14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPEECH/MUSIC DISCRIMINATION USING A WARPED LPC-BASED FEATURE AND A FUZZY EXPERT SYSTEM FOR INTELLIGENT AUDIO CODING

J.E. Muñoz-Expósito, S. Garcia-Galán, N. Ruiz-Reyes, P. Vera-Candeas, F. Rivas-Peña

Electronic and Telecommunication Engineering Department, University of Jaén
Polytechnic School, 23700 Linares, Jaén, SPAIN
phone: + (34) 953648554, fax: + (34) 953648508, email: {jemunoz, sgalan,nicolas,pvera,rivas}@ujaen.es

ABSTRACT

Automatic discrimination of speech and music is an important tool in many multimedia applications. This paper presents an evolutionary fuzzy rules-based speech/music discrimination approach for intelligent audio coding. A low complexity but effective feature, called Warped LPC-based Spectral Centroid (WLPC-SC), is defined for the analysis stage of the discrimination system. The final decision is made by a fuzzy expert system, which improves the accuracy rate provided by a Gaussian Mixture Model (GMM) classifier taking into account the audio labels assigned by the GMM classifier to past audio frames. Comparison between WLPC-SC and most timbral features proposed in [8] is performed, aiming to assess the good discriminatory power of the proposed feature. The accuracy rate improvement due to the fuzzy expert system is also reported. Experimental results reveal that our speech/music discriminator is robust and fast, making it suitable for intelligent audio coding.

1. INTRODUCTION

Automatic discrimination between speech and music has become a research topic of interest in the last few years. Several approaches have been described in the recent literature for different applications [1][2] [3][4][5]. Each of these uses different features and pattern classification techniques and describes results on different material.

Saunders [1] proposed a real-time speech/music discriminator, which was used to automatically monitor the audio content of FM audio channels. Four statistical features on the zero-crossing rate and one energy-related feature were extracted, and a multivariate-Gaussian classifier was applied, which resulted in an accuracy of 98%.

In Automatic Speech Recognition (ASR) of broadcast news, it's desirable to disable the input to the speech recognizer during the non-speech portion of the audio stream. Scheirer and Slaney [2] developed a speech/music discrimination system for ASR of audio sound tracks. Thirteen features to characterize distinct properties of speech and music, and three classification schemes (MAP Gaussian, GMM and k-NN classifiers) were exploited, resulting in an accuracy of over 90%.

Automatic discrimination of speech and music is an important tool in many multimedia applications. Khaled El-Maleh et al. [3] combined line spectral frequencies and zero-crossings for frame-level narrowband speech/music discrimination. The classification system operates using only a frame delay of 20 ms, making it suitable for real-time multimedia applications. An emerging multimedia application is content-based indexing and retrieval of audiovisual data. Audio content analysis is an important task for such application [6].

Comparative view of the value of different types of features in speech/music discrimination is provided in [7], where four types of features (amplitudes, cepstra, pitch and zero-crossings) are compared. Experimental results showed cepstra and delta cepstra bring the best performance. Mel Frequencies Spectral or Cepstral Coefficients (MFSC or MFCC) are very often used features for audio classification tasks, providing quite good results. In [4], MFSC's first order statistics are combined with neural networks to form a speech/music classifier that is able to generalize from a little amount of learning data. MFCC are a compact representation of the spectrum of an audio signal taking into account the nonlinear human perception of pitch, as described by the mel scale. They are one of the most used features in speech recognition and have recently proposed in musical genre classification of audio signals [8][9].

Another application that can benefit from distinguishing speech from music is low bit-rate audio coding. Designing an universal coder to reproduce well both speech and music is the best approach. However, it is not a trivial problem. An alternative approach is to design an intelligent audio coding scheme composed of a speech/music discriminator and a multi-mode coder that can accommodate different signals. The appropriate module is selected using the output of the speech/music classifier [10] [11].

