IPSJ Journal Vol. 50 No. 3 1100–1110 (Mar. 2009)

Regular Paper

Novel Tonal Feature and Statistical User Modeling for Query-by-Humming

Motoyuki Suzuki,[†1] Takuto Ichikawa,[†2]
Akinori Ito[†2] and Shozo Makino[†2]

This paper describes a query-by-humming (QbH) music information retrieval (MIR) system based on a novel tonal feature and statistical modeling. Most QbH-MIR systems use a pitch extraction method in order to obtain tonal features of an input humming. In these systems, pitch extraction errors inevitably occur and degrade the performance of the system. In the proposed system, a cross-correlation function between two logarithmic frequency spectra is calculated as a tonal feature instead of a difference of two successive pitch frequencies, and probabilistic models are prepared for all tone intervals existing in the database. The similarity scores between an input humming and musical pieces in a database are calculated using the probabilistic models. The advantages of this system are that it can obtain more appropriate tonal features than the pitch-based method, and it is also robust against inaccurate humming by the user thanks to its statistical approach. From experimental results, the top-1 retrieval accuracy given by the proposed method was 86.8%, which was more than 10 points higher than the conventional single pitch method. Moreover, several integration methods were applied to the proposed method with several conditions. The majority decision method showed the highest accuracy, and 5% reduction of retrieval error was obtained.

1. Introduction

Recently, several MIR (Music Information Retrieval) systems using QbH (Query by Humming) have been developed (e.g., CubyHum [1], Super MBox [2]). In these systems, an input humming is split into notes, and then the pitch and duration of the notes are extracted. Musical features (tone interval and inter-onset interval) are calculated from the pitch and duration information, and the musical features are compared with the features of musical pieces in a database.

†1 Institute of Technology and Science, The University of Tokushima
†2 Graduate School of Engineering, Tohoku University

One of the main problems of the conventional QbH-MIR systems is pitch extraction errors. Such errors, especially double-pitch and half-pitch errors, cannot be avoided even if we use an accurate method such as Yin [3] or Praat [4]. Tone interval is one of the most popular tonal features, and it is calculated as the difference between the pitches of two successive notes. Therefore, when a pitch extraction error on a note occurs, two tone intervals concerning the note become incorrect. This is a critical problem that directly degrades the performance of the system.

In order to overcome this problem, a QbH-MIR system using multiple pitch candidates has been developed [5]. This system extracts multiple pitch candidates of each note from the input humming, and matches them with melodies in the database considering all combinations of the pitch candidates. When three pitch candidates were used, the accuracy of pitch extraction became 99.7%, and the retrieval accuracy increased from 74.2% to 86.5%. However, the retrieval time became about n^2 times longer (where n denotes the number of pitch candidates) than the conventional QbH-MIR systems because the system must consider all combinations of pitch candidates. Another solution to avoid pitch extraction errors without incurring much computational effort is required.

We therefore consider what kind of information is *really* needed by QbH-MIR systems. QbH-MIR systems do not need "pitches" but need "tone intervals". In conventional systems, a tone interval is calculated using pitch information of two successive notes. If a tone interval could be calculated directly from the input humming, pitch extraction would not be needed. By considering only the relative pitch, the tone interval feature is expected to be more robust against half-pitch or double-pitch errors than the method based on pitch extraction.

Another problem is that most conventional MIR systems do not consider inaccurate humming by the user, yet most users are not professional singers and cannot hum very accurately. In other words, there are fluctuations in the user's humming, and the amount of fluctuation depends on the tone interval. For example, even if a user can hum the tone interval "+200 cent" stably, he or she might hum the tone interval "+1,300 cent" with a large fluctuation.

However, the conventional MIR systems do not consider the fluctuation in the user's humming. In these systems, the registered songs are represented by

sequences of tone intervals numerically such as "+200 cent", or symbolically such as "Up" and "Down". The similarity between the input humming and a song in the database is then calculated based on Euclidean or binary distance. These distances cannot deal with the fluctuation in the user's humming.

In order to solve this problem, Shih, et al.[6] proposed a statistical approach for automatic transcription of humming data. In this method, time sequences of pitch frequencies are extracted from the input humming, and Hidden Markov Models (HMM) are trained using these pitch sequences. Each HMM corresponds to each tone interval, and after the HMMs have been trained, the input humming is decoded using the Viterbi algorithm. The decoding process is similar to a speech recognition algorithm based on HMMs.

In this method, pitch sequences in the database are represented as a concatenation of the HMMs. All states in HMMs have a two-mixture Normal distribution as an output probability distribution. This method can consider the fluctuation of the humming because HMMs can represent the fluctuation as a variance of the Normal distribution. While the method is intended for automatic transcription of humming, it is also useful for a QbH-MIR system.

In this paper, we propose a QbH-MIR system that does not use pitch extraction. In this system, a cross-correlation function between log-scaled spectra of two successive notes is used as a tonal feature instead of the difference of pitch frequencies. The system can extract a tonal feature more robustly than a single pitch method, and it requires less computational effort than the multiple pitch method.

