Supplementary Material

Guitar Music Transcription from Silent Video

Shir Goldstein, Yael Moses

For completeness, we present detailed results and analysis of tests presented in the paper, as well as implementation details. Additionally, we present a demonstration on polyphonic data.

1 Temporal Segmentation - Implementation Details

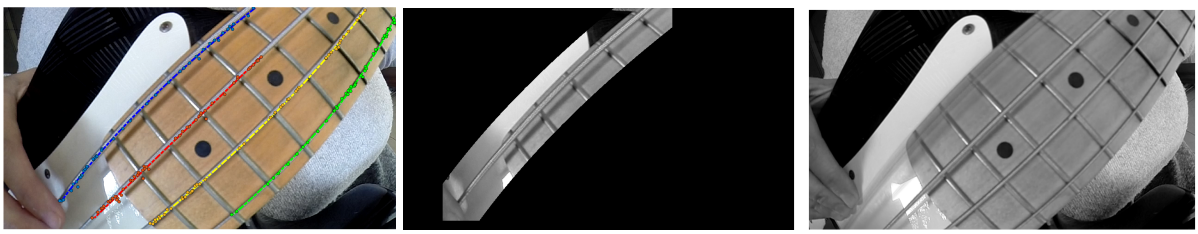
In this section, we present additional details about the temporal segmentation method. The temporal segmentation was computed using a thresholding of the spectrogram of each string-pixel. The threshold was defines using the assumption that notes were played around roughly 50% of the frames. Although this assumption holds in practice, we also would like to consider cases where significantly less frames are played on a string or when no notes are played. In this case, a different threshold should be defined based on the percent of played frames. (See Section 3.1.)

We proposed a spatial method for determining the set of frames in which a note is played for each given string. The basic assumption we use is that when a string is played it vibrates fast and hence no edges will be detected along the string on the average of a small set of frames (we consider 10 frames), see Fig. 1 (c).

We thus compute the edge image of the first frame of the video using Sobel edge detection. We then consider each string separately by applying a mask calculated by a dilation of a curve fitted though the string-pixels of each string (Fig. 1 (a) ,(b)). Then, for every 50 frames we conclude that a note is played if the first 10 frames and the last 10 frames has substantially less edges (Fig. 2). We then can estimate how many frames consist of played notes and calculate the threshold for the spectrogram accordingly. For example, if no notes where detected, the thresholded spectrogram would be empty.

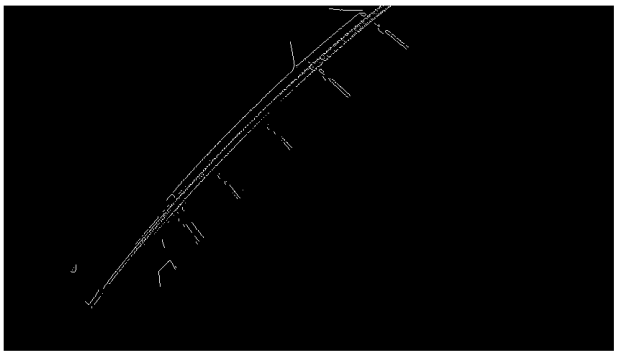
We test this algorithm on 8 videos, 2 for each string, that capture the bass guitar playing the chromatic scale on a single string but evaluate all strings, including unplayed ones (120 played notes in total).

Evaluating this data only for played strings without using the proposed spatial method yielded 105 TP and 45 FP resulting in a recall of 0.875 and precision of 0.7. Evaluating this data for all strings, including the unplayed ones, yielded similar recall of 0.83 (100 TP). Precision slightly drops to 0.61, since 63 FP where detected, mostly for the unplayed strings.