In this paper, we present our contribution to the design of a robust and real-time implemented speech/music discriminator, which can be integrated into an intelligent audio coder with application to internet audio streaming. For such goal, we define a simple but effective feature, called Warped LPC-based Spectral Centroid (WLPC-SC), which will be used as the only one feature in the analysis stage. Other speech/music discrimination approaches based on only one type of feature are presented in [12] and [5], which result in fast and robust classification systems. The final decision will be made by a quite simple fuzzy expert system, which is designed to improve the accuracy rate provided by a GMM classifier

2. SPEECH/MUSIC DISCRIMINATION FOR INTELLIGENT AUDIO CODING

Speech/music discrimination involves a suitable processing for two main tasks: audio feature extraction and classification of the extracted parameters. In this work, contributions in both directions are done. The resulting speech/music discriminator can be integrated into an intelligent multi-mode audio coder. This system must first perform an intelligent segmentation of the audio signals, so that they become se-

quences of audio frames labelled as speech or music, according to the decisions of the speech/music discriminator. Once the audio frames have been labelled, they are applied to coders adapted to the characteristics of each frame (i.e. a HVXC coder could be used for speech frames and an AAC coder for music frames). The underlying speech/music discriminator will assign a different cost to the two error possibilities, being much higher the cost of classifying a music frame as speech than the opposite. It aim to achieve high quality low bit rate audio coding for all types of audio signals, and must be applicable to last generation mobile phone systems and internet audio streaming.

2.1 Analysis stage: New Warped LPC-based feature

In our system, an *analysis frame* of 23 ms (1024 samples at 44100 Hz sampling rate), a *long texture frame* of 1 s (43 analysis windows) and a *short texture frame* of 250 ms are defined. Overlapping with a hop size of 512 samples is performed. Hence, the vector for describing the proposed feature, when using long texture frames, consists of 85 values, which are updated each 23 ms-length analysis frame. This large dimensional feature vector is difficult to be handled for classification tasks, giving rise to two main drawbacks: 1) too much computational cost, 2) possible too high misclassification rate. Therefore, it is required reducing the feature space to a few statistical values each 1 s-length long texture frame. In this work, the mean and variance of each feature vector are only computed.

Usually, speech signals has a low centroid frequency, which varies sharply at a voiced-unvoiced boundary. Instead, music signals show a quite changing behavior. There is no a specific pattern for such signals. We compute the centroid frequency by a one-pole lpc-filter. Geometrically, the lpc-filter minimizes the area between the frequency response of the filter and the energy espectrum of the signal. The one-pole frequency tells us where the lpc-filter is frequency-centered. Therefore, someway, the one-pole frequency informs us where most of the signal energy is frequency-localized.

However, the human auditory system is nonuniform in relation to the frequency. According to this statement, the Mel, the Bark and the ERB (Equivalent Rectangular-Bandwidth) scales [13] are defined for audio processing. For speech/music discrimination, it would be desirable to use a feature that works directly on some of these auditory scales, resulting in frequency-warped audio processing.

The transformation from frequency to Bark scale is a well studied problem [13] [14]. Generally, the Bark scale is performed via the all-pass transformation defined by the substitution in the z domain:

$$z = A_\rho(\zeta) \equiv \frac{\zeta + \rho}{1 + \zeta\rho} \tag{1}$$

which takes the unit circle in the z plane to the unit circle in the ζ plane, in such a way that, for $0 < \rho < 1$, low frequencies are stretched and high frequencies are compressed. Parameter ρ depends on the sampling frequency of the original signal [14]. Applying (1), the Bark scale values can be approximated from frequency positions as follows [13]:

$$b = 13arctan(0.76f(kHz)) + 3.5arctan(\frac{f(kHz)}{7.5})^2 \tag{2}$$

We propose the use of a one-pole warped-lpc filter based on this bilinear transformation to compute the WLPC-SC feature each 23 ms-length analysis frame.

The implementation of this filter can be downloaded from: *http://www.acoustics.hut.fi/software/warp* [13].

As can be seen in Fig. 1, the WLPC-SC feature shows clear differences between voiced and unvoiced phonemes due to the frequency-warped processing. Besides, these differences are bigger than in a drum-based music signal. The results in Fig. 1 suggest us that WLPC-SC could be a profitable low complexity feature to design a robust music/speech discriminator. It will be assessed in section 3.

