Drum sound detection in polyphonic music with hidden Markov models

Jouni Paulus and Anssi Klapuri*
Department of Signal Processing
Tampere University of Technology
Korkeakoulunkatu 1, Tampere, Finland

Abstract

This paper proposes a method for transcribing drums from polyphonic music using a network of connected hidden Markov models (HMMs). The task is to detect the temporal locations of unpitched percussive sounds (such as bass drum or hi-hat) and recognise the instruments played. Contrary to many earlier methods, a separate sound event segmentation is not done, but connected HMMs are used to perform the segmentation and recognition jointly. Two ways of using HMMs are studied: modelling combinations of the target drums, and a detector-like modelling of each target drum.

Acoustic feature parametrisation is done with mel-frequency cepstral coefficients and their first order temporal derivatives. The effect of lowering the feature dimensionality with principal component analysis and linear discriminant analysis is evaluated. Because the acoustic properties of drum sounds may vary between the training and target signals, unsupervised acoustic model parameter adaptation with maximum likelihood linear regression is evaluated. The performance of the proposed method is evaluated on a publicly available data set containing signals with and without accompaniment (non-drum instruments) and compared with two reference methods. The results suggest that the transcription is possible using connected HMMs, and that using detector-like models for each target drum provides a better performance than modelling drum combinations.

Keywords: hidden Markov model, maximum likelihood linear regression, mel-frequency cepstral coefficient, linear discriminant analysis

*This work was supported by the Academy of Finland, (application number 129657, Finnish Programme for Centres of Excellence in Research 2006–2011).

1 Introduction

This paper applies connected hidden Markov models (HMMs) to the transcription of drums from polyphonic musical audio. For brevity, the word "drum" is here used to refer to all the unpitched percussions met in Western pop/rock music, such as bass drum, snare drum, and cymbals. The word "transcription" is used to refer to the process of locating drum sound onset instants and recognising the drums played. The analysis result enables several applications, such as using the transcription to assist beat tracking [11], drum track modification in the audio [30], reusing the drum patterns from existing audio, or musical studies on the played patterns.

Several methods have been proposed in the literature to solve the drum transcription problem. Following the categorisation made in [5] and [10], majority of the methods can be viewed to be either *segment and classify* or *separate and detect* approaches. The methods in the first category operate by segmenting the input audio into meaningful events, and then attempt to recognise the content of the segments. The segmentation can be done by detecting candidate sound onsets or by creating an isochronous temporal grid coinciding with most of the onsets. After the segmentation a set of features is extracted from each segment, and a classifier is employed to recognise the contents. The classification method varies from a naive Bayes classifier with Gaussian mixture models (GMMs) [19] to support vector machines (SVMs) [24, 10] and decision trees [21].

The methods in the second category aim at segregating each target drum into a separate stream and to detect sound onsets within the streams. The separation can be done with unsupervised methods like sparse coding [26] or independent subspace analysis (ISA) [25], but these

require recognising the instruments from the resulting streams. The recognition step can be avoided by utilising prior knowledge of the target drums in the form of templates, and applying a supervised source separation method. Combining ISA with drum templates produces a method called prior subspace analysis (PSA) [4]. PSA represents the templates as magnitude spectrograms and estimates the gains of each template over time. The possible negative values in the gains do not have a physical interpretation and require a heuristic post-processing. This problem was solved using non-negative matrix factorisation (NMF) restricting the component spectra and gains to be non-negative. This approach was shown to perform well when the target signal matches the model (signals containing only target drums) [18].

Some methods cannot be assigned to either of the categories above. These include *template matching and adaptation* methods operating with time-domain signals [33], or with a spectrogram representation [31].

The main weakness with the "segment and classify" methods is the segmentation. The classification phase is not able to recover any events missed in the segmentation without an explicit error correction scheme, e.g., [29]. If a temporal grid is used instead of onset detection, most of the events will be found, but the expressivity lying in the small temporal deviations from the grid is lost, and problems with the grid generation will be propagated to subsequent analysis stages.

To avoid making any decisions in the segmentation, this paper proposes to use a network of connected HMMs in the transcription in order to locate sound onsets and recognise the contents jointly. The target classes for recognition can be either combinations of drums or detectors for each drum. In the first approach, the recognition dictionary consists of combinations of target drums with one model to serve as the background model when no combination is played, and the task is to cover the input signal with these models. In the detector approach, each individual target drum is associated with two models: a "sound" model and a "silence" model, and the input signal is covered with these two models for each target drum independently from the others.

