was obtained when preserving the 99% of the variance, giving a representation of 134 dimensions. Even though our optimized representation is also based on wavelets, and the same problem could be expected, no post-processing was necessary for GWP. Thus, we can assume that the band integration, besides reducing dimension, produced coefficients with probability densities more appropriate to gaussian mixture modeling.

Table 5 shows the phoneme classification results obtained with HMM, comparing GWP and other state-of-the-art speech representations. The optimized representation is the same from Table 4, which was evolved using an LVQ-based classifier. It can be noticed that the best results are those obtained by means of the optimized representation and MFCC, similar to the case of the validation with LVQ (Table 4). Even though the fitness was measured by means of an LVQ based classifier in the optimization, the evolved representation provided satisfactory results when using HMM. This means that the optimized representation captures information which is relevant for the discrimination, regardless of the type of classifier. Moreover, using this low-cost classifier, we have successfully saved significant computational time in the optimization. It should be taken into account that if we had used an HMM based classifier and, therefore, considered all the possible frames within a phoneme example, the evaluation of each individual would have taken approximately ten times more.

It is also important to note that the proposed GWP representation, which yielded the best classification result when using an LVQ based classifier, provides relatively lower performance with HMM. This is because, as explained before, the probability densities of the coefficients provided in wavelet based representations are not entirely suitable for gaussian mixture models [56]. Then, different alternatives remains to be explored besides band integration and PCA post-processing, in order to obtain coefficients more suitable to gaussian mixture models. Also, it should be considered that the dimensionalities of the wavelet based representations are much higher than those of the cepstral representations, which makes the training of the classifier more difficult.

Despite the previous considerations, the results from Tables 4 and 5 show that the proposed method is useful for the optimization of wavelet based representations. Moreover, the results obtained with the GWP features shown that using this evolutionary methodology it is possible to improve the performance of the classical representations in ASR.

3.4. Evaluation in noise conditions

In order to evaluate the robustness of the optimized representation, white noise was added to the original utterances. The tests were made at several noise levels, and the *mismatch training and test* (MMTT) condition was considered, which means that the classifiers were trained with clean signals and tested with noisy signals. In general, the input of an ASR system would consist on speech signals with different SNR to those in the training set. For this reason, the evaluation of the recognition performance in MMTT conditions is more realistic than the case where the SNR is the same in both training and test sets.

Each test consisted in a ten-fold cross-validation process with random partitions, which consisted of 2484 and 621 patterns for training and testing, respectively. The process of training and testing was repeated for these ten partitions and results were averaged, ensuring that the resultant accuracy would not be biased because of a particular data partition. Figure 5 shows the average results, as well as the estimated standard deviations, showing that the GWP improves the MFCC in all cases. At 0 and 5 dB SNR the DWT+PCA representation gives better results, however, for most of the noise conditions its performance is noticeably worse than GWP. It can be noticed that, on average, the results given by GWP are significantly better when compared to the other representations.

In order to evaluate the statistical significance of these results, we estimated the probability that GWP is better than each of the reference representations for the given tasks. To perform this test, statistical independence of the

