IET Signal Processing

Research Article

Multi-objective optimisation of wavelet features for phoneme recognition

ISSN 1751-9675
Received on 29th December 2015
Revised on 25th February 2016
Accepted on 13th March 2016
doi: 10.1049/iet-spr.2015.0568
www.ietdl.org

Leandro Daniel Vignolo[1] ✉, *Hugo Leonardo Rufiner*[1,2], *Diego Humberto Milone*[1]

[1]*Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET, Ciudad Universitaria UNL, (S3000) Santa Fe, Argentina*
[2]*Laboratorio de Cibernética FI-UNER, (E3100) Oro Verde, Entre Ríos, Argentina*
✉ *E-mail: ldvignolo@sinc.unl.edu.ar*

Abstract: State-of-the-art speech representations provide acceptable recognition results under optimal conditions, though their performance in adverse conditions still needs to be improved. In this direction, many advances involving wavelet processing have been reported, showing significant improvements in classification performance for different kinds of signals. However, for speech signals, the problem of finding a convenient wavelet-based representation is still an open challenge. This study proposes the use of a multi-objective genetic algorithm for the optimisation of a wavelet-based representation of speech. The most relevant features are selected from a complete wavelet packet decomposition in order to maximise phoneme classification performance. Classification results for English phonemes, in different noise conditions, show significant improvements compared with well-known speech representations.

1 Introduction

One of the most important issues in automatic speech recognition involves the pre-processing stage, which is meant to produce a manageable set of significant features. The pre-processing should be able to reveal the key-features of phonemes, in order to exploit the capabilities of the classification phase [1]. The most widely used features for speech recognition, and also applied for different tasks involving speech and music signals, are the mel-frequency cepstral coefficients (MFCCs) [2]. The MFCC are based on the linear model of voice production and a psycho-acoustic frequency mapping according to the mel scale [1].

Even though these features provide acceptable performance under laboratory conditions, recognition rates degrade significantly in presence of noise. This has motivated many advances in the development of alternative feature extraction approaches. Particularly, concepts from the psychophysics of hearing were exploited in the development of techniques such as perceptual linear prediction (PLP) [3] and relative spectra [4], which provide robust features based on an estimate of the auditory spectrum. More recently, speech processing techniques based on computational intelligence tools have been developed [5]. For example, a methodology for learning specialised filter banks using deep neural networks was proposed in [6]. Moreover, several approaches based on evolutionary computation have been proposed for the search of optimal speech representations [7–10].

Wavelet-based processing provides useful tools for the analysis of non-stationary signals [11], which have been found suitable for speech feature extraction [12–14]. The wavelet packet transform (WPT) offers a wide range of possibilities for the representation of a signal in the time-scale plane [11]. Hence, in order to build a representation based on the WPT, frequently a particular orthogonal basis is selected among all the available basis [12]. However, for speech recognition there is no evidence showing the convenience of the use of orthogonal basis. Furthermore, it is known that the analysis performed at the level of the auditory cortex is highly redundant [15]. Therefore, removing the orthogonality restriction the complete WPT decomposition offers a highly redundant set of coefficients, some of which can be selected to build an optimal representation.

The optimisation of wavelet decompositions for feature extraction has been studied in many different ways, though it is still an open challenge in speech processing. For example, an entropy-based method for best wavelet packet basis was proposed for electroencephalogram classification [16]. The use of wavelet-based decompositions has also been applied to the development of features for speech and emotion recognition [17, 18]. Other interesting proposals involve the use of evolutionary computing for the optimisation of over-complete decompositions for signal approximation [19], for the design of finite impulse response filters [20] and for the extraction frequency-domain features [21]. Also, in [22] a genetic algorithm (GA) was employed for the selection of an appropriate wavelet packet basis for image watermarking. Furthermore, the optimisation of wavelet decompositions by means of evolutionary algorithms was proposed for signal denoising [23].

It is important to note, however, that the WPT decomposition offers great flexibility, which has not been fully explored for feature extraction. Usually the search for an optimal decomposition is restricted to non-redundant representations, reducing drastically the number of possible solutions. Without this restriction, a hard combinatorial problem arises due to the availability of a large number of non-orthogonal dictionaries.

In the previous work, we presented a novel approach for the optimisation of over-complete decompositions from a WPT dictionary, using a genetic wrapper [7]. The classification performance was used to guide the optimisation, relaying on a classifier based on learning vector quantisation, and the task involved a set of Spanish phonemes. This wrapper was focused only on classification accuracy improvement, overlooking other important issues such as the dimensionality of the representation. To obtain a more proper representation for speech recognition, here we propose a multi-objective GA (MOGA) [24], which allows to maximise the classification accuracy while minimising the number of features. In this case, for the purpose of obtaining appropriate features for state-of-the-art speech recognisers, a classifier based on hidden Markov models (HMMs) [25] is used to estimate the capability of candidate solutions, using on a set of English phonemes. The proposed method, which we refer to as *evolutionary wavelet packets* (EWPs), exploits the benefits provided by multi-objective evolutionary optimisation in order to

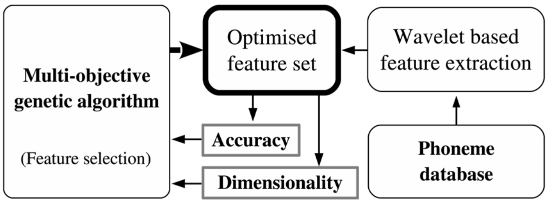

Fig. 1 *General scheme of the proposed multi-objective optimisation method*

find a better speech representation. Fig. 1 illustrates the general scheme of this approach.

