Probabilistic Spectrum Envelope: Categorized Audio-features Representation for NMF-based Sound Decomposition

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 useful techniques for audio signal analysis in recent years. In particular, supervised NMF, in which a large number of samples is used for analyzing a signal, is garnering much attention in sound source separation or noise reduction research. However, because such methods require all the possible samples for the analysis, it is hard to build a practical system based on this method. In this paper, we propose a novel method of signal analysis that combines the NMF and probabilistic approaches. In this approach, it is assumed that each audio-source category (such as phonemes or musical instruments) has an environment-invariant feature, called a probabilistic spectrum envelope (PSE). At the start, the PSE of each category is learned using a technique based on Gaussian Process Regression. Then, the observed spectrum is analyzed using a combination of supervised NMF and Genetic Algorithm with pre-trained PSEs.

Index Terms: signal analysis, source separation, non-negative matrix factorization, probabilistic spectrum envelope, Gaussian process, genetic algorithm

1. Introduction

Source separation from a single-channel signal has been recognized as a challenging task in signal processing. To achieve this, many approaches have been proposed so far; for example, a method based on factorial HMM [1], an ICA-based method [2], 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. Many of these techniques adapt a NMF algorithm to the decomposition of the observed spectrogram matrix into two matrices. One is a basis matrix, whose rows roughly indicate spectrums corresponding to each acoustical event (phonemes in a speech signal, or musical tones in a musical signal, etc. In this paper, we focus on musical signal analysis.) The other is called an activity matrix, which shows temporal information of each basis vector.

The analyzing methods based on NMF are broadly divided into two categories: an unsupervised approach [3, 4, 5] and a supervised approach [6, 7, 8]. Because 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 the mixed signal using the spectral templates of each acoustical event, which are learned beforehand. Compared to an unsupervised approach, this technique tends to produce preferable results in terms of analysis speed and the accuracy. However, if

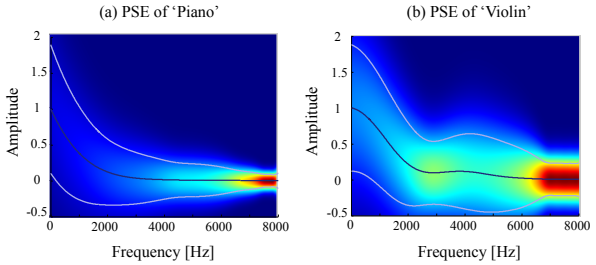

Figure 1: Examples of probabilistic spectrum envelope; the left is Piano and the right is Violin. The red and blue color in the figure indicate the large and small values of probability, respectively. The black line is the mean envelope, and the white lines are the mean envelope plus and minus variance envelope.

unlearned sounds are contained in the test signal, the accuracy may deteriorate because there are very many models that belong to the same category of musical instruments. For example, the "Piano" category includes different models made: "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 audio analysis, which uses the model-invariant features (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 the spectrum envelope, which does not depend on the pitch. The feature is defined as a set of the *mean* spectrum envelope and *variance* spectrum envelope in the time-frequency domain as shown in Figure 1. Once the PSE is estimated, any spectrum belonging to the category can be obtained by multiplying a set of comb filters and randomly-generated spectrum envelopes from the PSE.

Figure 2 shows a system flowchart of mixed sound analysis using PSE representation. In our approach, unsupervised NMF and extended Gaussian Process (SPGP+HS [9]) are employed to estimate the PSE features of each category on the training stage. When analyzing a test signal, we use semi-supervised

NMF with the basis matrix changed to fit the signal using Genetic Algorithm based on the pre-trained PSE.

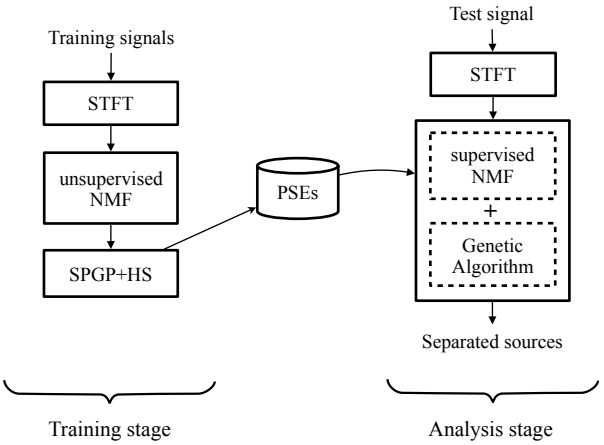

Figure 2: Flowchart of proposed method. Modeling of probabilistic spectrum envelopes and analyzing mixed music signals using the envelopes.

