Estimation of Music Recording Quality to Predict Automatic Music Transcription Performance

Markus Schwabe[1(✉)], Thorsten Hoffmann[1], Sebastian Murgul[2], and Michael Heizmann[1]

[1] Karlsruhe Institute of Technology, Institute of Industrial Information Technology,
Hertzstraße 16, 76187 Karlsruhe, Germany
{markus.schwabe,michael.heizmann}@kit.edu

[2] Klangio GmbH, Alter Schlachthof 39, Karlsruhe, Germany
sebastian.murgul@klangio.com

Abstract. Music signals can nowadays be recorded and further processed by lots of different devices in order to extract additional information like instruments and genre or use parts of those signals in various applications. Thereby, music recording quality has a big impact on all kinds of Music Information Retrieval (MIR) signal processing and their results. In this work, the recording quality of piano music is estimated by three separate neural network approaches for background noise, sound disturbances, and reverberation. The approaches for background noise and sound disturbances estimate the resulting Signal to Noise Ratio (SNR) of the music piece, the first for constant SNR and the latter for the time-dependent case. Reverberation is estimated by means of the two room parameters reverberation time and early decay time. Exemplarily, the SNR estimation results are validated in the field of piano music transcription, where the impact of the estimated recording quality on the automatic transcription results is analysed. According to those results, the piano music transcription performance can be predicted by means of the recording quality parameters.

Keywords: Recording quality · Piano music · Music transcription · Neural networks

1 Introduction

Automatic music transcription (AMT), which is one part of the Music Information Retrieval (MIR) task, tries to create a human readable sheet of music from an input audio signal. Commercial products like 'Piano2Notes'[1] tend to output transcription results of varying quality, depending on the musical complexity and the quality of the recording. These products are available as mobile device applications and are used in different scenarios by both professional and amateur users. In most cases, the user can influence the recording quality for example by

[1] https://piano2notes.com.

the distance to microphone, the reduction of environmental sounds, or the choice of the recording room. Therefore, it is useful to estimate the recording quality for direct user feedback in order to give hints for possible improvements. Moreover, further processing algorithms like noise suppression could be used in advance of the MIR task in case of a low estimated recording quality to improve the recorded signal or reduce problematic interferences. Since the recording quality generally affects AMT results as well as other MIR tasks like music source separation or beat tracking, the approach based on the estimated recording quality for the AMT task in this work can be transferred to other MIR tasks as well.

Degraded music signal quality and its impact on MIR task performance has been investigated by Mauch and Ewert [14] by a toolbox with 14 controlled degradation units. Their experiments showed that no general relationship between music degradation and all MIR task performances can be found, but that performance strongly depends on the methods and degradations used. They analysed audio ID, score-to-audio alignment, beat-tracking, and chord detection as MIR tasks and suggested the development of more robust algorithms by means of their audio degradation toolbox [14]. Especially for data-driven approaches, robustness is achieved by the incorporation of diverse training examples, which was highlighted by Serizel et al. [20] for the case of sound event detection with noise and signal degradation. Additionally, robustness against adversarial attacks can be improved by simple methods like compression or addition of white noise [21]. Beside degradation, audio compression is a second impact on MIR results that was investigated by Hamawaki et al. [8] for content-based MIR and by Uemura et al. [23] for chord recognition. While chord recognition is not strongly affected by compressed input signals, the effects of different bit rates could be reduced by normalizing the MFCC feature in case of content-based MIR results.

Quality evaluation of audio signals is often achieved by human perception and judgement in literature, e.g. for compressed music [4] as well as for telephone speech signals [16]. Even if they aim to develop an objective framework for the quality evaluation, the human perception is not important for further signal processing algorithms. Therefore, objective criteria like Signal to Noise Ratio (SNR) suit better for this aim. In case of music signals, no SNR estimation approach is known by the authors, but for speech signals, the NIST SNR measurement [2] and the WADA-SNR algorithm [11] are used to estimate the SNR by exploiting the statistical characteristics of speech like the amplitude density and gamma distributions. As there are significant characteristical differences between speech and music, proven approaches for speech SNR estimation unfortunately lead to big errors in music SNR estimation, even for white noise.

Besides SNR estimation, Kendrick et al. [10] tried to rate the room influence by means of important room acoustic parameters that are calculated under the premise of a known speech or music signal. For unknown signals, blind estimation algorithms of the reverberation time have been presented only for the speech case. Eaton et al. [6] achieved a noise-robust estimation and Diether et al. [5] developed a real-time algorithm suitable to mobile applications. According to the different characteristics of speech and music, those algorithms are not suited for reverberation time estimation in music signals.

In this work, the recording quality of music signals and its impact on an MIR task is estimated by means of relevant objective quality parameters. Consequently, subjective human perceptions are not included in that quality definition. As the estimation should identify possible opportunities for improving recording quality, several quality parameters are estimated for the relevant signal degradation sources. Empirically, three main sources for a reduced AMT task performance caused by the recording quality have been identified: room reverberation (incl. echos), noise, and short interferers. These sources lead for example to inaccuracies in active notes' time estimation and increase the chance of false positives in case of AMT. Other audio degradations and audio compression only have very small impact, so they are neglected in this work.

