Bird Sounds Classification by Large Scale Acoustic Features and Extreme Learning Machine

Kun Qian, Zixing Zhang, Fabien Ringeval, Björn Schuller

Session "Biological and Biomedical Signal Processing"
December 16, 2015

Outline

- Motivation
- Approach
- Database
- Experiments
- Conclusion

Motivation

- Monitoring CLIMATE CHANGE and HABITAT LOSS.
- Classification of bird species by their sounds is less expensive and superior in bad weather condition than telescope.
- Interdisciplinary Study: Ecology, Zoology, Bioacoustics, Signal Processing, Machine Learning, Big Data, etc.

Motivation

- Systematic Framework
  - Syllables Detection: How to find the suitable units for further feature extraction and machine learning? (Supervised or Unsupervised, Semi-supervised)
  - Feature Extraction: How to define the capable descriptors for feeding the learning model? (Speech-like or New)
  - Feature Selection: How to re-generate or modify the original Lower Level Descriptors (LLDs) for reducing the feature dimensions? (Classical Methods or Deep Neural Network)
  - Machine Learning: How to set up feasible learning architecture? (Extreme Learning Machine)

Approach

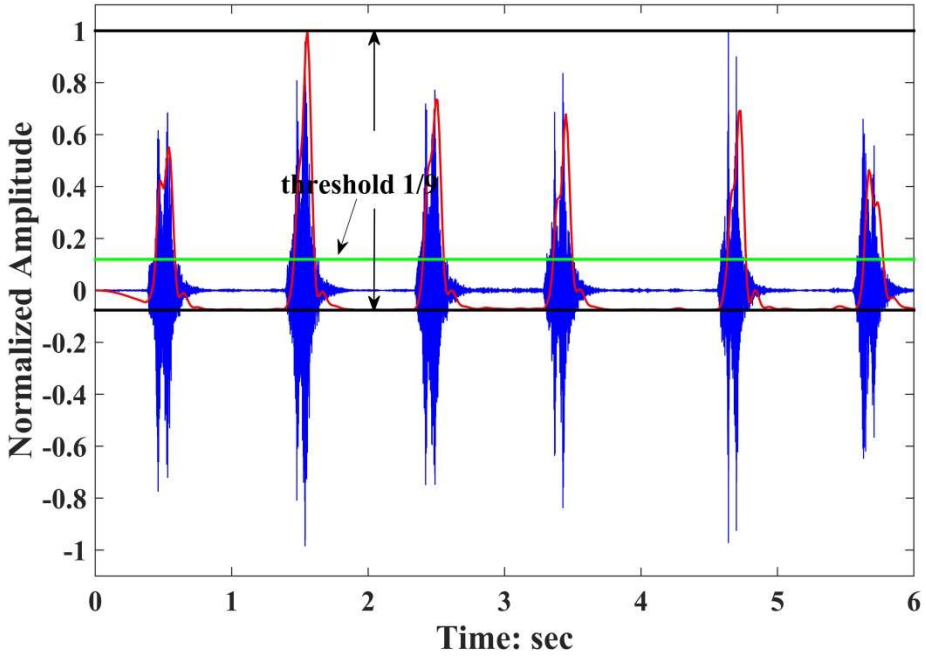
- **Syllables Detection**: Unsupervised method based on p-center detector.
- **Large Scale Acoustic Features Extraction**: openSMILE toolkit (INTERSPEECH 2009 Emotion Challenge feature set).
- **Feature Selection**: ReliefF algorithm (ranking features by their performance on classification).
- **Machine Learning**: Extreme Learning Machine (ELM).

P-center Detector

- Originated from estimating the values of entropy, the average frequency, and the centroid with the rhythmic envelope.
- No needs for data training phase, which is usually time-consuming and taking much more human works than unsupervised methods.
- Adaptive to current processing audio recording (e.g., the quality of audio signals, the background noise level, and the specific bird sound characters, etc.).

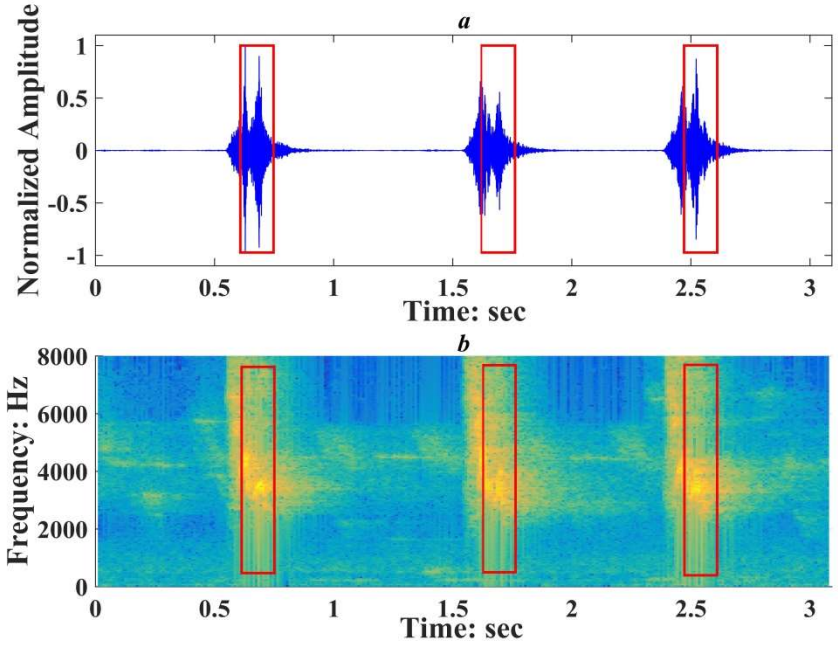
S. Tilsen and K. Johnson, “Low-frequency fourier analysis of speech rhythm,” *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. EL34–EL39, 2008.

