Investigating Combinations of Visual Audio Features and Distance Metrics in the Problem of Audio Classification

Paweł Forczmański and Tomasz Maka

Abstract The article addresses a problem of audio signal classification employing image processing and recognition methods. In such an approach, vectorized audio signal features are converted into a matrix representation (feature map), and then processed, as a regular image. In the paper, we present a process of creating a low-dimensional feature space by means of two-dimensional Linear Discriminant Analysis and projecting input feature maps into this subspace using two-dimensional Karhunen–Loeve Transform. The classification is performed in the reduced feature space by means of voting on selected distance metrics applied for various features. The experiments were aimed at finding an optimal (in terms of classification accuracy) combination of six feature types and five distance metrics.The found combination makes it possible to perform audio classification with high accuracy, yet the dimensionality of resulting feature space is significantly lower than input data.

Keywords Audio classification · Feature extraction · Image recognition · 2DLDA · 2DKLT · Distance metrics

1 Introduction

Analysis of audio signals is one of the most interesting and challenging tasks of multimedia systems. While there are many issues that are fairly good solved, e.g., the identification of persons on the basis of the registered voice and speech recognition in terms of content; there are also many problems that are still left for analysis. This includes the automatic classification of audio scene in terms of recognizing background sounds in certain environments, e.g., restaurant, factory, street, etc. The

P. Forczmański (✉) · T. Maka
Faculty of Computer Science and Information Technology, West Pomeranian University
of Technology, Szczecin, Żołnierska Str. 52, 71–210 Szczecin, Poland
e-mail: pforczmanski@wi.zut.edu.pl

T. Maka
e-mail: tmaka@wi.zut.edu.pl

R. Burduk et al. (eds.), *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, Advances in Intelligent Systems and Computing 403, DOI 10.1007/978-3-319-26227-7_69

sensitivity, which characterizes human sense of hearing is already achievable for machines—each of the physical quantities characterizing audio signal can now be specified much more precisely using computerized analyzers than using human sense of hearing. On the other hand, human beings are still able to use the acquired information from the audio signal in a more effective manner. It should be noted also that background sounds are extremely complex, when it comes to the formal description. The automation of audio scene classification process can have multiple purposes, e.g., improving speech-based services that depend on properties of environment, content-based audio retrieval in large multimedia databases, intelligent surveillance, etc. The audio signal can be described by a number of parameters [19, 21]. So, it seems that there is no need to seek for new methods of representation of audio signal, but we should focus on the selection and use of existing ones. As it was shown in [16, 21], most of the developed methods use physical characteristics of sound in a form of low-level feature vectors. The classification of such data is based mainly on a one-dimensional approach [1, 15]. In contrast, developed algorithm adopts a two-dimensional approach to audio signal and is based on the observation that the visual representation of audio signal may carry more useful information that in case of one-dimensional vector representation [5]. In such case, certain methods aimed at image processing may obtain higher accuracy of classification in comparison to established methods [18]. Recently, many applications of Support Vector Machines to audio scene classification have been proposed (e.g., [10, 11]). The results are promising, however, such methods are complex in terms of training. Therefore, this work is devoted to the task of evaluation of the possibility of application of pattern recognition methods in the context of automated classification of audio scene. The main focus is put on finding a combination of different audio features and adequate, simple distance metrics (used as classifiers), that gives the highest possible classification accuracy.