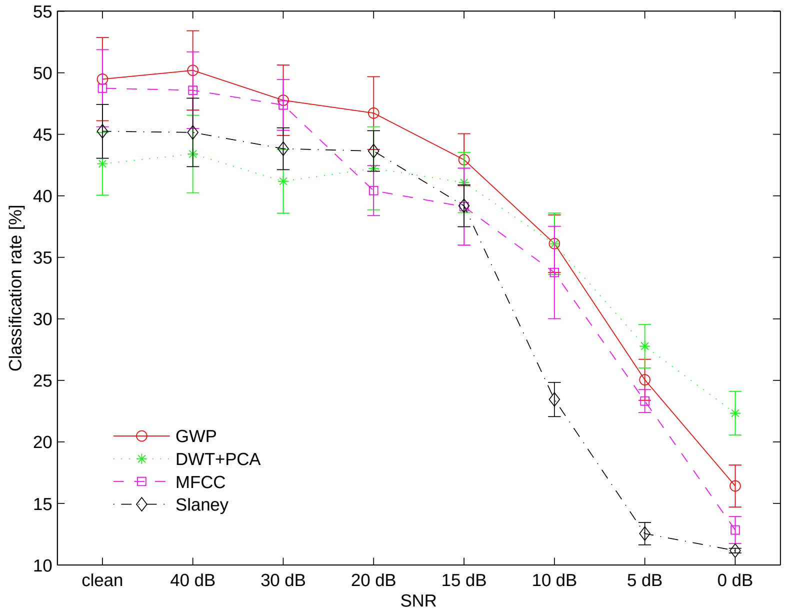

Figure 5: Classification results obtained with an HMM based classifier evaluated in MMTT conditions.

classification errors for each phoneme has been assumed, and the binomial distribution of the errors was approximated by means of a Gaussian distribution. Table 6 shows the statistical significance of the classification rates obtained with the HMM based classifier: results improved by GWP with statistical significance higher than 97% are indicated with the superscript $\triangle$. It can be noticed that most of the classification improvements obtained with the proposed optimized representation are statistically significant. For instance, the probability that GWP performs better than the cepstral coefficients obtained with Slaney's filterbank is higher than 0.99 for all SNRs.

Table 7 shows more detailed information about the classification performance comparing MFCC and GWP, for the case of 40 dB SNR in MMTT conditions. In these confusion matrices, the rows correspond to the actual phoneme and the columns to the predicted phoneme, while the percentages of correct classification are on the diagonal. These matrices show coincidences between the phonemes which are most confused with MFCC and the ones that are confused with GWP. For example, in both cases phoneme /t/ was confused with /p/ and vice versa, which is reasonable as these two plosive

Table 6: Statistical significance of classification results obtained with an HMM based classifier evaluated in MMTT conditions. Superscript $\triangle$ indicates that the statistical significance of the improvement of GWP is higher than 97%.

SNR	Cepstral coefficients		Wavelets	
	MFCC(13)	Slaney(18)	DWT+PCA(134)	**GWP**(104)
clean	48.74	45.24$^{\triangle}$	42.60$^{\triangle}$	**49.48**
40 dB	48.58$^{\triangle}$	45.15$^{\triangle}$	43.40$^{\triangle}$	**50.19**
30 dB	47.38	43.82$^{\triangle}$	41.18$^{\triangle}$	**47.77**
20 dB	40.42$^{\triangle}$	43.64$^{\triangle}$	42.23$^{\triangle}$	**46.72**
15 dB	39.12$^{\triangle}$	39.19$^{\triangle}$	41.08$^{\triangle}$	**42.93**
10 dB	33.77$^{\triangle}$	23.45$^{\triangle}$	36.10	**36.11**
5 dB	23.32$^{\triangle}$	12.55$^{\triangle}$	**27.78**	25.05
0 dB	12.84$^{\triangle}$	11.17$^{\triangle}$	**22.34**	16.42

Table 7: Confusion matrix obtained from the validations with MFCC and GWP in MMTT conditions and 40 dB SNR.