2 Materials and methods

2.1 Wavelet packets decomposition

Wavelet bases are simultaneously localised in both time and frequency, and this property is essential for the analysis of signals which show transient and stationary behaviours. Wavelets are defined as centred functions with zero mean and unitary norm [11], which are translated and scaled in order to obtain the time–frequency atoms. The computation of the continuous wavelet transform involves the inner product of a signal with the family of time–frequency atoms. The discretisation of scaling and translation parameters, particularly with scaling factor 2^j, gives the *discrete dyadic wavelet transform* (DWT). In the fast implementation of the DWT, this is obtained by convolving the signal with a pair of quadrature mirror filters (low-pass and high-pass) to decompose the signal into detail and approximation coefficients [11]. The approximation is further decomposed within an iterative process, in which the frequency resolution is increased on each step. The WPT extends the DWT decomposition by applying low-pass and high-pass filters in each level to detail coefficients, as well as the approximation, offering more flexibility for frequency band selection. This results in the full WPT decomposition tree (Fig. 2), which provides an over-complete dictionary, and is obtained by

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

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

where $g[n]$ and $h[n]$ are the impulse responses of the high-pass and low-pass filters associated to the wavelet and scaling functions, respectively, j is the depth of the node and r is an index for the nodes which lay on the same depth. Then, c_j^{2r} is referred to as the approximation of c_{j-1}^r and c_j^{2r+1} is referred to as the detail.

The decomposition offered by the WPT allows to analyse a signal in a much more flexible time-scale plane, in which different sub-trees can be selected to extract the desired information from the full decomposition. Choosing one among all the possible combinations for a particular application is a challenging problem, which is usually solved by restricting the search to orthogonal basis using diverse criteria [16, 19]. The most common paradigm for signal compression using WPT is based on entropy measures and it is known as *best orthogonal basis* [26]. Another alternative is the *local discriminant basis* algorithm, which selects basis maximising a discriminant measure [27]. However, for the classification problem, the convenience of an orthogonal basis has not been proved. Moreover, previous studies conclude that the redundancy in a representation provides robustness for the classification of noisy signals [10], suggesting that a thorough search within the full decomposition provided by the WPT worth to be studied.

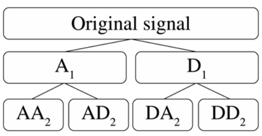

Fig. 2 *Wavelet packets tree with two decomposition levels ('A' stands for approximation and 'D' for detail coefficients)*

2.2 GAs with multiple objectives

Inspired by the natural process of evolution, the GA emerged as meta-heuristic optimisation methods, capable of finding global optima in complex search spaces [28]. To conduct the search these algorithms need to evaluate an objective function, according to the problem under study. It is important to note, however, that in real-world problems usually more than one objective need to be satisfied. In general, the solution of an optimisation problem with more than one objective consists not in a single point but a set of points known as the Pareto-optimal front [29].

The most common and basic approaches to tackle multi-objective problems using evolutionary computation consider all but one objective as constraints or the combination of the individual objective functions into a single aggregative function [24]. Other more powerful approaches attempt to determine a Pareto-optimal or non-dominated set of solutions [24]. This means a set of candidate solutions offering different objective trade-offs, and for which none of the objectives can be improved without detriment of other objective function.

Many alternatives and modifications to the classical GA have been proposed to find the Pareto front in multi-objective problems [29]. Particularly, in [30] a variation of the classical GA was proposed, the MOGA, capable of directing the search toward the true Pareto front while maintaining population diversity. The MOGA differs from the classical GA only in the way fitness is obtained for each individual in the population. A rank is first assigned to each solution, according to the number of chromosomes in the population by which it is dominated [24]. Then, a fitness is assigned to every solution based on its rank [30].

A common problem that usually prevents multi-objective evolutionary algorithms converging to the true Pareto-optimal is the fact that the population tends to scatter around the existing optima, in stable sub-populations or niches. To overcome this problem, fitness sharing techniques enforce the search in unexplored sections within the Pareto front, and contribute to maintain population diversity [30]. This is accomplished by the penalisation of solutions that are located close to each other.

2.3 Evolutionary wavelet features

In the feature extraction process, we used 256-sample windows, which are 32 ms at 8 kHz sampling frequency. The WPT process of filtering and decimation was performed to obtain a wavelet packet tree of six levels, consisting of 1536 coefficients. To reduce the search space, the coefficients corresponding to each frequency band were integrated by groups, meaning that the frequency bands were subdivided in order to obtain an energy coefficient for each group. The proposed integration scheme for a half of the WPT tree is depicted in Fig. 3, whereas the other half is integrated in the same manner. In this figure, dark grey rectangles represent the

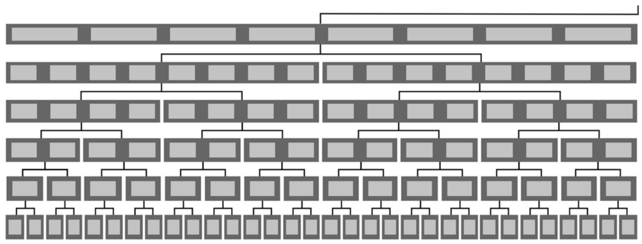

Fig. 3 *Illustration of the frequency band integration scheme (half tree)*

Table 1 Integration scheme applied to the WPT decomposition tree (256-sample signal)

Level	1	2	3	4	5	6
nodes	2^1	2^2	2^3	2^4	2^5	2^6
integration groups per node	2^3	2^3	2^2	2^1	2^0	2^0
wavelet coefficients per group	2^4	2^3	2^3	2^3	2^3	2^2
integration coefficients	2^4	2^5	2^5	2^5	2^5	2^6

nodes at the six levels of the decomposition tree. Light grey squares represent integration groups, which cover a variable number of wavelet coefficients. Also, Table 1 exhibits the number of integration groups in each node and the number of coefficients in each group. This integration scheme was designed according to the most relevant frequency bands in speech. In [7], the integration coefficient k in the feature vector corresponding to window p, $w_p[k]$, was normalised by $\hat{w}_p[k] = (w_p[k]/\arg\max_i w_i[k])$. Here, instead, for frame p the integration coefficients were normalised by its maximum coefficient value, $\hat{w}_p[k] = (w_p[k]/\arg\max_j w_p[j])$. In this way, the resulting normalised coefficients are independent of the signal energy. It should be noted that each training and testing pattern is composed of a variable number of $\hat{\boldsymbol{w}}_p$ vectors, each corresponding to a different temporal frame.

Wavelet families have been compared in order to determine which one is the most convenient for speech recognition [14]. On the basis of the literature, preliminary analysis included the wavelet families Meyer, Daubechies, Symmlets, Coiflets y Splines [11, 31]. As a result, the fourth-order Coiflet family was selected for the optimisation experiments.