Statistical models are also introduced to represent the tonal feature, and similarities between the input humming and registered songs are calculated based on statistical distance.

2. A New Tonal Feature without Pitch Extraction

As mentioned above, a QbH-MIR system does not need the absolute pitch frequency of a note. Rather, the system actually needs a tone interval, which is a log-scaled ratio of the pitch frequencies of two successive notes. In this section, a new tone interval feature is proposed. It is derived directly from the spectrum of two notes without determining the absolute pitch frequencies.

2.1 Peak Frequency of Cross-correlation Function

The spectrum of a vowel part of an input humming has several peaks, which correspond to a fundamental frequency (F_0) and its integral multiples. **Figure 1** shows an example of a vowel spectrum. If an F_0 becomes higher, the spectrum is stretched along the linear frequency axis. When the x-axis represents the logarithm of frequency, a change of F_0 causes translation of the whole spectrum. In this case, the amount of translation corresponds to the tone interval of two fundamental frequencies. The amount of translation can be estimated by extracting the peak of the cross-correlation function between two log-scaled spectra[7].

Let the input signal be $x(t)$, and its power spectrum be $X(w)$. Then the log-frequency power spectrum is $X(\xi)$, where $\xi = \log_2(w)$. Next, the signal $y(t)$ that is 1200α cent higher than $x(t)$ can be approximated as $Y(w) \approx X(2^{-\alpha}w)$, and its log-frequency power spectrum is $Y(\xi) \approx X(\xi - \alpha)$. By calculating a cross-correlation function of X and Y, we obtain

$$C_{XY}(l) = \sum_{\xi} X(\xi)Y(\xi + l) \approx \sum_{\xi} X(\xi)X(\xi - \alpha + l) \quad (1)$$

which has a peak at $l = \alpha$.

Figure 2 shows an example of a cross-correlation function. In this figure, the

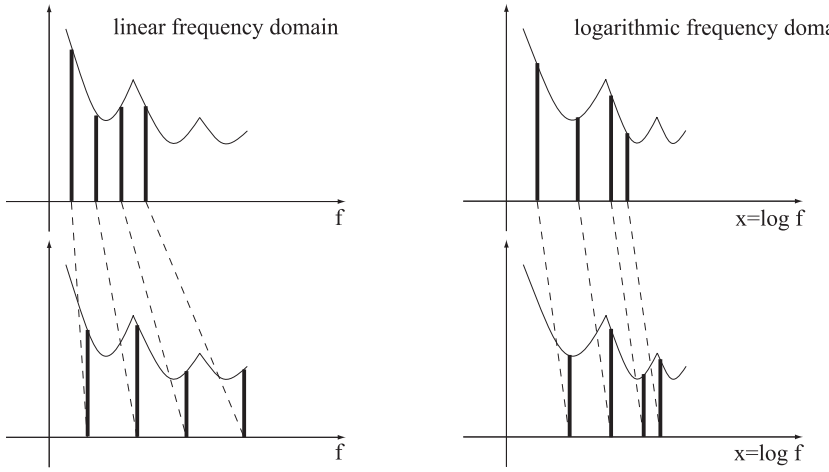


Fig. 1 Vowel spectra with different fundamental frequencies.

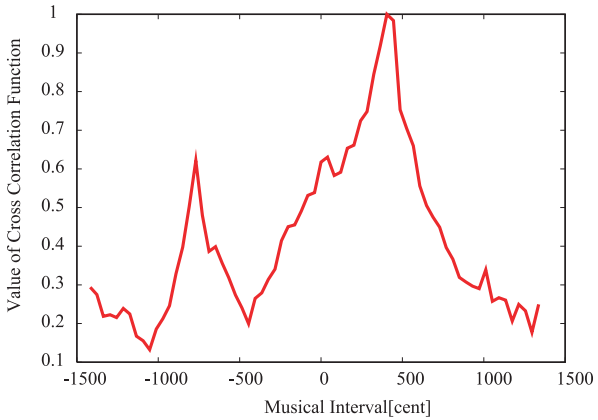


Fig. 2 An example of cross-correlation function.

vertical axis represents the "normalized" correlation value, which is calculated by dividing the correlation value by the maximum correlation value. This example was calculated from two hummed notes with a +400 cent interval. It has two large peaks around +400 cent and −800 cent. The peak around +400 cent corresponds to the correct tone interval, and the peak around −800 cent corresponds to the tone interval when the pitch frequency of the previous or following note is estimated with a double or half pitch error. It can be seen that the wrong peak is much smaller than the correct peak, demonstrating the robustness of the proposed feature against double/half pitch errors. In the proposed system, the highest peak is extracted from $C_{XY}(l)$, and is used as a tonal feature.

2.2 Whole Shape of Cross-correlation Function

We also tried to use a whole $C_{XY}(l)$ shape as a tonal feature instead of the highest peak of $C_{XY}(l)$.

