AUTOMATIC TIMBRE CLASSIFICATION OF ETHNOMUSICOLOGICAL AUDIO RECORDINGS

Dominique Fourer, Jean-Luc Rouas, Pierre Hanna, Matthias Robine

▶ **To cite this version:**

Dominique Fourer, Jean-Luc Rouas, Pierre Hanna, Matthias Robine. AUTOMATIC TIMBRE CLASSIFICATION OF ETHNOMUSICOLOGICAL AUDIO RECORDINGS. International Society for Music Information Retrieval Conference (ISMIR 2014), Oct 2014, Taipei, Taiwan. hal-01095153

HAL Id: hal-01095153
https://hal.science/hal-01095153

Submitted on 17 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AUTOMATIC TIMBRE CLASSIFICATION OF ETHNOMUSICOLOGICAL AUDIO RECORDINGS

Dominique Fourer, Jean-Luc Rouas, Pierre Hanna, Matthias Robine
LaBRI - CNRS UMR 5800 - University of Bordeaux
{fourer, rouas, hanna, robine}@labri.fr

ABSTRACT

Automatic timbre characterization of audio signals can help to measure similarities between sounds and is of interest for automatic or semi-automatic databases indexing. The most effective methods use machine learning approaches which require qualitative and diversified training databases to obtain accurate results. In this paper, we introduce a diversified database composed of worldwide non-western instruments audio recordings on which is evaluated an effective timbre classification method. A comparative evaluation based on the well studied Iowa musical instruments database shows results comparable with those of state-of-the-art methods. Thus, the proposed method offers a practical solution for automatic ethnomusicological indexing of a database composed of diversified sounds with various quality. The relevance of audio features for the timbre characterization is also discussed in the context of non-western instruments analysis.

1. INTRODUCTION

Characterizing musical timbre perception remains a challenging task related to the human auditory mechanism and to the physics of musical instruments [4]. This task is full of interest for many applications like automatic database indexing, measuring similarities between sounds or for automatic sound recognition. Existing psychoacoustical studies model the timbre as a multidimensional phenomenon independent from musical parameters (e.g. pitch, duration or loudness) [7, 8]. A quantitative interpretation of instrument's timbre based on acoustic features computed from audio signals was first proposed in [9] and pursued in more recent studies [12] which aim at organizing audio timbre descriptors efficiently. Nowadays, effective automatic timbre classification methods [13] use supervised statistical learning approaches based on audio signals features computed from analyzed data. Thus, the performance obtained with such systems depends on the taxonomy, the size and the diversity of training databases. However, most of existing research databases (e.g. RWC [6], Iowa [5]) are only composed of common western instruments annotated with specific taxonomies. In this work, we revisit the automatic instrument classification problem from an ethnomusicological point of view by introducing a diversified and manually annotated research database provided by the *Centre de Recherche en Ethno-Musicologie* (CREM). This database is daily supplied by researchers and has the particularity of being composed of uncommon non-western musical instrument recordings from around the world. This work is motivated by practical applications to automatic indexing of online audio recordings database which have to be computationally efficient while providing accurate results. Thus, we aim at validating the efficiency and the robustness of the statistical learning approach using a constrained standard taxonomy, applied to recordings of various quality. In this study, we expect to show the database influence, the relevance of timbre audio features and the choice of taxonomy for the automatic instrument classification process. A result comparison and a cross-database evaluation is performed using the well-studied university of Iowa musical instrument database. This paper is organized as follows. The CREM database is introduced in Section 2. The timbre quantization principle based on mathematical functions describing audio features is presented in Section 3. An efficient timbre classification method is described in Section 4. Experiments and results based on the proposed method are detailed in Section 5. Conclusion and future works are finally discussed in Section 6.

2. THE CREM ETHNOMUSICOLOGICAL DATABASE

The CREM research database [1] is composed of diversified sound samples directly recorded by ethnomusicologists in various conditions (i.e. no recording studio) and from diversified places all around the world. It contains more than 7000 hours of audio data recorded since 1932 to nowadays using different supports like magnetic tapes or vinyl discs. The vintage audio recordings of the database were carefully digitized to preserve the authenticity of the originals and contain various environment noise. The more recent audio recordings can be directly digital recorded with a high-quality. Most of the musical instruments which com-

© Dominique Fourer, Jean-Luc Rouas, Pierre Hanna, Matthias Robine.
Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Dominique Fourer, Jean-Luc Rouas, Pierre Hanna, Matthias Robine. "Automatic timbre classification of ethnomusicological audio recordings", 15th International Society for Music Information Retrieval Conference, 2014.

