AUTOREGRESSIVE MFCC MODELS FOR GENRE CLASSIFICATION IMPROVED BY HARMONIC-PERCUSSION SEPARATION

Halfdan Rump, Shigeki Miyabe, Emiru Tsunoo, Nobukata Ono, Shigeki Sagama
The University of Tokyo, Graduate School of Information Science and Technology
{rump,miyabe,tsunoo,onono,sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

In this work we improve accuracy of MFCC-based genre classification by using the Harmonic-Percussion Signal Separation (HPSS) algorithm on the music signal, and then calculate the MFCCs on the separated signals. The choice of the HPSS algorithm was mainly based on the observation that the presence of harmonics causes the high MFCCs to be noisy. A multivariate autoregressive (MAR) model was trained on the improved MFCCs, and performance in the task of genre classification was evaluated. By combining features calculated on the separated signals, relative error rate reductions of 20% and 16.2% were obtained when an SVM classifier was trained on the MFCCs and MAR features respectively. Next, by analyzing the MAR features calculated on the separated signals, it was concluded that the original signal contained some information which the MAR model was capable of handling, and that the best performance was obtained when all three signals were used. Finally, by choosing the number of MFCCs from each signal type to be used in the autoregressive modelling, it was verified that the best performance was reached when the high MFCCs calculated on the harmonic signal were discarded.

1. INTRODUCTION

Music information retrieval (MIR) is a diverse research field with many different areas of interest, such as chord detection, melody extraction etc. One of the popular tasks is classifying music into genres, which not only serves to ease organization of large music databases, but also drives the general development of features for representing the various important aspects of music. The task of genre classification draws upon many different kinds of information which means that one can either use features capable of expressing the music as a whole, or use many different types of features, each describing specific aspects of the music, such as the beat, melody, timbre etc. A low level feature frequently used for modelling music is the Mel-Frequency Cepstral Coefficients (MFCC), originally proposed in [1], (see [2] for a comprehensive review). The MFCCs are often calculated on the unaltered spectrum, thus containing information of all aspects of the music. The MFCCs effectively function as a lossy compression of a short part of the music signal into a small number of coefficients. It may happen that certain characteristics of the music signal which could be useful for genre classification are blurred by the compression. A possible way to resolve this issue is to break down the music signal into several signals, each containing a specific kind of information about the signal, and then calculate the MFCCs on the new signals. An example could be to separate the instruments and then calculate the MFCCs for the signals, each containing only a single instrument. However, it is possible that such a separation will fail, thus generating unpredictable results which might actually be worse than just using the original signal for classification. In this work we have used a simple algorithm that separates the music signal into two signals, one containing harmonics and the other containing percussion. The choice of this algorithm is based on some observations about the nature of the MFCCs, discussed in section 2.

After the music signal has been separated, MFCCs can be calculated on all three signals (original signal, harmonics and percussion). A classifier can be trained directly on the MFCCs, or more elaborate models can be constructed and used for classification. In this paper we investigate if higher classification performance can be achieved by separating the music signal as described above. We train a multivariate autoregressive (MAR) model on the MFCCs from the three signal types, and use it in a classifier.

The MAR model has proven to be efficient for the task of genre classification. First of all, the MAR model integrates the short time feature frames temporally, and secondly it is capable of modelling the covariances between the MFCCs. Since the ultimate goal of genre classification algorithms is to reach an accuracy of 100%, it is most meaningful to analyse the model with the highest accuracy. Therefore the article will focus mostly on the results obtained when using the MAR model for classification. Furthermore, by comparing performance of the MAR features calculated on the different signal types, it can be inferred which aspects of the music the MAR model analyses.

2. THE MEL-FREQUENCY CEPSTRAL COEFFICIENTS

The Mel-Frequency Cepstral Coefficient (MFCC) feature extraction is a useful way of extracting timbre information. The music signal is divided into a number of short time frames. For each frame, N_m coefficients are calculated, thus yielding N_m time series to be modelled by the MAR model, described in section 3.

