Detection of musically relevant regions in multiresolution time-frequency representations evaluated on piano recordings

Nicolas Figueiredo **Marcelo Queiroz**
Computer Science Department, University of São Paulo
{nsf,mqz}@ime.usp.br

ABSTRACT

This paper investigates possible estimators of musical information in subregions of a time-frequency representation of an audio signal, in the context of obtaining multiresolution time-frequency representations through iterative refinements. An experiment was conducted comparing estimators extracted from STFT spectrograms of Disklavier performances to reference values extracted from the corresponding MIDI files. Different reference values were considered, capturing the presence of musical information in the time-frequency plane that are relevant to music information retrieval tasks such as automatic music transcription and timbre analysis. The impact of the size of the time-frequency subregions and initial resolution of the STFT were analyzed. The influence of the introduction of simple energy decay models in the reference values was also investigated. A second experiment was conducted evaluating chosen estimators as features in a predictive model for the binary detection of musically relevant regions of a time-frequency representation. Naive-Bayes models were trained using binary piano rolls (with and without harmonics) as ground truth. Results show that it is possible to detect musically relevant regions of a time-frequency representation with satisfactory results using Shannon and Rényi entropies.

1. INTRODUCTION

The spectrogram is a widely used analysis/representation tool in sound and music computing, despite the fact that the linear frequency resolution of the Short-Time Fourier Transform (STFT) is not particularly adequate for applications dealing with music events organized within log-frequency strata. The trade-off between time and frequency resolutions makes the STFT representation far from ideal for signals with both melodic and percussive events of interest, since either percussive events or melodic events will be poorly located in the time-frequency plane (TFP) according to the analysis window chosen. This is a burden for tasks such as automatic music transcription (AMT) that depend on the precise detection of both note onset times and melody contours, especially when the analysis of expressive elements with small-scale variations (e.g. vibrato) is also targeted.

These limitations motivate the development of adaptive transforms, which are representations that vary both time and frequency resolution in different regions of the TFP according to the contents of the analyzed signal. These representations allow more representation space to be used in subregions which are more relevant to the task at hand and should be more detailed, whilst reducing representation space for irrelevant subregions. Such structured representations lend themselves relatively easily to MIR tasks that depend on local TFP information (e.g. onset detection, frequency peak estimation). On the other hand, off-the-shelf machine learning methods that expect uniform TFP representations would probably require some form of decoding and segmentation prior to input.

In [1], a general structure for adaptive transforms is presented, where several spectrograms with different resolutions are precomputed for the entire signal, and then used to compose a multiresolution spectrogram. A specific algorithm is also shown, in which the TFP is divided into subregions, and for each subregion the sparsest representation among a set of alternatives is chosen, where sparsity is measured by an entropy-related metric. A formal mathematical framework for the analysis, transformation and resynthesis of a signal with adaptive time-frequency resolution based on nonstationary Gabor frames is developed in [2]. Rényi entropies are used as a sparsity measure for the choice between different sets of analysis windows at each time frame of a signal, and a resynthesis method is provided along with a theoretical upper bound for its error. A similar algorithm is given in [3]: the TFP is divided into rectangular regions, and for each one the sparsest (according to Rényi entropy) Gabor representation is chosen among two pre-computed representations. Then, this initial representation is subtracted from the original signal. The resulting residual signal is again approximated using the same adaptive algorithm, and the process is iterated until a certain criterion is met, resulting in a layered representation of the original signal.

This general structure for adaptive transforms is very inefficient: several regions of these multiresolution representations are discarded after their computation, causing a huge computational overhead. This motivates the investigation of a more efficient representation that avoids the unnecessary computation of prior high-resolution representations of the TFP, performing them only if there is

