COMPARISON OF AUDIO FEATURES FOR RECOGNITION OF WESTERN AND ETHNIC INSTRUMENTS IN POLYPHONIC MIXTURES

Igor Vatolkin **Günter Rudolph**
TU Dortmund, Department of Computer Science
{igor.vatolkin;guenter.rudolph}@tu-dortmund.de

ABSTRACT

Studies on instrument recognition are almost always restricted to either Western or ethnic music. Only little work has been done to compare both musical worlds. In this paper, we analyse the performance of various audio features for recognition of Western and ethnic instruments in chords. The feature selection is done with the help of a minimum redundancy - maximum relevance strategy and a multi-objective evolutionary algorithm. We compare the features found to be the best for individual categories and propose a novel strategy based on non-dominated sorting to evaluate and select trade-off features which may contribute as best as possible to the recognition of individual and all instruments.

1. INTRODUCTION

Instrument recognition in polyphonic audio signals, when acoustic properties of multiple simultaneously played sources contribute together to the spectrum, corresponds to a very challenging problem in music information retrieval. An unknown number of sound sources, instrument bodies with different characteristics, dissimilarities of overtone distribution across pitches, various playing styles, applied effects, etc. hinder the robust identification of playing instruments. Earlier works started with recognition of individual tones [6, 15]. Several years later first studies on polyphonic instrument recognition were published [5, 7]. In further works, many different and complex methods were proposed, like source separation [14], complex feature engineering [27], or deep neural networks [12].

However, most studies concentrate on the detection of Western instruments in Western classical or popular music. Recently, more attention was paid to analyse also ethnic/world music, for example for onset detection in Carnatic music [22] or rhythm analysis in Indian music [23], but only little work was reported on recognition of ethnic instruments, in particularly in polyphonic recordings, or in both Western and ethnic recordings. [10] presented a study on recognition of 10 Indian stringed, wind, and percussive instruments (sitar, edakkai, indian flute, etc.), however only two instruments were mixed together at the same time. Classification of solo recordings into three families (string, woodwind, percussion) of 9 Pakistani instruments (benju, bainsuri, tabla, etc.) was done in [17]. [2] examined properties of various acoustic features for recognition of five Hindustani instruments. In [25], the performance of models trained for recognition of Western instruments in Western mixtures was validated when applied for polyphonic mixtures with ethnic samples.

The goal of our study was to propose a strategy how audio features can be automatically validated for their ability to detect Western and/or ethnic instruments. We adopt the general experiment setup from [25], starting with a large feature set and applying feature selection for identification of the most relevant features. However, in contrast to [25], we aim at the recognition of not only Western, but also ethnic instruments, and incorporate datasets created with both Western and ethnic samples, making recognition tasks harder, but also allowing the classification models to be more robust and not restricted to data sets created with more similar instruments. Furthermore, we propose a novel strategy based on non-dominated sorting to identify features which are particularly well suited to classify either Western, ethnic, or both categories of instruments.

The remainder of this paper is organised as follows. In Section 2, we introduce the backgrounds of multi-objective feature selection. Section 3 describes the study setup. In Section 4, we discuss the results, compare the features, and present our strategy how the most relevant features for the recognition of Western, ethnic, and both groups of instruments can be identified. We conclude with Section 5.

2. MULTI-OBJECTIVE FEATURE SELECTION

The goal of feature selection (FS) is to identify relevant features and to remove irrelevant and redundant ones. Relevant features contribute to the "best" classification models, so that their removal would decrease the classification quality. Irrelevant features do not capture any important properties of classification categories; if too many of such features are contained in the data set, some of them may be identified by chance as relevant, leading to decreased generalisation ability of models. Redundant features can be removed from the feature set without decrease of classification quality for models trained with this set, because other features already describe the same properties. For a good introduction into feature selection, we refer to [11].

To evaluate feature sets, some criterion is needed, like classification accuracy, or correlation with the target. In the multi-objective feature selection (MO-FS), several of such criteria are optimised simultaneously:

$$\mathbf{q}^* = \arg\min_{\mathbf{q}} \{ m_i (\mathbf{y}, \hat{\mathbf{y}}, \Phi(\mathcal{F}, \mathbf{q})) : i = 1, ..., K \}, \quad (1)$$

where $\mathcal{F}$ is the complete feature set, $\Phi(\mathcal{F}, \mathbf{q})$ is the selected feature set, $\mathbf{q}$ is the binary vector which indicates features to be selected (zero entry at position i means that i-th feature is not selected), $\mathbf{y}$ are correct labels (categories to predict), $\hat{\mathbf{y}}$ are predicted labels, and $m_1, ..., m_K$ are K evaluation or *objective* functions, which may measure classification performance (accuracy, precision, recall, etc.) but also other relevant criteria (number of selected features, degree of internal redundancies across features, etc.)