We present three neural network approaches to estimate the influence of the identified degradation sources noise, short interferers, and reverberation in order to rate the recording quality of unknown piano music. Finally, we exemplarily analyse the impact of the recording quality on AMT algorithms using an implementation of 'Onsets and Frames' [9] in Sect. 6.

2 Music Data Processing

The pure recording process of music data can be described using three basic components: sound source x_S, sound transmission path $g(\cdot)$, and recorded sound y. It is assumed that the recording environment does not change. Therefore, the transmission can be modeled by a room impulse response (RIR) [13]. Mathematically, the discrete recording process can be described by

$$y[n] = x_\mathrm{S}[n] * g[n] + r_\mathrm{disturb}[n] + r_\mathrm{noise}[n] \tag{1}$$

with the convolution operator $*$, the RIR $g[n]$ for the music source transmission path, the background noise $r_\mathrm{noise}[n]$, and $r_\mathrm{disturb}[n]$ for all disturbing short interferers which are transmitted to the recorder. As the transmission of noise or disturbing sound is not of interest, only their overlapping signal portions at the recorder are considered. Consequently, $r_\mathrm{noise}[n]$ and $r_\mathrm{disturb}[n]$ include the effects of RIR between background or disturbing sound sources and the recorder.

For preprocessing, the constant-Q transform (CQT) [18] is widely used in music signal tasks, because it defines a time-frequency representation with logarithmic frequency scale of the discrete signal $x[n]$. It is calculated by

$$X_\mathrm{CQT}(m,k) = \sum_{n=m-\lfloor N_k/2 \rfloor}^{m+\lfloor N_k/2 \rfloor} x[n]\; a_k^*\left[n - m + \frac{N_k}{2}\right] \tag{2}$$

with time index m, frequency index k, frequency-dependent normalization factor N_k and the floor operator $\lfloor\cdot\rfloor$. The basis function $a_k[n]$ is defined by

$$a_k[n] = \frac{1}{N_k} w\left[\frac{n}{N_k}\right] e^{-j2\pi \frac{f_k}{f_\mathrm{A}}} \tag{3}$$

with sampling rate f_A and the time-dependent window $w[n]$ at time step n.

3 Quality Metrics

In order to evaluate the quality of a music signal, several metrics can be used. The most common metric is the SNR which describes the ratio of the signal power P_{signal} to the sum of all noise or disturbance powers P_{noise} in a logarithmic scale:

$$\text{SNR}_{\text{dB}} = 10 \cdot \log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right) . \tag{4}$$

Another metric for acoustic signals is the reverberation time t_{RT} of the recording room. It is calculated by means of the backwards integration

$$s_{\text{back}}[n] = \sum_{i=n}^{N_{\text{RIR}}} g^2[i] \ , \qquad 0 \leq n < N_{\text{RIR}} \tag{5}$$

of the squared discrete RIR $g[n]$ [19]. Instead of infinity in the continuous case, the upper bound N_{RIR} of the sum represents the number of samples of the discrete RIR describing the sound transmission as in (1). Similar to the SNR calculation in (4), a logarithmic ratio

$$s_{\text{dB}}[n] = 10 \cdot \log_{10}\left(\frac{s_{\text{back}}[n]}{s_{\text{back}}[0]}\right) \tag{6}$$

is calculated which describes the steady decay rate of the signal. Then, t_{RT} is defined as the time span for the decay from $s_{\text{dB}}[n] = -5\,\text{dB}$ to $s_{\text{dB}}[n] = -25\,\text{dB}$. An alternative for the reverberation time is the early decay time t_{EDT} which describes the time span for the decay from $s_{\text{dB}}[n] = 0\,\text{dB}$ to $s_{\text{dB}}[n] = -10\,\text{dB}$. This can be useful for a detailed analysis of the early behaviour of the signal. Both times are extrapolated to a decay of 60 dB for comparison, like in [10].

The mean absolute error (MAE) is used as the main evaluation metric. For two discrete signals $z_1[n]$ and $z_2[n]$ of length N, it is defined by

$$\text{MAE} = \frac{1}{N} \cdot \sum_{i=1}^{N} |(z_1[i] - z_2[i])| \, . \tag{7}$$

4 Datasets

Several datasets have been used due to the different sound sources for a reduced recording quality. They can be split into the three parts piano music, noise sounds, and RIR.

Piano music is the basis for all quality analysis of this work. It is taken from the MAPS dataset [7], from which the first 30 s of all 270 music pieces (no solo notes) are extracted to get an equal length for all recordings. Those 270 music pieces consist of 210 synthesized piano songs that are used for training and 60 real recorded piano songs which are used for testing.

The noise dataset consists of generated white noise, once and double low-pass filtered white noise (often called pink and brown noise) and an additional recording of high frequency radio noise from [3]. These sounds have a fairly steady characteristic. Additionally, canteen and factory noise from [3] and several sound classes of the 'UrbanSound Dataset' [17] are used as disturbance noises with higher variances and more distinct separate events.