	MFCC									**GWP**								
	/a/	/e/	/i/	/o/	/u/	/b/	/d/	/p/	/t/	/a/	/e/	/i/	/o/	/u/	/b/	/d/	/p/	/t/
/a/	***61.7***	08.8	00.2	11.9	02.8	04.5	09.6	00.3	00.3	***57.8***	08.3	00.0	13.2	11.6	05.1	04.1	00.0	00.0
/e/	11.0	***66.4***	06.5	04.8	03.9	03.7	03.8	00.0	00.0	09.1	***61.0***	08.4	04.4	01.2	02.8	11.7	00.3	01.2
/i/	00.2	24.6	***60.9***	02.6	06.0	03.8	02.0	00.0	00.0	00.2	06.8	***77.5***	00.7	03.1	00.6	09.4	00.7	01.0
/o/	15.5	17.1	03.2	***32.3***	13.6	13.3	03.5	00.3	01.2	08.6	10.3	01.2	***35.7***	29.3	05.5	05.7	02.6	01.3
/u/	04.8	04.2	07.4	19.6	***38.0***	15.9	09.0	00.9	00.3	02.3	01.8	02.9	16.1	***57.1***	13.8	04.1	02.1	00.0
/b/	04.7	01.3	03.4	09.9	05.9	***45.2***	18.6	09.9	01.3	04.5	00.3	00.0	11.5	25.8	***29.7***	17.0	06.2	05.1
/d/	06.1	37.0	00.9	03.1	02.8	07.4	***27.8***	05.7	09.4	06.1	23.2	09.4	05.4	07.7	05.4	***26.8***	07.4	08.7
/p/	00.0	00.4	00.0	00.0	01.3	06.0	16.7	***34.1***	41.6	00.0	00.0	00.0	00.2	00.7	03.2	10.0	***52.5***	33.5
/t/	00.0	00.0	02.3	00.2	00.2	02.6	07.0	16.8	***71.0***	00.0	01.0	00.7	00.4	00.9	00.6	12.0	30.7	***53.6***
	Average: 48.58%									Average: 50.19%								

consonants share many spectral features. In a similar way vowels /o/ and /u/, which are close in the formants map, are quite confused in both cases. Similarly, Table 8 shows the confusion matrices obtained in the classification with MFCC and GWP in the case of 15 dB SNR and MMTT conditions. It can be noticed that the vowels /a/ and /u/ are mostly misclassified with MFCC, but they are significantly better distinguished with GWP. The optimized features also introduced an important improvement for the vowel /i/, which is confused with the phoneme /d/ when using MFCC. It is also interesting to note that the phonemes which have their classification rates most affected by noise when using MFCC and GWP do not match. Though, when the noise level is increased, the number of confusions between phonemes /t/ and /d/ increases for both MFCC and GWP. Similarly, the number of confusions between phonemes /t/ and /p/ is also increased in both cases. Another interesting remark is that, for both representations, phoneme /t/ is

Table 8: Confusion matrix obtained from the validations with MFCC and GWP in MMTT conditions and 15 dB SNR.

	MFCC									**GWP**								
	/a/	/e/	/i/	/o/	/u/	/b/	/d/	/p/	/t/	/a/	/e/	/i/	/o/	/u/	/b/	/d/	/p/	/t/
/a/	***04.5***	31.0	00.0	20.1	00.0	03.2	32.8	08.1	00.3	***43.8***	17.3	00.5	18.1	09.8	03.4	07.3	00.0	00.0
/e/	00.0	***80.6***	03.3	00.0	00.0	00.3	15.8	00.0	00.0	04.2	***60.3***	17.7	03.5	01.3	00.9	11.0	00.0	01.2
/i/	00.0	19.1	***53.1***	00.0	00.2	00.2	24.4	00.0	03.2	00.0	11.0	***82.5***	01.2	03.4	00.3	01.7	00.0	00.0
/o/	00.0	25.2	01.8	***26.7***	01.3	09.6	24.6	05.4	05.5	07.1	13.6	00.8	***31.9***	29.6	02.5	12.5	00.2	02.1
/u/	00.0	08.4	11.6	16.8	***03.6***	27.8	29.4	01.3	01.0	02.3	04.4	06.1	19.1	***57.5***	05.2	05.4	00.0	00.0
/b/	00.0	03.7	04.1	03.5	00.4	***12.0***	56.7	08.7	11.0	01.5	09.3	08.6	18.1	29.9	***09.6***	14.9	01.6	06.7
/d/	00.0	29.6	01.2	00.6	00.0	02.8	***49.0***	03.4	13.6	04.5	27.7	12.3	02.1	11.4	00.9	***27.5***	02.2	11.5
/p/	00.0	00.5	00.0	00.0	00.0	00.0	05.1	***48.1***	46.4	00.0	00.5	06.8	01.6	08.7	01.6	19.4	***12.0***	49.4
/t/	00.0	00.0	00.0	00.0	00.0	00.0	13.3	12.2	***74.5***	00.0	01.9	06.7	00.0	03.2	00.7	15.7	10.6	***61.3***
	Average: 39.12%									Average: 42.93%								