In [25], it is proposed to minimise two criteria: the number of selected features and the balanced classification error, which is defined as follows:

$$e = \frac{1}{2} \left(\frac{FN}{TP + FN} + \frac{FP}{TN + FP} \right), \quad (2)$$

where TP is the number of true positives, TN true negatives, FP false positives, and FN false negatives.

When MO-FS is applied, some feature sets cannot be compared: consider a smaller feature set, which leads to a higher classification error, and a larger feature set, which leads to a lower error; none of these sets can be described as superior to another one. However, some sets may be worse with regard to both criteria, and can be identified with the help of non-dominance relation: feature set $\mathbf{q}_1$ dominates feature set $\mathbf{q}_2$ ($\mathbf{q}_1 \prec \mathbf{q}_2$), if and only if

$$\begin{aligned} &\forall i \in \{1, ..., K\} : m_i(\mathbf{q}_1) \leq m_i(\mathbf{q}_2) \text{ and} \\ &\exists j \in \{1, ..., K\} : m_j(\mathbf{q}_1) < m_j(\mathbf{q}_2). \end{aligned} \quad (3)$$

In other words, $\mathbf{q}_1$ dominates $\mathbf{q}_2$, when it is not worse than $\mathbf{q}_2$ with regard to all criteria, and is better with regard to at least one criterium. Here, we restrict us to minimisation of all criteria; criteria to be maximised can be simply redefined for minimisation (e.g., multiplying them with -1). The goal of MO-FS is to find a *non-dominated front* of incomparable feature sets, which are not dominated by any other feature set.

3. EXPERIMENTAL SETUP

In the experimental setup, we mostly follow [25]. Table 1 lists all instruments used in this study. The instruments which were recognised in classification experiments, are written in normal font. Further instruments, which were present in audio mixtures, but were not considered as classes to be identified, are written in italic font. The Western instrument samples are taken from MUMS [4], RWC [9], and University of Iowa [1] databases, and ethnic from Ethno World 5 Professional & Voices [2].

The chords are randomly mixed from individual samples as described in [25], however, in contrast to previous work, we have created heterogeneous data sets, so that in each mixture of three to four tones at least one Western and at least one ethnic sample is contained. The *experiment* set consists of 3000 chords. During feature selection, it is divided into the training and optimisation set by means of 5-fold cross-validation. *Training* set is used to train classification models, and *optimisation* set to estimate the balanced classification error e. Other independently mixed 3000 chords are used as *holdout* set to measure the effect of overfitting towards the experiment set.

The audio features comprise acoustic characteristics available for extraction with open-source AMUSE framework [26], including mel frequency cepstral coefficients (MFCCs) [21], root mean square (RMS) of the time signal, various spectral characteristics (centroid, bandwidth, kurtosis, skewness, flux, etc.), chroma [8] and chroma energy normalized statistics (CENS) [20], but also other less frequently used features like characteristics of ERB bands and Bark scale domain [19] or phase domain [18]. Before the extraction, the audio was downsampled to 22,1 kHz mono signal, and the most short-framed features were extracted from 512 samples without overlap; for exact details see [24]. For each feature, three dimensions are stored separately: a value from the middle of the attack interval, from the onset frame (the end of the attack interval equal to the beginning of the release interval), and from the middle of the release interval, where attack and release intervals were previously extracted with MIR Toolbox [16], leading to the complete set of 795 feature dimensions. Classification is done with random forest classifier [13].

The first FS strategy was minimum redundancy-maximum relevance (MRMR) [3], which aims at the minimisation of redundancy between selected features and maximisation of relevance to the target category. The second FS strategy was evolutionary multi-objective feature selection (EMO-FS) with $\mathcal{S}$-metric selection evolutionary multi-objective algorithm (SMS-EMOA) [1], for further details please see [25].

4. DISCUSSION OF RESULTS

4.1 Performance Analysis

Table 2 provides the summary of results after feature selection. $e_H(\Phi)$ denotes the baseline classification error using the complete feature set. $|\widehat{\Phi}_O|$ denotes the cardinality of the feature set with the smallest optimisation error $e_O(\widehat{\Phi}_O)$. $e_H(\widehat{\Phi}_O)$ denotes the holdout error for that feature set, and $e_H(\widehat{\Phi}_H)$ the best holdout error among all output feature sets after feature selection.

