Semi-Supervised NMF in the chroma Domain Applied to Music Harmony Estimation

Takuya Takahashi, Takeshi Hori, Christoph M. Wilk and Shigeki Sagayama
Meiji University, Tokyo, Japan
E-mail: {ev60552, hori, wilk, sagayama}@meiji.ac.jp

***Abstract*—In this paper, we discuss non-negative matrix factorization (NMF) applied to chroma feature sequences to reduce the chroma-specific noise in chord estimation from music signals using the hidden Markov model (HMM).**

Even in the case of single pitch sounds, the raw 12-dimensional chroma vectors obtained from the music signal by summing and normalizing the spectrum by octaves often contain irrelevant components such as non-octave overtones falling into different pitch classes and cause inaccuracies in estimation of harmonies. NMF applied to the chroma domain is expected to suppress such chroma components in the NMF activation matrix caused by overtones, and thus "purifies" the noisy chroma vectors. By reducing the dimensionality to 12 dimensions as opposed to NMF applied to the raw spectrum, we expect advantages with respect to statistical robustness as well as computational cost for pitch class estimation of single and multiple tones.

We use the "purified" chroma vectors in combination with a harmony progression model based on an HMM where the NMF activation distributions are modeled as observations associated with hidden harmonies, whose transition probabilities have been obtained statistically. We attempt to improve harmony estimation accuracy by combining suppression of irrelevant components and the HMM-based harmony model.

In the experimental evaluation, we demonstrate the reduction of irrelevant components in raw chroma vectors computed from recordings of musical instruments. In addition, using music audio data with harmony annotation from the RWC database, we compare the harmony estimation accuracies using our method and conventional chroma.

I. Introduction

In this paper, we propose a non-negative matrix factorization method for chroma sequences to reduce chroma specific noise for chord estimation, utilized in a hidden Markov model (HMM).

Techniques for analyzing music signals are used in tasks such as chord estimation, key estimation, and automatic transcription. Chroma vectors are commonly used for harmony estimation and contain magnitude values of the 12 pitch classes of western music [1]. To account for the time sequential nature of harmony progressions, HMMs are also often used for harmony estimation [2]. Futhermore, Saito et al. proposed a harmony estimation method using Specmurt chroma [3], which is a method suppress harmonic overtones in a music signal [4]. Also, as a method using music theory, Ueda et al. describe a chord estimation method taking into account key modulation using functional harmony [5], Uemura et al. proposed a chord estimation method based on harmony similarity using a doubly nested circle of fifths [6], Mauch et al. proposed a harmony estimation method focusing on the relationships between harmonies and their constituent notes and using a dynamic Bayesian network to account for the structure of music such as repetition patterns [7], [8], [9]. Kurauchi et al. proposed a chord estimation method based on the fact that valley frequencies (Spectrum Dip) differ between different harmonies [10], and Kurokawa et al. performed harmony estimation using that Spectrum Dip method with chroma vectors [11].

A lot of music analysis methods using NMF have been published. Raczyński et al. analyzed music signals using 12 NMF basis vectors, one for each pitch class [12]. Maruo et al. proposed a method combining harmony estimation using a Bayesian HMM with pitch estimation utilizing a Bayesian NMF [13].

Furthermore, harmony analysis from music audio signals using neural networks has been reported to record high harmony analysis accuracy. Shigtia et al. proposed a harmony analysis method in which harmony sequence HMM was replaced with a RNN [14]. Filip Korzeniowski and Gerhard Widmer proposed a harmony estimation method by using convolutional neural network and conditional random field [15].

Music audio signal analysis methods using chroma vectors are widely used, but conventional chroma vectors contain irrelevant spectrum components (e.g. harmonic overtones), which make harmony estimation more difficult. Therefore, in this paper, we propose a method to reduce irrelevant components of chroma vectors using NMF. By using this method, we expect to obtain activation distributions with reduced noise. Moreover, compared to conventional NMF analysis of music signals, we expect to improve calculation speed and to reduce statical error, because the basis is reduced to 12 dimensions. Furthermore, even with sound sources that contain a lot of noise, harmony estimation accuracy improvement could be expected, if the NMF basis vectors are learned for the noisy data.

