FREQUENCY COMPONENT RESTORATION FOR MUSIC SOUNDS USING A MARKOV RANDOM FIELD AND MAXIMUM ENTROPY LEARNING

Tomonori Izumitani and Kunio Kashino

NTT Communication Science Laboratories
3-1, Morinosato Wakamiya, Atsugi-shi,
Kanagawa, Japan

ABSTRACT

We propose a method that estimates frequency component structures from music signals with noise and restores them. Restoring frequency components hidden by other interfering sounds is a difficult problem but has become important in various music information processing systems for melody extraction and audio retrieval. The proposed method is based on a probabilistic model of a frequency component structure represented as a Markov random field. To design an appropriate model, we introduce a supervised learning technique based on the maximum entropy model. We tested the method using musical audio signals generated from notes played by real instruments and noises. For four of six instruments, the proposed method achieves F-measures greater than 0.44 even in periods where signals are replaced by noises. We also evaluated the method in terms of feature distortion recovery in audio fingerprint matching tasks. The results show that the proposed method clearly reduces the effect of noise on the similarity values.

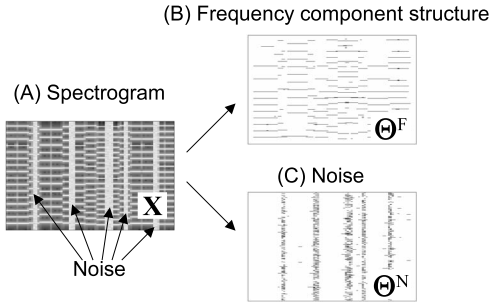

Fig. 1. Estimation of a frequency component structure by the proposed method. A spectrogram of a musical audio signal (A), a restored frequency component structure (B), and estimated noises (C).

1. INTRODUCTION

Frequency components, which comprise a fundamental frequency (F0) and its overtones, are time continuous peaks on a sound spectrogram. Their structure is an important property of instrumental and vocal sounds. Many systems for music information processing use this property. For example, systems for melody and bass-line extraction [1] and estimation of multiple F0s [2] often use harmonic structures.

Therefore, estimation of frequency component structure is an important task for various music information processing systems. However, the implementation is not straightforward because the structure can be easily broken by interfering factors caused by percussive sounds or noises.

In this study, we focus on the problem of estimating frequency component structures, including hidden or erased sections, from spectrograms of musical audio signals with noise.

Various methods which extract the frequency components have been developed for partial tracking tasks, e.g. methods using the Kalman filter [3] and Hidden Markov models [4], mainly in the context of analysis/synthesis. Based on a probabilistic scheme, we specifically address the case where the spectral peaks are intermittently hidden or erased by noise or other irregular sounds.

Frequency component structures have such a microscopic characteristic that the state of a point on a spectrogram is strongly affected by the states of adjacent points. We adopt a Markov random field (MRF)[5] to represent probabilistic properties of frequency component structures because it can naturally deal with the microscopic characteristic in a probabilistic form.

In MRF applications, it is essential to design a posterior density model of a state when its vicinity states are given. In addition, it is important to determine a set of appropriate model parameters. For this purpose, we introduce a supervised learning method. We use a maximum entropy model (MEM) [6] to represent and learn the posterior density. In the model, the posterior density is represented by a function proportional to the exponent of a linear combination of binary values.

Recently, Reyes-Gomez et al. proposed a method also using probabilistic models that can restore spectral powers when some regions on a spectrogram are missing [7]. Based on a learning scheme, our method does not require explicit transition models given in advance. In addition, ours estimates frequency component structures without prior knowledge on noise locations.

2. ESTIMATING FREQUENCY COMPONENT STRUCTURES

Many musical sounds have frequency component structures, typically harmonic structures, that reflect the characteristics of instruments or vocal mechanisms. These structures are represented as peaks in the frequency direction extending in the time direction on a sound spectrogram.

