Multi-objective optimisation of wavelet features for phoneme recognition

Leandro D. Vignolo[1]*, Hugo L. Rufiner[1] and Diego H. Milone[1]

Research Institute for Signals, Systems and Computational Intelligence, **sınc**(*i*), FCICH-UNL/CONICET. Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina.

One of the most important issues in speech applications involves the preprocessing stage, which is meant to produce a manageable set of significant features, exploiting the capabilities of the classification phase [5]. 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) [1,5]. These are based on the linear model of voice production and a psychoacoustic scale [5]. Even though MFCCs provide acceptable performance under laboratory conditions, recognition rates degrade significantly in presence of noise. This has motivated many advances in the development of robust feature extraction approaches, like perceptual linear prediction (PLP) and relative spectra [1]. More recently, speech processing techniques based on computational intelligence tools have been developed. For example, several approaches based on evolutionary computation have been proposed for the search of optimal speech representations [8]. Wavelet based processing provides useful tools for the analysis of nonstationary signals, which have been found suitable for speech feature extraction [6]. In order to build a representation based on the wavelet packet transform (WPT), frequently a particular orthogonal basis is selected among all the available basis [6]. However, for speech recognition there is no evidence showing the convenience of the use of orthogonal basis. 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, the optimisation of wavelet decompositions by means of evolutionary algorithms was proposed for image watermarking [4] and for signal denoising [2]. In [9] we proposed a novel approach for the optimisation of over-complete decompositions from a WPT dictionary based on a multi-objective genetic algorithm (MOGA). The MOGA allows to maximise the classification accuracy while minimising the number of features. For the purpose of obtaining appropriate features for state of the art speech recognizers, a classifier based on hidden Markov models (HMM) 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* (EWP), exploits the benefits provided

[*] Autor corresponsal: ldvignolo@sinc.unl.edu.ar

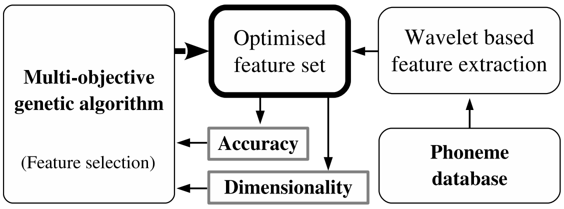

Fig. 1: General scheme of the proposed multi-objective optimisation method.

by multi-objective evolutionary optimisation in order to find a better speech representation. Fig. 1 illustrates the general scheme of this approach.

The MOGA was proposed 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 chromosomes composed of binary genes, each of which corresponds to a specific wavelet coefficient. The first target function evaluates the selected feature subset, providing a measure of classification performance. A classifier is used as the first objective function, so that the 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 $F_d = 1 - \frac{n_s}{l}$, where n_s is the number of selected coefficients and l is the chromosome length.

For the experiments phonetic data was extracted from the TIMIT English speech database [3], considering a set of isolated phonemes including /b/, /d/, /eh/, /ih/ and /jh/. The classification performance of the optimised representation was evaluated under several types of noise, comparing the evolutionary wavelet decomposition (EWP.b+DA[1]) 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.

The results (Table 1) show that the space of the optimised features increases class separation, providing important classification improvements in comparison to 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.

The optimisation led to a highly redundant representation, which is able to exploit redundancy in order to increase robustness. However, less than 25% of the

[1] DA means that the delta and acceleration coefficients are appended to the feature vector.

Table 1: Classification results in different noise conditions (Accuracy [%]).

		Dim.	-5 dB	0 dB	5 dB	10 dB	20 dB
WHITE	MFCC+DA	39	38.42	41.00	23.40	41.62	78.14
	PLP+DA	39	39.92	44.34	38.18	50.50	**78.68**
	EWP.b+DA	108	**43.14**	**62.86**	**70.36**	**74.14**	76.84
PINK	MFCC+DA	39	39.88	46.44	62.52	73.76	79.62
	PLP+DA	39	41.02	55.06	70.48	**77.16**	**81.30**
	EWP.b+DA	108	**56.40**	**64.92**	**70.62**	73.06	74.22
BUCCANEER	MFCC+DA	39	40.44	48.10	65.30	76.22	80.14
	PLP+DA	39	40.86	55.80	70.24	**77.74**	**81.86**
	EWP.b+DA	108	**47.50**	**57.88**	**67.12**	71.42	74.62
KEYBOARD	MFCC+DA	39	39.06	49.60	60.40	68.70	**78.32**
	PLP+DA	39	40.66	49.28	59.26	67.00	76.76
	EWP.b+DA	108	**49.16**	**58.62**	**66.78**	**70.80**	74.04
VIOLET	MFCC+DA	39	41.80	53.82	65.96	72.78	**79.30**
	PLP+DA	39	42.00	50.88	64.78	72.32	77.44
	EWP.b+DA	108	**51.08**	**64.14**	**71.18**	**72.98**	74.46

coefficients obtained from the WPT integration scheme. This means that the proposed MOGA achieved an important dimensionality reduction when compared to the decomposition optimised in [7].

The average number of generations required to obtain the optimised representations was 687 while the average time for each generation was 495 seconds, using an Intel Core I7 processor with 8GB RAM. Note that every run of the search algorithm provides an acceptable solution.

References

1. Cutajar, M., Gatt, E., Grech, I., Casha, O., Micallef, J.: Comparative study of automatic speech recognition techniques. IET Signal Proc. 7(1), 25–46 (Feb 2013)
2. El-Dahshan, E.S.: Genetic algorithm and wavelet hybrid scheme for ECG signal denoising. Telecommunication Systems 46, 209–215 (2011)
3. Garofalo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: DARPA TIMIT acoustic phonetic continuous speech corpus CD-ROM. Tech. rep., U.S. Dept. of Commerce, NIST, Gaithersburg, MD (1993)
4. Huang, C.L., Matsuda, S., Hori, C.: Feature normalization using MVAW processing for spoken language recognition. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific. pp. 1–4 (Oct 2013)
5. 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, 1st edn. (2001)
6. Montefusco, L., Puccio, L.: Wavelets: Theory, Algorithms, and Applications. Wavelet Analysis and Its Applications, Elsevier Science (2014)
7. Vignolo, L.D., Milone, D.H., Rufiner, H.L.: Genetic wavelet packets for speech recognition. Expert Systems with Applications 40(6), 2350–2359 (2013)
8. Vignolo, L.D., Rufiner, H.L., Milone, D.H., Goddard, J.C.: Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classification. EURASIP Journal on Advances in Signal Proc. 2011, 8:1–8:14 (2011)
9. Vignolo, L.D., Rufiner, H.L., Milone, D.H.: Multi-objective optimisation of wavelet features for phoneme recognition. IET Signal Processing (March 2016), `http://dx.doi.org/10.1049/iet-spr.2015.0568`