Here, we propose the use of a MOGA for the selection of the optimal feature set, based on the WPT decomposition, for phoneme recognition. The objective functions should evaluate the representation suggested by a given chromosome, providing measures which are relevant for this particular problem. The candidate solutions represented by the individuals in the population of the MOGA are defined by binary chromosomes composed of 208 genes, each one corresponding to a specific integration coefficient.

In the proposed MOGA, the first target function evaluates the selected feature subset, providing a measure of classification performance. An HMM-based phoneme classifier is used as the first objective function, so that the classification accuracy is obtained for each evaluated individual. This classifier is trained on a corpus of isolated phonemes, and the accuracy obtained on a test set is the return value of the first objective function (F_a). It is also desired to obtain a speech representation containing the smallest number of coefficients, which is known to be beneficial for the recognition with HMM based in Gaussian mixtures. Therefore, the second target function takes into account the number of selected coefficients, favouring smaller subsets. This objective function was defined as $F_d = 1 - (n_s/l)$, where n_s is the number of selected coefficients and l is the chromosome length. Fig. 4 shows example Pareto fronts obtained using F_a and F_d as objective functions. To locate the ideal optimum at the origin as usual, because the objective functions are increasing, the axes of these plots are $1 - F_a$ (the classification error) and $1 - F_d$. The plot shows the dominant solutions from the first and last generation in an optimisation experiment. It can be seen how the best individuals in the population moved in direction to the ideal optimum, improving according to both functions.

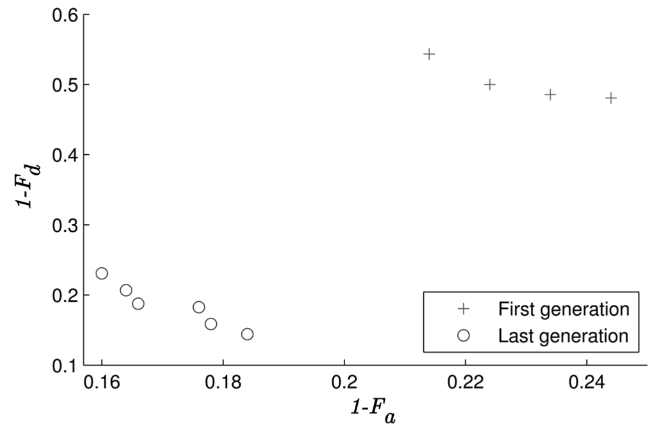

Fig. 4 *Example Pareto fronts obtained from a MOGA experiment, in the first and last generations*

3 Results and discussion

3.1 Speech data and experimental setup

Phonetic data was extracted from the TIMIT speech database [32] and selected randomly from all dialect regions including both male and female speakers. Utterances were phonetically segmented to obtain individual files with the temporal signal of every phoneme occurrence. To evaluate robustness, several types of noises were added to the signals considering signal-to-noise ratio (SNR) levels from −5 to 20 dB. The speech signals were downsampled to 8 kHz and frames were extracted using a Hamming window of 32 ms (256 samples) and a step size of 100 samples. All possible frames within a phoneme occurrence were extracted and padded with zeros where necessary. The set of English phonemes /b/, /d/, /eh/, /ih/ and /jh/ was considered. Occlusive consonants /b/ and /d/ were included because they are very difficult to distinguish in different contexts. Phoneme /jh/ presents special features of the fricative sounds. Vowels /eh/ and /ih/ are commonly chosen because they are close in the formants space. This phoneme set is a challenge for automatic recognition [33].

Our classifier is based on continuous HMM, using Gaussian mixtures with diagonal co-variance matrices for the observation densities. On the basis of state-of-the-art speech recognisers, we used a three-state HMM with mixtures of four Gaussian [9, 34]. To perform a fair comparison, the same classifier configuration was used for all the representations. Tools from the HMM Toolkit [35] were used for building and training the models. This toolkit implements the Baum–Welch algorithm [25] which is used to estimate the HMM parameters, and the Viterbi algorithm [25] to search for the most likely state sequence, given the observed events.

For the MOGA evolution an optimisation data set was used, while a separate evaluation set was left apart in order to estimate the generalisation performance. The optimisation data was split into training and validation sets, consisting of 2500 and 500 phonemes, respectively. We have set the size of these sets based on preliminary experiments, showing that fewer data caused overfitting of the optimisation process, while greater amounts of data caused the evolution to take impractical amount of time without improvements.

In the MOGA, the population size was set to 70 individuals, the cross-over rate was set to 0.8, the mutation rate was set to 0.2 and the niche size was set to 0.07. The termination criteria were to stop the optimisation after 700 generations. However, if no improvement was obtained during half of this number of generations, the optimisation was stopped earlier.

3.2 Phoneme classification results

At the end of every generation, the MOGA provides a set of individuals which dominate the actual population, in the sense that no other individual is closer to the Pareto front. Then, from the optimal set provided in the last generation, we have chosen the chromosome which achieved the best accuracy. For each of the optimisation experiments performed, the classification capabilities of the optimised feature set were evaluated. This evaluation was performed through cross-validation using the evaluation data set, composed of all the occurrences of the selected phonemes in all the TIMIT dialect regions (excluding the optimisation set). From

Table 2 Classification test results with white noise using static features (accuracy [%])

	Dim.	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
EWP.a	48	41.42	**59.76**	**67.78**	**72.16**	**74.58**	74.96
GWP	95	**42.94**	50.30	52.38	57.18	59.52	66.52
WPBI + TH	208	33.90	50.14	65.26	70.56	72.90	73.86
WP + TH + PCA	193	30.66	37.30	40.40	41.92	42.56	43.84
DWT + PCA	193	27.94	34.36	39.34	42.98	45.84	45.32
CEPLPC	12	24.80	35.60	41.24	44.52	49.24	53.92
LPC	14	22.46	24.62	36.12	41.76	45.50	46.02
MFCC	13	24.52	38.54	42.72	44.00	51.02	74.76
PLP	13	22.50	31.90	43.44	47.98	62.08	**77.42**
HFCC	16	20.24	25.98	47.26	62.78	67.68	70.54
EFB	16	20.56	36.88	60.30	68.32	68.70	69.82

Bold values are the highest accuracies for each column/SNR

this data, ten partitions where randomly sampled, each of which consisted of 2500 training signals and 500 test signals. To perform the validation tests close to real situations, we considered the mismatch training (MMT) condition. This means that the classifier was trained with clean signals only, while the tests were performed using noisy signals at different SNR levels. To compare the performance of the optimised feature set, the same HMM-based classifier was trained with different well-known speech features: MFCC [1], linear prediction coefficients (LPCs) [1], LPC cepstrum (CEPLPC) [3] and PLP [3]. The performances of the cepstral features obtained through evolutionary filter banks (EFBs) [10] and the human factor cepstral coefficients (HFCCs) [36] were also included in the comparison. For these representations typical parameters were used: orders 14 and 12 cepstral coefficients for CEPLPC, order 14 for LPC, 26 filters and 12 cepstral parameters for PLP and MFCC. For HFCC, 30 filters were considered and the bandwidth parameter E-factor was set to 5, based on the results shown in [10]. In the case of EFB, 18-filter configuration referred as C4 in [10] was used.

