Dublin Institute of Technology
ARROW@DIT

Conference papers

Audio Research Group

2009-01-01

Evaluating Ground Truth for ADRess as a Preprocess for Automatic Musical Instrument Identification

Joseph McKay
Dublin Institute of Technology, joey.mckay@dit.ie

Mikel Gainza
Dublin Institute of Technology, mikel.gainza@dit.ie

Dan Barry
Dublin Institute of Technology, dan.barry@dit.ie

Follow this and additional works at: http://arrow.dit.ie/argcon

Part of the Engineering Commons

Recommended Citation

McKay, J., Gainza, M. & Barry, D. Evaluating ground truth for ADRess as a preprocess for automatic musical instrument identification. Paper presented at the *126th Audio Engineering Society Convention, 7-10 May, Munich, Germany, 2009.*

This Conference Paper is brought to you for free and open access by the Audio Research Group at ARROW@DIT. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@DIT. For more information, please contact yvonne.desmond@dit.ie, arrow.admin@dit.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License

Audio Engineering Society

Convention Paper

Presented at the 126th Convention
2009 May 7–10 Munich, Germany

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Evaluating Ground Truth for ADRess as a Preprocess for Automatic Musical Instrument Identification

Joey McKay[1], Mikel Gainza[1], and Dan Barry[1]

[1] *Audio Research Group, Dublin Institute of Technology, Dublin, Ireland*

Correspondence should be addressed to Joey McKay (joey.mckay@comp.dit.ie)

ABSTRACT

Most research in musical instrument identification has focused on labeling isolated samples or solo phrases. A robust instrument identification system capable of dealing with polytimbral recordings of instruments remains a necessity in music information retrieval. Experiments are described which evaluate the ground truth of ADRess as a sound source separation technique used as a preprocess to automatic musical instrument identification. The ground truth experiments are based on a number of basic acoustic features, while using a Gaussian Mixture Model as the classification algorithm. Using all 44 acoustic feature dimensions, successful identification rates are achieved.

1. INTRODUCTION

Classifying musical instruments with a computational model has mainly been approached by considering the relevant acoustic features produced by the musical instrument and passing these features into a machine learning algorithm to obtain a condensed and representative model of each instrument class. The model can be interpreted as a timbre perception space, where the multi-dimensional nature of timbre is transposed into a multi-feature descriptor based system which represents the sensory information received by the ear.

The instrument classification learning models typically require the provision of a *training set*, which comprises of instrument recordings and the instrument labels associated to them. A large set of *features* is extracted from the recordings, which are transformed into *selected features* by the refinement algorithms. Then, a *learning algorithm* is utilised in order to train the *classifier*. The performance of the classifier is evaluated by using a *test set* of instrument recordings which are independent of the train-

ing set. The different parts that comprise such an automatic musical instrument identification system (AMIIS) are depicted in Fig. 1.

1.1. Instrument identification approaches

Research into automatic musical instrument identification (AMII) can be subdivided into the following categories; each category reflecting the instrument sound samples used as input to the system for classification: isolated notes, musical phrases, and polytimbral music. For a detailed literature review of both isolated notes approaches and musical phrases approaches, see [1] and [2].

1.2. Classification of polytimbral sounds

The recognition of musical instruments in polytimbral mixtures can be divided into two main approaches:

1. *Overlapped source recognition*: direct recognition of the instrument source taking into consideration the influence of the overlapping of multiple sound sources
2. *Prior source separation*: the mixture is preprocessed using a sound source separation algorithm to separate the instrument sources which are then classified by the system.

1.2.1. Overlapped source recognition

In [3], the authors address the fact that few researchers have dealt with identifying musical instruments in a polytimbral context. Kitahara's study addresses the problem facilitating score-based music annotation of polyphonic music. The main difficulty in identifying instruments in polyphonic music is the fact that acoustical features of each instrument could not be extracted without blurring because of the overlapping of partials. The system presented applies weights to features, where higher weighting values are utilised for features affected less from overlapping. Kitahara uses Discriminant Analysis with Mixed Sounds (DAMS) on training data obtained from polyphonic sounds, which generates a subspace where the influence of the partial overlapping problem is minimised. In addition, the author considers the temporal continuity of melodies. Thus, if the majority of identified sequential notes were for example a flute, a note classified as clarinet within this sequence could be considered erroneous. However, the system uses prior information of the correct fundamental frequencies. In addition, the testing procedure was performed by using synthetic music.