[1] CREM audio archives freely available online at: http://archives.crem-cnrs.fr/

pose this database are non-western and can be uncommon while covering a large range of musical instrument families (see Figure 1(a)). Among uncommon instruments, one can find the lute or the Ngbaka harp as cordophones. More uncommon instruments like Oscillating bamboo, struck machete and struck girder were classified by ethnomusicologists as idiophones. In this paper, we restricted our study to the solo excerpts (where only one monophonic or polyphonic instrument is active) to reduce the interference problems which may occur during audio analysis. A description of the selected CREM sub-database is presented in Table 1. According to this table, one can observe that this database is actually inhomogeneous. The aerophones are overrepresented while membranophones are underrepresented. Due to its diversity and the various quality of the composing sounds, the automatic ethnomusicological classification of this database may appear as challenging.

Class name	Duration (s)		#	
aerophones-blowed	1,383		146	
cordophones-struck	357		37	
cordophones-plucked	715	1,229	75	128
cordophones-bowed	157		16	
idiophones-struck	522		58	
idiophones-plucked	137	753	14	82
idiophones-clinked	94		10	
membranophones-struck	170		19	
Total	3,535		375	

Table 1. Content of the CREM sub-database with duration and number of 10-seconds segmented excerpts.

3. TIMBRE QUANTIZATION AND CLASSIFICATION

3.1 Timbre quantization

Since preliminaries works on the timbre description of perceived sounds, Peeters *et al.* proposed in [12] a large set of audio features descriptors which can be computed from audio signals. The audio descriptors define numerical functions which aim at providing cues about specific acoustic features (e.g. brightness is often associated with the spectral centroid according to [14]). Thus, the audio descriptors can be organized as follows:

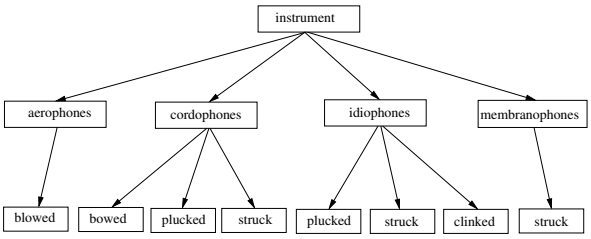
- Temporal descriptors convey information about the time evolution of a signal (e.g. log attack time, temporal increase, zero-crossing rate, etc.).
- Harmonic descriptors are computed from the detected pitch events associated with a fundamental frequency (F_0). Thus, one can use a prior waveform model of quasi-harmonic sounds which have an equally spaced Dirac comb shape in the magnitude spectrum. The tonal part of sounds can be isolated from signal mixture and be described (e.g. noisiness, inharmonicity, etc.).
- Spectral descriptors are computed from signal time-frequency representation (e.g. Short-Term Fourier Transform) without prior waveform model (e.g. spectral centroid, spectral decrease, etc.)
- Perceptual descriptors are computed from auditory-filtered bandwidth versions of signals which aim at approximating the human perception of sounds. This can be efficiently computed using Equivalent Rectangular Bandwidth (ERB) scale [10] which can be combined with gammatone filter-bank [3] (e.g. loudness, ERB spectral centroid, etc.)

In this study, we focus on the sound descriptors listed in table 2 which can be estimated using the timbre toolbox [2] and detailed in [12]. All descriptors are computed for each analyzed sound excerpt and may return null values. The harmonic descriptors of polyphonic sounds are computed using the prominent detected F_0 candidate (single F_0 estimation). To normalize the duration of analyzed sound, we separated each excerpt in 10-seconds length segments without distinction of silence or pitch events. Thus, each segment is represented by a real vector where the corresponding time series of each descriptor is summarized by a statistic. The median and the Inter Quartile Range (IQR) statistics were chosen for their robustness to outliers.

