AUDITORY SCENE CLASSIFICATION USING ENSEMBLE LEARNING WITH SMALL AUDIO FEATURE SPACE

Technical Report

Tomasz Maka

Faculty of Computer Science and Information Technology,
West Pomeranian University of Technology, Szczecin
Zolnierska 49, 71-210 Szczecin, Poland
tmaka@wi.zut.edu.pl

ABSTRACT

The report presents the results of an analysis of audio feature space for auditory scene classification. The final small feature set was determined by the selection of the attributes from various representations. Feature importance was calculated exploiting the Gradient Boosting Machine. A number of classifiers were employed to build the ensemble classification scheme, and majority voting was performed to obtain the final decision. In the result, the proposed solution uses 223 attributes and outperforms the baseline system by over 6 per cent.

Index Terms— audio features, auditory scene analysis, ensemble learning, majority voting

1. INTRODUCTION

An essential element in the audio classification system is the discriminatory features. Used feature space determines the complexity of the model used in classification stage. The features can be selected in supervised or unsupervised manner. In the first case, the attributes are designed using domain knowledge, whereas in the other, they are generated automatically. Nowadays, the vast majority of audio classification systems is built on top of deep neural networks. Especially in convolutional neural networks (CNN), the features are generated in the unsupervised learning process by convolutional layers. The CNN systems use audio data converted to spectrograms or melspectrograms as input data, therefore to create features with predictive power, a large dataset is required.

The size of the feature space is a crucial part of the model, defines the resources required for storing and transmission and may have a direct influence on improving the performance and accuracy of the system. Therefore, searching for compact and yet discriminative feature sets is still under research [1, 2, 3].

In this work, an analysis of various feature sets was performed to determine a representation of source audio signal with the highest predictivity power.

2. FRAMEWORK

The proposed framework consists of two parts – parameterisation and classification. The feature space is created by the first stage whose structure is shown in Fig. 1. The resulting feature vector is composed of 13 subsets with discriminative attributes selected in the feature importance analysis process. From time-frequency representations like spectrogram, melspectrogram and cochleagram the attributes were selected using Gradient Boosting Machine [4] using whole development dataset. The final individual sets can be briefly summarised as follows:

Binaural unit – interaural time difference, interaural intensity difference, interaural coherence, and azimuth.

Pitch properties – statistical properties of pitch contour.

Onset map – properties of onsets detected in all channels of cochleagram.

Binary map – attributes of the binary map obtained by thresholding channels of cochleagram.

Channel dependencies – energy differences between neighbouring channels of cochleagram.

Dominant bands – a selected number of bands with the highest energies in cochleagram.

Channels sparsity – Hoyer sparsity [5] computed for the individual channels of cochleagram.

Sub-band energies – energies calculated in 8 equally sized ranges of cochleagram, melspectrogram and spectrogram.

Spectrogram activations – attributes of activation matrix by computing the non-negative matrix factorisation of a spectrogram.

Melspectrogram activations – properties of activation matrix by computing the non-negative matrix factorisation of a melspectrogram.

ΔBIC trajectory – attributes of trajectory calculated as a difference between Bayesian Information Criterion (BIC) values of models used in audio segmentation [6].

Temporal envelope – properties of temporal envelope [7].

Histograms of feature contours – characteristic of histograms obtained for various [8] low-level feature contours.

Classification procedure employs an ensemble learning. In the first phase, accuracies were estimated for 62 individual classifiers. Then, a set of the best classifiers (with the accuracies higher than 50 per cent) were formed, and ensemble learning with majority/hard voting was executed. In the next steps, consecutive classifier combinations were removed from the set to maximise the accuracy. Finally, eight classifiers were selected for majority/hard voting scheme as depicted in Tab. 1.