The spectrum of a vowel is expressed as a multiplication of the spectrum of the glottal sound source (vertical impulses in Fig. 1) and the transfer function of a vocal tract (wave line in Fig. 1). The spectrum of the glottal sound source depends on F_0, whereas the transfer function does not. Therefore, the two shapes of log-frequency spectra corresponding to two different fundamental frequencies are not strictly identical, which violates the assumption of Eq. (1). Although the difference is considered to be small, it may cause an error in extracting the tone interval from $C_{XY}(l)$. On the other hand, the whole shape of $C_{XY}(l)$ seems to be more robust against the difference. In order to express the tonal features more robustly, we also propose a new tonal feature based on the whole shape of $C_{XY}(l)$.

At first, the values of $C_{XY}(l)$ are represented as a vector. Each dimension of the vector $\boldsymbol{v}$ is defined as the value of $C_{XY}(l)$ at each l. In the experiment described in Section 5, the vector $\boldsymbol{v}$ had 69 dimensions, which ranged from −1,400 cent to +1,400 cent. More specifically, v_0 was equal to $C_{XY}(-1{,}400)$, v_{34} was equal to $C_{XY}(0)$, and v_{68} was equal to $C_{XY}(+1{,}400)$.

After making the vectors from all training samples, the number of dimensions of $\boldsymbol{v}$ is reduced to k using principal component analysis (PCA) in order to estimate statistical parameters from a small amount of training data. Moreover, reducing the number of dimensions is expected to lead to both rapid calculation and robustness against noise.

2.3 Comparison of F_0 Estimation Method Based on Harmonic Analysis

In recent years, novel F_0 estimation methods based on harmonic analysis have been proposed for melody representation of polyphonic audio [8)–10)]. Although these methods are used for decomposing polyphonic audio signals, they can also be used for extracting the pitch of humming data [9)]. In other words, the same as a traditional pitch estimator, these F_0 estimators can be used for extracting a tone interval by calculating the difference between successive pitches. In this section, we compare the proposed tone interval extractor with these "new generation" F_0 estimators.

These F_0 estimators are based on harmonic analysis. If the harmonic structure of an input humming is collapsed, these methods cannot estimate pitch correctly. On the other hand, the proposed method is based on the similarity of whole shapes of two log-spectra, and so can estimate a tone interval from not only harmonic structured spectra but also collapsed spectra. If the whole shape of one of the log-spectra is different from that of the other, the proposed method cannot estimate a tone interval correctly, but in this case the F_0 estimators cannot estimate it correctly either. Thus the proposed method has higher performance

than the F_0 estimators based on harmonic analysis.

3. Statistical Modeling for Tone Interval

A user's humming contains fluctuations of pitch frequency. In order to absorb the fluctuations, a statistical model is introduced for modeling a tonal feature.

The fluctuations of pitch frequency depend on various factors, such as the ability of the singer, tone interval, absolute pitch frequency of a note, duration of a note, tempo, and so on. Even though it is better to construct statistical models for each factor independently, this is impossible because a huge amount of hummed data is needed for training a huge number of statistical models. In order to construct reasonable numbers of statistical models, we introduce several assumptions.

First, it is assumed that the fluctuations of pitch frequency corresponding to different tone intervals are characterized by different statistical distributions. And it is also assumed that all pairs of notes with the same tone interval are characterized by the same statistical distribution even if the absolute pitch frequencies of the notes are different.

From these assumptions, we can prepare statistical distributions corresponding to each 100 cent tone interval. A specific distribution is estimated using all of the feature vectors corresponding to the same tone interval. For example, the distribution corresponding to +200 cent is estimated using feature vectors calculated from "do" → "re", "sol" → "la", and so on. In this paper, these statistical distributions are called "tone interval models".

Figure 3 shows a histogram of tone interval data in the database corresponding to −200 cent. While there is a large peak at −200 cent, we can observe many data distributed around the correct tone interval, which seem to be caused by inaccurate humming by the user. The shape of the histogram looks similar to the Laplace distribution. Therefore, we tried to use both the M-mixture Normal distribution with diagonal covariance matrix and the Laplace distribution in the experiment. An M-mixture K-dimensional Normal distribution with diagonal covariance matrix is defined as

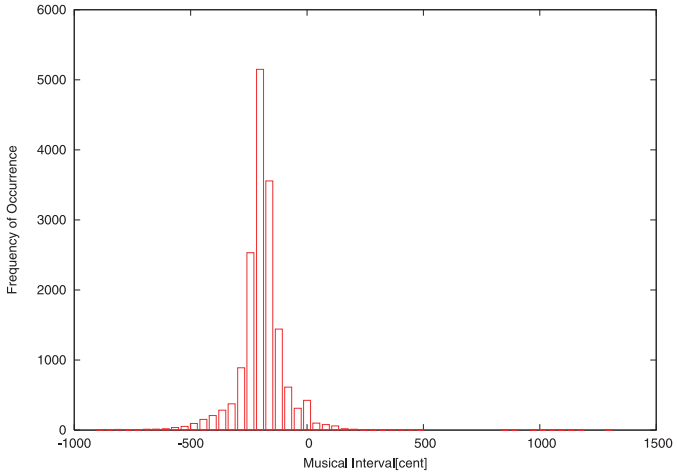


Fig. 3 Histogram of tone interval data.

