Cross-Dataset Validation of Feature Sets in Musical Instrument Classification

14 November 2015

Patrick Donnelly

Montana State University

MONTANA STATE UNIVERSITY

Numerical Intelligent Systems Laboratory

Motivation

Problem: Identify the musical instrument(s) present in recordings.

Important tasks in Music Information Retrieval (MIR):

- Recognize musical genre
- Catalog recordings by instrument
- Automatic score transcription from audio

We present a feature extraction approach for classification of single instruments that is extensible to multi-label classification of mixtures of two, three, and four musical instruments playing simultaneously.





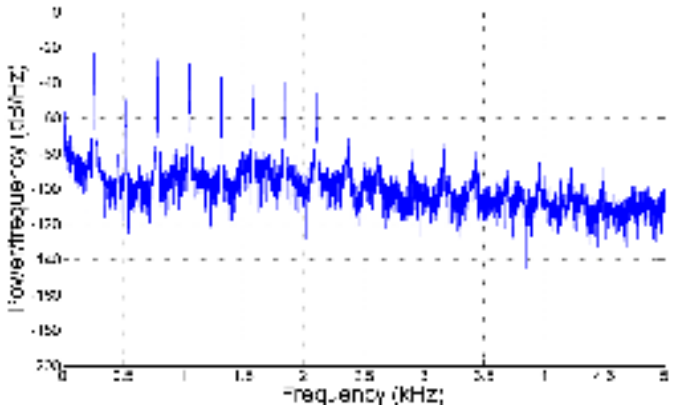
Timbre

- Timbre is also known as tone color or tone quality
- Timbre allows differentiation between instruments, even when playing the same note

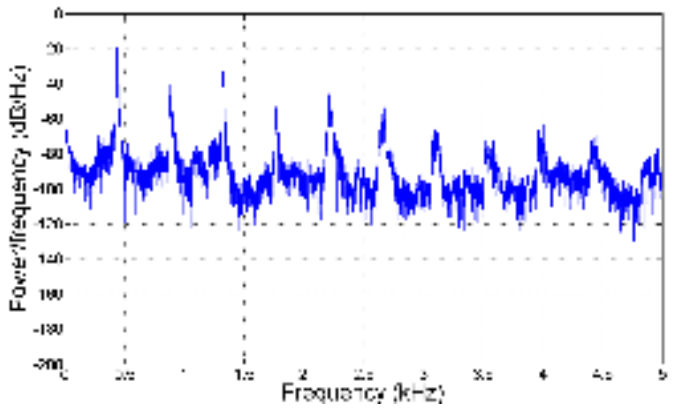
Psychoacoustic Definition

Timbre perception relies on the

- fine timing (*envelope*) of each
- harmonic constituent (*partial*).



Clarinet (261 Hz) Play



Violin (440 Hz) Play

Related Work: Multi-label classification

Multi-label algorithms:

k-nearest neighbor	Jiang *et al.*, 2009
decision trees	Jiang *et al.*, 2009
support vector machines	Duan *et al.*, 2014
hidden Markov models	Paulus *et al.*, 2009
neural networks	Hamel *et al.*, 2010
random ferns classifier	Kursa *et al.*, 2014

Source separation estimation:

modeling mixtures of prototypes	Leveau *et al.*, 2007
missing feature approach	Giannoulis *et al.*, 2014
source factorization	Rodriguez-Serrano *et al.*, 2012
computational modeling	Burred *et al.*, 2010

Limitations

There are large variation in feature extraction, classification algorithms, datasets, and evaluation metrics. Many studies are limited by use of synthetic data, evaluation of small sets of instruments, or dependence on temporal features.





Overview

- We design a system for instrument classification that satisfies the qualities of scalability, generalizability, and practicality; extensible to multi-label classification
- We present a data-driven approach to learn instrument specific areas of the spectra for feature extraction
- We develop a binary-relevance feature extraction approach to couple with binary-relevance classification
- We demonstrate our approach with experiments on the largest data repository in the literature
  - single- and multi-label classification
  - cross-validation for dataset bias
  - cross-dataset validation





Feature Extraction Approach

We developed a feature extraction scheme extensible for multi-classification of mixtures of musical instruments with the design goals of scalability, generalizability, and practicality.