Both MRMR and EMO-FS significantly outperform the baseline method which trains models with all features. This means, that FS explicitly makes sense. As it can be

[1] http://theremin.music.uiowa.edu/MIS.html

[2] http://www.bestservice.de

Category	Instruments
WESTERN	
Bowed	Cello, viola, violin
Key	Piano
Stringed	Acoustic guitar, electric guitar
Woodwind/brass	Flute, trumpet
ETHNIC	
Bowed	Dilruba, egyptian fiddle, erhu, *jinghu opera violin*, *morin khuur violin*
Key	Hohner melodica, scale changer harmonium
Stringed	Balalaika, bandura, banjolin, banjo framus, *bouzouki*, *ceylon guitar*, *cümbüs*, *domra*, *kantele*, *oud*, *sitar*, *tampura*, *tanbur*, *saz*, *ukulele*
Woodwind/brass	Bawu, dung dkar trumpet, fujara, *pan flute*, *pinkillo*, *pivana*, *shakuhachi*

Table 1: Instruments used in this study.

	No FS	MRMR				EMO-FS			
Task	$e_H(\Phi)$	$\lvert\widehat{\Phi}_O\rvert$	$e_O(\widehat{\Phi}_O)$	$e_H(\widehat{\Phi}_O)$	$e_H(\widehat{\Phi}_H)$	$\lvert\widehat{\Phi}_O\rvert$	$e_O(\widehat{\Phi}_O)$	$e_H(\widehat{\Phi}_O)$	$e_H(\widehat{\Phi}_H)$
WESTERN									
Acoustic guitar	0.4395	10	0.3962	0.3906	0.3906	46	0.3809	0.3885	0.3751
Cello	0.4696	12	0.4295	0.4382	0.4382	37	0.4358	0.4574	0.4404
Electric guitar	0.1704	309	0.1532	0.1424	0.1369	44	0.1238	0.1158	0.1084
Flute	0.4907	3	0.4485	0.4651	0.4516	45	0.4513	0.4675	0.4547
Piano	0.2531	7	0.2148	0.2411	0.2260	54	0.2112	0.2320	0.2237
Trumpet	0.3119	20	0.2516	0.2538	0.2506	64	0.2706	0.2604	0.2488
Viola	0.4968	13	0.4735	0.4857	0.4687	36	0.4417	0.4561	0.4406
Violin	0.4791	8	0.4518	0.4639	0.4632	55	0.4538	0.4605	0.4504
ETHNIC									
Balalaika	0.3976	34	0.3070	0.2987	0.2821	48	0.3113	0.3226	0.2931
Bandura	0.5000	2	0.4653	0.4893	0.4587	50	0.4713	0.4809	0.4689
Banjo framus	0.4909	11	0.3915	0.4252	0.4151	39	0.4184	0.4139	0.3994
Banjolin	0.4827	18	0.3504	0.3895	0.3895	32	0.3661	0.4012	0.3937
Bawu	0.3776	12	0.1814	0.2150	0.1836	66	0.2028	0.2138	0.1963
Dilruba	0.4492	32	0.3769	0.3987	0.3974	59	0.3297	0.3761	0.3503
Dung dkar	0.4213	47	0.3971	0.3487	0.3487	49	0.3478	0.3373	0.2983
Egyptian fiddle	0.3533	79	0.2387	0.2750	0.2420	57	0.1763	0.1984	0.1692
Erhu	0.4507	12	0.3719	0.3766	0.3672	42	0.3609	0.3489	0.3293
Fujara	0.3061	28	0.1945	0.1887	0.1840	51	0.1994	0.2236	0.1959
Melodica	0.3889	33	0.2721	0.3041	0.2926	48	0.2638	0.2898	0.2633
Scale changer harmonium	0.3192	22	0.2342	0.2590	0.2445	43	0.2263	0.2578	0.2149

Table 2: Results after feature selection, details are explained in Section 4.1.

expected, $e_H(\widehat{\Phi}_O)$ is often higher than $e_O(\widehat{\Phi}_O)$. However, this difference is significant only for ethnic instruments and EMO-FS. The null hypothesis that both errors come from the same distribution is rejected by means of Wilcoxon signed rank test for paired observations (MATLAB function SIGNRANK) only for ethnic instruments/EMO-FS with p-value of 0.0210. For combination Western/EMO-FS, p = 0.1484, for Western/MRMR p = 0.1094 and ethnic/MRMR p = 0.0922. Both MRMR and EMO-FS are comparable: each method leads to a smaller $e_O(\widehat{\Phi}_O)$ for exactly a half of Western and a half of ethnic categories, and there is no significant difference between these errors after the application of Wilcoxon rank sum test for unpaired observations (MATLAB function RANKSUM).