$$p(\boldsymbol{z}|\boldsymbol{\theta}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) \tag{2}$$

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(z_k - \mu_k)^2}{2\sigma_k^2}\right\} \tag{3}$$

$$\boldsymbol{\theta} = \{\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, w_0, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1, w_1, \cdots, \boldsymbol{\mu}_M, \boldsymbol{\sigma}_M, w_M\} \tag{4}$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)$ denotes a mean vector and $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_K)$ denotes a variance vector that corresponds to the diagonal element of the covariance matrix.

A K-dimensional Laplace distribution is defined as

$$p(\boldsymbol{z}|\hat{\boldsymbol{\mu}}, \boldsymbol{B}) = \prod_{k=1}^{K} \frac{1}{2B_k} \exp\left(\frac{-|z_k - \hat{\mu}_k|}{B_k}\right) \tag{5}$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \cdots, \hat{\mu}_K)$ denotes a median vector, and $\boldsymbol{B} = (B_1, \ldots, B_K)$ denotes a scale parameter vector which is calculated by the equation,

$$B_k = \frac{1}{N} \sum_{i=1}^{N} |x_{ik} - \hat{\mu}_k| \tag{6}$$

where N denotes the number of samples.

4. Melody Matching Method Using Continuous DP Matching

An overview of the melody matching method is shown in **Fig. 4**. It is assumed that we have information on the melody lines of all the songs in the database, which are represented as MIDI sequences. The matching method is the same as the conventional MIR systems based on continuous DP matching [11] with insertion/deletion penalty [12] except for the definition of a similarity score between an input humming and a song in the database.

At first, an input humming is split into notes by the event detection module. The event detection method [13] is based on power information. In order to avoid detecting respiratory sound, a band pass filter (600 Hz–1,500 Hz) was used to enhance the formant frequency of the vowel /a/. Then a differential filter was applied and the onset time was extracted.

The similarity score is based on the above-mentioned probability and inter-onset interval (IOI) ratio. Let $\boldsymbol{z}(i)$ be a tonal feature vector (a peak tone interval or a vector of the cross-correlation function compressed by the PCA) calculated from the i-th and $(i+1)$-th notes in the input humming, and T_j be a tone interval of the j-th and $(j+1)$-th notes of a song in the database.

Let $\Delta L_h(i)$ be the i-th log-IOI ratio in the input humming, and $\Delta L_m(j)$ be the j-th log-IOI ratio of the song in the database, which is calculated as

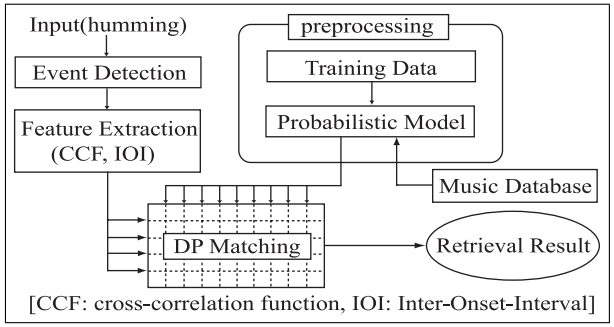


Fig. 4 Overview of the retrieving method.

$$\Delta L_h(i) = \log\left\{\frac{S_h(i+1)}{S_h(i)}\right\} \tag{7}$$

where $S_h(i)$ denotes an IOI of the i-th note in the humming, and is defined as the duration between the onset (start) time of the i-th note and the onset time of the $(i+1)$-th note.

The similarity score between the i-th interval in the input humming and the j-th interval of the song is calculated as

$$l(i,j) = w \log \frac{p(\boldsymbol{z}(i)|\boldsymbol{\theta}(T_j))}{\sum_{t\in\boldsymbol{T}} p(\boldsymbol{z}(i)|\boldsymbol{\theta}(t))} - (1-w)|\Delta L_h(i) - \Delta L_m(j)| \tag{8}$$

where $\boldsymbol{\theta}(T_j)$ is the parameter set of a tone interval model corresponding to the tone interval T_j, and w is a weighting factor. $\boldsymbol{T}$ denotes the set of all prepared tone intervals. The similarity score is calculated by subtracting the IOI distance from the tonal score because both a *higher* tonal score and a *lower* IOI distance indicate that an input humming and a song in the database are similar. In this similarity measure, IOI was not represented by statistical modeling, although it would be easy to do this in the same way as for the tonal feature. Statistical modeling of IOI was not employed in this paper in order to directly compare the effectiveness of the new tonal feature and its statistical modeling with conventional methods.

The best alignment between the input humming and a song in the database is calculated using continuous DP matching. If a matching path corresponding to the insertion or deletion of a note is selected, both the cross-correlation function and the IOI ratio are re-calculated using the appropriate pair of notes [5]. Note that the similarity score defined by Eq. (8) may become negative, but this is not a problem for the DP matching method which can find the optimum correspondence between two sequences regardless of the sign of the similarity.

