Anuran sound classification using MPEG-7 frame descriptors

Javier Romero[1], Amalia Luque[1*], Alejandro Carrasco[1]

[1]Escuela Politécnica Superior, Universidad de Sevilla, Spain
javier@romeroyromero.es,{amalialuque,acarrasco}@us.es

Abstract. Anuran sexual calls are highly influenced by temperature. So the existence of some species in a certain geographical area becomes a good indicator of climate change. For this purpose, biologists are recording huge amounts of animal sounds which have to be classified. In this paper we present an automatic anuran sound classification method based on MPEG-7 features. Up to ten data mining techniques are used to classify real field-obtained sounds, comparing their results. Quite good classification success is achieved, even considering the noisy ambient.

Keywords: sound classification · data mining ·MPEG-7 features · habitat monitoring

1 Introduction

Since Roger Revelle in 1950 [1] alerted about the consequences of the greenhouse gases, the international scientific community has been looking for indicators of human influence on climate. In 1972, Dennis Meadows showed in its report on the limits of growth [2] the first predictive computer model of the warming caused by fossil fuels. Since then many models have been proposed trying to explain the long-term evolution of different climate indicators.

One of the consequences of climate change is its impact on the development of basic physiological functions of various species [3,4]. Thus, for example, the sound produced during the mating call plays a central role in sexual selection and reproduction of numerous ectotherm species (those that regulate its temperature from ambient temperature), which include Anura (frogs and toads), fish and insects [5,6,7]. Various acoustic patterns are used to attract potential mates, as a defense, to ward off opponents, and to respond to the risks of predation. These sounds are therefore critical to adapting individuals to the environment.

However, sound production in ectotherms animals is strongly influenced by the ambient temperature [8] which can affect various features of acoustic communication system. In fact the ambient temperature, once exceeded a certain threshold, can restrict the physiological processes associated to the sound production even inhibiting behav-

adfa, p. 1, 2011.

iors call. As a result, the temperature may significantly affect the patterns of calling songs modifying the beginning, duration and intensity of calling episodes and, consequently, the anuran reproductive activity.

Therefore, the analysis and classification of the sounds produced by certain animal species have revealed as a strong indicator of temperature changes and therefore the possible existence of climate change. Especially interesting are the results provided by anuran sounds analysis [9].

This analysis requires first recording different sounds in their natural environment, where devices as the one described in [10] can be used. Processing of the recorded sounds can be locally performed in real time [11]; or in a remote center requiring, in this case, an adequate communication system (usually a wireless sensor networks), which generally requires information compressing technologies [12].

For the goal of obtaining climate change indicators, the in-field real time processing is usually not a requirement, so that sounds can be off-line analyzed from the available databases such as the sound library in the Museo Nacional de Ciencias Naturales [13].

Processing and classification of animal sounds is a recurrent issue for biologists. In [14] is presented a system specifically designed for anuran identification with the goal of providing open online searches.

The process of classifying sounds, one more specifically anuran sounds, can be split in two steps. In the first one the extraction of a more or less large set of features is performed. The second stage performs sound classification based on the aforementioned features.

Most of the used algorithms are based on spectral or temporal parameters such as, for example, the spectral centroid, the bandwidth, or the zero crossing rate [15]. Depending on the applications, sound types and, in many cases, authors' choice, these algorithms lack homogeneity, both in feature selection or even in how every parameter is defined.

A particular case of spectral parameters are the Mel Frequency Cepstral Coefficients: MFCCs. These parameters, widely used in sound classification processes, are uniquely defined and even standardized [16] so they are a good solution to cope with the heterogeneity described in the preceding paragraph.

While MFCCs are a set of standard parameters which can be applied to sound classification, they offer a one-dimensional view of the audio segment that they represent. Indeed, all parameters are derived from the same approach: the sound signal cepstrum (a function computed out of the power spectrum). Although using MFCCs has the advantage of standardization, they limit the search for more semantically expressive features.

An alternative is the use of MPEG-7 standard [17]. The goal of the standard is not mainly sound classification but a much broader objective: standardizing multimedia (text, images, audio, video...) description. In order to sounds representation the standard, in its Part 4, proposes a set of parameters and algorithms much richer from a semantic point of view than those provided by the MFCCs.

MPEG-7 parameters have therefore a double advantage: standardization and semantic richness. These features make them especially attractive to explore sound classification techniques [18]. Moreover, the comparison among the results obtained with MFCCs and MPEG-7 based techniques are not conclusive, depending on the application [19].

Furthermore, whatever the parameters used to characterize a sound, they are only the raw material for classification algorithms. Regarding this issue many solutions have been described, although the most common techniques are those based on Hidden Markov Models [20]. This is, indeed, the technique chosen for the MPEG-7 standard.

