Journal of Multimedia Information System VOL. 9, NO. 1, March 2022 (pp. 69-72): ISSN 2383-7632 (Online)
http://doi.org/10.33851/JMIS.2022.9.1.69

Brief Paper:
Salient Chromagram Extraction Based on the Savitzky-Golay Filter for Cover Song Identification

Jin Soo Seo[1*]

Abstract: Extraction of a salient chromagram is utmost important for cover song identification. Cover song refers to a live performance, a remix, or a new recording of a previously recorded track. This paper utilizes the Savitzky–Golay filters in chromagram extraction for suppressing timber-related components of a music signal, which is not preserved while generating cover songs. By removing the timber-related components, the discriminative tonal components, which are conducive for cover song identification, are emphasized in chromagram. Experiments on cover song identification over two datasets show that the Savitzky–Golay filters are more effective in reducing timber effects in chromagram than other types of filters.

Key Words: Savitzky-Golay Filter, Chromagram, Cover Song.

I. INTRODUCTION

Cover song refers to a live performance, a remix, or a new recording of a previously recorded track. Cover song identification (CSI) is difficult and challenging because changing timbre, rhythm, song structure, main key, and lyrics, occurred during cover song generation, may produce highly different cover versions [1-2]. The practical applications of the CSI are copyright protection and music-archive management.

One commonly used musical property for CSI is the tonal contents of music, such as chromagram or pitch class profiles, which are independent of timber and loudness and thus suitable for CSI [3]. The chromagram vector is extracted in a short-time interval (called a frame) by quantifying the spectral energy of the octave-folded subbands. By pooling spectral energy into one octave, chromagram features identify pitches that differ by an octave, which is necessary for the CSI. However; timber variations, which are caused by changing singer or instrumentation during covers song generation, cannot be dealt with. To cope with the large variations in timbre, severable approaches have been proposed. In [4], Muller and Ewert proposed the chroma DCT-reduced log pitch (CRP) by utilizing the upper-frequency discrete cosine transform (DCT) coefficients of the spectral energy in extracting chromagram with the assumption that the lower-frequency components of the spectral energy are closely related to the aspect of timber and should be reduced. In [5], the trend estimation filters, such as the moving average and the Hodrick-Prescott (HP) [6] filter, was used in removing the trend of the spectral energy, which is smoothly-varying component and assumed to be closely-related to the timber. Although the HP filter has improved the CSI accuracy [5], the HP filter has only one parameter for tuning its filtering characteristics, which is a limitation further enhancing the CSI performance. In an effort to find a trend-estimation filter which can be easily-adjusted for attaining best CSI performance, this paper employs the Savitzky-Golay (SG) filters [7]. Employing the least-square fit and a polynomial function as a filter kernel, the SG filter is able to reduce noises and find a trend line for a signal. The SG filters are optimal in the sense that they minimize the least-squares error in fitting a polynomial to frames of a noisy signal. The SG filter was originally proposed for analytical chemistry and has been utilized in a number of applications including digital control systems, speech recognition, denoising, and signal enhancement. In this paper, the CSI performance of the chromagram using the SG filter is experimentally compared with that using other types of filters.

II. PROPOSED CHROMAGRAM EXTRACTION METHOD

2.1. Chromagram Extraction by Removing Local Trend

The baseline of the chromagram used in this paper is the chroma log pitch (CLP) [3], whose extraction is based on a pitch-frequency scale as shown in Fig. 1(a). First, the input music signal is decomposed into 88 frequency bands with

Manuscript received February 23, 2022; Revised March 11, 2022; Accepted March 13, 2022. (ID No. JMIS-22M-02-003)

Corresponding Author (*): Jin Soo Seo, +82-33-640-2428, jsseo@gwnu.ac.kr

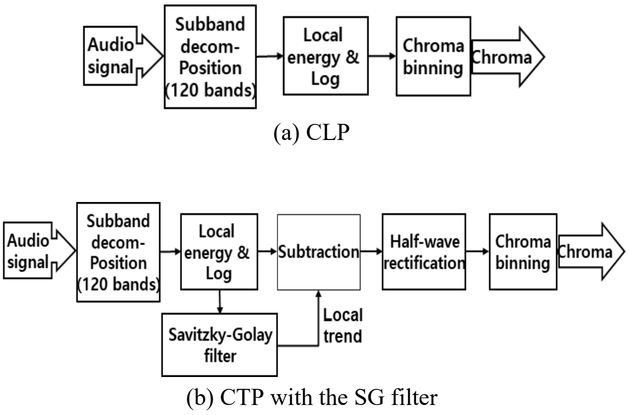
[1]Department of Electrical Engineering, Gangneung-Wonju National University, Gangneung, Korea, jsseo@gwnu.ac.kr