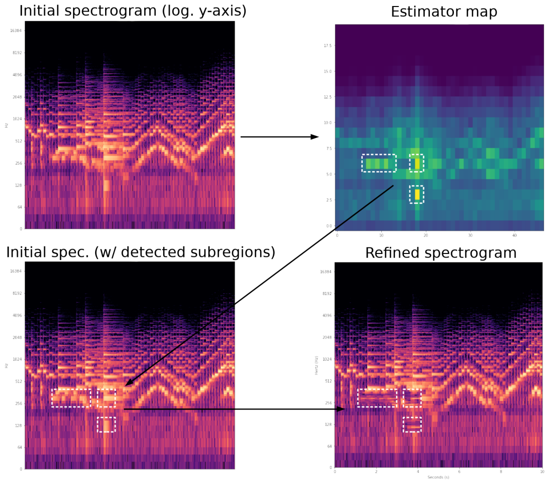

Figure 1. A single refinement step of an algorithm for adaptive multiresolution representation algorithm based on a relevant musical subregion estimator. From an initial generic spectrogram (plotted with a logarithmic Y-axis), three subregions (represented by the white dotted lines) are detected as musically interesting, and then transported to the original spectrogram in order to be refined via localized high resolution STFT computations.

evidence of the presence of musical events within a particular time-frequency subregion (this is what we call a *relevant* subregion). This way, such a multiresolution spectrogram could eventually use the same amount of data of a fixed-resolution spectrogram to provide a more precise representation, by using coarser resolutions in areas that do not contain musical events and higher resolutions in musically relevant regions, while at the same time reducing the computational overhead by not throwing away any computed data.

What differentiates our approach from the discussed related work is that our criterion for choosing local resolutions is not the comparison between pre-computed representations of different resolutions, but the use of a musical information estimator. Moreover, in the presented experiment we compare several candidates for such an estimator with respect to their actual ability to detect musically relevant regions in a TFP, a problem which was not addressed in the related literature. This is the focus of this paper, which characterizes, to our knowledge, a novel approach to multiresolution representation of audio signals.

In order to develop such a representation, we need to find a reliable way to detect musically relevant regions of a fixed-resolution spectrogram. These are here defined as regions that contain sound events (characterized by their pitch, loudness, duration and timbre [4]) as well as expressive events such as pitch bends, tremolo and vibrato. This article presents two experiments that intend to answer how well we can identify these regions and what would be the best estimators for this task. A motivational application for such estimators is presented in Fig. 1, which illustrates the application of a musical information estimator as part of an algorithm for obtaining an adaptive multiresolution time-frequency representation. The algorithm starts with an initial single-resolution spectrogram, and a relevance estimator (in this example, the energy density) is computed for subregions of the TFP (in this case, rectangles of 500 ms by 500 cent). The resulting map is interpreted as an index to the presence of music events that should be represented with a higher resolution. The three most prominent subregions in this example have been selected, within which high-resolution STFT representations were computed (using sub-band processing for computational efficiency) and inserted in the corresponding subregions of the initial spectrogram, resulting in a multiresolution representation. This refinement step could then be repeated in other subregions until a certain predefined criterion is met, e.g. a memory or time constraint, or a desired accuracy in the determination of further information from this spectrogram, such as onsets or instantaneous frequencies.

Section 2 presents an overview of both experiments conducted, as well as a justification behind the choice of estimators and reference values. Section 3 presents in detail technical information about the estimator evaluation experiment, and results are shown in Section 4. A technical description of the detection experiment is given in Section 5, followed by the results in Section 6. Finally, Section 7 discusses the outcomes of both experiments and their applicability in a multiresolution representation, as well as future work.

2. METHODOLOGY

The main objective of the proposed experiments is to investigate possible musical information estimators in the TFP. In the first experiment, we evaluate how well different estimators correlate to reference values relating to some MIR tasks. In the second experiment, we evaluate chosen estimators in a binary classification task to further evaluate their applicability as part of an algorithm for a multiresolution representation.

2.1 First experiment: estimator evaluation

For the first experiment, reference values were extracted from a MIDI file, while the estimators were extracted from an STFT spectrogram taken from a corresponding performance synchronized to such MIDI file. This is made possible through the use of piano performances recorded in an Yamaha Disklavier: for each performance in the utilized dataset, there are corresponding MIDI and audio files aligned within a 3 ms accuracy. Fig. 2 gives an overview of the steps involved in this experiment: from the MIDI file, we build a density map of musical events (MIDI notes, possibly including harmonic partials of such notes), to be used as ground truth; from the sound recording, we build a density map using a candidate estimator (such as entropies or spectrogram statistics, see Section 2.1.2). Both density maps are computed over rectangular TFP subregions measured in ms (width) by cents (height). Finally, groundtruth and estimator density maps are compared according to their Pearson correlation coefficient.

