AUDIO FEATURE SPACE ANALYSIS FOR ACOUSTIC SCENE CLASSIFICATION

Tomasz Maka

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

ABSTRACT

The paper presents a study of audio features analysis for acoustic scene classification. Various feature sets and many classifiers were employed to build a system for scene classification by determining compact feature space and using an ensemble learning. The input feature space containing different sets and representations were reduced to 223 attributes using the importance of individual features computed by gradient boosting trees algorithm. The resulting set of features was split into distinct groups partly reflected auditory cues, and then their contribution to discriminative power was analysed. Also, to determine the influence of the pattern recognition system on the final efficacy, accuracy tests were performed using several classifiers. Finally, conducted experiments show that proposed solution with a dedicated feature set outperformed baseline system by 6%.

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

1. INTRODUCTION

The classification of the acoustical environment plays an essential role in human-machine interaction systems and it becomes a very popular research area in the last decade. The process of acoustical scene analysis involves many auditory cues [1] to determine the components of the scene. These cues are exploited to decompose and grouping of acoustic streams based on perceptual mechanisms of human hearing [2]. Attributes like periodicity, onsets and offsets, amplitude and frequency modulation, discontinues in the frequency domain, time–frequency units are very often used in the process of forming auditory objects. The time–frequency structure of an acoustic scene is dependent on the number of sound sources, its properties and variability in time. Additionally, the knowledge of acoustical attributes and their perceptual and physical meaning allows to create more sophisticated features and facilitate the scene decomposition.

On the other hand, in the deep–learning paradigm [3] the features are computed using unsupervised learning with only minimal preprocessing of the raw audio data as an input. In particular, features estimated in convolutional neural networks [4] yield to high classification accuracy and considerably outperforms the traditional pattern recognition systems. The problem with such features is the difficulty in their acoustical interpretation which may be necessary for system adaptation to changing environmental conditions in the acoustical scene. The solution in such a case requires a lot of data, causing the model to become large and complex. Since the features are the critical element of audio analysis systems, its selection is not a trivial task. The type of features and the size of the feature space determine the model used at the pattern recognition stage. Model complexity directly affects the system implementation, it defines required memory, computational resources and is the component influencing on the classification accuracy.

The most audio analysis systems dedicated to events detection or scene classification and using low–level features generate large feature spaces often containing more than a thousand attributes. Authors in [5] proposed a system with a large number of cepstral, spectral, voicing and energy features with statistical functionals, delta and acceleration coefficients. The parametrisation stage operated on the feature space with 6669 elements. The approach to event detection described in [6] uses 4096 audio features derived from well–known MFCC [7] features. An approach using 2000 features based on non-negative supervised matrix factorisation with Gaussian kernel SVM classifier is presented in [8]. A dimensionality of feature space equal to 4096 with were proposed in [9]. The computed random features approximating three types of kernels with SVM classifier were used to acoustic scene classification task. A very low-dimensional feature space was presented in [10]. Only nine optimised AMS (Amplitude Modulation Spectrum) features together with the LDA classifier was employed to classify acoustic scenes.

This study is a part of the work being developed for the purpose of creating the hierarchy of the robust audio features and its high–level representations for extracting objects and their properties from an audio stream.

2. PROPOSED FRAMEWORK

Due to the attempt of audio features analysis dedicated to acoustic scene analysis, we decided to use the traditional machine learning scheme. Such an approach is realised in two steps, where at first stage an input signal is converted to a feature space, then the data is fed to the classifier at the second step. The initial set of features was inspired by the auditory cues proposed for scene analysis [1, 2]. Dimensionality of the source set of features was equal to 2861. Next, the number of attributes was reduced in the feature importance analysis process using Gradient Boosting Machine [11] for whole development dataset. The resulting feature vector is composed of 13 subsets with 223 discriminative attributes as depicted in Figure 1. The final subsets can be briefly summarised as follows:

Binaural unit (F_1) – interaural time difference, interaural intensity difference, interaural coherence, and azimuth.

Pitch properties (F_2) – statistical properties of pitch contour.