In the following we explain the motivation for including a separation step by considering how the MFCCs are calculated. In the Mel filter-bank analysis, the bandwidth of each filter is linear for frequencies under around 1 kHz, and thereafter grows logarithmically. Therefore each of the lower Mel coefficients is the mean of a relatively narrow frequency band. If the spectrum is characterized by narrow pitch spikes, the difference between two adjacent Mel coefficients is likely to be large. Since the MFCCs are obtained by applying the DCT transform, these differences will be described by the high MFCCs. In other words, the high MFCCs are capable of closely fitting the pitch present in the frame on which they are calculated. Pitch is usually not a very good indicator for music genre, and therefore the high MFCCs should be discarded. On the other hand, if the spectrum has a smooth envelope the high order MFCCs will not model pitch, and therefore may be usable for genre classification. Most music signals contain both harmonics (pitch spikes) and percussion (smooth spectral envelope). Since the presence of pitch is harmful to the information content of the high MFCCs, it seems feasible to separate harmonics from percussion.

Furthermore it is possible that the shape of the spectral envelope of harmonics and percussion when they have been separated is useful for genre classification, and that the information content of the lower MFCCs will be improved by separating the music signal.

3. THE MULTIVARIATE AUTOREGRESSIVE MODEL

The MAR model is similar to the normal autoregressive model, in that it predicts the next sample of a time series as a linear combination of past samples. The MAR model extends the capabilities of the normal AR, as it capable of making predictions for multiple time series and utilizes correlations between time series for prediction. The prediction of the n'th N_m time series is calculated as

$$\mathbf{x}_n = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{x}_{n-I(p)} + \mathbf{u}_n \quad (1)$$

where $\mathbf{x}_n$ is a $N_m \times 1$ vector containing the predictions, and n is the frame index. P is the model order which specifies the number of time lags used for prediction. The MAR model is not constrained to using only time lags $1 \dots P$, but an arbitrary set of time lags $I = \{\tau_1 \dots \tau_P\}$ can be chosen. $\mathbf{A}_1 \dots \mathbf{A}_P$ are the $N_m \times N_m$ weight matrices for time lags $\tau_1 \dots \tau_P$. Element $[A]_p^{i,j}$ is the weight that controls how much of signal j, time-lagged τ_p samples, is used for prediction of signal i. $\mathbf{u}_n$ is the offset vector and can be omitted if each time series is subtracted by it's mean before estimating the coefficient matrices. The model parameters can be estimated by using the least mean squares approach. The P weight matrices $\mathbf{A}_1 \dots \mathbf{A}_P$ and the the offset vector $\mathbf{u}_n$ are stacked into a $PN_m^2 + N_m$ dimensional vector, and this constitutes the feature vector used for classification.

A basic assumption of the MAR model is that the time series upon which it is calculated has a stationary distribution. At first glance this assumption does not seem to go well with the nature of the percussive signal since it does not have a smooth time envelope. However, over longer periods roughly the same percussion sounds and thus MFCCs will appear again and again, which can be interpreted as stationarity. On the other hand, even though the harmonic signal has a smooth time envelope for a given note, meaning that the MFCCs will have a stationary distribution during the note, the distribution will change as the next note is struck. Since the exact same combination of harmonics, or in other words the same pitch spikes which are modelled by the high order MFCCs, is unlikely to occur more than maybe a few times, the distribution cannot be assumed stationary.

High order models are characterized by a high variance which gives them the power to fit closely to a time series, but also makes them prone to over-fitting. Low order models are more dominated by bias which makes them more suitable in cases where the signal envelope is the desired target. In [3], the MAR model was found to perform best with $P = 3$ when the task was genre classification, but the optimal value might differ according to the application for the reasons listed above.