2. Estimating PSE

2.1. Spectral peaks extraction

The probabilistic spectrum envelope (PSE) of each category is estimated by SPGP+HS regression [9] 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, 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. These peaks are found by searching for the harmonic peaks of each basis vector.

2.2. PSE estimation using SPGP+HS

PSE is defined as a spectrum envelope with variance values. Therefore, we employed extended GP (SPGP+HS [9]), which can approximate the shape of any function with varying variance, for the estimation of the PSE.

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 $\boldsymbol{h}, \theta, \bar{\boldsymbol{f}}$ by using a gradient-based method (for more details, see [9]).

3. Analyzing mixed sound

3.1. Spectrums generation based on PSE

The spectrum envelope $e^c(f)$ based on the PSE $E^c(f, y; \mu_f^c, \sigma_f^c)$ of category c is randomly generated as the following:

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

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

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

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

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

$$\Psi(f; \nu) = \sum_l \exp\left\{-\frac{(f - \nu \cdot l)^2}{2\lambda_0^2}\right\} \tag{9}$$

where l is the index of Gaussian components, and λ_0 is a hyperparameter for determining the kurtosis of each component.

The above procedure can generate the spectrum of category c and fundamental frequency ν.

3.2. Fitness calculation with supervised NMF

We set an intended basis matrix $\tilde{\mathbf{W}}$ whose rows are randomly generated from PSEs of various categories. Various fundamental frequencies of each row vector $\tilde{w}(f)$ are not duplicated for each category.

Given an amplitude spectrogram $\mathbf{X}$ of a test signal, an activity matrix $\tilde{\mathbf{H}}$ can be calculated by applying supervised NMF with $\tilde{\mathbf{W}}$. That is, only each element of $\tilde{\mathbf{H}}$ is repeatedly updated by Eq. (4) while keeping $\tilde{\mathbf{W}}$ fixed.

Fitness $\Theta(\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$ of given $\tilde{\mathbf{W}}$ and $\tilde{\mathbf{H}}$ is defined as:

$$\Theta(\tilde{\mathbf{W}}, \tilde{\mathbf{H}}) = \frac{1}{D_{EUC}(\mathbf{X}, \tilde{\mathbf{W}}\tilde{\mathbf{H}})} \tag{10}$$

where, D_{EUC} is Euclidean distance. If $\tilde{\mathbf{W}}$ has *better* spectral rows for the test signal, the distance between $\mathbf{X}$ and $\tilde{\mathbf{W}}\tilde{\mathbf{H}}$ becomes smaller. Therefore, the *better* $\tilde{\mathbf{W}}$ for the test signal makes fitness $\Theta(\tilde{\mathbf{W}}, \tilde{\mathbf{H}})$ larger.

Table 1: GA keywords in proposed method.

keyword	meaning
individual	a basis matrix $\tilde{\mathbf{W}}$
gene	a basis vector $\tilde{w}(f)$
fitness	inverse of the distance $D_{EUC}(\mathbf{X}, \tilde{\mathbf{W}}\tilde{\mathbf{H}})$
crossover	search the optimum by combining multiple PSEs
mutation	search the optimal spectrum envelope from PSE

3.3. 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 signal. To do this, we introduce an optimization method, which combines PSE, supervised NMF and genetic algorithm (GA).

GA is a method for finding the optimum by repeating natural-evolution-inspired techniques: selection, crossover, mutation and inheritance. Table 1 summarizes the meanings of each GA keyword in our proposed method.