II. semi-supervised chroma-NMF

A. Chroma Feature Value

A chroma matrix, denoted as $\mathrm{ch}(k,t)$, and proposed by Fujishima[1] can be obtained from the squared values of a semitone Constant-Q Transform (CQT) $\Phi(f,t)$.

$$\mathrm{ch}(k,t) = \log(\sum_{i=0}^{n} \Phi(12i+k,t)),\ k = 0, \cdots, 11 \quad (1)$$

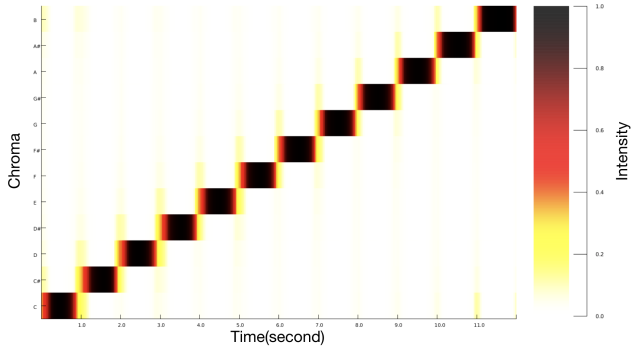

Fig. 1. Chroma of an audio signal containing the 12 pitches of the chromatic scale generated with sine waves in 1 second intervals.

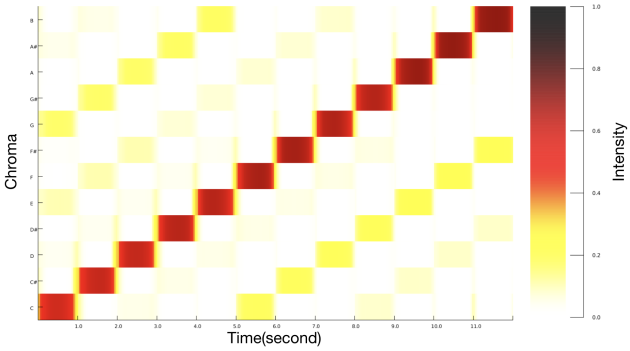

Fig. 2. Chroma of an audio signal containing the 12 pitches of the chromatic scale generated with sawtooth waves in 1 second intervals.

where n represents the number of octaves included, $\Phi(f,t)$ represents the magnitude of the CQT at frequency f and time t, and $\mathrm{ch}(k,t)$ represents the magnitude of the pitch class number k at time t.

Fig. 1 shows the chroma matrix of an audio signal containing sine waves with pitches ascending in semitones from C to B in one second intervals. Fig. 2 displays a chroma matrix similarly obtained using sawtooth waves. Comparing the two graphs, one can see the irrelevant components that result from the harmonic structure of the sawtooth waveform, even occurring in recorded monophonic sound.

B. KL-Divergence Standard Non-negative Matrix Factorization

Non-negative Matrix Factorization(NMF) is an algorithm that decomposes one non-negative matrix into two non-negative matrices. It is applied in various fields such as image, sound, biological signal analysis. Especially in the field of audio analysis, NMF can be applied easily, because a spectrum is a nonnegative matrix. In this paper, we define:

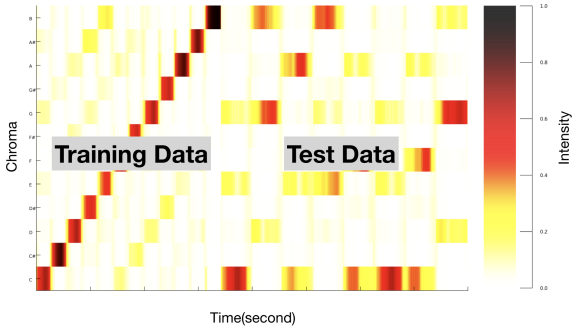

Fig. 3. The matrix that is concatenated the chroma matrices of labeled training data and test data to be analyzed.

Non-negative Matrix (spectrum): $\boldsymbol{Y} \in \mathbb{R}^{M \times N}$

Basis matrix: $\boldsymbol{W} \in \mathbb{R}^{M \times R}$

Activation matrix: $\boldsymbol{H} \in \mathbb{R}^{R \times N}$

The two non-negative matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ are computed to approximate the original matrix as follows.

$$\boldsymbol{Y} \approx \boldsymbol{W} \times \boldsymbol{H} \tag{2}$$

In this paper, we use the KL divergence in the NMF algorithm[12]. The basis matrix $\boldsymbol{W}$ and activation matrix $\boldsymbol{H}$ are obtained using the following update functions.