center frequencies corresponding to the MIDI pitches p = 21 to p = 108. Further details on the frequency band positions and bandwidth are described in [3]. At each of the 88 subbands, the short-time mean-square power (local energy) is calculated. As in [3], we add 20 zeros at the beginning and 12 at the end to construct a 120-dimensional feature vector where the entries correspond to MIDI pitches from p=1 to p=120. Then a logarithmic compression on the pitch representation is applied to account for the logarithmic sensation of sound intensity. Finally, the 12-dimensional chromagram is obtained by chroma binning, which adds up the corresponding values of the pitch representation that belong to the same chroma.

This work is an extension of the previous work in [5], where the chroma trend-removed log pitch (CTP) was proposed. Trend estimation tries to decompose a time-series signal into a medium-to-long term trend part and a short-term cycle part to detect and predict tendencies and regularities in the time series signal without knowing any information a priori about the signal. Mathematically, the decomposition of the given time series y_n into a trend x_n and a cycle c_n is expressed by

$$y_n = x_n + c_n, \qquad (1)$$

for n =1, 2, ..., N. The CTP is obtained by removing the trend of the spectral energy with an assumption that the trend of the spectral energy is closely related to the aspect of timber and thus should be removed for timber invariance. The overview of the CTP extraction is shown in Fig. 1(b): 1) estimating the trend of the 120-dimensional logarithmically compressed pitch representation; 2) subtracting the estimated trend from the pitch representation, which is equivalent to take the cycle part in (1); 3) taking the positive part of the trend-subtracted pitch by the half-wave rectification; and 4) performing the chroma binning. Since the local peaks of the spectral energy are related to music-specific harmony, emphasizing the peaky tonal components and reducing noise by removing trend along with the half-wave rectification are conducive in improving CSI accuracy. The performance of the CTP is contingent on the trend estimation [5], which is addressed using the SG filter in Section 2.2.



(a) CLP

(b) CTP with the SG filter

Fig. 1. Overview of the chromagram extraction from an audio.

2.2. Local Trend Estimation of Time Series Using the SG Filter

Throughout a number of different disciplines, such as macroeconomics, geophysics, biology, and social sciences, various trend estimation methods have been utilized. Among them, this paper considers the SG filter. Employing the least-square fit and a polynomial function as a filter kernel, the SG filter is able to reduce noises and find a trend line for a signal. The SG filters are optimal in the sense that they minimize the least-squares error in fitting a polynomial to frames of a noisy signal. Let y be a signal for a SG filtering. For a frame of length 2M + 1 from the signal y, denoted by $y_{-M}, y_{-M+1}, \dots, y_m, \dots, y_{M-1}, y_M$, we should find the coefficients, denoted by a_k, of a fitting polynomial with degree K given by

$$\mathrm{p(m)} = \textstyle\sum_{k=0}^{K} a_k m^k, \qquad (2)$$

with the objective of the minimization of the following mean-squared fitting error:

$$\begin{aligned} \varepsilon_K &= \textstyle\sum_{m=-M}^{M} (p(m) - y_m)^2 \\ &= \textstyle\sum_{m=-M}^{M} (\sum_{k=0}^{K} a_k m^k - y_m)^2. \end{aligned} \qquad (3)$$

The original paper by Savitzky and Golay [8] showed that at each position, the smoothed output value obtained by sampling the fitted polynomial is identical to a fixed linear combination of the local set of input samples; i.e., the set of 2M+1 input samples within the approximation interval are effectively combined by a fixed set of weighting coefficients that can be computed once for a given polynomial degree K and approximation interval of length 2M+1. Thus, the same weighting coefficients will be obtained at each group of 2M+1 input samples, and so we can think of least-squares smoothing as a shift-invariant discrete convolution process [7]. That is, the output samples of the SG filters can be computed by a discrete convolution instead of the polynomial fitting process in (2) and (3). Detailed analysis on the derivation and the characteristics of the SG filters is presented in [7].