2.1.1 Musical information reference values

Three different reference values are extracted from each MIDI file (see Fig. 3):

1. Piano roll: from the note-on/off and sustain pedal position events, a simple binary piano roll representation (no velocity information) is built;

2. Piano roll with harmonics: the simple binary piano roll is augmented with the inclusion of 7 harmonics for each note present;

3. Piano roll with harmonics and decay models: a piano roll with velocity information is built from the MIDI events contained in the file, along with 7 harmonics for each note. Then, linear energy decay models are used to attenuate each note over the time axis and each harmonic over the frequency axis. For modeling the energy decay over time and frequency, we used models based on measured data from acoustic pianos [5, 6]. For each note, a decay of 8 dB/s is considered, and seven harmonics are included with a decay of 4.3 dB per partial. MIDI velocity values in [5] were translated using estimates from [7], as decays of 28 velocity points per second and 7.86 velocity points per partial.

After their extraction from the MIDI file, each matrix is divided into subregions defined by dimensions given in cents by miliseconds. The mean value inside each of these subregions is considered in order to form the reference value matrix. It is important to note that the reference value (and thus our definition of a musically relevant subregion) is strongly dependant on the subregion size used: as implemented, it represents the overall "note content" of each subregion, and the size of these subregions greatly influences this.

Each estimator derived from the spectrogram was compared against each of these reference values, in order to give insight into their applicability under different circumstances. The correlation of an estimator with the simple piano roll should give us an idea of its applicability to the AMT task, since a piano roll would be a possible representation format for the transcription. In a representation built specifically for AMT, piano roll events should be given more importance, and should be represented with a finer resolution with respect to the remaining subregions of the TFP. The correlation of an estimator with the piano roll with harmonics allows us to evaluate its sensibility to the presence of harmonics, which are an integral part of the timbre of nearly every musical instrument, and would show up on a spectrogram, but not necessarily in a music score. Harmonics might be of interest to timbre analysis, and so estimators that correlate well with this augmented piano roll would be candidates for producing multiresolution time-frequency representations for timbre-related tasks. Finally, the introduction of decay models into the piano roll represents an attempt to include a very simple acoustic instrument model: this may be suited to test the invariability of each estimator in relation to specific instrument timbres, but also to look for estimators that would be useful to study dynamic (i.e. time-varying) aspects of timbre.

2.1.2 Estimators

Rényi entropies [2, 3, 8] and energy variance [9] have both been employed as musical information estimators in the TFP. For this experiment, the Rényi entropy was computed with $\alpha = 3$, a value justified by the discussion in [10]. The Shannon entropy was also considered in our evaluation. The standard deviation was used in place of variance, in order to preserve the same scale of the original energy or amplitude data. All of these estimators were extracted from amplitude, energy and dB energy STFT spectrograms. Finally, the amplitude, energy and dB energy densities (their mean value inside each subregion) were evaluated, totaling 12 estimators.

These estimators are motivated by the fact that musical events are characterized not only by an increase of energy or amplitude (which would be captured by their densities) but also by an increase of information complexity due to note onsets, spaces between harmonics and other observable events. This increase in complexity (possibly captured by entropy and standard deviation) also justifies a closer look at these regions: an adaptive transform should use a finer resolution to represent regions containing onsets, for example, but also regions containing entangled harmonics belonging to different harmonic series.

Although reference values were here defined as the mean values of several forms of the piano roll, considering the AMT application and the interpretation that relevant regions are regions containing events to be transcribed, these reference values do not translate directly into mean energy or amplitude of the spectrogram, which includes background noise, acoustics of the instrument, acoustics of the room, spectral leakage and other analysis artifacts. Moreover, the experiments presented in Sections 4 and 6 show that, even though there is a good correlation between the reference values obtained from the piano rolls and the corresponding estimators obtained from the STFT, these are not necessarily the same estimators that perform best when it comes to identifying musically relevant subregions.

2.2 Second experiment: detection of musically relevant regions of the TFP

In the second experiment, the best performing estimators were further evaluated as features to be used in a predictive model for the binary detection of musically relevant regions of the TFP. This experiment follows the same structure of the first one: ground truth values are extracted from a MIDI file and features are extracted from the corresponding audio recording. Single-feature predictive models and feature pairings were tested for this task.