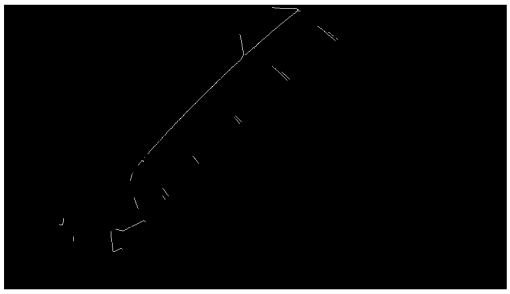


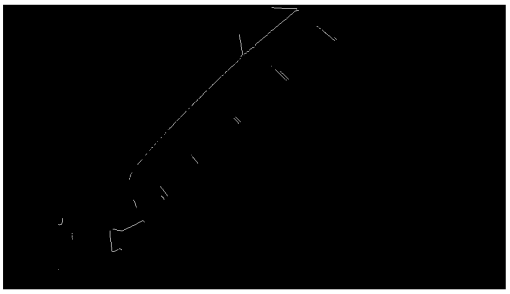
(a) (b) (c)

Figure 1: (a) The string-pixels for each string and the corresponding curves fitted. (b) The masked area for string E on the gray image. (c) The average of 10 frames when a note is played on the E string resulting in a blurred, edgeless string.



(a)





(b) (c)

Figure 2: (a) The edges images computed for the first video frame, and two frames that are the average of the first 10 frames in a section (b) and the 10 last ones (c). Clearly, when a note is played, less edges appear in the average images.

2 The String Pixels Choice

We next present the results of our method, when the string-pixels were manually chosen or computed by our calibration method using open string notes or high-fret notes (see Sec. 4 in the paper). Example of string-pixels of the three types are shown in Figure. 3. The results presented in the paper were obtained using the string-pixels chosen with high-fret notes. In Table 1 we compare the results of our method for the three types of string-pixels: a slight superiority of the results obtained by high-fret notes for calibrations can be observed.

(a) High-fret string-pixels (b) Open string string-pixels (c) Manual string-pixels

Figure 3: The string-pixels obtained using three different methods: **(a)** String-pixels obtained by using the temporal-spectral algorithm with a high-fret note (6, 7 or 8). **(b)** String-pixels obtained by using the temporal-spectral algorithm with an open string note. **(c)** String-pixels that were marked manually upon the string.

	Frame-by-Frame	Onsets F-measure	PD with GT int.
Manual	**79 %**	0.85	85 %
Auto. Open-String	73 %	0.85	84 %
Auto. High Fret	**79 %**	**0.87**	**86 %**

Table 1: Different string-pixels detection methods and their performances, evaluated by the different evaluation methods.

3 Polyphonic Demo

In order to demonstrate polyphony (test 5 in the paper), we played a polyphonic music that includes five chords played on an acoustic guitar, that is, played on 5 or 6 strings simultaneously. The sequence of chords played was: C (open), G (open), F (Barred), D (open), and G (open). For the temporal segmentation, we assume that the guitarist plays chords only. Hence, only temporal intervals that overlap in at least three strings were considered and temporal intervals lasting less than 48 frames (200 ms.) were discarded. The results are shown in Fig. 4 Note that in this case the offset is irrelevant. The time-overlapping temporal segments were shortened to the shortest interval in the chord to provide better visualization. In practice, when playing chords, a letter representing the chord will appear in the musical sheet (for example, Am7), and no offset notation is used.