There is no particular recommendation for the polynomial degree K and frame length 2M+1, which needs to be set beforehand. For most cases, the polynomial of degree up to 3 has been used for the SG filters [8]. Especially the polynomial with degree 0 is corresponding to the moving average filter. For a given polynomial degree, the larger the frame length, the more noise suppression and the less error variance of the filter output, but when the frame length is selected too large, the filter output, in comparison with the actual signal, becomes distorted and biased. The selection of the frame length and the polynomial degree of the SG filters with regard to the CTP extraction will be addressed in Section III.

III. EXPERIMENTAL RESULTS

The CSI accuracy of the proposed salient chromagram based on the SG filters was evaluated on two cover song datasets using the CSI method in [1]. The first cover song dataset (abbreviated as covers80) is the one that was used by Dan Ellis in his work [9]. The covers80 consists of 80 original and cover song pairs (160 songs in total), which are available online. The second cover song dataset (abbreviated as covers330) is composed of 1000 songs, where 330 songs are test data (30 original songs and 10 cover versions per each original song), and the other 670 songs were embedded as imposters. The covers330 was collected by the author. For the covers80 dataset, we calculated the precision at one, *P@*1, which is the rate of the covers correctly identified in top 1 when querying each song on the 160 songs in the dataset. For the covers330 dataset, we queried the 330 cover songs over the 1000 entire songs and computed the mean number of covers identified in top 10 (*MNCI*$_{10}$). For both datasets, we evaluated the average rank of the first correctly identified cover (*Rank*$_1$) and the mean of average precision (MAP). We follow the experimental procedures in the MIREX 2020 [10].

Each song in the datasets was converted to mono at a sampling frequency of 22050 Hz and then divided into frames of 200 ms overlapped by 100 ms where the 12-dimensional chromagram vector was computed as a low-level feature for each frame. The 12-dimensional chromagram vector was normalized with respect to the Euclidean norm to have unit length. In extracting chromagram, we utilized the pitch representation in the chroma toolbox [3] with the default parameter settings. From the pitch representation, we extracted three different types of the chromagram, CLP, CRP, and CTP, for evaluating the cover song identification performance. In extracting CTP, we utilize the SG filters and the HP filters. The HP filters performed best in the previous work [5].

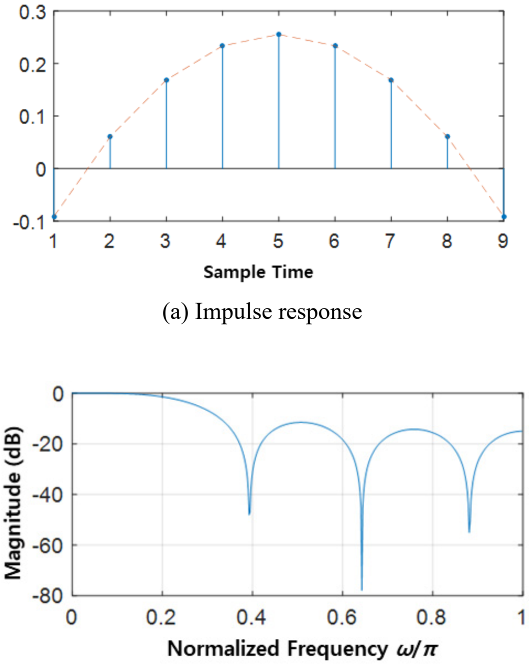
Table 1. Cover song identification performance of the covers80 dataset. Accuracy measures are the average rank of the first correctly identified cover, *Rank*$_1$, precision at one, *P@*1, and the mean of average precision, MAP.

Method	*M*	*K*	*Rank*$_1$	*P@*1	MAP
CTP - SG filter	1	0	17.72	0.588	0.643
	2	0	17.94	0.588	0.648
	2	2	20.34	0.556	0.609
	3	0	21.04	0.563	0.622
	3	2	**16.55**	0.588	0.651
	4	0	21.09	0.531	0.586
	4	2	18.07	**0.619**	**0.674**
	5	0	24.68	0.506	0.553
	5	2	18.74	0.569	0.641
	6	0	24.73	0.506	0.559
	6	2	24.61	0.550	0.611
CTP-HP filter [5]			17.23	0.613	0.669
CRP [4]			24.86	0.556	0.605
CLP [3]			24.14	0.588	0.631

Table 2. Cover song Identification performance of the covers330 dataset. Accuracy measures are the average rank of the first correctly identified cover, *Rank*$_1$, the mean number of covers identified within the ten first answers, *MNCI*$_{10}$, and the mean of average precision, MAP.

