X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

Computational vocalization recognition — an alternative to transect sampling in Tropical Forests

Jugurta Montalvão, Marcelo Cardoso
Universidade Federal de Sergipe (UFS), Instituto AMUIRANDE
jmontalvao@ufs.br, mcsousa@infonet.com.br

Resumo – Uma nova estratégia de apoio à estimação de densidade de primatas em regiões de floresta tropical densa é apresentada, baseada em biometria computacional, normalmente usada na verificação de orador. Na adaptação que é proposta, as vocalizações de primatas substituem a voz humana. Essa abordagem nova é motivada pela dificuldade existente na aplicação do método do Transecto em áreas de difícil acesso, bem como pela constatação de que os próprios primatas tendem a usar vocalizações, nesses ambientes densos, como forma de identificação individual à distância. Para tornar o processo de verificação resistente ao intenso ruído de fundo existente nas matas, um método de extração de características com mascaramento espectral foi usado. Os experimentos realizados com vocalizações de pares de Guigós, em 14 localidades do estado de Sergipe, resultaram numa taxa de erro de verificação de identidade de aproximadamente 11%.

Palavras-chave – Transecto; Biometria computacional; Mascaramento espectral; Guigós; Coimbra-filho?s titi monkey; Understory forest.

Abstract – A new auxiliary strategy for primate population density estimation in tropical forests is presented. This strategy is based on computational biometrics usually applied to speaker verification tasks. In this proposed adaptation, primate vocalizations replace human voices. The motivation behind this work is the difficulty associated to the application of transect-based methods in dense understory forest, along with the perception of the importance of primate vocalization as a natural communication tool, through which they also communicate their identity, by hypotheses, to other animals at distance (far beyond visual contact). Moreover, to make the new approach more robust to typical intense background noise in tropical forests, we deployed a feature extraction method based on (psychoacoustic) spectral masking. Experiments with couples's vocalizations from 14 locations in Sergipe (Brasil) yielded verification error rates of about 11%.

Keywords – Transect; Computational Biometrics; Spectral Masking; Coimbra-filho?s titi monkey; Understory forest.

1 INTRODUCTION

A usual method for estimating densities of terrestrial mammals (e.g. primates) in forested areas is the Transect. Strictly speaking, a transect is just a path along which one records and counts occurrences of the animal along the path, simultaneously estimating and registering the distance of the animal from the chosen path. This results in an estimate of the actual density of objects over that domain. There is a number of different types of transect methods, such as strip transects, line transects, belt transects, point transects and curved line transects.

Transect based methods have low operating costs and allow detection of a large number of species, but it is difficult to be applied in areas of secondary forest and areas of sharp terrain relief. In these locations, although many species are still detected through their vocalizations, they are rarely seen amid dense vegetation. In addition, some forest regions have too low of a density of individuals, thus making visual-based methods rather inefficient, especially when the time scheduled for field study is too restrictive.

On the other hand, according to [1], using play-backs to attract individuals shows advantages as compared to traditional visual transect, since it is not necessary to traverse the entire study area to make a population estimation. Indeed, by reproducing primate vocalization through recordings, it is possible to stimulate them to the point of increasing the chances of visual contacts. It is known, however, that the play-back can induce a wide variety of animal reactions, ranging from a simple back-vocalization to fear and silent reactions, depending on the species and/or the region where it is applied. Besides, there are some parameters that can affect the accuracy of play-back based approaches, such as average interval between plays, sound intensity (or distance from the targeted animal) and the "meaning" of specific recordings, from the point of view of the targeted animal.

This work proposes an alternative auxiliary approach for counting primates in tropical forests where high vegetation density and/or sharp relief hinders visual counting. We adapted computational algorithms already in use for human identity verification (computational biometrics) to non-human primates. We further hypothesize that most primate cries in forests are indeed aimed at individual identification. That is to say that primates probably adapt their vocalizations to compensate for environmental impairment, which may even explain why primate vocalizations in tropical forests concentrate acoustic energy in specific spectral bands, thus covering wide regions, and transmitting their (territorial) message as far as possible. Therefore, we believe that pri-

mates already use vocalizations as an important communication support in dense forests, where visual sight is almost impossible at long distances.