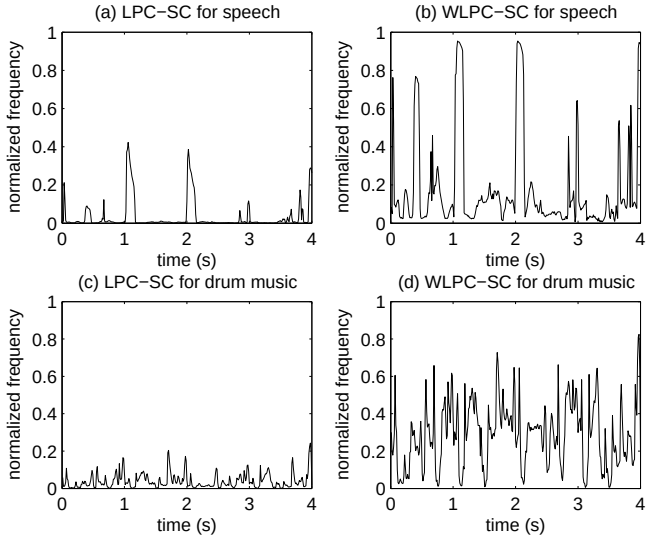

Figure 1: Example illustrating the values that LPC-SC and WLPC-SC takes for both speech and music signals.

2.2 Classification stage: Fuzzy rules-based expert system

For classification purposes, a number of standard Statistical Pattern Recognition (SPR) classifiers [15] were evaluated. Here, a three-component GMM classifier with diagonal covariance matrices is used because it showed a slightly better performance than other SPR classifiers. The performance of the system does not improve when using a higher number of components in the GMM classifier. The GMM classifier is initialized using the K-means algorithm with multiple random starting points. Modern classification techniques, such as Neural Networks (NN), Support Vector Machines (SVM) and dynamic programming, could also be used, but the complexity would increase to a great extent.

Nevertheless, we are interested in discriminating between speech and music for intelligent audio coding. A suitable coder must be selected each 23 ms-length analysis frame according to the decision of the speech/music discriminator (i.e. a HVXC coder can be applied to speech frames and an ACC coder to music frames). If coder selection is only based on current frame data, the GMM classifier we will obtain low success rate. It is very important to assure a robust performance of the speech/music discriminator for intelligent audio coding. Hence, we propose the use of a Fuzzy rules-based expert system for selecting the suitable coder each 23

ms-length analysis frame, which takes into account information not only of the current frame but also of past frames. The classification stage will consist of two components: the GMM classifier and the fuzzy expert system. It seems likely that the inclusion of the expert system within the classification stage implies an improvement of the classification accuracy rate obtained by the GMM classifier.

The fuzzy expert system takes the final decision from four input parameters. The input parameters (P_0, P_1, P_2 and P_3) represent the probabilities obtained by the 3-GMM classifier for four consecutive 250 ms-length short texture frames. The last of them includes the current 23 ms-length analysis frame just at the end.

Using these probabilities and a knowledge base, the Fuzzy rules-based expert system selects a suitable coder (a coder adapted to music or a coder adapted to speech) for intelligent audio coding. The fuzzy rules-based expert system structure appears in figure 2.

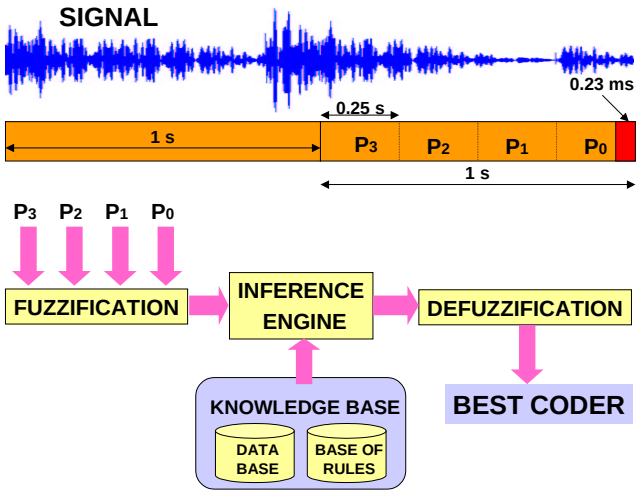

Figure 2: Expert System General Structure.

All inputs has been calculated using the GMM classifier from the mean and variance of the vector associated to each 250 ms-length short texture frame. These vectors consist of the WLPC-SC feature values corresponding to the 23 ms-length analysis frames that constitute each 250 ms-length short texture frame. All probabilities are [0,1] normalized. Input membership functions are represented in Figure 3.