Firstly, the binary piano roll was used as ground truth: as discussed in Section 2.1.1, a feature that predicts a piano roll with accuracy would be a good detector to be used as part of an AMT-motivated multiresolution representation. Secondly, in order to evaluate the influence of harmonics in

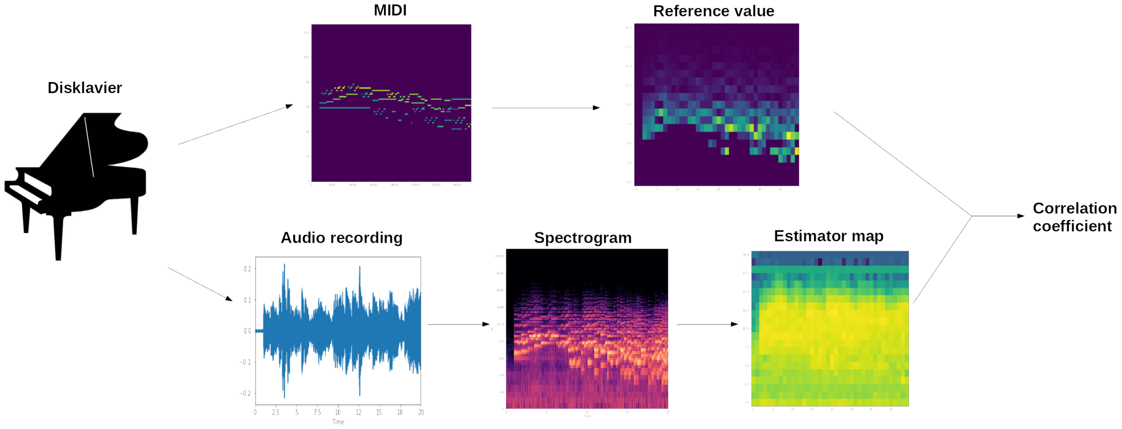

Figure 2. Overview of the experiment: from a single piano performance a MIDI-derived symbolic event density map and a spectrogram-derived estimator map are built and correlated.

this detection, a ground truth consisting of a binary representation of the piano roll with harmonics and decay models was used. For this reference value, a musical density threshold had to be chosen in order to use this map to produce binary (relevance) labels; further discussion is presented in Sec. 5. The comparison of these models with the previous ones should give us insight into the sensitivity of different estimators to harmonics, and how these harmonics aid or harm thes detection of musically relevant TFP regions.

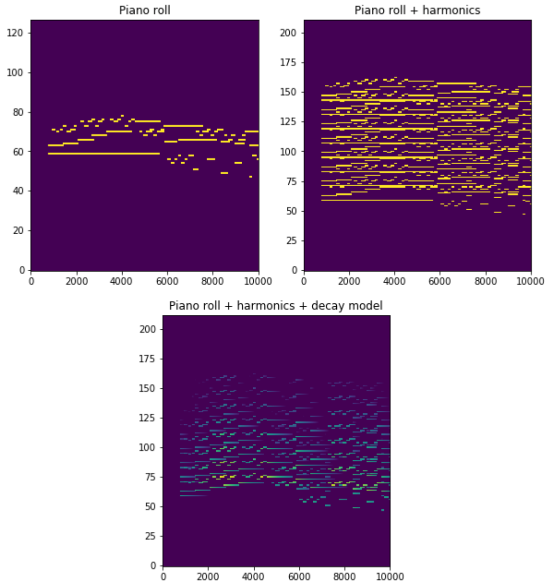

Figure 3. Alternative intermediate symbolic representations for the event density map.

3. ESTIMATOR EVALUATION EXPERIMENT

The MAESTRO dataset [11] was utilized for both experiments. It contains over 200 hours of piano performances with paired MIDI files and sound recordings. For the estimator evaluation, 16 minutes of performances were selected from the dataset, and used in the following analysis: from the corresponding MIDI files, the three reference maps described in 2.1.1 were extracted, according to a rectangular subregion size. From the corresponding audio recordings, an STFT analysis was performed, from which the estimators described in 2.1.2 were extracted. Finally, the Pearson correlation between each possible estimator/reference value pairing was computed.

The recordings in the MAESTRO dataset come from the yearly Piano-e-Competition. In order to represent different recording conditions in the 16 minutes of music selected for the experiment, each year in the dataset was sampled equally. For every recording/MIDI pair selected, 30 seconds were extracted from the halfway point of the performance and used in the experiment.

