COMPARING FEATURES FOR FORMING MUSIC STREAMS IN AUTOMATIC MUSIC TRANSCRIPTION

Yohei Sakuraba, Tetsuro Kitahara and *Hiroshi G. Okuno*

Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University Sakyo-ku, Kyoto 606-8501, Japan
{sakuraba, kitahara, okuno}@kuis.kyoto-u.ac.jp

ABSTRACT

In formating temporal sequences of notes played by the same instrument (referred to as *music streams*), timbre of musical instruments may be a predominant feature. In polyphonic music, the performance of timber extraction based on power-related features deteriorates, because such features are blurred when two or more frequency components are superimposed in the same frequency. To cope with this problem, we integrated timbre similarity and direction proximity with success, but left using other features as future work. In this paper, we investigate four features, *timbre similarity*, *direction proximity*, *pitch transition* and *pitch relation consistency* to clarify the precedence among them in music stream formation. Experimental results with quartet music show that direction proximity is the most dominant feature, and pitch transition is the secondary. In addition, the performance of music stream formation was improved from 63.3% by only timbre similarity to 84.9% by integrating four features.

1. INTRODUCTION

Automatic music transcription is important for many applications including music archival and music retrieval as well as for supports of composers and arrangers. It consists of two processes: note composition and music stream formation. The latter extracts a temporal sequence of notes played by the same instrument. This paper focuses on music stream formation and discusses dominant features for this formation.

In previous studies [1][2] timbre of musical instruments, which is extracted by the power envelope of the frequency component, relative power of the fundamental component, and so on, has been used for music stream formation. However when two or more frequency components are superimposed in the same frequency, these features are blurred. This makes it difficult to extract precise timbre. To solve this problem, some studies [3][4] have improved methods of extracting timbre, and other studies [5][6] have integrated other features with timbre. Eggink *et al.* [3] have used missing feature theory and Kinoshita *et al.* [4] have proposed a feature adaptation technique. Kashino *et al.* [5] have integrated music interval transitions with timbre similarity, and Sakuraba *et al.* [6] have integrated timbre similarity and direction proximity. However, the research investigating other features has not been done yet.

In this paper, to clarify the precedence among features and to improve the performance of music stream formation, four features, *timbre similarity*, *direction proximity*, *pitch transition* and *pitch relation consistency* are exploited. We evaluate which combination of four features is most useful for music stream formation.

The rest of this paper is organized as follows: Section 2 presents the four features in detail. Section 3 overviews the processing architecture. Section 4 reports experiments on music stream formation. Finally, Section 5 concludes this paper.

2. FEATURES FOR FORMING MUSIC STREAM

Music stream formation aims to generate a temporal sequence of notes played by each instrument. The main process of this formation is to determine whether a music stream s and a note n are played by the same instrument or not. To do this determination, we use four features, timbre similarity $TS(s,n)$, direction proximity $DP(s,n)$, pitch transition $PT(s,n)$ and pitch relation consistency $PRC(s,n)$. Timbre similarity is the most commonly used feature. We use a 23-dimensional feature vector used in our previous paper [6] for representing timbre. Direction proximity is a feature used in our previous paper [6]. It will be useful because each instrument is usually played at the same position from the beginning to the end of a musical piece. Pitch transition is a new feature we proposed. In tonal music, pitch transitions do not appear equally. Thus, if the transition is often seen in tonal music, we can determine the stream and the note are played by the same instrument. Pitch relation consistency is also a new feature. In general, the pitch relation (which music stream has a higher pitch) is maintained. This is suggested by the fact that in usual string quartets (1st violin, 2nd violin, viola and cello), most top and bottom notes are played by 1st violin and cello, respectively. If the pitch relation are maintained when the stream tracks the note, we can determine they are played by the same instrument.

The confidence measure of the music stream formation between s and n is defined as

$$L(s,n) = TS(s,n) \times DP(s,n) \times PT(s,n) \times PRC(s,n).$$

2.1. Timbre Similarity

Timbre similarity $TS(s,n)$ is defined by the mean of the timbre similarity between n and each note that belongs to s, that is,

$$TS(s,n) = \frac{1}{|s|}\sum_{n_i \in s} ts(n_i, n),$$

where $|s|$ represents the number of the notes of the stream s. The timbre similarity $ts(n_j, n_k)$ between two notes n_j and n_k is defined by the probability that the two notes are played by the same

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research (A), No.15200015, the Sound Technology Promotion Foundation, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan).