Furthermore, we compared the classification performance of genetic wavelet packets (GWPs) [7] and other wavelet-based representations. The same WPT decomposition with band integration employed for EWP but without feature selection and using soft thresholding for denoising [37], named WPBI + TH. It is important to remark that, when using the features obtained from WPT without performing the proposed band integration step, the training of the HMM classifier showed convergence problems. This is because the Gaussian mixtures are not able to model adequately the probability distributions of these coefficients [38]. Then, in order to obtain other wavelet-based features to compare their performance with the HMM classifier, a post-processing based on principal component analysis (PCA) [13] was applied. For the representation denoted as WP + TH + PCA, soft thresholding was applied to WPT coefficients and PCA was performed, preserving the 99% of the variance. The performance of the features based on the discrete DWT with PCA post-processing (DWT + PCA) was also compared.

Table 3 Classification test results with white noise using DA coefficients (accuracy [%])

	Dim.	−5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
EWP.a + DA	144	42.70	59.54	66.68	71.08	71.68	74.40
EWP.b + DA	108	43.14	**62.86**	**70.36**	**74.14**	**75.14**	76.84
EWP.c + DA	117	43.14	58.14	67.12	70.92	73.24	75.44
EWP.b + TH + DA	108	**43.58**	58.56	68.62	70.88	72.22	72.52
GWP + DA	285	41.68	53.58	49.66	48.78	50.10	59.46
WPBI + TH + DA	624	29.46	38.46	46.42	50.38	52.02	52.16
WP + TH + PCA + DA	579	33.44	37.34	38.46	40.90	41.84	43.24
DWT + PCA + DA	579	32.70	40.32	42.82	43.82	44.78	44.44
CEPLPC + DA	36	33.66	40.14	44.76	49.68	59.10	69.76
LPC + DA	42	20.72	23.20	35.80	41.98	45.10	46.00
MFCC + DA	39	38.42	41.00	23.40	41.62	50.00	78.14
PLP + DA	39	39.92	44.34	38.18	50.50	54.44	**78.68**

Bold values are the highest accuracies for each column/SNR

In the first optimisation experiment, we used clean signals in the train and test sets employed for the evaluation of candidate solutions. The MOGA converged to a subset of 48 coefficients, to which we will refer to as EWP.a. Table 2 shows the average classification results obtained through cross-validation, and considering different SNR levels in the test sets. It can be seen that the optimised representation EWP.a provides significant improvements in adverse noise conditions. From 0 to 15 dB SNR, the average accuracy of the optimised feature set outperforms all the other representations. Moreover, for 20 dB SNR the result obtained with the EWP.a is better than those of most of the other representations.

In the second experiment, we performed the optimisation including the delta and acceleration (DA) coefficients [1] in the representation. The result was a subset of 36 integration coefficients (a total of 108 features including DA), named EWP.b + DA. In a last experiment, also including DA coefficients, we used noisy signals at 5 dB SNR for the evaluation of the individuals during the optimisation. The MOGA converged to a subset of 39 coefficients (a total of 117 coefficients, EWP.c + DA). The cross-validation results are shown in Table 3, comparing the performances obtained with the reference representations including DA coefficients. In this comparison, we also included the performance of another representation consisting of the same set of coefficients selected for EWP.b + DA, in which soft thresholding was applied before the band integration, EWP.b + TH + DA. As in the previous table, all the optimised representations provided important improvements, specially at low SNR levels. Moreover, EWP.b + DA also outperforms all the classical representations on clean signals. It is interesting to note that, even though the feature set optimised using noisy signals (EWP.c + DA) provided improvements compared with state-of-the-art representations, the feature set optimised using only clean signals (EWP.b + DA) produced the best results for most of the noise levels. Note that EWP.b + TH + DA also performs better than the reference representations. However, without thresholding the optimised representation (EWP.b + DA) shows the best performance. This suggests that the evolutionary feature selection provides the more robust coefficients, without the need of an additional denoising step. The other wavelet representations show only minor improvements compared with state-of-the-art features. In these experiments, the average number of generations required to obtain the optimised representations was 687 while the average time for each generation was 495 s, using an Intel Core I7 processor with 8 GB random access memory. Note that every run of the search algorithm provides an acceptable solution.

Table 4 shows confusion matrices comparing the performance of PLP + DA and EWP.b + DA at low SNR levels. Rows correspond to the actual phoneme and columns to predictions, while the percentages of accuracy are shown on the diagonal. These matrices show coincidences between the phonemes which are most confused with PLP + DA and those confused with EWP.b + DA. For example, in both cases /eh/ was repeatedly confused with /ih/. Also, it can be noted that PLP + DA fails to discriminate phonemes /eh/ and /ih/ from /jh/ at lowest noise levels, and EWP.b + DA allows to improve their discriminability. Moreover, even if PLP + DA presents higher accuracy for some individual phonemes, EWP.b + DA achieves better balance providing important improvements in the total accuracy rate.

The classification performance of the optimised representations was also evaluated under several types of noises (Table 5), comparing the best evolutionary wavelet decomposition (EWP.b + DA) with the reference representations (MFCC + DA and PLP + DA). Even though EWP.b + DA was optimised using clean signals, it allowed to obtain important improvements at low SNR levels (from −5 to 5 dB) for four of the five noise types considered in these experiments.