In order to test the influence of the resolution of the STFT from which the estimators are extracted, the experiment was repeated for windows of 512, 1024, 2048 and 4096 samples. All recordings are sampled at 44100 kHz, and hop sizes were chosen as one-fourth of the size of the analysis window.

As discussed in Sec. 2.1.1, the dimensions of the subregions determine our reference value, and should be con-

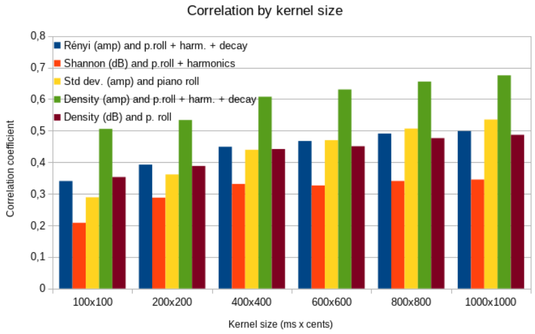

Figure 4. The effect of subregion size in the correlation coefficients for 5 estimator/reference value pairings. All estimator/reference value pairings presented (approximately) monotonically increasing behaviour.

sidered part of our definition of musical relevance. Since our work is strongly motivated by AMT, the best subregion size would be the one that produces the representation best suited for this task. Admittedly, even this motivation does not entail a single optimal subregion size: this would vary according to spectral characteristics of the audio being transcribed, such as the expected number of notes per second or the proximity in the frequency axis of simultaneous notes and harmonics per chord, so it is not appropriate nor possible to treat the subregion size as a variable to be optimized. With this discussion in mind, the experiment was repeated for subregion sizes of 100 ms by 100 cents, 200 ms by 200 cents, 400 ms by 400, 600 ms by 600 cents, 800 ms by 800 cents and 1000ms by 1000 cents, in order to observe, compare and discuss the results of the obtained representations under different conditions.

4. ESTIMATOR EVALUATION RESULTS AND DISCUSSION

Overall, the estimators that achieved the highest correlations were the density and standard deviation of the amplitude spectrogram (see table 1). Density of the dB spectrogram and Rényi entropy of the amplitude spectrogram also achieved fair results, although no estimator achieved correlation coefficients significantly above 0.5. Estimators extracted from the energy spectrogram (not in dB) did not achieve notable results, as well as Shannon entropies, that achieved the lowest correlation with each of the three reference values.

Fig. 4 shows that increasing subregion size tends to increase the correlation between reference value and estimator for all pairings. This is somewhat expected, given that using a bigger subregion size has the effect of making both matrices lose detail, favoring general trends in the data which are easier to estimate than minor local changes. No outliers were observed in this trend, which agrees with our discussion in 2.1.1 about the impossibility of finding an optimal subregion size. Further studies analyzing the impact of subregion size in a subsequent AMT task of selected pieces with similar spectral characteristics could provide more information about this behaviour.

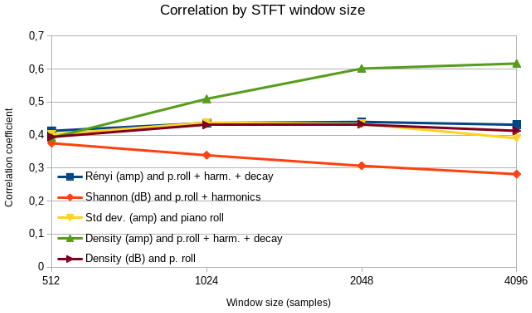

Figure 5. The effect of window size in the correlation coefficients for 5 estimator/reference value pairings.

The variation of correlation caused by STFT window size (see Fig. 5) seems to be fairly minimal for most estimators, although with some interesting exceptions and characteristics. The Shannon entropy is the only feature that presents monotonically decreasing performance with the increase of window size, while amplitude density seems to be the estimator mostly favored by an increase in window size. All other estimators are not as sensible to this variation.

Entropies in general performed better with the reference value of the piano roll with harmonics (see Fig. 6). This means that the introduction of decay models in the reference value actually served to (slightly) decorrelate these estimators with the reference. This could be related to the invariance of entropy with respect to data scaling, which pushes the reference and estimator maps in this case apart from each other as the reference decays. Standard deviation and amplitude density presented higher correlation with the reference value containing decay models, which agrees with its sensitivity to scaling and the adherence of the simple decay models to the amplitude behavior observed in the spectrograms. Surprisingly, the introduction of decay models did not improve the correlation of the density of the dB spectrogram and the reference value. This probably means that the linear decay models represent poorly the energy decay profiles exhibited in the dB spectrogram of the Disklavier recordings.

