Automatic Feature Weighting in Automatic Transcription of Specified Part in Polyphonic Music

Katsutoshi Itoyama, Tetsuro Kitahara, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno
Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{itoyama, kitahara, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

We studied the problem of automatic music transcription (AMT) for polyphonic music. AMT is an important task for music information retrieval because AMT results enable retrieving musical pieces, high-level annotation, demixing, etc. We attempted to transcribe a part played by an instrument specified by users (*specified part tracking*). Only two timbre models are required in the *specified part tracking* to identify the specified musical instrument even when the number of instruments increases. This transcription is formulated into a time-series classification problem with multiple features. We furthermore attempted to automatically estimate weights of the features, because the importance of these features varies for each musical signal. We estimated quasi-optimal weights of the features using a genetic algorithm for each musical signal. We tested our AMT system using trio stereo musical signals. Accuracies with our feature weighting method were 69.8% on average, whereas those without feature weighting were 66.0%.

Keywords: automatic music transcription, specified part tracking, feature weighting, genetic algorithm

1. Introduction

Recently, because of the growth of the digital music industry, demand for music information retrieval (MIR) and management of musical data has been increasing. Automatic music transcription (AMT) is needed to improve MIR because musical scores enable MIR by melody or musical instrument, etc. AMT for polyphonic music generally consists of two successive processes: *note formation*, which estimates the onset time and pitch of each note, and *stream formation*, which classifies the formed notes by their instruments (parts). The latter problem has not been studied enough, which made the AMT incomplete. Therefore, a method to form streams is strongly required to realize an AMT for polyphonic music.

Previous studies of stream formation were classified broadly into two approaches. One identifies the musical instruments of all parts and labels all the instruments given [1, 2, 3]. In this approach, training data for all instruments which could be contained in musical pieces is required to separate all instruments exactly. Another forms streams without information about musical instruments contained in musical pieces. This approach does not require training data [4]. However, users cannot extract streams they wanted because the obtained streams have no label of the target instrument.

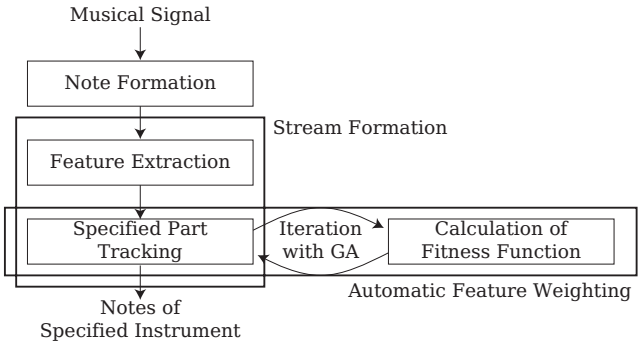

Figure 1. Overview of Specified Part Tracking

We developed a new approach in which the AMT system is given one of the musical instruments included in the musical pieces and transcribes that part of the musical pieces. By focusing on only one musical instrument a user specified, we require only two timbre models to identify the specified musical instrument. *Specified part tracking* is the stream formation based on our approach.

We also developed a method for automatically estimating weights of features which are used in the specified part tracking. The importance of these features depends on musical signals. For example, directional reliability by alignment of instruments, distortion of timbre features by noises. We develop a method for estimating quasi-optimal weights for each musical signal using a genetic algorithm.

2. Problem Specification

Specified part tracking classifies musical notes into the set of notes of a specified instrument N and that of other instruments $\bar{N}$. We defined a pair of them $\mathcal{H} = (N, \bar{N})$ as a hypothesis of the specified part tracking. The specified part tracking is performed as follows:

1. Generate two initial hypotheses $(\{n_1\}, \phi)$, $(\phi, \{n_1\})$ for the first note n_1.
2. Expand each hypothesis $\mathcal{H} = (N, \bar{N})$ on notes $n_1 \cdots n_k$ into two new hypotheses $\mathcal{H}_0 = (N \cup \{n_{k+1}\}, \bar{N})$ and $\mathcal{H}_1 = (N, \bar{N} \cup \{n_{k+1}\})$, and calculate the reliability (corresponding likelihood) of each hypothesis.
3. If the number of hypotheses exceeds a constant K, delete all hypotheses except those reliabilities are in the top K.
4. Iterate 2 and 3 for all notes.
5. After expanding hypotheses and calculating their reliability through a note list, output a hypothesis which has the maximum reliability as the result of the specified part tracking.

