# A Multiple-Expert Framework for Instrument Recognition

Mikus Grasis, Jakob Abeßer, Christian Dittmar, and Hanna Lukashevich

Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
mikus.grasis@idmt.fraunhofer.de
jakob.abesser@idmt.fraunhofer.de
christian.dittmar@idmt.fraunhofer.de
hanna.lukashevich@idmt.fraunhofer.de

**Abstract.** Instrument recognition is an important task in music information retrieval (MIR). Whereas the recognition of musical instruments in monophonic recordings has been studied widely, the polyphonic case still is far from being solved. A new approach towards feature-based instrument recognition is presented that makes use of redundancies in the harmonic structure and temporal development of a note. The structure of the proposed method is targeted at transferability towards use on polyphonic material. Multiple feature categories are extracted and classified separately with SVM models. In a further step, class probabilities are aggregated in a two-step combination scheme. The presented system was evaluated on a dataset of 3300 isolated single notes. Different aggregation methods are compared. As the results of the joined classification outperform individual categories, further development of the presented technique is motivated.

**Keywords:** instrument recognition, partial tracking, partial-wise features, overtones, classifier ensemble, decision fusion

## 1 Introduction

Each music genre is characterized by a typical range of music instruments, which have a major influence on the timbre of musical pieces. Algorithms for automatic instrument recognition are useful for a wide range of application scenarios. First, these algorithms allow for an efficient search, indexing, and recommendation of music pieces based on timbral similarity. Second, genre classification algorithms are likely to perform better if the presence of instruments and instrument groups can be applied as audio features. Third, the automatic recognition of instruments allows to select instrument-adaptive algorithms for source separation and automatic music transcription. Finally, the temporal progression of instrumentation and instrument density often correlates with perceptual time-continuous properties such as dynamic and tension.

**Musical redundancy towards instrumentation** In terms of instrumentation, music pieces show different levels of redundancies:

1. **Global redundancy**: If a particular instrument plays throughout a segment (e.g., the chorus), multiple note events can be detected and assigned towards that instrument.
2. **Local redundancy concerning partial envelopes**: Notes played on harmonic instruments consist of a fundamental frequency component and multiple overtones. Both the magnitude and the frequency envelopes of the harmonic component show a similar progression over time.
3. **Local redundancy concerning spectral frames**: Notes played on harmonic instruments usually show a very similar spectral distribution in adjacent spectral frames (e.g. in the beginning of the note decay part).

In this paper we propose an instrument-recognition framework that combines classification results from different feature categories—*note-wise, partial-wise* and *frame-wise* features. Although we only evaluated the framework with isolated note recordings so far, the approach is targeted towards the use on polyphonic material.

## 2 Previous Work

A number of solutions have been presented for the identification of musical instruments from a given audio signal. Aside from early experiments [1] many works have initially focused on the identification of monophonic sound sources. These concepts regard spectral properties of fixed-length segments [2], or take into account the temporal properties of isolated notes [3]. Later contributions cover aspects such as the pitch dependency of timbre [4] or the temporal integration of spectral features over time [5].

Over the latest years identification of instruments in polyphonic music recordings has received an increasing attention. Recognition of polyphonic sources is often performed as identification of dominant solo instruments. This has been done with feature-based approaches and an extensive training on real-world polyphonic training data [6]. Also the application of source separation techniques has been beneficially applied in this context [7]. Other approaches aim at a decomposition of the musical signal. For instance, Itoyama et al. presented a system for simultaneous separation and classification of sounds [8]. However, these algorithms tend to show heavy computational requirements. Another approach is the *selective* classification of signal-portions that show no interference from spectral overlaps. This concept has been applied by classifying instruments based on individual partials by Barbedo & Tzanetakis in [9].

We propose to extend the concept of partial classification proposed by Barbedo & Tzanetakis to a *multiple-expert instrument classification* scheme. The separate classification of observations from different feature categories allows the selective processing of unimpaired signal components. This strategy could ensure a robust classification that is applicable to polyphonic and multi-instrumental music signals, which are characterized by spectral overlap of different sound sources.