5. Experiments

In order to investigate the effectiveness of the proposed system, several experiments were carried out using our original humming database. Although two singing/humming databases (QBSH corpus [14] and ThinkIT corpus [15]) are publicly available, most of the queries in these corpora are singing data. Moreover, hummed queries are sung with various phonemes such as /ta/, /la/ and /ti/.

Because our system only accepts a humming query hummed with /ta/, we could not use these open databases.

5.1 Experimental Conditions

Table 1 shows the experimental conditions. Training and evaluation data were hummed using the phoneme /ta/. None of the singers were professional singers, and they had not received vocal music training. Training data were collected by "controlled humming". The hummed data consisted of five notes: base note, +200 cent, +400 cent, +200 cent, and base note again (e.g. "Do" "Mi" "Sol" "Mi" "Do"). A singer hummed 30 samples (6 patterns × 5 times), and all training data was manually split into notes. 150 notes were obtained from 30 hummed samples sung by a singer, and 22,500 tonal features were calculated from all combinations of two notes (150×150) in the training data. It would be preferable to calculate a tonal feature only from two successive notes. However, all combinations of two notes were used for calculating a tonal feature in order to increase the number of training samples from a few hummed samples. On the other hand, all of the evaluation data consisted of songs in the database hummed by different singers.

We assumed that the maximum tonal interval is ±1,200 cent. Therefore, 25 interval models from −1,200 cent to +1,200 cent were constructed, and a cross-correlation function was calculated from −1,400 cent to +1,400 cent. A log-scaled spectrum is represented by 256 points, which means that the gap between two successive points corresponds to 41 cent. Therefore, the positive side of the cross-correlation function can be represented by 34 points (from +41 cent to +1,394 (= 34×41) cent). As a result, the whole shape of a cross-correlation function can be represented by a 69-dimension vector.

Table 1 Experimental conditions.

Acoustical Analysis	Sampling rate	16 kHz
	Window size	64 ms
Interval modeling	Training data	Humming by 10 males (225,000 pieces)
	Interval class	25 classes from −1,200 to +1,200 cent
Evaluation data		Humming by 5 males (326 pieces)
Music database		156 pieces of Children's songs

5.2 Retrieval Experiments

Top-1 retrieval accuracies are shown in **Table 2**. In this table, the statistical distribution "None" means that the tone interval model was not used and the similarity was calculated using the Euclidean distance. The tonal feature "Pitch" with statistical distribution "None" corresponds to the conventional single pitch MIR system. The definition of "top-n retrieval accuracy" is described in Ref. 5). This definition considers the case that the n-th candidate has the same score as the $(n+1)$-th candidate.

Several controllable parameters (number of dimensions k used in the PCA, weighting factor w described in Eq. (8), and insertion/deletion penalty used in the DP algorithm [12]) were determined *a posteriori*. Various combinations of parameters were examined using the test data, and the maximum accuracy is shown as a result. **Figure 5** shows an example of the retrieval accuracies given by various parameters for the ("Laplace", "Whole") condition. From this figure, $k = 8$ and $w = 0.2$ were selected. The ("1-mix", "Whole") condition also selected $k = 8$, but the ("2-mix", "Whole") condition selected $k = 12$. The selected values of the weighting factor w were distributed from 0.19 to 0.45.

From Table 2, tonal features calculated from the cross-correlation function dramatically improved the retrieval performance. Most notably, the peak tone interval of the cross-correlation function was about 10 points higher than the pitch-based tonal features for both the "Laplace" and "None" conditions. The whole-shape feature was not effective compared with the peak tone interval.

The results when the peak tone interval was used as a tonal feature revealed that the Laplace distribution was more effective than the Euclidean distance. However, the Normal distribution was not effective. The two-mixture Normal distribution showed higher performance than that of the single mixture, how-

Table 2 Top-1 retrieval accuracy.

		Tonal features		
		Peak	Whole	Pitch
Statistical distribution	1-mix Normal	80.7%	81.9%	—
	2-mix Normal	81.9%	83.7%	—
	Laplace	86.8%	83.7%	75.5%
	None	83.4%	—	74.2%

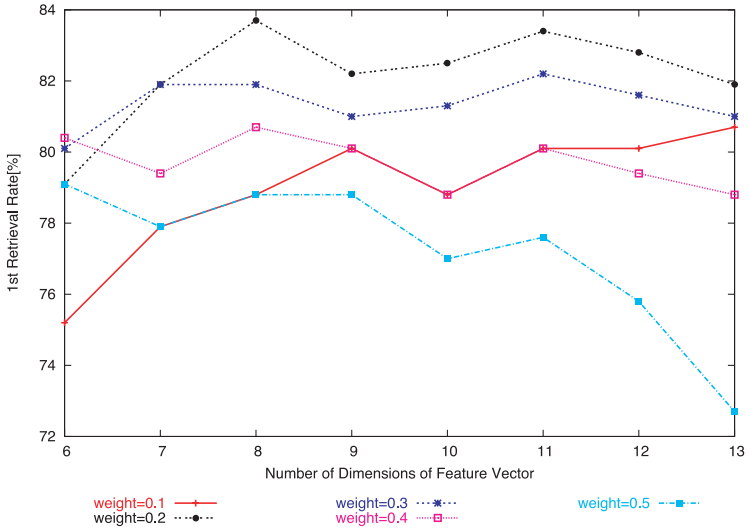


Fig. 5 Example of retrieval accuracy given by various settings.

