Automatic music transcription using Non-negative matrix factorization

Sang Ha Park, Seokjin Lee and Koeng-Mo Sung

Applied Acoustics Lab., INMC, School of Electrical Engineering, Seoul National University, Republic of Korea

PACS: 43.75.ZZ

ABSTRACT

This paper proposes an effective method for the automatic music transcription in polyphonic music. The method consists of a combination of Non-negative Matrix Factorization (NMF), subharmonic summation method and onset detection algorithm. We decompose the magnitude spectrum of a music signal into the spectral component and the temporal information of every note using NMF. Then, the accurate pitch of each note is calculated from the decomposed frequency components based on the subharmonic summation method. And an algorithm for detecting the onset is applied for estimating the temporal information of a musical note. Our method is simple and has a low computational cost, because the method is not a note training-based. The previous researches using NMF detect the pitch and the time duration 'manually', therefore the previous methods are difficult to use in the real engineering. Our proposed method improved this problem with 'automatically' detecting the fundamental frequency and the rhythm component. Furthermore, the proposed method automatically performed the indexing of the musical notes which is useful in the real engineering field. The transcription performance is evaluated with recorded polyphonic music signals, and the performance of the proposed method is better than the conventional NMF based methods in estimating both frequency component and time duration information.

1. INTRODUCTION

Music transcription is extracting the pitch components and the rhythm components from an audio source and then notating the music on the paper. Because only musically well-trained people can transcribe music, transcription of recorded music was done manually. Therefore, many researchers began a study of an automatic music transcription, various approaches have been proposed based on computational auditory models or probabilistic signal models [1]. These many knowledge-based approaches need high cost and are complex.

Recently, transcription methods using adaptive signal models such as Independent Component Analysis (ICA), Nonnegative Matrix Factorization (NMF) and Sparse Coding are suggested and performed well. Among the methods, NMF based music transcription is very efficient [2]. In this paper, we proposed an automatic music transcription algorithm based on NMF. We used subharmonic summation method and onset detection algorithm for the better performance than conventional NMF transcription methods.

2. PROPOSED METHOD

2.1. NMF And NNSC

Non-negative Matrix Factorization (NMF) is a matrix decomposition method proposed by Lee and Seung [3]. The goal of this algorithm is to decompose a given non-negative $M \times N$ matrix V into two low order matrices :

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu} \tag{1}$$

where W is basis matrix and H is coefficient matrix. Both of the matrix are non-negative and the dimensions of the matrix W and H are $M \times R$ and $R \times N$ respectively. The rank R is generally chosen so that $(M+N)R < MN$, and the product WH can be regarded as a compressed form of matrix V.

To find a pair of W and H which minimizes the error of reconstruction, we use the Kullback-Leibler divergence [4] [2].

$$D = \left\| X \otimes \ln\left(\frac{X}{WH}\right) - X + WH \right\|_F \tag{2}$$

where $\| \cdot \|_F$ is the Frobenius norm, $\otimes$ is the Hadamard product (an element-wise multiplication of the matrices) and the division is elementwise.

For the perfect decomposition, the rank R of the matrix W and H has to be selected to a proper value. But if the distribution of H is sparse, input matrix V can be represented by a small number of weighting matrix. In this case, Non-negative Sparse Coding (NNSC), a sparseness term is added to the objective function of NMF, is useful [5]. The cost function of NNSC is modified as eq. (3).

$$C(W,H)=\frac{1}{2}\|V-WH\|^2+\lambda\sum_{ij} f(H_{ij}) \qquad (3)$$

where the parameter λ constrolls the tradeoff between the sparseness and the accurate reconstruction. When the rank R is highly selected, NNSC decompose the residual values of the matrix W and H into low energy noise. Therefore, it is useful when we don't know the exact value of R. However, the drawback of this method is that an appropriate value of sparseness parameter λ is needed.