In our adaptation of biometrics to primates in tropical forests, we must take into account that (forest) background noise is an important disturbing factor. Therefore, we use a previously published feature extraction method (from recorded sounds) which, in former works, presented a strong resistance to additive noise. This method is based on the mimic of (psychoacoustic) masking effect in human listening. Its implementation is briefly explained in Section 2. Our preliminary experimental results are presented in Section 3, with Coimbra-Filho Titi monkey's (locally known as *Guigós*) vocalizations acquired from 14 locations in Sergipe (Brazil) [1].

2 Signal analysis and pattern recognition issues

Speaker identification/verification is a challenging modality of behavioral biometrics, in which robustness issues remain as open questions for researchers. Biometrics is the science of establishing the identity of an individual based on the physical, chemical or behavioral attributes of the person [2], and a biometric speaker recognition system is defined as a computer system capable of identifying a person based only on the information carried through his voice. In this work, we assume that this can be extended to non-human primates as well.

Frequently, Mel-frequency cepstral coefficients (MFCC) are deployed as a low-dimensional set of features to represent short-segments of speech. MFCC were first proposed in a technical report by Bridle and Brown [3], in 1974, as being the log spectrum transformed through a 19-channel filter bank, so that corresponding energies were, in turn, cosine transformed into 19 "spectrumshape" coefficients. Paul Mermelstein, through his book article titled "Distance Measures for Speech Recognition" [4], named this algorithm as mel-based cepstral parameters, thus using the MFCC acronym for the first time. In his work, he applied the algorithm to measure inter-word distances for a time-warping task in speech recognition. Since then, MFCC remains a powerful sound representation tool, for it partially mimics human perception of "sound color" [5], thus becoming widely popular in the signal processing community in its almost original form. Speaker verification is not an exception to this rule. For instance, in [6], 19 MFCC are extracted from overlapped short-frames of speech signals, whereas in [7] only the 12 lowest MFCC are used as acoustic features.

Unfortunately, it is well-known now that different operating conditions (during signal acquisition) severely affect MFCC (e.g. channel response, background noise), thus leading to feature mismatch across training and recognition. In order to cope with it, most approaches keep MFCC as features but introduce some kind of compensation. For example, in [7], Cepstral Mean Normalization has been used in order to remove linear channel distortion, along with RASTA filtering and feature warping, that have been used in order to achieve robustness against channel and noise effects. On the other hand, the authors of [7] also claim that "state-of-the-art text-independent speaker recognizers use mean subtraction at the utterance level, often referred to as cepstral mean subtraction (CMS)", even though CMS may degrade accuracy recognition of clean data (no channel mismatch).

Alternatively, noise compensation may be done directly during MFCC computation through spectral subtraction per band and/or by changing the band logarithmic energy compression with constant root functions (possibly with adaptive root parameters). In [8], four such strategies are cross-compared, including a new one proposed by the authors, where adaptive root energy compression and noise compensation in sub-bands, together, outperform all the others strategies in an isolated word recognition task.

In [9] we claimed that most efforts on channel compensation and sophisticated pattern recognizer design can be saved through a very simple change in MFCC computation, namely, the inclusion of spectral masking in its algorithm. We further showed that the Ensemble Interval Histogram (EIH) [10] and the Zero Crossings with Peak Amplitude (ZCPA) [11] (two main MFCC-related alternative features) do implement spectral masking, which may explain their reduced sensitivity to external noise.

Here, we use the MFCC with rectangular lateral masking (FastMask-R) proposed and detailed in [9], because it is remarkably robust to additive noise. This feature extraction method can be summarized as follows.

We start with an MFCC implementation, with:

- Short-time analysis: 25 ms per frame.
- Overlapping between frames: 82% (advance of 4.5 ms per frame)
- Blackman window instead of (typical) Hamming window
- Frequency scale: Mel
- Filter shapes: rectangular
- Filter Bandwidth: Constant, in Mel scale (403.2 Mel)

Given all overlapping short frames of signals in order to discard silent frames (or frames with too low acoustic energy), we set an adaptive energy threshold and systematically discard frames whose energy is below it, as illustrated in Figure 1, for a typical Guigó vocalization recording (approx. 3s).

