MUSIC RECOGNITION USING NOTE TRANSITION CONTEXT

Kunio Kashino and Hiroshi Murase

NTT Basic Research Laboratories
3-1 Morinosato-Wakamiya, Atsugi-shi, Kanagawa, 243-01 Japan.
kunio@ca-sun1.brl.ntt.co.jp, murase@apollo3.brl.ntt.co.jp

ABSTRACT

As a typical example of sound-mixture recognition, the recognition of ensemble music is addressed. Here music recognition is defined as recognizing the pitch and the name of an instrument for each musical note in monaural or stereo recordings of real music performances. The first key part of the proposed method is adaptive template matching that can cope with variability in musical sounds. This is employed in the hypothesis-generation stage. The second key part of the proposed method is musical context integration based on the probabilistic networks. This is employed in the hypothesis-verification stage. The evaluation results clearly show the advantages of these two processes.

1. INTRODUCTION

We have been addressing the music recognition task for real musical performances. Here music recognition is defined as the problem of recognizing the pitch and the name of an instrument for each musical note in a monaural or a stereo recording of a real music performance. It is expected that this music recognition technique will be applicable to automatic music transcription systems, signal-to-MIDI (Musical Instrument Digital Interface) conversion systems, and music-database indexing systems.

The approach towards the music recognition problem has had a long history. The early work inspired by frequency-analysis techniques concerns the transcription of a single-pitched melody such as a vocal solo[7, 5]. Later, recognition systems for multiple-pitched music performed by a single musical instrument (*e.g.* piano solos) were proposed[4]. However, few works have addressed the recognition of multiple-pitched music performed by *multiple* kinds of musical instruments (*e.g.* such as by a chamber ensemble), although several attempts can be found in the literature[1, 2].

In music signals, most frequency components that originate in different musical notes overlap. This results in an intrinsic ambiguity in the interpretation of input signals. In addition, the identification of musical instruments becomes an essential issue in the multiple-instrument case. Specifically, when the input music is a real recording rather than a sampler[1] performance[3], then the identification based on the conventional methods such as the discriminant analysis or the template matching becomes difficult due to the feature variability for each note. As a solution to these technical problems, two processing stages are proposed in this paper.

[1]A sampler is an electronic musical instrument that stores waveforms of real instruments on memory and playbacks them on receiving MIDI data from a computer.

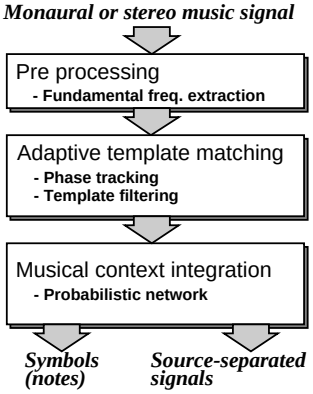


Figure 1: The overview of the proposed processing.

This paper is organized along the processing scheme presented in Figure 1. After a brief remark on the pre processing in Section 2, two processing stages mentioned above are described in Sections 3 and 4, respectively. The first processing called adaptive template matching is invented to cope with the variability of each musical note. The second processing called the music stream network method is designed to improve recognition accuracy by integrating musical context. Evaluations for these two methods are discussed in Section 5, followed by a concluding remark given in Section 6.

2. PREPROCESSING

As explained in the following section, the proposed adaptive template matching requires an average fundamental frequency for each note. Thus fundamental frequency extraction is first performed on the input signal as a preprocessing.

Since the input may include the unknown number of musical notes, it is not straightforward to extract all the fundamental frequencies; one cannot realize this extraction without prior knowledge on features of notes that may be included in the input. The extraction here is based on a gross matching between the input spectral pattern and those of many notes stored in a database. According to the preliminary experiments using three-part real ensemble performances, it is shown that both the "precision" (propor-

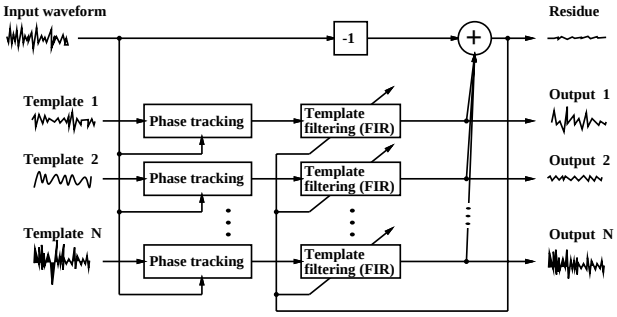


Figure 2: Block diagram of the adaptive template matching.

tion of correctly extracted notes in all the extracted notes) and the "recall" (proportion of correctly extracted notes in all the input notes) are approximately 80 % in the current implementation.

