A Comparison Study to Identify Birds Species Based on Bird Song Signals

Xin Guo[1], Qing-Zhong Liu[2]

[1]Department of Computer Science, Sam Houston State University, Huntsville, TX 77382, USA
[2]Department of Computer Science, Sam Houston State University, Huntsville, TX 77382, USA
[1] xxg004@shsu.edu, [2]liu@shsu.edu

Abstract: this paper presents a comparison study in automatically identifying bird species based on bird acoustic signals, using audio files from XENO-CANTO online database. The features including Mel-Frequency Cepstral Coefficients (MFCC), geo-related meta-features, and the integration are compared. The learning classifiers Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), k-Nearest Neighbor (kNN), and Ensemble Learning are examined. Our experimental results show that in the comparison study, ensemble learning using discriminant learner with the integration of MFCC features and geo meta-features obtains the best detection performance.

1. Introduction

Birds live worldwide and rank as the class of tetrapod with the most living species, at approximately ten thousands, in which more than half identified species being passerines, known as perching birds or songbirds. Bird vocalizations are traditionally divided into bird calls and bird songs. The distinction between songs and calls is based on complexity, length, and context and has a lot of exceptions. Songs are longer and more complex, mainly produced by males and associated with courtship and mating during the breeding season, while calls tend to be shorter, simpler and produced by both sexes throughout the year; serve such functions as alarms or keeping members of a flock in contact [1].

Bird watching is a traditional and popular activity focused on observation of birds. Due to fact that many birds are more easily heard than photographed, it is promising to rely on their sounds as a convenient and reliable method for species identification. With the rapid development of digital technology, portable devices such as mobile phones are equipped with outdoor recording functionality, adequate storage capacity, and computational power to do onsite recording analysis. It is easier than ever for bird watcher to record bird sounds during bird watching. On the other hand, professionals like ornithologist, ecologists, traditionally take advantage of long-term semi-automatic acoustical monitoring without human presence at recording site for scientific research or ecosystem evaluation purpose [2].

The motivation of bird species identification goes beyond bird watching. Bird is a good indicator of the state of their surrounding ecosystem; since they are widely distributed and react quickly to changes in environmental conditions such as habitat loss, declining biodiversity, and climate change. Acoustical monitoring for tracking bird migration and for estimating populations of bird species provides information to understand and evaluate the changes in environment. Recognition by bird sounds also would be a powerful tool for automatic identifying bird in cases such as in areas near airports to prevent collisions with aircraft [2].

Automatic bird species identification based on birdsongs is a recent application of machine learning, essentially of pattern recognition and classification. The challenge can be broken down into two main stages. First, each of the bird sounds recording should be analyzed, normally by signal processing tools, to produce a discriminative feature set that represent the original bird audio signal regarding its species. Techniques widely used in feature extraction include temporal and spectral measurements, Linear Predictive Coding, Mel Frequency Cepstral Coefficients (MFCC). Then these feature sets serve as input to a classification system. Several algorithms have been employed, including probabilistic and instance based classifier, neural networks and support vector machines [3].

Linear prediction coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal, using the information of a linear predictive model [4]. It is one of the most powerful speech analysis techniques. The basic idea behind this model is that a speech sample can be approximated with a linear combination of previous speech samples. LPC analysis tries to determine the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense.

The cepstral coefficients are the results of taking the inverse Fourier transform representation of the logarithmic magnitude spectrum of a signal. LPC derived cepstral coefficients (LPCCs), is a very effective representation for speech coding, analysis, synthesis, and recognition. A significant property of the LPC spectral modeling is that the LPC spectrum matches the signal spectrum closely near the spectral peaks. So the linear prediction cepstral coefficients (LPCCs) are more robust and reliable features for speech recognition and have been proven to be more relevant than LPCs [5, 8].

MFCC is a frame-level feature. It is computed by transforming the spectrum of a frame into the Mel scale, which approximates the human auditory system's response more closely than the linear frequency. The steps that are applied in the traditional method to get the coefficients are started with reemphasizing the sampled signal and then applying the framing and windowing on it, then taking the Fast Fourier Transform (FFT) for each windowed frame, the signal now is a power spectrum, this signal enters to a Mel filter bank and the length of the output is equal to the number of filters created, after taking a discrete cosine transform to the log of the filter bank's output, an array of features that describe the spectral shape of the signal [6,7,18].