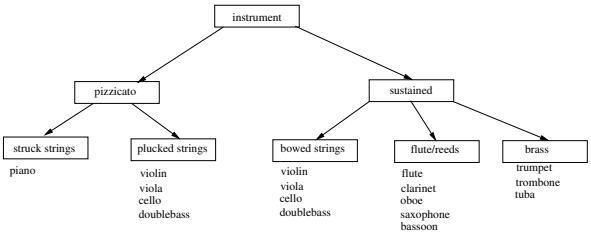
Acronym	Descriptor name	#
Att	Attack duration (see ADSR model [15])	1
AttSlp	Attack slope (ADSR)	1
Dec	Decay duration (ADSR)	1
DecSlp	Decay slope (ADSR)	1
Rel	Release duration (ADSR)	1
LAT	Log Attack Time	1
Tcent	Temporal centroid	1
Edur	Effective duration	1
FreqMod, AmpMod	Total energy modulation (frequency,amplitude)	2
RMSenv	RMS envelope	2
ACor	Signal Auto-Correlation function (12 first coef.)	24
ZCR	Zero-Crossing Rate	2
HCent	Harmonic spectral centroid	2
HSprd	Harmonic spectral spread	2
HSkew	Harmonic skewness	2
HKurt	Harmonic kurtosis	2
HSlp	Harmonic slope	2
HDec	Harmonic decrease	2
HRoff	Harmonic rolloff	2
HVar	Harmonic variation	2
HErg, HNErg, HFErg,	Harmonic energy, noise energy and frame energy	6
HNois	Noisiness	2
HF0	Fundamental frequency F_0	2
HinH	Inharmonicity	2
HTris	Harmonic tristimulus	6
HodevR	Harmonic odd to even partials ratio	2
Hdev	Harmonic deviation	2
SCent, ECent	Spectral centroid of the magnitude and energy spectrum	4
SSprd, ESprd	Spectral spread of the magnitude and energy spectrum	4
SSkew, ESkew	Spectral skewness of the magnitude and energy spectrum	4
SKurt, EKurt	Spectral kurtosis of the magnitude and energy spectrum	4
SSlp, ESlp	Spectral slope of the magnitude and energy spectrum	4
SDec, EDec	Spectral decrease of the magnitude and energy spectrum	4
SRoff, ERoff	Spectral rolloff of the magnitude and energy spectrum	4
SVar, EVar	Spectral variation of the magnitude and energy spectrum	4
SFErg, EFErg	Spectral frame energy of the magnitude and energy spectrum	4
Sflat, ESflat	Spectral flatness of the magnitude and energy spectrum	4
Scre, EScre	Spectral crest of the magnitude and energy spectrum	4
ErbCent, ErbGCent	ERB scale magnitude spectrogram / gammatone centroid	4
ErbSprd, ErbGSprd	ERB scale magnitude spectrogram / gammatone spread	4
ErbSkew, ErbGSkew	ERB scale magnitude spectrogram / gammatone skewness	4
ErbKurt, ErbGKurt	ERB scale magnitude spectrogram / gammatone kurtosis	4
ErbSlp, ErbGSlp	ERB scale magnitude spectrogram / gammatone slope	4
ErbDec, ErbGDec	ERB scale magnitude spectrogram / gammatone decrease	4
ErbRoff, ErbGRoff	ERB scale magnitude spectrogram / gammatone rolloff	4
ErbVar, ErbGVar	ERB scale magnitude spectrogram / gammatone variation	4
ErbFErg, ErbGFErg	ERB scale magnitude spectrogram / gammatone frame energy	4
ErbSflat, ErbGSflat	ERB scale magnitude spectrogram / gammatone flatness	4
ErbScre, ErbGScre	ERB scale magnitude spectrogram / gammatone crest	4
Total		164

Table 2. Acronym, name and number of the used timbre descriptors.

[2] MATLAB code available at http://www.cirmmt.org/research/tools



(a) Hornbostel and Sachs taxonomy (T1)



(b) Musician's instrument taxonomy (T2)

Figure 1. Taxonomies used for the automatic classification of musical instruments as proposed by Hornbostel and Sachs taxonomy in [16] (a) and Peeters in [13] (b).

3.2 Classification taxonomy

In this study, we use two databases which can be annotated using different taxonomies. Due to its diversity, the CREM database was only annotated using the Hornbostel and Sachs taxonomy [16] (T1) illustrated in Figure 1(a) which is widely used in ethnomusicology. This hierarchical taxonomy is general enough to classify uncommon instruments (e.g. struck bamboo) and conveys information about sound production materials and playing styles. From an another hand, the Iowa musical instruments database [5] used in our experiments was initially annotated using a musician's instrument taxonomy (T2) as proposed in [13] and illustrated in Figure 1(b). This database is composed of common western pitched instruments which can easily be annotated using T1 as described in Table 3. One can notice that the Iowa database is only composed of aerophones and cordophones instruments. If we consider the playing style, only 4 classes are represented if we apply T1 taxonomy to the Iowa database.