We have also analysed the performance of these optimised representations in the classification of a wider set of phonemes (apart from those included in the optimisation). In this test, we considered the phonemes with the greater number of examples in the train and tests sets from the TIMIT corpus, discarding those

IET Signal Process., pp. 1–7

Table 4 Confusion matrices showing the percentages of average classification from ten data partitions in MMT conditions, with white noise at differences SNR levels. PLP + DA and optimised feature set EWP.b + DA

		PLP + DA					EWP.b + DA				
		/b/	/d/	/eh/	/ih/	/jh/	/b/	/d/	/eh/	/ih/	/jh/
0 dB	/b/	***61.9***	30.5	0.0	0.0	7.6	***34.6***	64.7	0.0	0.1	0.6
	/d/	24.7	***56.0***	0.0	0.0	19.3	10.5	***81.5***	0.2	0.4	7.4
	/eh/	0.1	10.8	***0.5***	6.2	82.4	0.3	15.3	***48.7***	31.6	4.1
	/ih/	0.6	3.2	0.0	***5.7***	90.5	0.2	6.4	21.9	***63.6***	7.9
	/jh/	0.1	2.3	0.0	0.0	***97.6***	0.2	13.6	0.0	0.3	***85.9***
						Diagonal average: 44.34					Diagonal average: **62.86**
5 dB	/b/	***13.5***	8.4	0.0	2.5	75.6	***59.1***	39.5	0.3	0.9	0.2
	/d/	1.3	***8.7***	0.0	1.0	89.0	16.4	***73.0***	0.9	1.1	8.6
	/eh/	0.5	2.8	***11.8***	60.3	24.6	0.2	2.3	***53.2***	42.9	1.4
	/ih/	0.2	1.7	2.2	***57.1***	38.8	0.3	2.6	19.8	***74.8***	2.5
	/jh/	0.0	0.2	0.0	0.0	***99.8***	0.2	6.7	0.0	1.4	***91.7***
						Diagonal average: 38.18					Diagonal average: **70.36**
10 dB	/b/	***20.8***	13.9	2.1	10.1	53.1	***71.4***	26.8	0.8	0.8	0.2
	/d/	1.9	***12.8***	0.1	3.5	81.7	22.7	***68.3***	1.6	0.8	6.6
	/eh/	0.1	0.2	***30.6***	67.9	1.2	0.4	0.7	***61.0***	37.6	0.3
	/ih/	0.1	0.0	6.0	***88.9***	5.0	0.6	1.1	19.5	***77.1***	1.7
	/jh/	0.0	0.4	0.0	0.2	***99.4***	0.1	5.5	0.2	1.3	***92.9***
						Diagonal average: 50.50					Diagonal average: **74.14**
15 dB	/b/	***20.9***	17.6	5.2	19.7	36.6	***73.6***	24.7	1.1	0.4	0.2
	/d/	2.1	***17.0***	0.6	8.2	72.1	23.7	***66.1***	1.7	1.1	7.4
	/eh/	0.0	0.2	***42.6***	57.0	0.2	0.3	0.7	***69.8***	28.8	0.4
	/ih/	0.0	0.0	5.8	***93.4***	0.8	0.3	0.7	25.1	***72.6***	1.3
	/jh/	0.0	0.7	0.0	1.0	***98.3***	0.1	4.5	0.0	1.8	***93.6***
						Diagonal average: 54.44					Diagonal average: **75.14**

Bold/italics values are the diagonal of each confusion matrix, meaning that each of these values is the accuracy rate for each class, and these are the values used to compute the average

Table 5 Classification test results considering other noise types (accuracy [%])

		–5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
PINK[a]	MFCC + DA	39.88	46.44	62.52	73.76	78.10	79.62
	PLP + DA	41.02	55.06	70.48	**77.16**	**80.02**	**81.30**
	EWP.b + DA	**56.40**	**64.92**	**70.62**	73.06	74.04	74.22
BUCCANEER[a]	MFCC + DA	40.44	48.10	65.30	76.22	78.96	80.14
	PLP + DA	40.86	55.80	70.24	**77.74**	**80.74**	**81.86**
	EWP.b + DA	**47.50**	**57.88**	**67.12**	71.42	73.50	74.62
VOLVO[a]	MFCC + DA	76.06	77.54	78.32	79.10	79.62	79.70
	PLP + DA	**78.02**	**79.80**	**80.64**	**80.92**	**81.52**	**81.78**
	EWP.b + DA	70.64	73.66	74.48	74.62	74.74	74.70
KEYBOARD[b]	MFCC + DA	39.06	49.60	60.40	68.70	**74.34**	**78.32**
	PLP + DA	40.66	49.28	59.26	67.00	72.98	76.76
	EWP.b + DA	**49.16**	**58.62**	**66.78**	**70.80**	72.86	74.04
VIOLET[c]	MFCC + DA	41.80	53.82	65.96	72.78	**77.06**	**79.30**
	PLP + DA	42.00	50.88	64.78	72.32	75.16	77.44
	EWP.b + DA	**51.08**	**64.14**	**71.18**	**72.98**	74.20	74.46

Bold values are the highest accuracies for each SNR and for each noise type
[a]http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html
[b]http://www.ece.rochester.edu/~zduan/data/noise
[c]http://www.audiocheck.net

with <1000 examples in the test set. The resulting set, consisting of 21 phonemes, together with the corresponding number of training and test examples are listed in Table 6. As the classes are not balanced, the classification performance is measured with the unweighed accuracy rate (UAR) [39]. As it can be seen in Table 7, EWP.a + DA and EWP.b + DA provide increased robustness in comparison with MFCC + DA and PLP + DA at low SNR levels. Owing to the number of classes this is a complex classification task; however, the performances obtained with the optimised representations are far from the rate given by chance classification, even at 0 dB SNR. Fig. 5 shows the confusion matrices obtained with PLP + DA and EWP.b + DA at 20 dB SNR, in which lighter squares indicate higher accuracy. It can be noted, by comparing the diagonals, that the optimised features provide improved accuracy for most classes. Also, the values outside the diagonal (confusions) are lower for EWP.b + DA. This experiment

Table 6 Phoneme set and respective number of training and test examples used in the experiments of Table 7