However, pattern classification, whether sound or not, it is a field with a large tradition and where many solutions have been suggested. With the name of machine learning, data mining, business intelligence or some others, very powerful algorithms can be found specialized in general pattern classification [21]; and, more specifically, in sequential data classification [22] and in sound classification [23].

2 Sound classification

The digital processing of acoustic signals is performed through two successive and complementary processes:

1. Significant sound features selection and extraction.
2. Sound classification (species identification) based on the previous extracted features.

For testing purposes sound files provided by the Zoological Sound Library [13] have been used, corresponding to 2 species, the epidalea calamita (natterjack toad) and alytes obstetricans (common midwife toad), with a total of 63 recordings containing 3 types of sounds:

1. Epidalea calamita; mating call (23 records)
2. Epidalea calamita; release call (10 records)
3. Alytes obstetricans (30 records)

In total 6,053 seconds (1h: 40 ': 53 ") of recording have been analyzed, with a 96 seconds (1': 36") average file length, and a 53" median.

A common feature of all recordings is that they have been made in the natural habitat, with very significant surrounding noise (wind, water, rain, traffic, voice...), which meant an additional challenge in the signal processing.

Sound classification process begins obtaining a sequence of descriptors. For this purpose, temporal sound frames are considered and, for every frame, the following parameters are calculated:

1. Total Power
2. Relevant Power
3. Spectrum Power Centroid
4. Spectrum Power Dispersion
5. Spectrum Flatness
6. Pitch
7. Harmonic Ratio
8. Upper Frequency of Harmonicity
9. Formants (harmonic peaks)
   (a) Three first formants' frequencies
   (b) Three first formants' bandwidth
10. Harmonic Centroid
11. Harmonic Deviation
12. Harmonic Spread
13. Harmonic Variation

Using these features we get a good description of every frame through an $\mathbb{R}^{18}$ vector. Therefore, every sound is described using a set of points in an $\mathbb{R}^{18}$ space. The classification of each file is achieved by comparing its set of points to the pattern sounds sets of points.

This comparison, known as supervised classification in data mining realm, can be performed using many different techniques. A broad and representative selection of them has been used through this paper: minimum distance [24]; maximum likelihood [25]; decision trees [26]; k-nearest neighbor [27]; support vector machine (SVM) [28]; logistic regression [29]; neural networks [30]; discriminant function [31]; and Bayes classifiers [32]. Additionally, these algorithms are compared to the one proposed in the MPEG-7 [17] standard: Hidden Markov Models [20].

The results of applying one of these techniques (decision trees) to the set of available audio files are shown in Fig. 1. In the X-axis sound files sorted by type of sound are represented: 1) epidalea calamita mating call (blue zone); 2) epidalea calamita release call (green zone); 3) and alytes obstetricans (red zone). For each file there is a vertical line in a color corresponding to the classification made by the algorithm (the same color code applies). In a perfect classification each line color should match the graph area color. Each difference is a misclassification. Finally, the height of each line is the probability that the algorithm assigns to the classification made.

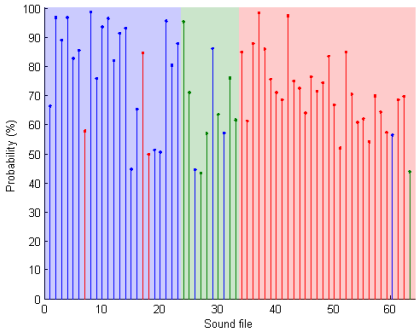

Fig. 1. Decision tree classification

Overall classification results using decision trees are summarized in Fig. 2. A more detailed analysis can be obtained through the confusion matrix as it is shown in Table 1.

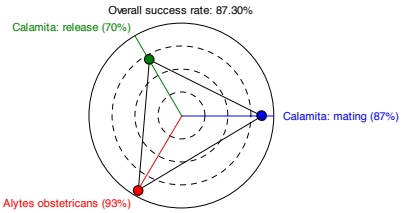

Fig. 2. Overall classification results using decision trees

Table 1. Confusion matrix

		Estimated sound		
		1	**2**	**3**
Real sound	**1**	86.96%	0.00%	13.04%
	2	30.00%	70.00%	0.00%
	3	3.33%	3.33%	93.33%

Algorithm results are also compared in Table 2. These results are graphically shown in Fig. 3. Bar's height reflects the overall success rate for each classifier. Points (with the usual color code) indicate the success rate for each type of sound.

