Bird Audio Detection using probability sequence kernels

Anshul Thakur, Jyothi Jain, Padmanabhan Rajan, A.D. Dileep

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi
E-mail: anshul_thakur@students.iitmandi.ac.in, padman@iitmandi.ac.in

February 3, 2017

The 2016 BAD challenge requires to determine if a given audio signal has a bird sound in it. The participants are provided with two labeled datasets namely Warblr and Freefield having 6045 and 1935 recordings. The evaluation dataset contains 8620 recordings from Chernobyl and Warblr datasets. Each recording is 10 seconds long. Our submission to the challenge focuses on minimizing the pre-processing effort and deals with audio recordings as they are provided.

Mel-frequency cepstral coefficients (MFCCs) are used as feature representations. Delta and delta-delta coefficients are appended to the 13 dimensional MFCCs to represent each audio frame by a 39 dimensional feature vector. Each feature is normalized to have zero-mean and unit variance. Channel effects due to recording devices or recording environment and additive noises lead to change in mean and variance of features from an audio recording respectively. The features become robust to channel effects and additive noise after mapping them to an ideal distribution like the standard normal distribution [1]. The distribution of each feature is made normal using short-term Gaussianization [1]. First, the features are decorrelated using a linear transform and then the features are warped by applying short-time windowed cumulative distribution function (CDF) matching so that their distribution becomes normal.

Dynamic kernel based support vector machines are used to handle variable length patterns like feature representations of speech or audio recordings. The probability sequence kernel (PSK), one such dynamic kernel is used in our framework. It has been used earlier for bird species identification task [2]. The PSK uses UBM/GMM to map a set of feature vectors to a higher dimensional fixed length vector [3]. The concatenation of responsibility terms of mixtures of UBM and adapted UBM (class-specific GMM) forms this mapped fixed length vector known as probability alignment vector.

Let $X = \{x_1, x_2, x_3, ..., x_T\}$ be a set of feature vectors. If UBM has Q components, then responsibility terms vector is of size $2Q$ and can be represented as, $\Psi(x) = [\gamma_1(x), \gamma_2(x), ...\gamma_{2Q}(x)]$. The set, X, of feature vectors is represented as a fixed length vector $\Phi_{\text{PSK}}(X)$, defined as

$$\Phi_{\text{PSK}}(X) = \frac{1}{T}\sum_{t=1}^{T}\Psi(x_t) \tag{1}$$

The $\Phi_{\text{PSK}}(X)$ is also of dimensions, $D = 2Q$. The PSK between sets of feature vectors, X_m and X_n can be calculated using equation 2.

$$K_{\text{PSK}}(X_m, X_n) = \Phi_{\text{PSK}}(X_m)^T S^{-1} \Phi_{\text{PSK}}(X_n) \tag{2}$$

S is a correlation matrix and is defined as:

$$S = \frac{1}{M} R^T R \tag{3}$$

where R is a $M \times D$ matrix whose rows are probabilistic aligned vectors for all the training feature vectors.

In our framework, we have modified the use of PSK as mentioned above. Instead of using UBM and class-specific GMMs, we have only used a GMM (having Q mixtures) built using the features of recordings labeled as bird sounds. The probability alignment vectors for both the bird and non-bird classes are calculated using this GMM. Hence $\Phi_{PSK}(X)$ is of Q dimensions. This has led to the decrease in dimensionality of $\Phi_{\text{PSK}}(X)$ vectors. Given that the GMM is made from recordings labeled as bird, the probabilistic alignment vectors for bird and non-bird class will be different. This makes it possible to distinguish the two classes. Since the recordings labeled as bird sounds may also contains non-bird sounds, the distinction between probabilistic alignment vectors of two classes may decrease. However, we also experimented with UBM/GMM framework as used in [3] instead of GMM. No significant difference in performance was observed.

The computation complexity to calculate $\Phi_{\text{PSK}}(X)$ for any test example is $\mathcal{O}(N \times Q)$ where N is number of feature vectors and Q is the number of mixtures in GMM. The computation complexity to calculate K_{PSK} is $\mathcal{O}(N^3)$.

GMM used in our framework has 128 mixtures having diagonal covariance matrices and is built using 100 randomly chosen audio recordings from Warblr database. Also, the PSK kernel gram matrix is built using 200 randomly chosen recordings from both bird and non-bird classes. The libsvm [4] is used for SVM implementation and MFCCs are extracted using voicebox [5]. In our first submission, we didn't use short term Gaussianization and got an AUC score of almost 73% whereas after applying short-term Gaussianization in our second submission, we got an AUC score of almost 77%. This indicates the significance of feature warping using short-term Gaussianization in the proposed framework.

References

[1] B. Xiang, U. Chaudhari, J. Navrátil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proc. Int. Conf. Acoust. Speech, Signal Process*, 2002.

[2] D. Chakraborty, P. Mukker, P. Rajan, and A. Dileep, "Bird call identification using dynamic kernel based support vector machines and deep neural networks," in *Proc. Int. Conf. Mach. Learn. App.*, 2016.

[3] K. Lee, C. You, H. Li, and T. Kinnunen, "A gmm-based probabilistic sequence kernel for speaker verification." in *INTERSPEECH*, 2007.

[4] "Libsvm," https://www.csie.ntu.edu.tw/ cjlin/libsvm/, accessed: 2017-01-10.

[5] "Voicebox," www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html/, accessed: 2017-01-10.