0-7803-8484-9/04/$20.00 ©2004 IEEE

instrument. Specifically, the difference x_{jk} of the feature vectors of the two notes is first calcluated, and then the probability defined by

$$ts(n_j, n_k) = p(\mathit{\Pi_0}|x_{jk}) = \frac{p(x_{jk}|\mathit{\Pi_0})}{p(x_{jk}|\mathit{\Pi_0}) + p(x_{jk}|\mathit{\Pi_1})},$$

are calculated. $\mathit{\Pi_0}$ and $\mathit{\Pi_1}$ are hypotheses that the two notes are played by the same instrument and by different instruments, respectively. Here, the prior probabilities of the two hypotheses are given the same value. The probability density function $p(x|\mathit{\Pi_i})$ of this distribution is defined by

$$p(x|\mathit{\Pi_i}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} exp\{-\frac{D_i^2(x;\mu_i)}{2}\}$$

where d is the number of dimensions of the feature space, μ_i is the mean of the distribution for each hypothesis $\mathit{\Pi_i}$, Σ is the covariance and D_i^2 is the squared Mahalanobis distance. D_i^2 is defined as follows:

$$D_i^2(x;\mu_i) = (x-\mu_i)^T \Sigma^{-1} (x-\mu_i),$$

where T is the transposition operator. We use the 23-dimensional feature vector used in our previous paper[6]. As training data for timbre similarity, a musical instrument sound database NTTMSA-P1, which consists of 1353 solo tones of five instruments, is used.

2.2. Direction Proximity

The direction of notes is calculated by using the two microphones in the following two steps:

1. Harmonic Structure Extraction
In every frame of the spectrogram, spectral peaks are extracted from the power spectrum. Then, the peaks which correspond to the harmonic structure are selected.

2. Localization
The interaural phase difference (IPD) of every selected peak is calculated as

$$IPD = \tan^{-1}\left(\frac{\Im[Sp(l)]}{\Re[Sp(l)]}\right) - \tan^{-1}\left(\frac{\Im[Sp(r)]}{\Re[Sp(r)]}\right)$$

where $Sp(l)$ and $Sp(r)$ are the spectra of the left and right channels, and $\Re[X]$ and $\Im[X]$ are the real and imaginary parts of X, respectively. The direction θ of the peak is calculated by

$$\theta = sin^{-1}\left(\frac{c}{2\pi f l}\,(IPD \pm 2n\pi)\right) \quad (n = 1, 2, \cdots)$$

where f, l, and c are the peak frequency, the distance between microphones, and the sonic speed, respectively. The direction of the note $D(n)$ is defined as the class mark of the highest frequency in the direction histogram of all peaks.

Direction proximity between a music stream s and a note n is defined as

$$DP(s,n) = 1 - \frac{|D(s) - D(n)|}{2\,T_d\,(D(s))},$$

where $T_d(x)$ is the threshold of the direction. $T_d(x)$ is designed in consideration of human hearing, and is defined as

$$T_d(x) = T_c + (T_o - T_c) \cdot \frac{|x|}{90},$$

where T_c and T_o are the thresholds to determine the proximity of two directions in the center (0 deg) and at either periphery (± 90 deg), respectively.

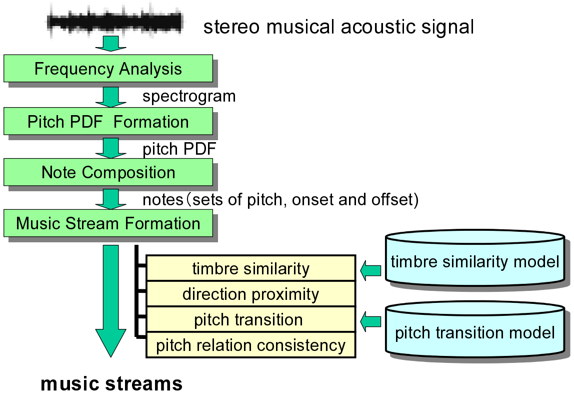

Fig. 1. Overview of our automatic music transcription system.