In practice, a musical sound may contain various kinds of transient and irregular components, as in percussive sounds and noises. In addition, the original instrumental or vocal sounds themselves have irregular components such as those appearing at the onset of a note. For simplicity, we treat these components as noises in this study.

These noises are represented as irregular patterns on a sound spectrogram [Fig. 1 (A)]. The objective of this study is to obtain the frequency component structure [Fig. 1 (B)] in a sound spectrogram of musical audio signals containing noise. We represent a frequency component structure as a frequency $\times$ time state matrix, $\mathbf{\Theta}^{\mathrm{F}}$, which represents the existence of frequency components by binary values 0/1. The size of the matrix $\mathbf{\Theta}^{\mathrm{F}}$ is the same as X, which is an observation matrix comprised of spectral powers of the original audio signal.

In order to recover hidden or erased frequency components, component interpolation is needed as well as noise component suppression. Thus, we formulate the following *a posteriori* probability maximization model :

$$\hat{\mathbf{\Theta}} \;=\; \arg\max_{\mathbf{\Theta}^{\mathrm{F}},\mathbf{\Theta}^{\mathrm{N}}} P(\mathbf{\Theta}^{\mathrm{F}},\mathbf{\Theta}^{\mathrm{N}}|\mathrm{X}). \tag{1}$$

In this formulation, we introduce an auxiliary matrix $\mathbf{\Theta}^{\mathrm{N}}$, which is a binary state matrix that represents components originating in noise, to reflect a difference in the generation mechanisms of observation X according to the existence of noise.

In this paper, we represent the elements of matrices X, $\mathbf{\Theta}^{\mathrm{F}}$, and $\mathbf{\Theta}^{\mathrm{N}}$ as x_{ij}, θ^{F}_{ij}, and θ^{N}_{ij}, respectively. Borrowing terminology from image processing, we call each pair comprising a frequency and time component, namely, a point (i, j) on the spectrogram, a "pixel". The two kinds of states that are represented by $\mathbf{\Theta}^{\mathrm{F}}$ and $\mathbf{\Theta}^{\mathrm{N}}$ are called the "frequency component state" and "noise state," respectively.

In the rest of this paper, we use notations $\mathbf{\Theta}$ and θ_{ij} to simultaneously represent two state matrices, $\mathbf{\Theta}^{\mathrm{F}}$ and $\mathbf{\Theta}^{\mathrm{N}}$, and their elements, θ^{F}_{ij} and θ^{N}_{ij}.

3. METHODS

3.1. Defining a probabilistic model using MRF

Focusing on a small region around a pixel, we find that a frequency component has following strong characteristics: if the pixel is involved in the frequency component structure, neighboring pixels in the frequency direction tend to be excluded from frequency components while those in the time direction tend to be included.

The MRF is suitable for representing such characteristics in the form of probabilistic model. It assumes that states of a pixel can be characterized only by the states of neighboring pixels and introduces "potential functions," V^o and V_c, to reflect such characteristics.

To define a model, we consider a neighborhood region around a pixel (i,j), which consists of the pixel (i,j) and its neighboring pixels. Suppose that G_{ij} denotes a set of neighboring pixels around (i,j), excluding (i,j) itself. Then, appropriate potential functions, $V^o(\theta_{ij},\theta_{i'j'},x_{i,j},x_{i'j'})$ and $V_c(\theta_{ij},\theta_{i'j'})$, $(i',j') \in G_{ij}$, are defined for all cliques that include the pixel at (i,j). These functions correspond to likelihood $P(\mathrm{X}|\mathbf{\Theta})$ and prior distribution $P(\mathbf{\Theta})$, respectively. A clique c means a subset of pixels within a neighborhood region such that every pixel in c is also included in the neighborhood region of the rest of the pixels in it.