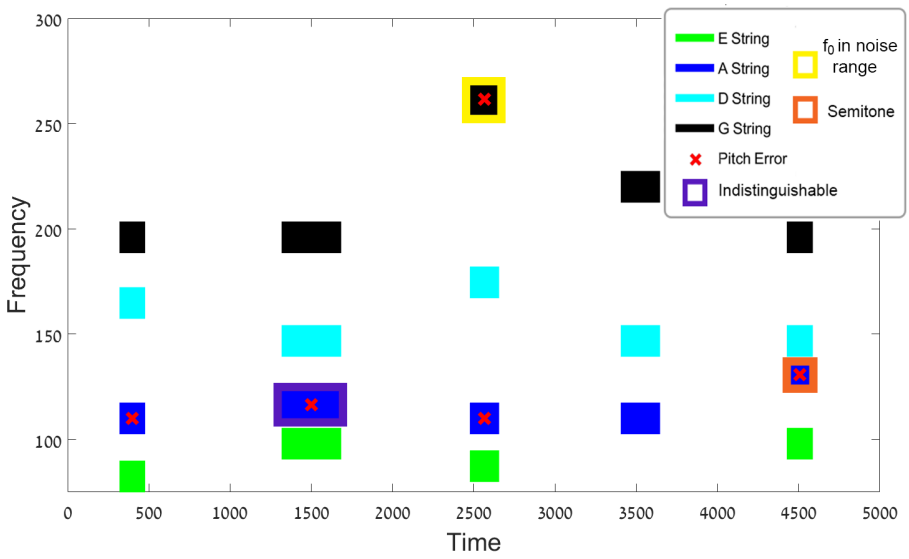


Figure 4: The method's output in "chord mode", for a video capturing the playing of the chords Cmaj-Gmaj-Fmaj-Dmaj-Gmaj. Notes detected with the wrong pitch are marked by x. The color of the rectangles around the detected notes indicate the type of error. The offsets are disregarded as in common chord notations.

4 Tests Results and Analysis

In this section we present a full detailed results for tests 1-3 classified by instrument, string and fret. We refer to the expected failures of the pitch detection presented in the paper at the end of Sec. 3.2 and in Fig. 4.

4.1 Temporal Note Segementaion (Test 1):

For the temporal note segmentation (Test 1), the results per string and instrument are presented in Table 2 and Table 3, and per string and fret in Table 4.

In Fig. 5a we present the recall is presented by frets.

In addition, we include in the appendix the onset results for additional three different thresholds, for all instruments 12 and per instrument 11. Note that the results mentioned throughout the paper are achieved with threshold = 80.

Discussion

First, as is mentioned throughout the paper, the bass guitar yields the best results, due to strings length and thickness, as well as the fact that bass guitar's notes rarely fall in the noise range. Additionally, better results were obtained for the lower strings (E and A) rather than for the higher ones (D and G). This may be explained by the weaker signal of the higher, thinner strings. In addition, 17% of the notes in the two higher strings has their f_0 in the noise range, as opposed to only 3% for the two lower ones. The better results obtained for the low-fret notes may be explained by their longer length, which causes a higher vibration amplitude.

	Acoustic			**Classical**			**Electric**		
	Onset		Offset TP*	Onset		Offset TP*	Onset		Offset TP*
Str.	Rec.	Prec.		Rec.	Prec.		Rec.	Prec.	
E(1)	0.98	0.81	0.73	0.90	0.65	0.60	1	0.96	0.81
A(2)	0.94	0.8	0.71	0.87	0.625	0.46	1	0.96	0.94
D(3)	0.71	0.64	0.41	0.58	0.61	0.55	0.92	0.87	0.81
G(4)	0.75	0.48	0.33	0.75	0.54	0.72	0.96	0.75	0.52

Table 2: Classical, acoustic and electric guitars onset and offset detection results out of 52 notes played per string on each instrument. Recall and precision are presented for each string and instrument. For offset we specify TP*, which is TP offsets out of the TP onsets.

	Bass		
	Onset		Offset TP*
String	Recall	Precision	
E(1)	1	0.97	0.97
A(2)	0.95	0.95	0.79
D(3)	0.95	0.68	0.79
G(4)	0.85	0.72	0.76

Table 3: Bass guitar onset and offset detection results out of 60 notes played per string on each instrument. Recall and precision are presented for each string and instrument. For offset we specify TP*, which is TP offsets out of the TP onsets.

Fret	Err. St. 1	Err. St. 2	Err. St. 3	Err. St. 4	Total Err.	% Total Err.
0	0	1	4	1	6	9%
1	0	0	0	1	1	2%
2	0	0	0	5	5	8%
3	0	0	0	7	7	11%
4	0	1	1	2	4	6%
5	0	0	1	5	6	9%
6	0	0	2	1	3	5%
7	0	0	4	2	6	9%
8	0	2	6	0	8	12.5%
9	0	1	11	2	14	22%
10	3	0	6	1	10	16%
11	1	2	4	4	11	17%
12	2	4	5	2	13	20%
Total Err.	6	11	44	33	94	**11%**
% Total Err.	3%	5%	21%	16%	**11%**	

Table 4: Summary of the errors of the temporal note segmentation (true-positive) per string and fret. The last two columns and rows show the total error per each fret and string respectively and the corresponding percent from the notes played per fret / string.