Addressing the limitations in [3], a later study by Kitahara et al. [4] describes a technique for recognising musical instruments in polyphonic music without relying on onset detection or fundamental frequency (F0) estimation. The idea behind the technique is visualising the probability that a target instrument exists at each time.

In [5], the authors identify frequency regions that are dominated by energy from an interfering tone. The regions are then considered unreliable and excluded from the GMM classification process. In [6] and [7], hierarchical methods which recognise musical instruments in polyphonic music without the requirement of prior source separation are presented. The methods identify combinations of instruments by detecting cues on the common structures of musical ensembles.

A method based on independent subspace analysis (ISA) is presented in [8], where the features are derived from the statistical independent components provided by the ICA part of ISA. In [9], non-linear ISA is utilised to model the short-term log-power spectra of polyphonic music as weighted non-linear combinations of typical note spectra. The spectra are learned from a training set of isolated notes or solo recordings from different instruments.

In [10], an instrument recognition system based on the calculation of the fundamental frequency and the onsets played by the instruments is presented. In this case, the classifier uses a set of neural networks.

1.2.2. Preprocess with source separation

Sound source separation refers to the problem of synthesising S_N source signals given a C_M channel mixture of those source signals. When there are fewer input mixtures than sources to be separated ($M < N$), we have the *degenerate* case. In the *non-degenerate* case ($M \geq N$), the basic problem is to estimate the mixing matrix to determine how the sources are combined into mixtures. This matrix may then be inverted to obtain the input sources. In digital audio, the case most frequently encountered is the two mixture degenerate case, as many or most currently available commercial digital

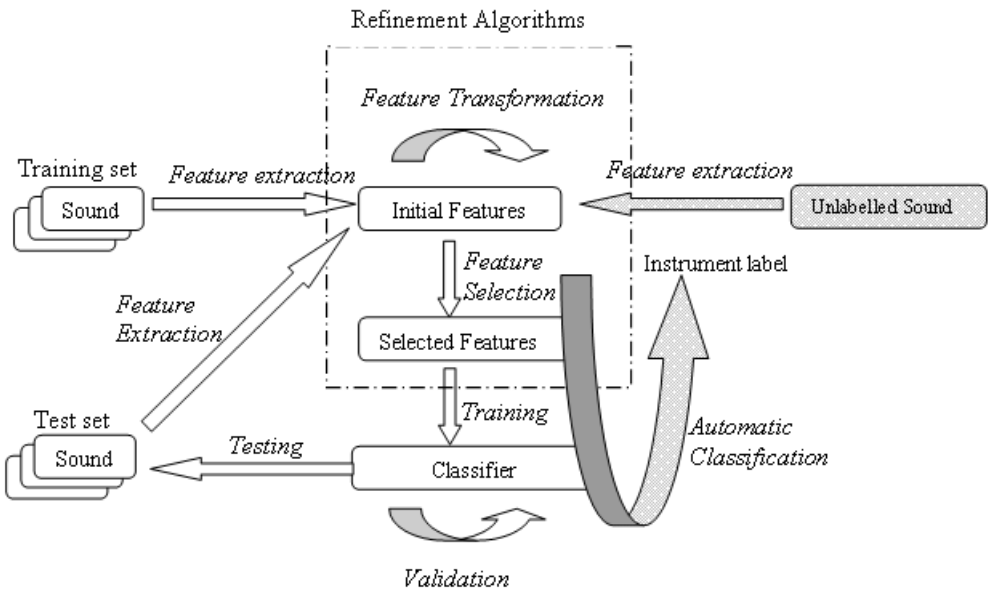

Fig. 1: An AMIIS overview. Adapted from [1]

recordings contain two channels (stereo) but more than two instruments, voices, or other sounds. A system which can separate the instrument sources within a polytimbral mixture has been described in the literature as ideal [11].