## 3 Proposed System

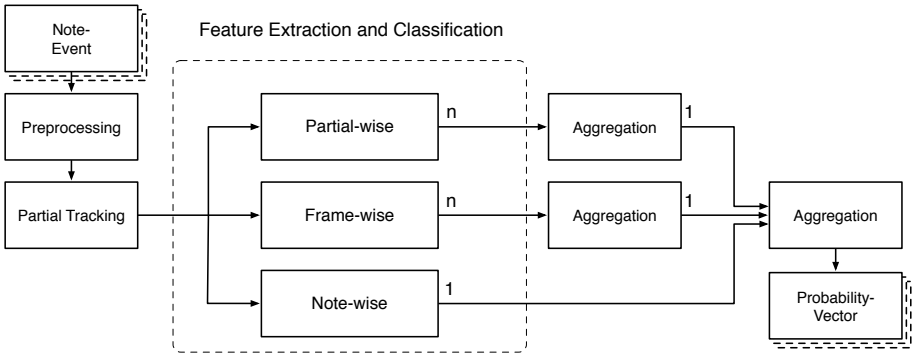



**Fig. 1.** Proposed Multiple-Expert System for Instrument Recognition. Each feature group results in one (1) or multiple (n) class probability vectors.

The proposed system operates on the basis of individual single notes. For that purpose, automatic music transcription algorithms must be applied. Classification of single notes bears the advantage that multiple notes can be evaluated in order to derive a final instrument class label for a given music segment (e.g. in case of a dominant main instrument). The transcription algorithms return a list of note events, which are characterized by the parameters *onset* (note start), *offset* (note end) and *pitch*. In the pre-processing stage of the framework (Section 3.2), the Short-time Fourier transform (STFT) of the analyzed signal is computed. In the partial tracking step (Section 3.3), the fundamental frequency and the first 9 overtones of each note events are tracked in the magnitude spectrogram. This results in a magnitude and frequency track for each harmonic component. As shown in Section 3.4 and Section 3.5 audio features are calculated from multiple categories and classified separately using Support Vector Machine (SVM) models. Finally, Section 3.6 describes a two-stage aggregation process that assigns a class of instrument labels for each note event.

### 3.1 Note Event Detection (Music Transcription)

The proposed algorithm is eventually to be applied on real-world polyphonic and multi-instrumental signals. In this case a prior detection of fundamental frequencies and note boundaries of individual note events must be performed. Current automatic transcription algorithms still yield typical detection error (note boundaries, pitch). In the case of defective note boundaries the presented approach can represent an improvement compared to a solely note-based feature extraction system: the classification of frame-wise features remains unaffected. However, pitch detection errors show their effect on the systems' partial tracking

procedure. The effects of the common *octave error*, where a pitch is detected an octave above or below the actual pitch, are limited to the number of correct partials detected, though. Let's say the given $f_0$ is an octave too high, in this case every other partial envelope is missed, the remaining are of undiminished quality. On the other hand, if the detected $f_0$ has been an octave below the true $f_0$, the system will try to find partial envelopes where none can be found. As will be explained in Section 3.3, such would be discarded due to low energy.

### 3.2 Pre-processing

The incoming samples are expected to be at a sampling rate of 44.1 kHz. All samples are normalized to their maximum value. Although this neglects dynamic information, we found in our experiments that normalization improves classification results slightly. Next, a 2048 point STFT with an overlap of 512 samples and a four-time zero-padding is performed, resulting in a time resolution of 86.13 frames per second and a window length of about 46.4 ms. The individual time frames have been weighted by a Hamming window function before transformation.

### 3.3 Partial Tracking

In order to identify the individual magnitude and frequency curves of the overtones, a tracking procedure is implemented. First, the frequency of the fundamental $f_0$ at a time shortly after the frame containing the envelope peak of the time-domain note event signal is chosen as the starting point for the search. Starting from this point, the loudest frequency bins within a pre-selected frequency band of the spectrogram are connected until a pre defined threshold value $L_{abort} = -39\,\mathrm{dB}$ is undershot.

To prevent from erratic search for partials in the noise, the search is also terminated when a detected noise floor is reached. This noise floor is determined as a mean value from a percentage $n$ of lowest energy magnitude bins in the spectrogram of the given note. In our experiments, we found $n = 15\,\%$ to work best.