The matrix W and H are iteratively updated until the objective function in eq. (3) converges to the minimum value. The multiplicative update rule is as follows :

$$\begin{aligned} H_{am} &\leftarrow H_{am}\frac{\sum_i W_{ia}X_{im}/(W\cdot H)_{im}}{\sum_k W_{ka}} \\ W_{ia} &\leftarrow W_{ia}\frac{\sum_m H_{am}X_{im}/(W\cdot H)_{im}}{\sum_v H_{av}} \end{aligned} \qquad (4)$$

Using the iteration update equation, the matrix V is finally decomposed into W and H .

2.2. Onset detection

In music transcription, the order of R means the number of musical notes (fundamental frequencies) used in a music score. The value of R should be given for performing the NMF or NNSC. To estimate the range of R, we used the number of note onset. It is also used in minimizing the error of the time composent in the NMF result.

Note onset detection algorithm based on the signal features can be divided into two groups : methods using the temporal features and methods using the spectral features [6]. The temporal features based methods include “envelope follower” which detects the increase of the signal's amplitude by rectifying and smoothing the signal.

$$E_0(n)=\frac{1}{N}\sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1}|x(n+m)|w(m) \qquad (5)$$

where $x(n)$ is time domain signal, $w(m)$ is an N-point window centered at $m=0$.

The second temporal features based method is modified method of “envelope follower” which finds the maximum in local energy :

$$E(n)=\frac{1}{N}\sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1}|x(n+m)|^2 w(m) \qquad (6)$$

However, these methods using the signal envelope or signal local energy can't detect the onset of low energy sound and the computational cost is high. So, we used the spectral feature based methods.

The firtst algorithm based spectral feature is the weighted spectrum energy method. After calculating the energy of the frequency component in the short-time Fourier transformed (STFT) singal, the high frequency components are weighted.

$$WE(n)=\frac{1}{N}\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1}W_k\,|X_k(n)|^2 \qquad (7)$$

where X_k is STFT of the signal and W_k is frequency dependent weighting. Because the energy of the signal concentrated at low frequencies, the performance degradation of onset detection algorithm about the high frequency component usually occur. The wighting W_k prevent this problem, and we use linear weighting function $W_k=|k|$ in this paper.

Another spectral based method is using the “spectral difference” which is detecting function as a distance between spectra. This method finds the time of rapid change in frequcncy values was arisen.

$$SD(n)=\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1}\{H(|X_k(n)|-|X_k(n-1)|)\}^2 \qquad (8)$$

where $H(x+|x|)/2$, i.e., zeros for negative values.

Among four onset detection algorithms, the performance of the “spectral difference” was the best. Therefore, this algorithm was used when detecting the note onset in our music transcription method.

2.3. Subharmonic summation

When the sound source is decomposed with NMF or NNSC, the decomposed matrix W and H contains frequency component and time component respectively. But the matrix W is constructed with not only the fundamental frequencies i.e. the pitch of the note used in the score, but also harmonic frequencies. To find the fundamental frequencies in the elements of matrix W , we used the subharmonic summation method [7].

The subharmonic summation is pitch estimation method which is calculating the fundmental frequency by grouping the harmonic frequencies to the corresponding fundamental frequencies. The process is as follows.

1. The Short Time Fourier Transform (STFT) is applied to the signal. Then, the range of the frequencies is restricted below 1250Hz for the pitch determination process.

2. Find the peak values of the amplitude spectrum below 1250Hz, and the frequencies around the peaks i.e. within 20Hz from the peaks are zero padded.

3. The resultant spectrum is smoothed using a Hanning filter.

$$B_k(n)=\frac{1}{4}A_{k-1}(n)+\frac{1}{2}A_k(n)+\frac{1}{4}A_{k+1}(n) \qquad (9)$$

where $A_k(n)$ and $B_k(n)$ are the unsmoothed spectrum and the smoothed spectrum respectively, n is time index and k is frequency index. The smoothed spectrum $B_k(n)$ represents the peak enhanced spectrum.

4. The spectral window $W_k(n)$, representing the auditory sensitivity is multiplied to the smoothed spectrum $B_k(n)$.

$$P_k(n) = W_k(n) B_k(n) \tag{10}$$

A rased arc-tangent function is used for $W_k(n)$.