The approach by [12] using DUET was limited to 5 musical instruments in a mixture composed of isolated notes. DUET, as a degenerate case, assumes a signal model of S_N sources in 2 channels (stereo input signals). DUET relies on an assumption referred to as '*W-disjoint orthogonality*'. This assumption approximately holds for mixtures of speech. However, polytimbral music signals do not obey this assumption. The harmonic structure of music and the harmonic nature of instruments and voices means DUET introduces some artifacts into the estimate of the original sources, which can compromise the accuracy of the identification system.

As a preprocess to AMII, [13] successfully applied ICA to separate piano and violin from a mixture of the two instruments. ICA relies on the important assumption that the sources must be statistically independent with non-Gaussian distributions. In addition to this, ICA assumes that there are at least as many observation mixtures as there are independent sources. When concerned with musical recordings, we will have at most only 2 observation mixtures, the left and right channels. This makes ICA unsuitable as a means to processing real-world data where the musical excerpt may be a mixture composed of numerous instruments.

1.3. Instrument identification using ADRess

As mentioned, neither DUET nor ICA have proven feasible for identification of instruments in modern day representations of music. The overlap of sources breaches DUET's assumption of W-disjoint orthogonality while popular modern music formats do not provide the same number of observation mixtures as there are sources. Stereo music for instance only has two channels, the left and right. Given that the majority of music collections used by the general public will be 'real-world' songs of various instrumentations, across a broad spectrum of musical genres, an AMIIS which can emulate the source separation stage of Auditory Scene Analysis [14] would provide significant scope in fields such as Musical Information Retrieval (MIR). ADRess enables separation of multiple instrument sources from real-world polytimbral stereo musical excerpts [15].

The ADRess algorithm achieves source separation

by taking advantage of destructive phase cancellation in the frequency domain. For each frame, m, of a short-time Fourier transform (STFT) representation of the signal, one channel is iteratively gain scaled and subtracted from the other in the complex frequency domain after which the absolute value is taken. The resulting array is of dimension $N \times ß$, where N is the number of frequency points and ß, the azimuth resolution, is the number of equally spaced gain scalars between 0 and 1. The operation reveals local minima, due to phase cancellation across the azimuth plane for each frequency component. Using a simple clustering technique, components belonging to a single source are seen to have their minima in a localised region about some gain scalar which ultimately refers to the intensity ratio between each channel, i.e., the pan position of the source in stereo space. By estimating the magnitude of each of the time-frequency minima and only resynthesising those with a desired intensity ratio, a single source may be reconstructed. The process can be summarised as follows with the iterative gain scaling process achieved using equation (1) where $X_j(k,m)$ is a complex frequency domain representation of the mth frame of the jth channel (left or right).

$$\begin{aligned} Az_1(k,m,i) &= |X_2(k,m) - g(i)\,X_1(k,m)| \\ Az_2(k,m,i) &= |X_1(k,m) - g(i)\,X_2(k,m)| \end{aligned} \quad (1)$$

where $1 \leq k \leq N$, N being the Fourier transform length, and where $g(i) = \frac{i}{ß}$, for all i where, $0 \leq i \leq ß$, and where i and ß are integer values. ß refers to the number of gain scalars to be used and ultimately gives rise to the resolution achieved in the azimuth plane. The resulting matrix represents the frequency-azimuth plane for the mth frame of the jth channel. Each of k frequency bins will exhibit a local minimum at some index i. It can be observed that the majority of frequency bins pertaining to a single source should exhibit their minima around a singular value for i. These local minima represent the points at which frequency components experience a reduction in energy due to destructive phase cancellation between the left and right channel. This energy reduction is directly proportional to the amount of energy which the cancelled source had contributed to the overall mixture and so to invert these minima around a single azimuth point should yield short-time magnitude spectra of the individual sources. To achieve this inversion, we simply subtract the minimum from the maximum of the function in (1) for each of k frequency bins as described in equation (2).