3. ADAPTIVE TEMPLATE MATCHING

The basic idea of the template matching here involves the matched filter. As widely accepted, a matched filter in the time-domain is a powerful tool to identify the signal of a specific sound in a mixture. In our scheme, a bank of matched filters is arranged in parallel. Each template corresponds to a musical note played by a specific kind of instrument with a specific musical pitch (*e.g.* Piano-C4, Piano-D4, $\cdots$). Then, by calculating correlation between the output from each filter and the input signal, a specific sound source is detected.

In this scheme, adaptation of templates is important, because the waveforms from musical notes in real performances differ significantly according to the individual instrument, the expression such as vibrato, and the player. Therefore we have devised the adaptive template matching method as shown in Figure 2. The method consists of two stages: (1) phase tracking and (2) template filtering by FIR adaptive filters.

3.1. Phase Tracking

The first step of template adaptation is phase filtering. Phase tracking absorbs the phase fluctuations of the fundamental frequency components.

If the input signal is not a mixture of multiple sounds but a single sound, adaptive pitch tracking methods as discussed in the literature can be used. However, such signal processing methods are not directly applicable to a sound mixture where multiple pitches are present. Thus we have developed a simple algorithm to realize the phase adaptation. The algorithm consists of the following six steps.

(1) At the pre-processing stage, perform frequency analysis on the input z, to extract the fundamental-frequency components. Because z may be a mixture of multiple sound signals, there may be multiple fundamental-frequency components.

(2) For each fundamental frequency component, choose r_i. Each r_i is a template of a possible sound included in z.

(3) Apply a narrow-band bandpass filter to r_i, using the average fundamental frequency of each r_i as the bandpass-filter center frequency. For each time sample, store the phase of the output waveform of the bandpass filter. Let $p_{r,i}(k)$ denote the phase at time k.

(4) Apply the same bandpass filter, as applied to r_i, to the input z, and store the phase information for each fundamental frequency as $p_{z,i}(k)$.

(5) Calculate the required time shift $\Delta k_{r,i}(k)$. Since the phase difference $\Delta p_{r,i}(k)$ is given as:

$$\Delta p_{r,i}(k) = p_{z,i}(k) - p_{r,i}(k)\,, \tag{1}$$

the time shift $\Delta k_{r,i}(k)$ is calculated by:

$$\Delta k_{r,i}(k) = \frac{f_s}{2\pi f_{c,i}} \Delta p_{r,i}(k), \tag{2}$$

where f_s is the sampling frequency and $f_{c,i}$ is the center frequency of the applied bandpass filter.

(6) The amplitude value r_i at time k is given as:

$$r_i(k) = r_i(\ k - \Delta k_{r,i}(k)\ ). \tag{3}$$

3.2. Template Filtering

The second step of template adaptation is template filtering. Template filtering absorbs the fluctuation in the amplitude of the fundamental frequency components, and both the amplitude and phases of the overtones.

We consider representing an input acoustic signal $z(k)$ with a sum of template waveforms, each of which is given by the convolution of the filter coefficients $h_n(m)$ and the phase-adjusted waveform $r_n(k)$. Then our problem can be formulated as the minimization of J in the equation:

$$J = E\left[\left\{z(k) - \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} h_n(m)\ r_n(k-m)\right\}^2\right]\ , \tag{4}$$

where k is the time sample, n enumerates the templates, N is the estimated number of sound sources (which is not predefined), M is impulse response length of the filters, and E denotes an average over time.

The necessary condition for J to hold the minimum value over $h_n(m)$ is that the values of the partial derivative $\partial J/\partial h_n(m)$ are 0 for all n and m. Using this condition, it is straightforward to derive $N \times M$ simultaneous linear equations as follows:

$$\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} E\left[r_i(k-j)\ r_n(k-m)\right]\ h_n(m) = E\left[r_i(k-m)\ z(k)\right]\ , \tag{5}$$

where $i = \{0, 1, \cdots, N-1\}$ and $j = \{0, 1, \cdots, M-1\}$. By solving this equation, the optimal filter coefficients $h_n(m)$ are obtained. The computation required in solving Equation 5 is third order with respect to M and N.

4. MUSICAL CONTEXT INTEGRATION

4.1. Creating Music Stream Networks

While the above described method absorbs the musical note fluctuations, the method uses only local information and the matching results can still be ambiguous. Thus, in our scheme shown in Figure 1, the output from adaptive template matching is treated as hypotheses rather than final results, and a hypothesis verification method is employed to integrate musical context.