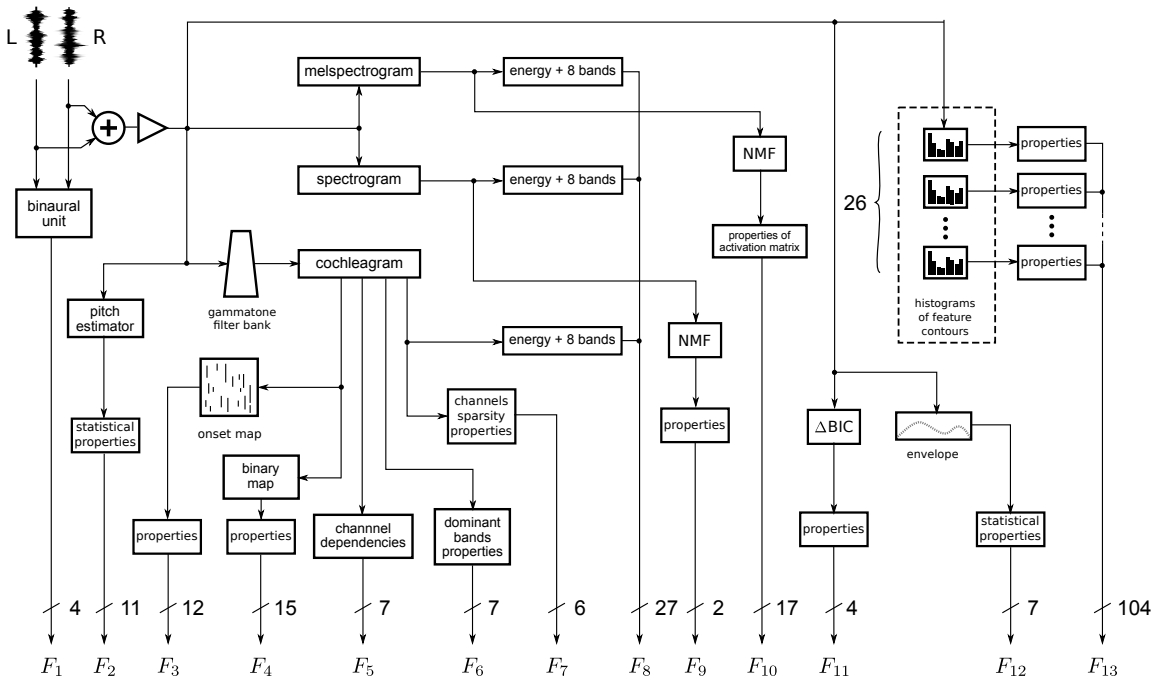

Figure 1: The diagram of the proposed system converting an audio signal to feature vector with 223 attributes.

Onset map (F_3) – properties of onsets detected in all channels of cochleagram.

Binary map (F_4) – attributes of binary map obtained by thresholding single channels of cochleagram.

Channel dependencies (F_5) – energy differences between neighbouring channels of cochleagram.

Dominant bands (F_6) – selected number of bands with the highest energies in cochleagram [12].

Channels sparsity (F_7) – Hoyer sparsity [13] computed for the individual channels of cochleagram.

Sub-band energies (F_8) – energies calculated in 8 equally sized ranges of cochleagram, melspectrogram and spectrogram.

Spectrogram activations (F_9) – attributes of activation matrix by computing non-negative matrix factorisation of spectrogram.

Melspectrogram activations (F_{10}) – properties of activation matrix by computing non-negative matrix factorisation of melspectrogram.

ΔBIC trajectory (F_{11}) – attributes of trajectory calculated as a difference between Bayesian Information Criterion (BIC) values of models used in audio segmentation [14].

Temporal envelope (F_{12}) – properties of temporal envelope [15].

Histograms of feature contours (F_{13}) – characteristic of histograms obtained for various [16] low–level feature contours.

In order to determine the variability of attributes between classes, we have averaged feature vectors over development set and mapped them clockwise onto unit circle as shown in Figure 2. Such visualisation highlights the similarities and differences between classes and can be used to determine the discriminative attributes. For example, in the case of the 'Tram' and 'Bus' classes the averaged feature vectors are quite similar and may be a reason for misclassification. After initial experiments with obtained feature space and

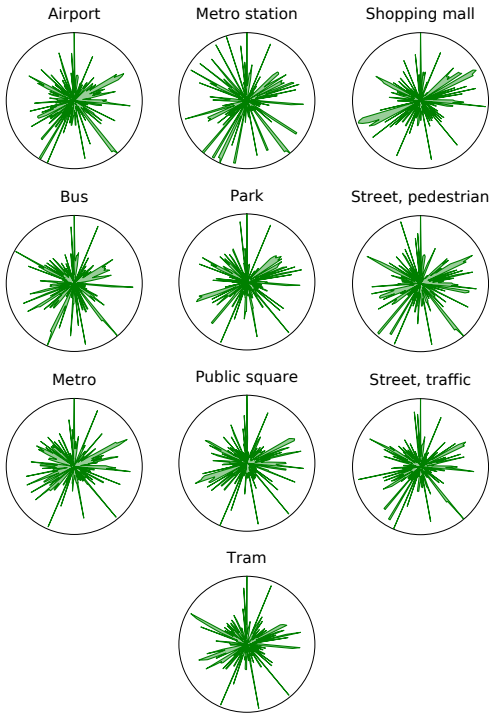

Figure 2: Averaged, normalised and mapped onto unit circle feature vectors for the whole development set.