It is important to note the high variance of the obtained correlation values. Although there are noticeable and useful trends in the data, this fluctuation means that the estimator values should be further analyzed with caution in the context of relevant segment classification, the performance of which is still unclear in the presence of these fluctuations. Our second experiment aims to assess this classification performance in the context of the iterative refinement of a multiresolution time-frequency representation.

5. DETECTION OF MUSICALLY RELEVANT REGIONS OF THE TFP EXPERIMENT

For this experiment, all recordings from 2004 to 2015 present in the MAESTRO dataset were utilized, totalling nearly

Estimators	**Piano roll**	**Piano roll + harmonics**	**P. roll + harm. + decay**
Rényi (amp. spec)	0.43 +- 0.08	0.51 +- 0.09	0.49 +- 0.09
Shannon (dB spec)	0.26 +- 0.05	0.34 +- 0.06	0.29 +- 0.05
Std. dev. (amp. spec)	0.51 +- 0.10	0.55 +- 0.10	**0.64 +- 0.08**
Std. dev. (en. spec)	0.36 +- 0.10	0.39 +- 0.10	0.50 +- 0.09
Density (amp. Spec)	**0.55 +- 0.12**	**0.57 +- 0.14**	**0.66 +- 0.11**
Density (dB spec)	0.48 +- 0.10	0.52 +- 0.12	0.51 +- 0.11

Table 1. Correlation results of the best performing estimator/reference value pairings, using a subregion size of 800ms per 800 cents and an STFT with an analysis window of 2048 samples. All correlations achieved significance values $p < 0.05$

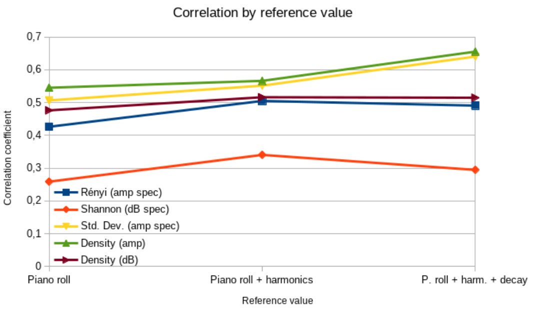

Figure 6. The effect of introducing harmonics and decay models in the reference values for 5 estimator/reference value pairings.

9 hours analyzed from 1043 recordings. Once again, for every selected recording/MIDI pair, 30 seconds were extracted from the halfway point of the performance and used in the experiment. 80% of the dataset was used for training and the remaining 20% for the evaluation.

From the estimator evaluation experiment, the best estimators from experiment 1 according to their Pearson correlation coefficient were selected, namely Rényi entropy, standard deviation and amplitude density of the amplitude spectrogram, and Shannon entropy and energy density of the dB spectrogram. An STFT window of 2048 samples was used, motivated by the results shown in Fig. 5. A subregion size of 800 ms per 800 cents was used while keeping in mind that the utilized dataset contains piano performances with heterogeneous spectral characteristics, and it would not be possible to choose an ideal subregion size for this experiment.

The first step of this experiment consisted of the training of Naive-Bayes models for the prediction of the binary piano roll reference, using the mentioned estimators as single-feature models and every possible feature pairing as two-feature models. In a second step, designed to test the sensibility of each feature to the presence of harmonics, the models were trained using a binary representation of the piano roll with harmonics and decay models as ground truth. The rationale for also including the piano roll with harmonics and decay model in this experiment is not to evaluate the model for itself, but to gain insight on whether false positives of the binary piano roll model in step 1 could be related to the presence of harmonics, and also to enquire whether there are estimators that behave differently in the identification of subregions containing harmonics.

