THE 9TH ANNUAL MLSP COMPETITION: SECOND PLACE

Hong Wei Ng and Thi Ngoc Tho Nguyen

Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore
{hongwei.ng, tho.nguyen}@adsc.com.sg

ABSTRACT

The MLSP 2013 Bird Classification Challenge requires participants to predict the set of bird species present in audio clips in a given test set, with the aim of maximizing the micro-AUC score computed from the predictions. This report summarizes the 2nd place solution by team Herbal Candy that achieved a micro-AUC score of 0.95050.

***Index Terms*—** Audio classification

1. INTRODUCTION

The MLSP 2013 Bird Classification Challenge requires participants to predict the set of bird species present in audio clips collected in the H. J. Andrews (HJA) Long-Term Experimental Research Forest. Participants are given 322 audio clips for training and another 323 for testing. Each audio clip in the training set is labelled with the set of bird species present or otherwise given no labels. The problem is formulated as a multi-instance multi-label machine learning problem where an audio clip can potentially contain multiple instances of bird vocalizations and the classifier has to predict probabilities indicating the presence of 19 species of birds in the clip. The goal is to maximize the micro-AUC score computed from the predictions made on the test set.

The proposed approach first detects segments of interest in each spectrogram computed from the given audio clips, $a^{(1)}, \ldots, a^{(645)}$, using a sub-band energy ratio-based method. Next, audio and image features are computed for each of the segments and summarized using a "bag-of-words" (BOW) model. These features are further augmented with features encoding statistics relating species of birds to the environment they appear in and the level of "interesting sound" in a clip. Finally, Extremely Randomized Trees [1] classifiers are trained using the binary relevance method to make the required predictions. We describe the details of our approach in the following sections.

This work is supported by the research grant for ADSC's Human Sixth Sense Programme from Singapore's Agency for Science, Technology and Research (A*STAR).

2. SEGMENTING SPECTROGRAMS

The goal of segmentation is to extract segments (either bird or unknown sounds) in the spectrogram of an audio clip that are significantly different from background noise, after which features are computed to describe them. Each clip is first passed through a pre-emphasis filter to highlight the high-frequency components before computing its power spectrogram. We discard the first 50 frequency bins based on observations that bird songs in the clips tend to occupy only high frequencies. The noise floor [2] of the audio clip is then subtracted from its spectrogram and a 2D median filter is further applied to remove salt and pepper noise residues. Following this, we compute the sub-band energy ratio (16 bands) of each frame in the spectrogram to produce a sub-band spectrogram with lower frequency resolution. This has the effect of joining up segments near each other and removing small ones. An adaptive threshold is applied on this new spectrogram to obtain a binary mask, which is then upsampled back to the size of the original spectrogram and used to segment the pre-emphasized spectrogram (Figure 1).

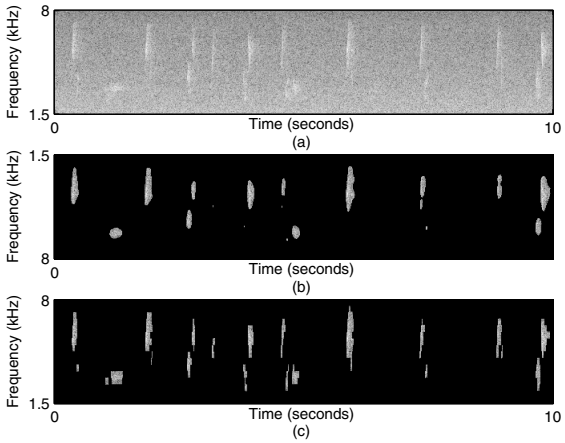

Fig. 1. Spectrograms of audio clip with id 2. Best viewed on a computer monitor. (a) Original spectrogram. (b) Segmented spectrogram (by [2]). (c) Segmented spectrogram (ours). Notice the faint signals detected by our algorithm.

978-1-4799-1180-6/13/$31.00

3. FEATURES

We compute three sets of features for each audio clip: instance, noise, and location features. Instance features are later used to build a "bag-of-words" model (see Section 3.2).

