CONSTRAINED SPECTRUM GENERATION USING A PROBABILISTIC SPECTRUM ENVELOPE FOR MIXED MUSIC ANALYSIS

Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki
Department of Computer Science and Systems Engineering, Kobe University, Japan
nakashika@me.cs.scitec.kobe-u.ac.jp, {takigu,ariki}@kobe-u.ac.jp

ABSTRACT

NMF (Non-negative Matrix Factorization) has been one of the most widely-used techniques for musical signal analysis in recent years. In particular, the supervised type of NMF is garnering much attention in source separation with respect to the analysis accuracy and speed. In this approach, a large number of spectral samples is used for analyzing a signal. If the system has a minimal number of samples, the accuracy deteriorates. Because such methods require all the possible samples for the analysis, it is hard to build a practical analysis system. To analyze signals properly even when short of samples, we propose a novel method that combines a supervised NMF and probabilistic search algorithms. In this approach, it is assumed that each instrumental category has a model-invariant feature called a probabilistic spectrum envelope (PSE). The algorithm starts with learning the PSEs of each category using a technique based on Gaussian Process Regression. Using the PSEs for spectrum generation, an observed spectrum is analyzed under the framework of a supervised NMF. The optimum spectrum can be searched by Genetic Algorithm using sparseness and density constraints.

1. INTRODUCTION

Mixed music analysis (estimating the pitch and instrument labels of each musical note from a single-channel polyphonic music signal with multiple instruments) has been recognized as one of the most challenging tasks in musical signal processing. To achieve this, many approaches have been proposed so far: ICA-based methods [1, 3], HTTC (Harmonic-Temporal-Timbral Clustering) [6], Instrogram [5], etc. Of all these techniques, the methods based on NMF (Non-negative Matrix Factorization) have attracted considerable attention lately as a way to analyze signals more effectively and more easily. In many of these techniques, an observed spectrogram matrix can be represented as a linear combination of two matrices: a basis matrix whose columns roughly indicate spectrums of each musical source with various pitches and instruments, and an activity matrix which shows temporal information of each basis vector.

NMF-based analysis methods are broadly divided into two categories: an unsupervised approach [4] and a supervised approach [2]. Since the former approach decomposes the spectrogram without the assumption of the spectral structures of audio sources, the unintended basis matrix and activity matrix will be obtained. Therefore, it is hard to analyze mixed-source audio correctly using an unsupervised approach.

On the other hand, a supervised approach decomposes a mixed musical signal using the spectral templates of each musical source, which are learned beforehand. Compared to an unsupervised approach, this technique tends to produce preferable results in terms of analysis speed and accuracy. However, if *unlearned* sounds are contained in the test signal, the accuracy may deteriorate because there are many different types (models) of instrument that belong to the same instrumental category. For example, the "Piano" category includes different models: "Piano1","Piano2", and so on. To improve the decomposition accuracy, many kinds of spectral templates (not only different categories but different models in the categories) should be trained. However, this is extremely difficult to build into a real system.

To solve this problem, we propose a novel method of mixed music analysis, which uses a model-invariant feature (probabilistic spectrum envelope; PSE) of each category. This feature is derived from the following idea. An instrument's spectrum can differ slightly due to various factors associated not only with the type of instrument (model) but also the manufacturer, the materials used, the temperature, humidity, and playing-style, etc. However, the way the spectrum fluctuates is not completely random, as it depends on the instrument's category. Therefore, we introduce the PSE feature that does not depend on the pitch, the model, the material, and other various factors. This is similar to a spectrum envelope feature, which does not depend on the pitch. The feature is defined as a set of the *mean* spec-

trum envelope and *variance* spectrum envelope in the time-frequency domain as shown in Figure 1 (a). Once the PSE is estimated, any spectrum belonging to the category can be obtained by multiplying various comb filters and randomly-generated spectrum envelopes from the PSE.

Figure 1 shows a system flowchart of mixed music analysis under the PSE framework. In our approach, unsupervised NMF and extended Gaussian Process (SPGP+HS [7]) are employed to estimate the PSE features of each category on the training stage. At the analysis stage, we use supervised NMF for the analysis, in which an optimum basis vector can be searched using a Genetic Algorithm with sparseness and density constraints.