4. HARMONIC-PERCUSSION SIGNAL SEPARATION

The Harmonic-Percussion Signal Separation (HPSS) algorithm proposed in [5], is a simple and fast method of dividing a musical signal, **N**, into two signals, **H** and **P**, each containing only the harmonic and percussive elements respectively. HPSS can be thought of as a two-cluster soft clustering, where each spectrogram grid-point is assigned a graded membership to a cluster representing harmonics and a cluster representing percussion. The algorithm uses the fact that percussion has a short temporal duration and is rich in noise, while harmonic elements have a long temporal duration with most of the signal energy concentrated in pitch spikes. Thus in the spectrogram, percussion appears as vertical lines of high power, whereas harmonic elements appear as horizontal lines.

In broad terms, the HPSS algorithm works by assuming independence between **H** and **P**, and using Bayes formula to calculate $p(\mathbf{H}, \mathbf{P}|\mathbf{N})$

$$\log p(\mathbf{H}, \mathbf{P}|\mathbf{N}) = \log p(\mathbf{N}|\mathbf{H}, \mathbf{P}) + \log p(\mathbf{H}) + \log p(\mathbf{P}) \quad (2)$$

The prior distributions $p(\mathbf{H})$ and $p(\mathbf{P})$ are defined as functions that measure the degree of smoothness in time and frequency respectively.

$$\log p(\mathbf{H}) = \sum_{\omega,\tau} \frac{-1}{2\sigma_H^2}(H^{\gamma}_{\omega,\tau-1} - H^{\gamma}_{\omega,\tau})^2 \quad (3)$$

$$\log p(\mathbf{P}) = \sum_{\omega,\tau} \frac{-1}{2\sigma_P^2}(P^{\gamma}_{\omega-1,\tau} - P^{\gamma}_{\omega,\tau})^2 \quad (4)$$

Where σ_H, σ_P and γ has been manually specified as in [5]. Thus the prior for **H** will be high when each row of the spectrogram is characterized by slow fluctuations, and similarly the prior for **P** will be high when this is the case for columns of the spectrogram. The likelihood function has been defined by measuring the I-divergence between **N** and **H** + **P**:

$$\log p(\mathbf{N}|\mathbf{H},\mathbf{P}) = \quad (5)$$
$$-\sum_{\omega,\tau}\left(N_{\omega,\tau}\log\frac{N_{\omega,\tau}}{H_{\omega,\tau}+P_{\omega,\tau}} - N_{\omega,\tau} + H_{\omega,\tau} + P_{\omega,\tau}\right)$$

and so the likelihood is maximized when $N_{\omega,\tau} = H_{\omega,\tau} + P_{\omega,\tau}$ for all ω and τ. The log-likelihood function is maximized by using the EM-algorithm. The update equations have been omitted in this work, but can be found in [5].

It is important to realize that since the HPSS algorithm is not a source separation algorithm but rather a decomposition of the original signal, no criteria of success has been defined, and so the algorithm cannot fail unless it fails to converge.

5. DATASET

We used the TZGENRE dataset proposed in [8]. The dataset has $N_s = 1000$ songs divided equally into 10 genres: blues, classic, country, disco, hip-hop, jazz, metal, pop, reggae and rock. Each song is a 30s sound snippet, and only one MAR model is calculated for the whole song. Other methods for calculating multiple MAR models on a single song and combining them afterwards can be found in [3] and [4].

6. EXPERIMENTAL SETUP

First the music signal was separated by using HPSS, and MAR features were calculated for each signal. If the MAR model is capable of using both harmonics and percussive elements at the same time, such a decomposition will not result in higher performance. However, if for instance the MAR model analyses the harmonic elements, then removing percussion will enable the MAR features to perform better. In the following, MAR features calculated on the harmonics, percussion and normal signals will be referred to as $\mathbf{m}_h$, $\mathbf{m}_p$, $\mathbf{m}_n$ respectively, whereas MFCCs will be referred to as $\mathbf{c}_h$, $\mathbf{c}_p$ and $\mathbf{c}_n$. In addition to the three single signal feature types, four combinations features of the MAR features and four combinations of the MFCCs were constructed: $\mathbf{m}_{hp}$, $\mathbf{m}_{hn}$, $\mathbf{m}_{pn}$, $\mathbf{m}_{hpn}$, $\mathbf{c}_{hp}$, $\mathbf{c}_{hn}$, $\mathbf{c}_{pn}$ and $\mathbf{c}_{hpn}$.