$$A\bar{z}_1(k,m,i) = \begin{cases} Az_1(k,m)_{max} - Az_1(k,m)_{min} & \text{if } A\bar{z}_1(k,m,i) = min \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

where 'min' and 'max' refer to the global minimum and maximum of the kth frequency-azimuth function. Note that the inverted frequency-azimuth plane for channel 2 is created in an identical fashion. Now, the instantaneous magnitude spectrum of a single source or subspace at pan position d, predominant in the jth channel can be approximated as in (3)

$$Y(k,m) = \sum_{i=d+H/2}^{i=d-H/2} A\bar{z}_j(k,m,i) \times \left(1 - \frac{2\,|d-i|}{H}\right) \quad (3)$$

where d is the azimuth index, i.e. the pan position of the source for separation and H is the azimuth subspace width which is simply a neighbourhood around the azimuth index. The second term in (3) simply creates a linear weighting function such that components further from the azimuth index are scaled down. This essentially creates a triangular separation window along the azimuth axis. $YR(k)$ is now an $Nx1$ array containing the short-time magnitude spectrum of a single source or azimuth subspace. For a detailed description of the ADRess algorithm, refer to [15].

While the future work of this research will focus on real world audio, a number of experiments were undertaken to establish just how valid ADRess is as a prior sound source separation technique while using synthetic examples. These experiments are described in the following sections.

2. GROUND TRUTH EXPERIMENTS

This section details the settings for the ground truth experiments.

2.1. Training and test data

To establish the ground truth, data from a stereo mixture of 5 MIDI instruments was created using the Universal Sound Module in Steinberg Cubase

Instrument	MIDI Volume	Pan Value
Acoustic Bass (ABs)	116	-10 (Left)
Violin (Vn)	78	-43 (Left)
Trumpet (Tr)	44	64 (Right)
Cello (Ce)	65	28 (Right)
Tuba (Tu)	96	15 (Right)

Table 1: Mixture settings for 5 instruments in Cubase SE

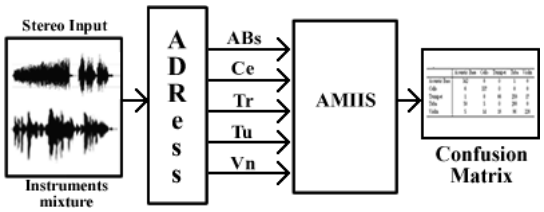

Fig. 2: Test process for ADRess separated instruments.

SE [16]. These instruments are: Acoustic Bass (**ABs**), Cello (**Ce**), Trumpet (**Tr**), Tuba (**Tu**) and Violin (**Vn**). The parameter settings for the 5 instruments stereo mixdown are detailed in Table 1. The data used in the experiments is as follows:

- **Training data**: The training data was extracted from the MIDI scales of each of the 5 musical instruments: ABs(F2-F4), Ce(C2-F5), Tr(E3-B5), Tu(F1-E4) and Vn(G3-F#7). Overlap of the pitch ranges was ensured.
- **Testing data**: The testing data consisted of three groups of instrument samples:
  1. **solo**: From the MIDI 5 instrument mixture in Cubase, each instrument is soloed in the composition and exported as a mono sound excerpt. Thus, the original instrument before the mixing process is available as test data.
  2. **sep**: ADRess is applied to the stereo mixture of the 5 instruments. Each instrument is extracted using ADRess, and saved as the separated source. The test process for ADRess separated instruments is shown in Fig. 2. The detected source positions using ADRess are illustrated in Fig. 3.

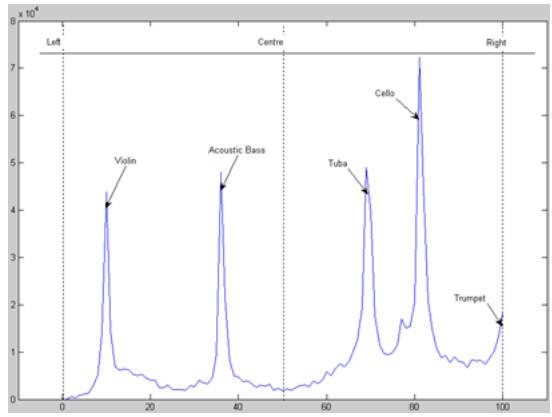

Fig. 3: Plot shows the sum of all energy in source positions for the entire mix. Peaks of panned instruments are clearly visible. (From left to right: violin, acoustic bass, tuba, cello & trumpet.)

  3. **mix**: The 5 instrument mixture is tested against the system to determine how the AMIIS performs on polytimbral mixtures.