3. Implementation

We implemented the specified part tracking with four features. The features were classified into two: we used "Timbre Similarities to the Model" to evaluate the similarity between note n and the specified instrument; "Timbre Similarities to the Specified Part," "Proximity of Localization," and "Pitch Transition Frequency" to evaluate the similarity between n and the specified part. The latter features were designed based on Sakuraba *et al.* [4].

Timbre Similarity to the Model (f_I) This feature represents the timbre similarity between a note n and the specified instrument. The timbre of note n is described by vector $\boldsymbol{x}(n)$ proposed by Kitahara *et al.* [5]. The distance of n to the model of the specified instrument M represents the similarity between n and the instrument, but features extracted from mixed sounds are frequently distorted. We used *global model* G, which does not depend on any musical instruments, and we used the distance of n to M ($d(n, M)$) divided by the distance n to G ($d(n, G)$) to evaluate the similarity. $f_I(n)$ is described as the statistical probability calculated by an F-test:

$$f_I(n) = \int_{d_I(n)}^{\infty} \frac{\xi^{m/2-1}}{B(m/2, m/2)(\xi+1)^m} \, \mathrm{d}\xi,$$

where $d_I(n) = d(n, M)/d(n, G)$, $m = \dim(\boldsymbol{x}(n))$, and $B(m_1, m_2)$ is the Beta function.

Timbre Similarity to the Specified Part (f_S) This feature represents the timbre similarity of a note n and the part N. The timbre features are described as the same as above. We used the distance n to the distribution of the timbre features of $\tilde{n} \in N$ ($d_S(n)$) to evaluate the similarity. $f_S(n)$ is calculated by a χ^2-test:

$$f_S(n) = \int_{d_S(n)}^{\infty} \frac{\xi^{m/2-1} e^{-\xi/2}}{2^{m/2}\Gamma(m/2)} \, \mathrm{d}\xi,$$

where $\Gamma(m)$ is the Gamma function.

Localizational Proximity (f_L) This feature represents the localizational proximity of n to N. The localization of the note is the mode value of the interaural phase difference (IPD) of every frame. We used the distance of n to the distribution of the localization of $\tilde{n} \in N$ ($d_L(n)$) to evaluate the proximity. $f_L(n)$ is calculated by a χ^2-test:

$$f_L(n) = \int_{d_L(n)}^{\infty} \frac{1}{\sqrt{2\pi}} \xi^{-1/2} e^{-x/2} \, \mathrm{d}\xi.$$

Pitch Transition Frequency (f_T) This feature represents the frequency of pitch transitions. This is the trigram probability that n follows N. We used the model of the pitch transition as a trigram model in which the pitch occurrence probability depends on the pitch of the adjacent two notes ($pitch(n_{c-1})$ and $pitch(n_c)$). $f_T(n)$ is described as a posterior probability under N:

$$f_T(n) = p(pitch(n) | pitch(n_{c-1}), pitch(n_c)).$$

We used two different timbre features f_I and f_S. There is the risk that the timbre features of each musical piece and training data and then the reliability of f_I becomes lower. Even if they are not similar, the reliability of f_S keeps up because f_S compares the timbre features between the musical notes in the same musical piece.

The reliability of a hypothesis $f(\mathcal{H})$ is calculated based on multiple features as:

$$f(\mathcal{H}) = f_I(\mathcal{H}) \times \left(\sum_{i \in S,L,T} w_i f_i(\mathcal{H}) \right),$$

$$f_i(\mathcal{H}) = \sum_{n \in N} f_i(n) - \sum_{n \in \bar{N}} f_i(n).$$

f_I is not given the weight and previleged, because our aim is tracking the part that the user specified and f_I is the only feature that evaluates the timbre similarity to the instrument that the user specified.

4. Automatic Weighting of Multiple Features

After evaluation of hypotheses, optimal weights of features differ depending on the recording conditions of acoustic signals, etc. Therefore, these weights must be automatically estimated from acoustic signals. To do this, we designed a fitness of the specified part tracking. Optimal weights can be estimated by searching for weights that maximize this fitness. We defined the two following conditions to estimate the fitness of $\mathcal{H} = (N, \bar{N})$ and designed quantitative measures.