The conditional distribution $P(\theta_{ij}|\theta_{i'j'},x_{i,j},x_{i'j'},(i',j') \in G_{ij})$ is represented as

$$\begin{aligned} &P(\theta_{ij}|\theta_{i'j'},x_{i,j},x_{i'j'},(i',j') \in G_{ij}) \\ &\propto \exp\left[-\frac{1}{T}\left(V^o(\theta_{ij},\theta_{i'j'},x_{i,j},x_{i'j'}) + \sum_{c:\ (i,j)\in c} V_c(\theta_{ij},\theta_{i'j'})\right)\right], \end{aligned} \tag{2}$$

where T is a parameter that controls the sharpness of the distribution. The neighborhood region may be defined by the Euclidean distance [5]. Here, we simply adopt a rectangular region composed of forward and backward m (n) pixels along the direction of frequency (time) [Fig. 2 (A)].

In order to estimate the state matrix $\hat{\mathbf{\Theta}}$ that has maximum posterior density, we use Gibbs sampling and simulated annealing, which iteratively generate θ_{ij} according to $P(\theta_{ij}|\theta_{i'j'},x_{i,j},x_{i'j'},(i',j') \in G_{ij})$, moving the focusing pixel and its neighborhood region. This process gradually sharpens peaks of the distribution.

In the image processing field, potential functions V_c and V^o are often defined manually using *a priori* knowledge about the objects. However, the justification of such models is not straightforward and empirical parameter tunings are unavoidable.

3.2. MEM

To avoid such problems, we used a supervised learning method, the MEM, which can directly estimate the posterior distribution $P(\omega|D)$ from training data. Here, D denotes a data sample and ω a class or category. In this study, category ω corresponds to θ_{ij} and each sample D corresponds to each pixel at (i,j) on a spectrogram. We assume sample D is characterized only by the states and observations within the neighborhood region except for the state at (i,j); namely, $\theta_{i'j'},\mathrm{x}_{i'j'},(i',j') \in G_{ij}$, and x_{ij} [Fig. 2 (A)]. In what follows, we outline how the MEM is employed.

Firstly, we introduce multiple feature functions $f_l(D,\omega),(l = 1,2,...,F)$ that take binary values (0 / 1). Each function is defined by a combination of data D and a class ω. Then, we want to estimate $P(\omega|D)$ using feature functions $f_l(D,\omega)$. When training data are given, we can calculate the expectation of each feature function, $\tilde{E}(f_l)$, by counting the number of cases where $f_l(D,\omega) = 1$ in the training data. The basic idea of the MEM is to find the optimal $P(\omega|D)$ that maximizes the entropy of joint distribution $P(\omega,D) = P(\omega|D)P(D)$ under the condition that the expectations of $f_l(D,\omega)$ calculated from $P(\omega|D)$ are equal to $\tilde{E}(f_l)$ for all l.

The solution to the above maximization can be written as

$$P_\Lambda(\omega|D) = \frac{1}{Z(D)}\exp\left(\sum_l \lambda_l f_l(D,\omega)\right), \tag{3}$$

where $Z(D)$ is a normalization term and $\Lambda = (\lambda_1,...\lambda_F)$ are model parameters to optimize. The Λ can be estimated by an improved iterative scaling method [6], which is a kind of hill-climbing method.

3.3. Defining feature function for MEM

A feature function $f_l(D,\omega)$ is defined with respect to ω and data D, where $\omega = \theta_{ij}$. To extract features for each pixel, information from $\theta^{\mathrm{F}}_{i'j'},\theta^{\mathrm{N}}_{i'j'},x_{i,j}$, and $x_{i',j'},(i',j') \in G_{ij}$ are available.

In preparation for defining f_l, we introduce a static feature function $g_{l'}(\theta_{i'j'},\mathrm{x}_{i'j'},\mathrm{x}_{ij}),(i',j') \in G_{ij}$, which also takes a binary value and does not depend on $\omega = \theta_{ij}$.