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W^T} \dfrac{\boldsymbol{Y}}{\boldsymbol{WH}}}{\boldsymbol{W^T 1}} \tag{3}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\dfrac{\boldsymbol{Y}}{\boldsymbol{WH}} \boldsymbol{H^T}}{\boldsymbol{1 H^T}} \tag{4}$$

where the element-wise product of $\boldsymbol{X}$ and $\boldsymbol{V}$ is represented by $\boldsymbol{X} \odot \boldsymbol{V}$. The division of matrices is performed for each element.

C. Semi-Supervised Chroma-NMF

"Semi-supervised chroma-NMF" is a method to decompose chroma matrices obtained by processing the power spectrogram of music audio signals into basis and activation matrices. In this paper, the number of basis vectors of the NMF is set to 12 corresponding to the 12 pitch classes. Before the training process, the chroma basis vectors are initialized such that the magnitude of the respective pitch class is significantly larger than all other magnitudes. A row of $\boldsymbol{W}$ encodes the magnitude distribution of each chroma vector (including overtones), and a column of $\boldsymbol{W}$ represents a basis vector corresponding to each of the 12 note classes. A row of $\boldsymbol{H}$ encodes development over time, and an activation column the magnitude of each chroma vector at the corresponding point in time.

As the first step of our method, a constant-Q transform is applied to an audio signal and the result converted into chroma vectors by summing magnitudes over all octaves. We assume that the harmonic structure of each instrument is similar. We exploit this by applying NMF on unknown data and data containing groundtruth labels simultaneously. Thereby the algorithm is guided to learn harmonic structures of notes as basis vectors due to the existing labels of the groundtruth part of the data, while it is also able to account for harmonic properties of the intruments in the music to be analyzed. We implement such a semi-supervised learning process by concatenating the chroma matrices of labeled training data and test data to be analyzed(Fig.3). NMF is then applied to the concatenated matrix, thereby being influenced by both data types when estimating optimal basis vectors. The concatenated chroma matrices are denoted as $\boldsymbol{Y}$.

The part of the activation matrix $\boldsymbol{H}$ in the time frame of the training data is initialized according to the sound attenuation model. The sound attenuation model assumes that the power of a single tone diminishes exponentially with time (like when hitting a piano key). The basis matrix $\boldsymbol{W}$ and activation matrix $\boldsymbol{H}$ are updated using equation (3) and (4). However, the training data portion of the activation is not updated.

We assume that the relative harmonic overtone structure is the same for every pitch class, independent from absolute pitch, i.e. the columns of the basis matrix can be matched by shifting. Therefore, we suppress variations between in basis vectors by averaging the correspondingly shifted basis vectors after each update step. The average vector is set as the first column of the basis matrix $\boldsymbol{W}$. In the second column, the mean vector shifted by one semitone. Accordingly, we shift the elements of the mean vector for each column and set the average vector to all twelve columns.

Updating is repeated until the likelihood in the KL-divergence becomes sufficiently small. In this paper, in order to avoid the arbitrariness of scale of the basis matrix $\boldsymbol{W}$, it was was normalized to $\sum_m W_{m,r} = 1$.

In our method, chroma matrices are obtained by decomposing power spectrogams of music audio signals into basis and activation matrices using a NMF algorithm. And advantage of our method is the improvement of note basis vector estimation using easily obtainable training data, which is however limited by the hypothesis that harmonic structures of different instruments are similar (in case of differing harmonic structures, estimation accuracy is expected to decrease). Therefore, method differs from previous research in that it facilitates the initialization of the algorithm and reduces computational cost. For intstance, a similar method proposed by Raczyński et al.[12] applies NMF for multi-pitch analysis, but does not reduce the spectrogram to chroma, instead defining 12 NMF basis vectors with 88 entries each and consequentially higher computational cost. Maruo et al.[13] proposed another similar method that first applies NMF with 88 basis vectors, which are supposed to learn harmonic structures of single notes and then computing chroma from the NMF result afterwards, which is the inversion of the two steps of our algorithm and requires more computational time. Lastly, Ueda et al.[5] have an objective very similar to ours, which is to reduce the irrelevant components in chroma vectors, but use a very different approach. They remove such irrelevant components by minimizing the non-diagonal entries of the chroma covariance matrix.