The sample-rate of the songs was 22.05 kHz. The MFCCs were calculated on 20 ms windows with an overlap of 10 ms. 40 filter-banks were used in the MFCC calculation. Since the number of MFCCs used to calculate the MAR features has a great influence on performance, each combination of features was evaluated with 19 different values of N_m. For each combination an $N_s \times D$ data matrix was created by stacking the N_s features vectors, each of dimension D. For features containing only MAR combinations, the dimension is $D = c(PN_m^2 + N_m)$, where $c \in \{1, 2, 3\}$ is the number of stacked MAR models.

The classifier used was a support vector machine with a Gaussian kernel. Kernel parameters σ and C were not tuned, but each column of the data matrix was normalized with respect to standard deviation. 500-fold cross validation was used for each of the 19 values of N_m, resulting in a $N_s \times 19$ matrix, where each column contained the average accuracy for each song for a given N_m. The overall performance for a given N_m was obtained by taking the mean of that column.

7. RESULTS

In this section the results of the experiments described in section 6 are presented and discussed.

7.1 Combining features from the separated signals

Figure 1 shows the classification performance of the seven combinations when the classifiers were trained directly on the MFCCs. The difference between the classifier trained on the MFCCs calculated on the original signal to the best performing feature, $\mathbf{c}_{hp}$, is 7.5%, corresponding to a relative error rate reduction of 20.0%. This is a significant improvement, and confirms that the MFCCs have problems expressing both harmonic and percussive information when present at the same time.

$\mathbf{c}_h$ reaches its near peak performance for low N_m. This means that for the harmonic signal, very little usable information is contained in the high MFCCs. The MFCCs are fairly low-dimensional which means that the SVM classifier is still able to achieve optimal performance, and thus performance only degrades slightly. Performance of $\mathbf{c}_p$ keeps increasing when including more MFCCs, meaning that the higher MFCCs in the percussion signal contains usable information. Furthermore, the performance gained by including higher MFCCs is more than for the harmonics signal but less than for the percussion signal. This confirms that the presence of harmonics degrades the information quality of the higher MFCCs.

Next, we use the MAR model for classification and test performance of $\mathbf{m}_h$, $\mathbf{m}_p$ and $\mathbf{m}_n$, and of the combinations of them. The performance of the seven combination features is shown on Figure 2. $\mathbf{m}_n$ is the most powerful of the three single model features peaking with a performance of 74.1%. Pleasingly, all three single model features have a lower performance than the combination features. $\mathbf{m}_{hnp}$ had a peak performance of 77.6%, a gain of 3.6% compared to the best single signal model.

As was also seen when using the MFCCs in the classifier, $\mathbf{m}_{hp}$ performs significantly better than $\mathbf{m}_n$. This shows that the autoregressive modelling of the MFCCs cal-