better classified when the SNR is 15 dB than when it is 40 dB. These results show that the classification performance of the classical representation can be improved. This suggest that, by means of this GWP methodology, additional robustness against white noise can be provided to a state-of-the-art ASR system.

In order to perform a qualitative analysis of the optimized representation, the tiling of the time-frequency plane was constructed from the selected decomposition using the criteria proposed in [58]. The result is shown in Figure 6, where each decomposition level is depicted separately for an easier interpretation. Each ellipse represents a selected group from the integration scheme (Table 1), then their widths and time localizations are determined by the time-frequency atoms corresponding to the integration group (Figure 3). Therefore, each element in the tiling represents a time-frequency atom that was obtained by adding the original wavelet atoms, according to the integration scheme. This explains why the atoms for levels 1 and 2 are the same time width, as the number of coefficients integrated in the groups for level 1 are twice the number of coefficients in the groups for level 2 (Figure 3). The same explanation applies for the width of the atoms in levels 5 and 6. Note that the whole time-frequency tiling is obtained by the superposition of these six sub-figures, yielding great overlapping between atoms from different decomposition levels. A first observation is that the optimization of the WPT-based decomposition led to a highly redundant non-orthogonal representation, which has been able to exploit this redundancy in order to increase robustness against additive noise. However, the optimized representation uses only 50% of the total of the coefficients obtained from the integration of the whole WPT-tree. This also suggest that it could be possible to achieve still further redundant and robust representations. It is also interesting to note that there are some selected atoms concentrated at the center of the time axis, which could be related to how the phonemes were sampled from the speech corpus, as only the frames extracted from the center of each utterance were considered for the optimization. Also, there are some atoms concentrated at the side parts of the time axis, which could be related to the plosive phonemes.

4. Conclusion and future work

A wrapper optimization strategy has been proposed, taking advantage of the benefits provided by evolutionary computation techniques, in order to carry on the search for an advantageous wavelet-based speech representation. The results, obtained in the classification of a group of nine phonemes from spanish, shown that the optimized representation provides important improvements in comparison to the classical features. This suggests that the task of a classifier is simplified when using this optimized representation, due to a better class separation in the features space. Therefore, the proposed strategy provides an alternative feature set for speech signals, which allows to improve the classification results in the presence of noise.

In future work we would design more specific genetic operators, so that more information about the problem in hand could be incorporated to the search. Pursuing the objective of finding a representation more suitable for HMM with gaussian mixture modeling, one interesting idea is to incorporate some measure about the gaussianity of the probability densities of the GWP coefficients to the fitness function. Besides, we would study different alternatives to the proposed band integration scheme and the use of temporal information, in regard to successive speech frames, like first and second derivatives.

In order to obtain a representation which allows to improve the results in a continuous speech recognition system, future experiments will include more phonemes in the data-sets used for the optimization. Also, the robustness of the representation will be evaluated in comparison with different state-of-the-art robust representations, considering different noise types.

References

[1] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice Hall, NJ, 1993.