2 Feature Maps

In typical audio analysis task, various features calculated in several domains can be employed. The feature space should be able to capture time-frequency structure of the signal. Therefore, in our approach we use feature space created from a set of feature vectors arranged into a feature map. In such an approach, the source audio signal of length N is decomposed into a set of M overlapping frames (the overlap between consecutive frames is equal to half of the frame length). For each frame a feature vector of length W is calculated and added as a new column to the feature matrix (map). The number of columns in the feature map is equal to $M = \lfloor (F_s \cdot (N - K)) / \lfloor F_s \cdot 0.5 \cdot K \rfloor \rfloor + 1$, where: N is the total length of input signal [s], K is a frame size [s] and F_s the sampling rate [Hz]. In our investigations, the size of feature vector extracted for each audio frame was constant and equal to $W = 24$. Resulting feature maps calculated for exemplary audio clip are shown in Fig. 1. In order to capture various characteristics of audio we have employed feature

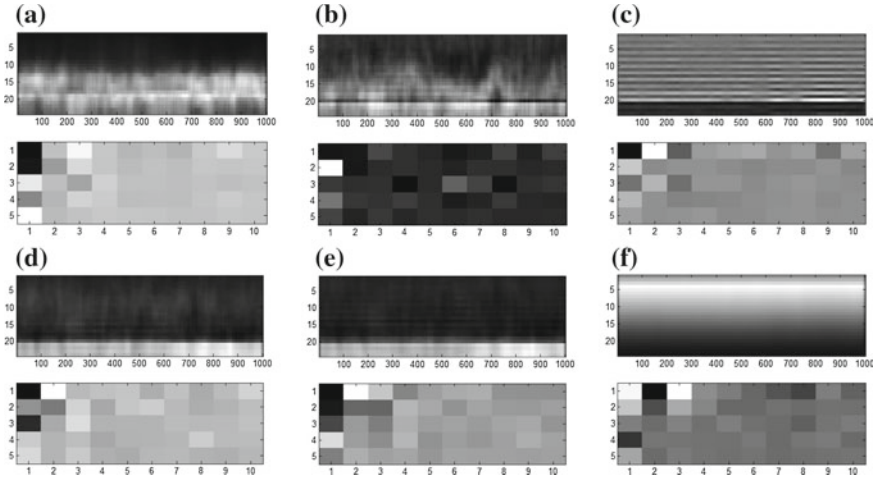

Fig. 1 Feature maps of exemplary audio signal for the following features ($W = 24$) and their 2DLDA-based representation ($s = 5, p = 10$): BFB (**a**), MFCC (**b**), LPC (**c**), LFCC (**d**), LPCC (**e**), LSF (**f**)

maps constructed on the basis of state-of-the-art feature sets, namely *Bark-Frequency Filter Bank* (BFB) [22], *Mel-Frequency Cepstral Coefficients* (MFCC) [3], *Linear Prediction Coefficients* (LPC) [19], *Linear-Frequency Cepstral Coefficients* (LFCC) [19], *Linear Prediction Cepstral Coefficients* (LPCC) [19], *Line Spectral Frequencies* (LSF) [12].

3 Algorithm Description

Created feature map is intended to preserve important low-level characteristics of audio signal. Its informative potential depends on the resolution of the image. Unfortunately, processing large image matrices requires large time and memory overhead and is not desirable. Therefore, in many similar tasks, certain dimensionality reduction is applied. Although, simple spatial resampling could be useful, it would lead to the elimination important details. Hence, we apply a more sophisticated approach, namely dimensionality reduction preceded by a data analysis stage. Our algorithm includes three components: data preparation (feature map creation), reduction of dimensionality (projection to the subspace), and classification (see Fig. 2). At the offline stage of building a reference database containing reduced forms of feature maps, we use a method employed, among others, in recognizing facial images, stamps and textures [6, 13, 17], namely Two-dimensional Linear Discriminant Analysis (2DLDA) [13], while at the recognition stage of test samples—Two-dimensional Karhunen–Loeve Transform (2DKLT) [7, 8, 13]. In order to obtain such reduc-