For statistical recognition systems it is of great importance to obtain the best possible representations of the classes to be selected. Thus all partials that are to be classified are selected by means of a required minimum energy and length. Successive overtones in the harmonic series show increasing variance in their temporal development. We limited the number of partials to the fundamental frequency and 9 overtones, as this yielded the best performance in a conducted cross-validation experiment (see Figure 2).

Figure 3 illustrates the tracked partials for two different instruments. Individual partials of the piano note exhibit irregularities in magnitude progression due to *string beating* phenomena.[1]

[1] The so-called *string beating* occurs with string instruments (e.g., guitar or piano) and is caused through superimpositions of closely pitched oscillation modes.

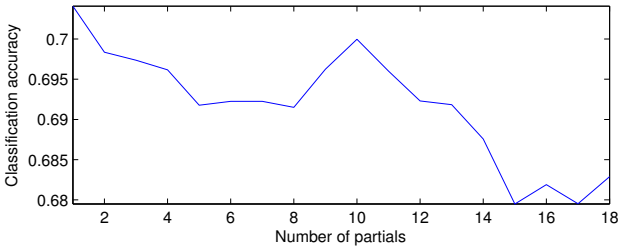



**Fig. 2.** Classification accuracy for partial-wise classification in dependence of the number of partials extracted.

### 3.4 Feature Extraction

The audio features that are extracted for each note are categorized into three groups:

1. *Note-wise features* - These global features are extracted once for each note event.
2. *Partial-wise features* - These features characterize the frequency course and the magnitude envelope of each detected partial of a note.
3. *Frame-wise features* - Theses features can be extracted in individual time frames both on spectrogram frames as well as on partial magnitude and frequency values.

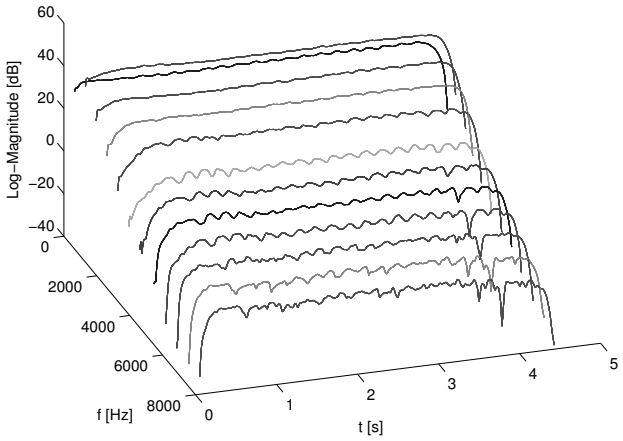
Figure 1 illustrates the different feature categories. In the following three sections, examples for all feature categories are explained in detail.

#### Note-wise Features

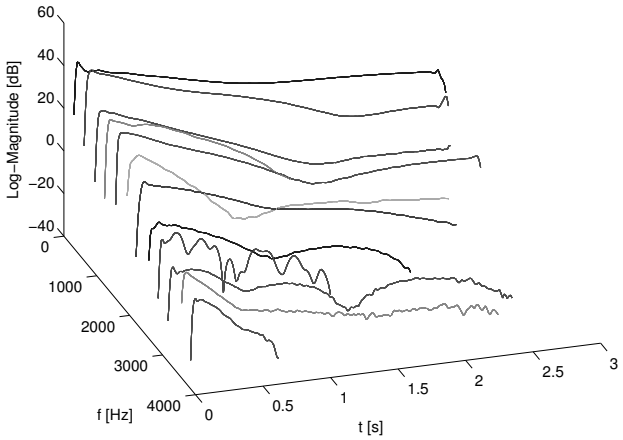
*Aggregated timbre-features* The first group of note-wise features is based on a simple two-stage envelope model that is used to partition the duration of a note event into an attack part and a decay part. Frame-wise timbre features such as spectral centroid, spectral crest factor, spectral decrease, spectral roll-off, and spectral slope are aggregated over both envelope parts by computing the statistical measures minimum, maximum, mean, median, and variance which are then used as features.

*Subharmonic energy* The second group of note-wise features characterizes the energy distribution in subharmonic regions, i.e., the spectral regions between the partials. First, a general measure of noisiness is computed by removing the harmonic components from the spectrum using a comb filter that is aligned to the $f_0$ curve. The aggregated energy in the remaining spectrogram is divided to the original spectrogram energy. Second, the presence of subharmonics[2] is

[2] Subharmonics can for instance be observed in flageolet tones from string instruments as well as in brass instrument notes.




