Feature design for multilabel bird song classification in noise (NIPS4B challenge)

Dan Stowell and Mark D. Plumbley
Centre for Digital Music, Queen Mary University of London, London, UK
dan.stowell@qmul.ac.uk

Bird vocalisations are highly varied, containing natural variation across a range of timescales. In recent work we have modelled the transitions between syllables [4], and combined this with representations which capture fine FM variations within syllables [2]. Following on from such work we are exploring feature design for the representation of temporal structure in sounds such as birdsong.

The 2013 NIPS4B bird song challenge is concerned with automatically recognising the presence of a number of species, from sound alone. For training, 687 audio clips are provided (each annotated as containing 0–6 species); for testing, 1000. The clips are around 5 seconds long, recorded by automated monitoring units, and often noisy and with the target sounds both distant and quiet.

Our submission to the challenge therefore focusses on feature design for two goals: noise robustness, and the representation of temporal structure. We first analyse each sound file into basic features, either MFCCs (13 MFCCs plus delta features) or our peak-chirplet representation [3]. Importantly, both of these feature algorithms are modified to apply noise reduction in their spectral analysis step, simply by median-filtering: taking a spectral median profile across time, subtracting this median from the values, and keeping only the positive values.

We then reduce the time-series data for each file down to an atemporal summary vector. We summarise our noise-reduced MFCCs by their mean and standard deviations, a simple and common baseline approach. We summarise our chirplets by a histogram of all the bigrams found in the file: in other words, for every transition from one packet of energy to another, we record the time separation as well as the frequency and chirp-rate values, and these parameters form the axes of the high-dimensional histogram we create (related to the method in [3]). The time-separation between bigram pairs is not constant: we examine all possible transitions shorter than 1 second. Note that we avoid any need to perform segmentation on the input audio files. The histogram represents the set of all transitions observed in the audio data, and is used directly for multilabel classification.

For multilabel classification we use Random Forests implemented in scikit-learn [1]. In variations of our submission, we use either MFCC statistics (52 dimensions), chirplet histograms (up to 20,000 dimensions), or both. We experimented with dimension reduction but found this unnecessary. We also experimented with other multilabel classifiers, but found they generally reduced performance relative to Random Forests. On the public leaderboard (evaluated on a test set of 333 audio files), we attain 89.5% by the Area Under the Curve (AUC) score.

Acknowledgments: DS & MP are supported by an EPSRC Leadership Fellowship EP/G007144/1.

References

[1] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[2] D. Stowell et al. Improved multiple birdsong tracking with distribution derivative method and Markov renewal process clustering. In *Proc ICASSP*, 2013. preprint arXiv:1302.3642.

[3] D. Stowell and M. D. Plumbley. Framewise heterodyne chirp analysis of birdsong. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 2694–2698, 2012.

[4] D. Stowell and M. D. Plumbley. Segregating event streams and noise with a Markov renewal process model. *Journal of Machine Learning Research*, 14:1891–1916, 2013.