standard classifiers, we decided to employ an ensemble learning. The reason was the classification accuracies we acquired for 64 individual classifiers because the average accuracy value was close to the baseline. The selection of classifiers was performed from a set of the classifiers with the accuracies higher than 50%. Then, an ensemble learning with majority/hard voting was executed. In the next step, successive classifier combinations were removed or replaced from the set to maximise the accuracy. The procedure ends when no improvements in classification accuracy occur.

In the result, a set of classifiers presented in Table 1 have been used in the majority voting scheme. There was no specific tuning of the classifiers, the parameters and configurations were selected randomly by the selection algorithm.

Table 1: The final set of classifiers 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. EXPERIMENTAL EVALUATION

The system performance was evaluated on the development dataset of DCASE'2018 competition (Task 1) [17]. The audio data was recorded in 10 acoustic scenes and consists of binaural, 8640 segments each 10 seconds long using 48 kHz sampling rate and 24-bit resolution. The recordings were captured in six European cities.

In the first experiment, we have verified classification accuracy for individual classifiers using a complete feature set. The results are presented in Figure 3. In three cases, the accuracy exceeded 60%: for classifier C_1 is equal 62.9%, for C_4 is 63.2% and for classifier C_5 is equal to 61.9%.

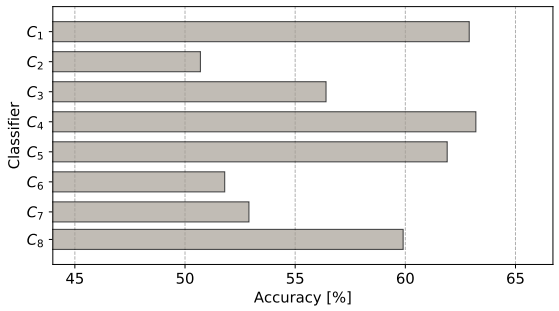

Figure 3: Influence of individual classifiers on the classification accuracy using the complete feature space.

Another experiment was to determine a discriminate power for subsets ($F_1 - F_{13}$) defined in Figure 1. The performance of acoustic scene recognition for individual subsets is presented in Figure 4.

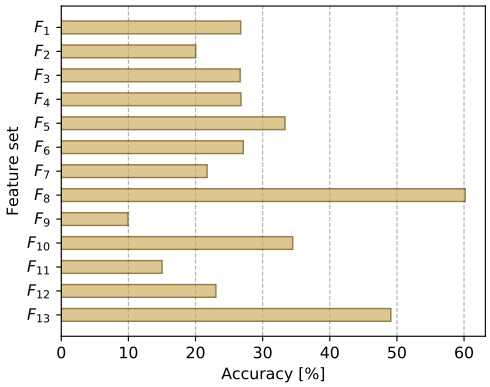

Figure 4: System performance using separate feature subsets.

The results show that the most discriminative subset F_8 gives the accuracy equal to 60.17%. Because this subset includes attributes from three different time-frequency representations in various frequency ranges, we examined the influence of each of the representations on the classification effectiveness. In Table 3 the results for three feature vectors computed from different representations are depicted. For each case the frequency range in further divided into eight equal bands to form the feature vector. The best-obtained accuracy is observed for cochleagram which may suggest that most discriminative data is located below 8kHz in the frequency domain.

The impact of individual subsets on the classification effectiveness was carried out in two subsequent experiments. In the first analysis, we started with the subset that has the highest discriminatory power (see Figure 4), then the main set was increased by the consecutive subsets as shown in Figure 5.

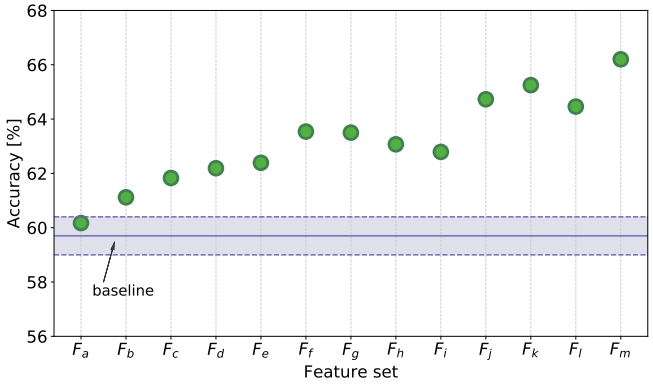

Figure 5: Classification results with combined subsets of attributes in order from the most to the least discriminative[1].