(a) trumpet note




(b) piano note

**Fig. 3.** Two examples for the tracked partials in the logarithmic magnitude spectrum: a trumpet note and a piano note.

investigated in a similar way, only that additional comb filters are applied at different fractional positions between the note partials.

### Partial-wise Features

These features capture characteristics of the individual frequency and magnitude envelopes that have been detected in the spectrum of a single note. A first group of features is derived from the magnitude function $a_k(t)$ of an individual partial:

*Temporal centroid* The temporal centroid of a magnitude envelope is a powerful feature to separate sounds with transient excitation from such with a continuous excitation. For plucked string instruments the temporal centroid is very close to the start of the note (because, after an initial displacement of the string a note will just fade). For bowed string instruments, and also for notes that are played by wind instruments, the centroid will tend to be placed in the middle of the note, as energy is supplied to the sound over the duration of the note.

*Envelope energy* The energy is determined as the integral of the envelope function and is well suited to distinguish between short and long notes played.

*Fluctuation of magnitude* To calculate a measure of the local fluctuation of magnitude, the moving average of the given envelope function is withdrawn by means of a moving average filter. Variance as well as the zero crossing rate are then determined from the remaining envelope and used as features.

*Duration of attack and decay parts* For many instruments magnitude envelopes can be approximated with a simple, two-stage model consisting of attack and decay parts. As suggested in [10] the logarithmic time durations of both parts are used as features. In addition, the ratio between the durations of both parts is calculated.

*Polynomial approximation of the magnitude envelopes* A linear function of the magnitude envelope is approximated by linear regression (linear magnitude in the attack and logarithmic magnitude of the decay part). The coefficients of this function are then used as features.

The following features can be calculated for a frequency course $f_k(t)$ as well as a magnitude function $a_k(t)$ of an individual partial:

*Modulation of frequency & magnitude* Different physical phenomena such as

- frequency modulations (Vibrato) and
- magnitude modulations (Tremolo, string beating)

can be approximated with a number of periods of a sinusoid function. This function is retrieved from the modulation spectrum of the current frequency or magnitude envelope $f_k(t)$ or $a_k(t)$ by searching for salient peaks in a range from 2 - 20 Hz.

The following features are then calculated:

- Modulation frequency in Hz

– Modulation lift in cents (for frequency modulation) respectively without unit (for magnitude modulation)
– Dominance measure for strength of modulation (normalized value between 0 and 1)
– Number of oscillation periods

**Frame-wise Features**

The frame-based features are calculated either over the entire spectral frame or on the magnitudes and frequencies of the harmonics within a frame. As feature extraction showed to be more robust for the entire spectral frame as for the individual harmonics this feature group was divided into two categories to be classified separately: *frames spectral* and *frames harmonic*.

The first group of frame-wise features (*frames spectral*) consists of timbre features such as spectral centroid, spectral crest factor, spectral decrease, spectral roll-off, and spectral slope.

The second group (*frames harmonic*) regard the frequency positions and magnitudes of individual partials in a given spectral frame. Features of this category include the harmonic relative magnitudes, inharmonicity and tristimulus.

### 3.5 Classification

SVM Classifiers have proven to show good performance in instrument recognition tasks, as in [11]. Since the focus of this work is to evaluate the multiple–classification approach, we use SVM's to classify the observations of all three categories of extracted features for reasons of simplicity[3]. Before classification a normalization of the feature vectors to zero mean and unit variance is performed. The RBF function (radial basis function) is applied as the kernel function for the SVM classifier.

For an unknown note event, which is to be classified, first all features are calculated. Using the trained classifiers, vectors of class probabilities are determined for each feature group. As indicated in Figure 1 the number of vectors from the classifiers for the frames and overtones depend on the number of detected frames or overtones. The classifier, which is based on note-wise features returns exactly one probability vector.

### 3.6 Result Aggregation

The aim of the framework is to end up having a single vector of class probabilities from which the most likely instrument can be derived. This aggregation is performed in a two-stage combination scheme.

[3] The libSVM implementation, as described in [12] was used.