Figure 2 (B) shows how features from states and observations are extracted. First, we simply use the value of the frequency component and noise state, $\theta^{\mathrm{F}}_{i'j'}$ and $\theta^{\mathrm{N}}_{i'j'}$, $(i',j') \in G_{ij}$, as shown in Fig. 2 (B)–(i) and –(ii). These two states yield $(2m+1)(2n+1)-1$ static feature functions.

Power peaks in the direction of frequency provide rich information about frequency components, especially in a clear sound without noise. We use binary information in order to represent whether each pixel is a power peak or not. We represent it as $\mathrm{x}^{\mathrm{P}}{}_{ij}$ and X^{P} in

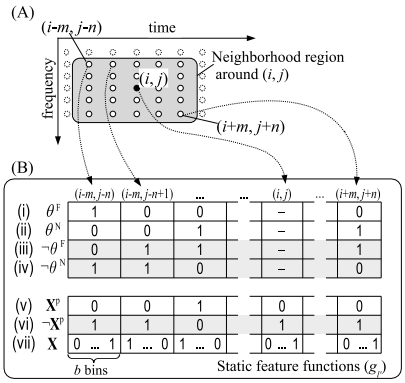

Fig. 2. A neighborhood region around a pixel at (i, j) (A) and extracted values of static feature functions ($g_{l'}$) from states and observations within the neighborhood region around a pixel at (i, j) (B).

a matrix form, which yields $(2m + 1)(2n + 1)$ static feature functions [Fig. 2 (B)–(v)].

Spectral powers x_{ij} and $\mathrm{x}_{i'j'}, (i', j') \in G_{i'j'}$ take real values. We allocate one of b bins to each x_{ij} or $\mathrm{x}_{i'j'}, (i', j') \in G_{i'j'}$ according to the spectral power [Fig. 2 (B)–(vii)]. This yields $b(2m+1)(2n+1)$ static feature functions.

For $\mathbf{\Theta}^{\mathrm{F}}$, $\mathbf{\Theta}^{\mathrm{N}}$ and X^{p}, we introduce the negative of each value [Fig. 2 (B)–(iii), –(iv), and –(vi)]. These values are used in order to keep the sum of all $f_l(D, \omega)$ through the training samples constant and thereby simplify the MEM learning [6].

From the above discussion, a feature function $f_l = f_{\omega^{\mathrm{F}}\omega^{\mathrm{N}}l'}$ can be defined for all four combinations of frequency component and noise states, $\omega^{\mathrm{F}} = 0/1$ and $\omega^{\mathrm{N}} = 0/1$, using static feature functions $g_{l'}$ as follows:

$$f_{\omega^{\mathrm{F}}\omega^{\mathrm{N}}l'}(D, \theta^{\mathrm{F}}_{ij}, \theta^{\mathrm{N}}_{ij}) = \begin{cases} g_{l'}(D) & \text{if } \theta^{\mathrm{F}}_{ij} = \omega^{\mathrm{F}} \text{ and } \theta^{\mathrm{N}}_{ij} = \omega^{\mathrm{N}}, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

4. EXPERIMENTS

We evaluated the effectiveness of the proposed method in terms of frequency component restoration accuracy. We performed two experiments. In the first experiment, precision and recall values with respect to frequency component restoration were measured, and in the second, recovery from audio feature distortion was measured.

4.1. Preparing training and test data

To generate training data, we composed artificial sounds of musical notes that are constituted only by harmonics with a constant amplitude.

For the test data, we used acoustic instrumental sounds of musical notes in the RWC Musical Instrument Sound Database (RWC-MDB-I-2001 No. 01-50) [8]. We chose five instrumental sounds and one vocal sound, namely, piano, violin, flute, trumpet, marimba, and alto (vocal), with the normal playing style.

First, we prepared 3.5-sec and 7.5-sec musical phrases, by connecting sound segments obtained above, for training and testing, respectively. Next, for both musical phrases, we made noisy sounds artificially by replacing certain sections with Gaussian noise. These sound data were used in order to generate observation matrices X in all experiments.

