NOTE RECOGNITION OF POLYPHONIC MUSIC BY USING TIMBRE SIMILARITY AND DIRECTION PROXIMITY

Yohei Sakuraba and *Hiroshi G. Okuno*

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

ABSTRACT

Note recognition in automatic music transcription consists of two processes: *simultaneous grouping* and *sequential grouping*. The former generates a note from frequency components, while the latter generates a temporal sequence of notes. Their main problem are disambiguation in note composition and design of features effective for music stream creation, respectively. For the simultaneous grouping, to cope with the problem note hypotheses are created based on overlap detection of frequency components. For the sequential grouping, timbre similarity and direction proximity are integrated. The result of experiments with quartet music recorded in an anechoic chamber showed that the proposed method improved the F-measure of each grouping by 0.10 and 0.14, respectively.

1. INTRODUCTION

Note recognition of polyphonic music is an important technology for many applications including automatic music transcription, auditory scene analysis, automatic arrangement, and automatic tagging for MPEG-7 as well as for supports for composers and arrangers.

Note recognition in automatic music transcription of polyphonic music consists of two processes: *simultaneous grouping* and *sequential grouping* [1]. The simultaneous grouping forms note candidates from frequency components, while the sequential grouping forms streams of notes candidates. Since each process cannot recognize notes without ambiguities due to common frequency components of polyphonic music, disambiguation of note recognition is critical in automaticmusic transcription.

Ambiguities in the simultaneous grouping are caused by polyphony, for example, when multiple instruments play at the same time. Usually, *harmonicity* is the most dominant feature in simultaneous grouping. However, more than one note may share some overtones, which makes it more difficult to detect fundamental frequency. Consider, for example, that notes C4 (262 Hz) and C5 (524 Hz) are simultaneously played. Spectral peaks extracted from the power spectrum are around 262 Hz, 524 Hz, 786 Hz, 1048 Hz, and so on. Since these spectral peaks are similar to those of C4, it is difficult to recognize a note of C5.

This kind of ambiguities due to common overtones is caused not only by notes in the octave relation (e.g. C4 and C5), but also by any combination of polyphonic music. Although the ambiguity is inevitable in the simultaneous grouping, most studies on automatic music transcription have not clarified this point [2]–[6].

Ambiguities in the sequential grouping are extensively studied by psychophysics [1]. The chord, rhythm, and timbre of musical instruments are the main clues for sequential grouping. However, when two or more components are superimposed in the same frequency, these features are blurred and thus extracting features precisely is difficult.

In this paper, to improve the performance of the simultaneous grouping, the overlap probability of frequency components is introduced and notes are formed based on this probability. In addition, to improve the performance of the sequential grouping, timbre similarity and direction proximity of notes are exploited.

2. SIMULTANEOUS GROUPING

The simultaneous grouping consists of the three stages as is shown in **Fig. 1**. This system assumes stereo musical acoustic signals as input. First, in the *frequency analysis and localization* stage, frequency components are extracted from input signals, and the direction of frequency components is obtained by calculating the interaural intensity difference (IID) and the interaural phase difference (IPD). In the *overlap estimation* stage, the overlap probability of each frequency component is obtained based on the variation of direction. Finally, in the *note hypothesis creation and evaluation* stage, note hypotheses are created based on the harmonicity with overlap probability. Then the notes of the highest score based on the overlap probability are created.