Phoneme	Train	Test	Phoneme	Train	Test	Phoneme	Train	Test
/aa/	3064	1133	/eh/	3853	1440	/n/	7068	2501
/ae/	3997	1407	/ih/	5051	1709	/q/	3590	1244
/ao/	2940	1156	/ix/	8642	2945	/r/	6539	2525
/ax/	3610	1346	/iy/	6953	2710	/s/	7475	2639
/ax-r/	3407	1383	/k/	4874	1614	/t/	4364	1535
/d/	3548	1245	/l/	5801	2356	/w/	3140	1239
/dh/	2826	1053	/m/	3903	1526	/z/	3773	1273

Table 7 Results obtained in the classification of the extended set of the 21 phonemes from Table 6 using white noise (UAR [%])

SNR	By chance	MFCC + DA	PLP + DA	EWP.a + DA	EWP.b + DA
0 dB	04.76	07.04	08.83	**12.98**	12.33
10 dB	04.76	16.30	20.49	29.73	**31.44**
20 dB	04.76	33.34	35.83	36.69	**40.08**

Bold values are the highest accuracies for each row/SNR

shows that EWP features are useful to discriminate other phonemes than those included in the optimisation. These results also suggest that the representations obtained with the proposed methodology could provide robustness to a continuous speech recognition system, even if only a reduced set of phonemes is considered in the optimisation. Even though it would be interesting to include more phonemes in the optimisation process, it should be taken into account that for several phonemes there is a reduced number of occurrences in the corpus, which could not allow to build proper train, test and validation sets.

To provide a qualitative analysis of the optimised decomposition, the tiling of the time–frequency plane was constructed using the criteria proposed in [40]. This is shown in Fig. 6, where each decomposition level is depicted separately for an easier interpretation. Each ellipse represents a group of coefficients from the integration scheme (Table 1); therefore, the widths and time localisations are determined by the corresponding time–frequency atoms. This means that each element in the tiling represents a time–frequency atom that was obtained by combining the original wavelet atoms, according to the integration scheme. Note that the number of coefficients in the groups of level 1 is twice the number of coefficients in the groups of level 2 (Fig. 3), which explains why the atoms for levels 1 and 2 are the same width in Fig. 6. This explanation also applies to the width of the atoms in levels 5 and 6. We remark that the optimisation of the decomposition based on the WPT has led to highly redundant representations, which are able to exploit redundancy in order to increase robustness. This characteristic is shared by all the EWP, showing redundancy at different regions of the time–frequency plane. For example, the optimised decompositions incorporate several time–frequency atoms below 1 kHz at every level. The results obtained suggest that the presence of redundant information in particular frequency bands allows to reduce the impact of noise. This could be thought as an enhancement technique, which reinforces the discriminative information. However, the optimised representations use <25% of the coefficients obtained from the WPT integration scheme. This means that the proposed MOGA achieved an important dimensionality reduction when compared with the decomposition optimised in [7], in which 50% of the available coefficients were selected.

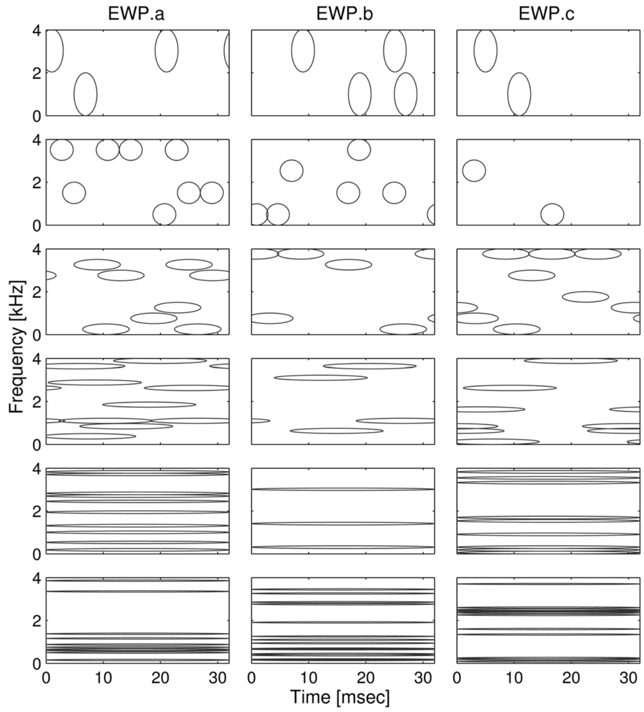

Fig. 6 *Tiling of the time–frequency plane obtained for the optimised decompositions. For a better visualisation, each level was schematised separately (from top: levels 1–6)*

Even though EWP.a, EWP.b and EWP.c provided similar results, they show differences in their time–frequency tilings. This could be due to the different conditions in which the decompositions were optimised, regarding the presence of noise and the use of DA coefficients, which might alter the search direction. For example, it is interesting to note that EWP.c shows less time–frequency atoms at the first and second decomposition levels, which may be due to the use of noisy signals. It is also interesting to note that the tilings presented in [7] show some atoms concentrated at the centre of the time axis, which could be related to the fact that only the frame extracted from the centre of each phone was considered in the optimisation. On the contrary, all the frames within a