In addition to the HMM baseline system, the use of model adaptation with maximum likelihood linear regression (MLLR) will be evaluated. MLLR adapts the acoustic models from training to better match the specific input.

The rest of this article is organised as follows: Section 2 describes the proposed HMM-based transcription method, Section 3 details the evaluation setup and presents the obtained results, and finally Section 4 presents the conclusions of the paper. Parts of this work have been published earlier in [16] and [17].

2 Proposed method

Figure 1 shows an overview of the proposed method. The input audio is subjected to sinusoids-plus-residual modelling to suppress the effect of non-drum instruments by using only the residual. Then the signal is subdivided into short frames from which a set of features is extracted. The features serve as observations in HMMs that have been constructed in the training phase. The trained models are adapted with unsupervised maximum likelihood linear regression [12] to match the transcribed signal more closely. Finally, the transcription is done by searching an optimal path through the HMMs with Viterbi algorithm. The steps are described in more detail in the following.

2.1 Feature extraction and transformation

It has been noted, e.g., in [8] and [31], that suppression of tonal spectral components improves the accuracy of drum transcription. This is no surprise, as the common drums in pop/rock drum kit contain a notable stochastic component and relatively little tonal energy. Especially the idiophones (e.g., cymbals) produce mostly noise-like signal, while the membranophones (skinned drums) may contain also tonal components [6]. The harmonic suppression is here done with simple sinusoids-plus-residual modelling [15, 22]. The signal is subdivided into 92.9 ms frames, the spectrum is calculated with discrete Fourier transform, and 30 sinusoids with the largest magnitude are selected by locating the 30 largest local maxima in the magnitude spectrum. The sinusoids are then synthesised and the resulting signal is subtracted from the original signal. The residual serves as the input to the following analysis stages. Even though the processing may remove some of the tonal components of the membranophones, the remaining ones and the stochastic components are enough for the recognition. Preliminary experiments also suggest that the exact number of removed components is not important, even doubling the number to 60 caused only an insignificant drop in the performance.

The feature extraction calculates 13 mel-frequency cepstral coefficients (MFCCs) in 46.4 ms frames with

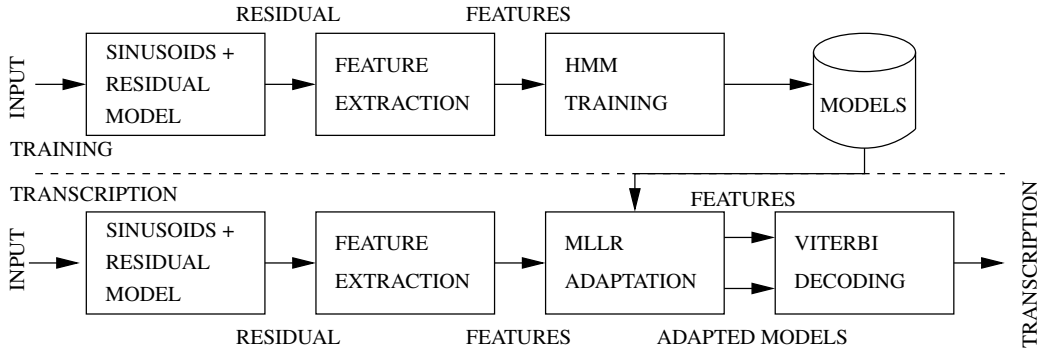

Figure 1: A block diagram of the proposed HMM transcription method including acoustic model adaptation.

75% overlap [1]. In addition to the MFCCs their first order temporal derivatives are estimated. The zeroth coefficient which is often discarded is also used. MFCCs have proven to work well in a variety of acoustic signal content analysis tasks including instrument recognition [2]. In addition to the MFCCs and their temporal derivatives, other spectral features, such as band energy ratios, spectral kurtosis, skewness, flatness, and slope used, e.g., in [24] were considered for the feature set. However, preliminary experiments suggested that their inclusion reduces the overall performance slightly and they are not used in the presented results. The reason for this degradation is an open question to be addressed in the future work, but is assumed that the features do not contain enough additional information compared to the original set to compensate the increased modelling requirements.

The resulting 26-dimensional feature vectors are normalised to have zero mean and unity variance in each feature dimension over the training data. Then the feature matrix is subjected to dimensionality reduction. Though unsupervised transformation with principal component analysis (PCA) has been successfully used in some earlier publications, e.g., [23], it did not perform well in our experiments. It is assumed that this is because PCA attempts only to describe the variance of the data without class information, and it may be distracted by the amount of noise present in the data.