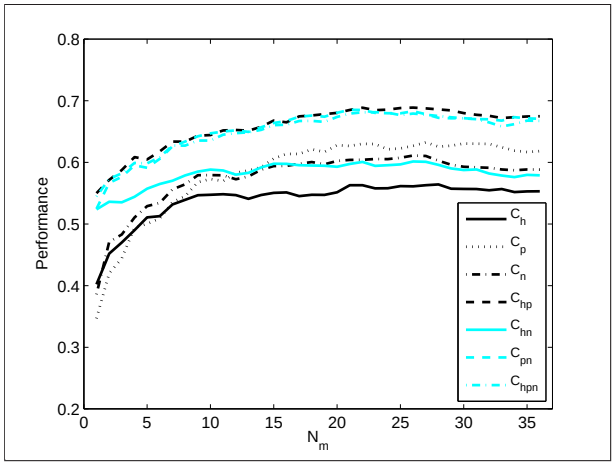

Figure 1. Performance curves for the classifier trained on MFCCs

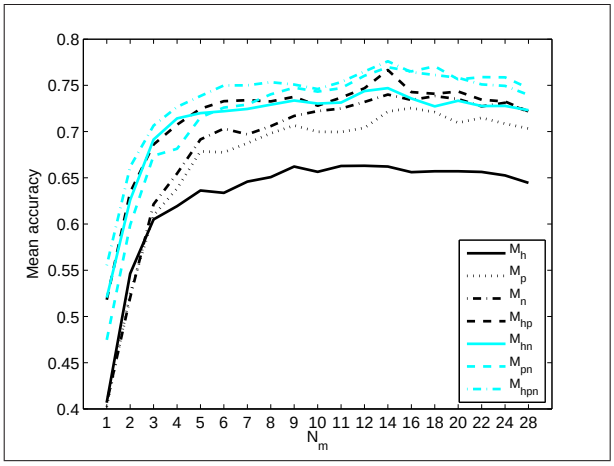

Figure 2. Performance curves for the classifier trained on MAR features

culated on the original signal cannot compensate for the MFCCs' inability to handle the mixture of harmonic and percussive information.

An important difference between using MFCCs or MAR features in the classifier is that $\mathbf{m}_{hpn}$ outperformed $\mathbf{m}_{hp}$, whereas $\mathbf{c}_{hpn}$ and $\mathbf{c}_{hp}$ had the same level of performance. Thus the MAR model is capable of modelling some properties of the original signal **N**, which are present in neither **H** nor **P**. More specifically, the MAR model can in some cases predict percussion from harmonics or vice versa, due to the autoregressive modelling. This is a reasonable claim when keeping in mind that the HPSS algorithm is not a source separation algorithm, and that some instruments will produce both harmonics and percussive sounds.

As an example, when a note is played on a piano the hammer hits the string causing it to vibrate, resulting in a sound with a high attack part and a slowly declining envelope. Since this will happen every time the piano is used, the MAR model can use the attack part to make a prediction about the rest of the sound. When using HPSS to separate the signal however, percussion is assumed to be independent from harmonics, and the attack part, which is rich in noise and has a short temporal duration, is assigned to the percussion signal while the rest of the sound is assigned to the harmonic signal. When this happens the MAR model can no longer model the dependencies, so including MAR features calculated on the original signal increases performance.

7.2 Differences between the signal type MAR features

In this section we analyse some of the differences between the MAR features calculated on each of the separated signals.

An important step towards understanding the MAR features and specify their application domain is to investigate to which degree features calculated on the different signal types classify the same songs or not. In the former case, classification accuracy with different signal types is largely genre dependent, and in the latter case there will be some

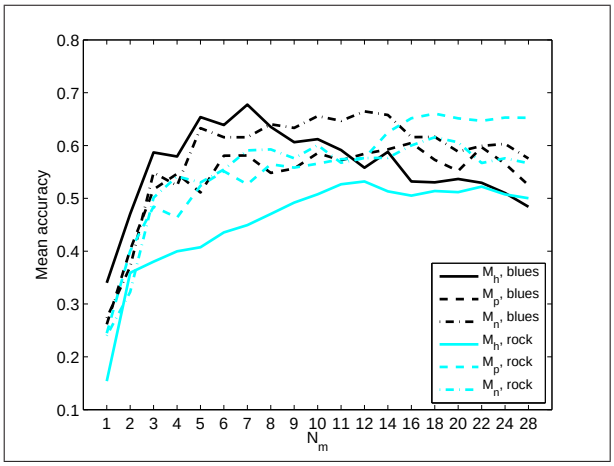

Figure 3. Examples of genre specific performance, only MAR features

easy songs which can be classified by all signal models, and some hard songs that only the features with an overall high performance can classify.