2.1. Frequency Analysis and Localization

Musical acoustic signals are first analyzed by short time Fourier transform, with Hamming windows (4096 points), and then spectral peaks are extracted from the power spectrum. Frequency components are obtained by connecting the peaks that have the same MIDI note number. Direction θ is obtained by using the IID and IPD of each peak [7].
(1) The IID and IPD are calculated as follows:

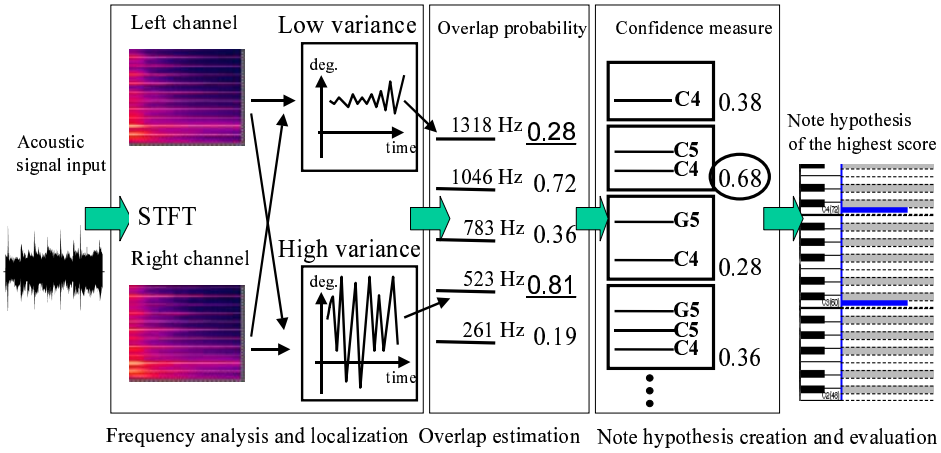

Fig. 1. Simultaneous grouping consists of the three stages.

$$IID = \frac{\sqrt{\Re[Sp(l)]^2 + \Im[Sp(l)]^2}}{\sqrt{\Re[Sp(r)]^2 + \Im[Sp(r)]^2}}$$

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

where $Sp(l)$ and $Sp(r)$ are the spectrum of the left and the right channel, and $\Re[X]$ and $\Im[X]$ are the real and imaginary part of X, respectively.
(2) The IID is used to determine whether a peak is at the center, on the left or right because a pair of microphones with the baseline of 20cm do not give a strong IID.
(3) θ is calculated by

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

where f, l, and c are the peak frequencies, the distance between microphones, and the sonic speed, respectively.

2.2. Overlap Estimation

The overlap probability of each frequency component is calculated by using the variance of direction of a peak for each frequency component obtained in the previous stage. The variance of direction is small if no overtones of other harmonics are superimposed on that peak, while large otherwise.

Since the variance of direction is represented by the standard deviation (σ) and the gradient of the least-square fitting of direction of each peak (g), the overlap probability is obtained by integrating all overlap probabilities based on the Dempster-Shafer theory. Dempster's rule of combination is represented as follows [8]:

$$m(A_k) = \frac{\sum_{A_i \cap A_j = A_k} m_1(A_i) m_2(A_j)}{1 - \sum_{A_i \cap A_j = \phi} m_1(A_i) m_2(A_j)}$$

where m_1 and m_2 are the basic probabilities of a hypothesis reasoned from a proof independent of each other, and A_i and A_j are the focal elements.

One hypothesis that a frequency component is superimposed, and a second, that it is not, are represented by S and $\overline{S}$, respectively. The three basic probabilities are defined as follows:

$m(S)$ basic probability of S.
$m(\overline{S})$ basic probability of $\overline{S}$.
$m(S, \overline{S})$ basic probability of unknown whether S or $\overline{S}$

The basic probalilities, m_σ and m_g, are defined for σ and g of the variance of direction, respectively. They are defined as follows:

$$m_x(S) = r \times \{1 - SIG_{T_x}(x)\} \quad x = \sigma, g$$
$$m_x(\overline{S}) = r \times SIG_{T_x}(x)$$
$$m_x(S, \overline{S}) = 1 - r$$
$$r = \frac{\text{number of peaks with direction} - 1}{\text{number of peaks belonging to frequency component} - 1}$$

$SIG_T(x)$ is the sigmoid function with the threshold T with $SIG_T(T)$=0.5 and $SIG_T(0)$=0.99. The thresholds T_σ and T_g are set to 5 degrees and 9.6 degrees/second, respectively.