P-center Detector



P-center represents the prominent part of the audio signal. Thus, the syllables can be detected when a suitable threshold and consecutive duration are set. (bird species: house sparrow)

P-center Detector



Detection of syllables by p-center and its corresponding spectrogram. (bird species: house sparrow)

Large Scale Acoustic Features Extraction

- INTERSPEECH 2009 EC standard feature set:
  12 functionals, 2 x 16 acoustic Low-Level Descriptors (LLDs) , with first order delta regression coefficients, totally 12 x 2 x 16 = 384 dimensions.
- Toolkit: openSMILE

  http://opensmile.sourceforge.net/

LLDs (16)	**Statistical functionals** (12)
MFCC 1–12	max, min, range, maxPos and minPos
RMS Energy	(absolue position of maximum/minimun
ZCR	value in frames), arithmetic mean, slope,
F0	offset and quadratic error for a linear approxi-
HNR	mation, standard deviation, skewness, kurtosis

Feature Selection

- *Feature Ranking* (to know which one is good or bad).
- ReliefF (can be regarded as an evaluator to rank features)

We can get the ranking weights $W_{(i)}$ of the i-th feature evaluated by ReliefF algorithm. In our study, we introduce *contribution rate* to select the better features for further machine learning phase:

$$contribution\ rate = \frac{\sum_{j=1}^{M} W_{(j)}^{+}}{\sum_{i=1}^{N} W_{(i)}^{+}},$$

where W^{+} represents t features evaluated by ReliefF.

M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," Machine Learning, vol. 53, no. 1-2, pp. 23–69, 2003.

Classifier: Extreme Learning Machine (ELM)

- Fast and Efficient
- A Feedforward Neural Network with a Single Hidden Layer
- Three-Step Learning Model

Parameters Setting:

Activation Function: 'radbas';

Number of Hidden Nodes: 30, 000.

codes available @:

http://www.ntu.edu.sg/home/egbhuang/elm_codes.html

G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

Database

- Free & Public Database @ (the picture below is also from:)

http://gallery.new-ecopsychology.org/en/voices-of nature.htm

(54 species of birds, recorded in real field with high audio quality)

Experimental Results

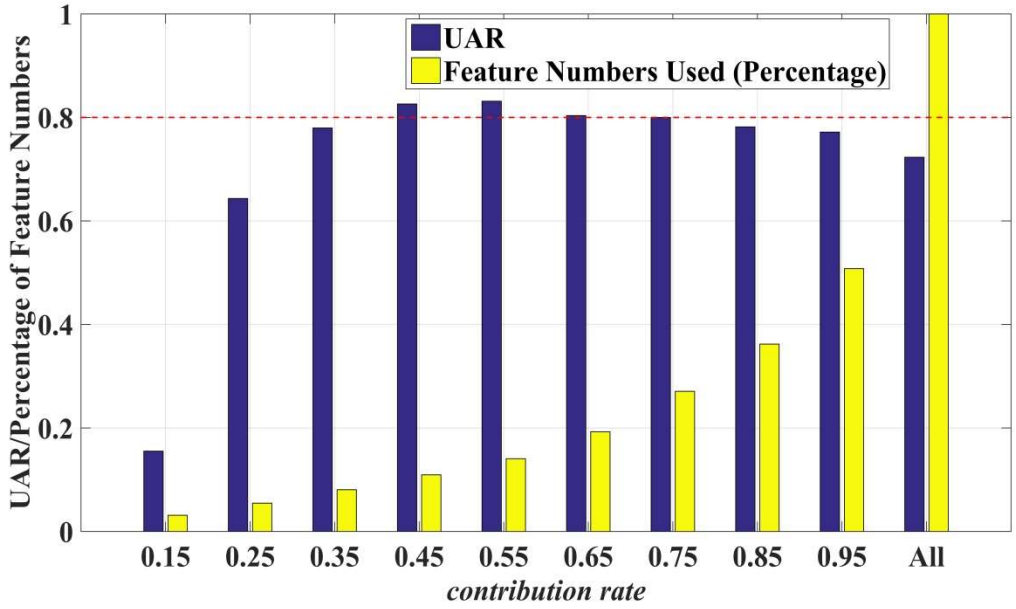
- A comparison with different classifiers for 54 species of birds classification

Classifiers	UAR %	Accuracy %
ELM	73.04	80.09
SVM	70.76	77.93
Ensemble	62.56	71.13
*k*NN	53.11	63.66

- **UAR** (Unweighted Average Recall): Calculated by the sum of recall values (class-wise accuracy) for all classes divided by the number of classes.
- **Accuracy**, i. e., WAR (Weighted Average Recall): Widely used, the correctly classified instance numbers divided by the total number of instances.