1. The number of notes derived from the specified instrument in N is greater than in $\bar{N}$. We designed the difference between N and $\bar{N}$ of the feature on the similarities of timbre to the model as:

$$E[f_I(N)] - E[f_I(\bar{N})].$$

2. The majority of notes included in N are derived from the same sound source. We designed the summation of the ratio of within-class variance to between-class variance as:

$$\sum_{i \in I,S,L,T} \frac{\left(E[f_i(N)] - E[f_i(\bar{N})]\right)^2}{Var[f_i(N)] + Var[f_i(\bar{N})]}.$$

We defined the product of these two values as the fitness of the specified part tracking. We used a genetic algorithm to search for quasi-optimal weights because theoretical calculation of the weights that maximize the fitness is difficult. The procedure of automatic weight estimation is as follows:

1. Generate initial genes randomly.
2. Track a specified part with the weights of each gene.
3. Calculate the fitness of each gene from the results of the specified part tracking.
4. Select genes by elite and roulette wheel selection.
5. Crossover between two randomly selected parents and generate a new gene which has a weight that is the mean of the weights of parents.
6. Mutate randomly selected genes into randomly calculated weights.
7. Output the weights of the gene with the highest fitness when above steps repeated L times (L is a constant.)

5. Experiment

We conducted three experiments on AMT for polyphonic music to show the effectiveness of our method:

1. We evaluated the effectiveness of automatic feature weighting. We used a trio musical signal including violin, flute and piano and tracked each instrument part.

2. We evaluated the robustness of the specified part tracking with automatic feature weighting to errors derived from automatic note formation. We used HTC [6] as a baseline method of note formation.

3. We evaluated whether automatic feature weighting can estimate appropriate weights: the estimated weights reflect the importance of the features. We tested the specified part tracking and automatic feature weighting, to see whether the weight for proximity of localization decreases according to the reliability of localization. We created the musical pieces with several reliabilities of localization by adding following deviation to the localization of each note:

   $$50 \times X \times \text{(Variance Rate of Localization)},$$

   where X is a random variable derived from $\mathcal{N}(0, 1)$.

We evaluated the accuracy F using the F-measure, which is defined as

$$P = \frac{\text{\# of notes which are correctly tracked}}{\text{\# of notes the system outputs}},$$

$$R = \frac{\text{\# of notes which are correctly tracked}}{\text{\# of notes which is on the score}} \text{ and,}$$

$$F = \frac{2 \times P \times R}{P + R}.$$

In experiment 2, correctly tracked notes mean that the notes have correct pitch and their onset time deviation is at most 10ms. We compared three feature weightings: even weights; weights estimated by our method; weights estimated by our method using the accuracy as the fitness (upper limit).

5.1. Data for Experiments

The polyphonic musical signal we used was "Auld lang syne," played for about 1 minute, which included 242 notes. This musical signal was generated by mixing audio data taken from RWC-MDB-I-2001 [7] according to a standard MIDI file (SMF) on a computer. To create the timbre model and the global model, we used mixed sound templates [5]. We used duo and trio musical pieces for mixed sound templates which were generated according to the SMFs from RWC-MDB-C-2001 (Piece Nos. 13, 16 and 17) [8]. We also used SMFs from RWC-MDB-C-2001 (Piece Nos. 1–50) to create the trigram model of pitch transition.

5.2. Experimental Results

The results of experiments 1 and 2 are listed in Tables 1 and 2, respectively. Using automatic feature weighting, we improved the accuracies from 66.0% to 69.8% in experiment 1 on average. This shows that the introduction of weights avoided incorrect part tracking (e.g., tracking the violin part even though the flute part was specified). We improved the accuracies from 44.5% to 55.3% in experiment 2 on average. This shows the robustness of our feature weighting method to errors derived from automatic note formation. The results of experiment 3 are listed in Table 3. This shows that the more a musical signal has variance of localization, the more the weight w_L decreases (i.e., appropriate weights were estimated according to the importance of features). The accuracies were also improved by feature weighting.

In experiment 1, the accuracy of the piano part decreased from 98.3% to 93.1%. This shows our feature weighting method cannot always estimate better weights than even weights. However, the results also show the number of false alarms decreased by feature weighting. This means the estimated weights can reject the notes of other part and noises derived from note formation.