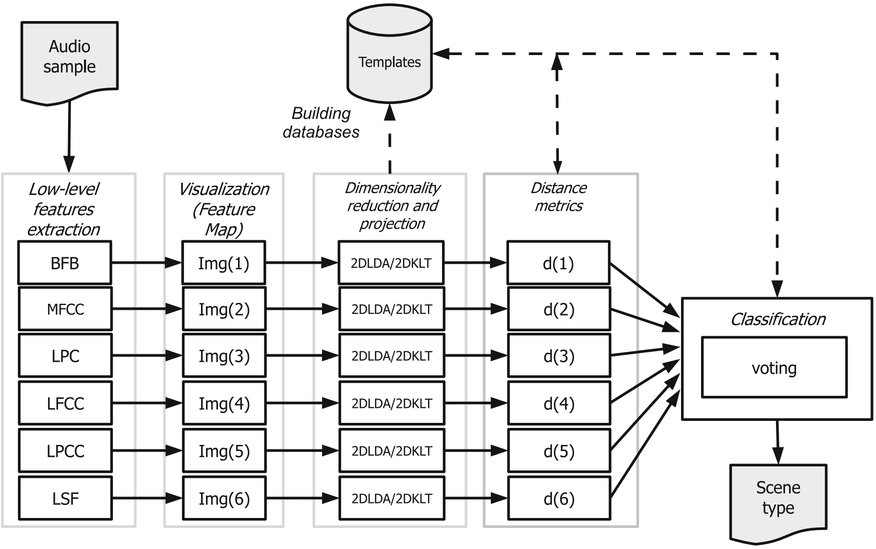

Fig. 2 Scheme of audio sample processing and classification

tion effect, together with clustering improvement, we propose to use a dimensionality reduction stage. The analysis of the research showed that Linear Discriminant Analysis may be successfully applied in this case. Since one-dimensional variant of LDA requires much more space to store covariance matrices and also does not cope with small-sample-size problem, we propose to use a two-dimensional Linear Discriminant Analysis. At the stage of classification, we use a combination of simple distance-based classifiers (1-Nearest Neighbor with different distance metrics).

3.1 Dimensionality Reduction

The aim of dimensionality reduction is to limit the volume of data describing input samples by selecting dimensions that represent them with a acceptably low error. It is typical for many methods involving subspace projection that all samples from the learning part of the database are subject to analysis (2D analysis), which leads to forming the transformation matrices. They are later used to project input images (2D projection) into a subspace forming a reference database. Feature maps calculated for audio samples are grouped according to G classes corresponding to the content. In our approach, each input feature map $X^{(g,l)}$ is represented as a grayscale image represented using matrix of dimensions $W \times M$ elements, where g is the class number and l is a respective feature map in each class ($l = 1, \ldots, L$). This approach is repeated for each feature type, which number, in our case is equal to six. In the first step, for each feature type, an average matrix $\bar{X}_{W \times M}$ of all the matrices X is calculated and a mean matrix in each class $g = 1, \dots, G$: $\bar{X}^{(g)}_{W\times M}$. Due to the large size of input feature maps (in experimental studies—even 24 × 1001 pixels) it is not possible to apply directly the one-dimensional LDA method, therefore, it was decided to use a two-dimensional variant—2DLDA. In this method, the covariance matrices for within-class scatter (Sw) and between-class scatter (Sb) are calculated, each two for row and column representation of feature map, respectively, [13]: $Sw^{(Row)}_{W\times W}$ and $Sb^{(Row)}_{W\times W}$, $Sw^{(Col)}_{M\times M}$ and $Sb^{(Col)}_{M\times M}$. Then, we calculate the corresponding matrices H, determining the total distribution of the classes in the feature space: $H^{(Row)}_{W\times W} = \left(Sw^{(Row)}_{W\times W}\right)^{-1} Sb^{(Row)}_{W\times W}$ and $H^{(Col)}_{M\times M} = \left(Sw^{(Col)}_{M\times M}\right)^{-1} Sb^{(Col)}_{M\times M}$. They are used to maximize the so-called Fisher criterion, the aim of which is to increase the dispersion between-class scatter in relation to the intra-class [9]. This gives the improvement of clustering and significantly increases the effectiveness of later classification. For the matrices $H^{(Row)}$ and $H^{(Col)}$, we solve the task of searching for the eigenvalues $\{\Lambda^{(Row)}, \Lambda^{(Col)}\}$ and eigenvectors $\{V^{(Row)}, V^{(Col)}\}$. In the final step of analysis, from the diagonals of $\Lambda^{(Row)}$ and $\Lambda^{(Col)}$ s and p maximal elements are selected, respectively, and their positions are recorded. From $\left(V^{(Row)}\right)^T$ s rows corresponding to selected elements are extracted, and from $V^{(Col)}$ p columns in the same way. Then, the two transformation matrices are constructed: $F^{(Row)}$ containing $s \times W$ elements and $F^{(Col)}$ having $M \times p$ elements. Projection of lth feature map from the gth class $X^{(g,l)}$ into a subspace is performed by matrix multiplication [13] $Y^{(k,l)}_{s\times p} = F^{(Row)}_{s\times W}\left(X^{(g,l)}_{W\times M} - \bar{X}_{W\times M}\right) F^{(Col)}_{M\times p}$. In order to balance between reduction degree and classification performance, it was assumed that the information contained in feature maps, arranged vertically (in columns of each image) is more important than horizontal information (stored in rows of image), so we used such relation as $s \leq p$.