4.2 Best Features for Individual Categories

To identify the most relevant features for each category, we may estimate feature ranks as follows. Let N_c be the number of solutions (feature sets) in the non-dominated front after the multi-objective optimisation for category c (recall that the non-dominated front contains the best incomparable solutions, cf. Section 2). Let $q_{k,i}$ be 1 when feature k in non-dominated solution i is selected, and 0 when this feature is not selected. We may count the number of occurrences of feature k in the front and normalise this number by the size of the front:

$$r(c,k) = \frac{1}{N_c} \cdot \sum_{i=1}^{N_c} q_{k,i}. \tag{4}$$

Table 3 lists two most relevant features for individual classification tasks using either MRMR (columns 2-5) or EMO-FS (columns 6-9). As MRMR starts with the most relevant feature and only adds further features during the iteration process, $r = 1$ for all 1st best features in that case. Because of the differences in operating methods (EMO-FS explores a significantly larger number of feature sets, but is slower), the two most relevant features are usually not the same. However, for cello, flute, and erhu the best feature is exactly the same for MRMR and EMO-FS. For scale changer harmonium and bawu the 1st best feature for MRMR is the same as the 2nd best feature for EMO-FS.

One observation is that MFCCs seem to play a more important role for the recognition of Western instruments: although all processed MFCC dimensions together account for appr. 17% of the complete feature set, they correspond to 56.25% of all 16 entries for 1st best features (8 entries for each MRMR and EMO-FS) and 18.75% for 2nd best features for Western instruments, but only to 20.83% for

	MRMR				EMO-FS			
Task	**1st best feature**	r	**2nd best feature**	r	**1st best feature**	r	**2nd best feature**	r
WESTERN								
Ac. guitar	A(MFCC 2)	1	O(Bark scale magn. 6)	0.83	A(Phase domain angles)	0.80	A(Chroma 11)	0.60
Cello	O(MFCC 4)	1	A(Spectral slope)	0.89	O(MFCC 4)	0.50	A(RMS)	0.42
El. guitar	A(1st period. ampl. peak)	1	A(RMS ERB band 2)	0.95	R(Bark scale magn. 19)	0.58	R(Delta MFCC 3)	0.42
Flute	A(MFCC 3)	1	R(MFCC 6)	0.67	A(MFCC 3)	0.55	O(LPC 2)	0.36
Piano	R(MFCC 1)	1	O(RMS ERB band 2)	0.83	O(RMS ERB band 1)	0.67	A(Sum corr. components)	0.56
Trumpet	R(Ampl. 4th spectr. peak)	1	R(Inharmonicity)	0.92	R(Ampl. 5th spectr. peak)	0.73	R(MFCC 2)	0.64
Viola	O(Low energy)	1	A(CENS chroma 11)	0.80	O(MFCC 9)	0.82	R(RMS ERB band 1)	0.55
Violin	A(MFCC 1)	1	A(Var. aver. dist. betw. ZC)	0.80	O(MFCC 5)	0.60	A(RMS ERB band 2)	0.50
ETHNIC								
Balalaika	O(Bark scale magn. 21)	1	A(LPC 3)	0.89	O(Bark scale magn. 21)	0.73	O(LPC 3)	0.53
Bandura	A(Sub-band energy rat. 4)	1	R (MFCC 1)	0.50	A(MFCC 3)	0.86	A(Spectral kurtosis)	0.71
Banjo framus	O(Spectral flux)	1	A(LPC 4)	0.83	A(LPC 4)	0.89	A(ZC rate ERB band 6)	0.67
Banjolin	O(Bark scale magn. 23)	1	O(MFCC 8)	0.92	A(LPC 2)	0.80	O(Sum corr. components)	0.80
Bawu	A(RMS peak num. above mean ampl.)	1	O(Bark scale magn. 6)	0.83	O(RMS peak number)	0.56	A(RMS peak num. above mean ampl.)	0.56
Dilruba	O(MFCC 3)	1	O(MFCC 1)	0.88	A(Spectral extent)	0.88	O(RMS ERB band 2)	0.75
Dung dkar	O(Spectral kurtosis)	1	R(ZC rate ERB band 2)	0.89	R(Ampl. 1st spectr. peak)	0.88	R(LPC 1)	0.63
Egypt. fiddle	A(Phase domain angles)	1	A(MFCC 7)	0.94	A(Spectral flux))	0.69	O(Bark scale magn. 23)	0.62
Erhu	R(MFCC 4)	1	A(RMS peak number)	0.80	R(MFCC 4)	0.70	R(RMS)	0.60
Fujara	A(Max. ampl. chroma)	1	O(Ampl. 4th spectr. peak)	0.91	R(RMS)	0.90	O(Spectral flatness 4)	0.70
Melodica	R(RMS ERB band 8)	1	A(Spectral bandwidth)	0.92	R(MFCC 2)	0.56	A(ZC rate ERB band 6)	0.44
Scale changer harmonium	R(LPC 5)	1	A(MFCC 1)	0.91	R(Phase domain angles)	0.60	R(LPC 5)	0.50

Table 3: Ranks of features for categorisation of individual instruments. A(·): features from middles of attack intervals; O(·): from onset frames; R(·): from middles of release intervals. LPC: linear prediction coefficient; ZC: zero-crossings.