1. **Signature Clustering** – data-driven approach to learning areas of spectral prominence for each instrument to inform feature extraction.
2. **Feature Extraction** – considering only the amplitude features in regions masked by the spectral filter for a given instrument hypothesis.
3. **Classification** – use of binary classifiers, each trained for single instrument, to predict if the hypothesized instrument is present in the mixture.





Datasets

Four Datasets

- 1 to 3 performers
- all notes in range
- multiple dynamic levels
- 1 second audio examples

Woodwind	Brass	String
Flute	French Horn	Violin
Clarinet	Trumpet	Viola
Alto Sax	Trombone	Cello
Oboe	Tuba	Contrabass
Bassoon		

13 instruments in common

Dataset	Abbr.	# Examples	Smallest	Largest
McGill University Master Samples	MUMS	1043	14	152
Iowa Musical Instrument Samples	MIS	3871	55	832
Real-World Computing Collection	RWC	24,216	480	3117
Philharmonia Orchestra Samples	PHO	9522	557	954





Learning Spectral Signatures

Binary-Relevance Feature Extraction

Binary-relevance feature extraction uses a different feature set for each binary-relevance classifier.

Goal: We present a data-driven approach to training instrument-specific spectral filters for use in feature extraction.

Advantages of Approach

- trained on solo instrument data
- learns important locations for feature extraction
- learns specific locations for each instrument
- results in variable feature space per instrument



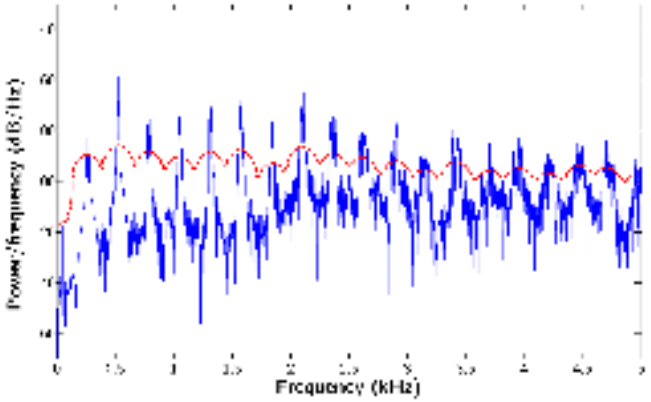


Learning Spectral Signatures

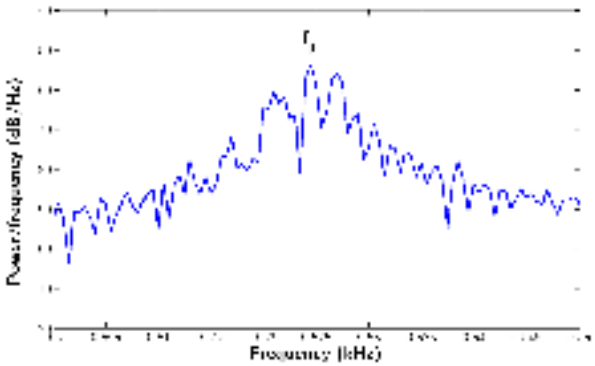
Approach

- identify f_0
- identify significant peaks in spectra of signal
- extract location as ratio to fundamental
- cluster locations by instrument with k-means cluster with adaptive k
- extract μ and σ for each cluster
- clusters correspond to important locations in the spectra

Donnelly and Sheppard. **Clustering Spectral Filters for Extensible Feature Extraction in Musical Instrument Classification.** In 27^{th} *International FLAIRS Conference*, 2014.