3.2 Distance Metrics

As it was mentioned, audio pattern classification is based on the operations in the reduced feature space. Reference database contains reduced feature maps calculated for all learning samples. The class assignment of a query feature map (its reduced form $Y^{(q)}$) is done using a distance-based Nearest-Neighbor Classifier on all reference objects $Y^{(r)}$. We applied five popular distances, described below. The Euclidean distance (L_2 norm) [4] is given as

$$d_{\text{Euclid}}(Y^{(q)}, Y^{(r)}) = \sqrt{\sum_{s,p}\left(y^{(q)}_{s,p} - y^{(r)}_{s,p}\right)^2}. \tag{1}$$

The Chebyshev distance (L_{max} norm) is defined as [2]:

$$d_{\text{Cheb}}(Y^{(q)}, Y^{(r)}) = \max_{s,p}\left(\left|y^{(q)}_{s,p} - y^{(r)}_{s,p}\right|\right). \tag{2}$$

The taxicab distance (cityblock/Manhattan metric or L_1 norm) is calculated as [20]:

$$d_{\text{City}}(Y^{(q)}, Y^{(r)}) = \sum_{s,p} \left| y_{s,p}^{(q)} - y_{s,p}^{(r)} \right| . \tag{3}$$

The Minkowski distance of order m (in our case $m = 5$) is defined as:

$$d_{\text{Mink}}(Y^{(q)}, Y^{(r)}) = \left(\sum_{s,p} \left| y_{s,p}^{(q)} - y_{s,p}^{(r)} \right|^m \right)^{\frac{1}{m}} . \tag{4}$$

The Canberra distance (a weighted version of Manhattan distance) is given as follows [14]:

$$d_{\text{Canb}}(Y^{(q)}, Y^{(r)}) = \sum_{s,p} \frac{\left| y_{s,p}^{(q)} - y_{s,p}^{(r)} \right|}{\left| y_{s,p}^{(q)} \right| + \left| y_{s,p}^{(r)} \right|} . \tag{5}$$

Above distances have been chosen on a basis of computational simplicity and proved efficiency in many pattern recognition problems.