1st and 20.83% for 2nd best ethnic features. Among ethnic instruments, the half of all MFCC occurrences belongs to bowed instruments (dilruba, egyptian fiddle, erhu), and other two belong to key instruments (melodica and scale changer harmonium). This leads to a careful suggestion that the mel spectrum is probably not the best feature domain for ethnic stringed and brass instruments, which deserves further investigations. Among two most relevant features for ethnic instruments, particularly LPCs occur rather frequently (8 times / 16.7% of all entries against 1 time / 3.13% for Western instruments).

With regard to attack/onset/release-envelope, we may observe, that all three extraction frame categories appear frequently in Table 3. However, the attack phase seems to be generally more relevant for all instruments (for Western instruments, 43.75% of all entries belong to features stored from middles of attack intervals, for ethnic, 39.58%). Onset features correspond to 28.13% of Western and 33.33% of ethnic entries, and release features to 28.13% of Western and 27.08% of ethnic entries.

4.3 Best Features for Western and Ethnic Instruments

To compare the significance of features for all Western instruments against all ethnic instruments, we may estimate mean feature ranks across all categories of the same "world". Let M_W be the number of Western instruments ($M_W = 8$) and M_E of ethnic instruments ($M_E = 12$). Let $q_{c,k,i}$ be 1, when feature k is selected in i-th non-dominated solution of the category c, and 0, when it is not selected. Then, the accumulated rank of feature k for all Western instruments is calculated as:

$$r(W,k) = \frac{1}{M_W} \cdot \sum_{c=1}^{M_W} \left(\frac{1}{N_c} \cdot \sum_{i=1}^{N_c} q_{c,k,i} \right), \qquad (5)$$

and, similarly, the accumulated rank of feature k for all ethnic instruments as:

$$r(E,k) = \frac{1}{M_E} \cdot \sum_{c=1}^{M_E} \left(\frac{1}{N_c} \cdot \sum_{i=1}^{N_c} q_{c,k,i} \right), \qquad (6)$$

The accumulated rank corresponds to the relative share of selections of a feature k among all non-dominated solutions for all categories of the same world.

Table 4 lists top 10 features for Western and ethnic categories. Additionally to non-dominated fronts from the optimisation set (columns 1,2,5,6), we analyse the importance of features for the independent holdout set (columns 3,4,7,8). As we can observe, the best features for the optimisation set are often the same as for the holdout set, which supports the suggestion, that those features are well suitable for different data sets. Again, we see, that with regard to accumulated ranks, MFCCs appear rather often for Western categories (8 entries) than for ethnic categories (6 entries), and LPCs rather often for ethnic categories (6 entries vs. no entry). EMO-FS selected often RMS for ERB bands among top 10 Western features (9 of 20 corresponding entries vs. no entry for ethnic categories).

To measure the statistical difference between importances of features for both worlds, we validate the following statistical hypothesis H0: given the accumulated ranks of top 20 features of one world, we assume that the ranks of the same features but for another world belong to the same distribution. As Western and ethnic categorisation tasks are independent, we validate this hypothesis by means of Wilcoxon rank sum test (RANKSUM function in