In order to simulate different recording conditions, a dataset with recorded RIRs of nine different rooms [22] is used. Within this dataset, two rooms (R112 and CR2) will be used exclusively for testing while the other seven rooms are used for training. Time intervals of $t_{\mathrm{RT}} \in [0.4\,\mathrm{s}, 2.2\,\mathrm{s}]$ and $t_{\mathrm{EDT}} \in [0.2\,\mathrm{s}, 3.0\,\mathrm{s}]$ for the training rooms and of $t_{\mathrm{RT}} \in [0.4\,\mathrm{s}, 2.0\,\mathrm{s}]$ and $t_{\mathrm{EDT}} \in [0.3\,\mathrm{s}, 1.5\,\mathrm{s}]$ for the test rooms have been calculated as ground truth room parameters.

5 Recording Quality Estimation

As room reverberation and background noise influence the whole music recording by different effects and short interferers are only present during a defined time interval, the estimation of the quality metrics is split up into three separate regression algorithms based on neural networks. Its schematic overview with the respective outputs is illustrated in Fig. 1. The neural network architectures were determined experimentally with focus on small but powerful networks. Therefore they are composed of several fully connected (FC) layers and some additional convolutional layers at the beginning if a dimension reduction is necessary.

All algorithms use the CQT of the music signal with 84 frequency bands, a minimum frequency of 32.70 Hz and a hop size of 512 as input for their preprocessing. The sampling frequency 22 050 Hz is common in audio processing. Furthermore, all networks are trained using Adam optimizer [12] and mean squared error loss with a batch size of 1024. The training is executed for 50 epochs. ReLU is used as activation function in the hidden layers and all fully connected layers are followed by a 40 % dropout to minimize overfitting. Each output layer is a single neuron with linear activation.

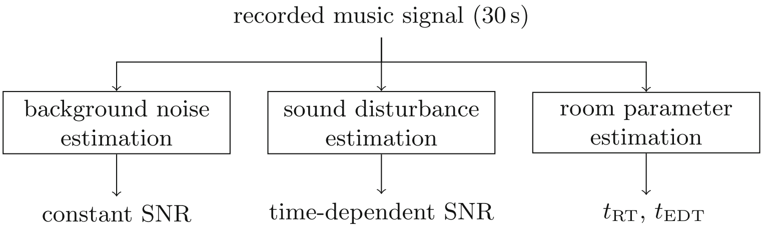
recorded music signal (30 s)
background noise estimation
sound disturbance estimation
room parameter estimation
constant SNR
time-dependent SNR
t_{RT}, t_{EDT}

Fig. 1. Schematic overview of the recording quality estimation.

5.1 Background Noise Estimation

The first network estimates the SNR of a 30 s input song superposed by different background noise types. Therefore, white, pink, brown, or high frequency noise is scaled and overlapped with the original music to reach an SNR level in the interval $[-5\,\text{dB}, 20\,\text{dB}]$ with steps of 2.5 dB.

For the training dataset, every combination of MAPS piano song, noise type, and SNR level is created. In order to increase the amount of training samples, every recording is also resynthesized from its respective MIDI file with sound profiles of nine different instruments from the GM 1 sound set [1], followed by a similar data generation with all noise types and SNR levels. Acoustic grand piano (PC_1), church organ (PC_{20}), acoustic guitar (PC_{25}), acoustic bass (PC_{33}), viola (PC_{42}), trumpet (PC_{57}), tenor sax (PC_{67}), flute (PC_{74}), and banjo (PC_{106}) are chosen as synthesized instruments, for which the indices represent their respective MIDI program change (PC) numbers. For the test dataset, the 60 real recorded MAPS songs and the 9×60 resynthesized variants of them are considered in two separate cases with the same data generation as described above. In total, this yields 83 150 samples for training, 2640 samples for testing with real recordings, and 23 760 samples for testing with resynthesized songs.

On the basis of the CQT of each 30 s dataset sample, mean and variance are calculated for each of the 84 frequency bands during preprocessing. Thus, the input of the neural network is reduced to only 168 values which enables a very fast inference. The network consists of three hidden FC layers with 128, 64, and 32 neurons respectively which yields a network with 32 001 parameters. Its architecture is illustrated in Fig. 2.

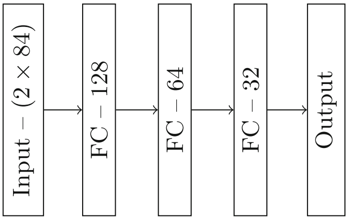

Fig. 2. Network structure for background noise estimation.

Table 1 shows the MAE results of the SNR estimation for the real piano recordings and the resynthesized test dataset for different noise types. The best results are obtained for brown noise and the worst ones for high frequency noise while all errors are very close within one test dataset. Between the two datasets, there is a distinct difference for all considered cases. For the real recorded songs, MAE values of 0.96 dB can be achieved, but for their resynthesized variants, the MAE almost doubles to 1.69 dB. One reason for this difference is the difficult noise estimation in case of specific instruments like church organ (PC_{20}) or acoustic bass (PC_{33}). In Table 2, the results of all synthesized instruments are compared to the real piano test recordings by means of the MAE and the mean standard deviation (STD). Since the STD lies within a single SNR step of 2.5 dB in most cases, the SNR estimation for background noise performs reliably. The best results can be achieved for the real recorded piano. Consequently, resynthesizing will not be considered for the following networks as it did not show better results and might also not represent a realistic scenario, because piano music synthesized by various instruments was used.