4.2 Pitch Detection (Tests 2 & 3):

We present our pitch detection algorithm results (Test 2 and Test 3). In Table 5, Table 6, Table 8 and Table 9 we analyse the different types of errors, per string and fret. Table 7 and Table 10 summarize the results. We show the results first for manually obtained temporal intervals, and then for intervals obtained by our temporal note segmentation algorithm.

Discussion

The performance on the two lower strings was better than on the two higher ones. However, the best results were achieved for the second lowest string (A) rather than the lowest one (E). This can be explained by the sparse fundamentals on string A, as opposed to the relatively dense fundamentals of string E (Fig. 4(b) in the paper). This property affects the pitch detection and not the temporal segmentation; thus string E had the best temporal note segmentation results.

Using the classification of expected failures, we found that 2% are harmonic errors, 6% are semitone, 4% occurred in indistinguishable notes, and 13% of the errors were obtained in noise-incident notes. Fig. 5b presents the errors obtained per fret and those obtained only for noise-incident notes. The two sets of errors are correlated for notes played up to the 8th fret. The errors for higher-fret notes correlate with the errors of their temporal segmentation in Test 1; signals that are hard to detect are also hard to analyze.

In those tests again, the Bass had superior results.

Err. / Str.	Total Err.	$S_{expected}$			$S_{noise\_incident}$		
		HE	SE	IN	N_{f0}	N_{h2}	N_{f0} & N_{h2}
E(1)	70 (32%)	10 (7%)	0	12 (6%)	0	13 (6%)	0
A(2)	54 (25%)	0	6 (3%)	15 (7%)	5 (2%)	19 (9%)	0
D(3)	86 (40%)	4 (2%)	17 (8%)	4 (2%)	12 (6%)	4 (2%)	24(11%)
G(4)	82 (38%)	0	27 (12.5%)	4 (2%)	11 (5%)	4 (2%)	24 (11%)
Total	292 (34%)	14 (2%)	50 (6%)	35 (4%)	28(3%)	40 (5%)	48 (6%)

Table 5: Pitch detection errors given GT temporal intervals, per string. The first column summarize the total errors of each string and the respective percent from the total number of played notes, 216 (per string). The other columns present the number of errors for each type and set defined above, and the respective percent from the total played notes.

Err. / Fret	Total Err.	$S_{expected}$			S_{noise_range}		
		HE	SE	IN	N_{f0}	N_{h2}	N_{f0} & N_{h2}
0	4 (6%)	1 (2%)	0	0	0	3 (5%)	0
1	0	0	0	0	0	0	0
2	23 (36%)	0	0	12 (19%)	8 (12.5%)	12 (19%)	0
3	23 (36%)	0	0	0	0	0	12 (19%)
4	21 (33%)	0	7 (11%)	4 (6%)	0	4 (6%)	12 (19%)
5	11 (17%)	0	2 (3%)	0	0	1 (2%)	0
6	11 (17%)	0	3 (5%)	0	0	0	0
7	28 (44%)	0	2 (3%)	12 (19%)	12 (19%)	12 (19%)	0
8	27 (42%)	0	1 (2%)	0	0	0	12 (19%)
9	34 (53%)	0	10 (16%)	4 (6%)	0	4 (6%)	12 (19%)
10	34 (53%)	0	14 (22%)	0	0	0	0
11	27 (42%)	0	6 (9%)	0	0	0	0
12	37 (58%)	13 (20%)	1 (2%)	0	5 (8%)	0	0
13*	2(12.5%)	0	1 (6%)	0	0	0	0
14*	10 (62.5%)	0	3 (19%)	3 (19%)	3 (19%)	4 (25%)	0
Total	292 (34%)	14 (2%)	50 (6%)	35 (4%)	28(3%)	40 (5%)	48 (6%)

Table 6: Pitch detection errors given GT temporal intervals, per fret. Note, *frets 13 and 14 are only applicable for the bass guitar. The first column summarize the total errors of each fret and the respective percent from the total number of played notes, 64 for frets 0-12 and 16 for frets 13-14. The other columns present the number of errors for each type and set defined above, and the respective percent from the total played notes.