The feature transformation used here is calculated with linear discriminant analysis (LDA). LDA is a class-aware transformation attempting to minimise intra-class scatter while maximising inter-class separation. If there are N different classes, LDA produces a transformation to $N-1$ feature dimensions.

2.2 HMM topologies

Two different ways to utilise connected HMMs for drum transcription are considered: drum sound combination modelling and detector models for each target drum. In the first case, each of the 2^M combinations of M target drums is modelled with a separate HMM. In the latter case, each target drum has two separate models: a "sound" model and a "silence" model. In both approaches the recognition aims to find a sequence of the models providing the optimal description of the input signal. Fig. 2 illustrates the decoding with combination modelling, while Fig. 3 illustrates the decoding with drumwise detectors.

The main motivation for the combination modelling is that in popular music multiple drums are often hit simultaneously. However, the main weakness is that as the number of target drums increases, the number of combinations to be modelled also increases rapidly. Since only the few most frequent combinations cover most of the occurrences, as illustrated in Fig. 4, there is very little training data for the more rare combinations. Furthermore, it may be difficult to determine whether or not some softer sound is present in a combination (e.g., when kick and snare drums are played, the presence of hi-hat may be difficult to detect from the acoustic information) and a wrong combination may be recognised.

With detector models, the training data can be utilised more efficiently than with combination models, because all combinations containing the target drum can be used to train the model. Another difference in the training phase is that each drum has a separate silence (or background) model.

As will be shown in Sec. 3, the detector topology generally outperforms the combination modelling which was found to have problems with overfitting the

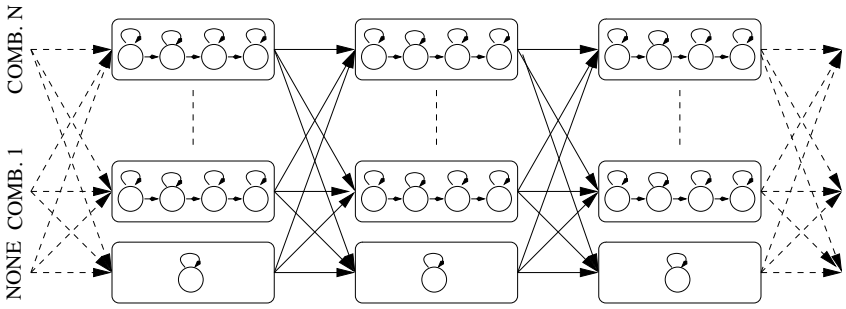

Figure 2: Illustration of the basic idea of drum transcription with connected HMMs for drum combinations. The decoding aims to find the optimal path through the models given the observed acoustic information.

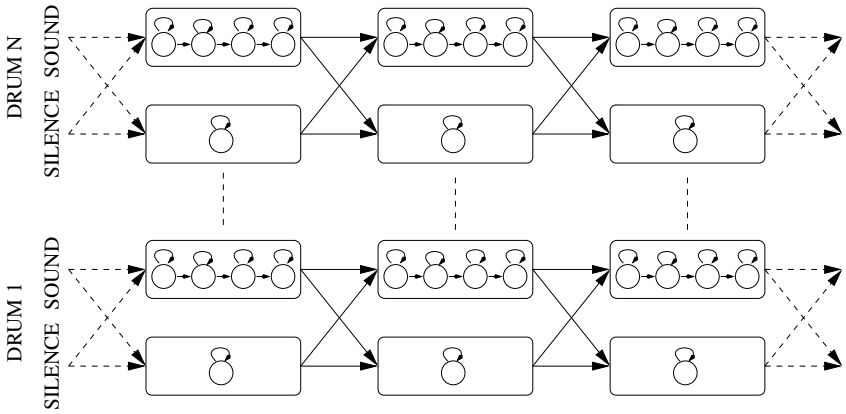

Figure 3: Illustration of the basic idea of drum transcription with HMM-based drum detectors. Each target drum is associated with two models, "sound" and "silence", and the decoding is done for each drum separately.

limited amount of training data. This was indicated by the following observations: performance degradation with increasing the number of HMM training iterations and acoustic adaptation, and slight improvement in the performance with simpler models and reduced feature dimensions. Because of this, the results on acoustic model adaptation and feature transformations is presented only for the detector topology (similar choice has been done, e.g., in [10]). For the sake of comparison, however, results are reported also for the combination modelling baseline.