All audio signals were sampled at $Fs = 11,025$ Hz. Each sampled signal was divided into frames, each having 1,024 samples. A fast Fourier transform (FFT) was calculated for each frame with the Hanning window. Spectral powers in decibels were used in composing the observation matrix X.

Table 1. Performance of the proposed method. Values in parentheses indicate the results for simple peak extractions without the proposed method.

(A) Whole period (7.5 sec.)

instrument	precision	recall	F-measure
piano	0.60 (0.26)	0.80 (0.88)	0.69 (0.41)
violin	0.63 (0.35)	0.64 (0.87)	0.64 (0.50)
flute	0.58 (0.32)	0.60 (0.88)	0.59 (0.47)
trumpet	0.61(0.32)	0.60 (0.87)	0.61(0.47)
alto	0.50 (0.31)	0.36 (0.86)	0.42 (0.45)
marimba	0.28 (0.12)	0.28 (0.85)	0.28 (0.21)

(B) Periods where audio signals are replaced by noises (1.1 sec. in total)

instrument	precision	recall	F-measure
piano	0.44 (0.07)	0.52 (0.58)	0.47 (0.12)
violin	0.48 (0.11)	0.40 (0.57)	0.44 (0.18)
flute	0.47 (0.08)	0.42 (0.57)	0.44 (0.14)
trumpet	0.50 (0.09)	0.41 (0.56)	0.45 (0.15)
alto	0.33 (0.09)	0.23 (0.55)	0.27 (0.15)
marimba	0.09 (0.03)	0.13 (0.47)	0.11 (0.05)

4.2. Evaluation of the accuracy

For the training data, two parts of the original 3.5-sec musical sound were replaced by 200- and 300-msec-long Gaussian noises. For the test data, we used five short Gaussian noises, which have various durations of 150 to 300 msec.

To make the ground truth for frequency component states $\mathbf{\Theta}^{\mathrm{F}}$, power peaks along the frequency axis having power greater than a threshold were extracted from spectrograms of the original sound without noise.

Power peaks in FFT frames that overlap noise sections and those that do not appear in the spectrograms of the original sound without noise were extracted for the ground truth for noise states Θ^{N}.

To create feature functions associated with spectral powers, X in Fig. 2 (B)–(vii), the value of spectral powers were normalized in the direction of frequency into a range of [0 1] using the powers of 100 neighboring frequency components.

In this experiment, the size of the neighborhood m and n was set at 3 and the number of bins, b, used to create feature functions for MEM was 3.

We examined 509 (frequency) $\times$ 90 (time) pixels and calculated the accuracy of estimated θ^{F}_{ij}. The performance was measured by precision, recall, and the F-measure. The F-measure is the harmonic mean of the precision and the recall.

Table 1 (A) shows the performance of the proposed method for the whole 7.5-sec sound. Table 1 (B) shows the results calculated only for the sections replaced by noises. Parenthetic values are the results from simple peak extraction.

For four instruments, namely, piano, violin, flute, and trumpet, the precision, recall, and F-measure are greater than 0.44 even in the noise sections. And the proposed method improved the perfor-

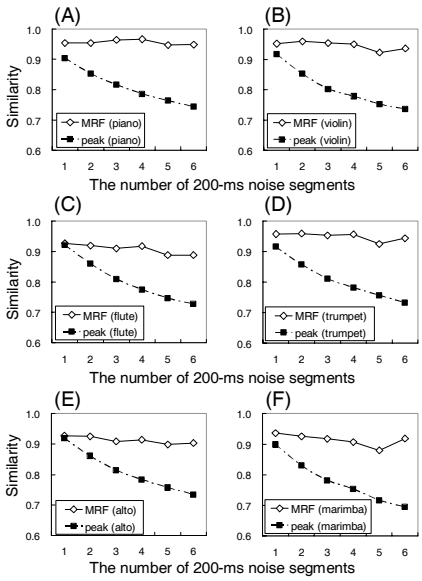

Fig. 3. Relationship between the number of 200-ms noise segments and the similarity to musical sounds without noise for piano (A), violin (B), flute (C), trumpet (D), alto (E), and marimba (F). "MRF" indicates the proposed method and "peak"indicates the peak extraction method.

mance for all instruments in the precision and F-measure.