Table 3 Comparison of retrieval accuracies with conventional methods.

Method	Top-1	Top-10	Calculation time
Proposed	86.8%	93.6%	1.28
Single pitch	74.2%	89.0%	1
Multiple pitches	86.5%	94.1%	9.28

ever, it did not outperform the conventional method. This result seems to have been caused by a mismatch of the shape of the distribution between the Normal distribution and the actual distribution of users' humming.

Table 3 shows the top-1 and top-10 retrieval accuracy of the proposed method (the peak tone interval of the cross-correlation function with the Laplace distribution) and the conventional methods. The calculation time normalized by the single pitch method is also shown. In this table, the "Multiple pitches" method [5] used three pitch candidates.

The proposed method showed almost the same performance as the system based on multiple pitches, and the retrieval time hardly increased compared with the single pitch method.

5.3 Retrieval Accuracy with a Large Database

In this experiment, only 156 songs were registered in the database, but for practical use, ten thousand or more songs should be registered. Let us discuss applying the proposed method to such a practical situation.

The main problem is the retrieval time. The proposed method is based on continuous DP matching, so the retrieval time is proportional to the number of songs in the database. In the experiment, the retrieval time for a query was about 300 ms. If the number of songs in the database increases to 10,000, the retrieval time would become 19 seconds, which is very slow compared with index-based MIR systems [16),17)]. Although our programming code is not optimized for retrieval speed, some extra methods should be introduced to speed up the retrieval.

One idea is a pre-selection method. First, a query is input into a rapid (but lower performance) method, and the method outputs several hundred candidates. The proposed method uses this list of output candidates as a database, and retrieves the final result.

Another idea is to combine the proposed method with an index-based MIR system. The proposed tonal feature and its statistical modeling idea could be introduced to an existing index-based MIR system, or an indexing approach could be introduced into our proposed method.

In any case, the main problem to be solved in the next step is to use the proposed method with a large database.

6. Integration of Retrieval Results

In the proposed method, we can choose any combination of tonal feature and statistical distribution. From Table 2, the combination of the peak tone interval and the Laplace distribution showed the highest retrieval accuracy. Although other combinations showed lower retrieval accuracy than the best combination, some of them gave correct retrieval results for input humming queries for which the system with the best combination could not give the correct results. Therefore, integration of the retrieval results may improve the final retrieval accuracy.

Methods of integrating results have been proposed [18)–20)] in many research

Table 4 Number of methods *vs.* samples which can be retrieved.

All methods	Two methods	One method	No methods	Total
259	15	22	30	326

fields. In these methods, each recognizer outputs a recognition result, and the final result is calculated using all of the output results with some information such as rank, recognition score, and confidence score.

In this section, we investigate the possibility of improving the retrieval accuracy, and try to apply several simple integration methods to the retrieval results.

6.1 Investigation of the Retrieval Results

In order to investigate the possibility of improvement, the retrieval results obtained in Section 5.2 were analyzed in detail for the following three combinations:

- The peak tone interval with the Laplace distribution
- The whole-shape feature with the Laplace distribution
- The whole-shape feature with the two-mixture Normal distribution

Table 4 shows the number of test samples that were retrieved correctly by one, two, or three methods. From this table, about 80% of the test samples were retrieved correctly by all methods. On the other hand, 11% of the test samples were retrieved correctly by at least one method, but not all methods. If an appropriate retrieval method were selected for each test sample, the retrieval accuracy could be improved to 91%. Though it is not known how to select an appropriate method, this finding suggests that improvements could be achieved by integrating the retrieval results. In this section, several integration methods are applied to retrieval results.

6.2 Integration Methods

6.2.1 Majority Decision

The first integration method is the majority decision. Each retrieval method votes for the first candidate song, then the song that gathers the most votes is selected as the final result. In addition, we assign a priority to each of the retrieval methods according to the retrieval accuracy of the method. If more than one song gathers the maximum number of votes, the result generated by the method with the highest priority becomes the final result.

6.2.2 Rank-based Rescoring

The majority decision method considers only the first candidate of each method. In order to consider the second or lower-ranked candidate, a rank-based rescoring method is introduced.

Each retrieval method outputs all candidates with their rank number. Let $r_j(i)$ be the rank number of the song i output by the retrieval method j. The rank-based score $s_j(i)$ is calculated from $r_j(i)$, and the total score $t(i)$ of the song i is calculated as follows:

$$t(i) = \sum_j s_j(i) \tag{9}$$

Finally, the song with the highest score is selected as the final result.