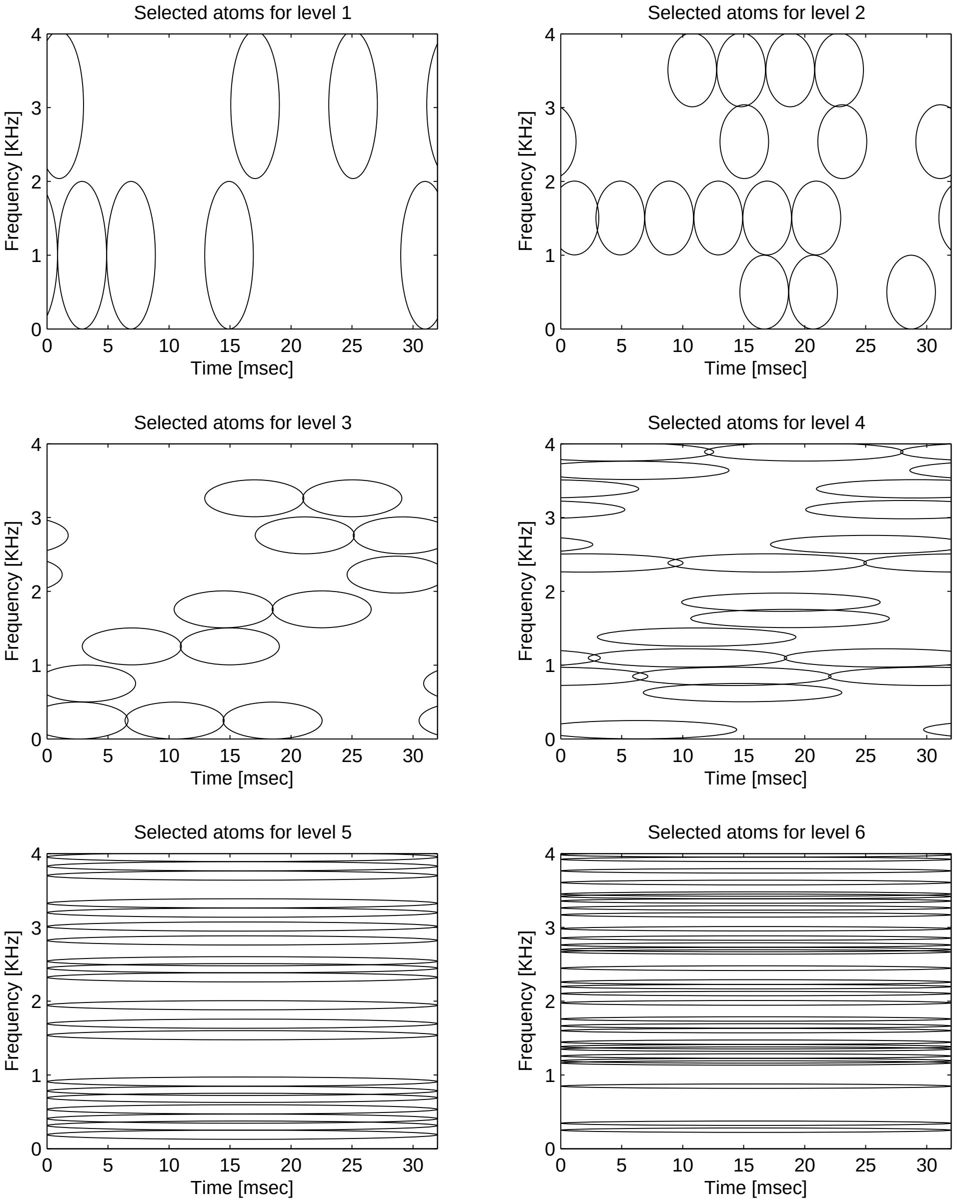

Figure 6: Tiling of the time-frequency plane obtained for the optimized decomposition. For a better visualization, each decomposition level was schematized separately.

[2] L. Rabiner, R. Schafer, Digital Processing of Speech Signals, Prentice Hall, NJ, 1978.

[3] S. V. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech and Signal Processing 28 (1980) 57–366.