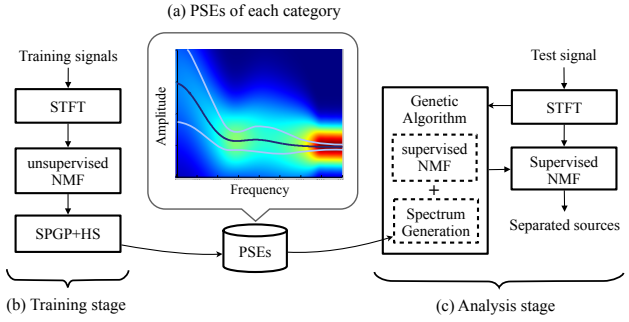

Figure 1. Flowchart of mixed music analysis using propabilistic spectrum envelope (PSE). The red and blue color indicate the large and small values of probability, respectively. The black and white lines are the mean envelope, and mean plus/minus variance envelope.

2. PSE ESTIMATION

2.1 Spectral peaks extraction

The probabilistic spectrum envelope (PSE) of each category is estimated by SPGP+HS regression [7] in this paper. In this section, we will discuss the way spectral peaks (input samples used for the regression) are obtained.

First, we prepare some acoustic signals, each of which contains only the needed musical sources of the instrumental category. The various sources do not sound at the same time. In this paper, 12 half-tone sources sound in sequence every octave. Employing NMF to the amplitude spectrogram $\mathbf{V}$ ($\in \mathbb{R}^{F\times T}$) of the signal, $\mathbf{V}$ is approximately decomposed into the product of a basis matrix $\mathbf{W}$ ($\in \mathbb{R}^{F\times R}$) and an activity matrix $\mathbf{H}$ ($\in \mathbb{R}^{R\times T}$) as follows:

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

$$\forall i,j,k, \mathbf{W}_{ij} \geq 0,\ \mathbf{H}_{jk} \geq 0 \tag{2}$$

where F, T and R are the numbers of bins of frequency, time and bases, respectively (here, $R = 12$).

$\mathbf{W}$ and $\mathbf{H}$ can be obtained by iteratively calculating update rules based on Euclidean divergence. The update rules for each matrix element are:

$$\mathbf{W}_{ij} \leftarrow \mathbf{W}_{ij}\frac{(\mathbf{VH}^T)_{ij}}{(\mathbf{WHH}^T)_{ij}} \tag{3}$$

$$\mathbf{H}_{jk} \leftarrow \mathbf{H}_{jk}\frac{(\mathbf{W}^T\mathbf{V})_{jk}}{(\mathbf{W}^T\mathbf{WH})_{jk}}. \tag{4}$$

From the updated matrix $\mathbf{W}$, a set of N spectral peaks $\mathbb{P} = (\boldsymbol{f}, \boldsymbol{y}) = \{(f_n, y_n)\}_n$ are exploited, where f_n and y_n are frequency and amplitude of the n-th peak, respectively. These peaks are found by searching for the harmonic peaks of each basis vector.

2.2 PSE estimation using SPGP+HS

In this paper, the PSE of each category can be estimated by extended Gaussian Process (SPGP+HS [7]), which can approximate the shape of any function with varying variance more accurately than the standard Gaussian Process.

By giving a set of peaks, $\mathbb{P}$, to one-dimensional SPGP+HS, we obtain PSE mean envelope μ_f and PSE variance envelope σ_f, as follows:

$$\mu_f = \mathbf{K}_{ff_m}\mathbf{Q}\mathbf{K}_{f_mf_n}\mathbf{\Lambda}^{-1}\boldsymbol{y} \tag{5}$$

$$\sigma_f = \mathbf{K}_{ff} - \mathbf{K}_{ff_m}(\mathbf{K}_{f_mf_m}^{-1} - \mathbf{Q})\mathbf{K}_{ff_m}^T \tag{6}$$

where, $\mathbf{Q} = \left(\mathbf{K}_{f_mf_m} + \mathbf{K}_{f_mf_n}\mathbf{\Lambda}^{-1}\mathbf{K}_{f_mf_n}^T\right)^{-1}$ and $\mathbf{\Lambda} = diag(\mathbf{K}_{f_nf_n} - \mathbf{K}_{f_mf_n}^T\mathbf{K}_{f_mf_m}^{-1}\mathbf{K}_{f_mf_n})$. $\mathbf{K}_{ab}$ is a gram matrix between a and b with a parameter θ. Pseudo-inputs $\bar{\boldsymbol{f}} = \{\bar{f}_m\}_{m=1}^M$ indicate the representatives of any inputs $\boldsymbol{f}$, satisfied $M \ll N$. $h_m \in \boldsymbol{h}$ denotes an uncertainty parameter to the pseudo-input $\bar{f}_m$. We can find the optimum parameters $\boldsymbol{h}, \theta, \bar{\boldsymbol{f}}$ based on a gradient-based method (for more details, see [7]).