Extracting significant peaks

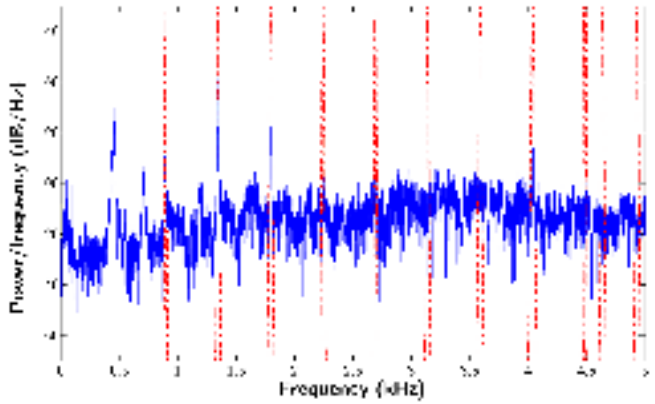


Zoomed view of peak

Feature Extraction

Signature as a Spectral Mask

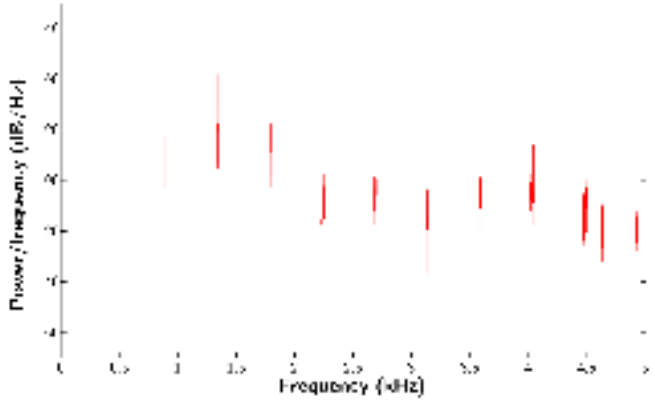
- instrument specific signatures
- relative to hypothesized f_0
- consider amplitudes under mask
- simple max amplitude feature
- amplitude normalization to f_0



Spectra with signature

dynamic	f_0	f_1	f_2	f_3	f_4
f	-69.53 —	-51.27 1.36	-51.42 1.35	-58.81 1.18	-61.14 1.14
mf	-73.94 —	-54.24 1.36	-52.26 1.40	-57.38 1.29	-62.01 1.19
p	-74.84 —	-53.72 1.39	-53.61 1.39	-58.56 1.28	-61.66 1.21

Amplitude normalization example



Masked spectra

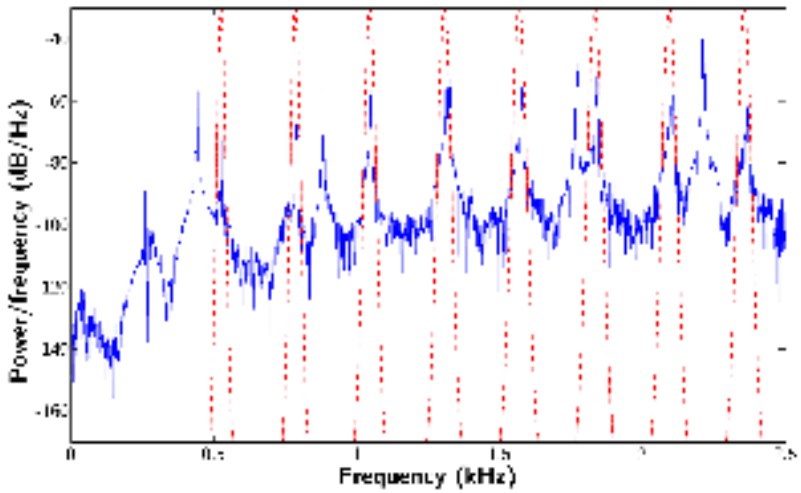
Feature Extraction from Mixtures

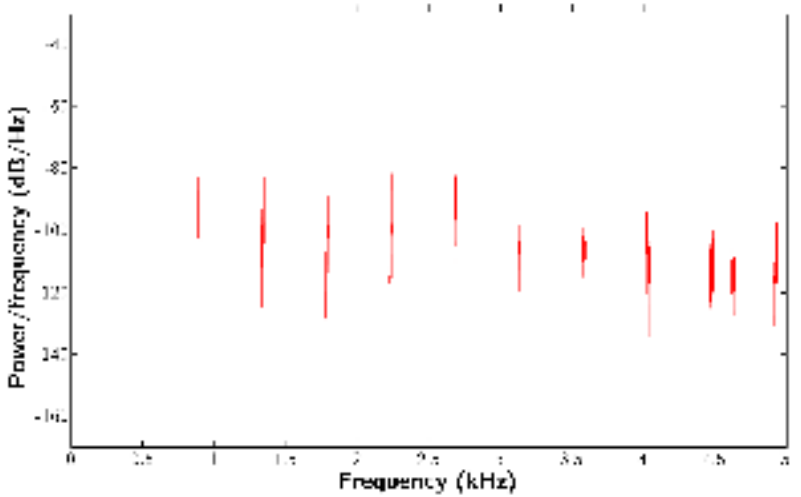
Consider only the spectra under the signature