[4] H. Hermansky, Perceptual linear predictive (plp) analysis of speech, J Acoust Soc Am 87 (4) (1990) 1738–1752. doi:10.1121/1.399423.

[5] H. Hermansky, N. Morgan, Rasta processing of speech, IEEE Trans Speech Audio Process 2 (1994) 578–589. doi:10.1109/89.326616.

[6] A. Hassanien, G. Schaefer, A. Darwish, Computational intelligence in speech and audio processing: Recent advances, in: X.-Z. Gao, A. Gaspar-Cunha, M. Kppen, G. Schaefer, J. Wang (Eds.), Soft Computing in Industrial Applications, Vol. 75 of Advances in Intelligent and Soft Computing, Springer Berlin / Heidelberg, 2010, pp. 303–311, 10.1007/978-3-642-11282-9-32.

[7] N. Nehe, R. Holambe, DWT and LPC based feature extraction methods for isolated word recognition, EURASIP Journal on Audio, Speech, and Music Processing 2012 (1) (2012) 7. doi:10.1186/1687-4722-2012-7.
URL `http://asmp.eurasipjournals.com/content/2012/1/7`

[8] S. Patil, M. Dixit, Speaker independent speech recognition for diseased patients using wavelet, Journal of the Institution of Engineers (India): Series B 93 (2012) 63–66, 10.1007/s40031-012-0010-3.
URL `http://dx.doi.org/10.1007/s40031-012-0010-3`

[9] E. Avci, Z. H. Akpolat, Speech recognition using a wavelet packet adaptive network based fuzzy inference system, Expert Systems with Applications 31 (3) (2006) 495 – 503. doi:10.1016/j.eswa.2005.09.058.

[10] J.-D. Wu, B.-F. Lin, Speaker identification using discrete wavelet packet transform technique with irregular decomposition, Expert Systems with Applications 36 (2, Part 2) (2009) 3136 – 3143.

[11] M. Vetterli, C. Herley, Wavelets and filter banks: Theory and design, IEEE Trans. Signal Proc. 40 (10) (1992) 2207–2232.

[12] N. Hess-Nielsen, M. V. Wickerhouser, Wavelets and Time-Frequency Analisys, Proceedings of the IEEE 84 (4) (1996) 523–540.

[13] R. Munkong, B.-H. Juang, Auditory perception and cognition, Signal Processing Magazine, IEEE 25 (3) (2008) 98–117. doi:10.1109/MSP.2008.918418.

[14] S. Ray, A. Chan, Automatic feature extraction from wavelet coefficients using genetic algorithms, in: Proceedings of the 2001 IEEE Signal Processing Society Workshop, Neural Networks for Signal Processing XI, 2001, pp. 233–241.

[15] D. Wang, D. Miao, C. Xie, Best basis-based wavelet packet entropy feature extraction and hierarchical eeg classification for epileptic detection, Expert Systems with Applications 38 (11) (2011) 14314 – 14320. doi:10.1016/j.eswa.2011.05.096.

[16] H. Rufiner, J. Goddard, A method of wavelet selection in phoneme recognition, in: Circuits and Systems, 1997. Proceedings of the 40th Midwest Symposium on, Vol. 2, 1997, pp. 889 –891 vol.2.

[17] P. Kumar, M. Chandra, Hybrid of wavelet and mfcc features for speaker verification, in: Information and Communication Technologies (WICT), 2011 World Congress on, 2011, pp. 1150 –1154. doi:10.1109/WICT.2011.6141410.

[18] V. Tiwari, J. Singhai, Wavelet Based Noise Robust Features for Speaker Recognition, International Journal of Signal Processing 5 (2) (2011) 52 – 64.

[19] A. Daamouche, L. Hamami, N. Alajlan, F. Melgani, A wavelet optimization approach for ecg signal classification, Biomedical Signal Processing and Control. In Press. doi:10.1016/j.bspc.2011.07.001.