[1] $F_a = F_8$; $F_b = F_a \cup F_{13}$; $F_c = F_b \cup F_{10}$; $F_d = F_c \cup F_5$; $F_e = F_d \cup F_6$; $F_f = F_e \cup F_4$; $F_g = F_f \cup F_1$; $F_h = F_g \cup F_3$; $F_i = F_h \cup F_{12}$; $F_j = F_i \cup F_7$; $F_k = F_j \cup F_2$; $F_l = F_k \cup F_{11}$; $F_m = F_l \cup F_9$

Table 2: The class-wise accuracy of the development set: confusion matrix (a), comparison with the baseline (b).

Category \ Estimate	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**

(a)

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 %**

(b)

Table 3: Performance of individual representations of set F_8.

Representation	Frequency range [Hz]	Bands	Accuracy
Cochleagram	50 – 8000	128	51.99 %
Melspectrogram	0 – 12000	128	46.58 %
Spectrogram	0 – 24000	1024	42.65 %

In the second experiment, the attempts were made to remove further subsets to assess the impact of the resulting feature set on the classification effectiveness. According to the results depicted in Figure 6, the subset F_8 is a crucial part of the feature vector. Its contribution is similar as in case of the experiment which results are presented in Figure 4. Interestingly, the smallest decreasing of accuracy is observed for the pitch related features (subset F_2) although such attributes are discriminative for parts contained speech in an audio signal.

Finally the classification experiments were performed using the proposed framework and development dataset. The confusion matrix is presented in Table 2a. 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 'Tram'. The comparison of our system with the baseline is shown in Table 2b, where in case of classes 'Airport' and 'Public square' decrease in accuracy was observed.

Due to the length of the recordings, many of the segments have a similar acoustical structure for different classes which caused misclassification. Unfortunately, is such case, low-level audio features are ineffective.

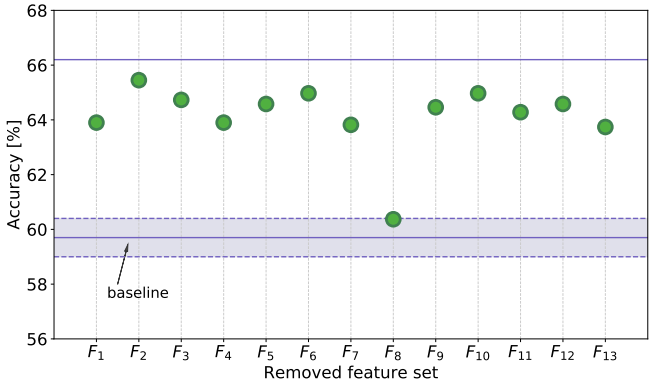

Figure 6: Obtained accuracy in the situation of removing individual subsets of the final feature set.

4. CONCLUSION

In this paper, we have presented an approach to classify of acoustic scenes with the dedicated set of audio features and ensemble learning classification stage. An advantage of our system is a small set of features which can be used in systems with low resources. At the current stage of development, our system has worse efficiency in comparison to deep-learning based solutions, but similar to human hearing abilities for the development set. In future work, we intend to design hierarchical audio features dedicated to specific acoustic scenes including events and background noise. For this purpose, we have designed and implemented a dedicated application[2]. It can be used to browse various audio representations and to support the process of developing a new hybrid features.

[2] http://quefrency.org/dcase2018

5. REFERENCES

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis – Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds. IEEE Press / Wiley–Interscience, 2006.

[2] A. S. Bregman, *Auditory Scene Analysis – The Perceptual Organization of Sound*, 1st ed. The MIT Press, 1994.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[4] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *IEEE International Symposium on Circuits and Systems – ISCAS'2010*. IEEE, 2010, pp. 253–256.

[5] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics – WASPAA'2013*, New Paltz, NY, USA, October 20–23 2013, pp. 1–4.

[6] F. Metze, S. Rawat, and Y. Wang, "Improved audio features for large-scale multimedia event detection," in *IEEE International Conference on Multimedia and Expo – ICME'2014*. IEEE, 2014, pp. 1–6.

[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

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

[9] 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.

[10] 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.

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

[12] T. Maka, "Auditory scene classification based on the spectro-temporal structure analysis," West Pomeranian University of Technology, Szczecin, Tech. Rep., 2017.

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

[14] 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.

[15] 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.

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

[17] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," 2018, submitted to DCASE2018 Workshop.