2.3. Pitch Transition

To obtain the pitch transition, we analyzed 50 scores in a large classical music database RWC-MDB-C-2001 [7] and generated the trigram model of the pitch by using CMU-Cambridge Toolkit. The number of notes was 167,179.

Pitch transition between a music stream s and a note n is defined as

$$PT(s,n) = p(n|n_{|s|-1}, n_{|s|}),$$

where $n_i (1 \leq i \leq |s|)$ is the ith note that belongs to s. Since the frequency of each pitch transition depends on the tonality of the musical piece, the transition was normalized in the tonality.

2.4. Pitch Relation Consistency

Pitch relation consistency $PRC(s,n)$ is defined by the mean of the pitch relation consistency between s and each stream s_i $(1 \leq i \leq N)$ that plays simulaneously with s, that is,

$$PRC(s,n) = \frac{1}{N}\sum_{i}^{N} prc(s, s_i).$$

The pitch relation consistency $prc(s, s_i)$ between two streams s and s_i is defined by the history of the pitch relation, that is,

$$prc(s, s_i) = \frac{1}{2} + (q_i - \frac{1}{2})M(t_i),$$

where q_i is the ratio of the time that s_i has higher pitch than s, to the total time they play, t_i is the time that s_i and s play (at the same time). $M(t)$ is the function whose values increase as two streams play at the same time and is defined as

$$M(t) = 1 - exp(-c \cdot t),$$

where c is the constant. Here, we set $c = 3.0$.

3. IMPLEMENTATION OF AUTOMATIC MUSIC TRANSCRIPTION SYSTEM

Fig. 1 is the flowchart of the automatic music transcription system. First, the frequency analysis analyzes, musical acoustic signals, and then spectral peaks are extracted from the power spectrum. Second, in the pitch probability density function formation stage, a probability density function (PDF) of pitch, which represents the relative dominance of every possible harmonic structure, is formed. Third, in the note composition stage, multiple agents track the temporal trajectories of salient peaks in the PDF and notes are generated. Finally in the music stream formation stage, streams of notes are generated.

3.1. Frequency Analysis

First, stereo musical acoustic signals (sample rate: 48kHz) are analyzed by short time Fourier transform, with Hamming windows (4,096 points) for every 21ms, and then spectral peaks are extracted from the power spectrum. The PDF $p^{(t)}(x)$ of the frequency components is defined as follows:

$$p^{(t)}(x) = \frac{\psi^{(t)}(x)}{\int_{-\infty}^{\infty} \psi^{(t)}(x)dx}$$

where $\psi^{(t)}(x)$ is the power spectrum at frequency x (in Hz).

3.2. Pitch Probability Density Function Formation

In order to represent how predominant each harmonic structure is in the power spectrum, the PDF of the pitch is calculated according to [8]. We assume that the observed PDF $p^{(t)}(x)$ consists of a weighted mixture of harmonic-structure tone models. The tone model indicates where the harmonics of the pitch (e.g. C4, G5) tend to occur. When the PDF of each tone model whose pitch is F is denoted as $p(x|F)$, the mixture density $p(x;\theta^{(t)})$ is defined as

$$p(x;\theta^{(t)}) = \int_{F_L}^{F_H} w^{(t)}(F)p(x|F)dF,$$

$$\theta^{(t)} = \{w^{(t)}(F)|F_L \leq F \leq F_H\},$$

where F_L and F_H denote the lower and upper limits of the possible pitch range and $w^{(t)}(F)$ is the weight of a tone model $p(x|F)$ that satisfies

$$\int_{F_L}^{F_H} w^{(t)}(F)dF = 1.$$

If we can estimate the model parameter $\theta^{(t)}$ such that $p^{(t)}(x)$ is likely to have been generated from $p(x;\theta^{(t)})$, $p^{(t)}(x)$ can be considered to be decomposed into harmonic-structure tone models and $w^{(t)}(F)$ can be interpreted as the PDF of pitch:

$$p^{(t)}_{\text{pitch}}(F) = w^{(t)}(F) \quad (F_L \leq F \leq F_H)$$

Therefore the problem to be solved is to estimate the model parameter $\theta^{(t)}$ when we observe $p^{(t)}(x)$. The maximum likelihood estimator of $\theta^{(t)}$ is obtained by maximizing the mean log-likelihood defined as

$$\int_{-\infty}^{\infty} p^{(t)}(x) \log p(x;\theta^{(t)})dx.$$

For this maximization, the Expectation-Maximization (EM) algorithm is used. By introducing a hidden variable F describing which tone model was responsible for generating each observed frequency component at x, we can specify the two steps as follows:

1. **E-step**: Compute the following conditional expectation of the mean log-likelihood:

$$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} p^{(t)}(x)E_F[\log p(x,F;\theta^{(t)})|x;\theta'^{(t)}]dx.$$

   where $E_F[a|b]$ denotes the conditional expectation of a with respect to the hidden variable F with the probability distribution determined by the condition b.