Analysis is carried out by finding the point where all signal models have approximately the same accuracy, and calculating the correlation between the $N_s \times 1$ song accuracy vectors. It was observed that there is a low correlation between which songs $\mathbf{m}_h$ and $\mathbf{m}_p$ classify. This suggests that the two signal models contains different information which allows for the classification of different songs, and thus are efficient with different kinds of music. For most genres $\mathbf{m}_n$ is slightly better than $\mathbf{m}_p$, with $\mathbf{m}_h$ being the worst performing of the three. However, for some genres $\mathbf{m}_h$ achieves the best performance when the high MFCCs were discarded, as can be seen on Figure 3. Furthermore, the fact that the correlation of the song classification vectors of $\mathbf{m}_p$ and $\mathbf{m}_n$ was high, means that they classify more of the same songs than $\mathbf{m}_h$ and $\mathbf{m}_n$, which is consistent with the fact that $\mathbf{m}_{hn}$ and $\mathbf{m}_p$ classify more of the same songs than $\mathbf{m}_{pn}$ and $\mathbf{m}_h$. These results suggest that MAR

Feature	Performance	Relative ERR
$\mathbf{c}_n$	61.1%	N/A
$\mathbf{c}_{hp}$, Constr.	68.9%	20.0%
$\mathbf{m}_n$	74.1%	N/A
$\mathbf{m}_{hpn}$, Constr.	77.6%	13.5%
$\mathbf{m}_{hpn}$, N.Constr.	78.3%	16.2%

Table 1. Overview of the best performing features. Constr. or N.Constr. refer to the constraint on N_m.

features calculated on the original music reflect the percussive elements to a higher degree than the harmonics elements.

The fact that $\mathbf{m}_{pn}$ is even higher than $\mathbf{m}_{hp}$ seems like a contradiction to the statement made earlier that $\mathbf{m}_n$ is more correlated with $\mathbf{m}_p$ than with $\mathbf{m}_h$. The explanation to this is most likely that the gains from combining uncorrelated features, i.e. $\mathbf{m}_{hp}$ and $\mathbf{m}_{hn}$, cannot match the penalty caused by the low performance of $\mathbf{m}_h$. Although $\mathbf{m}_p$ and $\mathbf{m}_n$ are somewhat correlated, there are still some differences in what songs they classify, and this seems to results in a performance gain when combined.

7.3 Selecting N_m for each signal type

Figure 2 in section 7.1 shows that the MAR features calculated on the different signal types perform best for different values of N_m. In this section we investigate if performance can be improved by removing the constraint that the number of MFCCs used to calculate the MAR model must be the same for all signal types. Since it is possible that simply combining the best performing models does not achieve the highest performance, the five best models of each signal type were used to form a number of combination features.

Figure 4 shows the performance plotted versus the dimensionality of the feature vector, using the same number of MFCCs, and with different number of MFCCs. The figure makes it easy to compare feature efficiencies, as a point that is situated higher and on the left side of another point of the same type, means that a feature of lower dimensionality had higher performance.

From Figure 4 it seems that the method of selecting N_m for each single MAR model is not particularly capable of producing low dimensional features, but the method do achieve the highest overall performance. However, since it is in general infeasible to try all combinations of N_m before selecting the best one, a general tendency must be discovered. In section 2 it was suggested that the high MFCCs calculated on the harmonics signal should be discarded, whereas high MFCCs from the percussion signal could be used. This was the case when the classifier was trained directly on the MFCCs, and when the classifier was trained on the MAR features. It is not surprising therefore, that the best performance of 78.3% was obtained by discarding the high MFCCs for the harmonic signal and using high MFCCs from the percussion signal.