D. Semi-Supervised "Chroma-NMF" Applied to Chord Estimation

In this section, we describe a method to apply semi-supervised chroma NMF discussed in section II to chord estimation from audio signals.

Fujishima[1] proposed a method for harmony estimation by computing the Euclidean distance between chroma and harmony template vectors. However, the harmony template vectors used in the experiment of [1] are ideal vectors (12-dimensional vectors with the magnitudes of the pitch classes contained in the respective harmony set to 1, and all others set to 0) and the irrelevant components in the chroma vectors were not considered. Activation matrices obtained by chroma-NMF make it easier to determine where each note sounds, by removing irrelevant components. Therefore, improvement of the accuracy of harmony estimation using NMF harmony template vectors can be expected.

We use an HMM for smoothing of the estimation results and to reduce the influence of occasional nonharmonic tones. The hidden states of the HMM are the harmony labels. As a limitation of the model, high recognition accuracy can not be expected if the music audio signal includes many nonharmonic tones. Furthermore, if recognition of erroneous harmony labels is made due to nonharmonic tones, the recognition accuracy of the harmony of the entire song is degraded due to error propagation in the HMM.

The emission probabilities of the HMM are set according to template vectors of harmonies derived heuristically with reference to [1]. Each harmony template vector has twelve elements corresponding to each note class. A chord template vector t_i is denoted as follows.

$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_{12} \end{pmatrix} \tag{5}$$

where h denotes the harmony corresponding to the template vector. Each pitch class contained in a harmony is set to $1/N$ where N is the number of pitch classes in the harmony.

To compute the probability of a specific harmony h_k given an observed activation vector x, we first calculate the squared distance d between the harmony template vector and the observation.

$$d(h_k) = \sum (t_i - x_i)^2 \tag{6}$$

We then compute the probability from this distance as follows.

$$p(h_k|\mathbf{x}) = \frac{e^{-d(h_k)/2}}{\sum_i e^{-d(h_i)/2}} \tag{7}$$

TABLE I
PERCENTAGE OF IRRELEVANT COMPONENT MAGNITUDES IN CHROMA

	chroma	chroma-NMF
sawtooth	32.8%	22.2%
piano	24.2%	17.3%
trumpet	40.0%	19.8%

Percentages computed from an audio data generated from MIDI data containing Pachelbel's Canon's harmony sequence using different virtual instruments.

where the normalization factor is the sum over all possible harmonies. By considering $p(h_k|\mathbf{x})_t$ as the emission probability of the HMM at time t and performing maximum likelihood estimation, a harmony label sequence is obtained. The transition probabilities of the HMM are obtained by extracting harmony sequence statistics from an existing music corpus.

III. Evaluation Experiment on "Semi-Supervised Chroma-NMF"

In this section, we evaluate how well irrelevant components in chroma are suppressed when applying semi-supervised chroma NMF by comparing the results with conventional chroma.

A. Experimental Conditions

We prepared MIDI data which contains the following chord sequence (Pachelbel's Canon transposed to C).
$\{Cmaj, Gmaj, Amin, Emin, Fmaj, Cmaj, Fmaj, Gmaj\}$
MIDI velocity and tempo (one second per chord) is constant and all pitches are contained in the same octave. The MIDI data was then turned into audio data using simple sawtooth waveforms as well as virtual piano and trumpet instruments. In all three cases, audio data of a note scale (from C to B, one note per second) played on a piano was used as training data (as described in section II-C). The correct answer matrix of the training data was prepared taking not only onset times, but also the sound attenuation model into account, i.e. the assumption that the piano sound diminishes exponentially with time. Both training and test data were analyzed using CQT with a window shift of 40 ms. As a method of evaluation, we compare the ratio of irrelevant components (combined magnitude of pitch classes not in the MIDI data at the respective point in time) in chroma vectors obtained with and without NMF processing as shown in table I.