MRMR				EMO-FS			
Optimisation set	r	**Holdout set**	r	**Optimisation set**	r	**Holdout set**	r
TOP 10 FEATURES FOR WESTERN CATEGORIES							
A(1st period. ampl. peak)	0.228	R(1st period. ampl. peak)	0.259	R(RMS ERB band 1)	0.182	A(RMS ERB band 2)	0.258
A(MFCC 1)	0.211	A(1st period. ampl. peak)	0.251	A(RMS ERB band 2)	0.177	R(RMS ERB band 1)	0.221
O(Low energy)	0.178	A(MFCC 1)	0.227	A(Max. ampl. chroma)	0.172	A(Max. ampl. chroma)	0.185
R(1st period. ampl. peak)	0.177	O(Low energy)	0.195	A(Phase domain angles)	0.169	R(Var. aver. dist. betw. ZC)	0.176
R(ZC rate ERB band 1)	0.169	A(MFCC 3)	0.191	O(Bark scale magnitude 1)	0.156	R(RMS ERB band 2)	0.173
A(MFCC 3)	0.164	A(Spectral slope)	0.178	A(MFCC 3)	0.152	A(Ampl. 1st spectral peak)	0.167
A(Spectral slope)	0.151	R(MFCC 3)	0.159	A(Sum corr. components)	0.147	O(RMS ERB band 1)	0.157
R(Inharmonicity)	0.135	R(Spectr. centroid ERB 8)	0.150	O(RMS ERB band 1)	0.148	A(Phase domain angles)	0.144
A(ZC rate ERB band 10)	0.135	R(Low energy)	0.144	A(Ampl. 1st spectr. peak)	0.143	O(RMS ERB band 2)	0.143
R(MFCC 1)	0.132	R(MFCC 1)	0.141	R(RMS ERB band 2)	0.141	O(CENS chroma 6)	0.142
ETHNIC							
A(Low energy)	0.203	A(Low energy)	0.155	R(Ampl. 1st spectral peak)	0.199	R(Ampl. 1st spectral peak)	0.203
R(LPC 6)	0.158	A(Sub-band energy ratio 4)	0.148	O(Bark scale magnitude 21)	0.195	O(Bark scale magnitude 21)	0.178
A(Sub-band energy ratio 4)	0.146	R(LPC 6)	0.137	O(RMS peak number)	0.170	O(Spectral extent)	0.161
O(LPC 7)	0.130	A(Phase domain angles)	0.137	R(RMS)	0.170	O(LPC 4)	0.157
A(Phase domain angles)	0.129	O(MFCC 3)	0.135	A(Spectral bandwidth)	0.159	R(Bark scale magnitude 3)	0.154
O(Bark scale magnitudes 6)	0.125	O(Bark scale magnitude 6)	0.128	O(Spectral flux))	0.143	R(RMS)	0.153
O(MFCC 3)	0.121	R(ZC rate for ERB band 2)	0.119	R(RMS peak num. above mean ampl.)	0.141	A(Strength of 6.major key)	0.145
O(Bark scale magnitude 21)	0.104	O(Bark scale magnitude 21)	0.111	R(MFCC 2)	0.137	A(Inharmonicity)	0.135
A(LPC 3)	0.097	O(Spectr. centroid ERB 10)	0.109	R(MFCC 4)	0.137	A(MFCC 6)	0.130
O(Spectr. centroid ERB 10)	0.095	A(LPC 4)	0.108	A(Bark scale magnitude 17)	0.136	O(MFCC 1)	0.130

Table 4: Accumulated ranks of features for categorisation of Western and ethnic instruments. A($\cdot$): features from middles of attack intervals; O($\cdot$): from onset frames; R($\cdot$): from middles of release intervals. LPC: linear prediction coefficient; ZC: zero-crossings.

MATLAB). H0 is rejected in all cases for both feature selection strategies and both sets (optimisation/holdout). Table 5 contains p-values. This means that top 20 features which are particularly good for recognition of Western instruments are not similarly good for the recognition of ethnic instruments, and vice versa. However, please note that H0 is rejected only for a limited set of 8 Western and 12 ethnic instruments, even if they were carefully chosen to represent different instrument categories. Further studies with a significantly larger number of instruments may support or weaken this statement.

4.4 Best Features for All Categories

To provide generic recommendations on features which are particularly useful for the recognition of both Western and ethnic instruments, Figure 1 plots the accumulated ranks $r(W, k)$ and $r(E, k)$ of all features. Upper subfigures contain results for MRMR, bottom subfigures for EMO-FS, left subfigures correspond to optimisation set, and right subfigures to holdout set. Dashed lines divide the rank space in three regions. For features in the bottom right region, $r(W, k)$ is at least twice as large as $r(E, k)$. For features in the top left region, $r(E, k)$ is at least twice as large as $r(W, k)$. The ranks of features in the middle region are comparable for Western and ethnic instruments. As we are interested to identify features which are best suited for for the classification of all instruments, we marked the first non-dominated front with large filled circles and the second non-dominated front with small filled circles, supported with feature IDs. The mapping of IDs to feature names is provided in Table 6. For MRMR, the features belonging to first non-dominated fronts are low energy, MFCC 1, and 1st periodicity amplitude peak. For EMO-FS, these features are RMS, Bark scale magnitude 3, RMS for ERB bands 1 and 2, maximal amplitude in the chromagram, and amplitude of the 1st spectral peak.

It is worth to mention that even if our feature vector contains almost 800 dimensions, the features can be extracted from various frame lengths or with varying parameters, and further signal descriptors can be added. Further work is necessary to identify better features for instrument recognition, and our framework provides an automatic strategy to evaluate the suitability of features or their extraction parameters to classify instruments of different categories by means of non-dominance relation.