The earlier works for birdsongs identification focused on template matching usually used in conjunction with dynamic time warping (DTW). First one manually obtains a collection of templates including intervals with no bird sound, and then slides them across a target spectrogram. DTW algorithm is used to stretches either the template or the target spectrogram to calculate some measure of similarity. This can be viewed as simultaneous segmentation and classification [9].

Lopes et al. [3] presented a comparison of the performance of 3 feature sets originally implemented for music analysis combined with a series of machine learning algorithms applied to the bird species identification problem. Experiments were conducted in order to evaluate various combinations of feature sets and classifiers in a database composed by 101 audio records from 3 bird species. Somervuo et al. [10] found that MFCC outperforms sinusoidal features and a collection of spectral features such as spectral centroid, bandwidth, roll-off, flux, etc. Lee etc. designed image shape features to identify bird species based on the recognition of fixed-duration birdsong segments where their corresponding spectrograms are viewed as gray-level images [11]. Chou and Liu [12] applied a wavelet transformation to transform sections of the bird songs. Then the first five order MFCCs are computed, and same order MFCC are aligned. Neal et al. [13] proposed a supervised time-frequency audio segmentation method using a Random Forest classifier. Springer [14] addressed the issue when multiple species of birds sing concurrently in the same recording. Zhao et al [15] designed acoustic environment signatures that can be used for background noise recognition.

In this paper, we make a comparison study by comparing MFCC features and geo-metadata features with the use of several machine learning classifiers.

2. Methodology

2.1 Data Set

XENO-CANTO is a collaborative database containing more than 192k audio records that cover 9120 bird species observed all around the world by more than 2000 contributors at the time of writing and these numbers keep growing during each day. Recordings in the dataset are not consisting only bird song. A substantial part of the elements contain background noise such as sounds from other animals, wind and machine noise, or electric hum, which is near to the real world applications. Contributors also provide the metadata of each audio file including geographical information, date and /or time of the day and/or presence of background species.

The lifeclef2014 Bird Identification Task is based on a subset of XENO-CANTO database. It contains 14027 audio recordings belonging to the 501 bird species in the area of South American centered on Brazil [16]. Additional information includes the audio file associated metadata in XML format.

2.2 Feature Extraction

The dataset is sourced from a large online collection of user-submitted recordings, and therefore suffers from inconsistent audio quality. The recordings have varying audio quality due to atmospheric conditions like wind and rain, interfering bird and insects calls, quality of the recording equipment, and varying professionalism of the recordist. In order to avoid bias in the evaluation related to recording devices, the organizers preprocessed the whole audio data to normalize frequency sample to 44.1.kHz wav format. A temporal signal is first transformed into a serie of frames where each frame consists in 16 mfcc (Mel-filter cepstral coefficients), including energy (first coefficient). Each frame represents a duration of 11.6 ms (e.g 512 temporal bins of a signal sampled at 44 100 Hz). Two successive frames overlap of 33% i.e. 3.9 ms.

The mel scale is a means of mapping the physical frequency to the perceptual representation. The mapping between the physical frequency scale (Hz) and perceptual frequency scale (mel) is approximately linear below 1000 Hz and logarithmic at higher frequencies. The relation between the physical frequency scale and the mel frequency scale can be described as:

The mel scale maps the physical frequency to the perceptual representation. The mapping between the physphysicalquency scale (Hz) and perceptual frequency scale (mel) is approximately linear below 1000 Hz and logarithmic at higher frequencies. The mapping between mel frequency scale and physical frequency scale can be described as:

$$mel = 2595\log\left(1+\frac{f}{700}\right) \tag{1}$$

Where f is the spectral frequency of the input bioacoustics signal using short-time Fourier transform.

In the process of MFCCs feature extraction, the Fourier spectrum is filtered by a set of mel-scale filters. The MFCCs are computed by performing DCT on the logarithmic energy output by every bandpass filter

$$c_m = \sum_{k=0}^{K-1} \log(E_k)\cos\left(m\frac{\pi}{K}(k+0.5)\right) \tag{2}$$

$$0 \le m \le L-1 .$$

Where K is the number of bandpass filters, L is the desired length of MFCCs, and Ek is the energy of the output of the k-th bandpass filter. L is set to 16 according to the study in the reference [17].