2.2. Features

8 features in total were chosen for these ground truth experiments. Feature vectors (FV) were calculated for each frame of an STFT, with an FFT size of 1024 samples and a hop size of 512. The solo testing phrases for each instrument are of equal length, resulting in 343 FVs. The dimension of a solo FV, then equals (343 × dimension of the features). This compares to the ADRess separated instruments, which have 336 frames. The features applied in this research are described as follows:

- *Zero Crossing Rate* (ZC): In discrete time signals, a zero crossing occurs if there is a sign difference between successive samples. ZC is a measure of the frequency content of a signal. In the calculation, each pair of samples is checked to determine where zero crossings occur and then the average is computed over N consecutive samples. ZC is computed for each time frame of the signal. Low ZC values indicate periodic signals while noisy signals have high ZC values. ZC were succesfully implemented by [17] and ZC as a feature was selected 18 times (out of 45) in a separate study [18].

- *Spectral Rolloff* (SR): measures the spectral shape and is defined as the frequency below which a percentage of the magnitude distribution is concentrated [19].

- *Spectral Centroid* (SC): is correlated with the perceived brightness from Multidimensional Scaling studies [20]. SC can be thought of as the center of gravity for the frequency components of a signal and is calculated as:

$$SC = \sqrt{\frac{\sum_{k=1}^{N/2} f(k)P(k)}{\sum_{k=1}^{N/2} P(k)}} \tag{4}$$

  where, $f(k)$ is the frequency at index k, N is the size of the FFT and $P(k)$ is the power spectrum, i.e. magnitude spectrum squared.

- *Bandwidth* (BW): BW is defined as the width of the range of frequencies that the signal occupies. BW is the square root of the power-weighted average of the squared difference between the spectral components and frequency centroid. In general, the BW range of speech is from 0.3KHz to 3.4KHz. For music the range is much wider, ordinarily BW is 22.05KHz. BW has shown effectiveness in many audio classification systems [21]. BW is calculated using the following equation:

$$BW = \sqrt{\frac{\sum_{k=1}^{N/2} \left(f(k) - SC\right)^2 P(k)}{\sum_{k=1}^{N/2} P(k)}} \tag{5}$$

- *Spectral Flux* (SF): measures the average variation value of the spectrum between two adjacent frames and measures the amount of local spectral change [22].

- *Mel-frequency Cepstral Coefficients* (MFCCs): are perceptually motivated features originally developed for the classification of speech, which have proven successful for various sound source classification tasks including instrument classification [23].

- *Delta Coefficients*: calculate the rate of change of the MFCCs as delta coefficients (speed) and delta-delta (acceleration) coefficients.

The overall feature dimension when all features are included is 44.

2.3. Classification

Gaussian mixture models (GMMs), are frequently employed in musical instrument recognition research [24]. GMM were chosen as the classification technique for these ground truth experiments. As is the case with GMMs, choosing a *model order* to model each musical instrument adequately is an important but difficult problem. There is no theoretical way to estimate the number of mixture components a priori. The objective is to choose the minimum number of components necessary to adequately model an instrument for good identification. For the purposes of these experiments, a standard of 5 mixture components has been used. The k-means algorithm is used to determine the GMM centres. The priors are computed from the proportion of examples belonging to each cluster. The covariance matrices are calculated as the sample covariance of the points associated with (i.e. closest to) the corresponding centres. Full covariance matrices are used in the GMM initialisation. The EM algorithm is used for training the GMMs. A constant of 500 iterations of the EM algorithm is used.

3. RESULTS

This section discusses the classification results for the ground truth experiments carried out. Results are expressed via a 'confusion matrix', a frequently applied measure for assessing the quality of a classifier. The columns correspond to the predicted musical instrument and the rows to the actual instrument. The diagonal values show the true positives for the classifier, showing the number of frames correctly classified as their true instrument class.