5. CONCLUSIONS

In this paper, we have applied two feature selection methods for recognition of Western and ethnic instruments in polyphonic audio mixtures. Both methods lead to a significant reduction of the classification error compared to models trained with all features. To measure the relevance of features for individual categories as well as for a set of 8 Western and 12 ethnic categories, we proposed a simple rank measure based on feature occurrence in non-dominated fronts, with the aim to simultaneously minimise the number of features and the classification error. Even if larger feature sets with a smaller error are usually preferable for classification scenarios, also small feature sets with higher errors give valueful insights into relevance of individual features. The statistical comparison of features best suited for recognition of Western instruments against features best suited for recognition of ethnic instruments showed that their performance is significantly different. This empirically supports the suggestion, that many acoustic descriptors developed and optimised for music instru-

	MRMR		EMO-FS	
H0	**Optimisation set**	**Holdout set**	**Optimisation set**	**Holdout set**
Top 20 Western features are similarly good for ethnic categories	5.69e-08	5.69e-08	2.21e-07	6.70e-08
Top 20 ethnic features are similarly good for Western categories	1.99e-04	1.11e-04	6.67e-06	2.66e-06

Table 5: p-values for comparison of top 20 Western and top 20 ethnic features represented by their accumulated ranks.

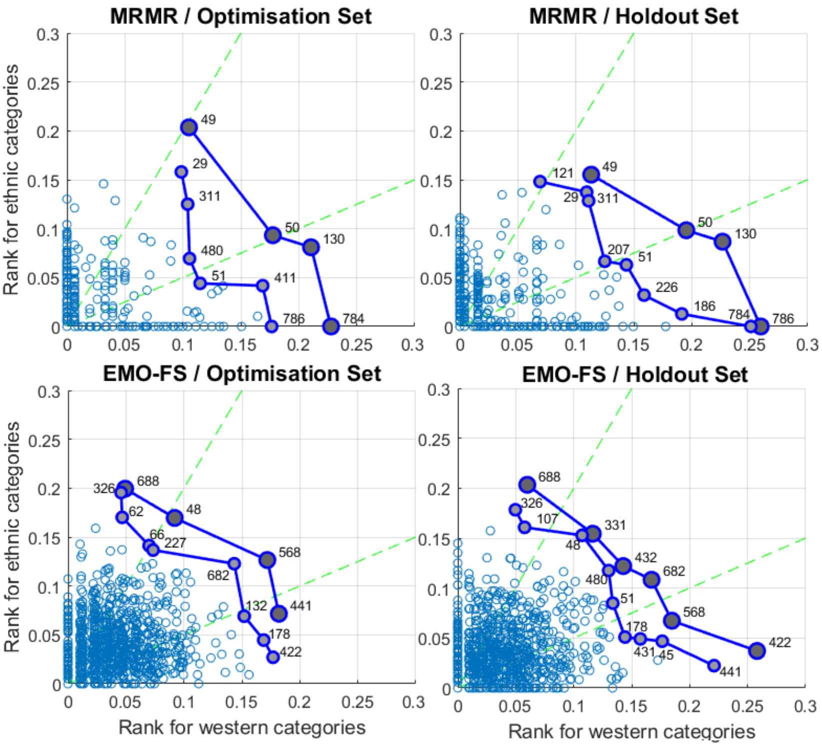

Figure 1: Best (large circles) and 2nd best (small circles) non-dominated features for both Western and ethnic categories. The fronts were estimated for the maximisation of accumulated ranks.

No.	Name	No.	Name	No.	Name
29	R(LPC 6)	132	A(MFCC 3)	422	A(RMS for ERB band 2)
45	R(Var. of aver. dist. between ZC)	178	A(Phase domain angles)	431	O(RMS for ERB band 1)
48	R(RMS)	186	A(MFCC 3)	432	O(RMS for ERB band 2)
49	A(Low energy)	207	O(MFCC 4)	441	R(RMS for ERB band 1)
50	O(Low energy)	226	R(MFCC 3)	480	R(Spectral centroid ERB band 10)
51	R(Low energy)	227	R(MFCC 4)	568	A(Max. ampl. chroma)
62	O(RMS peak number)	311	O(Bark scale magnitude 6)	682	A(Ampl. 1st spectral peak)
66	R(RMS peak number above mean ampl.)	326	O(Bark scale magnitude 21)	688	R(Ampl. 1st spectral peak)
107	O(Spectral extent)	331	R(Bark scale magnitude 3)	784	A(1st periodicity ampl. peak)
121	A(Sub-band energy ratio 4)	411	R(ZC rate for ERB band 1)	786	R(1st periodicity ampl. peak)
130	A(MFCC 1)				

Table 6: Names of features from two best fronts of Figure 1. A(·): features from middles of attack intervals; O(·): from onset frames; R(·): from middles of release intervals. LPC: linear prediction coefficient; ZC: zero-crossings.

ment recognition in Western music are not best suited for the recognition of ethnic instruments.

Another focus of our investigation was to identify those features which are particularly well suited for the recognition of both Western and ethnic instruments. This can be done by means of non-dominated sorting in the two-dimensional rank space. Even if the goal of identifying the best “compromise” features is somewhat contrary to the identification of the best specific features for Western and ethnic instruments, both approaches make sense. Keeping a nearly unlimited number of possible combinations of many world instruments with different effects and playing styles in mind, a good strategy is to start with a sufficiently large set of audio descriptors. In the second time-consuming optimisation step, more efforts can be spent for refining the extraction parameters of these features and development of further ones, which are particularly relevant for a concrete instrument class. With the help of our framework, both tasks can be executed and analysed automatically.