The sound models consist of a four-state left-to-right HMM where a transition is allowed to the state itself and to the following state. The observation likelihoods are modelled with single Gaussian distributions. The silence model is a single-state HMM with a 5-component GMM for the observation likelihoods. This topology was chosen because the background sound does not have a clear sequential form. The number of states and GMM components were empirically determined.

The models are trained with expectation maximisation algorithm [20] using segmented training examples. The segments are extracted after annotated event onsets using a maximum duration of 10 frames. If there is another onset closer than the set limit, the segment is truncated accordingly. In detector modelling, the training instances for the "sound" model are generated from the segments containing the target drum, and the remaining frames are used to train the "silence" model. In combination modelling, the training instances for each combination are collected from the data, and the remaining frames are used to train the background model.

2.3 Acoustic adaptation

Unsupervised acoustic adaptation with maximum likelihood linear regression (MLLR) [12] has been successfully used to adapt the HMM observation density parameters, e.g., in adapting speaker independent models to speaker dependent models in speech recogni-

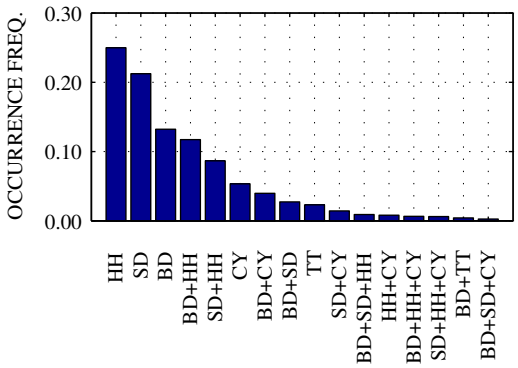

Figure 4: Relative occurrence frequencies of various drum combinations in "ENST drums" [9] data set. Different drums are denoted with BD (bass drum), CY (all cymbals), HH (all hi-hats), SD (snare drum), and TT (all tom-toms). Two drum hits were defined to be simultaneous if their annotated onset times differ less than 10 ms. Only the 16 most frequent combinations are shown.

tion [12], language adaptation from Spanish to Valencian [13], or to utilise a recognition database trained for phone speech to recognise speech in car conditions [3]. The motivation for using MLLR here is that it is assumed that the acoustic properties of the target signal always differ from those of the training data, and the match between the model and the observations can be improved with adaptation. The adaptation is done for each target signal independently to provide models that fit the specific signal better. The adaptation is evaluated only for the detector topology, because for drum combinations, the adaptation was not successful, most likely due to the limited amount of observations.

In single variable MLLR for the mean parameter, a transformation matrix

$$\boldsymbol{W} = \begin{pmatrix} w_{1,1} & w_{1,2} & 0 & \dots & 0 \\ w_{2,1} & 0 & w_{2,3} & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ w_{n,1} & 0 & \dots & 0 & w_{n,n+1} \end{pmatrix} \quad (1)$$

is used to apply a linear transformation to the GMM mean vector μ so that the likelihood of the adaptation data is maximised. The mean vector μ with the length n is transformed by

$$\mu' = \boldsymbol{W}[\omega, \mu^{\mathsf{T}}]^{\mathsf{T}}, \quad (2)$$

where the transformation matrix has the dimensions of $n \times (n+1)$, and $\omega = 1$ is a bias parameter. The non-zero elements of $\boldsymbol{W}$ can be organised into a vector

$$\hat{\mathbf{w}} = [w_{1,1}, \dots, w_{n,1}, w_{1,2}, \dots, w_{n,n+1}]^{\mathsf{T}}. \quad (3)$$

The value of the vector can be calculated by

$$\hat{\mathbf{w}} = \left[\sum_{s=1}^{S} \sum_{t=1}^{T} \gamma_s(t) \boldsymbol{D}_s^{\mathsf{T}} \boldsymbol{C}_s^{-1} \boldsymbol{D}_s \right]^{-1} \left[\sum_{s=1}^{S} \sum_{t=1}^{T} \gamma_s(t) \boldsymbol{D}_s^{\mathsf{T}} \boldsymbol{C}_s^{-1} \mathbf{o}(t) \right], \quad (4)$$

where t is frame index, $\mathbf{o}(t)$ is the observation vector from frame t, s is an index of GMM components in the HMM, $\boldsymbol{C}_s$ is the covariance matrix of GMM component s, $\gamma_s(t)$ the occupation probability of s^{th} component in frame t (calculated, e.g., with the forward-backward algorithm), and matrix $\boldsymbol{D}_s$ is defined as a concatenation of two diagonal matrices

$$\boldsymbol{D}_s = [\boldsymbol{I}\omega, \text{diag}(\mu_s)], \quad (5)$$

where μ_s is the mean vector of the s^{th} component and $\boldsymbol{I}$ is a $n \times n$ identity matrix [12]. In addition to the single variable mean transformation, also full matrix mean transformation [12] and variance transformation [7] were tested. In the evaluations, the single variable adaptation performed better than the full matrix mean transformation, and therefore the results are presented only for it. Variance transformation reduced performance in all cases.