Table 2. Algorithm classification results

Algorithm	**Sound 1 (23) Successes**		**Sound 2 (10) Successes**		**Sound 3 (30) Successes**		**Total (63) Successes**	
Minimum distance	14	61%	10	100%	0	0%	24	38.10%
Maximum likelihood	22	93%	5	50%	23	77%	50	79.37%
Decision trees	20	87%	7	70%	28	93%	55	87.30%
k-nearest neighbor	23	100%	5	50%	18	60%	46	73.02%
SVM	23	100%	2	20%	21	70%	46	73.02%
Logistic regression	23	100%	4	40%	14	47%	41	65.08%
Neural networks	11	48%	0	0%	29	97%	40	63.49%
Discriminant function	23	100%	4	40%	15	50%	42	66.67%
Bayes classifiers	15	65%	1	10%	29	97%	45	71.43%
Hidden Markov Models	18	78%	1	10%	29	97%	48	76.19%

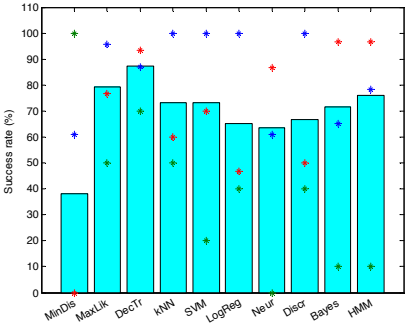

Fig. 3. Classifiers overall success rate

In order to select a classification algorithm, it is important not only to the overall success rate but also the balance for different types of sounds. To take into account this second factor we use as an indicator the error rate range R, defined as

$$R = \max_i E_i - \min_i E_i \quad (1)$$

where E_i represents the classification algorithm error rate for the i-th type of sound. Fig. 4 graphically shows the error rate range vs. error rate. A good sound classification algorithm should have both a low error rate $\bar{E}$ and a low error rate range R. Regarding points in Fig. 4, the closer to the origin, the better the algorithm is. A good measure combining both values is thus the distance D to the origin, calculated as

$$D = \sqrt{\bar{E}^2 + R^2} \quad (2)$$

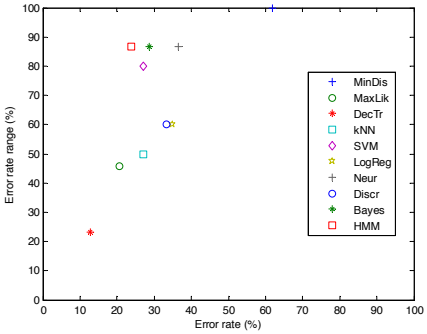

Fig. 4. Error rate range vs. error rate (%)

From the above equation we can easily derive a normalized figure of merit (between 0 and 1) given by the expression

$$M \equiv 1 - \frac{D}{D_{max}} = 1 - \frac{D}{\sqrt{2}} \quad (3)$$

The figures of merit for every classifier are presented in Table 3 . As can be seen, the decision tree is the classification algorithm presenting the best performance according to these criteria.

Table 3. Classifiers figure of merit

Algorithm	Success	Error	Range	Distance to origin	Figure of merit
Minimum distance	38.10%	61.90%	100%	1.18	16.84%
Maximum likelihood	79.37%	20.63%	43%	0.48	66.28%
Decision trees	87.30%	12.70%	23%	0.26	81.42%
k-nearest neighbor	73.02%	26.98%	50%	0.57	59.83%
SVM	73.02%	26.98%	80%	0.84	40.30%
Logistic regression	65.08%	34.92%	60%	0.69	50.91%
Neural networks	63.49%	36.51%	97%	1.04	26.71%
Discriminant function	66.67%	33.33%	60%	0.69	51.47%
Bayes classifiers	71.43%	28.57%	87%	0.92	35.25%
Hidden Markov Models	76.19%	23.81%	87%	0.90	36.22%

The aforementioned algorithms have been prototyped using built-in MATLAB® functions. For the selected technique (decision tree) the "fitctree" function with defaults parameters is employed. This function derives the classification tree from the frame parameters of some sound patterns (supervised classification). The MATLAB "predict" function is used for classifying the remaining sound frames.

3 Conclusions

The main motivation of the work has been to obtain an automatic sound classification tool to be applied in highly noisy natural environments. Sound classification leads us to detect certain anuran species which is later used for biologists as climate change indicators. On the other hand, the main novelty of the paper relies on the comparison of several sound classifiers which are based on MPEG-7 features.

From the above results, more detailed conclusions can be derived. In first place, features extraction using MPEG-7 parameters show very good results describing sound frames for classification purposes appearing as a serious competitors to MFCC features. Although extracting MPEG-7 features could require more computational effort, they are semantically richer and show remarkable classification performance.

Additionally, non-sequential sounds description (splitting the sound in frames, and not taking into account frames order) shows quite good classification result, in some cases significantly better than the sequentially oriented HMM.

Having thousands of frames to be classified, where every frame is described using an n-dimensional parameters space, highly resembles data mining problems. In the paper up to ten usual data mining techniques have been explored, and they have proven to be useful quite useful for sounds classification purposes.

Among these techniques, the best performance has been obtained using the decision tree classification algorithm. Its results clearly overcome the performance obtained through the Hidden Markov Models, the MPEG-7 proposed technique.