[20] A. R. Ferreira da Silva, Approximations with evolutionary pursuit, Signal Processing 83 (3) (2003) 465–481.

[21] R. Behroozmand, F. Almasganj, Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis, Computers in Biology and Medicine 37 (4) (2007) 474 – 485. doi:10.1016/j.compbiomed.2006.08.016.

[22] A. R. Ferreira da Silva, Wavelet denoising with evolutionary algorithms, Digital Signal Processing 15 (4) (2005) 382 – 399. doi:10.1016/j.dsp.2004.11.003.

[23] E.-S. El-Dahshan, Genetic algorithm and wavelet hybrid scheme for ecg signal denoising, Telecommunication Systems 46 (2011) 209–215, 10.1007/s11235-010-9286-2.

[24] R. Salvador, F. Moreno, T. Riesgo, L. Sekanina, Evolutionary Approach to Improve Wavelet Transforms for Image Compression in Embedded Systems, EURASIP Journal on Advances in Signal Processing 2011 (2011). doi:10.1155/2011/973806.

[25] W. Zhao, C. E. Davis, Swarm intelligence based wavelet coefficient feature selection for mass spectral classification: An application to proteomics data, Analytica Chimica Acta 651 (1) (2009) 15 – 23. doi:10.1016/j.aca.2009.08.008.

[26] A. Daamouche, F. Melgani, Swarm intelligence approach to wavelet design for hyperspectral image classification, Geoscience and Remote Sensing Letters, IEEE 6 (4) (2009) 825 – 829. doi:10.1109/LGRS.2009.2026191.

[27] W. Pedrycz, S. S. Ahmad, Evolutionary feature selection via structure retention, Expert Systems with Applications 39 (15) (2012) 11801 – 11807. doi:10.1016/j.eswa.2011.09.154.

[28] S. Chatterjee, A. Bhattacherjee, Genetic algorithms for feature selection of image analysis-based quality monitoring model: An application to an iron mine, Engineering Applications of Artificial Intelligence 24 (5) (2011) 786 – 795. doi:10.1016/j.engappai.2010.11.009.

[29] Y.-X. Li, S. Kwong, Q.-H. He, J. He, J.-C. Yang, Genetic algorithm based simultaneous optimization of feature subsets and hidden markov model parameters for discrimination between speech and non-speech events, International Journal of Speech Technology 13 (2010) 61–73, 10.1007/s10772-010-9070-4.

[30] L. D. Vignolo, H. L. Rufiner, D. H. Milone, J. C. Goddard, Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classifica-

tion, EURASIP Journal on Advances in Signal Processing Volume 2011, doi:10.1155/2011/284791.

[31] L. D. Vignolo, H. L. Rufiner, D. H. Milone, J. C. Goddard, Evolutionary Cepstral Coefficients, Applied Soft Computing 11 (4) (2011) 3419 – 3428. doi:10.1016/j.asoc.2011.01.012.

[32] L. Vignolo, H. Rufiner, D. Milone, J. Goddard, Genetic optimization of cepstrum filterbank for phoneme classification, in: Proceedings of the Second International Conference on Bio-inspired Systems and Signal Processing (Biosignals 2009), INSTICC Press, Porto (Portugal), 2009, pp. 179–185.

[33] L. Vignolo, D. Milone, H. Rufiner, E. Albornoz, Parallel implementation for wavelet dictionary optimization applied to pattern recognition, in: Proceedings of the 7th Argentine Symposium on Computing Technology, Mendoza, Argentina, 2006.

[34] S. Durbha, R. King, N. Younan, Wrapper-based feature subset selection for rapid image information mining, Geoscience and Remote Sensing Letters, IEEE 7 (1) (2010) 43 –47. doi:10.1109/LGRS.2009.2028585.

[35] H.-H. Hsu, C.-W. Hsieh, M.-D. Lu, Hybrid feature selection by combining filters and wrappers, Expert Systems with Applications 38 (7) (2011) 8144 – 8150. doi:10.1016/j.eswa.2010.12.156.