First, during the *within-ensemble* aggregation, the class probabilities within the feature categories are aggregated, so that from all classifiers ultimately one single vector is returned. Second, in the *between-ensemble* aggregation these vectors from the different feature categories are fused to a single vector, which is then returned by the framework. Additionally in this second stage of the aggregation process, *classifier weighting* is applied, in order to mirror relevance as supplied by the classifier accuracies.

The following methods of combination were examined for the two levels of aggregation:

*Mean* The aggregation result is formed by the mean values of the individual class probabilities.

*Highest Maximum* From the given observations the vector containing the highest single class probability is chosen.

*Best-to-Second* From the given observations the vector with the best difference between highest and second class probability is chosen.

*Majority Voting Hard* A majority voting scheme is conducted. A voice by weight of 1 is awarded to the class with highest probability in each probability vector.

*Majority Voting Soft* A majority voting scheme is conducted. A voice by weight of the highest probability is awarded for each probability vector.

For the combination of the final class probability vectors of the individual categories *classifier weighting* was applied. Each classifier was given a weight previously obtained from the cross-validation accuracies in the individual categories. Using this weight relevant decisions are given more significance during the aggregation process and the final classification result can be improved, as will be discussed in the results section.

The final output of the aggregation framework is a vector of class probabilities that has the length of the number of instrument classes.

## 4 Evaluation

For evaluation experiments a dataset consisting of 3300 isolated single note recordings of the 11 instruments shown in Table 1 was compiled. The samples were chosen randomly from three publicly available large-scale databases of instrument sounds: RWC Musical Instrument Sound Database [13], McGill University Master Samples [14] and IOWA Musical Instrument Samples [15]. The selected notes cover the entire pitch ranges of the instruments above a MIDI-pitch of 45 (A2). This constraint was made for spectral resolution to be ensured as sufficient for the partial tracking procedure. In order to obtain maximum

variance of the training data, all dynamic levels and a wide range of playing styles found in the individual databases were used in the dataset. For example, the flute samples cover the playing styles *normal, staccato, vibrato*, and *flutter* and the violin samples cover the playing styles *normal, staccato, ponticello*, and *vibrato*.

The prepared data was used to perform a 10-fold cross validation over all the included samples. Particular attention was paid to keep the partitions of the individual cross-validation folds alike for the different classification categories.

**Table 1.** Instruments used in the evaluation experiments.

| **Instrument** | RWC | Iowa | McGill |
|---|---|---|---|
| Piano | X | | X |
| Violin | X | X | X |
| Flute | X | X | X |
| Trumpet | X | X | X |
| Sax | X | | X |
| Ac.-Guitar | X | | |
| El.-Guitar | X | | |
| B3-Organ | X | | |
| Vocal | X | | |
| Cello | X | X | X |
| Clarinet | X | X | X |

## 5 Results

### 5.1 Aggregation Experiments

For the prepared dataset all combinations of methods for the two levels of aggregation were evaluated. For this procedure, all class probabilities of the different classification categories were stored for each fold of the cross-validation. Then, aggregation was performed and the accuracies of the individual folds were averaged.

Table 2 shows the accuracies obtained for the tested combination schemes. As we can see combination of class probabilities using the mean value in both aggregation stages led to the highest mean class accuracy of 91.2 %. The aggregation methods *Highest Maximum* and *Best-to-Second* scored very similar results, whereas hard majority voting between the individual feature categories brought less accurate classifications.

### 5.2 Results for the Individual Features Categories

As shown in Table 3 note-wise classification obtained the best accuracy of .85 . The applied spectral features and their statistical evaluation for the attack and

**Table 2.** Aggregated mean class accuracy values in percent for different combinations of result aggregation methods.

| **Inbetween-ensemble:** Mean<br>**Within-ensemble:** | Highest Max | Best to Second | Majority Hard | Majority Soft |
|---|---|---|---|---|
| Mean | **91.2** | 88.3 | 88.6 | 87.8 | 89.5 |
| Highest Max | 90.5 | 88.6 | 88.4 | 87.1 | 88.7 |
| Best-to-Second | 90.7 | 88.8 | 88.5 | 87.3 | 88.9 |
| Majority Hard | 89.8 | 87.2 | 85.7 | 87.8 | 88.5 |
| Majority Soft | 89.9 | 87.6 | 85.8 | 87.8 | 88.8 |

decay segments therefore capture characteristic properties of the instruments well. The classification of individual harmonics has surprisingly achieved an accuracy which exceeds the detection performance of individual spectral frames (accuracy of .68 for the overtones in comparison to an accuracy .60 for the frame-wise classification). This underlines the importance of the temporal information of instrument sounds. Furthermore, it appears that a classification can be successfully performed on the basis of isolated overtones for several instruments.