Table 1. MAE (in dB) of the background noise estimation for different noise types and test datasets (recorded and resynthesized).

	White	**Pink**	**Brown**	**High freq.**	**Average**
Real piano recordings	0.95	0.95	0.85	1.09	0.96
Resynthesized songs	1.80	1.67	1.54	1.77	1.69

Table 2. MAE and mean STD (in dB) of the background noise estimation for different instrument types of [1].

	Real	**$\mathbf{PC_1}$**	**$\mathbf{PC_{20}}$**	**$\mathbf{PC_{25}}$**	**$\mathbf{PC_{33}}$**	**$\mathbf{PC_{42}}$**	**$\mathbf{PC_{57}}$**	**$\mathbf{PC_{67}}$**	**$\mathbf{PC_{74}}$**	**$\mathbf{PC_{106}}$**
MAE	0.96	1.19	3.06	1.15	2.14	1.41	1.65	1.75	1.12	1.78
STD	0.99	1.17	2.64	1.16	2.59	1.34	1.72	1.80	1.23	1.57

5.2 Sound Disturbance Estimation

The second network estimates the presence and the SNR values of overlapped impulsive noise sounds. As it is assumed that the disturbances are time-variant, short parts of 2 s length are analysed. For the dataset construction ten parts of each MAPS piano song are extracted and combined with a randomly chosen disturbance sound and SNR level in the range [−5 dB, 20 dB] with steps of 2.5 dB. This leads to a total of 4.6×10^5 training and 13 200 test samples.

In order to consider time-dependency, the input of the neural network is the CQT of each 2 s dataset sample with 87 time bins. Figure 3 shows the network structure in which 'FC-Stack' consists of three fully connected layers with 256, 64, and 32 neurons respectively. The network has 71 473 parameters.

All results for the SNR estimation with different time-dependent noise types are listed in Table 3. The worst MAE value of 2.8 dB is detected in case of the air conditioner sound, the best MAE result of 1.75 dB is achieved for factory noise

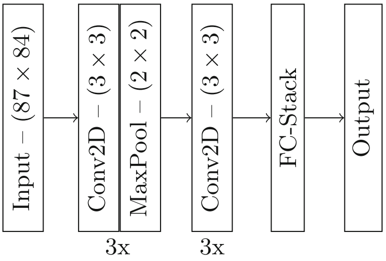

Fig. 3. Network structure for sound disturbance estimation.

disturbance. With an average MAE of 2.3 dB, most estimation errors remain within one SNR step of 2.5 dB. As the average STD of about 2.46 dB is also lower than one step, the estimation performs reliable.

Table 3. MAE and mean STD (in dB) of the sound disturbance estimation for air conditioner (a), car horns (b), playing children (c), dog bark (d), canteen (e), and factory (f) noise types.

	(a)	**(b)**	**(c)**	**(d)**	**(e)**	**(f)**
MAE	2.80	2.75	2.40	2.23	1.87	1.75
STD	3.05	2.51	2.68	2.48	2.04	1.97

In Fig. 4, the time-variant estimation is illustrated by the time-dependent average SNR values of several 30 s recordings with three dog barking disturbances and an overlap of 1 s between consecutive 2 s samples. The estimation shows a distinct break-in in fully disturbed parts and a slightly increase in partly disturbed samples (50 % is disturbed). Undisturbed samples generally show higher SNR values, which is expected, but the maximum SNR value of 20 dB cannot be reached in most cases. This effect could be explained by the real test recordings which maybe have included some additional noise caused by the recording conditions. Furthermore, higher SNR levels than 15 dB can hardly be discriminated which is illustrated exemplarily for dog barking in Fig. 5. Most SNR estimates are within one SNR step of 2.5 dB, but at higher SNR values this variance is enlarged. In those cases of high SNR, the overlapped sounds are too low for the neural network to detect the exact SNR level, so it estimates a value below the trained maximum of 20 dB. But although high SNR values are slightly underestimated, time periods of reduced SNR can be detected properly.

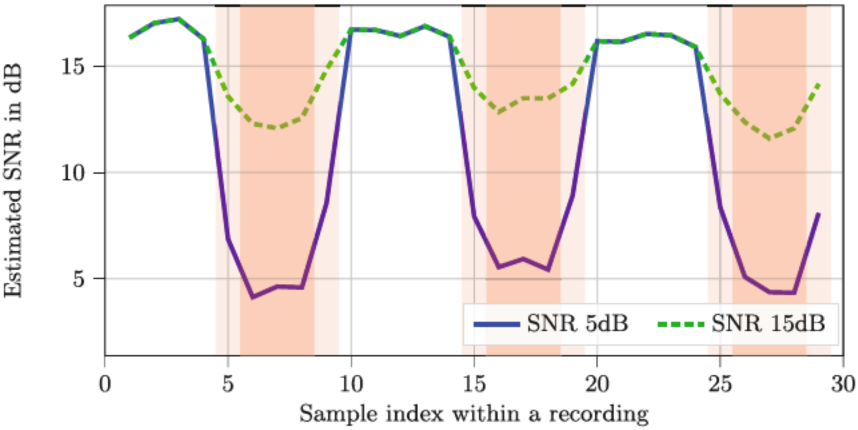

Fig. 4. Average estimated SNR values for fully (red), partly (50 % disturbed, light red), and undisturbed samples for overlapped dog barking at two SNR levels. The SNR levels specify those during the red area.