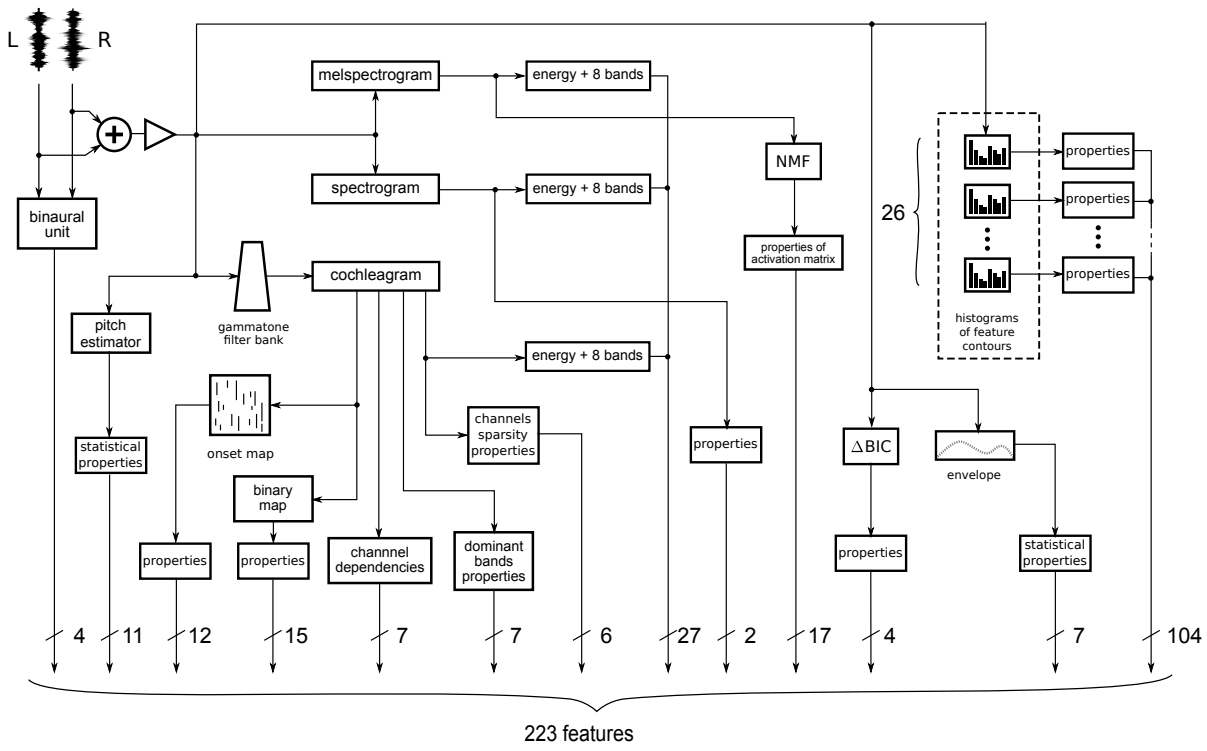

Figure 1: The diagram of the proposed system converting an audio signal to feature space.

Table 1: The set of classifiers used in the majority voting scheme.

Classifier	Description
C_1	Linear Discriminant Analysis
C_2	Quadratic Discriminant Analysis
C_3	Random Forest classifier with 10 trees using Gini impurity as splitting metric.
C_4	Random Forest classifier with 100 trees using Gini impurity as splitting metric.
C_5	Random Forest classifier with 100 trees using entropy to compute information gain.
C_6	Multi-layer perceptron classifier. It uses 3 hidden layers with 30 hidden units each.
C_7	K-nearest neighbors classifier with K=20.
C_8	Bagging classifier with 500 linear support vector classification estimators.

3. EVALUATION

The system performance was evaluated on the development dataset of DCASE'2018 competition (Task 1). The audio data consists of binaural, 10-seconds recordings from 10 acoustic scenes captured in six European cities.

