Acoustic detection of multiple birds in environmental audio by Matching Pursuit

Dan Stowell and Mark D. Plumbley DAN.STOWELL@EECS.QMUL.AC.UK
Centre for Digital Music, Queen Mary, University of London

Abstract

We describe a submission to the ICML 2013 Bird Challenge, in which we explore the use of sparse representations as an advance on the standard technique of cross-correlation template matching in time-frequency representations. The Matching Pursuit algorithm is used to represent the signal as a sparse set of activations of templates derived from the challenge training audio.

Given an audio recording, it is a challenging task to detect automatically which bird species are represented, and a task that is relevant to practical applications in bioacoustics (Stowell & Plumbley, 2010). Recent research developments go beyond single-label classification and can identify multiple species simultaneously present in a recording (Briggs et al., 2012), or track multiple birds through an audio scene (Stowell & Plumbley, submitted). The ICML 2013 Bird Challenge stimulates developments in the field by challenging researchers to identify algorithmically which of 35 bird species are present in a public dataset of 90 audio recordings.[1] The present note describes a contribution to the challenge which explores the use of *sparse representations* in a multi-label classification problem.

In signal processing, a *sparse representation* is recovered by assuming that the signal is composed from some "dictionary" of atomic elements, with only a small number of those elements being active (nonzero) for any given signal of interest (Plumbley et al., 2010). This approach is motivated by the discovery that neural coding often makes use of such sparsity, and also by the engineering prospect of representing signals in highly compact form. Sparse representations are currently the subject of much research activity, and have been used in audio and music for tasks such as audio compression and transcription (Plumbley et al., 2010).

Our submission explores sparse representation to improve on the common technique of cross-correlation template matching in time-frequency representations (such as spectrograms). In the standard cross-correlation scenario, we have one or more templates per species, and each template is separately cross-correlated against the spectrogram in question. Peaks in the cross-correlation function are taken as detections for the corresponding species. However, when there is a large number of species to be detected, and some of these potentially have very similar templates, there is a problem: a single region of energy in the spectrogram (e.g. a single birdsong syllable) could independently match against multiple templates, giving spurious detection of many species from a single sound. Sparse decompositions can overcome this, by finding a representation of the signal as a sum of activations from all elements considered together as a dictionary.

1. Method

In the present case we use the simple and widespread *Matching Pursuit* (MP) algorithm to find a sparse representation of a sound spectrogram as a sum of templates. We use a fast optimised implementation provided by the free and open-source Matching Pursuit Toolkit (MPTK) (Krstulovic & Gribonval, 2006). Our script is implemented in Python, using the Python bindings recently available in MPTK 0.7.

Preprocessing: Each audio file is converted into a standard spectrogram representation (log-magnitudes of STFT, frame size 512). We apply median-based background subtraction to help counter stationary background noise.

Training: Training the system consists of creating a dictionary of spectrogram "patches" from the training audio files. Each file is divided into segments using

[1] `www.kaggle.com/c/the-icml-2013-bird-challenge`

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

simple power thresholding, and long segments are further divided into multiple segments of maximum duration one second. We then discard segments which do not contain much structure and are broadly flat, since these are not strongly discriminative and can match against any noise. This decision is based on the crest factor of pixels within the patch. Each segment is then normalised in magnitude (to unit L_2 norm).

Testing: To analyse an audio file, we apply MP to its spectrogram using the dictionary of time-frequency segments, via MPTK. This produces a list of activations associated with elements in the dictionary, which can be used to reconstruct the signal or for further analysis. In the present case the required output is a list of probabilities per species. We derive the probability for each species heuristically as proportional to the total energy that MP has allocated to activations associated with that species.

2. Results

At time of writing, the AUC score on the annotated development data is 70.3%, and the AUC score evaluated on $\frac{1}{3}$ of the full data (this is the method used on the Kaggle website to give results-in-progress) is 64.6%. This demonstrates that the approach generalises satisfacorily. Results evaluated on the full dataset will be available after the challenge deadline.

3. Discussion

Note that this submission does not make use of any information other than the training and test audio provided by the challenge. In particular, for a real working system we would advocate the use of much larger audio collections to build the training data. However we wanted to explore how well the approach could make inferences from the provided data. Also, we do not perform adaptation of the system to the differing weather conditions, nor to the very different (reverberant) acoustic environment of the test audio.

In classical template-matching approaches, Dynamic Time Warping is commonly used to match templates against signals which have similar shape but with local differences in the length/speed of subregions. The MP approach we have deployed has no direct equivalent of this, which is a drawback when analysing sounds such as birdsong with natural variability in their production. The incorporation of such flexibility into sparse representations is an open topic; alternatively, novel signal representations may counter this issue.

We also note that template-matching approaches generally do not consider any time-sequencing of sounds at larger timescales, e.g. the grammatical sequencing of syllables. We are exploring how to combine syllable-by-syllable detection with methods which make inferences from temporal sequencing of birdsong, such as the Markov renewal process method we recently introduced (Stowell & Plumbley, submitted).

Future refinements of this approach could include more advanced approaches to creating the dictionary. For example, one might apply *dictionary learning*, a technique in sparse representations which directly optimises the dictionary so as to represent its inputs sparsely. We are also currently working with alternative signal representations which can provide detail of fine frequency modulations (Stowell et al., accepted).

Supported by EPSRC Fellowship EP/G007144/1.

References

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X. Z., Raich, R., Hadley, S. J. K., Hadley, A. S., and Betts, M. G. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *J Acoustical Society of America*, 131:4640–4650, 2012. doi: 10.1121/1.4707424.

Krstulovic, S. and Gribonval, R. MPTK: Matching Pursuit made tractable. In *Proc ICASSP 2006*, volume 3, pp. 496–499, Toulouse, France, May 2006. doi: 10.1109/ICASSP.2006.1660699.

Plumbley, M. D., Blumensath, T., Daudet, L., Gribonval, R., and Davies, M. E. Sparse representations in audio and music: From coding to source separation. *Proceedings of the IEEE*, 98(6):995–1005, 2010. doi: 10.1109/JPROC.2009.2030345.

Stowell, D. and Plumbley, M. D. Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers. Technical Report C4DM-TR-09-12, Centre for Digital Music, Queen Mary University of London, Aug 2010. URL `http://www.elec.qmul.ac.uk/digitalmusic/papers/2010/Stowell2010-C4DM-TR-09-12-birdsong.pdf`.

Stowell, D. and Plumbley, M. D. Segregating event streams and noise with a Markov renewal process model. submitted. preprint arXiv:1211.2972.

Stowell, D., Muševič, S., Bonada, J., and Plumbley, M. D. Improved multiple birdsong tracking with distribution derivative method and Markov renewal process clustering. In *Proc Int Conf Audio and Acoustic Signal Processing (ICASSP) 2013*, accepted. preprint arXiv:1302.3642.