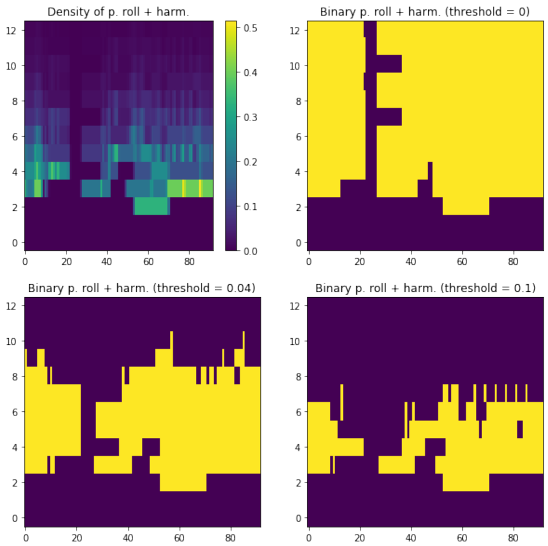

Figure 7. The effect of the threshold in transforming the piano roll with harmonics reference value into a binary one. The piano roll with harmonics density map is shown, along with 3 binary maps obtained using thresholds of 0, 0.04 and 0.1 respectively.

In order to transform the reference value explained in 2.1.1 into a binary map, a “musical density” threshold had to be chosen above which a bin was considered as a positive example of “musical activity”. If this threshold is set to 0, the presence of even the weakest harmonic would signal a positive sample, and if it is set close to 1, only the “musically densest” regions would be counted as a positive sample (see Fig. 7). Several different thresholds were tested leading to the results presented in the following section.

6. CLASSIFICATION RESULTS AND DISCUSSION

Among the single-feature models for the binary piano roll prediction, the Rényi entropy performed best according to F-Score, achieving a recall of 0.78 and precision of 0.62. The Shannon entropy achieved a notable recall of 0.92 but a precision of 0.42. Among the feature pairings, Shannon entropy and standard deviation achieved the highest F-

Estimator	F-Score	Precision	Recall	Avg. precision
Rényi (amp.)	**0.69**	0.62	0.78	0.69
Shannon (dB)	0.58	0.42	0.92	0.45
Std. dev. (amp)	0.52	**0.82**	0.38	0.72
En. density (dB)	0.65	0.61	0.69	0.67
Rényi + Shannon	0.62	0.47	**0.93**	0.67
Rényi + Std. dev.	0.63	0.76	0.53	**0.74**
Rényi + en. density	0.66	0.55	0.83	0.67
Shannon + en. density	0.63	0.49	0.91	0.66

Table 2. Selected classification results for the trained Naive-Bayes models using the binary piano roll as ground truth.

Score of 0.66 and the highest accuracy of 0.76 was achieved by the Rényi entropy and standard deviation pairing. All notable results are shown in Table 2.

In order to interpret these results and what they mean for a possible multiresolution representation, we must first discuss the different implications of high precision and high recall in this setting. If a positive subregion of the TFP is expected to be represented in higher resolution by our algorithm, false positives can be interpreted as spending computing power in unimportant regions, while false negatives are interpreted as withholding computing power in musically relevant regions that should be refined. Ideally, in a TFP representation aimed at AMT, we would like to represent in detail all regions containing musical information, even if this means spending a bit of computing power where this is not needed. Thus, when choosing between features with similar F-Scores, we favor the ones with higher recall over the ones with higher precision. Taking this into account, our results indicate that both Rényi and Shannon entropies are good musical information estimators.

Several thresholds were tested for the introduction of harmonics in the ground truth label. A threshold of 0 leads to a percentage of 78% positive training samples in relation to all samples - in practice, nearly every region of the TFP above 200 Hz is labeled as positive for all training samples. When using the simple piano roll with no harmonics as ground truth, 30% of the training samples are labeled as positive. As a middle ground, a threshold of 0.04 was chosen (49% of positive samples). For this threshold, the performance of the Rényi and Shannon entropies improve (see Table 3), while standard deviation and amplitude density suffer a small drop in F-Score. The improvement in precision of the Shannon entropy model is of special importance: it means that some of the false positives detected by the model trained with the simple piano roll were actually regions occupied by harmonics, and thus regions that actually contained some musical information (which would be relevant e.g. to timbre analysis).

Taking all classification results and our discussion of precision and recall into consideration, both Rényi and Shannon entropies present encouraging results for their usage in an algorithm for producing a mutiresolution time-frequency representation for tasks such as AMT and timbre analysis. The pairing of both entropies, as well as the pairing of Shannon entropy and energy density could also be useful for this detection. These results also further validate the usage of Rényi entropy as a time-frequency information content estimator as seen in [10].