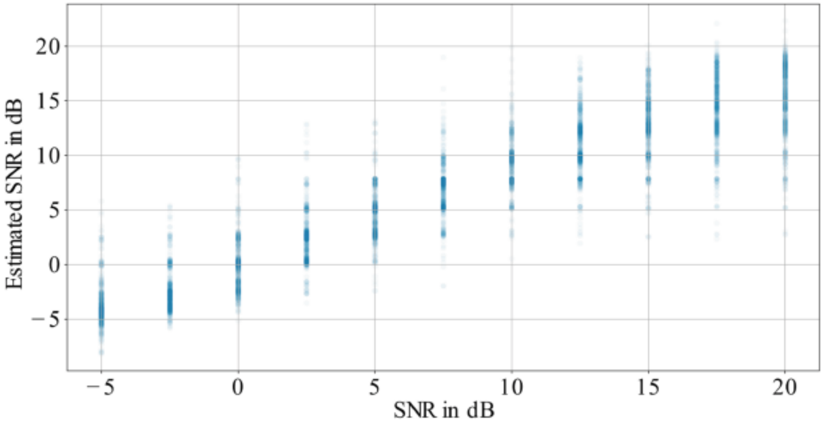

Fig. 5. Estimated vs. real SNR values for overlapped dog barking.

5.3 Room Parameter Estimation

Both room parameters reverberation time t_{RT} and early decay time t_{EDT} are estimated by the third neural network. To simulate recordings in different conditions, the MAPS piano songs are convolved with various RIRs of the RIR dataset. It is assumed that the room conditions may only slightly vary within a 30 s recording. Therefore, only a single estimation for each music piece is sufficient. As the data generation process is executed 500 times for each training song, about 9×10^4 unique samples are generated. For evaluation, each test song is convolved with randomly chosen RIRs out of the training and the test rooms to generate 1500 samples for each of the training and the test room dataset.

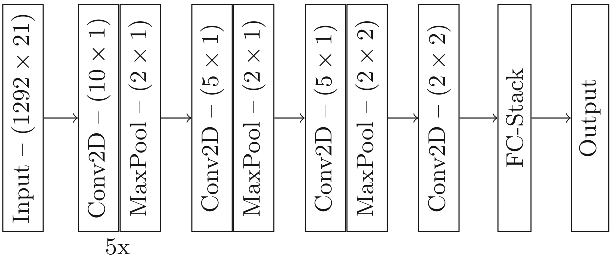

Fig. 6. Network structure for room parameter estimation.

Since the signal decay over time is influenced by the RIR, each 30 s dataset sample is preprocessed similarly to an onset detection [15] by a time differentiation. First, the CQT is transformed to logarithmic amplitude scale and then the time differentiation is performed which results in a decay value per time step. Moreover, the 84 frequency bands are reduced to 21 by summarizing blocks of 4 frequency bands respectively to get a more compact representation. The network has 35 745 parameters. Its structure is illustrated in Fig. 6 in which the same FC-Stack as in Sect. 5.2 is used.

Table 4 shows the results $t_{\mathrm{RT},60}$ and $t_{\mathrm{EDT},60}$ which represent the extrapolated estimation of t_{RT} and t_{EDT} to a 60 dB decay. The test room errors are smaller than the training room errors for both parameters because the ranges of their ground truth values are smaller. Furthermore, the estimation result errors are generally lower for $t_{\mathrm{EDT},60}$ because t_{EDT} is assumed to be both smaller in absolute numbers and easier to estimate. Figure 7 illustrates the distribution of all estimated $t_{\mathrm{RT},60}$ in relation to the real $t_{\mathrm{RT},60}$ values. The network generally underestimates higher values while lower values can be estimated decently. One reason could be the unbalanced training dataset which incorporates more rooms with moderate reverberation and therefore smaller time values.

Table 4. MAE of the estimated $t_{\mathrm{RT},60}$ and $t_{\mathrm{EDT},60}$ (in s) for training and test rooms extrapolated to a decay of 60 dB.

	Training dataset	**Test dataset**
$t_{\mathrm{RT},60}$	0.316	0.288
$t_{\mathrm{EDT},60}$	0.224	0.201

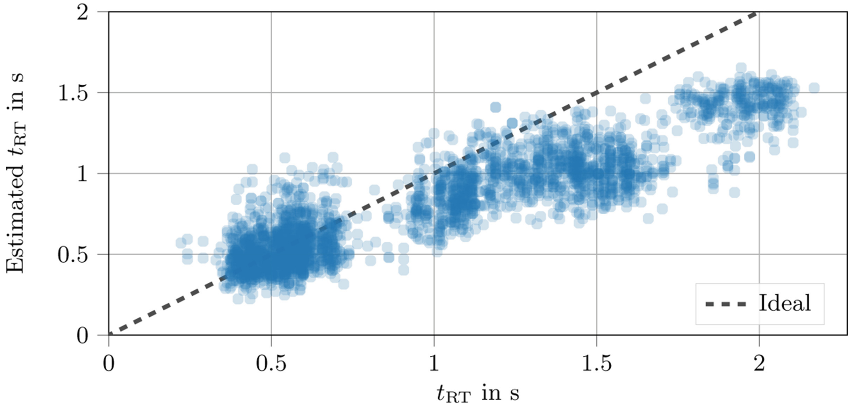

Fig. 7. Estimated vs. real tRT;60 for RIRs of all nine rooms of [22].