3.1. Experiment 1

13 MFCC feature coefficients were used for the first experiment, thus resulting with a FV of dimension 13. Referring to Table 2, the confusion matrices detail the results for this experiment. For solo instruments, the best classification result is for acoustic bass(99.7%). The worst performing instrument class is the trumpet(26.7%). As the MIDI volume (see

Experiment No.1: Solo Instruments

	AB	Ce	Tr	Tu	Vn	%Correct
AB	**342**	0	0	1	0	99.7
Ce	116	**182**	0	3	42	53
Tr	64	0	**102**	163	14	29.7
Tu	98	0	0	**245**	0	71.4
Vn	83	0	0	5	**255**	74.3
Overall classification rate						65.65%

Experiment No.1: ADRess Separated Instruments

	AB	Ce	Tr	Tu	Vn	%Correct
AB	**336**	0	0	0	0	100
Ce	331	**0**	2	1	2	0
Tr	214	0	**111**	10	1	33
Tu	330	0	0	**6**	0	1.8
Vn	312	4	9	5	**6**	1.8
Overall classification rate						27.39%

Experiment No.1: 5 instruments polytimbral mixture

	AB	Ce	Tr	Tu	Vn	
AB	343	0	0	0	0	
Overall classification rate						20%

Table 2: Confusion matrices for experiment no.1

Experiment No.2: Solo Instruments

	AB	Ce	Tr	Tu	Vn	%Correct
AB	**342**	0	0	1	0	99.7
Ce	6	**337**	0	0	0	98.3
Tr	1	0	**66**	259	17	19.2
Tu	50	5	0	**288**	0	84
Vn	5	14	18	86	**220**	64.1
Overall classification rate						73.06%

Experiment No.2: ADRess Separated Instruments

	AB	Ce	Tr	Tu	Vn	%Correct
AB	**183**	124	2	19	8	54.5
Cello	27	**285**	1	19	4	84.8
Tr	0	35	**47**	229	25	14
Tu	62	64	4	**201**	5	60
Vn	8	31	30	194	**69**	20.5
Overall classification rate						46.84%

Experiment No.2: 5 instruments polytimbral mixture

	AB	Ce	Tr	Tu	Vn	
AB	49	0	0	291	0	
Overall classification rate						20%

Table 3: Confusion matrices for experiment no.2

Table 1) is constant throughout these experiments, the low MIDI volume for the tuba(44) could account for its poor classification accuracy. The correlation between MIDI volume, pan value and classifier performance remains part of future work. The overall mean classifcation accuracy for the GMM classifier is 65.65%. For the ADRess separated instruments, the best performing class is the acoustic bass(100%). While ADRess outperforms the solo test for acoustic bass, it fails in comparison for the other instruments. Classification of the cello fails completely, confusing the cello with the acoustic base 331/336. Overall, classifcation for ADRess separated instruments using basic MFCC features, was a poor 27.39%. Using the complete polytimbral mixture, the overall accuracy is 20%, i.e. 100/#classes. Each instrument is confused 100% with the acoustic bass, meaning the classifier fails to classify the given class.

3.2. Experiment 2

13 MFCC coefficients, 13 delta and 13 delta-delta coefficients were the features used in experiment 2, with resulting FV of dimension 39. The confusion matrices in Table 3 detail the results. For solo instruments, the best classification result is for acoustic bass(99.7%). A high accuracy rate for cello(98.3%), outperforms the first experiment, suggesting the added features capture more of the salient characteristics of the cello. Better classification for the tuba(84%) is also achieved. The mean accuracy for the classifier is 73.06%, an overall improvement on experiment no.1 which suggests the delta coefficients describe these instruments better than MFCCs alone. Using ADRess, the best performing class is the cello(84.8%), which surpasses that of experiment no.1(0%). Overall classification accuracy for ADRess(46.84%) improves on that of experiment no. 1(27.39%).

3.3. Experiment 3

5 features were used in experiment 3: ZC, SR, SC, BW and SF, with a resulting of dimension 5. The confusion matrices in Table 4 detail the results. For solo instruments, the best classification result is for violin(98%). This is the best classification rate over all experiments, implying the 5 features capture the salient characteristics of the violin. Contrastingly, these features fail to describe the other 4 instruments with both trumpet and tuba not classified correctly at all. The mean classifier accuracy rate is 24.84%,

Experiment No.3: Solo Instruments

	AB	Ce	Tr	Tu	Vn	%Correct
AB	**8**	0	0	11	324	2.3
Ce	0	**82**	0	1	260	23.9
Tr	0	2	**0**	0	341	0
Tu	0	0	0	**0**	343	0
Vn	0	7	0	0	**336**	98
Overall classification rate						24.84%