3.3 Voting Schemes

We employed two schemes of joining elementary classifiers' results. Both of them are based on voting. The first one comes from the observation that different distant metrics applied for a single feature can have significantly different accuracy. Thus, we classify an unknown sample in parallel using five distances presented above and remember the results. When we join these five results and use majority voting scheme, we may obtain higher classification accuracy. It is also known as a mode, when we chose value that appears most often in a set of data. In case when all values appear with the same frequency or there is no possibility to chose one single *winner*, we choose arbitrarily the first result. The second approach is based on the similar assumption, that each feature may have a preferred distance metrics. It should be noted that this relation is unknown to us. In such case, during a classification of an unknown sample, we collect 30 individual results (combination of five distances and six features). After a majority voting, we have the final classification result. Both approaches are slightly different, since in the first case we set a feature and calculate distances for this particular feature, while in the second one we calculate all the combinations of features and distances. Although, the first approach is less computationally intensive, its accuracy depends on the selected features.

4 Experiments

4.1 Database Characteristics

The benchmark dataset used for the evaluation of algorithm's performance contains six classes ($g = 1, \dots, 6$) of the environmental sounds. The sounds represent the following acoustic scenes called: 'passing cars' ($g = 1$), 'rain' ($g = 2$), 'restaurant' ($g = 3$), 'shopping centre' ($g = 4$), 'beach' ($g = 5$), and 'factory' ($g = 6$). Each class consists of 42 audio recordings divided into 30 items used for learning phase and 12 items—for testing. Every audio recording is monaural, 10 s long with 22050 Hz sampling rate—total length of audio data is 42 min long. In order to demonstrate the complexity of the problem, Fig. 1 shows exemplary feature maps for each of the classes. It should be noted that their distinguishability in original attributes space is very limited. In the same figure, reduced representations of each feature map are also given. As it can be seen, most of the energy is condensed in the upper left corner of each matrix.

4.2 Experimental Results

In order to evaluate the performance of the proposed algorithm of classification, we investigated combinations of features and classifiers on the same dataset (classification schemes were presented above). The reduction parameters at the stage of 2DLDA/2DKLT is $s = 5$ and $p = 10$, hence we classified objects using 50 features, which is lower than in case of other known methods. The results are presented in Tables 1 and 2. We provide results for different number of classes ($G = \{3, 4, 5, 6\}$), since the first three classes are rather easy to distinguish, i.e. for $G = 3$ we take into consideration classes $g = \{1, 2, 3\}$ and for $G = 6$, classes $g = \{1, 2, 3, 4, 5, 6\}$, respectively. As it can be seen, the recognition accuracy decreases with the increase of number of classes. The other observation is that it is impossible to select one pair of feature-distance for which the accuracy is the highest for all the cases. In general, LPCC and LFCC perform better, no matter which distance metric we choose. Moreover, the voting on five distances calculated on the same feature gives slight increase in the classification accuracy over the individual case. The second voting scheme was also investigated. The results of experiments are provided in Table 3. This table presents recognition accuracy for different number of classes in two variants: maximum accuracy is the highest accuracy from Tables 1 and 2, while the voted accuracy means an accuracy for voting involving provided set of six pairs feature/distance (all features have been used), e.g., BFB/d_{City}, MFCC/d_{Cheb}, etc. As it can be seen, there is a significant progress in comparison to the first scheme. After applying the second scheme we are able to classify the simplest case of $G = 3$ classes with 100 % accuracy. On the other hand, the case of $G = 6$ classes still causes problems in terms of efficient classification. This is due to the characteristics of sound clips gathered in

Table 1 Classification accuracy as a function of distance metrics and feature type, for $G = 3$ and $G = 4$

Distance	$g = \{1, 2, 3\}$						$g = \{1, 2, 3, 4\}$					
	BFB	MFCC	LPC	LFCC	LPCC	LSF	BFB	MFCC	LPC	LFCC	LPCC	LSF
d_{Cheb}	0.81	0.72	0.64	0.81	0.75	0.5	0.63	0.44	0.31	0.69	0.6	0.38
d_{Euclid}	0.81	0.69	0.67	0.78	0.81	0.61	0.52	0.5	0.33	0.5	0.58	0.42
d_{City}	0.83	0.69	0.58	0.81	0.81	0.67	0.56	0.46	0.4	0.58	0.58	0.48
d_{Mink}	0.78	0.69	0.64	0.83	0.78	0.5	0.58	0.5	0.35	0.69	0.65	0.4
d_{Canb}	0.75	0.58	0.58	0.72	0.67	0.83	0.6	0.4	0.44	0.44	0.54	0.48
Voted	0.81	0.75	0.67	0.83	0.81	0.61	0.56	0.56	0.33	0.56	0.56	0.42