Cannot assume f_0

Must consider:

- each peak as f_0
- each instrument





Implications on Complexity

- Fast Fourier transform computed only once
- Adds an additional feature extraction step for each binary classifier

Overview of Experiments

Experimental Design

- k-NN with $k = 5$
- BR feature extraction
- BR classification
- Trained on solo instruments
- Tested on solo instruments
- Tested on mixtures of two, three, and four instruments

Paradigms

Single-label classification:

- Cross-validation
- Cross-dataset signature
- Cross-dataset

Multi-label classification:

- Self-classify
- Cross-validation
- Cross-dataset validation



Single Instrument: Cross-Validation

Experimental Design

Solo training set; solo test set; 10-fold cross-validation; k-NN; $q = 13$

Instrument	MUMS	MIS	RWC	PHO
French Horn	0.64	0.75	0.77	0.80
Trumpet	0.75	0.91	0.73	0.77
Trombone	0.51	0.74	0.76	0.77
Tuba	0.65	0.88	0.86	0.90
Flute	0.77	0.73	0.77	0.82
Clarinet	0.73	0.87	0.88	0.87
Alto Saxophone	0.53	0.79	0.74	0.75
Oboe	0.54	0.83	0.80	0.87
Bassoon	0.73	0.67	0.87	0.85
Violin	0.72	0.87	0.84	0.82
Viola	0.71	0.74	0.78	0.82
Cello	0.79	0.78	0.83	0.80
Contrabass	0.80	0.89	0.91	0.90

F_1 score by instrument and dataset.





Single Instrument: Cross-Dataset

Experimental Design

- Solo training set
- Solo test set
- k-NN
- Four datasets
- $q = 13$ instruments
- 4×4 cross-dataset validation
  - if training set is same as test set, 10-fold cross-validation
  - otherwise, train with one, test on other





Single Instrument: Cross-Dataset

Inst	MUMS	MIS	RWC	PHO
FH	**0.66**	0.65	0.59	0.59
TR	**0.79**	0.67	0.65	0.59
TB	**0.62**	0.66	0.65	0.61
TU	**0.69**	0.50	0.66	0.71
FL	**0.78**	0.73	0.72	0.63
CL	**0.81**	0.75	0.78	0.80
AS	**0.59**	0.38	0.47	0.44
OB	**0.68**	0.67	0.68	0.71
BS	**0.77**	0.72	0.70	0.68
VN	**0.73**	0.58	0.67	0.67
VA	**0.68**	0.63	0.66	0.65
VC	**0.78**	0.62	0.77	0.63
CB	**0.83**	0.74	0.84	0.77

Inst	MUMS	MIS	RWC	PHO
FH	0.78	0.75	**0.78**	0.67
TR	0.75	0.74	**0.72**	0.64
TB	0.78	0.74	**0.76**	0.72
TU	0.59	0.36	**0.87**	0.73
FL	0.77	0.73	**0.78**	0.67
CL	0.78	0.82	**0.89**	0.75
AS	0.78	0.75	**0.79**	0.53
OB	0.80	0.79	**0.82**	0.79
BS	0.83	0.81	**0.86**	0.77
VN	0.72	0.69	**0.87**	0.77
VA	0.83	0.60	**0.80**	0.61
VC	0.88	0.67	**0.84**	0.70
CB	0.93	0.85	**0.92**	0.81

Inst	MUMS	MIS	RWC	PHO
FH	0.65	**0.77**	0.65	0.62
TR	0.61	**0.91**	0.66	0.61
TB	0.65	**0.74**	0.69	0.68
TU	0.44	**0.88**	0.42	0.54
FL	0.66	**0.77**	0.74	0.66
CL	0.63	**0.88**	0.83	0.76
AS	0.66	**0.81**	0.64	0.55
OB	0.70	**0.85**	0.67	0.69
BS	0.81	**0.77**	0.73	0.68
VN	0.66	**0.87**	0.75	0.74
VA	0.69	**0.78**	0.58	0.61
VC	0.67	**0.80**	0.67	0.66
CB	0.91	**0.90**	0.88	0.81