Let take a particular MFCC features, denoted by the matrix C= $\{c_{m,t}\}$ (m=0, 1, ..15; t=1,2,…n) wherein n is the number of frames. The Following 16 MFCC features are extraction based on equation (3),

$$\bar{c}_m = \frac{\sum_{t=1}^{n}|c_{m,t}|}{n} \tag{3}$$

Additionally, we retrieve the following 16 MFCC features regarding the first derivative based on equation (4),

$$\bar{c}_m^1 = \frac{\sum_{t=1}^{n-1}|c_{m,t+1} - c_{m,t}|}{n-1} \tag{4}$$

And the following 16 MFCC features are designed based equation (5)

$$\bar{c}_m^2 = \frac{\sum_{t=1}^{n-2}|c_{m,t+2} + c_{m,t} - 2 * c_{m,t+1}|}{n-2} \tag{5}$$

Based on equations (3), (4) and (5), a total of 48 MFCC features are extracted.

Additionally, the three geo-meta features Latitude, Longitude and Elevation were extracted from XML file.

3. Experiments

We select the learning classifiers including Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) with linear kernel, and quadratic kernel, k-Nearest Neighbor (kNN), and Ensemble Learning [19] using discriminant analysis and kNN learners respectively for our comparison study. The following section discusses results obtained using the subset of 2014 BirdCLEF dataset when trying to find which classifier is scalable as bird species number increases, all results are obtained using 5-fold cross validation.

Table 1 lists the detection accuracy by applying the six classifiers to the three feature sets, 48-dimensional MFCC, 3-dimensional META, and the integration of MFCC and META on four different numbers of bird species 5, 10, 50 and 100 bird species (with 275, 494, 1892, or 3263 instances respectively). The ROC curves and confusion matrix are provided in Figures 1 to 5 by using the ensemble learning with the subspace discriminant classifier.

The accuracies of all 6 classifiers decrease dramatically as number of bird species increase from 5, 10, 50, to 100. SVM and Ensemble learning with Discriminant learner for random subspace have comparable highest accuracy but in our experiments the training time of SVM is much longer than ensemble learning. The integration of MFCC with geo-meta features obtains the best detection accuracy.

The experimental results show that ensemble learning with discriminant has the best performance. To compare the detection performance under different parameter of subspace dimensionalities, we adjust the subspace parameter from the default value 26 to 30, 40, and 50, respectively. Table 2 shows the detection accuracy under different subspace parameters with discriminant ensemble learning. Figure 6 shows the ROC curve in detecting all 501 species birds by using MFCC and geo-meta features together.

4. Conclusions

In this study, we conducted a comparison study identifying bird species based on bird songs. Our study shows that the integration of MFCC features and geo-meta features obtains the best detection accuracy with the use of discriminant ensemble learning in our comparison study.

Our future study includes develop more reasonable features and adopt deep learning techniques to identify the performance under different input parameters. The optimization of the parameters will be explored too.

Table 1. Validation Accuracy (%) On Different Species With Different Learning Classifiers

	MFCC Features				Geo-Meta Features				MFCC+Geo-Meta Features			
Number of species	*5*	*10*	*50*	*100*	*5*	*10*	*50*	*100*	*5*	*10*	*50*	*100*
Linear Discriminant	46.2	31.8	15.3	10.5	34.2	22.3	6.2	3.6	44.4	33.6	16.8	11.9
Linear SVM	52.4	39.7	18.8	13.3	43.6	25.5	8.6	4.7	56	43.1	22.1	15.2
Quadratic SVM	44.4	39.9	19.5	14	45.1	28.7	9.7	7	52.4	42.5	21.9	15.7
Medium kNN	42.5	30.4	11.2	8.7	44.7	27.9	10	7.3	41.8	28.7	12.4	8.9
Ensemble: Subspace Discriminant	56.7	43.1	23.3	16.4	33.5	21.1	6	3.6	59.6	46.2	26.3	18.8
Ensemble: Subspace kNN	42.9	35	15.5	12	36.4	25.5	10.6	7.4	52.7	36.2	18	14.6