Finally, the overall classification result reaches a remarkable success rate (87.30%), even more relevant considering the low quality of the analyzed sounds. This figure could possibly be increased if the frames order were considered using proper sequential classification methods.

References

1. Revelle, R., & Suess, H. E. (1957). Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric CO2 during the past decades. *Tellus*, *9*(1), 18-27.
2. Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W. (1972). The limits to growth. *New York*, *102*.
3. Deutsch, C. A., Tewksbury, J. J., Huey, R. B., Sheldon, K. S., Ghalambor, C. K., Haak, D. C., & Martin, P. R. (2008). Impacts of climate warming on terrestrial ectotherms across latitude. *Proceedings of the National Academy of Sciences*, *105*(18), 6668-6672.
4. Duarte, H., Tejedo, M., Katzenberger, M., Marangoni, F., Baldo, D., Beltrán, J. F., ... & Gonzalez-Voyer, A. (2012). Can amphibians take the heat? Vulnerability to climate warming in subtropical and temperate larval amphibian communities. *Global Change Biology*, *18*(2), 412-421.
5. Bradbury, J. W., & Vehrencamp, S. L. (1998). *Principles of animal communication*. Sinauer Associates.
6. Fay, R. R. & Popper, A. N. (Ed.). (2012). *Comparative hearing: fish and amphibians* (Vol. 11). Springer Science & Business Media.
7. Gerhardt, H. C., & Huber, F. (2002). *Acoustic communication in insects and anurans: common problems and diverse solutions*. University of Chicago Press.
8. Márquez, R., & Bosch, J. (1995). Advertisement calls of the midwife toads Alytes (Amphibia, Anura, Discoglossidae) in continental Spain. *Journal of Zoological Systematics and Evolutionary Research*, *33*(3-4), 185-192.
9. Llusia, D., Márquez, R., Beltrán, J. F., Benitez, M., & Do Amaral, J. P. (2013). Calling behaviour under climate change: geographical and seasonal variation of calling temperatures in ectotherms. *Global change biology*, *19*(9), 2655-2674.
10. Cambron, M. E., & Bowker, R. G. (2006, December). An automated digital sound recording system: the Amphibulator. In *Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on* (pp. 592-600). IEEE.
11. Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, *1*, e103.
12. Diaz, J. J., Nakamura, E. F., Yehia, H. C., Salles, J., & Loureiro, A. (2012, November). On the Use of Compressive Sensing for the Reconstruction of Anuran Sounds in a Wireless Sensor Network. In *Green Computing and Communications (GreenCom), 2012 IEEE International Conference on* (pp. 394-399). IEEE.
13. Fonozoo.com (2015). Retrieved from http://www.fonozoo.com/

14. Huang, C. J., Yang, Y. J., Yang, D. X., & Chen, Y. J. (2009). Frog classification using machine learning techniques. *Expert Systems with Applications*, *36*(2), 3737-3743.
15. Fulop, S. (2011). *Speech spectrum analysis*. Springer Science & Business Media.
16. ETSI, E. (2002). 202 050 v1. 1.3: Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms. *ETSI standard*.
17. ISO (2001). *ISO 15938-4:2001 (MPEG-7: Multimedia Content Description Interface), Part 4: Audio*. ISO
18. Wang, J. F., Wang, J. C., Huang, T. H., & Hsu, C. S. (2003, December). Home environmental sound recognition based on MPEG-7 features. In *Circuits and Systems, 2003 IEEE 46th Midwest Symposium on* (Vol. 2, pp. 682-685). IEEE.
19. Kim, H. G., & Sikora, T. (2004, June). How efficient is MPEG-7 for general sound recognition?. In *Audio Engineering Society Conference: 25th International Conference: Metadata for Audio*. Audio Engineering Society.
20. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.
21. Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.
22. Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, *45*(1), 12.
23. Gopi, E. S. (2014). Digital Speech Processing Using Matlab. Springer India.
24. Wacker, A. G., & Landgrebe, D. A. (1971). The minimum distance approach to classification. Purdue University. Information Note 100771
25. Le Cam, L. M. (1979). *Maximum likelihood: an introduction*. Statistics Branch, Department of Mathematics, University of Maryland.
26. Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc
27. Cover, T. M., & Hart, P. E. (1967). *Nearest neighbor pattern classification*. Information Theory, IEEE Transactions on, 13(1), 21-27.
28. Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.
29. Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.
30. Du, K. L., & Swamy, M. N. S. (2013). *Neural Networks and Statistical Learning*. Springer Science & Business Media.
31. Härdle, W. K., & Simar, L. (2012). *Applied multivariate statistical analysis*. Springer Science & Business Media.
32. Hastie, T., Tibshirani, R., & Friedman, J. (2005). The elements of statistical learning: data mining, inference and prediction. Springer-Verlag.