The overlap probability of the frequency component is obtained by integrating m_σ and m_g based on Dempster's rule of combination. The plausibility P^* and belief function P_* are obtained from $m(S)$, $m(\overline{S})$, and $m(S, \overline{S})$ as follows:

$$P^* = m(S) + m(S, \overline{S}), \qquad P_* = m(S)$$

The integrated overlap probability is defined as follows:

$$P = \frac{1}{2}(P^* + P_*)$$

2.3. Note Hypotheses Creation and Evaluation

Note hypotheses are created based on harmonicity and overlap estimation. Even if most overtones of a note overlap other tones, note hypotheses for it are also created in order to continue further hypothetical reasoning. The score of likelihood of note hypothesis L is defined by using the overlap probability $P(f_i)$ and direction $Pan(f_i)$ as follows:
(1) if frequency component f_i is overlapped,

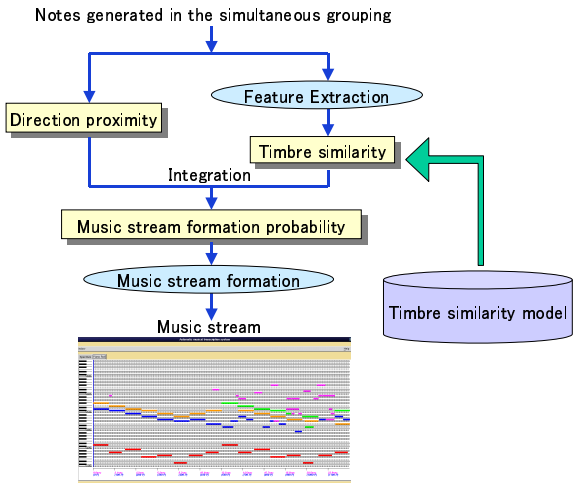

Fig. 2. Sequential grouping

$Score(f_i) = P(f_i)$

(2) if frequency component f_i is not overlapped,

$Score(f_i) = SIG_{T_\sigma}(|Pan(f_i) - Pan(n)|)$

where n is note with direction closest to f_i.

The note(s) with the greatest score L (average of score(f_i)) is given to the sequential grouping to form music streams.

3. SEQUENTIAL GROUPING

The procedure of the sequential grouping is shown in **Fig. 2.** Notes obtained in simultaneous grouping are the input. The system calculates the *music-stream probability* by integrating the timbre similarity and the direction proximity based on the Dempster-Shafer theory. Based on the probability, a music stream is formed from the notes.

3.1. Music Stream Probability

One hypothesis that two notes belong to the same music stream, and a second, that they belong to a different music stream, are represented by P and $\overline{P}$, respectively. The three basic probabilities are defined as follows.

$m(P)$ basic probability of P.
$m(\overline{P})$ basic probability of $\overline{P}$.
$m(P, \overline{P})$ basic probability of unknown whether P or $\overline{P}$

The m_t and m_p are the basic probabilities representing the timbre similarity and the direction proximity, respectively.

The key to music stream formation are timbre similarity and direction proximity. We think that it is reasonable to use, because musical instruments are moved seldom or within a narrow range.

For timbre similarity, the discrimination function is designed by using the support vector machine (SVM) with linear function as a kernel. A set of 23 features for the SVM include the spectral and temporal features which are designed according to the literature [9, 10].