2. **M-step**: Maximize Q($\theta^{(t)}|\theta'^{(t)}$) as a function of $\theta^{(t)}$ to obtain $\overline{\theta^{(t)}}$:

$$\overline{\theta^{(t)}} = \arg\max_{\theta^{(t)}} Q(\theta^{(t)}|\theta'^{(t)}).$$

Here, we need to assume $p(x|F)$ indicates where the harmonics of the pitch F tend to occur. We use the same model as the literature [8].

3.3. Note Composition

To select the pitch trajectory that is dominant and stable from the viewpoint of global pitch estimation, the method sequentially tracks peak trajectories in the temporal transition of the PDF of pitch, and outputs the notes that are the dominant and stable trajectories. A multiple-agent architecture[8] is used to track the pitch trajectories. It consists of a salience detector and multiple agents. The salience detector picks up salient peaks in the PDF of pitch, and agents track their trajectories according to the peaks. Each agent has its pitch, a confidence measure of the trajectory CM_{Note} and a penalty. These values are updated as follows.

1. The salience detector picks up salient peaks of $p^{(t)}_{\text{pitch}}(F)$ that are higher than the dynamic threshold adjusted according to the maximum peak.
2. The salient peaks are allocated to the agent that has the same pitch. If the salient peaks have not been allocated, a new agent for tracking it is generated.
3. Each agent has a penalty, and an agent whose penalty exceeds the threshold is terminated. An agent to which a salient peak has not been allocated or which cannot find its next peak in the PDF of pitch is penalized. The penalty of the agent to which a peak is allocated is reset to 0.
4. Each agent evaluates its own confidence measure, that is, the mean of the allocated peak.
5. The agents whose confidence measure exceeds a threshold T_{CM} are outputted as notes.

3.4. Music Stream Formation

In this stage, streams of notes are generated. To implement, we adopt a multiple-agent architecture that enables the tracking process to be controlled dynamically and flexibly. It consists of the salience detector and multiple agents that are dynamically generated and terminated. The salience detector picks up salient notes in the note hypotheses (that are generated in the previous stage), and agents track their trajectories by integrating the four features, that is, timbre similarity, direction proximity, pitch transition and pitch relation consistency.

Each agent has its confidence measure of the trajectory CM_{Stream} and a penalty. They behave at each block (the time corresponds to the 32nd notes) as follows.

1. The salience detector picks up the notes that are higher than the threshold T_{CM}.
2. The agent interact to allocate the notes among themselves according to the confidence measure of the music stream formation $L(s,n)$. If more than one agent claims the same note, the note is allocated to the agent that has highest CM_{Stream}. If the note has not been allocated, a new agent for tracking it is generated.

Table 1. Music used to evaluate the system

Title	Instruments	Playing time	#Notes
Pachelbel's Canon	Vn, Fl, Tp, Pf	6 min 30 s	5,868
Auld Lang Syne	Vn, Fl, Pf	1 min	726

Vn: Violin, Fl: Flute, Tp: Trumpet, Pf: Piano

3. Each agent has a penalty, and an agent whose penalty exceeds a threshold is terminated. An agent to which a note has not been allocated or which cannot find its next note is penalized. The penalty of the agent to which a note is allocated is reset.
4. Each agent evaluates its own confidence measure CM_{Stream} that is the mean of CM_{Note} of the allocated notes.
5. The agents are outputted as music streams.