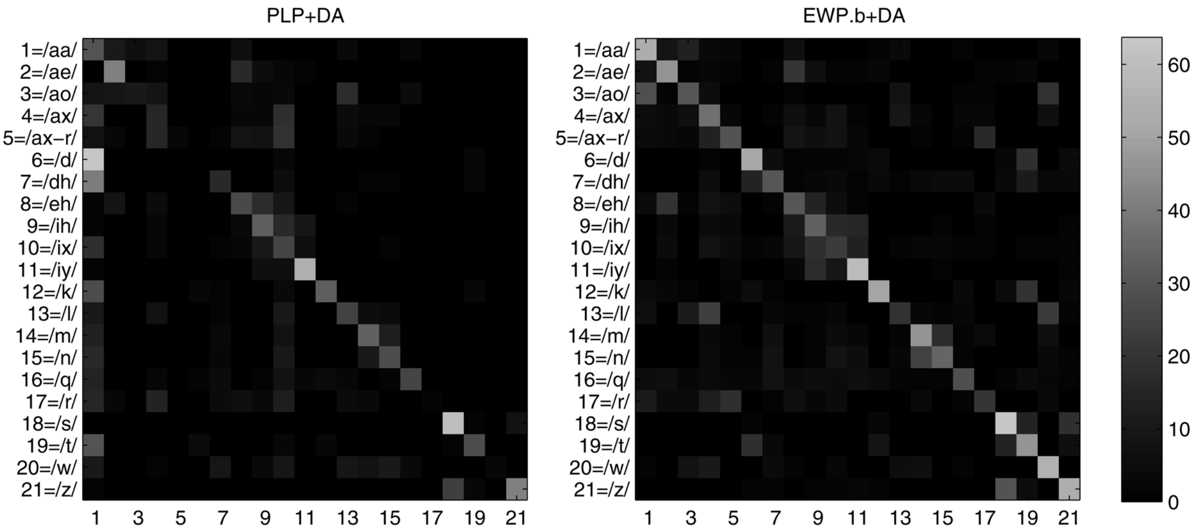

Fig. 5 *These confusion matrices show the classification rates for each of the 21 phonemes, obtained for PLP + DA and EWP.b + DA with white noise at 20 dB SNR*

phoneme were considered in this paper; thus, a different distribution of atoms could be expected.

4 Conclusion and future work

In this paper, a methodology for the optimisation of wavelet-based speech representations was proposed, taking advantage of the power of evolutionary computation techniques to explore large and complex search spaces. A multi-objective strategy was designed in order to maximise the discrimination capability of the representation while minimising the number of features. Following this methodology, relevant features have been selected from a wavelet packet decomposition, finding a good trade-off between redundancy and dimensionality to provide robustness in phoneme classification. The classification performance was evaluated using a set of phonemes taken from the TIMIT corpus, considering different noise conditions. The results show that the space of the optimised features increases class separation, providing important classification improvements in comparison with state-of-the-art robust features. Therefore, the proposed strategy stands as an alternative pre-processing methodology to obtain robust speech features, allowing to improve the classification performance in the presence of noise. Moreover, the results obtained in the classification with the extended set of phonemes suggest that the optimised representations could provide robustness to speech recognisers in tasks where the acoustic model has the primary role such as number or letter dictation.

In future work, it would be interesting to inquire into the design of new genetic operators, so that other specific constraints related to the problem could be taken into account. Also, in order to obtain a representation more suitable for HMM with Gaussian mixture modelling, one interesting idea is to include another objective function in the MOGA, in order to measure the Gaussianity of the EWP.

5 Acknowledgments

The authors acknowledge the support provided by the Agencia Nacional de Promoción Científica y Tecnológica (with projects PICT 2011-2440, PICT 2014-1442 and PICT 2014-2627), the Universidad Nacional de Litoral (with projects CAID 2011-519, -525 and PACT 2011-058) and the Consejo Nacional de Investigaciones Científicas y Técnicas from Argentina.