The hypothesis verification method is based on "music stream networks (MSN)". The MSN is a Bayesian probabilistic network that represents the stream of a melody. The network is constructed as shown in Figure 3. Let us consider two musical notes n_k, n_{k-1} (k denotes the order of the onset times of these notes, n_{k-1} precedes n_k). We define $Z(n_k, n_{k-1})$ using Equation (6):

$$Z(n_k, n_{k-1}) = W \sum_i \Big\{ - \ w_i \ \log P_i(n_k, n_{k-1}) \Big\} \ , \qquad (6)$$

where i is a suffix that enumerates the factor of Z, P_i is a conditional probability of the occurrence of the n_{k-1} to n_k transition in a given musical context, and w_i (> 0) is a weight for each factor. Since the component $-\log P_i$ is self-information delivered by the transition from n_{k-1} to n_k, Z can be viewed as a weighted sum of self-information. Thus Z reflects the infrequency of the transition for these two notes. Therefore we define the "music stream" as the sequence of musical notes that gives a locally minimum Z.

The term W is a time window that is defined as:

$$W(\delta t) = \exp\left(\frac{\delta t}{\tau}\right) \ , \qquad (7)$$

where δt is the difference between onset times for these two notes, and τ is a time constant. Unlike ordinary time windows, W becomes greater as δt increases.

Currently the following three factors of Z have been considered: (1) the transition of musical intervals, (2) the similarity of timbres, and (3) the consistency of musical roles.

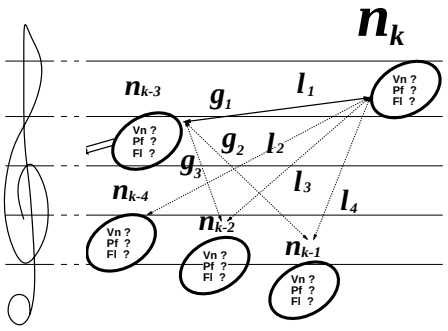
Firstly, the pitch transition probability in a melody can be used as P_1 in Equation (6). To obtain P_1, we analyzed 397 melodies extracted from 196 pop scores and 201 jazz scores, and calculated the probabilities of musical intervals. The number of note transitions was 62,689.

Secondly, it is reasonable to suppose that a sequence of notes tends to be composed of notes that have similar timbres. To incorporate this tendency, we define a distance measure between the timbres for two notes. We then estimate the probability that two notes a given distance apart sequentially appear in a music stream. This probability is used as P_2 in Equation (6). The distance between timbres is defined as the Euclidean distance between the timbre vectors. A timbre vector is a vector whose elements are the correlation values between the output from the template filters and the corresponding portion of the input signal. The distances between successive notes in a sequence are translated into probabilities using a histogram. This histogram models the distribution of timbre vectors for notes.

Finally, in ensemble music, a sequence of notes can be regarded as carrying a musical role such as a principal melody or a base-line. To introduce such musical semantics, we evaluate the probability that a note plays a musical role in a sequence of notes. This probability is used as P_3 in Equation (6). Although P_3 can be determined using a statistical analysis, we introduce a simplified approximation for P_3:

$$P_3 = ar + b \ , \qquad (8)$$

where a and b are constants, and r is the proportion of the highest (or lowest) notes in the musical stream under consideration.



When a new node (n_k) is created, the system first chooses the link that gives the minimum Z value (l_1) among the candidate links ($l_1, \cdots, l_4$). The system then evaluates Z values for the link candidates ($g_1, \cdots, g_3$) from the selected node (n_{k-3}), to choose the link with the minimum Z value (g_1). If g_1 and l_1 are identical, the link composes a music stream. If a music stream from n_{k-3} is already formed in a direction other than g_1, then the stream is cut, and the direction of the music stream is changed to $g_1 (= l_1)$.

Figure 3: A procedure showing the MSN creation.

4.2. Information Propagation on the MSN

Once the MSNs are created, the next task is to choose the most likely set of hypotheses, taking advantage of the contextual links. This can be done by regarding the MSN as the Bayesian network[6]. The Bayesian network is a tool for calculating the *a posteriori* probability when a series of events related to each other is observed. The information propagation scheme described in [3] enables us to choose the best-balance set of hypotheses each time an observation (*i.e.* matching result of a new incoming note) is made. The amount of computation required in this propagation is the linear order with respect to the number of notes.

5. EVALUATIONS

We have tested the proposed method using recordings of real ensemble performances listed in Table 1. These songs were arranged as three-part ensembles and each part was single-pitched.

Templates used in the adaptive template matching stage were played by different manufacturers' instruments from the ones used in the recording of the test songs. The number of taps in the template filtering was 20. We stored

Table 1: Test songs used in the evaluation experiments.