T1 class name	T2 equivalence	Duration (s)	#
aero-blowed	reed/flute and brass	5,951	668
cordo-struck	struck strings	5,564	646
cordo-plucked	plucked strings	5,229	583
cordo-bowed	bowed strings	7,853	838
Total		24,597	2,735

Table 3. Content of the Iowa database using musician's instrument taxonomy (T2) and equivalence with the Hornbostel and Sachs taxonomy (T1).

4. AUTOMATIC INSTRUMENT TIMBRE CLASSIFICATION METHOD

The described method aims at estimating the corresponding taxonomy class name of a given input sound.

4.1 Method overview

Here, each sound segment (cf. Section 3.1) is represented by vector of length $p = 164$ where each value corresponds to a descriptor (see Table 2). The training step of this method (illustrated in Figure 2) aims at modeling each timbre class using the best projection space for classification. A features selection algorithm is first applied to efficiently reduce the number of descriptors to avoid statistical overlearning. The classification space is computed using discriminant analysis which consists in estimating optimal weights over the descriptors allowing the best discrimination between timbre classes. Thus, the classification task consists in projecting an input sound into the best classification space and to select the most probable timbre class using the learned model.

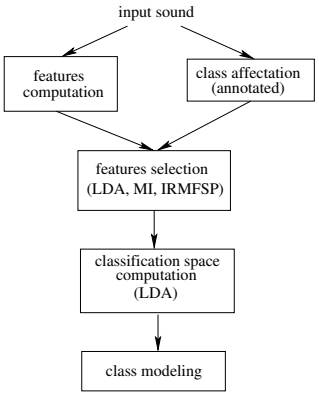


Figure 2. Training step of the proposed method.

4.2 Linear discriminant analysis

The goal of Linear Discriminant Analysis (LDA) [1] is to find the best projection or linear combination of all descriptors which maximizes the average distance between classes (inter-class distance) while minimizing distance between individuals from the same class (intra-class distance). This method assumes that the class affectation of each individual is a priori known. Its principle can be described as follows. First consider the $n \times p$ real matrix M where each row is a vector of descriptors associated to a sound (individual). We assume that each individual is a member of a unique class $k \in [1, K]$. Now we define W as the intra-class variance-covariance matrix which can be estimated by:

$$W = \frac{1}{n} \sum_{k=1}^{K} n_k W_k, \tag{1}$$

where W_k is the variance-covariance matrix computed from the $n_k \times p$ sub-matrix of M composed of the n_k individuals included into the class k.

We also define B the inter-class variance-covariance matrix expressed as follows:

$$B = \frac{1}{n} \sum_{k=1}^{K} n_k (\mu_k - \mu)(\mu_k - \mu)^T, \tag{2}$$

where μ_k corresponds to the mean vector of class k and μ is the mean vector of the entire dataset. According to [1], it can be shown that the eigenvectors of matrix $D = (B + W)^{-1}B$ solve this optimization problem. When the matrix $A = (B + W)$ is not invertible, a computational solution consists in using pseudoinverse of matrix A which can be calculated using $A^T(AA^T)^{-1}$.

4.3 Features selection algorithms

Features selection aims at computing the optimal relevance of each descriptor which can be measured with a weight or a rank. The resulting descriptors subset has to be the most discriminant as possible with the minimal redundancy. In this study, we investigate the three approaches described below.

4.3.1 LDA features selection

The LDA method detailed in Section 4.2 can also be used for selecting the most relevant features. In fact, the computed eigenvectors which correspond to linear combination of descriptors convey a relative weight applied to each descriptor. Thus, the significance (or weight) S_d of a descriptor d can be computed using a summation over a defined range $[1, R]$ of the eigenvectors of matrix D as follows:

$$S_d = \sum_{r=1}^{R} |v_{r,d}|, \tag{3}$$

where $v_{r,d}$ is the d-th coefficient of the r-th eigenvector associated to the eigenvalues sorted by descending order (i.e. $r = 1$ corresponds to the maximal eigenvalue of matrix D). In our implementation, we fixed $R = 8$.