3. ANALYSIS METHOD

3.1 Spectrums generation based on PSE

The spectrum envelope $e^c(f)$ based on the PSE of category c is randomly generated as follows:

$$e^c(f) \sim \mathcal{N}(\mu_f^c, \sigma_f^c). \tag{7}$$

$\mathcal{N}(\mu, \sigma)$ shows the normal distribution of mean μ and variance σ.

Spectrum $p(f)$, with a fundamental frequency f_0 along the envelope, $e^c(f)$ can be specifically calculated in Eq. (8).

$$p(f) = \max\left(e^c(f), 0\right) \cdot \Psi(f; f_0) \tag{8}$$

The reason for the maximum expression in Eq. (8) is that a spectrum cannot have negative values. $\Psi(f; f_0)$ is a comb filter with a fundamental frequency f_0, calculated as:

$$\Psi(f; f_0) = \sum_l \exp\left\{-\frac{(f - f_0 \cdot l)^2}{2\nu^2}\right\} \tag{9}$$

where l is the index of Gaussian components, and ν is a hyper-parameter to determine the kurtosis of each component.

Using the above procedure, we can obtain an intended basis matrix $\tilde{\mathbf{W}}$ whose columns (spectrums) are randomly generated for various categories and fundamental frequencies.

3.2 Basis matrix optimization using Genetic Algorithm

What we want to do in the analysis stage is to find the optimum NMF matrices $\hat{\mathbf{W}}$ and $\hat{\mathbf{H}}$ for a given test signal. To do this, we introduce an optimization method based on genetic algorithm (GA), which is a method for finding the optimum by repeating natural-evolution-inspired techniques: selection, crossover, mutation and inheritance.

Given an amplitude spectrogram $\mathbf{X}$ of a test signal and a randomly-generated basis matrix $\tilde{\mathbf{W}}$, the activity matrix $\mathbf{H}$ can be calculated by applying supervised NMF with $\tilde{\mathbf{W}}$. That is, each element of $\mathbf{H}$ is repeatedly updated by Eq. (4) while keeping $\tilde{\mathbf{W}}$ fixed. Since $\tilde{\mathbf{W}}$ determines $\mathbf{H}$ in this calculation, $\mathbf{H}$ can be considered as a function of $\tilde{\mathbf{W}}$. If $\tilde{\mathbf{W}}$ has better (more suited) spectral columns for the test signal, the distance between $\mathbf{X}$ and $\tilde{\mathbf{W}}\mathbf{H}$ must become smaller. Therefore, the minimization of Euclidean distance $D_{EUC}(\mathbf{X}, \tilde{\mathbf{W}}\mathbf{H})$ can be used as a criterion for finding the candidate $\tilde{\mathbf{W}}$. In addition to the distance criterion, we give two constraints $sp(\mathbf{H})$ and $den(\mathbf{H})$. The former $sp(\mathbf{H})$ leads the matrix $\mathbf{H}$ to be sparse, which is

$$sp(\mathbf{H}) = \frac{\#\{(j,k)|\mathbf{H}_{jk} \leq \epsilon\}}{R \times T} \tag{10}$$

where, $\epsilon(\geq 0)$ is a small value (in our experiments, $\epsilon = 0.1$).

The other constraint $den(\mathbf{H})$ represents the "density" of the elements in $\mathbf{H}$. This idea is inspired by the fact that musical notes of each instrument tend to group together in regard to time and tone. We define the constraint $den(\mathbf{H})$ as:

$$den(\mathbf{H}) = \frac{\sum_{k,l,l'} \exp\left\{-\frac{(s_{k,l}-s_{k+1,l'})^2}{2\rho^2}\right\}}{\sum_k N_k} \tag{11}$$

$$\{s_{k,l}\}_{l=1}^{N_k} = \{\, j \,|\mathbf{H}_{jk} \geq \epsilon\} \tag{12}$$

where ρ is a constant factor for determing the allowance for distant tones (in our test, $\rho = 3$).