Inst	MUMS	MIS	RWC	PHO
FH	0.70	0.65	0.66	**0.82**
TR	0.62	0.83	0.68	**0.79**
TB	0.78	0.73	0.66	**0.79**
TU	0.86	0.67	0.83	**0.91**
FL	0.49	0.74	0.68	**0.85**
CL	0.85	0.80	0.79	**0.88**
AS	0.57	0.55	0.62	**0.79**
OB	0.76	0.76	0.81	**0.88**
BS	0.70	0.73	0.77	**0.87**
VN	0.78	0.66	0.75	**0.84**
VA	0.68	0.63	0.65	**0.84**
VC	0.74	0.63	0.76	**0.83**
CB	0.90	0.87	0.87	**0.91**





Multi-label Classification: Problem Difficulty

The difficulty of a multi-label problem depends on the number of possible labels q and the number of true labels $|Y|$.

(Bernardini *et al.*, 2014)

$$\text{Cardinality} = \frac{1}{n}\sum_{i=1}^{n}|Y_i| \qquad \text{Density} = \frac{\text{Cardinality}}{q}$$

q	Polyphony	Cardinality	Density	Chance
5	Two	2	0.400	0.1
	Three	3	0.600	0.1
	Four	4	0.800	0.2
13	Two	2	0.154	0.01282
	Three	3	0.231	0.00349
	Four	4	0.307	0.00014





Multi-label Classification: Ranking Algorithm

- Adapt k-NN to return weighted average distance of neighbors for each class
- Rank instruments by this average distance

Example

Four Instruments

- Trumpet, *m*, F4 (354 Hz)
- Trombone, *f*, A5 (450 Hz)
- Flute, *p*, F5 (703 Hz)
- Alto Sax, *p*, G♭5 (746 Hz)

1. Viola
2. **Trumpet**
3. French Horn
4. **Flute**

---

5. **Trombone**
6. Tuba
7. Clarinet
8. **AltoSaxophone**
9. Oboe
10. Bassoon
11. Cello
12. Contrabass
13. Violin





Multi-label Evaluation Measures

Type	Metric	Ideal Score	Worst Score
Example-based	Subset Accuracy	1.0	0.0
	Hamming Loss	0.0	1.0
	Accuracy	1.0	0.0
	Precision	1.0	0.0
	Recall	1.0	0.0
	F_1 Measure	1.0	0.0
Label-based	Macro-Precision	1.0	0.0
	Macro-Recall	1.0	0.0
	Macro-F_1	1.0	0.0
	Micro-Precision	1.0	0.0
	Micro-Recall	1.0	0.0
	Micro-F_1	1.0	0.0
Rank-based	One-Error	0.0	1.0
	Coverage	0.0	$q - 1$
	Ranking Loss	0.0	$\lvert Y_i \rvert$
	Average Precision	1.0	0.0





Polyphonic Mixtures: Reduced Instrument Set ($q = 5$)

Five instruments: Cello, Violin, Clarinet, Flute, and Pianoforte

Type	Metric	2-mix	3-mix	4-mix
Example-based	Subset Accuracy	0.200	0.181	0.279
	Hamming Loss	0.143	0.151	0.112
	Accuracy	0.420	0.538	0.705
	Precision	0.536	0.677	0.828
	Recall	0.522	0.663	0.802
	F_1 Measure	0.527	0.668	0.812
Label-based	Macro-Precision	0.572	0.705	0.841
	Macro-Recall	0.522	0.665	0.802
	Macro-F_1	0.528	0.663	0.802
	Micro-Precision	0.537	0.677	0.828
	Micro-Recall	0.522	0.663	0.802
	Micro-F_1	0.530	0.670	0.815
Rank-based	One-Error	0.414	0.246	0.114
	Coverage_1	0.609	0.291	0.115
	Coverage_2	2.301	1.683	1.271
	Coverage_3	–	3.097	2.458
	Coverage_4	–	–	3.599
	Ranking Loss	0.332	0.355	0.248
	Average Precision	0.742	0.819	0.907