The first step of the analysis stage is to generate L number of basis matrices $\{\tilde{\mathbf{W}}_l\}_{l=1}^{L}$ from pre-trained PSEs (See 3.1.), and calculate the fitness for each matrix (See 3.2.). Every matrices have the same fundamental frequencies in their basis rows. To update the whole set, the following process is repeated G times:

1. Copy the best basis matrix (the one with the highest fitness) 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 vector 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" in the above, a probability q_l, at which the basis $\tilde{\mathbf{W}}_l$ is selected, is defined as:

$$q_l = \frac{\Theta(\tilde{\mathbf{W}}_l, \tilde{\mathbf{H}}_l)}{\sum_{l=1}^{L} \Theta(\tilde{\mathbf{W}}_l, \tilde{\mathbf{H}}_l)} \tag{11}$$

Eq. (11) shows that the *better* $\tilde{\mathbf{W}}_l$ tends to be selected more. p_{cross} and p_{mut} in steps 2 and 3 are the probabilities of crossover and mutation, respectively. They satisfy $p_{cross} + p_{mut} = 1$.

Furthermore, the above GA steps have the following conditions and constraints in this paper:

- "Crossover" in step 2 is a uniform crossover with a probability of 0.5. Each new vector is either of the two parents.
- In the "mutate" step, each basis vector mutates with a probability of λ_{mut} (here, $\lambda_{mut} = 0.9$) without altering the fundamental frequency. In other words, the 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.

Because of these constraints, each basis vector of each basis matrix 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. Because $\hat{\mathbf{W}}$ contains a category index c, a test signal can be decomposed into each source.

4. Experiments

4.1. Robustness in regard to changes in environment

To verify how robust the proposed method is in regard to environmental changes, a "model-variant" test and a "reverb-added" test were conducted. In these test, only a piano-category PSE was trained. Training signals were recorded using MIDI piano sound ("Piano1") at a 16 kHz sampling rate. The MIDI file contains 6-octave-half-tones sources ($R = 72, N = 2705$) from "C1" to "B6". In the experiments, 5 types of recordings were prepared as various test signals:

(a) Played with "Piano1"

(b) Played with "Piano2" (another model than "Piano1")

(c) Played with 'Piano3' (another model than "Piano1")

(d) Recorded with a reverb level 40 (approximately 0.5 sec.)

(e) Recorded with a reverb level 100 (approximately 1.0 sec.)

All the test signals were recorded using a MIDI file, a part of "RWC-MDB-C-2001 No. 43: Sicilienne op.78" from RWC Music Database [1]. When analyzing, we set the number of GA individuals $L = 5$ and generations $G = 20$.

We compared the proposed method with the following:

(1) s-NMF: supervised NMF (given only "Piano1" basis)

(2) us-NMF: unsupervised NMF

(3) ex. s-NMF: supervised NMF (given all bases (a) $\sim$ (e))

Binarizing obtained activity matrices $\hat{\mathbf{H}}$ with an adequate threshold, and we obtained the final results of automatic musical transcription for each method.

Figure 3 illustrates transcription accuracies $acc[\%]$ for each method. The accuracy is calculated as:

$$acc = \frac{N_{all} - (N_{ins} + N_{del})}{N_{all}} \cdot 100 \tag{12}$$

where N_{all}, N_{ins} and N_{del} mean the number of all 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, $\tau = 0.2$).

According to the results of s-NMF, although the accuracy of (a) is relatively high, when it comes to the other environments, the accuracy deteriorates. Meanwhile, the decline cannot be seen so much with ex. s-NMF. This means that supervised NMF is not very robust to the sounds it does not know. The results of the proposed method show comparatively high accuracy to other environments (b) $\sim$ (e) even if the proposed system does not know their sounds either. The preferable result is due to the fact that 1) the proposed method can estimate PSE from various pitches, not just various models or playing-styles, and 2) it can cover spectrum envelopes of unknown models from the PSE. For this reason, it can be said that the proposed method has robustness to unknown sounds.

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

As well as the proposed method, accuracies of unsupervised NMF differ little among various environments. However, its results show the lowest accuracies due to the occurrence of unintended bases.

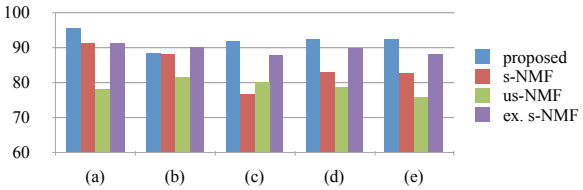

Figure 3: Accuracy rates of each method.