Finally, we set the criteria for the optimum search of the candidate $\tilde{\mathbf{W}}$ as follows:

$$\Theta(\tilde{\mathbf{W}}) = D_{EUC}(\mathbf{X}, \tilde{\mathbf{W}}\mathbf{H}) - \alpha \cdot sp(\mathbf{H}) - \beta \cdot den(\mathbf{H}) \tag{13}$$

where, α (≥ 0) and β (≥ 0) are weight parameters that reflect the effects of sparseness and density constraints, respectively.

In our analysis method, the optimum basis matrix $\hat{\mathbf{W}}$ is obtained using GA to minimize the objective function (13). The first step of GA is to generate U $(= 12$, in our tests) basis matrices $\{\tilde{\mathbf{W}}_u\}_{u=1}^{U}$ from pre-trained PSEs (See 3.1.), and evaluate the objective function for each matrix by Eq. (13). Note that fundamental frequency of each column in the u-th basis matrix $\tilde{\mathbf{W}}_u$ is different from the others, but the fundamental frequency of the l-th column for all basis matrices has the same fundamental frequency. To update the whole set, the following process is repeated G $(= 100$, in this paper) times:

1. Copy the best (smallest-objective) basis matrix of the previous generation to the current generation.
2. With a probability p_{cross}, exchange two selected basis matrices according to the uniform crossover.
3. With a probability p_{mut}, mutate a selected basis matrix based on PSE.
4. Repeat step 2 and 3 until the number of basis matrices of the current generation reaches L.

Concerning the expression "select" above, the probability of u-th candidate selection is defined as $\frac{\Theta(\tilde{\mathbf{W}}_u, \tilde{\mathbf{H}}_u)}{\sum_{u=1}^{U} \Theta(\tilde{\mathbf{W}}_u, \tilde{\mathbf{H}}_u)}$. This shows that the *better* $\tilde{\mathbf{W}}_u$ tends to be selected more. p_{cross} and p_{mut} in steps 2 and 3 are respectively the probabilities of crossover and mutation, which satisfy $p_{cross} + p_{mut} = 1$ (in this paper, $p_{cross} = 0.9, p_{mut} = 0.1$). Furthermore, our GA has the constraints that each basis matrix mutates without altering the fundamental frequencies. In other words, the mutated new vector is calculated by multiplying the randomly-generated spectrum envelope from PSE by the comb filter that has the same fundamental frequency as the original one. Therefore, basis matrices of each generation can be generated without changing the information on the fundamental frequency and category we set at first.

The final analysis result is the optimum NMF matrices $\hat{\mathbf{W}}$ and $\hat{\mathbf{H}}$, which are the best matrices in G-th generation ($\hat{\mathbf{H}}$ is obtained by supervised NMF with the optimum basis matrix $\hat{\mathbf{W}}$). Because $\hat{\mathbf{W}}$ contains a category index c, a test signal can be decomposed into each instrument.

4. EXPERIMENTS

To evaluate our proposed method, "wav-to-mid" tests were conducted. In these experiments, an acoustic data synthesized with MIDI sounds is automatically converted into MIDI format. A part of "RWC-MDB-C-2001 No. 43: Sicilienne op.78" from RWC Music Database[1] was used for the test (Figure 3 (a)). The monaural test signal was recorded at a 16 kHz sampling rate using multiple MIDI instruments: Piano and Flute (exactly, "Piano1" and "Flute1" instrumental

[1] http://staff.aist.go.jp/m.goto/RWC-MDB/

models of MIDI, respectively). Before the test, PSEs for the two categories were trained using the different sounds from the test signal ("Piano2" for "Piano" PSE and "Flute2" for "Flute" PSE). Using the PSEs, GA found the optimum matrices $\hat{\mathbf{W}}$ and $\hat{\mathbf{H}}$. By binarizing $\hat{\mathbf{H}}$ with an adequate threshold, we obtained the final results of MIDI format. The results were compared for the cases in which the objective function of GA has sparseness and density constraints and when it does not ("sp+den", "sp","den","w/o"). Since the results depend on the initial values of $\{\tilde{\mathbf{W}}_u\}_{u=1}^{U}$, we repeated each method by 100 times and computed the mean, maximum, and minimum values of accuracy. We also compared the results with the conventional method, supervised NMF ("s-NMF" given the basis matrix of "Piano2", "ideal" given that of "Piano1").