B. Experimental Result

Fig.4 and 7 display the chroma of the Canon chord sequence played by trumpet and piano, respectively. Fig. 5 and 8 show the activation of chroma-NMF played by trumpet and piano, respectively. Furthermore, the basis on which the sound source of the trumpet or the piano has been learned is shown in Fig. 6, 9. As shown in Table I, it was found that the magnitude of irrelevant components was reduced by using this method, in comparison with conventional chroma. As expected, the harmonic overtone component distributions of instruments

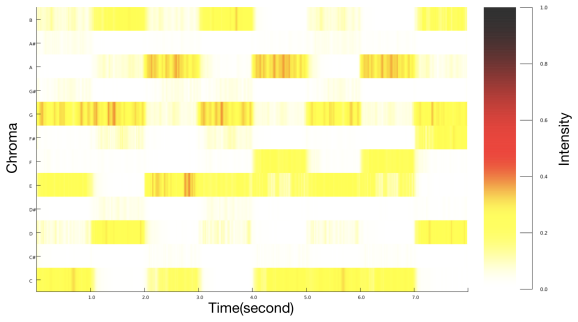

Fig. 4. Unprocessed chroma of Pachelbel's Canon (transposed to C) played by a virtual trumpet instrument.

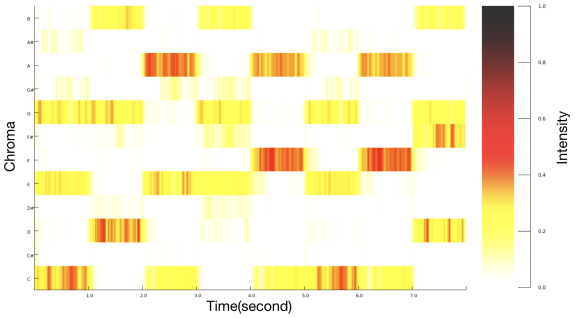

Fig. 5. Activation of semi-supervised chroma NMF of Pachelbel's Canon (transposed to C) played by a virtual trumpet instrument.

were learned in form of the vectors of the basis matrix $\mathbf{W}$, and consequently, the resulting activation matrix $\mathbf{H}$ contained less harmonic noise. Furthermore, the basis matrix $\boldsymbol{W}$ can be learned without requiring a lot of training data as opposed to methods based on neural networks. A disadvantage is that pitch classes can be mistakenly treated as overtones, e.g. when a C note and a G note are played at the same time. In this case, the G pitch class is a major overtone of the C pitch class. Consequently, The magnitude of actually played notes (the G note in the example) is also reduced as can be clearly seen in Fig. 7.

IV. Experimantal Evaluation of "Semi-Supervised Chroma NMF" Applied to Chord Estimation

A. Experimental Conditions

In this section, we discuss the evaluation of chord estimation based on the method described in section II-D. We chose some songs from the RWC database and our algorithm estimated chords from the activation matrix $\boldsymbol{H}$ obtained using semi-supervised chroma-NMF, finally applying a HMM for smoothing. At this time, taking into account the limitations of the model, songs without clear harmonies (e.g., containing a lot of parts where only single notes are played) were excluded

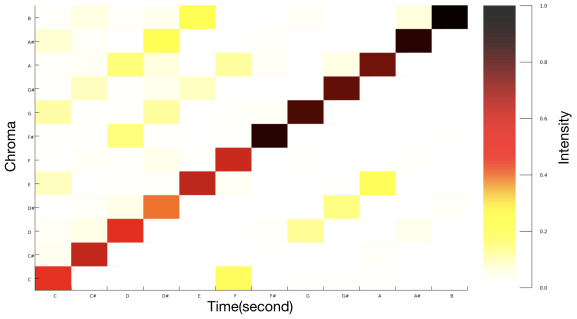

Fig. 6. Basis vectors of semi-supervised chroma NMF obtained from Pachelbel's Canon (transposed to C) played by a virtual trumpet instrument.

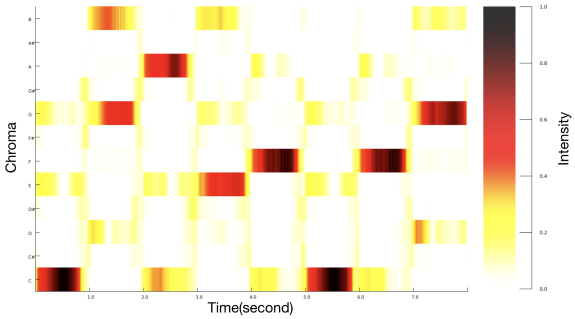

Fig. 8. Activation of semi-supervised chroma NMF of Pachelbel's Canon (transposed to C) played by a virtual piano instrument.