[1] Coimbra-Filho Titi monkeys, locally known as *Guigós*, are endangered primates.

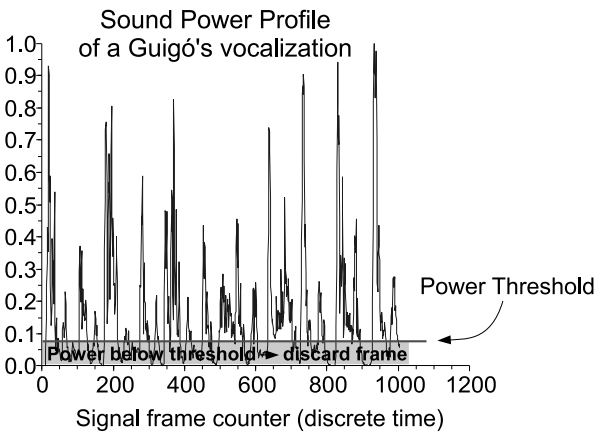

Figura 1: First signal segmentation through power profile: frames with signal energy below threshold are discarded.

This power thresholding improves Signal-to-Noise Ratios by selecting frames with higher "foreground" sound power (since the background noise in tropical forests is almost constant, on average, during recording sessions). Selected frames are then frequency analyzed and their spectra are split into 145 overlapping sets of spectral coefficients, thus simulating 145 rectangular filters with constant-bandwidth in Mel scale (i.e. non-constant in Hz scale). In order to include the masking effect, unlike usual MFCC, *we do not sum up spectral coefficients in each set*, but we just find the frequency of the maximum energy value from each band. These peak frequencies (one per band) are then represented in a histogram, i.e. a counter of how many times a given frequency was detected as a spectral peak.

This histogram plays the role of a spectral representation where irrelevant spectral energy is discarded (masked). Finally, the inverse co-sinus transform of this histogram (seen here as a real-valued vector) provides a vector of coefficients hereafter called FastMask-R.

3 Experimental Results

Coimbra-Filho titi monkey's vocalizations were acquired from 14 locations in the states of Sergipe (Brazil). Figure 2 gives a rough idea of location distribution, each one corresponding to a patch of forest where at least one titi monkey was previously observed. Recordings were made randomly in the early hours of the day, when higher activity of primates is frequently observed. Moreover, available sounds were acquired during the years 2003, 2005, 2008 and 2011, with the very same devices, namely: a Sony TCM 5000 recorder and directional microphone Sennheiser ME66. All recordings were digitalized with 16 bits per sample and sampling frequency of 44100Hz. Long recorded segments were further manually split into short files of about 3 s. Figure 3 illustrates this signal acquisition and digitalization process.

Among all the locations, we were able to assure that location Parui was that were only a single pair of titi monkeys (a couple) took part in the recordings. Since titi monkeys frequently vocalize in duets, or even in groups, and since their individual vocalizations are difficult to be unmixed, we simplify our goal by assuming that the couple in Parui location is "one individual". Therefore, we test whether biometrics is able to discern Parui's couple from other couples, from another 13 locations.

Accordingly, the computational experiment protocol can be summarized as follows:

(a) 5 short files ($\approx 3s$ each one) are randomly chosen from those made in Parui's location. These 5 short files are processed to provide a single matrix of 19-FastMask coefficient vectors, one vector per frame of 25ms.

(b) 1 short file is randomly chosen from:

  (b.1) recordings made in Parui's location

  (b.2) recordings made elsewhere

(c) Vector coefficients from (a) and (b) are regarded as sets of randomly generated vectors. Thus, the two "random sources" of vectors are compared with a K-Nearest Neighbourhood based classifier, with K=5, slightly modified in order to provide scores between 0 and 1, instead of averaged distances between feature vectors.