6 References

1 Huang, X., Acero, A., Hon, H.W.: 'Spoken language processing: a guide to theory, algorithm, and system development' (Prentice-Hall PTR, Upper Saddle River, NJ, USA, 2001, 1st edn.)
2 Ratanpara, T., Patel, N.: 'Singer identification using MFCC and LPC coefficients from Indian video songs'. Proc. 49th Annual Convention of the Computer Society of India (CSI), 2015, vol. 1, pp. 275–282, doi: 10.1007/978-3-319-13728-5_31
3 Cutajar, M., Gatt, E., Grech, I., *et al.*: 'Comparative study of automatic speech recognition techniques', *IET Signal Process.*, 2013, **7**, (1), pp. 25–46, doi: 10.1049/iet-spr.2012.0151
4 Hermansky, H., Morgan, N.: 'RASTA processing of speech', *IEEE Trans. Speech Audio Process.*, 1994, **2**, pp. 578–589, doi: 10.1109/89.326616
5 Hassanien, A., Schaefer, G., Darwish, A.: 'Computational intelligence in speech and audio processing: recent advances', in Xiao-Zhi Gao, António Gaspar-Cunha, Mario Köppen, Gerald Schaefer and Jun Wang (Eds.): 'Soft computing in industrial applications' (Springer Berlin/Heidelberg, Germany, 2010), vol. 75, pp. 303–311, doi: 10.1007/978-3-642-11282-9-32
6 Sainath, T.N., Kingsbury, B., Mohamed, A.R., *et al.*: 'Learning filter banks within a deep neural network framework'. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 2013, pp. 297–302, doi: 10.1109/ASRU.2013.6707746
7 Vignolo, L.D., Milone, D.H., Rufiner, H.L.: 'Genetic wavelet packets for speech recognition', *Expert Syst. Appl.*, 2013, **40**, (6), pp. 2350–2359, doi: 10.1016/j.eswa.2012.10.050
8 Li, Y.X., Kwong, S., He, Q.H., *et al.*: 'Genetic algorithm based simultaneous optimization of feature subsets and hidden Markov model parameters for discrimination between speech and non-speech events', *Int. J. Speech Technol.*, 2010, **13**, pp. 61–73, doi: 10.1007/s10772-010-9070-4
9 Vignolo, L.D., Rufiner, H.L., Milone, D.H., *et al.*: 'Evolutionary splines for cepstral filterbank optimization in phoneme classification', *EURASIP J. Adv. Signal Process.*, 2011, **2011**, pp. 8:1–8:14
10 Vignolo, L.D., Rufiner, H.L., Milone, D.H., *et al.*: 'Evolutionary cepstral coefficients', *Appl. Soft Comput.*, 2011, **11**, (4), pp. 3419–3428, doi: 10.1016/j.asoc.2011.01.012
11 Mallat, S.: 'A wavelet tour of signal processing' (Academic Press, London, 2008, 3rd edn.)
12 Montefusco, L., Puccio, L.: 'Wavelets: theory, algorithms, and applications', in Laura Montefusco and Luigia Puccio (Eds.): 'Wavelet analysis and its applications' (Academic Press London, UK, 2014)
13 Kotnik, B., Kačič, Z.: 'A noise robust feature extraction algorithm using joint wavelet packet subband decomposition and AR modeling of speech signals', *Signal Process.*, 2007, **87**, (6), pp. 1202–1223
14 Long, Y., Gang, L., Jun, G.: 'Selection of the best wavelet base for speech signal'. Proc. of 2004 Int. Symp. on Intelligent Multimedia, Video and Speech Processing, October 2004, pp. 218–221, doi: 10.1109/ISIMP.2004.1434039
15 Munkong, R., Juang, B.H.: 'Auditory perception and cognition', *IEEE Signal Process. Mag.*, 2008, **25**, (3), pp. 98–117, doi: 10.1109/MSP.2008.918418
16 Wang, D., Miao, D., Xie, C.: 'Best basis-based wavelet packet entropy feature extraction and hierarchical EEG classification for epileptic detection', *Expert Syst. Appl.*, 2011, **38**, (11), pp. 14314–14320, doi: 10.1016/j.eswa.2011.05.096
17 Biswas, A., Sahu, P., Bhowmick, A., *et al.*: 'Admissible wavelet packet sub-band-based harmonic energy features for Hindi phoneme recognition', *IET Signal Process.*, 2015, **9**, (6), pp. 511–519, doi: 10.1049/iet-spr.2014.0282
18 Huang, Y., Wu, A., Zhang, G., *et al.*: 'Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition', *IET Signal Process.*, 2015, **9**, (4), pp. 341–348, doi: 10.1049/iet-spr.2013.0446
19 Ferreira da Silva, A.R.: 'Approximations with evolutionary pursuit', *Signal Process.*, 2003, **83**, (3), pp. 465–481
20 Boudjelaba, K., Ros, F., Chikouche, D.: 'Adaptive genetic algorithm-based approach to improve the synthesis of two-dimensional finite impulse response filters', *IET Signal Process.*, 2014, **8**, (5), pp. 429–446, doi: 10.1049/iet-spr.2013.0005
21 Rivero, D., Guo, L., Seoane, J., *et al.*: 'Using genetic algorithms and k-nearest neighbour for automatic frequency band selection for signal classification', *IET Signal Process.*, 2012, **6**, (3), pp. 186–194, doi: 10.1049/iet-spr.2010.0215
22 Huang, H.C., Chen, Y.H.: 'Application of genetic-based wavelet packet watermarking for copyright protection', in Jeng-Shyang Pan, Hsiang-Cheh Huang, Lakhmi C. Jain, Yao Zhao (Eds.): 'Recent advances in information hiding and applications' (Springer Berlin Heidelberg, Berlin, Germany, 2013), vol. 40, pp. 139–153, doi: 10.1007/978-3-642-28580-6_7
23 El-Dahshan, E.S.: 'Genetic algorithm and wavelet hybrid scheme for ECG signal denoising', *Telecommun. Syst.*, 2011, **46**, pp. 209–215, doi: 10.1007/s11235-010-9286-2
24 Coello Coello, C.A.: 'Multi-objective evolutionary algorithms in real-world applications: some recent results and current challenges', in David Greiner, Blas Galván, Jacques Périaux, Nicolas Gauger, Kyriakos Giannakoglou, Gabriel Winter (Eds.): 'Advances in evolutionary and deterministic methods for design, optimization and control in engineering and sciences' (Springer International, Cham, Switzerland, 2015), vol. 36, pp. 3–18, doi: 10.1007/978-3-319-11541-2_1
25 Huang, X.D., Ariki, Y., Jack, M.A.: 'Hidden Markov models for speech recognition' (Edinburgh University Press, Edinburgh, 1990)
26 Coifman, R., Wickerhauser, M.V.: 'Entropy-based algorithms for best basis selection', *IEEE Trans. Inf. Theory*, 1992, **38**, (2), pp. 713–718
27 Saito, N., Coifman, R.: 'Local discriminant bases and their applications', *J. Math. Imaging Vis.*, 1995, **5**, (4), pp. 337–358, doi: 10.1007/BF01250288
28 Lin, C.D., Anderson-Cook, C.M., Hamada, M.S., *et al.*: 'Using genetic algorithms to design experiments: a review', *Qual. Reliab. Eng. Int.*, 2015, **31**, (2), pp. 155–167, doi: 10.1002/qre.1591
29 Deb, K.: 'Multi-objective optimization', in Edmund K. Burke, Graham Kendal (Eds.): 'Search methodologies' (Springer US, New York, US, 2014), pp. 403–449, doi: 10.1007/978-1-4614-6940-7_15
30 Fonseca, C.M., Fleming, P.J.: 'Genetic algorithms for multiobjective optimization: formulation discussion and generalization'. Proc. Fifth Int. Conf. on Genetic Algorithms, 1993, San Francisco, CA, USA, 1993, pp. 416–423
31 Rufiner, H., Goddard, J.: 'A method of wavelet selection in phoneme recognition'. Proc. 40th Midwest Symp. on Circuits and Systems, August 1997, vol. 2, pp. 889–891
32 Garofalo, J.S., Lamel, L.F., Fisher, W.M., *et al.*: 'DARPA TIMIT acoustic phonetic continuous speech corpus CD-ROM'. Technical report, U.S. Department of Commerce, NIST, Gaithersburg, MD, 1993
33 Stevens, K.N.: 'Acoustic phonetics' (MIT Press, London, 2000)
34 Demuynck, K., Duchateau, J., Van Compernolle, D., *et al.*: 'Improved feature decorrelation for HMM-based speech recognition'. Proc. Fifth Int. Conf. on Spoken Language Processing (ICSLP 98), November–December 1998
35 Young, S., Evermann, G., Gales, M., *et al.*: 'The HTK book (for HTK version 3.4)' (Cambridge University, England, 2006)
36 Skowronski, M., Harris, J.: 'Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition', *J. Acoust. Soc. Am.*, 2004, **116**, (3), pp. 1774–1780
37 Donoho, D.L., Johnstone, I.M.: 'Adapting to unknown smoothness via wavelet shrinkage', *J. Am. Stat. Assoc.*, 1995, **90**, (432), pp. 1200–1224
38 Milone, D.H., Di Persia, L.E., Torres, M.E.: 'Denoising and recognition using hidden Markov models with observation distributions modeled by hidden Markov trees', *Pattern Recognit.*, 2010, **43**, (4), pp. 1577–1589, doi: 10.1016/j.patcog.2009.11.010
39 Rosenberg, A.: 'Classifying skewed data: importance weighting to optimize average recall'. INTERSPEECH 2012, 2012
40 Lewicki, M.: 'Efficient coding of natural sounds', *Nat. Neurosci.*, 2002, **5**, (4), pp. 356–363