6 Experimental Results for AMT

In order to validate the quality estimation in a realistic MIR application, the relation between estimated SNR values and piano AMT results with the algorithm 'Onsets and Frames' [9] is investigated. As in the previous sections, real piano songs of the MAPS test dataset were superposed and convolved by different levels of noise or RIRs. Other MIR applications could benefit from the quality estimation as well, but are not considered in this work. The AMT result is given by the relative F1-Score

$$\mathrm{F1}_{\mathrm{rel}} = \frac{\mathrm{F1}_{\mathrm{d}}}{\mathrm{F1}_{\mathrm{p}}} = \frac{\mathrm{TP}_{\mathrm{d}} \cdot (\mathrm{TP}_{\mathrm{p}} + 0.5\,(\mathrm{FP}_{\mathrm{p}} + \mathrm{FN}_{\mathrm{p}}))}{(\mathrm{TP}_{\mathrm{d}} + 0.5\,(\mathrm{FP}_{\mathrm{d}} + \mathrm{FN}_{\mathrm{d}})) \cdot \mathrm{TP}_{\mathrm{p}}} \tag{8}$$

which is the resulting F1-Score for a disturbed recording $\mathrm{F1}_{\mathrm{d}}$ in relation to its undisturbed 'pure' version $\mathrm{F1}_{\mathrm{p}}$. Both F1-Scores $\mathrm{F1}_{\mathrm{d}}$ and $\mathrm{F1}_{\mathrm{p}}$ are calculated by means of their respective correctly detected notes (true positives TP), falsely detected notes (false positives FP), and missed notes (false negatives FN). Consequently, a relative F1-Score of 100 % means that the analysed recording achieves the same transcription quality as the undisturbed recording.

The resulting relative F1-Scores for the background noise types are illustrated in Fig. 8 in relation to the estimated SNR values. As the SNR estimation has achieved appropriate results in Table 1 and 2, only the results for the estimated SNR values are given. Furthermore, the mean relative F1-Scores in Fig. 8 and those of the true SNR values showed similar characteristics in early experiments. AMT results are only marginally decreased in case of intense brown noise, whereas white, pink, and especially high frequency noise have a high impact on the investigated AMT performance. Due to the data distribution, outliers with an atypical relative SNR are possible. Those outliers can be explained by the different music pieces and their level of difficulty for AMT.

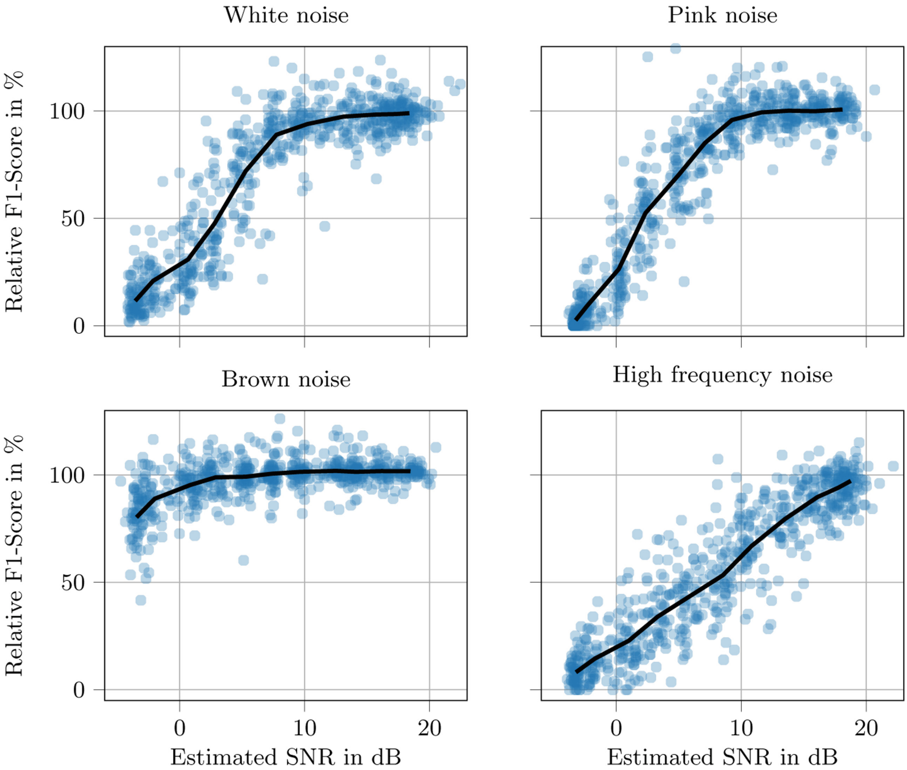

Fig. 8. Relative F1-Scores for piano AMT over estimated SNR values for dierent background noise types. The solid line represents the corresponding relative F1-Scores over the mean of all estimated SNR values for each ground truth SNR step.