(d) If the single recording comes from Parui's location (b.1), the resulting score is labeled as 'True', otherwise it is labeled as 'False'.

(e) After a number of scores are obtained and labeled, we adjust a decision threshold, T, so that the balance between False Accept Rate (FAR) and False Rejection Rate (FRR) is minimized.

Where

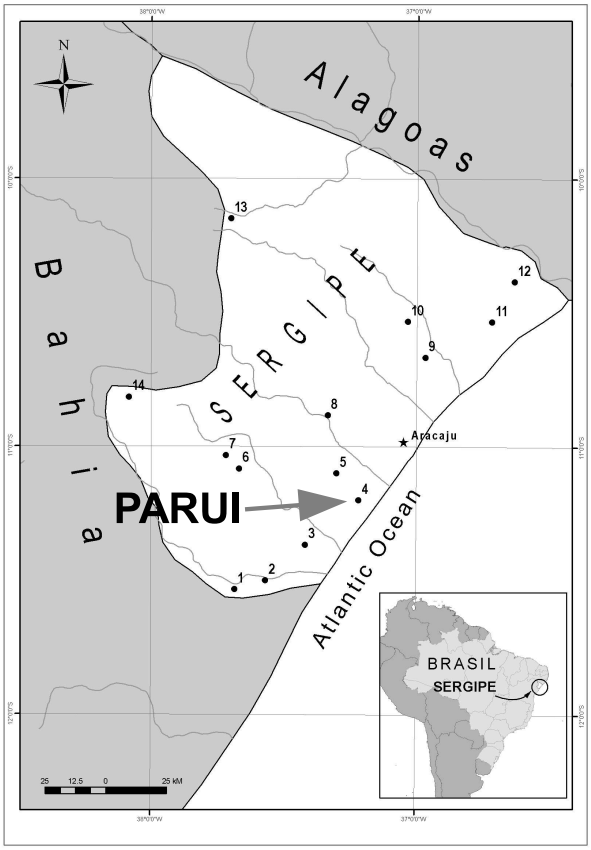

Figura 2: Signal acquisition locations in Sergipe (Brazil).

- FAR: the number of scores above T and labeled 'False', divided by the total of scores labeled 'False'
- FRR: the number of scores below T and labeled 'True', divided by the total of scores labeled 'True'

After 1092 independent random samplings and score comparisons, we found an Equal Error Rate (EER) — i.e. the operational point where FAR equals FRR — of $EER \approx 11\%$, which means that the Parui's couple was correctly detected in 11 out of 100 short recordings of their vocalizations, given a model build up with approximatively 15 seconds (i.e. $5 \times 3s$) of vocalization. Figure 4 illustrates this result.

4 DISCUSSIONS AND CONCLUSIONS

A new auxiliary strategy for primate population density estimation in tropical forests was presented. This strategy, based on computational biometrics, was adapted to primate vocalizations instead of human voices. The motivation behind this work was two fold: first because it is difficult to apply transect-based methods in dense understory forest, and secondly because we hypothesize that primate's vocalizations are naturally aimed at broadcasting their identification to others of its species. Evidences in favour of this hypothesis can be easily gathered. For instance, titi monkey's vocalizations are powerful, in terms of acoustic energy, and it seems to be fitted to overpower forest noise. Moreover, it is used to intimidate territorial invaders, as well as

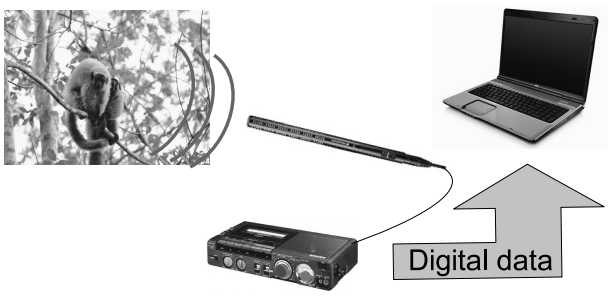

Figura 3: Signal acquisition illustration.