[36] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1-2) (1997) 273 – 324. doi:10.1016/S000437029700043X.

[37] S. Mallat, A Wavelet Tour of signal Processing, 3rd Edition, Academic Press, 2008.

[38] S. Mallat, A theory of multiresolution of signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Machine Intell. 11 (7) (1989) 674–693.

[39] R. Coifman, M. V. Wickerhauser, Entropy-based algorithms for best basis selection, IEEE Transactions on Information Theory 38 (2) (1992) 713–718.

[40] N. Saito, Local feature extraction and its applications using a library of bases, Ph.D. thesis, Yale University, New Haven, USA, director-Ronald R. Coifman (1994).

[41] S. N. Sivanandam, S. N. Deepa, Introduction to Genetic Algorithms, Springer London, Limited, 2008.

[42] H. Youssef, S. M. Sait, H. Adiche, Evolutionary algorithms, simulated annealing and tabu search: a comparative study, Engineering Applications of Artificial Intelligence 14 (2) (2001) 167 – 181. doi:10.1016/S0952-1976(00)00065-8.

[43] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag, 1992.

[44] T. Kohonen, Improved versions of learning vector quantization, in: Proc. of the Int. Joint Conf. on Neural Networks, San Diego, 1990, pp. 545–550.

[45] A. Echeverría, J. Tejedor, D. Wang, An evolutionary confidence measure for spotting words in speech recognition, in: Y. Demazeau, F. Dignum, J. Corchado, J. Bajo, R. Corchuelo, E. Corchado, F. Fernández-Riverola, V. Julián, P. Pawlewski, A. Campbell (Eds.), Trends in Practical Applications of Agents and Multiagent Systems, Vol. 71 of Advances in Intelligent and Soft Computing, Springer Berlin / Heidelberg, 2010, pp. 419–427.

[46] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Marino, C. Nadeu, Albayzin speech database design of the phonetic corpus, Tech. rep., Universitat Politecnica de Catalunya (UPC), Dpto. DTSC (1993).

[47] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, P. Woodland, The HTK book, HTK version 3.1, Cambridge University (2001).
URL http://htk.eng.cam.ac.uk

[48] A. Quilis, Tratado de Fonología y Fonética Españolas, Biblioteca Románica Hispánica, Editorial Gredos, Madrid, 1993.

[49] H. Rufiner, Comparación entre análisis onditas y fourier aplicados al reconocimiento automático del habla, Master's thesis, Universidad Autónoma Metropolitana, Iztapalapa (1996).

[50] I. Daubechies, Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[51] M. Slaney, Auditory Toolbox, Version 2, Technical Report 1998-010, Interval Research Corporation, Apple Computer Inc. (1998).

[52] T. Ganchev, N. Fakotakis, G. Kokkinakis, Comparative evaluation of various MFCC implementations on the speaker verification task, in: Proceedings of the SPECOM-2005, 2005, pp. 191–194.

[53] X. D. Huang, Y. Ariki, M. A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, 1990.

[54] K. Demuynck, J. Duchateau, D. Van Compernolle, P. Wambacq, Improved Feature Decorrelation for HMM-based Speech Recognition, in: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98), Sydney, Australia, 1998.

[55] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, Cambrige, Masachussets, 1999.

[56] D. H. Milone, L. E. D. Persia, M. E. Torres, Denoising and recognition using hidden Markov models with observation distributions modeled by hidden markov trees, Pattern Recognition 43 (4) (2010) 1577 – 1589. doi:10.1016/j.patcog.2009.11.010.

[57] C. M. Bishop, Pattern Recognition and Machine Learning, 1st Edition, Springer, 2007.

[58] M. Lewicki, Efficient coding of natural sounds, Nature Neuroscience 5 (4) (2002) 356–363.