**Table 3.** Accuracies for individual classification categories.

| Aggregated | Note | Partials | Frames Spectral | Frames Harmonic |
|---|---|---|---|---|
| .91 | .85 | .68 | .60 | .32 |

### 5.3 Results for the Aggregated Classification

Figure 4 shows the confusion matrix for the best configuration of the aggregated classification. For each line, the classified samples of an instrument distribute to the available range of instruments, which are applied from left to right. A number of corresponding confusions can be found within the result. These include instruments with similar mechanisms of sound production. For example piano, acoustic guitar and electric guitar are all instruments with a transient sound excitation (piano notes are *struck* and guitar notes are *plucked* or *picked*), and are therefore confused more likely. The representatives of the woodwind family, flute, saxophone, and clarinet also form such a corresponding confusion group. If we have a closer look at the row for the violin, clarinet or saxophone it can be inferred that virtually no confusion is made between these instruments and such with a transient excitation. A rather unexpected result however, occurs with the (mostly) bowed instruments: violin and cello. On the one hand these show mutual confusions as could be expected, but also a noticeable tendency towards the woodwind-family instruments can be observed. This may be explained by

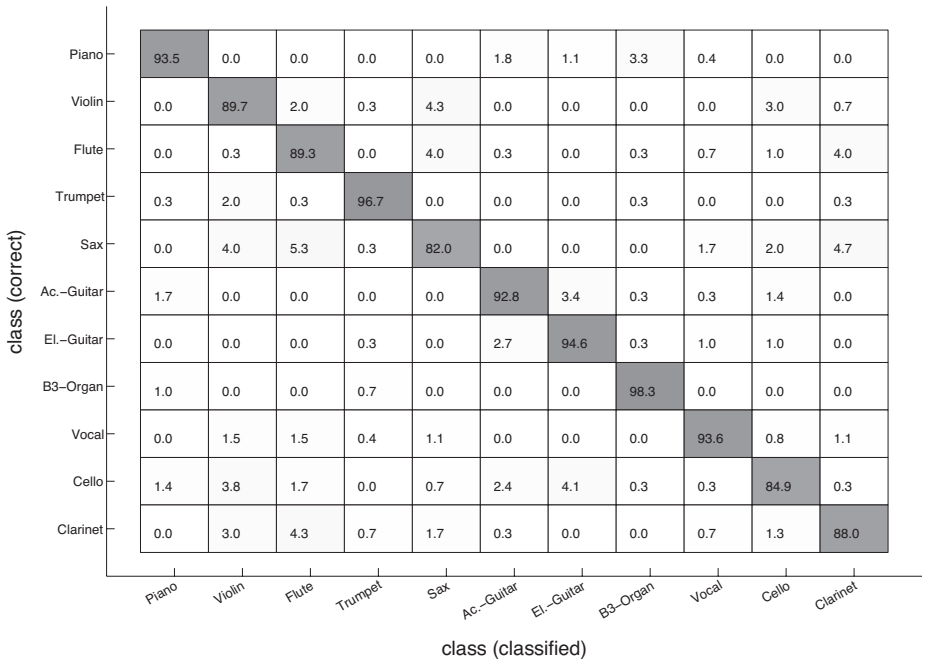



**Fig. 4.** Confusion matrix for the best configuration of the presented system.

the fact that both instrument families show considerable amounts of noise in their respective spectral distributions.

An examination of the individual classification categories reveals insights towards beneficial effects of the combined classification process:

1. *Improvements towards the best classification category* As we can see in Figure 4, 2.0 % of the violin samples have been misclassified as flute. For the individual feature categories the misclassification rates for this particular confusion were 2.7|8.3|1.9|14.4 % respectively[4]. The results of the best classification category (frame-wise spectral classification) have made a positive effect for the combined classification result.
2. *Improvements through synergistic effects* An ideal constellation for *classifier fusion* is to have classifiers that make different mistakes on a given sample. It is presumed that in this case correct results from an ensemble can overrule the mistakes of a single classifier. Final classification result for the Hammond B3 organ was 98.3 % (Figure 4) compared to 92.9|80.0|54.0|40.2 % for the individual feature categories. The frame-wise spectral classification showed strong confusions towards the trumpet (20.3 %), and the partial-wise

[4] note|partials|frames spectral|frames harmonic

classification tended towards cello (8.1 %) and the piano (6.7 %). As we see the combined result has been improved significantly, when compared with the best feature category.