Instrument	% **Success**
Bass	86
Acoustic	64
Classic	56
Electric	55
Total	**66**

Table 7: The pitch detection success percentage using GT temporal intervals, per instrument, counting the number of correctly detected pitches of the total notes played on each instrument - 240 for the bass guitar and 208 for all other guitars.

Err. / String	Matched	Total Err.	$S_{expected}$			S_{noise_range}		
			HE	SE	IN	N_{f0}	N_{h2}	N_{f0} & N_{h2}
E(1)	210 (97%)	68 (32%)	10 (5%)	0	12 (6%)	0	13 (6%)	0
A(2)	206 (95%)	55 (27%)	0	5 (2%)	13 (6%)	6 (3%)	18 (9%)	0
D(3)	196 (91%)	74 (38%)	5 (3%)	5 (3%)	4 (2%)	12 (6%)	4 (2%)	13 (7%)
G(4)	197 (91%)	74 (38%)	0	24 (12%)	4 (2%)	7 (4%)	5 (3%)	17 (10%)
Total	809 (94%)	271 (33%)	15 (2%)	33 (4%)	33 (4%)	25 (3%)	40 (5%)	30 (4%)

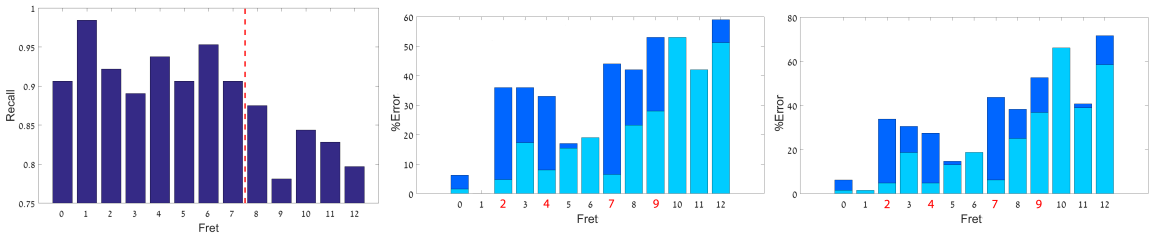
Table 8: Pitch detection errors given automatically segmented temporal intervals, per string. The first column is the number of intervals that were matched with a GT interval and corresponding percent of the total number of the GT notes, 216 (per string). The second column presents the errors per string in pitch detection only for the matched detected notes. The other columns present the number of errors for each type and set defined above, and the respective percent from the total played notes.

Err. / Fret	Matched	Total Err.	$S_{expected}$			S_{noise_range}		
			HE	SE	IN	N_{f0}	N_{h2}	N_{f0} & N_{h2}
0	64 (100%)	4 (6%)	1 (2%)	0	0	0	3 (5%)	0
1	64 (100%)	1 (2%)	0	0	0	0	0	0
2	62 (97%)	21 (34%)	0	1 (2%)	11 (18%)	6 (10%)	12 (19%)	0
3	59 (92%)	18 (31%)	0	1(2%)	0	0	0	7 (12%))
4	62 (97%)	17 (27%)	0	6 (10%)	4 (6%)	0	4 (6%)	10 (16%)
5	61 (95%)	9 (15%)	0	1 (2%)	0	0	1 (2%)	0
6	64 (100%)	12 (19%)	0	1 (2%)	0	0	0	0
7	64 (100%)	27 (42%)	0	1 (2%)	12 (19%)	12 (19%)	12 (19%)	0)
8	60 (94%)	23 (39%)	0	0	0	0	0	8 (13%)
9	57 (89%)	30 (53%)	0	3 (5%)	4 (7%)	0	4 (7%)	5 (9%)
10	59 (92%)	39 (66%)	0	17 (29%)	0	0	0	0
11	54 (84%)	22 (41%)	0	0	0	0	1(2%)	0
12	53 (83%)	38 (72%)	14 (26%)	1 (2%)	0	6 (11%)	1(2%)	0
13*	15 (94%)	3(20%)	0	1 (7%)	0	0	0	0
14*	11 (69%)	7 (64%)	0	1 (9%)	2 (18%)	1 (9%)	2 (18%)	0
Total	809 (94%)	271 (33%)	15 (2%)	34 (4%)	33 (4%)	25 (3%)	40 (5%)	30 (4%)