5. The low frequency weighting factor h_i is applied to the $P_k(n)$ because the higher harmonic have little affect to the pitch determination. Then, the weighted frequency values are summed.

$$H_k(n) = \sum_{i=1}^{N} h_i P_i(n) \tag{11}$$

where i is the compression factor, and $h_i = 0.84^{n-1}$ is used experimentally.

The function $H(n)$ is the subharmonic sum spectrum and is representing the added subharmonics.

3. MUSIC TRANSCRIPTION ALGORITHM

The proposed algorithm using NMF (or NNSC), subharmonic summation and onset detection algorithm is as follows. First of all, we applied STFT to the audio signal. In some music transcription algorithms, Equivalent Rectangular Bandwidth (ERB) based time-frequency representation which gives an approximation to the bandwidths of the filters in human hearing is used instead of STFT. But we used STFT in this paper because the performance using STFT and ERB representation are similar in transcription [1]. Then, the magnitude spectrum of the signal is composed with the frequencies and its magnitudes according to the time. Next, by performing NMF, the magnitude spectrum is decomposed into two matrices which contain frequency components and time components respectively.

The music transcription with NMF is simple, low computational cost and high performance. However, the pitch and rhythm components are not exactly extracted from the decomposed matrices because some errors are exist in the matrix W and H. And we have no information about the number of notes used in the score. Therefore, we have to expect the rank of R, and postprocessing after NMF is needed.

For the better performance, we performed the onset detection algorithm to the magnitude spectrum. Then, we can get the total number of the note onsets. We put the rank of R equals to the total number of the note onsets and put the maximum threshold to 70. And the onset result is used for eliminating the error of time component matrix H at the end. With this R, NNSC is performed and frequency component matrix W and time component matrix H are acquired. After normalize the values of matrix W, threshold is adjusted for the error elimination. Then, with the subharmonic summation method the pitch values are calculated. By comparing the pitch values with equal temperament scale, the note codes are detected and sorted in ascending order.

And from the matrix H, the peak values are detected and the threshold is adjusted for the error elimination as in matrix W. Then the onset times of W are compared with the total onset times which are obtained before. By elminating the error of W, the rhythm components can be acquired.

The block diagram of our proposed algorithm is as following Fig. 1.

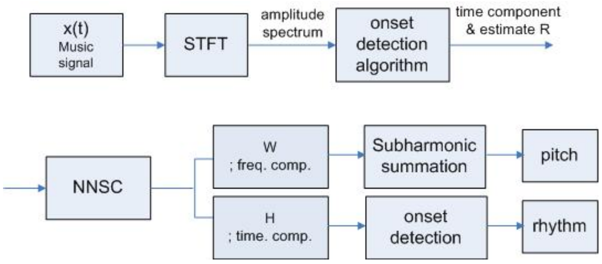

Figure 1. The block diagram of music transcription system

4. EVALUATION

We performed the music transcription with the recorded piano music. If the information about the exact number of notes except overlapping the same pitch notes is given, the music transcription is performed by NMF with excact R very successfully. The errors are reduced by subharmonic summation and onset detection steps.

We performed the transcription with proposed method (NMF + subharmonic summation + onset detection algorithm) and compared the performance with conventional method (only NMF). The test data consists of recorded piano sources with 10 s duration each. The sound sources are from a real piano recording of ‘J. S. Bach’s The Well-Tempered Clavier’.

Table 1. Piano transcription result

	Total note number	Conventional method	Proposed method
Music 1	21	19	19
Music 2	45	39	43
Music 3	73	69	73
Music 4	4	2	4
Music 5	24	22	22
Music 6	10	10	10
Total	177	161 (90.9%)	171 (96.6%)

In the 6 test data, total 177 events are occurred. Total 161 and 171 notes are successfully transcribed with conventional method and proposed method respectively. The performance is improved about 6% .

And then, we also performed the transcription with NNSC, subharmonic summation and onset detection algorithm. Fig. 2 is a part of the music score ‘J. S. Bach’s The Well-Tempered Clavier, Book 1 No. 6 in d minor, BWV 875 :

Fuga' which is a score of sound source used in NNSC transcription.