3.1. Raw features

Instance features for an audio clip $a^{(i)}$ consist of features derived from the binary mask described above and audio features from the segmented spectrogram. The binary mask features of each segment, computed as in [2], are: minimum frequency, maximum frequency, bandwidth, duration, area, perimeter, non-compactness and rectangularity. Likewise for audio features, we compute MFCC, delta MFCC, LPCC, and spectral properties consisting of sub-band energy, sub-band entropy, centroid, bandwidth, roll-off, and flux. These features are then averaged over the frames for each segment.

Noise features for $a^{(i)}$ consist of entropy, $e^{(i)}$, and correlation coefficient between averaged noise frame and average signal frame, $cr^{(i)}$. Together they characterize the amount of "interesting" sound present in the audio clip.

To encode the intuition that the appearance of bird species tend to be correlated to the environment they are found in, we compute a fixed vector of probabilities for each location l, $p_l \in R^{19}$, with each entry indicating the empirical probability that one of the 19 species of birds will appear at that location. The audio clip $a^{(i)}$ is assigned the location feature $p^{(i)}_{loc(i)}$ where $loc(i)$ gives the location clip i was recorded at.

3.2. Bag-Of-Words

We use the "bag-of-words" [3] approach to "reduce" the instance features (see Section 3.1) of each audio clip to a single feature vector. Two codebooks of sizes 100 and 50 are created from the same set of instance features using k-means clustering. Their sizes are tuned to maximize our classifier's micro-AUC score (see Section 4). For the audio clip $a^{(i)}$, the first codebook (size 100) is used to encode the number of times its instance features are closest to each of its codewords. This is represented as a vector $b^{(i)}_1 \in R^{100}$. The second codebook (size 50) plays a similar role as the first but encodes the distance to its corresponding codewords instead of counts. This gives a second vector $b^{(i)}_2 \in R^{50}$.

4. CLASSIFICATION

The audio clip $a^{(i)}$ is represented by a single feature vector formed by concatenating all the above features: $x^{(i)} := \langle b^{(i)}_1, b^{(i)}_2, e^{(i)}, cr^{(i)}, p^{(i)}_{loc(i)} \rangle \in R^{171}$. We set these values to all zeros for clips that we cannot detect any instance features. Further, we introduce an additional "noise" class and assign all the unlabelled clips (i.e., clips with none of the 19 species of birds) to it. The purpose of this "noise" class is to have it represent everything except birds that our segmentation algorithm finds (e.g., rain sounds), as well as static background noise when it detects no segments. This allows our classifier to explicitly model non-bird sounds *and* the absence of any sounds that stand out from the background noise, thereby improving its predictions for clips having these characteristics. Note that we do not modify existing labels in the training set to account for this new class as this is unfeasible and likely not allowed. We construct a single-instance multi-label classifier using the binary relevance method with Extremely Randomized Trees [1] as classifiers and train it using 5-folds cross-validation to maximize the micro-AUC score on this training set. Our classifier actually predicts 20 probabilities for each clip: 19 for the bird species, and 1 for "noise". When reporting our result, we only output the 19 probability values corresponding to the 19 bird species.

5. DISCUSSION AND RESULTS

We believe our segmentation algorithm, which was able to segment very faint signals, contributed significantly to our final result. Figure 1 compares one of our segmentation with that provided by the organizers (computed using the method described in [2]) and shows the improvement made by our algorithm in detecting segments. However, an unwanted side effect of this was an increase in the number of segments not belonging to bird calls. We handled this by introducing a "noise" class (see Section 4) to model these sounds, which we believe also helped our classifier make better predictions for audio clips that contained no bird songs. Adding this "noise" class improved our cross-validation score by 3-4%.

Our experiments also showed that the type of spectrogram used for performing segmentation and computing features matters. We found that using power spectrogram, rather than magnitude, gave better results as it produced signals which were more different from background noise.

Lastly, using location features led to a 1-2% improvement in our cross-validation score. This was unsurprising as the species of birds that appear at a location is expected to be highly correlated to the ecological conditions of that place.

6. REFERENCES

[1] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[2] F. Briggs et al., "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *Journal of Acoustical Society of America*, vol. 131, no. 6, pp. 4640–50, 2012.

[3] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification.," in *INTERSPEECH*, 2012.