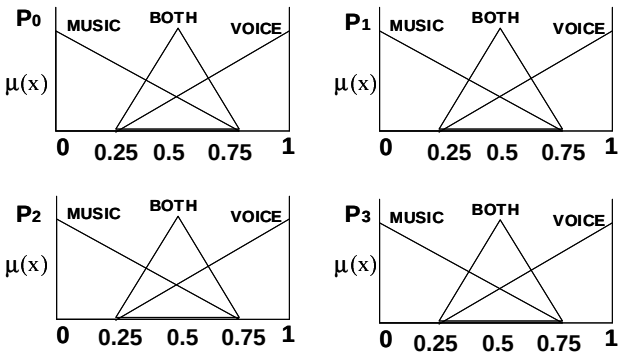

Figure 3: Membership functions for input variables.

There is only one output variable, called 'Coder', which is [0,1] normalized. If the output value is higher than 0.5, a speech coder is selected. Otherwise, a music coder is selected. Membership functions for the output variable is shown in Figure 4.

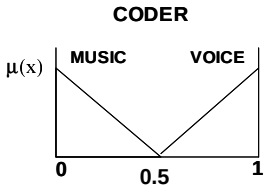

Figure 4: Membership functions for the output variable.

Next, we briefly outline the performance of the expert system and the methodology for building new knowledge base.

2.2.1 *Fuzzy rules-based expert system*

Fuzzificator transforms the input values for the inference engine process. Inference engine obtains an output fuzzy set using inputs and relationships defined in a fuzzy rules base.

The expert system takes the decision about the suitable coder for processing the audio signal from the input probabilities, a data base (input and output membership functions information) and a base of rules.

Finally, defuzzificator transforms the output fuzzy set in a value that allows select the suitable audio coder each 23 ms-length analysis frame.

2.2.2 *Building knowledge base*

The new rules added to the expert system knowledge base have been calculated using evolutionary computation. The methodology has been used with success in other research works [16][17]. The algorithm for knowledge acquisition is based on random rules generation and later insertion of the new rules into the knowledge base whether an improvement in the classification accuracy rate is achieved. In order to assess this improvement, it is required to compare the performance of the evaluated system (in our case, the intelligent audio coder) with and without each new rule.

The fuzzy expert system takes a decision every 23 ms. Two types of error can happen: an audio frame is labelled as speech when it is a music frame and the opposite. The first one (Music as Speech Error, MSE) is considered more serious than the second one (Speech as Music Error, SME), since it gives rise to a great loss of audio quality. The Speech as Music Error is less critical, because it implies an increase in the bandwidth necessary to transmit the signal, but no loss of audio quality is produced.

In order to design and evaluate the fuzzy expert system, a fitness function which considers both types of error is used:

$$Ev = 0.8 \cdot MSE + 0.2 \cdot SME \qquad (3)$$

For designing and testing the system, different audio files with 23 ms-length analysis frames labelled as speech or music are available. The system uses these files to train and build the knowledge base, which includes all learn rules that allow to improve the intelligent audio coder performance.

3. EXPERIMENTAL EVALUATION

First of all, the audio test database is carefully prepared. The speech data come from news programs of radio and TV stations, as well as dialogs in movies, with different levels of noise and music background, especially in news programs, and the languages involve English, Spanish, French and German. The speakers involve male and female with different ages. The length of the whole speech data is about an hour. The music consists of songs and instrumental music. The songs cover as more styles as posible, such as rock, pop, folk and funky, and they are sung by male and female in English and Spanish. The instrumental music we have chosen covers different instruments (piano, violin, cello, pipe, clarinet) and styles (symphonic music, chamber music, jazz, electronic music). Some music pieces in movies are also included, which are played by multiple different instruments. The length of the whole music data is also about an hour.

Next, we intend to assess the speech/music discrimination capability of the proposed feature. To achieve such goal, the WLPC-SC feature is compared to most timbral texture features proposed in [8]. The following timbral features are considered for comparison purposes: Spectral Centroid (SC), Spectral Rolloff (SR), Spectral Flux (SF), Time Domain Zero Crossings (ZC) and Mel-Frequency Cepstral Coefficients (MFCC)[1]. Comparison results are obtained evaluating the different features each 23 ms-length analysis frame with a 3-GMM classifier. The classifier takes a decision each 23 ms according to the mean and variance of the 85 samples-length vector corresponding to the last 1 s-length long texture frame. Table 1 shows the classification accuracy percentages when WLPC-SC is compared to the timbral features in [8].