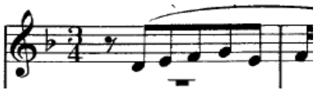

Figure 2. The Music score

Total 6 events are occurred and 4 notes except the repetition notes are used in the score. Among 4 kinds of onset algorithms "spectral difference" is used for the onset detection and 6 events are successfully detected.

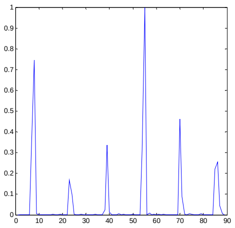

Figure 3. The result of onset detection

Then, the proposed algorithm (NNSC + subharmonic summation + onset detection algorithm) and the conventional algorithm are implemented with the predicted R from onset detection. The result is as following.

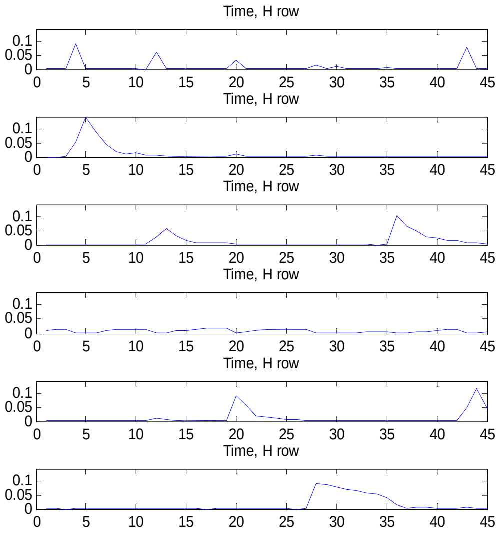

Figure 4. The result of conventional method

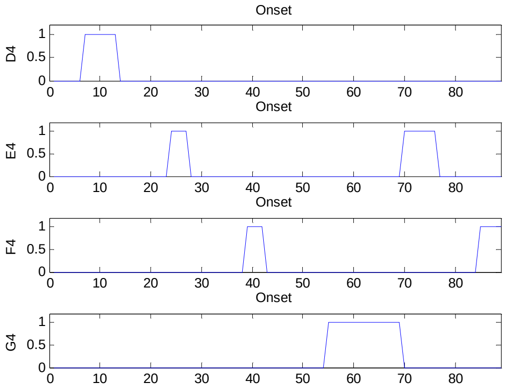

Figure 5. The result of proposed method

With conventional method as shown in Fig. 4, 2nd, 3rd, 5th, 6th rows are correctly decomposed but 1st and 4th rows are mistaken results because the rank R is overestimated. So when we use NNSC with no information about R, the error correction is needed and is done manually. However, in our system the errors are removed automatically in a series of process of subharmonic summation, onset detection and the thresholding in W and H. It can be confirmed in fig. 5 where the wrong result rows are eliminated.

With proposed method in both NMF and NNSC, the transcription results are improved than conventional method.

5. CONCLUSION

We proposed an effective method for the automatic music transcription based on NMF. The subharmonic summation method and onset detection algorithm are used for pitch and onset detection. The performance with the proposed algorithm is better than NMF or NNSC only method. In the futere, we will investigate the offset detection for the exact rhythm detection. And we will improve the prediction of R because it is roughly choosed in present algorithm.

REFERENCES

1. E. Vincent, N. Bertin and R. Badeau, "Two nonnegative matrix factorization methods for polyphonic pitch transcription" in *Proc. Music Information Retrieval Evaluation eXchange (MIREX),* (2007)

2. P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription" in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* 177-180 (2003)

3. D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization" *Nature*, **401**, 788-791 (1999)

4. D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization" in *Advances in Neural Information Processing 13 (Proc. NIPS*2000),* MIT Press (2001)

5. P. Hoyer, "Non-negative sparse coding" in *Neural Networks for Signal Processing,* 557-565 (2002)

6. J. P. Bello et. al., "A tutorial on onset detection in music signals" *IEEE Transactions on Speech and Audio Processing*, **13**, 1035-1047 (2005)

7. D. Hermes, "Measurement of pitch by subharmonic summation" *J. Acoust. Soc. Am.* **83**, 257-264 (1988)