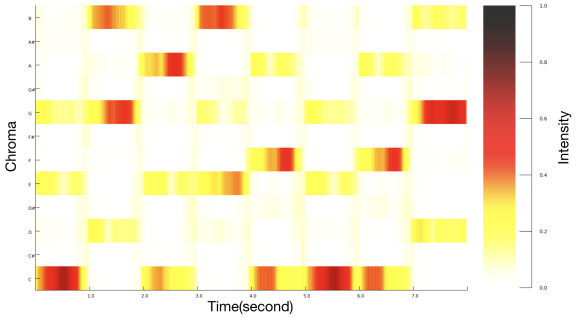

Fig. 7. Unprocessed chroma of Pachelbel's Canon (transposed to C) played by a virtual piano instrument.

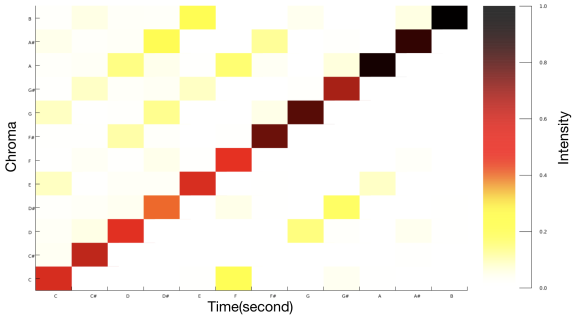

Fig. 9. Basis vectors of semi-supervised chroma NMF obtained from Pachelbel's Canon (transposed to C) played by a virtual piano instrument.

from the test data. We selected songs containing harmony for the most part from the RWC database. Specifically, these songs are shown in Table II.

For this paper, harmony estimation was performed with the following settings.

- CQT analysis was performed with a window shift of 40 ms.
- The harmony transition probabilities of the HMM were statistically obtained from the harmony-labeled data of the Isophonics database.
- Harmony labels were estimated for every 40 ms time-frame.
- Only the chord qualities major and minor were considered.

Similarly, chord estimation was performed based on minimum Euclidean distance between chroma-NMF activation vectors and harmony template vectors as described in section II-D. The harmony template vectors were defined by setting values of the N pitch classes contained in the harmony to $1/N$ and all other vector elements to 0. For comparison, chord estimation was also performed using conventional chroma as in [1]. The same harmony template vectors were used. For evaluation, harmony annotation data in [16] was used as the groundtruth. The estimation accuracy is shown in table II.

B. Experimental Results

As shown in table II, an estimation accuracy decrease of about 6% was observed in in comparison with estimation based on conventional chroma. In this experiment, we could not confirm significant improvement of estimation accuracy when applying chroma-NMF in combination with HMM smoothing for chord estimation. A reason for the accuracy decrease could be the confusion of played notes with overtones, resulting in reduction of not only overtone noise but also of harmony tone magnitudes. A further weakness, common to most chroma based harmony estimation methods is the problem that non-harmonic tones pose. Even if chroma allows to correctly identify sounding notes, chord estimation based on simple template vectors is unable to discern harmonic tones from non-harmonic ones, and therefore often fails to correctly identify the current harmony. Consequently, in order to improve the accuracy of harmony estimation, it is necessary to develop a harmony estimation method considering non-harmonic tones in real music. In particular, since a harmony template vectors are set heuristically, it is not possible to capture pitch class distributions of actual songs. Thus, if it

TABLE II
HARMONY ESTIMATION ACCURACY

	chroma	chroma-NMF	chroma-NMF &HMM
RWC-C2	5.6%	6.2%	6.9%
RWC-C6	27.5%	26.4%	31.3%
RWC-C22	52.3%	43.3%	34.4%
RWC-C23A	25.5%	16.4%	15.6%
RWC-C23E	40.7%	33.6%	35.7%
RWC-C28	41.1%	41.3%	38.0%
RWC-C29	62.9%	55.3%	61.9%
RWC-C30	63.0%	52.6%	63.0%
RWC-C32	41.0%	33.5%	31.7%
RWC-C33	30.7%	28.4%	25.5%
RWC-C35A	28.0%	44.2%	35.5%
RWC-C35B	36.8%	37.9%	28.0%
RWC-C35C	40.5%	26.0%	20.8%
Overall result	36.2%	31.6%	31.2%

was possible to accurately obtain harmony template vectors by another method, improvement of harmony recognition accuracy could be expected.