Method	*M*	*K*	*Rank*$_1$	*MNCI*$_{10}$	MAP
CTP - SG filter	1	0	4.75	7.042	0.729
	2	0	4.23	7.300	0.757
	2	2	8.15	6.536	0.675
	3	0	4.18	7.076	0.734
	3	2	4.52	7.309	0.755
	4	0	4.42	6.682	0.689
	4	2	4.36	**7.479**	**0.774**
	5	0	4.61	6.424	0.661
	5	2	4.28	7.391	0.761
	6	0	4.59	6.415	0.663
	6	2	5.02	7.055	0.729
CTP-HP filter [5]			4.24	7.430	0.767
CRP [4]			5.70	6.888	0.710
CLP [3]			**4.04**	7.164	0.739

Table 1 and Table 2 show the CSI performance of the CTP using the SG filters with different values of the frame length up to 13 was considered. We note that the impulse response of the SG filters with an odd degree, *K*, is the same as that with *K*-1 [7]. Thus we consider the SG filters with even degrees up to second order. For both SG and HP filters, the CSI accuracy of the CTP was better than that of the previous chromagrams; CLP and CRP. We note that the performance of the CTP using the HP filter in Table 1 and 2 is



(a) Impulse response

(b) Frequency response

Fig. 2. Savitzky-Golay filter with *M*=4 and *K*=2.

the best CSI accuracy achieved using the HP filter by adjusting the smoothing parameter. Regarding the type of the trend filter, the SG filter with *M*=4 and *K*=2 showed best performance for both datasets. The SG filter was more effective than the HP filter in boosting identification accuracy. Depending on the value of *M* and *K*, the SG filters possess different frequency responses. The impulse and the frequency response of the SG filter with *M*=4 and *K*=2 are shown in Fig. 2. The 3-dB cutoff frequency of the SG filter with *M*=4 and *K*=2 was 0.243 [7].

IV. CONCLUSION

In this paper, the SG filters have been utilized in removing the trend of the spectral energy for extracting a salient chromagram. The removal of trend emphasizes tonal contents of music, which are preserved against wide range of the possible distortions which may occur during cover song generation process. Appropriate choice of the trend-estimation filter is utmost important for attaining best performance, which is addressed in this paper. Experimental results on two datasets show that the use of the SG filter with appropriately-chosen parameters is effective in improving CSI accuracy. Further study includes a filter design for more discriminant and resilient chromagram extraction.

ACKNOWLEDGEMENT

The author is grateful to Gangneung-Wonju National University for granting sabbatical leave for this work.

REFERENCES

[1] J. Serra, E. Gomez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1138-1151, May 2008.

[2] J. S. Seo, "Improving cover song search accuracy by extracting salient chromagram components," *Journal of Korea Multimedia Society*, vol. 22, no. 6, pp. 639-645, Mar. 2019.

[3] M. Müller and S. Ewert, "Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features," in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Miami, FL, Oct. 2011, pp. 215-220.

[4] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 649-662, Mar. 2010.

[5] J. S. Seo, "Salient chromagram extraction based on trend removal for cover song identification," *IEICE Transactions on Information and Systems*, vol. 104, no. 1, pp. 51-54, Jan. 2021.

[6] R. J. Hodrick and E. C. Prescott, "Postwar U.S. business cycles: An empirical investigation," *Journal of Money, Credit and Banking*, vol. 29, no. 1, pp. 1-16, Feb. 1997.

[7] R. W. Schafer, "What is a Savitzky-Golay filter?," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 111-117, Jun. 2011.

[8] M. Sadeghi, F. Behnia, and R. Amiri, "Window selection of the Savitzky–Golay filters for signal recovery from noisy measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 8, pp. 5418-5427, Jan. 2020.

[9] D. P. W. Ellis and G. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking, " in *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, HI, Apr. 2007, pp. 1429-1432.

[10] Music Information Retrieval Evaluation eXchange (MIREX), https://www.music-ir.org/mirex/wiki/2020:Audio_Cover_Song_Identification, 2020.