FEATURE	SPEECH (%)	MUSIC (%)	GLOBAL (%)
SC	93.90	86.55	90.20
SR	95.93	65.60	80.67
SF	68.24	63.96	66.10
ZC	31.52	68.20	49.98
MFCC	98.11	84.54	91.28
WLPC-SC	97.33	82.31	89.77

Table 1: Classification accuracy percentage. WLPC-SC vs. timbral features

At the sight of the results in table 1, we can say that the proposed feature performs better than most of the timbral features in [8] for speech/music discrimination. The Spectral Centroid (SC) performs as well as the Warped LPC-based Spectral Centroid (WLPC-SC), while the Mel-Frequency Cepstral Coefficients (MFCC) give slightly better classification accuracy percentages. The good discrimination capability provided by the SC and MFCC features is achieved at the cost of a complexity increase regarding the WLPC-SC feature, which is much higher in the case of the MFCC feature. Note that the WLPC-SC feature does not require a DFT computation, while both SC and MFCC features require this computation. As shown in table 1, the proposed feature achieves high accuracy percentages while maintaining the complexity at a reduced degree.

We are also interested in comparing MFCC with all the rest timbral features in table 1 and all the rest timbral features in table 1 plus WLPC-SC. We intend to know if the proposed feature improves the classification accuracy percentage when it is added to all timbral features (except MFCC) for speech/music discrimination. Table 2 shows how the inclusion of the WLPC-SC feature within the feature set entails a discrimination capability improvement. The classification accuracy percentage grows about a 2%. However, it must be noted that no improvement is accomplished when all the rest timbral features in table 1 are used for speech/music discrimination regarding the case of using only the MFCC feature.

FEATURE	SPEECH (%)	MUSIC (%)	GLOBAL (%)
MFCC	98.11	84.54	91.28
All the rest timbral features	96.05	86.72	91.33
All the rest timbral features + WLPC-SC	98.86	87.95	93.15

Table 2: Discrimination capability improvement when the WLPC-SC feature is included within the feature set.

Now, we are also interested in knowing how much warping transformation influences in speech/music discrimination. Table 3 compares the classification accuracy results for both the proposed feature (WLPC-SC) and the same feature without warping transformation (LPC-SC).

FEATURE	SPEECH (%)	MUSIC (%)	GLOBAL (%)
WLPC-SC	97.33	82.31	89.77
LPC-SC	90.72	68.36	79.54

Table 3: Classification accuracy percentage. WLPC-SC vs. LPC-SC

From the results in table 3, it can be said that warping transformation is a very important operation for the good performance of the feature proposed in this paper, because it entails psychoacoustic information is taken into account. Table 3 shows an improvement in the speech/music discrimination capability higher than 10% regarding the case of not using the warping transformation.

Finally, table 4 shows the improvement in the classification accuracy rate due to the inclusion within the classification stage of the fuzzy expert system with regard to the case of using only the GMM classifier. Results in table 4 are accomplished by evaluating only one feature (the proposed WLPC-SC feature). Remember that the fuzzy system receives every 23 ms the decisions of the GMM classifier corresponding to the last four 250 ms-length short texture frames. The fuzzy system takes the final decision according to these decisions provided by the GMM classifier and a set a properly defined fuzzy rules.

We can see from table 4 that the fuzzy rules-based expert system implies a better performance of the speech/music discriminator. The global accuracy percentage grows about a 6%. Besides, the fuzzy expert system gives rise to an important decrease of the MSE errors (about a 13%), which improves the audio coder performance, since this type of error is considered critical for audio coding purposes.

[1] Only the first five cepstral coefficients are taken for classification

CLASSIFIER	SPEECH (%)	MUSIC (%)	GLOBAL (%)
3-GMM	97.33	82.31	89.77
3-GMM + Fuzzy system	96.17	95.08	95.62

Table 4: Classification accuracy percentage. 3-GMM vs. 3-GMM + Fuzzy system

Because of its simplicity and robustness, the resulting speech/music discriminator can be integrated into a real-time intelligent audio coder with different applications (i.e. internet audio streaming).