Two Instrument Mixtures ($q = 13$)

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	0.022	0.028	0.030	0.028
	Hamming Loss	0.250	0.246	0.237	0.241
	Accuracy	0.132	0.143	0.164	0.155
	Precision	0.188	0.201	0.231	0.218
	Recall	0.188	0.201	0.231	0.218
	F_1 Measure	0.188	0.201	0.231	0.218
Label-based	Macro-Precision	0.189	0.204	0.229	0.208
	Macro-Recall	0.187	0.201	0.231	0.220
	Macro-F_1	0.151	0.174	0.213	0.185
	Micro-Precision	0.188	0.201	0.231	0.218
	Micro-Recall	0.188	0.201	0.231	0.218
	Micro-F_1	0.188	0.201	0.231	0.218
Rank-based	One-Error	0.816	0.797	0.764	0.789
	Coverage_1	2.993	2.948	2.563	2.693
	Coverage_2	6.921	7.065	6.622	6.711
	Ranking Loss	0.712	0.701	0.667	0.677
	Average Precision	0.363	0.372	0.400	0.388





Three Instrument Mixtures ($q = 13$)

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	0.006	0.007	0.008	0.007
	Hamming Loss	0.342	0.340	0.327	0.331
	Accuracy	0.172	0.175	0.195	0.189
	Precision	0.259	0.263	0.292	0.283
	Recall	0.259	0.263	0.292	0.283
	F_1 Measure	0.259	0.263	0.292	0.283
Label-based	Macro-Precision	0.276	0.271	0.299	0.286
	Macro-Recall	0.259	0.263	0.291	0.284
	Macro-F_1	0.207	0.224	0.267	0.246
	Micro-Precision	0.259	0.263	0.292	0.283
	Micro-Recall	0.259	0.263	0.292	0.283
	Micro-F_1	0.259	0.263	0.292	0.283
Rank-based	One-Error	0.763	0.745	0.703	0.727
	Coverage_1	2.226	2.173	1.912	2.000
	Coverage_2	5.356	5.334	5.072	5.120
	Coverage_3	8.606	8.684	8.414	8.437
	Ranking Loss	1.018	1.01	0.953	0.984
	Average Precision	0.404	0.409	0.435	0.424





Four Instrument Mixtures ($q = 13$)

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	0.002	0.002	0.003	0.003
	Hamming Loss	0.413	0.406	0.398	0.400
	Accuracy	0.216	0.224	0.234	0.231
	Precision	0.329	0.340	0.353	0.350
	Recall	0.329	0.340	0.353	0.350
	F_1 Measure	0.329	0.340	0.353	0.350
Label-based	Macro-Precision	0.354	0.353	0.360	0.357
	Macro-Recall	0.328	0.339	0.351	0.348
	Macro-F_1	0.259	0.296	0.321	0.309
	Micro-Precision	0.329	0.340	0.353	0.350
	Micro-Recall	0.329	0.340	0.353	0.350
	Micro-F_1	0.329	0.340	0.353	0.350
Rank-based	One-Error	0.704	0.664	0.641	0.654
	Coverage_1	1.709	1.597	1.461	1.500
	Coverage_2	4.318	4.197	4.049	4.065
	Coverage_3	6.938	6.855	6.750	6.723
	Coverage_4	9.638	9.641	9.562	9.503
	Ranking Loss	1.268	1.247	1.173	1.219
	Average Precision	0.453	0.464	0.481	0.474





Instrument Results: Four Instrument Mixtures ($q = 13$)

Instrument	MUMS	MIS	RWC	PHO
Flute	0.445	0.470	0.426	0.442
Oboe	0.454	0.039	0.092	0.257
Clarinet	0.012	0.347	0.426	0.206
Bassoon	0.117	0.249	0.198	0.146
AltoSax	0.216	0.298	0.383	0.234
FrenchHorn	0.066	0.321	0.202	0.358
Trumpet	0.137	0.461	0.337	0.032
Trombone	0.341	0.182	0.328	0.391
Tuba	0.266	0.411	0.159	0.488
Violin	0.036	0.290	0.369	0.342
Viola	0.373	0.237	0.321	0.347
Cello	0.451	0.136	0.446	0.284
Bass	0.448	0.410	0.490	0.490

F_1 score by instrument label





Investigating Potential Dataset Bias

It is common practice to use solo instruments to create mixtures of multiple instruments (Duan *et al.*, 2014).