The adaptation is done so that the signal is first analysed with the original models. Then it is segmented to examples of either class ("sound" / "silence") based on the recognition result, and the segments are used to adapt the corresponding models. The adaptation can be repeated using the models from the previous adaptation iteration for segmentation. It was found in the evaluations that applying the adaptation repeatedly for three times produced the best result even though the obtained improvement after the first adaptation was usually very small. Further increment of the number of adaptation iterations from this started to degrade the results.

2.4 Recognition

In the recognition phase, the (adapted) HMM models are combined into a larger compound model, see Figs. 2 and 3. This is done by concatenating the state transition matrices of the individual HMMs and incorporating the inter-model transition probabilities in the same matrix. The transition probabilities between the models are estimated from the same material that is used for training

the acoustic models, and the bigram probabilities are smoothed with Witten-Bell smoothing [28]. The compound model is then used to decode the sequence with Viterbi algorithm. Another alternative would be to use token passing algorithm [32], but since the model satisfies the first order Markov assumption (only bigrams are used), Viterbi is still a viable alternative.

3 Results

The performance of the proposed method is evaluated using the publicly available data set "ENST drums" [9]. The data set allows adjusting the accompaniment (everything else but the drums) level in relation to the drum signal, and two different levels are used in the evaluations: a balanced mix and a drums-only signal. The performance of the proposed method is compared with two reference systems: a "segment and classify" method by Tanghe et al. [24], and a supervised "separate and detect" method using non-negative matrix factorisation [18].

3.1 Acoustic data

The data set "ENST drums" contains multichannel recordings of three drummers playing with different drum kits. In addition to the original multichannel recordings, also two downmixes are provided: "dry" with minimal effects, mainly having only the levels of different drums balanced, and "wet" resembling the drum tracks on commercial recordings, containing some effects and compression. The material in the data set ranges from individual hits to stereotypical phrases, and finally to longer tracks played along with an accompaniment. These "minus one" tracks played on accompaniment have the synchronised accompaniment available as a separate signal allowing to create polyphonic signals with custom mixing levels. The ground truth for the data set contains the onset times for the different drums, and was provided with the data set.

The "minus one" tracks are used as the evaluation data. They are naturally split into three subsets based on the player and kit, each having approximately the same number of tracks (two with 21 tracks and one with 22). The lengths of the tracks range from 30 s to 75 s with mean duration of 55 s. The mixing ratios of drums and accompaniment used in the evaluations are drums-only and a "balanced" mix. The former is used to obtain a baseline result for the system with no accompaniment.

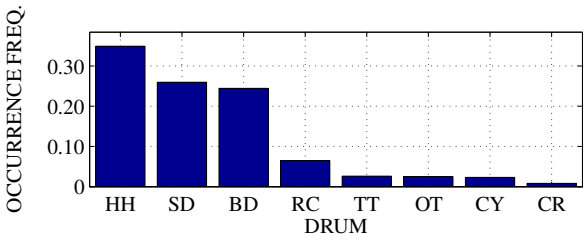

Figure 5: Occurrence frequencies of different drums in "ENST drums" data set. The instruments are denoted by: BD (bass drum), CR (all crash cymbals), CY (other cymbals), HH (open and closed hi-hat), RC (all ride cymbals), SD (snare drum), TT (all tom-toms), and OT (other unpitched percussion instruments, e.g., cow bell).

The latter, corresponding to applying scaling factors of $2/3$ for the drum signal and $1/3$ for the accompaniment, is used then to evaluate the system performance in realistic conditions met in polyphonic music.[1]

3.2 Evaluation setup

Evaluations are run using a three-fold cross-validation scheme. Data from two drummers are used to train the system and the data from the third are used for testing, and the division is repeated three times. This setup guarantees that the acoustic models have not seen the test data and their generalisation capability will be tested. In fact, the sounds of the corresponding drums in different kits may differ considerably (for example, depending on the tension of the skin, the use of muffling in case of kick drum, or the instrument used to hit the drum that can be a mallet, a stick, rods, or brushes) and using only two examples of a certain drum category to recognise a third one is a difficult problem. Hence, in real applications the training should be done with as diverse data as possible.