4. CONCLUSIONS

This paper presents a simple but robust approach to discriminate speech and music. The method exploits only one feature in the analysis stage, called Warped LPC-based Spectral Centroid (WLPC-SC), and a Gaussian Mixture Model improved by a fuzzy rules-based expert system in the classification stage. The new feature is an important contribution of the paper. Its simplicity and robustness make its application scope very wide, especially for applications where low computational cost is strongly demanded. Its performance is assessed by different experimental tests, which compare the proposed feature to other features commonly used in audio classification tasks, and also try to assess the improvement due to the warping transformation. We achieve an improvement in the discrimination capability higher than 10% with regard to the case of not using the warping transformation. The classification stage consists of a three-component GMM classifier and a fuzzy rules-based expert system, which takes the final decision from the probabilities of the GMM classifier and the fuzzy system knowledge base. The fuzzy system achieves an improvement about 6% regarding the case of using only the GMM classifier. Experiment results also demonstrate the robustness of the system. The classification accuracy percentage is higher than 95% for a wide range of audio samples. At the same time, its simplicity brings obvious advantages in constructing low cost systems.

REFERENCES

[1] Saunders, J. "Real-time discrimination of broacast speech/music", *Proc. IEEE ICASSP'96*, Atlanta, USA, pp. 993-996, 1996.

[2] Scheirer, E. and Slaney, M. "Construction and evaluation of a robust multifeature speech/music discriminator", *Proc. IEEE ICASSP'97*, Munich, Germany, pp. 1331-1334, 1997.

[3] El-Maleh, K., Klein, M., Petrucci, G. and Kabal, P. "Speech/music discrimination for multimedia applications", *Proc. IEEE ICASSP'2000*, vol. 6, pp. 2445-2448, 2000.

[4] Harb, H. and Chen, L. "Robust speech music discrimination using spectrum's first order statistics and neural networks ", *Proc. IEEE Int. Symp. on Signal Processing and Its Applications*, vol. 2, pp. 125-128, 2003.

[5] Wang, W.Q., Gao, W., Ying, D.W. "A fast and robust speech/music discrimination approach", *Proc. 4th Pacific Rim Conference on Multimedia.*, vol. 3, pp. 1325-1329, 2003.

[6] Zhang, T. and Kuo, J. "Audio content analysis for online audiovisual data segmentation and classification", *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, 2001.

[7] Carey, M.J., Parris, E.S. and Lloyd-Thomas, H. "A comparison of features for speech, music discrimination", *Proc. IEEE ICASSP'99*, Phoenix, USA, pp. 1432-1435, 1999.

[8] Tzanetakis, G. and Cook, P. "Musical genre classification of audio signals", *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, 2002.

[9] Burred, J.J. and Lerch, A. "Hierarchical automatic audio signal classification", *Journal of the Audio Eng. Soc.*, vol. 52, pp. 724-739, 2004.

[10] ISO-IEC. "MPEG-4 Overview (ISO/IEC JTC1/SC29/WG11 N2995 Document", 1999.

[11] Tancerel, L., Ragot, S., Ruoppila, V.T. and Lefebvre, R. "Combined speech and audio coding by discrimination", *Proc. IEEE Workshop on Speech Coding*, pp. 17-20, 2000.

[12] Karneback, S. "Discrimination between speech and music based on a low frequency modulation feature", *European Conf. on Speech Comm. and Technology*, Alborg, Denmark, pp. 1891-1894, 2001.

[13] Härmä, A., Karjalainen, M., Savioja, L., Välimäki, V., Laine, U.K. and Huopaniemi, J. "Frequency-warped signal Processing for audio applications", *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011-1031, November 2000.

[14] Smith III, J.O. and Abel, J.S. "Bark and ERB bilinear transforms", *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 697-708, November 1999.

[15] Duda, R., Hart, P. and Stork, D. "Pattern classification", Wiley, New York, 2000.

[16] Cordon, O., Herrera, F., Hoffmann, F. and Magdalena, L. "Genetic fuzzy systems. Evolutionary tuning and learning of fuzzy knowledge bases", *Advances in fuzzy systems. Applications and theory*, vol. 19, 2001.

[17] Galan, S.G., Bago, J.C., Aguilera, J., Velasco, J.R. and Magdalena, L. "Genetic fuzzy systems in stand- alone photovoltaic systems", *I International Workshop in Genetic Fuzzy Systems*, Granada, March 2005.