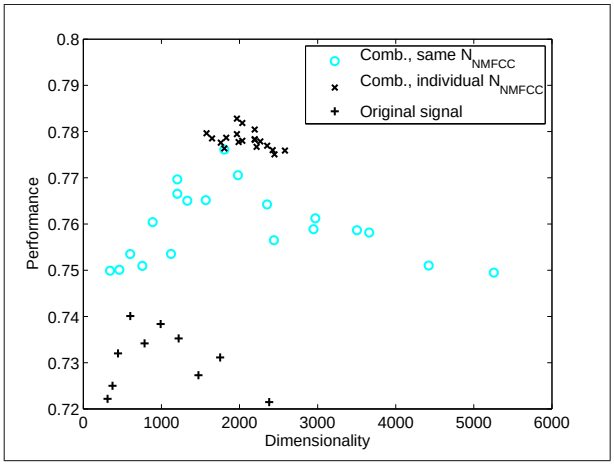

Figure 4. Performance and dimensionality of combination models

8. PERFORMANCE DEMONSTRATION

This section contains a short demonstration of the performance obtained when combining the improved features with two other features types, each describing different aspects of music. The first type is the Rhythm Map features, proposed in [6], which are calculated on the percussion signal. A song is represented as a ten dimensional vector, each element describing the membership to a rhythmic template extracted from the entire dataset. The second feature type, henceforth referred to as TZ-features, represents a song as an 68-dimensional vector containing a set of timbre related features proposed in [8]. The Rhythm Map is of special interest since it is calculated on the percussive signal provided by the HPSS algorithm, and thus provide no information about the harmonics. The TZ-features were chosen because they were tested in combination with Rhythm Map (see [7]), where it was shown that the two feature types compliment each other well. An accuracy of 75.0% was obtained on the dataset by the combination of Rhythm Map and TZ-features. When the MAR features calculated on the original signal were included as well, a performance of 80.1% was achieved. Finally, by separating the signal with HPSS and calculating MAR features on the three signals as proposed, a performance of 82.46% was obtained, corresponding to a relative error rate reduction of 12.0%.

9. CONCLUSION

In this work we proposed that separating the music signal into more signals, each containing certain characteristics of the original signal, could produce better features, leading to increased performance in the task of music genre classification. Based on the observation that the presence of harmonics causes the high MFCCs to be noisy, we used the HPSS algorithm to separate the signal into two signals, one containing harmonics and the other containing percussion. The separation increased performance significantly, both when the classifier was trained on the MFCCs and when it was trained on the MAR features. The best perfor-

mance obtained with the MAR features was 78.3%, corresponding to a relative error rate reduction of 16.2%. It was seen that the MAR model uses both harmonic and percussive information to make predictions, but that the percussive information seems to be the dominating. The fact that the best performance was reached when the MAR features from the separated signals were combined with the original signal showed us that the MAR-model could, to some extend, model dependencies between harmonic and percussive elements. The combination of MFCCs calculated on the harmonics signal and MFCCs calculated on the percussion signal performed better than MFCCs calculated on the original signal, and this was interpreted as an inability of the MFCCs to model the presence of both harmonics and percussion in the same signal. An important conclusion of this is that separating the music signal as proposed simply creates better low level features, which means that models trained on these features will also be improved.

10. REFERENCES

[1] S. B. Davis and P. Marmelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, Signal Proces.*, Vol. 28, No. 4, pp. 357–366, 1980

[2] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," *Proc. ISMIR*, 2000

[3] A. Meng, P. Ahrendt, J. Larsen and L. K. Hansen, "Temporal Feature Integration for Music Genre Classification," *IEEE Trans. Audio, Speech, Lang. Proces.*, Vol. 15, No. 5, pp. 1654–1664, 2007

[4] J. S. Shawe-Taylor and A. Meng, "An Investigation of Feature Models for Music Genre Classification Using the Support Vector Classifier," *Proc. ISMIR*, pp. 604–609, 2005

[5] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive componets in music signals," *Proc. ISMIR*, pp. 139–144, 2008

[6] E. Tsunoo, N. Ono and S. Sagayama, "Rhythm map: Extraction of unit rhythmic patterns and analysis of rhythmic structure from music acoustic signals," *Proc. ICASSP*, pp. 185–188, 2009

[7] E. Tsunoo, G. Tzanetakis, N. Ono and S. Sagayama, "Audio genre classification using percussive pattern clustering combined with timbral features," *Proc. ICME*, pp. 382–385, 2009

[8] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech, Audio Processing*, Vol. 10, No. 5, pp. 293–302, 2002