The variables m_t and m_p are defined as follows:

$$m_t(P) = \int_{-\infty}^{SVMscore} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

$$m_t(\overline{P}) = \int_{SVMscore}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

$$m_t(P, \overline{P}) = 0$$

$$m_p(P) = r_1 \times r_2 \times SIG_{T_p}(|Pan(n_1) - Pan(n_2)|)$$

$$m_p(\overline{P}) = r_1 \times r_2 \times \{1 - SIG_{T_p}(|Pan(n_1) - Pan(n_2)|)\}$$

$$m_p(P, \overline{P}) = 1 - r_1 \times r_2$$

where Pan is the direction of the note. The variables r_1 and r_2 are the reliability of the direction of notes n_1 and n_2, respectively, and they are defined as the sum of the overlap probability of the note's frequency components. T_p is designed as follows: the value is large in the center and small outside the center.

$$T_p = sin^{-1}\left(sin(Pan(n)) + \frac{c}{l} \cdot \frac{4.0}{f_s}\right) - Pan(n).$$

The variables c, l, and f_s are the sonic speed, the distance between the microphones, and the sample rate, respectively. The music-stream probability is calculated from the plausibility and belief function based on the Dempster-Shafer theory.

3.2. Music Stream Formation

Music-stream probability $P(S_i, n)$ $(1 \leq i \leq C)$ of music stream S_i with input note n is defined as follows:

$$P(Part_i, n) = \frac{1}{c_i} \sum_{j=1}^{c_i} P(n_{ij}, n)$$

where c_i is the number of notes in S_i. The music stream is formed according to the following algorithm:

1. Calculate probability $P(S_i, n)$ of music stream S_i with an input note n.
2. A note n is added in the stream S_i where $P(S_i, n)$ has the maximum value.
3. If the maximum value is less than 0.5, a new music stream is generated.

The above process is repeated for all notes so that all notes are grouped into some music streams.

4. EXPERIMENTS AND RESULTS

To evaluate the improvement of note recognition performance, stereo musical acoustic signals recorded in an anechoic chamber were used. The benchmark was the quartet in Pachelbel's Canon whose playing time was about 6 minutes and 30 seconds. The quartet music was played via four loud speakers using a MIDI sampler with musical instrument sound database NTTMSA-P1. Two microphones with the baseline of 20cm were used for recording. The distance from each loud speaker to the center of the pair of microphones was 1m.

The layout of the musical instruments (loud speakers) from left to right was a violin, a flute, a trumpet, and a piano. Three interval patterns were used for the instruments;

Table 1. Results of simultaneous grouping (class 1)

interval	without overlap probability	with overlap probability
20 deg.	0.63 (0.46, 0.98)	0.78 (0.76, 0.79)
40 deg.	0.63 (0.46, 1.00)	0.76 (0.80, 0.72)
60 deg.	0.60 (0.43, 0.97)	0.61 (0.66, 0.57)

F-measure (recall, precision)

Table 2. Results of simultaneous grouping (class 2)

interval	without overlap probability	with overlap probability
20 deg.	0.44 (0.31, 0.79)	0.58 (0.51, 0.66)
40 deg.	0.44 (0.31, 0.79)	0.57 (0.53, 0.62)
60 deg.	0.45 (0.31, 0.82)	0.49 (0.41, 0.61)

F-measure (recall, precision)

Table 3. Results of whole process (class 1)

interval	timbre similarity	direction proximity	Both
20 deg.	0.65 (0.64, 0.65)	0.68 (0.68, 0.69)	0.71 (0.70, 0.71)
40 deg.	0.60 (0.63, 0.56)	0.66 (0.71, 0.63)	0.69 (0.73, 0.65)
60 deg.	0.52 (0.56, 0.48)	0.52 (0.57, 0.49)	0.56 (0.61, 0.53)

F-measure (recall, precision)

Table 4. Results of whole process (class 2)

interval	timbre similarity	direction proximity	Both
20 deg.	0.40 (0.35, 0.46)	0.50 (0.45, 0.57)	0.51 (0.45, 0.58)
40 deg.	0.35 (0.32, 0.39)	0.49 (0.46, 0.53)	0.50 (0.46, 0.54)
60 deg.	0.34 (0.28, 0.43)	0.34 (0.28, 0.44)	0.38 (0.31, 0.50)

F-measure (recall, precision)