In case of time-dependent sound disturbances, the relative F1-Scores are illustrated in Fig. 9 for various sound classes. The data distributions are comparable to those of high frequency noise in Fig. 8, so only the mean values are given here. During each music piece of 30 s, three disturbances of 4 s have been analysed, which results in 40 % disturbance per recording. All investigated sound disturbance classes have a comparable and nearly proportional effect on the AMT results. Consequently, the SNR estimations of background noise and sound disturbances can be used for predictions on AMT result declines and therefore AMT performance reduction due to lower recording quality.

The AMT results with room parameter estimation are illustrated in Fig. 10 for the early decay time because it has got slightly better results than for $t_{\mathrm{RT},60}$ in Sect. 5.3. Although a clear correlation between relative F1-Score and $t_{\mathrm{EDT},60}$ can be stated, the data variance is very high and no reliable AMT performance prediction is possible. One reason for that is the piano characteristic that includes controlled reverberation in its sound production, so the influence of low reverberation for piano AMT performance is insignificant. Therefore, only a classification

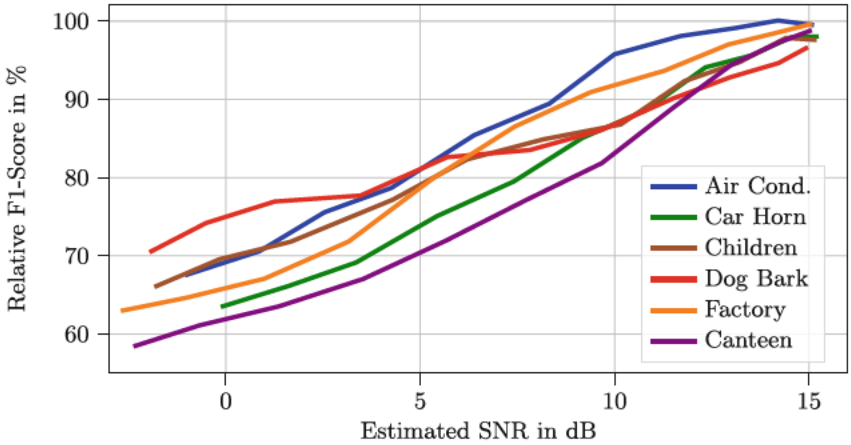

Fig. 9. Relative F1-Score for piano AMT over mean estimated SNR in case of disturbed samples (40% of the recording is disturbed).

of rooms with high or low $t_{\mathrm{EDT},60}$ is investigated here. The classification threshold is defined as 1.2 s according to the results of Fig. 10.

In Table 5, piano music transcription results classified by the predicted $t_{\mathrm{EDT},60}$ are presented for different SNR levels.

Table 5. Relative F1-Scores for piano AMT dependent on $t_{\mathrm{EDT},60}$ estimation and different SNR levels. Sound disturbance is present for 40 % of the recording, background noise and reverberation for the whole duration.

	SNR w.r.t. background noise	**SNR w.r.t. sound disturbance**		
		15 dB	7.5 dB	0 dB
$t_{\mathrm{EDT},60} < 1.2$ s	15 dB	65 %	58 %	49 %
	7.5 dB	46 %	43 %	40 %
	0 dB	9 %	9 %	12 %
$t_{\mathrm{EDT},60} \geq 1.2$ s	15 dB	34 %	31 %	25 %
	7.5 dB	19 %	18 %	15 %
	0 dB	4 %	4 %	4 %

The correlation of the estimated SNR values on piano AMT results is confirmed by those results. As different sound degradations have been used ensemble in this analysis, the relative SNR values are smaller than with only one degradation type. Additionally, a relation of the piano AMT performance and a high estimated early decay time can be stated because of the lower relative F1-Score for $t_{\mathrm{EDT},60}$ values above 1.2 s. Consequently, the performance reduction of piano

AMT due to lower recording quality and possible reasons for it can be predicted by the estimated quality parameters for background noise, short sound disturbances, and reverberation.

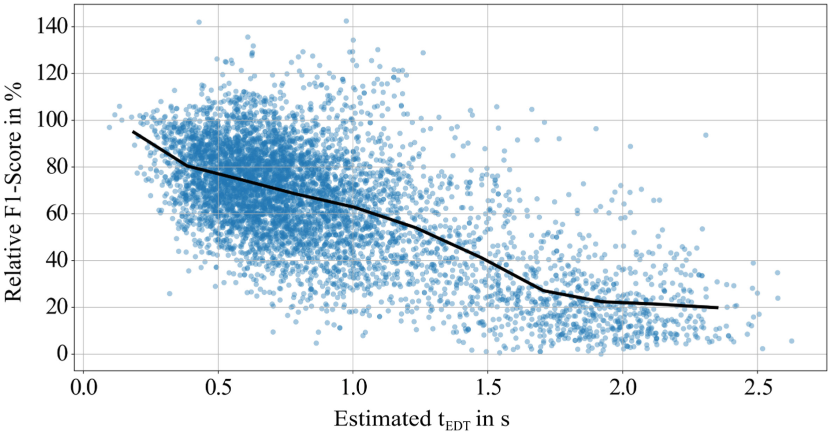

Fig. 10. Relative F1-Score for piano AMT over estimated tEDT;60. The solid line represents the corresponding sliding average.