These examples acknowledge for the effectiveness of the multiple expert classification approach on the instrument recognition task.

### 5.4 Comparison against other musical instrument recognition systems for isolated notes

The comparison to other state-of-the-art instrument recognition systems poses some difficulties, as taxonomies vary in terms of instruments and playing styles employed. A benchmark score has been reported by Tjoa and Liu in 2010 [16]. On a taxonomy of 24 instruments, including drum sounds, a mean class accuracy of 92.3 % was achieved. The average number of items per class is similar to the system presented in this paper, although it has to be pointed out that the number of items per class strongly varied.

## 6 Preliminary Experiment for Polyphonic Evaluation

A system for instrument recognition is of most use if it can be applied to a polyphonic data in the real-world case. Therefore we perform an additional evaluation experiment on artificial mixtures in this section. A set of training and test segments in two voices is assembled by which the classifiers for the individual feature categories are trained and evaluated. All individual single notes used in the training pieces are stored separately and then used to compare classifier performance on the basis of different training sets. Finally, we perform the aggregation of results with the winner method obtained from a prediction on the polyphonic training set.

*Selected Instruments* We selected seven classical orchestra instruments that are available in all three of the before mentioned publicly available large scale instrument databases: piano, violin, flute, cello, trumpet, saxophone, and clarinet. In RWC database selected instruments are present in three different instances, i.e. different manufacturer, recording studio, and executing musician. Together with the representations from IOWA, and McGill database we get a number of five individual instances for each instrument in total.

*Preparation of Artificial Mixtures* For the preparation of training and test mixtures we first divide the set of available samples by individual instrument instances. Selected samples for the experiment have the following origin:

- Training: IOWA, RWC1, and RWC2
- Test: McGill and RWC3

We avoid combining all samples for either training or test from only RWC database to beware of biassing effects.

The individual audio mixtures are then synthesized by placing an individual tone sample with required pitch at the time position of the notes that are played. All selected samples are normalized to their maximum absolute value.

*Melodic Material* The melodic material is based on midi-files of two-part inventions from J. S. Bach that are available for public download[5]. For each midi-file we consider all possible permutations of the instruments that were selected for the experiment, resulting in 49 different instrument combinations. This way every instrument happens to be playing either the left or right hand of the selected piano piece with every other instrument in combination. The midi-file segments are limited to a playing length of 12 seconds total and the first bar is omitted as it presents the theme of either composition for one hand only. BPM count of the selected midi-files was reduced by 50 % to ensure less problematic minimum note length.

*Experimental Setup* Figure 5 shows an overview for the setup of the conducted experiment.

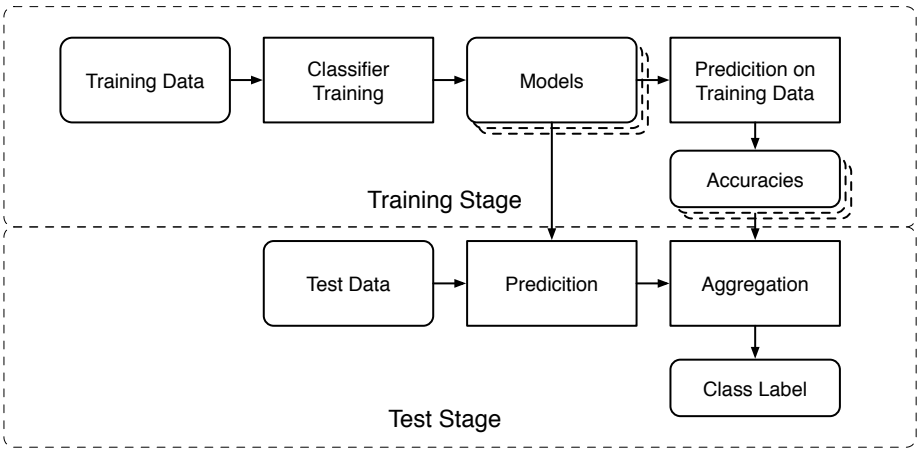



**Fig. 5.** Setup for polyphonic evaluation.

Results for individual feature categories and aggregated classification are shown in Table 4. We can see that also for classification on harmonically overlapped material the aggregated classification process can outperform the classifiers from single feature categories.