6. REFERENCES

[1] N. Beume, B. Naujoks, and M. Emmerich. Sms-emoa: Multiobjective selection based on dominated hypervolume. *European Journal of Operational Research*, 181(3):1653–1669, 2007.

[2] A. K. Datta, S. S. Solanki, R. Sengupta, S. Chakraborty, K. Mahto, and A. Patranabis. *Automatic Musical Instrument Recognition*, pages 167–232. Springer Singapore, Singapore, 2017.

[3] C. H. Q. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal on Bioinformatics and Computational Biology*, 3(2):185–206, 2005.

[4] T. Eerola and R. Ferrer. Instrument library (MUMS) revised. *Music Perception*, 25(3):253–255, 2008.

[5] J. Eggink and G. J. Brown. A missing feature approach to instrument identification in polyphonic music. In *Proc. of 2003 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 553–556. IEEE, 2003.

[6] A. J. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Proc. of IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 753–756. IEEE, 2000.

[7] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music. In *Proc. of 2005 IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 245–248. IEEE, 2005.

[8] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *Proc. of the International Computer Music Conference (ICMC)*, pages 464–467, 1999.

[9] M. Goto, H. Hashiguchi, T.i Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. In *Proc. of the 4th Int'l Conf. on Music Information Retrieval (ISMIR)*, pages 229–230, 2003.

[10] S. Gunasekaran and K. Revathy. Fractal dimension analysis of audio signals for indian musical instrument recognition. In *Proc. of the Int'l Conf. on Audio, Language and Image Processing (ICALIP)*, pages 257–261, 2008.

[11] I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, editors. *Feature Extraction. Foundations and Applications*, volume 207 of *Studies in Fuzziness and Soft Computing*. Springer, Berlin Heidelberg, 2006.

[12] Y. Han, J.-H. Kim, and K. Lee. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 25(1):208–221, 2017.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2009.

[14] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proc. of the 10th Int'l Society for Music Information Retrieval Conference (ISMIR)*, pages 327–332, 2009.

[15] I. Kaminsky and A. Materka. Automatic source identification of monophonic musical instrument sounds. In *Proc. of IEEE Int'l Conf. on Neural Networks*, volume 1, pages 189–194 vol.1, 1995.

[16] O. Lartillot and P. Toiviainen. MIR in Matlab (II): A toolbox for musical feature extraction from audio. In *Proc. 8th Int'l Conf. on Music Information Retrieval (ISMIR)*, pages 127–130, 2007.

[17] S. A. Lashari, R. Ibrahim, and N. Senan. Soft set theory for automatic classification of traditional pakistani musical instruments sounds. In *Proc. of Int'l Conf. on Computer Information Science (ICCIS)*, volume 1, pages 94–99, 2012.

[18] I. Mierswa and K. Morik. Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58(2-3):127–149, 2005.

[19] B. C. J. Moore and B. R. Glasberg. A revision of Zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345, 1996.

[20] M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *Proc. of the 6th Int'l Conf: on Music Information Retrieval (ISMIR)*, pages 288–295, 2005.

[21] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Upper Saddle River, 1993.

[22] J. Sebastian and H. A. Murthy. Onset detection in composition items of carnatic music. In *Proc. of the 18th Int'l Society for Music Information Retrieval Conf.*, pages 560–567, 2017.

[23] A. Srinivasamurthy. *A Data-driven Bayesian Approach to Automatic Rhythm Analysis of Indian Art Music*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2016.

[24] I. Vatolkin. *Improving Supervised Music Classication by Means of Multi-Objective Evolutionary Feature Selection*. PhD thesis, Dep. of Computer Science, TU Dortmund, 2013.

[25] I. Vatolkin. Generalisation performance of western instrument recognition models in polyphonic mixtures with ethnic samples. In *Proc. of the 6th Int'l Conf. on Computational Intelligence in Music, Sound, Art and Design (EvoMUSART)*, volume 10198 of *Lecture Notes in Computer Science*, pages 304–320, 2017.

[26] I. Vatolkin, W. Theimer, and M. Botteck. AMUSE (Advanced MUSic Explorer) - a multitool framework for music data analysis. In J. S. Downie and R. C. Veltkamp, editors, *Proc. of the 11th Int'l Society on Music Information Retrieval Conf. (ISMIR)*, pages 33–38, 2010.

[27] A. Zlatintsi and P. Maragos. Multiscale fractal analysis of musical instrument signals with application to recognition. *IEEE Transactions on Audio, Speech & Language Processing*, 21(4):737–748, 2013.