The classification experiments were performed using the proposed framework, and the confusion matrix is presented in Tab.2. The best result was obtained for 'Shopping mall' (89.9%) and the worst for 'Public square' (38.9%) with overall system performance equal to 66.2%. According to the confusion matrix, analogies can be noticed between classes with sound sources sharing similar physical properties. For example, such a situation is visible for classes 'Bus', 'Metro' and 'Train'.

The comparison of our system with the baseline is shown in Tab. 3, where in case of classes 'Airport' and 'Public square' no accuracy improvement occurred.

Table 2: Confusion matrix for the proposed system evaluated on the development dataset.

	Airport	Bus	Metro	Metro station	Park	Public square	Shopping mall	Street, pedestrian	Street, traffic	Tram
Airport	**47.2**		2.6	4.9	0.4	8.3	18.1	18.5		
Bus		**63.6**	5.4		1.2	0.4				29.3
Metro		5.4	**60.5**	9.2		1.1		0.4	3.4	19.9
Metro station	6.2	1.2	7.7	**55.6**	1.2	2.3	3.9	10.7	4.6	6.6
Park			0.4	2.1	**88.8**	3.7		3.3	0.9	0.8
Public square	0.5	1.9	1.3	3.7	16.7	**38.9**	1.9	19.0	14.7	1.4
Shopping mall	5.0			1.1		0.7	**89.6**	3.6		
Street, pedestrian	4.0			0.9	3.2	20.6	6.5	**57.5**	5.7	1.6
Street, traffic			0.4	1.6		5.3		6.5	**86.2**	
Tram	1.1	9.2	11.9	1.1			1.9	0.8		**73.9**

Table 3: The class-wise accuracy compared with the baseline.

Scene class	Accuracy	
	Baseline	**Proposed**
Airport	[72.9] %	47.2 %
Bus	62.9 %	[63.6] %
Metro	51.2 %	[60.5] %
Metro station	55.4 %	[55.6] %
Park	79.1 %	[88.8] %
Public square	[40.4] %	38.9 %
Shopping mall	49.6 %	[89.6] %
Street, pedestrian	50.0 %	[57.5] %
Street, traffic	80.5 %	[86.2] %
Tram	55.1 %	[73.9] %
Average	**59.7 %** (+/- 0.7)	**66.2 %**

4. CONCLUSION

The proposed framework for auditory scene recognition was built with small feature space which includes temporal-frequency properties and uses a majority voting ensemble classification scheme. The final performance is relatively similar to human hearing performance for development dataset. Despite lower performance than a typical CNN-based system, the proposed feature set can be used to design more sophisticated mid-level features reflecting acoustical properties of objects in the scene. In contrast, CNN feature maps are somewhat vague and quite hard in the interpretation.

5. REFERENCES

[1] S. Agcaer, A. Schlesinger, F.-M. Hoffmann, and R. Martin, "Optimization of amplitude modulation features for low-resource acoustic scene classification," in *23rd European Signal Processing Conference – EUSIPCO'2015*, Nice, France, August 31 – September 4 2015, pp. 2556–2560.

[2] A. Rakotomamonjy, "Enriched supervised feature learning for acoustic scene classification," Normandie Universite, Tech. Rep., 2016.

[3] A. Jimenez, B. Elizalde, and B. Raj, "Dcase 2017 task 1: Acoustic scene classification using shift-invariant kernels and random features," *arXiv preprint arXiv:1801.02690*, 2018.

[4] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[5] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, no. Nov, pp. 1457–1469, November 2004.

[6] S. S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, USA, February 8–11 1998, pp. 127–132.

[7] J. Gillard and M. Schutz, "The importance of amplitude envelope: Surveying the temporal structure of sounds in perceptual research," in *10th Sound and Music Computing Conference – SMC'2013*, Stockholm, Sweden, July 30 – August 3 2013, pp. 62–68.

[8] A. Lerch, *An Introduction to Audio Content Analysis – Applications in Signal Processing and Music Informatics*. John Wiley & Sons, Inc., 2012.