Table 2 Classification accuracy as a function of distance metrics and feature type, for $G = 5$ and $G = 6$

Distance	$g = \{1, 2, 3, 4, 5\}$						$g = \{1, 2, 3, 4, 5, 6\}$					
	BFB	MFCC	LPC	LFCC	LPCC	LSF	BFB	MFCC	LPC	LFCC	LPCC	LSF
d_{Cheb}	0.47	0.42	0.22	0.53	0.52	0.25	0.32	0.38	0.15	0.42	0.4	0.25
d_{Euclid}	0.42	0.48	0.27	0.48	0.5	0.35	0.36	0.36	0.26	0.33	0.38	0.26
d_{City}	0.45	0.43	0.33	0.55	0.62	0.37	0.38	0.32	0.26	0.42	0.44	0.31
d_{Mink}	0.47	0.42	0.28	0.53	0.53	0.28	0.36	0.38	0.21	0.43	0.36	0.25
d_{Canb}	0.43	0.42	0.33	0.38	0.55	0.32	0.4	0.32	0.26	0.32	0.33	0.29
Voted	0.45	0.48	0.32	0.5	0.58	0.32	0.38	0.33	0.25	0.38	0.43	0.22

Table 3 Classification accuracy for voting scheme with different features/distances

Number of classes (G)	Maximum accuracy	Voted accuracy	Features					
			BFB	MFCC	LPC	LFCC	LPCC	LSF
3	0.83	1.0	d_{City}	d_{Cheb}	d_{Cheb}	d_{Canb}	d_{Mink}	d_{Canb}
4	0.69	0.77	d_{Mink}	d_{Euclid}	d_{Canb}	d_{Mink}	d_{Mink}	d_{Canb}
5	0.62	0.67	d_{Canb}	d_{Euclid}	d_{Canb}	d_{Cheb}	d_{City}	d_{Canb}
6	0.44	0.53	d_{Canb}	d_{Euclid}	d_{Euclid}	d_{Mink}	d_{Cheb}	d_{Canb}

Table 4 Combination of feature/distance metric giving the highest possible classification accuracy for $G = 6$ and majority voting

No.	1	2	3	4	5	6
Feature	LFCC	MFCC	BFB	LPC	LPCC	LPCC
Distance	d_{Mink}	d_{Euclid}	d_{Canb}	d_{City}	d_{City}	d_{Canb}

the benchmark database and the presence of various, nonstationary sound events in the clips. An additional experiment was devoted to finding a combination of six pairs of feature/distance metric which give the highest possible accuracy. In this experiment, we assumed that not all features are used and there may be also situations, when the same feature is used more than one time (with different distance metrics). In order to find it, 30^6 combinations were examined using brute-force strategy. The classification accuracy found for $G = 6$ is close to 0.59, and the combination is as follows (see Table 4).

5 Summary

An algorithm for automatic audio scene classification was presented. The experiments were aimed at finding an optimal combination of features and distance metrics in terms of classification accuracy. Conducted experiments on a benchmark database showed the high efficiency of developed method. In comparison to other more complex methods (e.g., [10, 11]), the presented approach has good recognition accuracy together with simple implementation and low dimensionality of feature space, especially in comparison to other techniques. In future, certain more sophisticated classifiers and distance matrices may be used, which could lead to the further increase in classification accuracy, especially in case of higher number of audio classes.