Table 5. Results of sequential grouping

interval	timbre similarity	direction proximity	Both
20 deg.	0.84, 0.69	0.89, 0.88	0.92, 0.88
40 deg.	0.79, 0.60	0.89, 0.87	0.91, 0.87
60 deg.	0.85, 0.67	0.86, 0.68	0.92, 0.76

Left and right column are results of classes 1 and 2.

the intervals of 20, 40, and 60 degrees. The performance of the simultaneous grouping was evaluated by comparing the method with and without the overlap probability, while that of the sequential grouping was evaluated by comparing the results of first using only timbre similarity, next only direction proximity, and finally integrating the two. The accuracy was measured by the F-measure obtained by the recall rate (R) and the precision rate (P) as follows:

$$R = \frac{\text{number of correctly generated notes}}{\text{actual number of notes}}$$

$$P = \frac{\text{number of correctly generated notes}}{\text{number of generated notes}}$$

$$\textit{F-measure} = \frac{2 \times R \times P}{R + P}$$

In the simultaneous grouping, a note that has the correct note name and the correct onset time is defined *correct*. In the sequential grouping, a note that has the correct note name, the correct onset time, and the correct music stream is defined *correct*. If the onset-time error is less than a thirty-second note, the onset time is defined *correct*. If the notes in the output music stream originate in the same music stream in the score, the music stream is defined *correct*.

Since the performance is affected severely by fast tempo playing, the benchmark is divided into two classes. Musical measures without sixteenth notes are classified as *class 1*, while the others are classified as *class 2*. Musical measures with thirty-second notes are not included in the evaluation. The experimental results for the simultaneous grouping are shown in **Tables 1** and **2**. Results for the sequential grouping after simultaneous grouping are shown in **Tables 3** and **4**.

In the simultaneous grouping, the proposed method improved the F-measure by 0.10 on average. The F-measure for class 2 was poorer than that of class 1, because variation of direction of notes with short duration tend to be influenced by other notes. Coping with the problem of short-duration notes will be an important subject for future work. In the sequential grouping, the proposed method performed better than using one similarity. The results for the sequential grouping are shown in **Table 5**. The proposed method improved the accuracy in classes 1 and 2 to 91% and 83% on average, respectively.

5. CONCLUSION

In this paper, disambiguation in the simultaneous and sequential grouping of note recognition was attained by timbre similarity and direction proximity. The experimental results with quartet recordings in an anechoic chamber showed that the proposed method is quite effective in both groupings. Future work includes direct interaction between the simultaneous and sequential grouping. Such interaction may be triggered by the chord, melody line, and note transition probability.

6. REFERENCES

[1] A.S. Bregman: *Auditory Scene Analysis*, MIT Press, 1990.

[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] 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.

[4] R. Mani and S. H. Nawab: Integration of DSP Algorithms and Musical Constraints for the Separation of Partials in Polyphonic Music, *Proc. of ICASSP*, pp.1741–1744, 1998.

[5] A. Klapuri, T. Virtanen, A. Eronen, and J. Seppänen: Automatic transcription of musical recordings, *CRAC-01*, 2001.

[6] K.D. Martin: Automatic Transcription of Simple Polyphonic Music: Robust Front End Processing, MIT Media Lab Perceptual Computing Technical Report #399, 1996.

[7] K. Nakadai, H.G. Okuno, and H. Kitano: "Real-Time Sound Source Localization and Separation for Robot Audition", *Proc. of ICSLP*, pp.193–196, 2002.

[8] G. Shafer: *A Mathematical Theory of Evidence*, Princeton Univ. Press, 1976.

[9] A. Eronen and A. Klapuri: Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features, *Proc. of ICASSP*, pp.753–756, 2000.

[10] T. Kitahara, M. Goto, and H.G. Okuno: Musical Instrument Identification Based on F0-dependent Multivariate Normal Distribution, *Proc. of ICASSP, to appear*, 2003.