The target drums in the evaluations are bass drum (BD), snare drum (SD), and hi-hat (HH). The target set is limited to these three for two main reasons. First, they are found practically in every track in the evaluation data and they cover a large portion of all the drum sound events, as can be seen from Fig. 5. Secondly, and more importantly, these three instruments convey the

[1]The mixing levels are based on personal communication with O. Gillet, and result into an average of -1.25 dB drums-to-accompaniment ratio over the whole data set.

main rhythmic feel of most of the popular music songs, and occur in a relatively similar way in all the kits.

In the evaluation of the transcription result, the found target drum onset locations are compared with the locations given in the ground truth annotation. The hits are matched to the closest hit in the other set so that each hit has at most one hit associated to it. A transcribed onset is accepted as correct if the absolute time difference to the ground truth onset is less than 30 ms.[2] When the number of events is G in the ground truth and E in the transcription result, and the number of missed ground truth events and inserted events are m and i respectively, the transcription performance can be described with precision rate

$$P = (E - i)/E \tag{6}$$

and recall rate

$$R = (G - m)/G. \tag{7}$$

These two metrics can be further summarised by their harmonic mean, F-measure

$$F = (2PR)/(P + R). \tag{8}$$

3.3 Reference methods

The system performance is compared with two earlier methods: a "segment and classify" method by Tanghe et al. [24], and a "separate and detect" method by Paulus and Virtanen [18]. The former, referred to as *SVM* in the results, was designed for transcribing drums from polyphonic music by detecting sound onsets and then classifying the sounds with binary SVMs for each target drum. An implementation of the original author is used [14]. The latter, referred to as *NMF-PSA*, was designed for transcribing drums from a signal without accompaniment. The method uses spectral templates for each target drum and estimates their time-varying gains using NMF. Onsets are detected from the recovered gains. Also here the original implementation is used. The models for the SVM method are not trained specifically for the data used, but the generic models provided are used instead. The spectral templates for NMF-PSA are calculated from the individual drum hits in the data set used here. In the original publication the mid-level representation used spectral resolution of five bands. Here they are replaced with 24 Bark bands for improved frequency resolution.

[2]When comparing the results obtained with the same data set in [10], it should be noted that there the allowed deviation was 50 ms.

Table 1: Evaluation results for the tested methods using the balanced drums and accompaniment mixture as input.

Method	Metric	BD	SD	HH	Total
HMM	$P(\%)$	84.7	65.3	84.9	80.0
	$R(\%)$	77.4	44.9	78.5	68.0
	$F(\%)$	80.9	53.2	81.6	**73.5**
HMM+MLLR	$P(\%)$	80.2	66.3	84.7	79.0
	$R(\%)$	81.5	45.3	82.6	70.9
	$F(\%)$	80.8	53.9	83.6	**74.7**
HMM comb	$P(\%)$	54.9	38.8	73.0	55.0
	$R(\%)$	66.4	47.0	58.7	57.4
	$F(\%)$	60.1	42.5	65.1	**56.1**
NMF-PSA [18]	$P(\%)$	69.9	57.0	58.2	62.0
	$R(\%)$	57.9	16.7	53.5	43.6
	$F(\%)$	63.4	25.9	55.8	**51.2**
SVM [24]	$P(\%)$	80.9	65.9	47.1	54.3
	$R(\%)$	38.4	14.2	69.5	43.8
	$F(\%)$	51.1	23.4	56.1	**48.5**

3.4 Results

The evaluation results are given in Tables 1 and 2. The former contains the evaluation results in the case of the "balanced" mixture as the input, while the latter contains the results for signals without accompaniment. The methods are referred to as

- *HMM*: The proposed HMM method with detectors for each target drum without acoustic adaptation.
- *HMM+MLLR*: The proposed detector-like HMM method including the acoustic model adaptation with MLLR.
- *HMM comb*: The proposed HMM method with drum combinations without acoustic adaptation.
- *NMF-PSA*: A "separate and detect" method using NMF for the source separation, proposed in [18].
- *SVM*: A "segment and classify" method proposed in [24] using SVMs for detecting the presence of each target drum in the located segments.

The results show that the proposed method performs best among the evaluated methods. In addition, it can be seen that the acoustic adaptation slightly improves the recognition result. All the evaluated methods seem to have problems in transcribing the snare drum (SD),

Table 2: Evaluation results for the tested methods using signals without any accompaniment as input.