It was notable that the accuracies of the flute part in experiment 1 and 2 were reversal toward the accuracies of the violin and piano parts. We assumed this as follows. The timbre features of the flute notes were often distorted in polyphonic music, because the power of the flute notes

Table 1. Results of Experiment 1

Tracking Part	F with Weights		Upper Limit of F
	Even	Estimated	
Vn	85.7%	91.9%	92.5%
Fl	14.2%	24.6%	24.6%
Pf	98.3%	93.1%	99.1%
total	66.0%	69.8%	72.1%

Table 2. Results of Experiment 2

Tracking Part	F with Weights		Upper Limit of F
	Even	Estimated	
Vn	30.6%	48.7%	50.0%
Fl	36.9%	40.7%	43.2%
Pf	66.0%	76.6%	77.6%
total	44.5%	55.3%	56.9%

Table 3. Results of Experiment 3

Variance Rate of Localization	Estimated Weights			F with Weights	
	w_S	w_L	w_T	Even	Est.
0	0.40	0.55	0.05	92.6%	97.4%
1	0.59	0.20	0.21	85.7%	91.9%
2	0.28	0.12	0.60	74.4%	81.5%

at their onset time is smaller than the power of other instruments. However, the timbre features of the flute notes were hardly distorted if its onset time varies slightly because the flute notes have gradual power envelope at onset time. The note formation detects a strong attack as the onset time, and the onset time of flute notes was estimated slightly late. The distortion of the timbre features of the flute notes caused by the onset time deviation was smaller than by mixed sounds. Therefore, the transcription was more correct with automatic note formation.

6. Conclusion

We developed the specified part tracking and automatic feature weighting, and showed that our method can estimate better weights than even weights in many cases. We also confirmed the robustness to the error derived from automatic note formation. We need to improve our feature weighting to bring the estimated weights close to optimal weights, specifically by investigating the fitness in the GA.

We did not refer to conventional methods of note formation. Since accuracies of note formation were different among the parts, the results of experiment 2 were affected by note formation. Many studies have been done on note formation, and we need to examine several note formation methods. We are also planning to evaluate more complex musical pieces (e.g., including drums and commercial CD music).

We designed four features for the specified part tracking. Specifically, we used two different features about timbre similarity because humans can often distinguish instrument sounds by previous contents of musical pieces if they have not listened to the instruments. In addition, we used only two timbre models to identify musical instruments: the model of the specified instrument and the global model. Although conventional studies on musical instrument identification have been using models of the all instruments that a musical piece contains, our method is a new approach.

Many studies on musical instrument identification require that all instruments are known. However, this approach has several weak points: when the number of instruments increases, new data of those instruments must be created, etc. By contrast, the specified part tracking is scalable on the number of instruments that the system needs to prepare the data of instruments that users want to track. Because we did not evaluate the number of instruments of musical pieces, this is part of our future work.

7. Acknowledgements

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan). This research used the RWC Music Database (Classic, Musical Instrument Sound) [7, 8], and we thank everyone who contributed this database.

References

[1] K. Kashino and H. Murase, "A Sound Source Identification System for Ensemble Music Based on Template Adaptation and Music Stream Extraction," *Speech Communication*, vol. 27, pp. 337–349, Mar. 1999.

[2] T. Kinoshita, S. Sakai and H. Tanaka, "Musical Sound Source Identification Based on Frequency Component Adaptation," *Proc. IJCAI CASA Workshop*, pp. 18–24, Aug. 1999.

[3] E. Vincent and X. Rodet, "Instrument Identification in Solo and Ensamble Music Using Independent Subspace Analysis," in *Proc. ISMIR*, pp. 576–581, 2004.

[4] Y. Sakuraba, T. Kitahara and H. G. Okuno, "Comparing Features for Forming Music Streams in Automatic Music Transcription," in *Proc. ICASSP*, vol. IV, pp. 273–276, 2004.

[5] T. Kitahara, M. Goto, K. Komatani, T. Ogata and H. G. Okuno, "Instrument Identification in Polyphonic Music: Feature Weighting with Mixed Sounds, Pitch-Dependent Timbre Modeling and Use of Musical Context," in *Proc. ISMIR*, pp. 558–563, 2005.

[6] H. Kameoka, T. Nishimoto and S. Sagayama. "Harmonic-Temporal Structured Clustering via Deterministic Annealing EM Algorithm for Audio Feature Extraction," in *Proc. ISMIR*, pp. 115–122, 2005.

[7] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," in *Proc. ISMIR*, pp. 229–230, 2002.

[8] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Popular, Classical, and Jazz Music Databases," in *Proc. ISMIR*, pp. 287–288, 2002.