Experiment No.3: ADRess Separated Instruments

	AB	Ce	Tr	Tu	Vn	%Correct
AB	**31**	14	0	10	281	9.2
Ce	0	**69**	0	0	267	20.5
Tr	0	3	**0**	0	333	0
Tu	0	0	1	**1**	334	0.3
Vn	0	19	16	27	**270**	80.4
Overall classification rate						22.14%

Experiment No.3: 5 instruments polytimbral mixture

	AB	Ce	Tr	Tu	Vn	
AB	0	0	3	0	340	
Overall classification rate						20%

Table 4: Confusion matrices for experiment no.3

Experiment No.4: Solo Instruments

	AB	Ce	Tr	Tu	Vn	%Correct
AB	**340**	2	0	1	0	99.1
Ce	0	**295**	0	0	48	86
Tr	0	1	**2**	39	301	0.6
Tu	21	28	0	**294**	0	85.7
Vn	0	0	0	235	**108**	31.5
Overall classification rate						60.58%

Experiment No.4: ADRess Separated Instruments

	AB	Ce	Tr	Tu	Vn	%Correct
AB	**201**	96	1	36	2	60
Ce	11	**302**	2	15	6	90
Tr	0	43	**41**	154	98	12.2
Tu	52	91	0	**189**	4	56.2
Vn	12	4	0	146	**170**	50.6
Overall classification rate						53.88%

Experiment No.4: 5 instruments polytimbral mixture

	AB	Ce	Tr	Tu	Vn	
AB	2	58	3	119	161	
Overall classification rate						20%

Table 5: Confusion matrices for experiment no.4

a drop in performance from experiments 1 and 2. The overall classification rate for ADRess (22.14%) is slightly less than the solo classifier(24.84%) suggesting that overall the 5 features are poor descriptors for these 5 instruments.

3.4. Experiment 4

All the features were used in experiment 4, with a resulting FV of dimension 39. The confusion matrices in Table 5 detail the results. For solo instruments, classification results for both acoustic bass(99.1%) and tuba(85.7%) were high. Classification for tuba(85.7%) outperforms all other experiments. The overall classification rate for ADress separated instruments, 53.88%, is the highest rate in this test category from all the ground truth experiments. Indeed, ADRess improves on the classification of solo cello, solo trumpet and solo violin. This improvement warranted further investigation. Through informal testing, by reducing the azimuth width in ADRess, it was noted that ADRess performs feature reduction while maintaining the most salient attributes. Future work is required to confirm this hypothesis.

4. DISCUSSION AND FUTURE WORK

We have investigated the use of ADRess as a source separator for the purpose of AMII. The experiments detailed have been evaluated by means of confusion matrices which represent a valid method for analysing the performances of the GMM classifier from a qualitative perspective. The variation in classification rates between the experiments would suggest the importance of feature selection. Future work will investigate various feature selection and transformation algorithms. While the results show there is definite room for improvement in terms of classification rates, for instance the trumpet(12.2%) in experiment 4, overall classification rates compare favourably with those of the soloed instruments.

It must be noted that there were strict limitations on these experiments. The limitations of ADRess were inherited, to include: 1) overlapping in panned sources reduces the accuracy of the synthesised sources, 2) the mixture must be stereo samples, and 3) the sources must be manually separated. Future work includes developing a system to enable automatic separation of the instrument sources.

The system implemented in these ADRess ground truth experiments used basic features without any fine tuning of the GMM classifier. Further limitations include using an ideal number of 5 clusters (given there are 5 instruments) and the covariance structure of each component is set to 'full'. An investigation into other classification algorithms warrants future work. The improvement in classification rate for separated cello in experiment 4 warrants investigating the hypothesis that ADRess performs feature selection.

As a final remark, these experiments were applied to a limited training and testing set of synthetic samples. Future work will be to resolve the above mentioned limitations and work with 'real-world' sound samples. The detailed results of these experiments establish a positive ground truth in terms of using ADRess as a source separation technique for the automatic identification of musical instruments from polytimbral mixtures.