Method	Metric	BD	SD	HH	Total
HMM	$P(\%)$	95.7	68.7	82.7	82.5
	$R(\%)$	88.1	57.7	80.9	75.9
	$F(\%)$	91.8	62.7	81.8	**79.1**
HMM+MLLR	$P(\%)$	94.1	75.0	83.8	84.8
	$R(\%)$	92.1	56.7	84.9	78.4
	$F(\%)$	93.1	64.6	84.4	**81.5**
HMM comb	$P(\%)$	71.5	41.3	63.8	57.5
	$R(\%)$	74.2	54.3	55.3	60.4
	$F(\%)$	72.8	46.9	59.3	**58.9**
NMF-PSA [18]	$P(\%)$	85.0	75.6	57.1	68.5
	$R(\%)$	80.1	38.1	67.7	62.2
	$F(\%)$	82.5	50.7	61.9	**65.2**
SVM [24]	$P(\%)$	95.4	62.9	61.1	68.2
	$R(\%)$	54.0	37.9	72.3	56.6
	$F(\%)$	69.0	47.3	66.2	**61.9**

even without the presence of accompaniment. One reason for this is that the snare drum is often played in more diverse ways than, e.g., the bass drum. Examples of these include producing the excitation with sticks or brushes, or playing with and without the snare belt, or by producing barely audible "ghost hits".

When analysing the results of "segment and classify" methods it is possible to distinguish between errors in segmentation and classification. However, since the proposed method aims to perform these tasks jointly, acting as a specialised onset detection method for each target drum, this distinction cannot be made.

An earlier evaluation with the same data set was presented in [10, Table II]. The table section "Accompaniment $+0$ dB" in there corresponds to the results presented in Table 1 and section "Accompaniment $-\infty$ dB" corresponds to the results in Table 2. In both cases, the proposed method clearly outperforms the earlier method in bass drum and hi-hat transcription accuracy. However, the performance of the proposed method on snare drum is slightly worse.

The improvement obtained using the acoustic model adaptation is relatively small. Measuring the statistical significance with two-tailed unequal variance Welch's t-test [27] on the F-measures for individual test signals produces p-value of approximately 0.64 for the balanced mix test data and 0.18 for the data without accompaniment suggesting that the difference in the results is not statistically significant. However, the adaptation seems to provide a better balance on precision and recall rates. The performance differences between the proposed detector-like HMMs and the other methods are clearly in favour of the proposed method.

Table 3: Effect of feature transformation on overall F-measure (%) of detector HMMs without acoustic model adaptation.

	none	PCA 90%	LDA
Plain drums	63.6	66.0	79.1
Balanced mix	59.6	60.9	73.5

Table 3 provides the evaluation results with different feature transformation methods while using detector-like HMMs without acoustic adaptation. The results show that PCA has a very small effect on the overall performance while LDA provides a considerable improvement.

4 Conclusions

This paper has studied and evaluated different ways of using connected HMMs for transcribing drums from polyphonic music. The proposed detector-type approach is relatively simple with only two models for each target drum: a "sound" and a "silence" model. In addition, modelling of drum combinations instead of detectors for individual drums was investigated, but found not to work very well. It is likely that the problems with the combination models are caused by overfitting the training data. The acoustic front-end extracts mel-frequency cepstral coefficients (MFCCs) and their first order derivatives to be used as the acoustic feature. Comparison of feature transformations suggests that LDA provides a considerable performance increase with the proposed method. Acoustic model adaptation with MLLR is tested, but the obtained improvement is relatively small. The proposed method produces a relatively good transcription of bass drum and hi-hat, but snare drum recognition has some problems that need to be addressed in future work. The main finding is that it is not necessary to have a separate segmentation step in a drum transcriber, but the segmentation and recognition can be performed jointly with an HMM even in the presence of accompaniment and with bad signal-to-noise ratios.

References

[1] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, Aug. 1980.

[2] A. Eronen. Comparison of features for musical instrument recognition. In *Proc. of 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 19–22, New Platz, N.Y., USA, Oct. 2001.

[3] A. Fischer and V. Stahl. Database and online adaptation for improved speech recognition in car environments. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 445–448, Phoenix, Ariz., USA, 1999.

[4] D. FitzGerald, B. Lawlor, and E. Coyle. Prior subspace analysis for drum transcription. In *Proc. of 114th Audio Engineering Society Convention*, Amsterdam, The Netherlands, Mar. 2003.