7 Summary

Three neural network approaches for the estimation of piano music recording quality have been proposed. Each network concentrates on one of the recording quality degradation sources background noise, sound disturbances, or reverberation and estimates the respective SNR or room parameters. The results have been validated successfully in a realistic scenario of piano music transcription for which the quality estimation can be used to predict the performance reduction due to a lower recording quality.

In future works, quality estimation should be enlarged to other music genres than piano music. Furthermore, the presented quality estimation can be validated for other MIR tasks like music source separation or beat tracking.

References

1. GM 1 sound set. https://www.midi.org/specifications-old/item/gm-level-1-sound-set. Accessed 02 Sep 2021
2. NIST speech signal to noise ratio measurements. https://www.nist.gov/itl/iad/mig/nist-speech-signal-noise-ratio-measurements. Accessed 02 Sep 2021
3. Signal Processing Information Base (SPIB). https://spib.linse.ufsc.br/noise.html. Accessed 02 Sep 2021

4. Croghan, N.B.H., Arehart, K.H., Kates, J.M.: Quality and loudness judgments for music subjected to compression limiting. J. Acoust. Soc. America **132**(2), 1177–1188 (2012). https://doi.org/10.1121/1.4730881
5. Diether, S., Bruderer, L., Streich, A., Loeliger, H.A.: Efficient blind estimation of subband reverberation time from speech in non-diffuse environments. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 743–747. IEEE (2015). https://doi.org/10.1109/ICASSP.2015.7178068
6. Eaton, J., Gaubitch, N.D., Naylor, P.A.: Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 161–165. IEEE (2013). https://doi.org/10.1109/ICASSP.2013.6637629
7. Emiya, V., Badeau, R., David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Trans. Audio Speech Lang. Process. **18**(6), 1643–1654 (2009). https://doi.org/10.1109/TASL.2009.2038819
8. Hamawaki, S., Funasawa, S., Katto, J., Ishizaki, H., Hoashi, K., Takishima, Y.: Feature analysis and normalization approach for robust content-based music retrieval to encoded audio with different bit rates. In: Huet, B., Smeaton, A., Mayer-Patel, K., Avrithis, Y. (eds.) MMM 2009. LNCS, vol. 5371, pp. 298–309. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92892-8_32
9. Hawthorne, C., et al.: Onsets and frames: dual-objective piano transcription. arXiv preprint arXiv:1710.11153 (2017)
10. Kendrick, P., Cox, T.J., Zhang, Y., Chambers, J.A., Li, F.F.: Room acoustic parameter extraction from music signals. In: IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP), vol. 5, pp. V801–V804 (2006). https://doi.org/10.1109/ICASSP.2006.1661397
11. Kim, C., Stern, R.: Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In: Ninth Annual Conference of the International Speech Communication Association. pp. 2598–2601 (2008)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Kuttruff, H.: Room acoustics. CRC Press (2016). https://doi.org/10.1201/9781315372150
14. Mauch, M., Ewert, S.: The audio degradation toolbox and its application to robustness evaluation. In: International Society for Music Information Retrieval Conference (ISMIR), pp. 83–88 (2013)
15. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: audio and music signal analysis in Python. In: Proceedings of the 14th Python in Science Conference. vol. 8, pp. 18–25 (2015). https://doi.org/10.25080/MAJORA-7B98E3ED-003
16. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 2, pp. 749–752. IEEE (2001). https://doi.org/10.1109/ICASSP.2001.941023
17. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 1041–1044 (2014). https://doi.org/10.1145/2647868.2655045
18. Schörkhuber, C., Klapuri, A.: Constant-Q transform toolbox for music processing. In: 7th Sound and Music Computing Conference, Barcelona, Spain, pp. 3–64 (2010)
19. Schroeder, M.R.: New method of measuring reverberation time. J. Acoustical Soc. America **37**(6), 1187–1188 (1965). https://doi.org/10.1121/1.1939454

20. Serizel, R., Turpault, N., Shah, A., Salamon, J.: Sound event detection in synthetic domestic environments. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 86–90. IEEE (2020). https://doi.org/10.1109/ICASSP40776.2020.9054478
21. Subramanian, V., Benetos, E., Sandler, M.: Robustness of adversarial attacks in sound event classification. In: 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), pp. 239–243 (2019)
22. Szöke, I., Skácel, M., Mošner, L., Paliesek, J., Černockỳ, J.H.: Building and evaluation of a real room impulse response dataset. IEEE J. Selected Top. in Signal Process. **13**(4), 863–876 (2019). https://doi.org/10.1109/JSTSP.2019.2917582
23. Uemura, A., Ishikura, K., Katto, J.: Effects of audio compression on chord recognition. In: Gurrin, C., Hopfgartner, F., Hurst, W., Johansen, H., Lee, H., O'Connor, N. (eds.) MMM 2014. LNCS, vol. 8326, pp. 345–352. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-04117-9_34