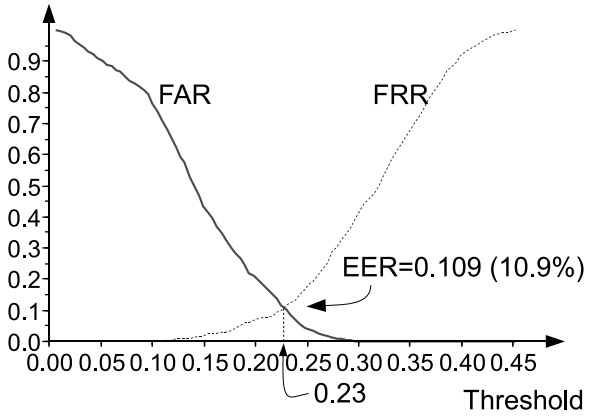

Figura 4: FAR and FRR empirical analysis – Parui's couple vocalization against other couples from 13 other locations.

to attract sexual partners in environments where visual contact is difficult. Therefore, we may conclude that cry features and modulations are aimed at transmitting individual messages across long distances (as long as possible). We also believe that cry features evoluted to compensate for environmental acoustic absorption and noise.

Besides, in biometric human identity verification, voice individualization is partially due to physiologic aspects (vocal tract dimensions). In the case of non-human primates we do not have reasons to believe that it should be different.

To make the new approach more robust to intense background noise in tropical forests, we deployed a feature extraction method based on (psychoacoustic) spectral masking. Experiments with couples's vocalizations from 14 locations in Sergipe (Brasil) yielded verification error rates of about $\approx 11\%$, under $SNR \approx 10dB$. We highlight that similar results were obtained in [9]. This corroborates the idea that computational biometrics can be successfully adapted to non-human primates.

Nonetheless, results presented here are preliminary, rather playing the role of an introductory step to induce further works on this new, possibly fruitful, application research domain. Accordingly, to allow further comparisons between the results reported here and performances of other approaches, with the same database, samples used in this work (short 'wav' files) are freely available to download at www.biochaves.ufs.br (Internet web site).

REFERENCES

[1] W. P. Martins. "Distribuição Geográfica e Conservação do Macaco-Prego-de-crista , Cebus robustus (Cebidae, Primates)". Master's thesis, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Agosto 2005.

[2] A. K. Jain, P. Flynn and A. A. Ross. *Handbook of Biometrics*. Springer Verlag, New York, 2007.

[3] J. S. Bridle and M. D. Brown. "An Experimental Automatic Word-Recognition System". Technical Report 1003, Joint Speech Research Unit (JSRU) Report, , Ruislip, England, 1974.

[4] P. Mermelstein. "Distance measures for speech recognition, psychological and instrumental". *Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed.*, pp. 374–388, 1976.

[5] H. Terasawa. "A Hybrid Model for Timbre Perception: Quantitative Representations of Sound Color and Density". Ph.D. thesis, Stanford University, Stanford, 2009.

[6] D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez and J. Ortega-Garcia. "Speaker verification using speaker- and test-dependent fast score normalization". *Pattern Recognition Letters*, vol. 28, pp. 90–98, 2007.

[7] V. Hautamaki, T. Kinnunen and P. Franti. "Text-independent speaker recognition using graph matching". *Pattern Recognition Letters*, vol. 29, pp. 1427–1432, 2008.

[8] B. Nasersharif and A. Akbari. "SNR-dependent compression of enhenced Mel sub-band energies for compensation of noise effects on MFCC features". *Pattern Recognition Letters*, vol. 28, pp. 1320–1326, 2007.

[9] J. M. ao and M. Araujo. "Is Masking a Relevant Missing Aspect of MFCC? A Speaker Verification Perspective". *Submitted to Pattern Recognition Letters (ELSEVIER) in Nov. 2010*, 2011.

[10] O. Ghitza. "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition". *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, pp. 115–131, 1994.

[11] D.-S. Kim, S.-Y. Lee and R. Kil. "Auditory Processing of Speech Signals for Robust Speech Recognition for Real-World Noisy Environments". *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, 1999.