One of the most important points is how to calculate $s_j(i)$. In this paper, the definition proposed in the weighting models rank (WMR) method [20] is used. $s_j(i)$ is calculated as follows:

$$s_j(i) = \exp\{A - Br_j(i)\} \tag{10}$$

In this definition, both A and B are constant, and these are defined as satisfying the following equations:

$$s_j(i) = N \quad \text{if} \quad r(i) = 1 \tag{11}$$

$$s_j(i) = 1 \quad \text{if} \quad r(i) = N \tag{12}$$

where N denotes the number of songs in the database.

We also introduce another definition:

$$s_j(i) = A - Br_j(i) \tag{13}$$

This definition gives a higher score than Eq. (10) to lower candidates. Note that the constants A and B are also defined using Eq. (11) and Eq. (12).

6.2.3 Rescoring Based on Retrieval Score

We also try to apply the rescoring method based on the retrieval score. Each retrieval method outputs all candidates with their retrieval scores. The retrieval score $L_j(i)$ is normalized by an average score, and the total score $t(i)$ of the song i is calculated in the same manner as Eq. (9):

$$\hat{L}_j(i) = L_j(i) - \frac{1}{N}\sum_{k=1}^{N} L_j(k) \tag{14}$$

Table 5 Retrieval accuracy of integration methods.

Individual results			Integrated results			
Peak	Whole-shape		Majority	WMR		Score
Laplace		Normal	decision	exp	linear	based
86.8%	83.7%	83.7%	87.4%	81.9%	82.2%	82.8%

$$t(i) = \sum_j \hat{L}_j(i) \tag{15}$$

6.3 Experiments of Integration Methods

In order to investigate the effectiveness of the integration methods, music information retrieval experiments were carried out. The three methods listed in Section 6.1 were used for all integrated methods, and experimental conditions were the same as in Section 5.

Table 5 shows the retrieval accuracy of each integration method. In this table, "WMR with exp" denotes the integration method with Eq. (10), and "WMR with linear" denotes the integration method with Eq. (13).

From this table, the majority decision method showed the highest performance of all, and it outperformed the best combination method. The reduction of retrieval error was about 5% (from 13.2% to 12.6%). However, integration was not effective with other methods.

Both the WMR method and score-based method are effective when all retrieval methods output the correct song with a higher rank. However, when a retrieval method output the correct song with a very low rank, these methods did not work well, even though the majority decision was not influenced by such a situation. Moreover, the majority decision outputs the result given by the highest priority retriever when the majority cannot be decided. An analysis of the results showed that this rule is effective.

7. Conclusion

In this paper, we proposed a new MIR system with QbH that needs no pitch extraction. A tone interval between two successive notes is directly calculated based on the cross-correlation of log-scaled spectra. Moreover, these tonal features are modeled by a statistical distribution in order to represent the fluctuation of the user's humming.

From the experimental results, the peak tone interval of the cross-correlation function with the Laplace distribution showed the highest performance of all. It has almost the same performance as the system based on multiple pitches, and the calculation time was hardly increased from the single pitch method.

Several integration methods were introduced in order to improve the retrieval accuracy. These methods integrate the retrieval results of the proposed methods with three combination settings. The majority decision method showed the highest accuracy. The proposed method with the best combination setting reduced the error by about 5%.

References

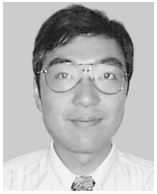
1) Pauws, S.: CubyHum: A Fully Operational Query by Humming System, *Proc. ISMIR*, pp.187–196 (2002).
2) Jang, J.S.R., Lee, H. and Chen, J.: Super MBox: An Efficient/Effective Content-based Music Retrieval System, *The 9th ACM Multimedia Conference* (*Demo paper*), pp.636–637 (2001).
3) de Cheveigne, A. and Kawahara, H.: YIN, a fundamental frequency estimator for speech and music, *The Journal of the Acoustical Society of America*, Vol.111, No.4, pp.1917–1930 (2002).
4) Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled sound, *Proc. Institute of Phonetic Sciences University of Amsterdam*, Vol.17, pp.97–110 (1993).
5) Heo, S.-P., Suzuki, M., Ito, A. and Makino, S.: An Effective Music Information Retrieval Method Using Three-Dimensional Continuous DP, *IEEE Trans. Multimedia*, Vol.8, No.3, pp.633–639 (2006).
6) Shih, H.-H., Narayanan, S.S. and Kuo, C.-C.J.: A Statistical Approach to Humming Recognition, *Proc. ICASSP 2002* (2002).
7) Sagayama, S., Takahashi, K., Kameoka, H. and Nishimoto, T.: Specmurt Anasylis: A Piano-Roll-Visualization of Polyphonic Music Signals by Deconvolution of Log-Frequency Spectrum, *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing* (2004).
8) Goto, M.: 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, No.4, pp.311–329 (2004).
9) Song, J., Bae, S.Y. and Yoon, K.: Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System, *Proc. ISMIR*, pp.133–139 (2002).
10) Kameoka, H., Nishimoto, T. and Sagayama, S.: A Multipitch Analyzer Based on