4.3.2 Mutual information

Features selection algorithms aim at computing a subset of descriptors that conveys the maximal amount of information to model classes. From a statistical point of view, if we consider classes and feature descriptors as realizations of random variables C and F. The relevance can be measured with the mutual information defined by:

$$I(C, F) = \sum_{c} \sum_{f} P(c, f) \frac{P(c, f)}{P(c)P(f)}, \tag{4}$$

where $P(c)$ denotes the probability of $C = c$ which can be estimated from the approximated probability density functions (pdf) using a computed histogram. According to Bayes theorem one can compute $P(c, f) = P(f|c)P(c)$ where $P(f|c)$ is the pdf of the feature descriptor value f into class c. This method can be improved using [2] by reducing simultaneously the redundancy by considering the mutual information between previously selected descriptors.

4.3.3 Inertia Ratio Maximisation using features space projection (IRMFSP)

This algorithm was first proposed in [11] to reduce the number of descriptors used by timbre classification methods. It consists in maximizing the relevance of the descriptors subset for the classification task while minimizing the redundancy between the selected ones. This iterative method ($\iota \leq p$) is composed of two steps. The first one selects at iteration ι the non-previously selected descriptor which maximizes the ratio between inter-class inertia and the total inertia expressed as follow:

$$\hat{d}^{(\iota)} = \arg\max_{d} \frac{\sum_{k=1}^{K} n_k(\mu_{d,k} - \mu_d)(\mu_{d,k} - \mu_d)^T}{\sum_{i=1}^{n}(f_{d,i}^{(\iota)} - \mu_d)(f_{d,i}^{(\iota)} - \mu_d)^T}, \tag{5}$$

where $f_{d,i}^{(\iota)}$ denotes the value of descriptor $d \in [1, p]$ affected to the individual i. $\mu_{d,k}$ and μ_d respectively denote the average value of descriptor d into the class k and for the total dataset. The second step of this algorithm aims at orthogonalizing the remaining data for the next iteration as follows:

$$f_d^{(\iota+1)} = f_d^{(\iota)} - \left(f_d^{(\iota)} \cdot g_{\hat{d}}\right) g_{\hat{d}} \ \forall d \neq \hat{d}^{(\iota)}, \tag{6}$$

where $f_{\hat{d}}^{(\iota)}$ is the vector of the previously selected descriptor $\hat{d}^{(\iota)}$ for all the individuals of the entire dataset and $g_{\hat{d}} = f_{\hat{d}}^{(\iota)} / \| f_{\hat{d}}^{(\iota)} \|$ is its normalized form.

4.4 Class modeling and automatic classification

Each instrument class is modeled into the projected classification space resulting from the application of LDA. Thus, each class can be represented by its gravity center $\hat{\mu}_k$ which corresponds to the vector of the averaged values of the projected individuals which compose the class k. The classification decision which affect a class $\hat{k}$ to an input sound represented by a projected vector $\hat{x}$ is simply performed by minimizing the Euclidean distance with the gravity center of each class as follows:

$$\hat{k} = \arg\min_{k} \|\hat{\mu}_k - \hat{x}\|_2 \ \ \forall k \in [1, K], \tag{7}$$

where $\|v\|_2$ denotes the l_2 norm of vector v. Despite its simplicity, this method seems to obtain good results comparable with those of the literature [12].

5. EXPERIMENTS AND RESULTS

In this section we present the classification results obtained using the proposed method described in Section 4.

5.1 Method evaluation based on self database classification

In this experiment, we evaluate the classification of each distinct database using different taxonomies. We applied the 3-fold cross validation methodology which consists in partitioning the database in 3 distinct random subsets composed with 33% of each class (no collision between sets). Thus, the automatic classification applied on each subset is based on training applied on the remaining 66% of the