[5] `http://www.bachcentral.com/midiindexcomplete.html`

**Table 4.** Classification accuracies depending on training base for model training.

| | isolated and poly | polyphonic | isolated |
|---|---|---|---|
| Note | .37 | .38 | .25 |
| Frames Spectral | .33 | .34 | .28 |
| Frames Harmonic | .41 | .41 | **.36** |
| Partials | .33 | .39 | .22 |
| Aggregated | **.44** | **.45** | .28 |

## 7 Conclusions

We have presented a new approach towards instrument recognition making use of redundancies in harmonic structure and temporal development of a note. Features from multiple categories have been extracted and classified separately with SVM models. In a two-step combination scheme, class probabilities have been fused *within* and also *in-between* individual feature categories. As we have seen in the discussed results (Sections 5, 6) the aggregated classification outperforms the best classification result of the individual categories. Relevant decisions are obviously given more authority during the combination and classifier weighting process.

## 8 Outlook

Further development of the system shall include development of confidence measures for the individual observations of each classification category. This approach seems very beneficial for instrument recognition in polyphonic, multitimbral music. In this case unaffected signal portions, i.e. frames, that suffer from no temporal overlapping, or overtones, that are not spectrally overlapped, could be assigned with a high relevance in the classification process. Also, classification techniques such as source separation, training with real world data or the combination of adjacent single note decisions could be successfully combined with the presented strategy.

## 9 Acknowledgments

The authors would like to thank the reviewers for their comments and suggestions. This research work is a part of the SyncGlobal project. It is a 2-year collaborative research project between piranha womex AG from Berlin and Bach Technology GmbH, 4FriendsOnly AG, and Fraunhofer IDMT in Ilmenau, Germany. The project is co-financed by the German Ministry of Education and Research within the framework of an SME innovation program (FKZ 01/S11007).

## References


1. T. Kinoshita, S. Sakai, and H. Tanaka, "Musical sound source identification based on frequency component adaptation," in *Proceedings of the Workshop on Computational Auditory Scene Analysis (IJCAI-CASA)*, Stockholm, Sweden, 1999, pp. 18–24.
2. J. C. Brown, O. Houix, and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1064–1071, 2001.
3. A. Eronen, "Comparison of features for musical instrument recognition," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, New York, USA, 2001, pp. 19–22.
4. T. Kitahara, M. Goto, and H. G. Okuno, "Musical instrument identification based on F0-dependent multivariate normal distribution," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, China, 2003, vol. 5, pp. 421–424.
5. C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification With Application to Musical Instrument Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 174–186, Jan. 2009.
6. F. Fuhrmann, M. Haro, and P. Herrera, "Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 2009, pp. 321–326.
7. J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, 2012, pp. 559–564.
8. K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling," in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, number 3, pp. 3816–3819.
9. J. G. A. Barbedo and G. Tzanetakis, "Instrument identification in polyphonic music signals based on individual partials," in *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010, pp. 401–404.
10. G. Peeters, "A Large Set of Audio Features for Sound Description (similarity and classification) in the CUIDADO project," Tech. Rep., IRCAM, Paris, France, 2004.
11. F. Fuhrmann and P. Herrera, "Polyphonic instrument recognition for exploring semantic similarities in music," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010.
12. C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," Tech. Rep., Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2013.
13. M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database," in *Proceedings of the 4th International Society for Music Information Retrieval Conference (ISMIR)*, Baltimore, Maryland, USA, 2003, pp. 229–230.
14. F. Opolko and J. Wapnick, "The McGill University Master Samples Collection on DVD (3 DVDs)," 2006.
15. L. Fritts, "University of Iowa Musical Instrument Sample Database," 1997.

16. S. Tjoa and K. J. R. Liu, "Musical instrument recognition using biologically inspired filtering of temporal dictionary atoms," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 435–440.