4.2. Analyzing mixed sound with multiple categories

In the experiment with multiple musical instruments, the PSE of the violin, in addition to the piano, were trained. The song used for the test was the same as that of the previous experiments, but multiple instruments played, as shown in Figure 4 (c). In the figure, the red and purple parts indicate piano and violin tones, respectively. Instrumental sources for the test and for the training are the same. In this experiment, the number of GA generations G was set to 500.

Analysis results using the proposed method are shown in Figure 4. Figure 4 (a) and (b) are the results of initial and final updating by GA, respectively. Category labels in error at the initial updating (found at $6 \sim 19$ seconds, G$\sharp$4 $\sim$ E5 tones of both instruments) are corrected almost completely by the 500th generation. This is because NMF matrices gradually get close to the test as GA updating proceeds. The results are shown Figure 4 (b), and these instrumentally-mixed sources can be partially separated. It is considered that there are definitive differences between piano PSE and violin PSE, and the differences improve the separation accuracy.

The main advantage of our method is that it can separate sound sources without having all possible knowledge about the instruments, unlike the supervised NMF and unsupervised NMF approaches. However, a larger number of generations G is required to raise the accuracy. In order to make the fullest possible use of PSE and reduce the computational time, future works include designing an EM-algorithm-based approach, instead of using GA.

5. Conclusions

In this paper, we proposed an algorithm for monaural sound source decomposition. The method categorizes some spectrum envelopes for a certain musical or phoneme 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. Both of them can be simultaneously estimated by SPGP+HS regression as described in this paper. In the analysis stage, Genetic Algorithm (GA) with supervised-NMF-based fitness was employed for an optimum search in all the spectrum envelopes that can be generated from the PSE.

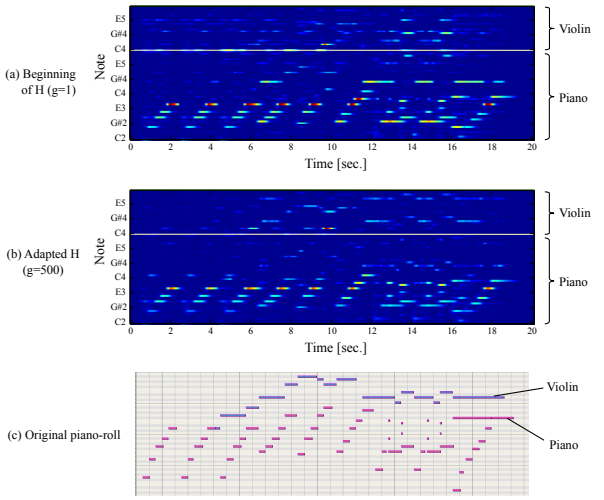

Figure 4: Results of the experiment with multiple instruments.

The simulation experiments using MIDI sources show that the proposed method is robust to environmental changes such as different models of instruments and reverb addition. When multiple categories are in a test signal, the separation worked to some extent.

6. References

[1] S. T. Roweis, "One microphone source separation," in *In Advances in Neural Information Processing Systems 13*. MIT Press, 2000, pp. 793–799.

[2] G. jin Jang and T. won Lee, "A maximum likelihood approach to single-channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, 2003.

[3] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2004, pp. 177–180.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, 2007.

[5] O. Dikmen and A. Cemgil, "Unsupervised single-channel source separation using bayesian nmf," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*. IEEE, 2009, pp. 93–96.

[6] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *In International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.

[7] A. Cont, S. Dubnov, and D. Wessel, "Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints," in *Proceedings of Digital Audio Effects Conference (DAFx)*, 2007, pp. 10–12.

[8] A. Cont, "Realtime multiple pitch observation using sparse non-negative constraints," in *International Symposium on Music Information Retrieval (ISMIR)*, 2006.

[9] E. Snelson and Z. Ghahramani, "Variable noise and dimensionality reduction for sparse Gaussian processes," in *Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence*. Citeseer, 2006.