Experimental Results

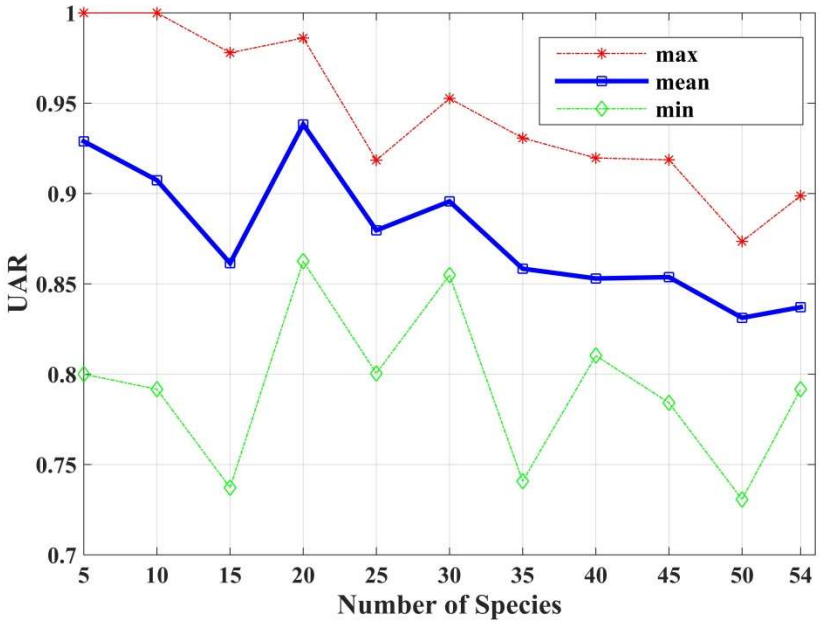
- Feature Selection



- Nearly 10% improvement of UAR, and with less than 15% features used.

Experimental Results

- Classification Results with Different Scales of Species



Species	UAR %	Accuracy %
10	90.74	94.71
20	93.82	93.91
30	89.56	89.56
40	85.30	89.03
50	83.12	85.60
54	83.71	86.57

- Excellent (species below 45), Good (species up to 54).

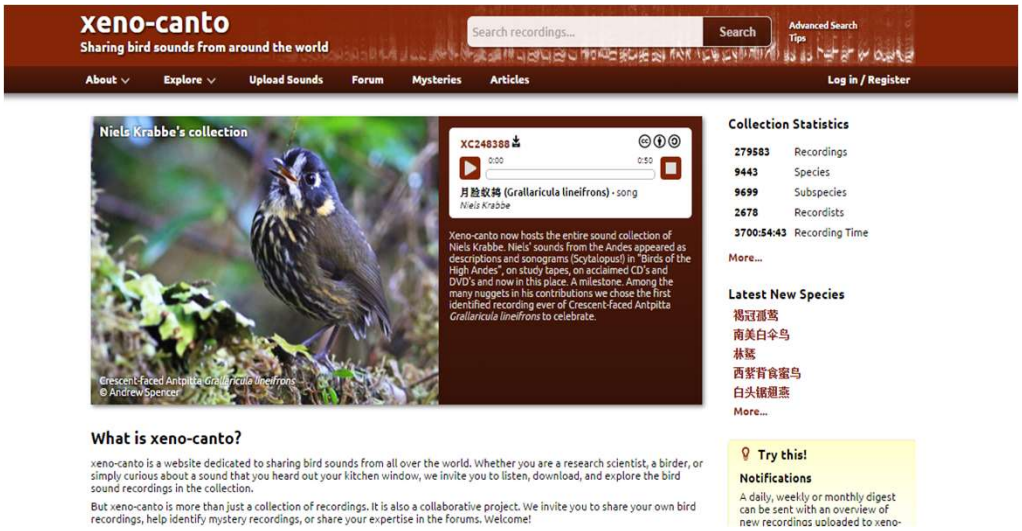
Conclusions

- The whole framework proposed is efficient and feasible.
- P-center based detector can be applied to the unsupervised syllables detection phase.
- openSMILE toolkit can be used in other areas beyond the speech emotion recognition.
- Feature selection is a necessary phase in the classification system.
- ELM-based classifier can be regarded as an efficient and robust model.

Future Works

- Large Database Needed:

Like the database collected by "Xeno-Canto", a website dedicated to sharing bird sounds from all over the world. (includes 279,583 recordings, 9,443 species of birds, more than 3,700 hours of recording time)



Note: this picture is coming from: http://www.xeno-canto.org/

Future Works

- REAL Large Scale Features Needed:

Our openSMILE toolkit can extract up to more than 6,000 dimensions of features for machine learning.

- Syllables Detection Methods:

Some other unsupervised techniques should be tested.

- Classifiers:

Deep Neural Networks (DNNs) or Advanced ELMs.

Thank you!