V. CONCLUSIONS

In this paper, we have proposed a method to analyze musical acoustic signals using semi-supervised chroma-NMF and discussed harmony estimation based on this method. By using semi-supervised chroma-NMF, it has been shown possible to suppress the magnitude of pitch classes of irrelevant overtones. However, the use of chroma-NMF combined with HMM could not demonstrate significant improvement of harmony estimation, possibly due to suppression of harmonic tones mistakenly treated as overtones. One of major reasons causing the low estimation precision is non-harmonic tones in the melody.

In the future research, we intend to implement more sophisticated emission probabilities for the harmony estimation HMM that can take non-harmonic tones into account. Furthermore, we plan to develop an algorithm which allows its user to manually add harmony labels to a small part of a song, which are then used for training in order to estimate the harmonies of the complete song.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 17H00749.

REFERENCES

[1] Takuya Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," *Proc. ICMC1999*, pp. 464–467, 1999.

[2] Kyogu Lee and Malcolm Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 2, pp. 291–301, 2008.

[3] Shoichiro Saito, Haruto Takeda, Takuya Nishimoto and Shigeki Sagayama, "Key Detection of Music Audio Signals via HMM using chroma Vector through Specmurt Analysis" *Proc. MUS*, Vol. 2005, No. 82 (2005-MUS-061), pp. 85–90, 2005.

[4] Shoichiro Saito, Hirokazu Kameoka, Keigo Takahashi, Takuya Nishimoto, and Shigeki Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE transactions on audio, speech, and language processing*, Vol. 16, No. 3, pp. 639–650, 2008.

[5] Yushi Ueda, Yuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama, "Hmm-based approach for automatic chord detection using refined acoustic features," *Proc. Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 5518–5521. IEEE, 2010.

[6] Aiko Uemura and Jiro Katto, "Chord Recognition using DNCOF Vector and chroma Vector," *Proc. MUS*, Vol. 2011, No. 5, pp. 1–6, 2011.

[7] Matthias Mauch and Simon Dixon, "Approximate note transcription for the improved identification of difficult chords," *Proc. ISMIR 2010*, pp. 135–140, 2010.

[8] Matthias Mauch and Simon Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 6, pp. 1280–1289, 2010.

[9] Matthias Mauch, "Automatic chord transcription from audio using computational models of musical context," 2010.

[10] Yuki Kurauchi, Masaki Matsubara, Masaki Oono and Hiroaki Saito, "Chord estimation based on valley points of frequency spectrum," *Proc. MUS*, Vol. 2008, No. 78 (2008-MUS-076), pp. 125–130, 2008.

[11] Naoko Kurokawa, Hiroaki Saito, "Chord Estimation by Combination of chroma Vector and Spectrum Dip," *The Institute of Electronics, Information and Communication Engineers and Information Processing Society of Japan*, Vol. 11, No. 2, pp. 165–170, 2012.

[12] Stanisław Raczyński, Nobutaka Ono, and Shigeki Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," *Proc. ISMIR 2007, 8th International Conference on Music Information Retrieval*. Citeseer, 2007.

[13] Satoshi Maruo, Kazuyoshi Yoshii, Katsutoshi Itoyama, Matthias Mauch and Masataka Goto, "Chord Recognition based on Approximate Note Transcription Using NMF with Chord Constraints," *IPSJ-MUS*, Vol. 2015, No. 1, pp. 421–422, 2015.

[14] Siddharth Sigtia, Nicolas Boulanger-Lewandowski, and Simon Dixon, "Audio chord recognition with a hybrid recurrent neural network," *Proc. ISMIR 2015*, pp. 127–133, 2015.

[15] Filip Korzeniowski and Gerhard Widmer, "A Fully Convolutional Deep Auditory Model for Musical Chord Recognition," *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing, Salerno, Italy*, September 2016.

[16] Kaneko, Hitomi and Kawakami, Daisuke and Sagayama, Shigeki, "Functional harmony annotation database for statistical music analysis," *Proc. the International Society for Music Information Retrieval Conference (ISMIR): Late Breaking session*, 2010.