[5] D. FitzGerald and J. Paulus. Unpitched percussion transcription. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*, pages 131–162. Springer, 2006.

[6] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer-Verlag New York Inc., New York, N.Y., USA, second edition, 1998.

[7] M. J. F. Gales, D. Pye, and P. C. Woodland. Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation. In *Proc. of the Fourth International Conference on Spoken Language Processing*, pages 1832–1835, Philadelphia, Pa., USA, Oct. 1996.

[8] O. Gillet and G. Richard. Drum track transcription of polyphonic music using noise subspace projection. In *Proc. of 6th International Conference on Music Information Retrieval*, pages 156–159, London, England, UK, Sept. 2005.

[9] O. Gillet and G. Richard. ENST-Drums: an extensive audio-visual database for drum signal processing. In *Proc. of 7th International Conference on Music Information Retrieval*, pages 156–159, Victoria, B.C., Canada, Oct. 2006.

[10] O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):529–540, Mar. 2008.

[11] M. Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, 30(2):159–171, 2001.

[12] J. Leggetter, C and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185, 1995.

[13] M. Luján, C. D. Martínez, and V. Alabau. Evaluation of several maximum likelihood linear regression variants for language adaptation. In *Proc. of Sixth International Language Resources and Evaluation Conference*, Marrakech, Morocco, May 2008.

[14] MAMI. Musical audio-mining, drum detection console applications, 2005. `http://www.ipem.ugent.be/MAMI/`.

[15] R. J. McAulay and F. Quatieri, Thomas. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, Aug. 1986.

[16] J. Paulus. Acoustic modelling of drum sounds with hidden Markov models for music transcription. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 241–244, Toulouse, France, May 2006.

[17] J. Paulus and A. Klapuri. Combining temporal and spectral features in HMM-based drum transcription. In *Proc. of 8th International Conference on Music Information Retrieval*, pages 225–228, Vienna, Austria, Sept. 2007.

[18] J. Paulus and T. Virtanen. Drum transcription with non-negative spectrogram factorisation. In *Proc. of 13th European Signal Processing Conference*, Antalya, Turkey, Sept. 2005.

[19] J. K. Paulus and A. P. Klapuri. Conventional and periodic N-grams in the transcription of drum sequences. In *Proc. of IEEE International Confer-*

ence on Multimedia and Expo, volume 2, pages 737–740, Baltimore, Md., USA, July 2003.

[20] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–289, Feb. 1989.

[21] V. Sandvold, F. Gouyon, and P. Herrera. Percussion classification in polyphonic audio recordings using localized sound models. In *Proc. of 5th International Conference on Music Information Retrieval*, pages 537–540, Barcelona, Spain, Oct. 2004.

[22] X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Picialli, and G. De Poli, editors, *Musical Signal Processing*, pages 91–122. Swets & Zeitlinger, 1997.

[23] P. Somervuo. Experiments with linear and nonlinear feature transformations in HMM based phone recognition. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume I, pages 52–55, Hong Kong, 2003.

[24] K. Tanghe, S. Dengroeve, and B. De Baets. An algorithm for detecting and labeling drum events in polyphonic music. In *Proc. of First Annual Music Information Retrieval Evaluation eXchange*, London, England, UK, Sept. 2005. Extended abstract.

[25] C. Uhle, C. Dittmar, and T. Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. of 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 843–848, Nara, Japan, Apr. 2003.

[26] T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *Proc. of International Computer Music Conference*, pages 231–234, Singapore, Oct. 2003.

[27] B. A. Welch. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, Jan. 1947.

[28] I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transcations on Information Theory*, 37(4):1085–1094, 1991.

[29] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. An error correction framework based on drum pattern periodicity for improving drum sound detection. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 237–240, Toulouse, France, May 2006.

[30] K. Yoshii, M. Goto, and H. G. Okuno. INTER:D: A drum sound equalizer for controlling volume and timbre of drums. In *Proc. of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, pages 205–212, London, England, UK, Nov. 2005.

[31] K. Yoshii, M. Goto, and H. G. Okuno. Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):333–345, Jan. 2007.

[32] S. J. Young, N. H. Russell, and J. H. S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Department, Cambridge, UK, July 1989.

[33] A. Zils, F. Pachet, O. Delerue, and F. Gouyon. Automatic extraction of drum tracks from polyphonic music signals. In *Proc. of 2nd International Conference on Web Delivering of Music*, pages 179–183, Darmstadt, Germany, Dec. 2002.