References

1. Abe, M., Matsumoto, J., Nishiguchi, M.: Content-based classification of audio signals using source and structure modelling. In: Proceedings of the IEEE Pacific Conference on Multimedia, pp. 280–283 (2000)
2. Cantrell, C.D.: Modern Mathematical Methods for Physicists and Engineers. Cambridge University Press, Cambridge (2000)
3. Davis, S., Mermelstein, P.: Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. ASSP **28**(4), 357–366 (1980)
4. Deza, E., Deza, M.M.: Encyclopedia of Distances. Springer, Berlin (2009)
5. Forczmański, P.: Evaluation of singer's voice quality by means of visual pattern recognition. J. Voice. doi:10.1016/j.jvoice.2015.03.001 (in press, 2015)
6. Forczmański, P., Frejlichowski, D.: Classification of elementary stamp shapes by means of reduced point distance histogram representation. Mach. Learn. Data Min. Pattern Recognit., LNCS **7376**, 603–616 (2012)
7. Forczmański, P., Labedz, P.: Recognition of occluded faces based on multi-subspace classification. In: 12th IFIP TC8 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), LNCS, vol. 8104, pp. 148–157 (2013)
8. Forczmański, P., Kukharev, G., Shchegoleva, N.: Simple and robust facial portraits recognition under variable lighting conditions based on two-dimensional orthogonal transformations. In: 17th International Conference on Image Analysis and Processing (ICIAP), LNCS, vol. 8156, pp. 602–611 (2013)
9. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, New York (1990)
10. Geiger, J.T., Schuller, B., Rigoll, G.: Large-scale audio feature extraction and SVM for acoustic scene classification. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–4 (2013)
11. Jiang, H., Bai, J., Zhang, S., Xu, B.: SVM-based audio scene classification, natural language processing and knowledge engineering. In: Proceedings of 2005 IEEE International Conference on IEEE NLP-KE'05, pp. 131–136 (2005)
12. Kleijn, W., Backstrom, T., Alku, P.: On line spectral frequencies. IEEE Signal Process. Lett. **10**(3), 75–77 (2003)
13. Kukharev, G., Forczmański, P.: Face recognition by means of two-dimensional direct linear discriminant analysis. In: Proceedings of the 8th International Conference PRIP 2005 Pattern Recognition and Information Processing. Republic of Belarus, Minsk, pp. 280–283 (2005)
14. Lance, G.N., Williams, W.T.: Computer programs for hierarchical polythetic classification ("similarity analysis"). Comput. J. **9**(1), 60–64 (1966)
15. Maka, T.: Attributes of audio feature contours for automatic singing evaluation. In: 36th International Conference on Telecommunications and Signal Processing (TSP), pp. 517–520. Rome, Italy, 2–4 July 2013
16. McKinney, M., Breebaart, J.: Features for audio and music classification. In: Proceedings of the International Symposium on Music Information Retrieval, pp. 151–158. Baltimore, Maryland (USA), 26–30 Oct 2003
17. Okarma, K., Forczmański, P.: 2DLDA-based texture recognition in the aspect of objective image quality assessment. Ann. Univ. Mariae Curie-Sklodowska. Sectio AI Informatica **8**(1), 99–110 (2008)
18. Paraskevas, I., Chilton, E.: Audio classification using acoustic images for retrieval from multimedia databases. In: 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications. EC-VIP-MC, pp. 187–192. Zagreb, Croatia, 2–5 July 2003
19. Rabiner, L., Schafer, W.: Theory and Applications of Digital Speech Processing. Prentice-Hall, Englewood Cliffs (2010)
20. Sammut, C., Webb, G.: Encyclopedia of Machine Learning. Springer, Berlin (2010)

21. Schuller, B., Wimmer, M., Moesenlechner, L., Kern, C., Arsic, D., Rigoll, G.: Brute-forcing hierarchical functionals for paralinguistics: a waste of feature space? In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4501–4504 (2008)
22. Smith III, J.O.: Spectral Audio Processing. W3K Publishing, Stanford (2011)