Simple peak extraction produced many extra peaks, especially in the noise regions. This resulted in comparatively higher recall values, but the precision values were severely degraded. In "marimba", all measures are very low. This is due to the evaluation method: the marimba has few harmonics and includes many irregular components. Therefore, the irregular components remained in the ground truth.

4.3. Evaluation in terms of feature distortion

Assuming that the proposed method will be applied to preprocessing in audio fingerprinting systems, we examined how it reduces the effects of noise on the similarity between the restored sounds and the original noiseless sounds.

To calculate the similarity, features completely different from the feature functions in MEM have to be extracted from sound data. We used the feature extraction method developed by Kashino et al [9]. That is, feature vectors were composed of the power in several frequency bands with a constant width along the frequency axis in a log-scale.

The experimental procedure was as follows:

- Spectrogram X was masked by an estimated frequency component state matrix $\mathbf{\Theta}^{\mathrm{F}}$ and the estimated frequency components from X was extracted.
- Seven frequency bands having a constant width along the frequency axis in a log-scale were made using the 300 - 3000 Hz region of the masked spectrograms.
- For each time frame, the average power of each band was calculated and a feature vector consisting of the average power at each frequency band and the time frame was generated.

Here, we simply used these feature vectors for the calculation of similarities. The cosine measure was used as the similarity.

The evaluation was performed using data where from one to six 200-msec-long Gaussian noises were substituted in the original sound, and the same experimental conditions as in the accuracy evaluations were used.

Figure 3 shows the relationship between the number of noise sections and the similarity to the sounds without Gaussian noise. It is clearly shown that with the proposed method the similarity measures for all instruments remain nearly unaffected by the amount of noise, whereas, without the proposed restoration, they decrease monotonically when the amount of noise increases. From this result, we anticipate that the proposed method will greatly improve the robustness of audio search based on fingerprinting.

5. CONCLUSION

We have proposed a method that restores frequency component structures from musical sound signals containing noise. The method exploits a probabilistic model represented by a Markov random field. Using a supervised learning method, based on a maximum entropy model, we resolved the difficulties in designing potential functions and in tuning parameters. The experimental results showed that the method greatly improves the accuracy of frequency component estimation in terms of F-measures, especially in sections where the original signals are replaced by noise. In addition, the results for similarity measures clearly indicated that the proposed method has the potential to improve the performance of music information retrieval systems based on fingerprinting.

6. REFERENCES

[1] M. Goto, "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, pp. 311–329, 2004.

[2] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, 2003.

[3] A. Sterian and G. H. Wakefield, "A model-based approach to partial tracking for musical transcription," in *Proc. of SPIE98*, 1998, pp. 171–182.

[4] Ph. Depalle, G. García, and X. Rodet, "Tracking of partials for additive sound synthesis using hidden markov models," in *Proc. of ICASSP93*, 1993, pp. I225–I228.

[5] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, 1984.

[6] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *Proc. of IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999, pp. 61–67.

[7] M. Reyes-Gomez, N. Jojic, and D. Ellis, "Deformable spectrograms," in *Proc. of AI & Statistics 2005*, Robert G. Cowell and Zoubin Ghahramani, Eds., 2005, pp. 285–292.

[8] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. of ISMIR2003*, 2003, pp. 229–230.

[9] K. Kashino, T. Kurozumi, and H. Murase, "A quick search method for audio and video signals based on histogram pruning," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 348–357, 2003.