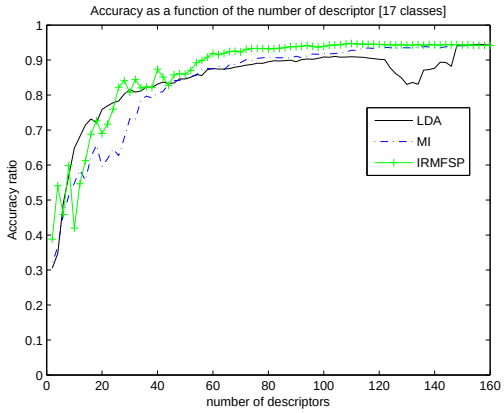
database. Figure 5.1 compares the classification accuracy obtained as a function of the number of used descriptors. The resulting confusion matrix of the CREM database using 20 audio descriptors is presented in Table 4 and shows an average classification accuracy of 80% where each instrument is well classified with a minimal accuracy of 70% for the aerophones. These results are good and seems comparable with those described in the literature [11] using the same number of descriptor. The most relevant feature descriptors (selected among the top ten) estimated by the IRMSFP and used for the classification task are detailed in Table 7. This result reveals significant differences between the two databases. As an example, harmonic descriptors are only discriminative for the CREM database but not for the Iowa database. This may be explained by the presence of membranophone in the CREM database which are not present in the Iowa database. Contrarily, spectral and perceptual descriptors seems more relevant for the Iowa database than for the CREM database. Some descriptors appear to be relevant for both database like the Spectral flatness (Sflat) and the ERB scale frame energy (ErbFErg) which describe the spectral envelope of signal.

	aero	c-struc	c-pluc	c-bowed	i-pluc	i-struc	i-clink	membr
aero	**70**	3	9	5		7		5
c-struc	6	**92**		3				
c-pluc	5	8	**73**	4		8		1
c-bowed			13	**80**	7			
i-pluc					**79**	14		7
i-struc	9	2	5		2	**79**		4
i-clink							**100**	
membr			11			17		**72**

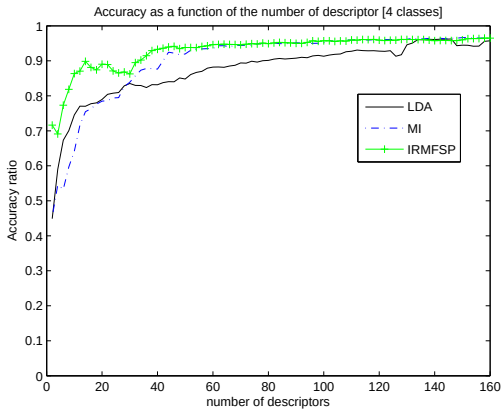
Table 4. Confusion matrix (expressed in percent of the sounds of the original class listed on the left) of the CREM database using the 20 most relevant descriptors selected by IRMSFP.

5.2 Cross-database evaluation

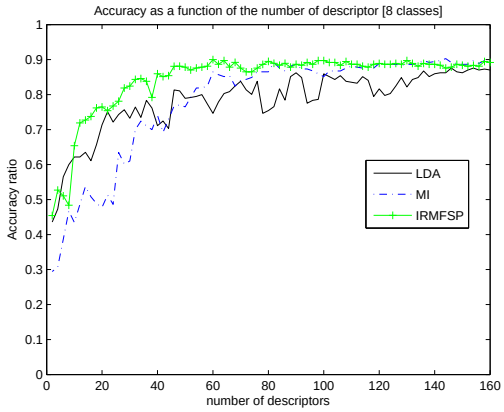
In this experiments (see Table 5), we merged the two databases and we applied the 3-fold cross validation method based on the T1 taxonomy to evaluate the classification accuracy on both database. The resulting average accuracy is about 68% which is lower than the accuracy obtained on the distinct classification of each database. The results of cross-database evaluation applied between databases using the T1 taxonomy are presented in Table 6 and obtain a poor average accuracy of 30%. This seems to confirm our intuition that the Iowa database conveys insufficient information to distinguish the different playing styles between the non-western cordophones instruments of the CREM database.



(a) Iowa database using T2



(b) Iowa database using T1



(c) CREM database using T1

Figure 3. Comparison of the 3-fold cross validation classification accuracy as a function of the number of optimally selected descriptors.

6. CONCLUSION AND FUTURE WORKS

We applied a computationally efficient automatic timbre classification method which was successfully evaluated on an introduced diversified database using an ethnomusicological taxonomy. This method obtains good classification results ($> 80\%$ of accuracy) for both evaluated databases which are comparable to those of the literature. However, the cross-database evaluation shows that each database cannot be used to infer a classification to the other. This can be explained by significant differences between these databases. Interestingly, results on the merged database obtain an acceptable accuracy of about 70%. As shown in previous work [11], our experiments confirm the efficiency of IRMFSP algorithm for automatic features selection applied to timbre classification. The interpretation of the

	aero	c-struc	c-pluc	c-bowed	i-pluc	i-struc	i-clink	membr
aero	**74**	14	5	3	2	1		
c-struc	12	**69**	10	5	1			2
c-pluc	1	7	**58**	29	1	2		2
c-bowed	3	6	33	**52**	1	3		
i-pluc		7		14	**79**			
i-struc	2	2	4	11	2	**51**		30
i-clink	11						**89**	
membr				6		17		**78**