4. EXPERIMENTAL RESULTS

To evaluate the improvement of the music stream formation performance, stereo musical acoustic signals are used. The benchmark was the quartet music in Pachelbel's Canon (*Canon*) and trio music in Auld Lang Syne (*ALS*), listed in Table 1. The music was played via four (or three) loud speakers using a MIDI sampler with real instruments sound database and recorded by two microphones (baseline: 20cm) in an anechoic room. The layout of the instruments was violin, flute, trumpet and piano from left to right in *Canon*, and flute, violin, piano in *ALS*. The current implementation uses the following parameter values: F_L = D2, F_H = F#6, T_c = 10 deg., T_o = 20 deg. and T_{CM} = 0.06.

The performance of the note composition was evaluated by recall rate (R) and precision rate (P).

$$R = \frac{\#\text{correctly generated notes}}{\#\text{actual notes}} \quad P = \frac{\#\text{correctly generated notes}}{\#\text{generated notes}}$$

This is because there are two types of errors. The first type is caused by generating a note that does not exist in the score, and the second type is caused by qnot generating a note that actually exists in the score. The note was determined correct when it had the same pitch as the score and its onset time error was less than a 32nd note. R and P were 66.4% and 76.0%, respectively.

To evaluate music stream formation, the notes that belong to the same music stream in the score are determined correct. The performance of the music stream formation is evaluated by

$$R = \frac{\#\text{correctly formed notes}}{\#\text{actual notes outputted in note composition}}.$$

The accuracies are listed in Table 2. Direction proximity was the most effective feature for music stream formation, and pitch transition is the secondary. The highest performance was the case of using timbre similarity, direction proximity and pitch transition. Timbre similarity was not effective in the case of *ALS*. This is because the features which represent the precise timbre are not clarified in the case of polyphonic music.

5. CONCLUSIONS

In this paper, to improve the performance of music stream formation, we integrated four features, that is, timbre similarity, direction proximity, pitch transition and pitch relation consistency. Experimental results with quartet music showed that direction proximity is the most dominant feature, and pitch transition is the second dominant feature. The performance of music stream formation improved from 63.3% to 84.9%. Future work includes correction of note composition error through the use of the music stream.

Table 2. The results of music stream formation

feature1	feature2	feature3	feature4	*Canon*	*ALS*
○	—	—	—	63.3%	79.9%
—	○	—	—	77.4%	84.9%
—	—	○	—	66.5%	75.2%
—	—	—	○	57.0%	63.8%
○	○	—	—	77.2%	84.2%
○	—	○	—	66.5%	79.5%
○	—	—	○	62.4%	80.2%
—	○	○	—	84.4%	91.6%
—	○	—	○	77.5%	84.6%
—	—	○	○	66.6%	76.5%
○	○	○	—	85.0%	90.3%
○	○	—	○	77.4%	84.9%
○	—	○	○	66.5%	79.2%
—	○	○	○	84.6%	91.6%
○	○	○	○	84.9%	90.3%

feature1: timbre similarity feature2: direction proximity
feature3: pitch transition feature4: pitch relation consistency

Acknowledgments: We thank everyone who has contributed to building and distributing the RWC Music Database: RWC-MDB-C-2001) [7]. We also thank Prof. Haruhiro Katayose, Dr. Kunio Kashino and Dr. Kazuhiro Nakadai for their valuable comments.

6. REFERENCES

[1] G. J. Brown and Martin Cooke: Perceptual Grouping of Musical Sounds: A Computational Model, *Journal of New Music Research*, pp.107-132, 1994.

[2] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka: Application of the Bayesian Probability Network to Music Scene Analysis, *Computational Auditory Scene Analysis*, D.F. Rosenthal and H. G. Okuno (eds.), Lawrence Erlbaum Associates, pp.115–137, 1998.

[3] J. Eggink and G. J. Brown: A Missing Feature Approach to Instrument Identification in Polyphonic Music, *Proc. of ICASSP*, pp.553–556, 2003.

[4] T. Kinoshita, S. Sakai, H. Tanaka: Musical Sound Source Identification Based on Frequency Component Adaptation, *Proc. of IJCAI*, pp.18–24, 1999.

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

[6] Y. Sakuraba and H. G. Okuno: Note Recognition of Polyphonic Music by Using Timbre Similarity and Direction Proximity, *Proc. of ICMC*, pp.167–170, 2003.

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

[8] Masataka Goto: A Robust Predominant-F0 Estimation Method for Real-time Detection of Melody and Bass Lines in CD Recordings, *Proc. of ICASSP*, pp.757–760, 2000.