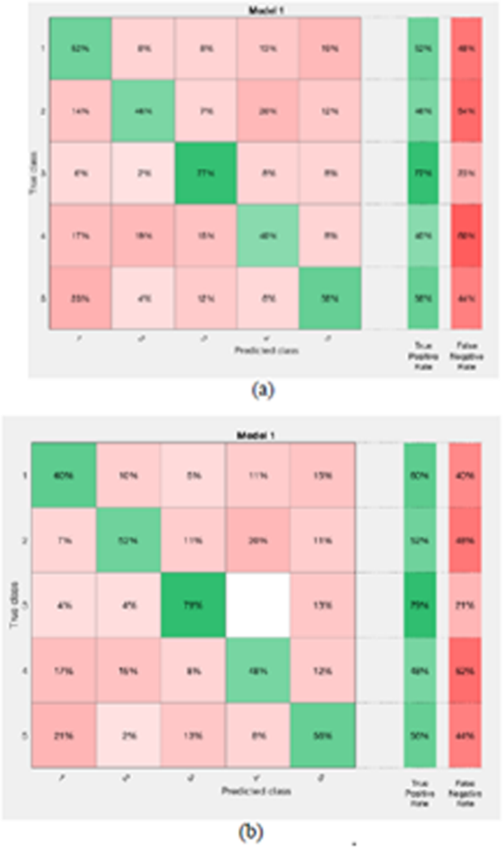


Figure 1: Confusion matrix of 5 species subsets with Ensemble: Subspace Discriminant classifiers using MFCC features only (a) and using MFCC and META features (b)

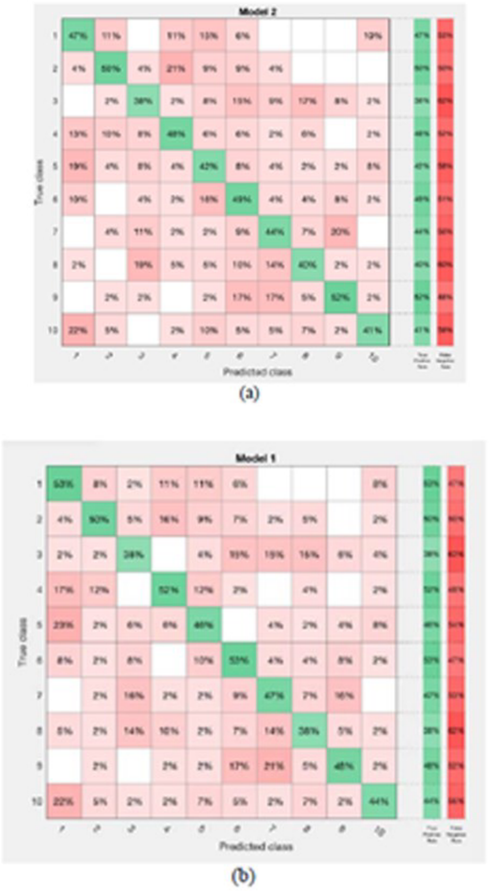
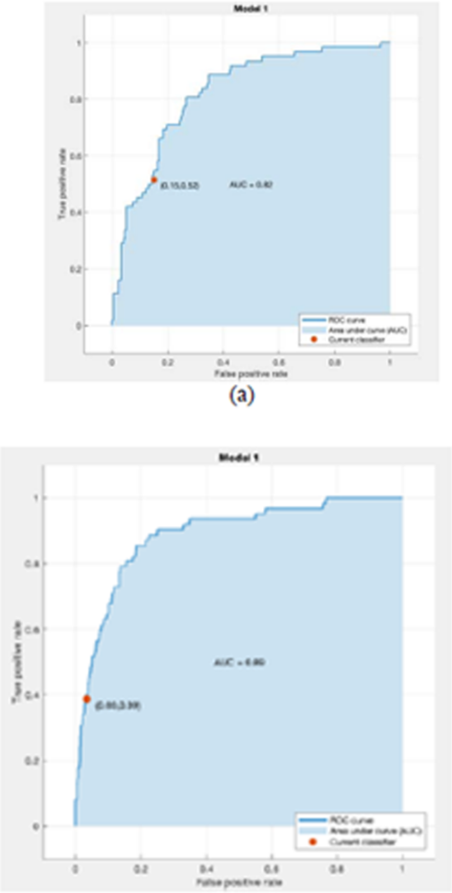


