# Acoustic Identification of Bird Species Using Probabilistic Latent Component Analysis

**Emmanouil Benetos** EMMANOUIL.BENETOS.1@CITY.AC.UK
Department of Computer Science, City University London, London, UK.

## Abstract

This submission for the ICML 2013 Bird Challenge uses the Probabilistic Latent Component Analysis (PLCA) method for identifying bird species in continuous audio recordings. A birdsong dictionary is created using pre-extracted spectral templates from provided training set. Sparsity constraints are also enforced in the symmetric PLCA model in order to lead to more meaningful solutions.

## 1. Introduction

These working notes for the ICML 2013 Bird Challenge[1] present a submitted system for identifying bird species in continuous audio recordings using Probabilistic Latent Component Analysis (PLCA). Section 2 presents the PLCA method while section 3 presents the proposed bird identification system. Finally, possible model extensions are discussed in section 4.

## 2. PLCA

Probabilistic latent component analysis (PLCA) is a spectrogram factorization technique that was first proposed in (Smaragdis et al., 2006). It approximates an input spectrogram $V_{\omega,t}$ as a bivariate probability distribution $P(\omega, t)$, where $\omega$ is the frequency index and $t$ the time index, and attempts to factorize $P(\omega, t)$ as a series of spectral components and component activations. It is closely related to non-negative matrix factorization (NMF) (Lee & Seung, 1999), where PLCA can be viewed as a special case of NMF using the Kullback-Leibler cost function. However, contrary to NMF, PLCA provides a probabilistic framework that is extensible as well as easy to interpret. PLCA and related spectrogram factorization techniques have been used extensively in audio and image signal processing research, namely for source separation, multi-pitch detection, acoustic event detection, and action recognition.

The symmetric PLCA model can be formulated as:

$$V_{\omega,t} \approx P(\omega, t) = \sum_z P(z)P(\omega|z)P(t|z) \qquad (1)$$

where $P(\omega|z)$ are the spectral templates corresponding to component $z$, $P(t|z)$ are the time-varying component activations, and $P(z)$ is the prior probability for the components. For estimating $P(z)$, $P(\omega|z)$, and $P(t|z)$, iterative update rules are employed, which are derived from the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

## 3. Proposed Method

### 3.1. Time-frequency Representation

As a time-frequency representation, the constant-Q transform (CQT) with a spectral resolution of 60 bins/octave is used (Schörkhuber & Klapuri, 2010). The lowest frequency bin is at 330Hz and the highest bin is at 12.5kHz, while the time step is 10ms. Afterwards, a simple noise suppression procedure is applied to the log-frequency spectrogram $V_{\omega,t}$ using a $\frac{1}{3}$-octave span median filter.

### 3.2. Extracting Spectral Templates

For each species recording in the training set, a dictionary of 20 atoms is extracted using PLCA ($z = 1, \cdots, 20$). The resulting spectral templates $P(\omega|z)$ are stacked together for all species, resulting in a matrix of dimensions $\Omega \times 700$ (where $\Omega = 295$ is the number of log-frequency bins and $700 = 20 \cdot 35$, with 35 being the number of bird species in the challenge). An example of extracted templates is given in Fig. 1.

[1] http://sabiod.univ-tln.fr/icml2013/

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

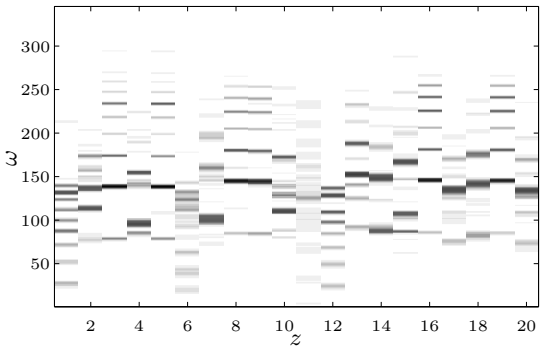


*Figure 1.* Pre-extracted spectral templates for the *Branta canadensis* species.

### 3.3. Bird Species Identification

For each recording from the test set, its normalized log-frequency spectrogram $V_{\omega,t}$ is fed into the model of (1). This time, the number of components $Z = 700$ and the update rules are only applied to $P(z)$ (which represents the mixture probability of the spectral components) and $P(t|z)$ (which represents the activation of each component over time), while $P(\omega|z)$ are the pre-extracted templates, which remain fixed. 70 iterations are used per test recording for estimating the unknown parameters.

Since the proposed model is overcomplete (it contains more information than in the input), it can converge to non-meaningful solutions. To that end, sparsity is enforced using the entropic prior proposed in (Smaragdis, 2009). In specific, sparsity constraints are applied to $P(t|z)$ (implying that only few dictionary components are active at a given time frame) and to $P(z)$ (implying that only few components should be present in the whole recording). The sparsity constraint in $P(z)$ also implies that only few bird species should be present in the recording.

Finally, the output of the PLCA model is $P(z)$, which is used to compute the probability of a bird species being present in a test recording:

$$P(bird_C) = \sum_{j \in C} P(z_j) \tag{2}$$

where $bird_C$ denotes a bird class and $C$ the set of components that belong to that class.

## 4. Model Extensions

The proposed model is fairly simple yet so far has reached solid results (with AUC=0.625), surpassing the baseline system by more than 9%. Its biggest drawback is the lack of any temporal modeling, which however can be supported by PLCA-based methods. One such example is Shift-invariant PLCA (Smaragdis & Raj, 2007) which supports time-frequency patches instead of spectral templates. Another option would be to add temporal constraints to the one-dimensional spectral templates, using the Non-negative Hidden Markov Model (Mysore, 2010), which combines PLCA with Hidden Markov Models. Finally, the constant number of basis can become variable, e.g. by performing segmentation on the training data and extracting one basis per segment.

## Acknowledgments

E.B. is supported by a City University London Research Fellowship.

## References

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

Lee, D. and Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401: 788–791, October 1999.

Mysore, G. *A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures*. PhD thesis, Stanford University, USA, June 2010.

Schörkhuber, C. and Klapuri, A. Constant-Q transform toolbox for music processing. In *7th Sound and Music Computing Conf.*, Barcelona, Spain, July 2010.

Smaragdis, P. Relative-pitch tracking of multiple arbitary sounds. *Journal of the Acoustical Society of America*, 125(5):3406–3413, May 2009.

Smaragdis, P. and Raj, B. Shift-invariant probabilistic latent component analysis. Technical report, Mitsubishi Electric Research Laboratories, December 2007. TR2007-009.

Smaragdis, P., Raj, B., and Shashanka, Ma. A probabilistic latent variable model for acoustic modeling. In *Neural Information Processing Systems Workshop*, Whistler, Canada, December 2006.