Table 5. Confusion matrix (expressed in percent of the sounds of the original class listed on the left) of the evaluated fusion between the CREM and the Iowa database using the 20 most relevant descriptors selected by IRMSFP.

	aero	c-struc	c-pluc	c-bowed
aero	**72**	9	10	9
c-struc	12	**12**	34	42
c-pluc	23	47	**28**	3
c-bowed	28	34	24	**14**

Table 6. Confusion matrix (expressed in percent of the sounds of the original class listed on the left) of the CREM database classification based on Iowa database training.

CREM T1	Iowa T1	Iowa T2	CREM+Iowa T1
Edur **Acor**	**AttSlp** Dec	**AttSlp** **Acor** ZCR	AmpMod **Acor** RMSenv
Hdev Hnois HTris3			
Sflat	SFErg ERoff	**Sflat** SRoff SSkew	**Sflat** SVar SKurt Scre
ErbGKurt	**ErbKurt** **ErbFErg** **ErbRoff** ErbSlp ErbGCent	ErbSprd **ErbFErg** ErbGSprd	 **ErbFErg** **ErbRoff**

Table 7. Comparison of the most relevant descriptors estimated by IRMFSP.

most relevant selected features shows a significant effect of the content of database rather than on the taxonomy. However the timbre modeling interpretation applied to timbre classification remains difficult. Future works will consist in further investigating the role of descriptors by manually constraining selection before the classification process.

7. ACKNOWLEDGMENTS

This research was partly supported by the French ANR *(Agence Nationale de la Recherche)* DIADEMS *(Description,Indexation, Acces aux Documents Ethnomusicologiques et Sonores)* project (ANR-12-CORD-0022).

8. REFERENCES

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Blackwell, New York, USA, 1958.

[2] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. on Neural Networks*, 5(4):537–550, Jul. 1994.

[3] E.Ambikairajah, J. Epps, and L. Lin. Wideband speech and audio coding using gammatone filter banks. In *Proc. IEEE ICASSP'01*, volume 2, pages 773–776, 2001.

[4] N. F. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer-Verlag, 1998.

[5] L. Fritts. Musical instrument samples. Univ. Iowa Electronic Music Studios, 1997. [Online]. Available: http://theremin.music.uiowa.edu/MIS.html.

[6] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. Rwc music database: Music genre database and musical instrument sound database. In *Proc. ISMIR*, pages 229–230, Oct. 2003.

[7] J. M. Grey and J. W. Gordon. Perceptual effects of spectral modifications on musical timbre. *Journal of Acoustic Society of America (JASA)*, 5(63):1493–1500, 1978.

[8] S. McAdams, S. Winsberg, S. Donnadieu, G. Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, 1995.

[9] N. Misdariis, K. Bennett, D. Pressnitzer, P. Susini, and S. McAdams. Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres. In *Proc. ICA & ASA*, volume 103, Seattle, USA, Jun. 1998.

[10] B.C.J. Moore and B.R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74:750–753, 1983.

[11] G. Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with intertia ratio maximization. In *115th convention of AES*, New York, USA, Oct. 2003.

[12] G. Peeters, B. Giordano, P. Susini, N. Misdariis, and S. McAdams. The timbre toolbox: Audio descriptors of musical signals. *Journal of Acoustic Society of America (JASA)*, 5(130):2902–2916, Nov. 2011.

[13] G. Peeters and X. Rodet. Automatically selecting signal descriptors for sound classification. In *Proc. ICMC*, Göteborg, Sweden, 2002.

[14] E. Schubert, J. Wolfe, and A. Tarnopolsky. Spectral centroid and timbre in complex, multiple instrumental textures. In *Proc. 8th Int. Conf. on Music Perception & Cognition (ICMPC)*, Evanston, Aug. 2004.

[15] G. Torelli and G. Caironi. New polyphonic sound generator chip with integrated microprocessor-programmable adsr envelope shaper. *IEEE Trans. on Consumer Electronics*, CE-29(3):203–212, 1983.

[16] E. v. Hornbostel and C. Sachs. The classification of musical instruments. *Galpin Society Journal*, 3(25):3–29, 1961.