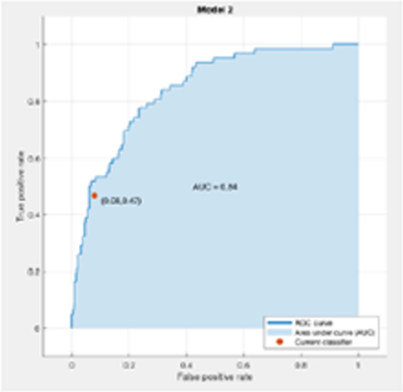
Figure 2: Confusion matrix of 10 species subsets with Ensemble: Subspace Discriminant classifiers using MFCC features only (a) and using MFCC and META features (b)



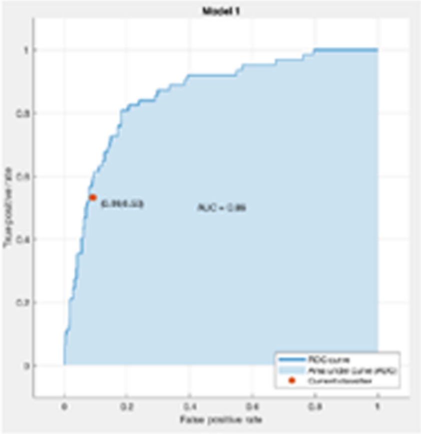
(a)

(b)

Figure 3: ROC curve of 5 species subsets with Ensemble: Subspace Discriminant classifiers using MFCC features only (a) and using MFCC and META features (b)

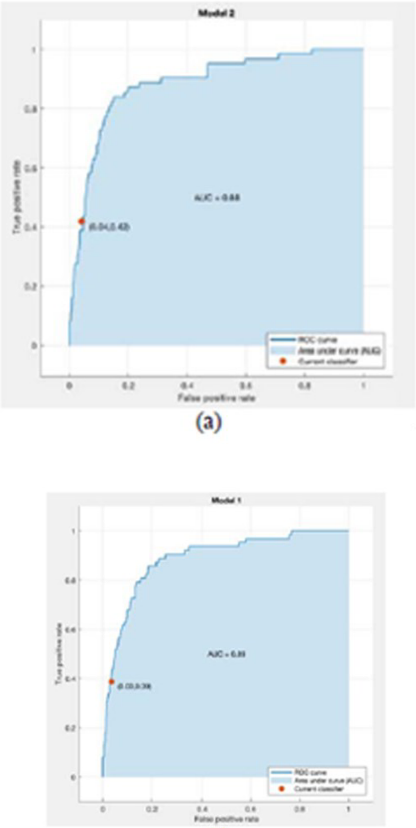


(a)



(b)

Figure 4: ROC curve of 10 species subsets with Ensemble: Subspace Discriminant classifiers using MFCC features only (a) and using MFCC and META features (b)



(a)

(b)

Figure 5: ROC curve of 50 species subsets with Ensemble: Subspace Discriminant classifiers using MFCC features only (a) and using MFCC and META features (b)

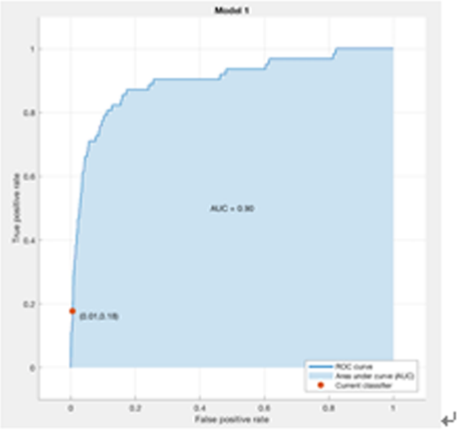


Figure 6: ROC curve of 501 species subsets by applying Ensemble: Subspace Discriminant classifier to MFCC and META

Table 2. Validation Accuracy (%) On Different Species By Discriminant Ensemble

Subspace dimension	Number of species			
	5	10	50	100
26	58.9	42.7	24.8	19.4
30	60	43.7	26.4	20.0
40	61.1	44.3	27.0	21.4
50	60.7	44.1	27.3	21.8

Acknowledgements

The support for this study from the college of science and engineering technology of the sam houston state university (shsu), and from the shsu computer science department is highly appreciated.