Although the solo training set and test sets of mixtures differ, they are not independent.

Experiment

- 5×2 dataset partition; independent training and tests set
- cross-validation experiment; averaged results over folds

Mixture	MUMS		MIS		RWC		PHO	
	DB	**CV**	**DB**	**CV**	**DB**	**CV**	**DB**	**CV**
2-mix	0.188	0.194	0.201	0.207	0.231	0.212	0.218	0.232
3-mix	0.259	0.267	0.259	0.283	0.292	0.283	0.283	0.298
4-mix	0.329	0.337	0.329	0.339	0.353	0.348	0.350	0.357

Comparison of F_1 score for two, three, and four instruments with dataset bias **DB** and cross-validation **CV**





Cross-Dataset Classification: Instrument Mixtures

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	0.029	0.029	**0.031**	0.025
	Hamming Loss	0.237	0.243	**0.236**	0.245
	Accuracy	0.163	0.151	**0.165**	0.145
	Precision	0.229	0.212	**0.232**	0.205
	Recall	0.229	0.211	**0.232**	0.204
	F_1 Measure	0.229	0.211	**0.232**	0.205

Mixtures of two instruments; trained with RWC

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	0.005	0.006	**0.009**	0.007
	Hamming Loss	0.325	0.333	**0.326**	0.335
	Accuracy	0.198	0.185	**0.196**	0.182
	Precision	0.296	0.278	**0.294**	0.274
	Recall	0.296	0.278	**0.294**	0.274
	F_1 Measure	0.296	0.278	**0.294**	0.274

Mixtures of three instruments; trained with RWC

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	0.004	0.002	**0.003**	0.002
	Hamming Loss	0.399	0.405	**0.397**	0.407
	Accuracy	0.233	0.225	**0.235**	0.222
	Precision	0.351	0.341	**0.355**	0.339
	Recall	0.351	0.341	**0.355**	0.339
	F_1 Measure	0.351	0.341	**0.355**	0.339

Mixtures of four instruments; trained with RWC





Cross-Dataset Classification: Instrument Mixtures

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	**0.023**	0.016	0.020	0.020
	Hamming Loss	**0.249**	0.252	0.253	0.250
	Accuracy	**0.135**	0.126	0.125	0.133
	Precision	**0.191**	0.180	0.177	0.189
	Recall	**0.191**	0.180	0.177	0.189
	F_1 Measure	**0.191**	0.180	0.177	0.189

Mixtures of two instruments; trained with MUMS

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	**0.006**	0.005	0.004	0.003
	Hamming Loss	**0.343**	0.348	0.346	0.343
	Accuracy	**0.169**	0.162	0.165	0.169
	Precision	**0.256**	0.246	0.251	0.257
	Recall	**0.256**	0.246	0.251	0.257
	F_1 Measure	**0.256**	0.246	0.251	0.257

Mixtures of three instruments; trained with MUMS

Type	Metric	MUMS	MIS	RWC	PHO
Example-based	Subset Accuracy	**0.001**	0.002	0.002	0.003
	Hamming Loss	**0.413**	0.416	0.417	0.408
	Accuracy	**0.215**	0.212	0.210	0.222
	Precision	**0.329**	0.324	0.322	0.337
	Recall	**0.329**	0.324	0.322	0.337
	F_1 Measure	**0.329**	0.324	0.322	0.337

Mixtures of four instruments; trained with MUMS





Contributions

- Binary-relevance feature extraction scheme
- Data-driven approach to learning spectral masks
- Large data source: four datasets; 13 instruments
- 4×4 cross-dataset validation of approach
- Comprehensive reporting of multi-label evaluation measures





Future Work

- Explore other training sets
  - combination of different datasets
  - mixtures of instruments (noisy)
- Combine our spectral masks with other types of features
- Incorporate temporal features
- Further optimize feature sets specific to each instrument
- Combine our procedure with a multi-pitch finding algorithm
- Analysis of confusions based on musical intervals and dynamic levels in mixtures