5. REFERENCES

[1] P. Herrera-Boyer, A. Klapuri, M. Davy, *Automatic Classification of Pitched Musical Instrument Sounds*, Signal Processing Methods for Music Transcription, Springer US, 163-200, 2006

[2] P. Herrera, G. Peeters, S. Dubnov, *Automatic Classification of Musical Instrument Sounds*, Journal of New Music Research, 32, 1, pp.3-21, vol.32, Num.1 2003

[3] T. Kitahara, M. Goto, K. Komatani, T. Ogata, Hiroshi G. Okuno, *Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps*, EURASIP Journal on Applied Signal Processing 2007, pp.1-15, 2005

[4] T. Kitahara, *Instrogram: A New Musical Instrument Recognition Technique Without Using Onset Detection NOR F0 Estimation*, International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006 Proceedings. IEEE, pp.229-232, vol.5, 2006

[5] J. Eggink, G. J. Brown, *A missing feature approach to instrument identification in polyphonic music*, Applications of Signal Processing to Audio and Acoustics, IEEE Workshop, 2003

[6] S. Essid, G. Richard, B. David, *Instrument recognition in polyphonic music*, Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.245-248 vol.3, March, 2005

[7] S. Essid, G. Richard, B. David, *Instrument recognition in polyphonic music based on automatic taxonomies*, IEEE Transactions on Audio, Speech, and Language Processing, 2006

[8] P. Jincahitra, *Polyphonic instrument identification using independent subspace analysis*, Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, (ICME), Taipei, Taiwan pp.1211-1214, vol.2, 2004

[9] E. Vincent, X. Rodet, *Instrument identification in solo and ensemble music using independent subspace analysis*, ISMIR 2004, 5th International Conference on Music Information Retrieval, pp.576-581, 2004

[10] T. Zhang, *Instrument classification in polyphonic music based on timbre analysis*, Proceedings of SPIE Conference on Internet Multimedia Management Systems II, pp.136-147, vol.4519, 2001

[11] N. Chetry, *Computer Models for Musical Instrument Identification*, PhD thesis, 2006

[12] Xiang, LI, *Musical Instruments Identification System*, Master of Science Project, Queen Mary University of London, 2006

[13] M. R. Bai, Menh-Chun, Chen, *Intelligent Preprocessing and Classification of Audio Signals*, Audio Engineering Society, vol.55, num.5, 2007

[14] A. S. Bregman,*Auditory Scene Analysis*, MIT Press, 1990

[15] D. Barry, *Sound Source Separation: Azimuth Discrimination and Resynthesis*, 7th International Conference on Digital Audio Effects (DAFX 04), Naples, Italy, 2004

[16] Steinberg Media Technologies, http://www.steinberg.net, 2008

[17] S. Essid, G. Richard, B. David, *Hierarchical Classification of Musical Instruments on Solo Recordings*, IEEE International Conference on Acoustics, Speech and Signal Processing, 2006, pp.V-V, vol. 5, 2006

[18] S. Essid, G. Richard, B. David, *Musical instrument recognition by pairwise classification strategies*, IEEE Transactions on Audio, Speech and Language Processing, pp.1401-1412, vol.14, num.4, 2006

[19] E. Benetos, M. Kotti, C. Kotropoulos,J.J. Burred, G. Eisenberg, M. Haller, T. Sikora, *Comparison of Subspace Analysis-Based and Statistical Model-Based Algorithms for Musical Instrument Classification*, 2nd Workshop on Immersive Communication and Broadcast Systems (ICOB), Berlin, Germany, 2005

[20] S. Handel, *Timbre Perception and Auditory Object Identification*, Hearing (Handbook of Perception and Cognition), Academic Press, 2nd Edition, 1995

[21] L. Bai, Y. Hu, S. Lao, A. Chen Jianyun, A. Wu Lingda, *Feature analysis and extraction for audio automatic classification*, 2005 IEEE International Conference on Systems, Man and Cybernetics, pp.767-772, vol.1, 2005

[22] E. Benetos, M. Kotti, C. Kotropoulos, *Musical Instrument Classification using Non-Negative Matrix Factorization Algorithms and Subset Feature Selection*, 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings, pp.V-V, vol.5, 2006

[23] C. Simmermacher, D. Deng, S. Cranefield, *Feature analysis and classification of classical musical instruments: an empirical study*, 2006

[24] Judith C. Brown, *Computer identification of musical instruments using pattern recognition with cepstral coefficients as features*, J. Acoust. Soc. Am, 105, pp.1933-1941, 1999