Estimator	Recall (w. harmonics)	Precision (w. harmonics)
Rényi (amp.)	0.82 (+0.04)	0.61 (-0.01)
Shannon (dB)	0.92	**0.60 (+0.18)**
Std. dev. (amp)	0.31 (-0.07)	0.84 (+0.02)
Amp. density (amp)	0.26 (-0.06)	0.81 (+0.03)
En. density (dB)	0.69	0.61

Table 3. Classification results using the binary piano roll with harmonics (threshold set to 0.04) as ground truth. In parentheses the score change in relation to the piano roll model (without harmonics) using the same feature is shown.

7. CONCLUSION

In this paper, two experiments investigating musical information estimators in the TFP were conducted. We evaluated the adherence of complexity estimators (entropy and standard deviation) and intensity estimators (density of amplitude and energy) to different reference values relating to tasks such as AMT and timbre analysis. The obtained results show important distinctions in the behavior of estimators in the presence or absence of decay models in the reference values.

A binary classification experiment was conducted in order to investigate the applicability of these estimators in producing an adaptive multiresolution time-frequency representation. Selected estimators were evaluated in the labeling of subregions as musically relevant or not, using two different ground truth references taken from a piano roll, with encouraging results. Specifically, Rényi and Shannon entropies achieved very high recall and good F-Scores, signaling their applicability as musical information estimators and further validating their use in the context of adaptive multiresolution time-frequency representations. Their sensitivity to the presence of harmonics was also evaluated, as well as possible feature pairings. Overall, the results show that it is possible to detect musically relevant regions of a TFP representation with satisfactory results using Rényi and Shannon entropies.

Future work includes the combination of these estimators with a subband processing algorithm for computing high-resolution STFT representations of musically interesting subregions. With a reliable detection mechanism, a computationally efficient adaptive multiresolution time-frequency representation can be obtained by iterating detection and subband processing and computing high resolution STFT representations within subregions actually containing relevant musical information.

Since the dataset used is comprised solely of piano performances, it is also important to evaluate how well these detection methods perform in other conditions, such as performances of instruments with very different spectral characteristics from the piano and more complex multi-instrument performances.

There is also interest in considering separately the frequency and time axes in our detection of relevant subregions in future work. Since the inherent tradeoff between time and frequency resolution in the STFT motivates our development of a multiresolution transform, a detection algorithm that distinguishes between regions that contain relevant time or frequency information could better guide the refinement step towards a better representation for AMT, representing, for instance, onset regions with higher temporal resolution and melodic lines with higher frequency resolution.

Acknowledgments

Tthe first author acknowledges having received financial support from CAPES, and the second author from CNPq grant 307389/2019-7.

8. REFERENCES

[1] A. Lukin and J. Todd, "Adaptive time-frequency resolution for analysis and processing of audio," in *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.

[2] M. Liuni, A. Robel, E. Matusiak, M. Romito, and X. Rodet, "Automatic adaptation of the time-frequency resolution for sound analysis and re-synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 959–970, 2013.

[3] F. Jaillet and B. Torrésani, "Time-frequency jigsaw puzzle: Adaptive multiwindow and multilayered gabor expansions," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 5, no. 02, pp. 293–315, 2007.

[4] A. Klapuri and M. Davy, *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

[5] T. Cheng, S. Dixon, and M. Mauch, "Modelling the decay of piano sounds," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 594–598.

[6] E. D. Blackham, "The physics of the piano," *Scientific american*, vol. 213, no. 6, pp. 88–99, 1965.

[7] R. Bresin, A. Friberg, J. Sundberg *et al.*, "Director musices: The kth performance rules system," *IPSJ Report Music Information Science (MUS)*, vol. 2002, no. 63 (2002-MUS-046), pp. 43–48, 2002.

[8] J. Brynolfsson, I. Reinhold, J. Starkhammar, and M. Sandsten, "The matched reassignment applied to echolocation data," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8236–8240.

[9] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 299–309, 1977.

[10] R. G. Baraniuk, P. Flandrin, A. J. Janssen, and O. J. Michel, "Measuring time-frequency information content using the rényi entropies," *IEEE Transactions on Information theory*, vol. 47, no. 4, pp. 1391–1409, 2001.

[11] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7