Harmonic Temporal Structured Clustering, *IEEE Trans. Audio, Speech and Language Processing*, Vol.15, No.3, pp.982–994 (2007).
11) Oka, R.: Continuous word recognition with Continuous DP, Technical Report S78–20, Acoustic Society of Japan (1978). (in Japanese).
12) Ito, A., Heo, S.-P., Suzuki, M. and Makino, S.: Comparison of Features for DP-matching based Query-by-humming System, *Proc. ISMIR*, pp.297–302 (2004).
13) Heo, S.-P., Suzuki, M., Ito, A., Makino, S. and Chung, H.-Y.: Multiple Pitch Candidates based Music Information Retrieval Method for Query-by-Humming, *Proc. AMR*, pp.189–200 (2003).
14) Jang, J.-S.R.: QBSH: A Corpus for Designing QBSH (Query by Singing/Humming) Systems. available at the QBSH corpus for query by singing/humming (http://www.cs.nthu.edu.tw/˜jang).
15) MIR group, ThinkIT Speech Lab., Chinese Academy of Sciences: ThinkIT Query-by-Humming Corpus. xwu@hccl.ioa.ac.cn.
16) Kosugi, N., Nishihara, Y., Sakata, T., Yamamoto, M. and Kushima, K.: A Practical Query-By-Humming System for a Large Music Database, *ACM Multimedia 2000*, pp.333–342 (2000).
17) Sonoda, T. and Muraoka, Y.: A WWW-based Melody-Retrieval System — An Indexing Method for A Large Melody Database, *Proc. International Computer Music Conference*, pp.170–173 (2000).
18) Fiscus, J.G.: A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), *Proc. Automatic Speech Recognition and Understanding*, pp.347–354 (1997).
19) Schwenk, H. and Gauvain, J.-L.: Combining Multiple Speech Recognizers using Voting and Language Model Information, *Proc. ICSLP*, Vol.II, pp.915–918 (2000).
20) Markov, K.P. and Nakagawa, S.: Frame level likelihood normalization for text-independent speaker identification using Gaussian mixture models, *Proc. ICSLP*, pp.1764–1767 (1996).

(Received May 13, 2008)
(Accepted November 5, 2008)
(Original version of this article can be found in the Journal of Information Processing Vol.17, pp.95–105.)

Motoyuki Suzuki was born in Chiba, Japan, in 1970. He received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1993, 1995, and 2004, respectively. Since 1996, he has worked with the Computer Center, the Information Synergy Center, and Graduate School of Engineering, Tohoku University, as a Research Associate. From 2006 to 2007, he worked with the Centre for Speech Technology Research, University of Edinburgh, UK, as a Visiting Researcher. He is now an Associate Professor of Institute of Technology and Science, University of Tokushima, Tokushima, Japan. He has been engaged in spoken language processing, music information retrieval, and pattern recognition using statistical modeling. He is a member of the Acoustical Society of Japan and the Institute of Electronics, Information and Communication Engineers.

Takuto Ichikawa was born in Tokyo, Japan in 1984. He received the B.E. and M.E. degrees from Tohoku University, Sendai, Japan, in 2006 and 2008 respectively. He is now working for Mobile Communication Business Development Group, SHARP corporation. He has engaged music information processing and software development of mobile phones.



Akinori Ito was born in Yamagata, Japan in 1963. He received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1984, 1986 and 1992 respectively. Since 1992, he has worked with Research Center for Information Sciences and Education Center for Information Processing, Tohoku University. He joined Faculty of Engineering, Yamagata University from 1995 to 2002. From 1998 to 1999 he worked with College of Engineering, Boston University, MA, USA, as a visiting scholar. He is now an Associate Professor of Graduate School of Engineering, Tohoku University. He has engaged in spoken language processing and statistical text processing. He is a member of the Acoustical Society of Japan, the Institute of Electronics, Information and Communication Engineers, Human Interface Society and the IEEE.

Shozo Makino was born in Osaka, Japan, on January 3, 1947. He received the B.E., M.E., and Dr.Eng. degrees from Tohoku University, Sendai, Japan, in 1969, 1971, and 1974, respectively. Since 1974, he has been working with the Research Institute of Electrical Communication, Research Center for Applied Information Sciences, Graduate School of Information Science, Computer Center, and Information Synergy Center, as a Research Associate, an Associate Professor, and a Professor. He is now a Professor of Graduate School of Engineering, Tohoku University. He has been engaged in spoken language processing, CALL system, autonomous robot system, speech corpus, music information processing, image recognition and understanding, natural language processing, semantic web search, and digital signal processing. He is a member of IEEE, ISCA, the Institute of Electronics, Information and Communication Engineers, Japan Society of Artificial Intelligence, Association for Natural Language Processing and Japan Society for Educational Technology.