Figure 2 illustrates MIDI-conversion accuracies for each method. The accuracy is calculated as $\frac{N_{all}-(N_{ins}+N_{del})}{N_{all}} \times 100$, where N_{all}, N_{ins} and N_{del} mean the total number of notes, insertion errors, and deletion errors, respectively. Because onset time and the duration of each sound source are not necessarily correct in the above binarizing process, we permitted the duration to differ and the onset time to shift τ seconds (in this paper, $\tau = 0.3$). The bar values of our methods in the figure are average accuracies for 100 tries, and the error bars indicate maximum and minimum values of the tries. Concerning the results of conventional methods, if the system knows exactly the same sounds as the test signal, it yields high performance (ideal). However, if the system does not know, the accuracy deteriorates dramatically (s-NMF). Meanwhile, each of our approaches maintains high accuracy even when the system does not learn the sounds of the test data. The preferable results are due to the fact that each PSE can be estimated by only various pitches, and it can cover spectrum envelopes of unknown models. Comparing within our approaches, the system with sparseness or density constraints achives better accuracy, and when both constraints were added ("sp+den"), for the tests with the best results, there were cases when the accuracy even exceeded the ideal value.

An analysis example of "sp+den" tries is shown in Figure 3 (b). Almost all the notes were estimated correctly, but parts of them were mistaken as octave-different notes. Therefore, we will improve the accuracy by adding other constraints to avoid octave differences in the future.

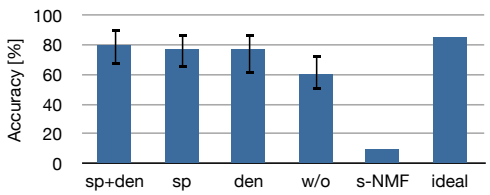

Figure 2. Accuracy rates of each method.

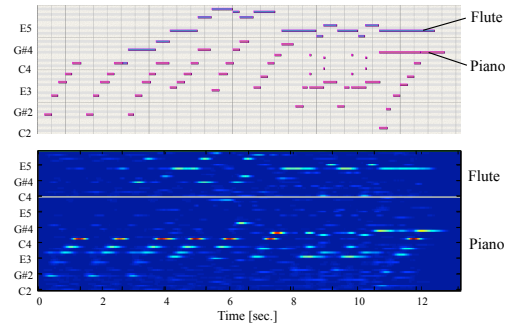

Figure 3. (above) Piano-roll representation of test MIDI data. The red and purple parts indicate piano and violin tones, respectively. (below) An example of analysis results with sparseness and density constraints.

5. CONCLUSIONS

In this paper, we proposed an algorithm for monaural sound source decomposition and multiple-pitch estimation. The method categorizes several spectrum envelopes for each musical category, inspired by invariance of spectral fluctuation in a category. This categorized envelope, called the probabilistic spectrum envelope (PSE), has a characteristic of being able to absorb differences between models, pitches, manufactures, playing-style, and so on. PSE consists of a mean envelope and variance envelope which can be simultaneously estimated by SPGP+HS regression as described in this paper. In the analysis stage, Genetic Algorithm (GA) with supervised-NMF-based objective and sparseness/density constraints was employed for an optimum search in all the spectrum envelopes that can be generated from the PSE.

The simulation experiments using MIDI sources show that the proposed method is robust to instrumental model changes. Since the results depend on the initial values, however, future research will include designing a directly optimum search method, such as ML (Maximum likelihood) or MAP (Maximum a posteriori) estimations.

6. REFERENCES

[1] M.A. Casey and A. Westner: In *Proceedings of the International Computer Music Conference*, pp. 154–161, 2000.

[2] A. Cont, S. Dubnov, and D. Wessel: In *Proceedings of Digital Audio Effects Conference (DAFx)*, pp. 10–12, 2007.

[3] Gil jin Jang and Te won Lee: *Journal of Machine Learning Research*, pp. 1365–1392, 2003.

[4] M. Kim and S. Choi: *Independent Component Analysis and Blind Signal Separation*, pp 617–624, 2006.

[5] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno: *Information and Media Technologies*, pp. 279–291, 2007.

[6] K. Miyamoto, H. Kameoka, T. Nishimoto, N. Ono, and S. Sagayama: *ICASSP 2008*, pp. 113–116, 2008.

[7] E. Snelson and Z. Ghahramani: In *Proceedings of the 22nd Uncertainty in Artificial Intelligence*, 2006.