Table 9: Pitch detection errors given automatically segmented temporal intervals, per fret. Note, *frets 13 and 14 are only applicable for the bass guitar. The first column is the number of intervals that where matched with a GT interval and corresponding percent of the total number of the GT notes (64 for frets 0-12 and 16 for frets 13-14). The second column presents the errors per fret in pitch detection only for the matched detected notes. The other columns present the number of errors for each type and set defined above, and the respective percent from the total played notes.

Instrument	% **Matched**	% **Success**
Bass	96	88
Acoustic	88	65
Classic	93	55
Electric	98	55
Total	**94**	**67**

Table 10: Total pitch detection success percentage using automatically obtained temporal intervals, per instrument. The first column is the percent of matched intervals from the notes played - 240 for the bass guitar and 208 for all other guitars. The second column is the percent of the successful pitch detection from the total notes matched.



(a) (b) (c)

Figure 5: Results per fret (up to the 12th fret). (a) Onset detection recall. (b) Errors of pitch detection using the GT temporal interval and (c) when using the computed temporal intervals. The number of errors in incident-noise notes is marked in dark-blue and for other notes in light-blue. Frets with indistinguishable notes are marked with the fret number in red.

Bass guitar

	Threshold = 12		Threshold = 24		Threshold = 36	
String	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**
E (1)	0.62	0.6	0.82	0.79	0.98	0.95
A (2)	0.8	0.8	0.92	0.92	0.93	0.93
D (3)	0.33	0.24	0.63	0.45	0.8	0.57
G (4)	0.33	0.29	0.55	0.46	0.68	0.58
Total	0.52	0.45	0.73	0.63	0.85	0.74

Electric guitar

	Threshold = 12		Threshold = 24		Threshold = 36	
String	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**
E (1)	0.48	0.46	0.88	0.85	0.96	0.93
A (2)	0.48	0.46	0.83	0.8	0.94	0.91
D (3)	0.48	0.45	0.85	0.8	0.92	0.87
G (4)	0.56	0.43	0.79	0.6	0.90	0.7
Total	0.5	0.45	0.84	0.76	0.93	0.84

Acoustic guitar

	Threshold = 12		Threshold = 24		Threshold = 36	
String	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**
E (1)	0.5	0.41	0.71	0.75	0.96	0.8
A (2)	0.5	0.43	0.75	0.64	0.83	0.7
D (3)	0.29	0.26	0.48	0.43	0.58	0.52
G (4)	0.21	0.13	0.4	0.26	0.46	0.29
Total	37.5	0.3	0.63	0.5	0.71	0.56

Classic guitar

	Threshold = 12		Threshold = 24		Threshold = 36	
String	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**
E (1)	0.29	0.21	0.54	0.39	0.73	0.53
A (2)	0.38	0.28	0.58	0.42	0.63	0.46
D (3)	0.27	0.29	0.37	0.39	0.46	0.49
G (4)	0.25	0.18	0.35	0.25	0.50	0.36
Total	0.3	0.23	0.46	0.36	0.58	0.46

Table 11: Onset detection results for all instruments, using different thresholds. A total of 240 notes are played on the bass guitar (60 per string) and 208 on each of the other guitars (52 per string).

Threshold = 12			Threshold = 24			Threshold = 36		
Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
0.36	0.43	0.39	0.56	0.68	0.61	0.64	0.79	0.71

Table 12: Precision, recall and F-measure evaluations for onset detection for all instruments, using different thresholds.