References

[1] http://web.stanford.edu/group/stanfordbirds/SUFRAME.html, accessed on April 25, 2017.

[2] M. Lopes, L. Gioppo, T. Higushi, C. Kaestner, C. Silla, A. Koerich, "Automatic bird species identification for large number of species," 2011 IEEE International Symposium on Multimedia (ISM), pp.117-122, 5-7 Dec. 2011.

[3] M. Lopes, A. Lameiras Koerich, C. Nascimento Silla, C. Alves Kaestner, "Feature set comparison for automatic bird species identification," in 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp.965-970, 9-12 Oct. 2011

[4] L. Deng and D. O'Shaughnessy. Speech processing: a dynamic and optimization-oriented approach. Marcel Dekker. pp. 41–48. ISBN 0-8247-4040-8, 2003.

[5] Speech Technology: A Practical Introduction. Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis,
http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

[6] C. Chou and H. Ko, “Automatic Birdsong Recognition with MFCC Based Syllable Feature Extraction”. In: Hsu CH., Yang L.T., Ma J., Zhu C. (eds) Ubiquitous Intelligence and Computing. UIC 2011. Lecture Notes in Computer Science, vol 6905. Springer, Berlin, Heidelberg.

[7] Practical Introduction to Frequency-Domain Analysis. MathWorks Online Documentation of Signal Processing Toolbox. Online Web Resources.

[8] C. Lee, Y. Lee, and R. Huang. "Automatic recognition of bird songs using cepstral coefficients." Journal of Information Technology and Applications 1,1, 17-23, 2006.

[9] S. E. Anderson, A. S. Dave, and D. Margoliash. Template-based automatic recognition of birdsong syllables from continuous recordings. J. Acoust. Soc. Am., 100(2):1209–1219, 1996. ISSN 0001-4966.

[10] P. Somervuo, A. Harma, S. Fagerlund. "Parametric representations of bird sounds for automatic species recognition", IEEE Transactions on Audio, Speech, and Language Processing 14,6: 2252-2263, 2006.

[11] C. Lee, S. Hsu, J. Shih and C. Chou, "Continuous birdsong recognition using Gaussian mixture modeling of image shape features." IEEE Transactions on Multimedia,15, 2: 454-464, 2013.

[12] C-H. Chou and P-H. Liu, “Bird Species Recognition by Wavelet Transformation of a Section of Birdsong”, Symp. and Workshop Ubiq., Auton. Trusted Comput., Brisbane, Australia, pp.189–193, July 2009.

[13] L. Neal, F. Briggs, R. Raich and X. Fern. "Time-frequency segmentation of bird song in noisy acoustic environments. " 2011 IEEE International Conference on. Acoustics, Speech and Signal Processing (ICASSP), 2011. DOI: 10.1109/ICASSP.2011.5946906

[14] J. Springer, Z. Duan, and B. Pardo. "Approaches to multiple concurrent species bird song recognition." The 2nd International Workshop on Machine Listening in Multisource Environments, 2013. http://www.ece.rochester.edu/~zduan/resource/SpringerEtal_BirdSongRecognition_ChiME13.pdf

[15] H. Zhao, and H. Malik. "Audio recording location identification using acoustic environment signature." IEEE Transactions on Information Forensics and Security 8, 11 (2013): 1746-1759.

[16] H. Goeau, H. Glotin, W. Vellinga, R. Planque, A. Rauber and A. Joly. LifeCLEF Bird Identification Task 2014. https://hal.inria.fr/hal-01088829/file/CLEF2014wn-Life-GoeauEt2014a.pdf

[17] O. Dufour, T. Artieres, H. Glotin and P. Giraudet (2014). “Clusterized mel filter cepstral coefficients and support vector machines for bird song identification”, In . Proc. 1st International Workshop on Machine Learning for Bioacoustics, joint to The 30th International Conference on Machine Learning (ICML 2013) Atlanta, USA, June, 2013, pages 89-93.

[18] D. P. W. Ellis. PLP and RASTA (and MFCC, and inversion) in Matlab, 2005. Online web resource.

[19] R. Polikar, "Ensemble learning." Ensemble machine learning. Springer US, 2012. 1-34.