Title	Instruments (Part order)	#Notes
Annie Laurie *	Fl, Vn, Pf	234
Lorelei **	Fl, Vn, Pf	297
Dreaming of Home and Mother ***	Vn, Fl, Pf	304
Auld Lang Syne *	Vn, Fl, Pf	242
Total		1077

Vn: Violin, Fl: Flute, Pf:Piano

Music by:
* Scotland air, ** Friedrich Silcher, *** J.P.Ordway

piano, flute, and violin templates; this means that the system presumed that each input note was played by either piano, flute, or violin. However, the number of simultaneous notes for each instrument was not given to the system. The number of parts were also unknown.

In order to clearly evaluate the two methods focused in this paper, we manually fed the system with the correct pitch information (MIDI note number) for each note, although this information is normally yielded by the preprocessing stage. Therefore the task of this evaluation tests was a sound source identification. The recognition rate R was simply defined as :

$$R = \frac{(\#\text{correctly recognized notes})}{(\#\text{output notes in total})} \quad (9)$$

The results are displayed in Figure 4. Here the "template filtering off" condition means that the number of taps in the template filtering (M in Equation (4)) was chosen to be 1. Therefore turning all the elements off is equivalent to the conventional matched filtering. Thus Figure 4 clearly shows that both of the adaptive template matching (PT and TF) and the musical context integration (NT, TS and CR) improves the source identification accuracy.

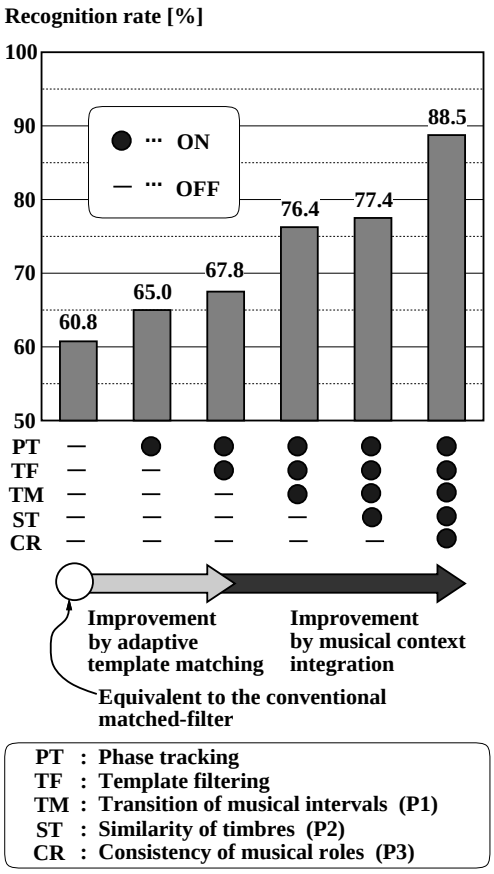


Figure 4: Summary of the experimental results.

6. CONCLUSIONS

We have presented a new processing method for ensemble music recognition. The method consists of two stages, adaptive template matching and musical context integration. Specifically, the evaluations using recordings of real ensemble performances have revealed that the integration of musical context improves the precision of source identification from 67.8 % to 88.5 % on average.

We are planning to evaluate the system using musical performances that have further varieties; for example, performances that include musical instruments different to those reported here, and also performances with more than three parts. In addition, application of the proposed method to an automatic transcription system and a music-database indexing system will also be considered in future work.

7. ACKNOWLEDGMENTS

The authors wish to thank Dr. Y. Tohkura, Dr. K. Ishii, Dr. H. G. Okuno, Dr. N. Osaka, Dr. T. Kawabata, and T. Nakatani for their help and encouragement in conducting this research.

8. REFERENCES

[1] Chafe, C. and Jaffe D.: Source Separation and Note Identification in Polyphonic Music. *Proc. of ICASSP-86*, pages 1289–1292, 1986.

[2] Kashino K. and Murase H.: A Music Stream Segregation System Based on Adaptive Multi-Agents. *Proc. of the IJCAI-97*, Vol.2, pages 1126–1131, 1997.

[3] Kashino K., Nakadai K., Kinoshita T., and Tanaka H.: Organization of Hierarchical Perceptual Sounds. *Proc. of the IJCAI-95*, Vol.1, pages 158–164, 1995.

[4] Katayose, H. and Inokuchi S.: An Intelligent Transcription System. *Proc. of Int'l. Conf. Music Perception and Cognition*, pages 95–98, 1989.

[5] Niihara T. and Inokuchi S.: Transcription of Sung Song. *Proc. of ICASSP-86*, pages 1277–1280, 1986.

[6] Pearl J.: Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence*, 29(3):241–288, 1986.

[7] Piszczalski M. and Galler B. A.: Automatic Music Transcription. *Comp. Music Journal*, 1(4):24–31, 1977.