Proceedings of 20th International Symposium on Music Acoustics
(Associated Meeting of the International Congress on Acoustics)

25-31 August 2010, Sydney and Katoomba, Australia

Automatic scoring of sung melodies in comparison with human performance

Masuzo Yanagida (1), Yumiko Mizuno (1) and Tomoya Matsunaga (2)

(1) Dept. of Informatics, Doshisha University, Kyo-Tanabe, Japan
(2) NTT Data Corporation, Koto-ku, Tokyo, Japan

PACS: 43.75.Cd, 43.75.Rs, 43.75.Xz

ABSTRACT

Performance of transcribing acoustic signals into music notation is compared between an automatic transcribing system recently developed by the authors' group and human scorers, focusing mainly on tone height using two kinds of singing sounds : one, sung by human singers and the other, synthesized with a commercially sold singer system. As test melodies should be unknown to human scorers, several melodies were designed for this particular experiment. Those melodies are divided into two sets of melody groups: one for human singing and the other for synthetic singing. The reason why the test melodies were not made common for human singing and synthetic singing is that each test molody should be unknown to human scorers. Each set is sub-classified into three tonality levels, highly tonal, less tonal and atonal. Significant difference is recognized in correct tone height transcription rate among three tonality levels both for human scorers and the automatic transcription system for melodies sung by human singers. Although results obtained for melodies sung by human singers are somewhat vague as the correct answers are hard to be defined, but results obtained for melodies sung by synthetic voice were just as expected that transcription performance of the system is far superior to that of human scorers for atonal melodies or tone series, and significantly superior for low tonality melodies, but no significant difference for high tonality melodies .

INTRODUCTION

"Automatic Scoring" or "Automatic Transcription" is a procedure of transforming music sounds into any of music notations [1]. The notation is not limited to the form of western staff notation. Somewhat higher-level description[2, 3] will be useful for theoretical handling, and conceptual description[4] will be required for composition or conceptual design of music. The current study, however, employs conventional staff notation or its equivalent MIDI notation as the target notation.

The most important and practical application of automatic transcription would be music retrieval or tune retrieval from singing voice [5, 6, 7]. In that case, however, we usually feel ourselves satisfied if the target tune is highly ranked in output list given by the retrieval system. That means accurate transcription is not necessary but rough transcription suffices for music retrieval. Actually a very simple "Parsons code" [8], describing only up/down or repeat, was tried in early stage music retrieval, though it was detailed to be 5-level contour description by Vercoe et al. [9]. Recent topics in Query-by-Humming seem to have moved to fast algorithms for evaluating similarity [10] and the substance of melodic similarity itself [11].

A transcription system [12, 13] can be a useful tool for storing musical sounds in a computer and for constructing a content-based music database [14] by accumulating transcribed data in a database [15, 16]. It also helps beginners to analyze [17], compose [18] and arrange [19] music besides providing means for music retrieval based on melodic similarity [5, 11, 20].

As automatic transcription is one of classic issues in musical information processing, many research works have been reported so far, such as extraction of fundamental frequency (f_0) from complex tones [21], tracking f_0 based on nature of musical sound [22], note identification of each part in polyphonic music [23], multipitch detection based on tied Gaussian mixture model [24], and so forth.

In contrast to abundance of research works in automatic transcription of sounds played on musical instruments, there have been few in sung music except those for "Query by humming" [5, 10, 25], maybe because of anticipated difficulty in evaluating transcription performance on unstable human singing. In contrast to transcription, performance of tune retrieval is easy to evaluate. Though automatic transcription of a single melody played on a musical instrument achieves high recognition rate, it is not the case for sung melodies due to inaccuracy and instability of f_0 and intentional or unconscious fluctuation of tempo in actual human singing [26].

It is difficult to make automatic transcription systems simulate performance of human scorers as they transcribe sounds into musical staff making full use of their abilities and sensibilities in music, but it is expected to be not so difficult to construct a system that can achieve better performance than human scorers at least for atonal tone sequences as far as they are sung correctly without artistic deviation. The ultimate goal of our current research is to construct an automatic transcription system that shows better performance than human scorers at least for correctly sung atonal tone sequences and hopefully even for tonal melodies. We hope that an expected system would show the same performance for any tone sequence independent of its tonality degree if the tone sequence is sung with the same proficiency or fidelity to an original score.

Our current system, that has been revised several times over the original version [27], well judges "syllable name" and "note value" of each tone, estimating singer's "standard f_0", or f_0 (Hz) of "Do" in the "movable Do" scale, and local "standard

tempo" (number of quarter notes in a minute) in input singing. These functions are necessary for realizing a robust transcription system that can achieve satisfactory performance even in case melodies are sung in free time-varying tempo on a scale on any arbitrary standard f_0, including cases of time-varying standard f_0 as far as the fluctuation is not so fast nor deep. The system introduces two templates to be matched to occurence frequency distribution on logarithmic scales: one is “Scale Template" to estimate singer’s standard f_0, and the other is “Note-value Template" to estimate physical time duration of IOI (Inter-Onset Interval) corresponding to the shortest note-value in the local tempo.

The following section explains strategies for comparing transcription performance of our current transcription system with that of human scorers as a necessary step for improving our system to know how the systen could be improved looking into performance differences between human scorers and the system. The subsequent section presents our basic idea of “template matching” on occurrence frequency distribution to estimate singer’s standard f_0 corresponding to “Do” on “movable Do” scale and the local standard IOI corresponding to the shortest note-value. The remainder sections describe the scheme of current experiments for performance comparison and the results, followed by discussions and conclusions.

EXPECTATION ON AUTOMATIC TRANSCRIPTION SYSTEMS

General requirements

Performamce of transcribing sung melodies into staff notation depends mainly on ability of the scorer, but it depends also on accuracy or fidelity of singing to the original score and degree of tonality of the melody itself. In case transcription is done by a human scorer, he/she would make full use of his/her sensibility and capability in music for judging the height of a tone in a given series of tones with his/her sense of absolute/relative pitch referring to his/her sense of musical scales.

Absolute height of the musical scale on which a singer sings is sometimes shifted from the standard pitch, or A_4=440$\sim$ 442Hz, in case a tune is sung by a singer who doesn’t have absolute pitch not hearing any reference tone.

Automatic transcription systems are expected to properly function as far as the song roughly keeps musical intervals, or frequency ratio among tones during a definite scope of singing even if the absolute height is shifted from the standard one. In addition, fluctuation in tempo makes it difficult to estimate note-values, or nominal tone duration on the score. Means to manage the arbitrariness of absolute height and flexible tempo are proposed in literature, our system solves the problems by employing a more reliable method based on a unified concept in statistics.

Quantization of tone height and tone duration

Automatic transcription systems are required to quantize both tone height and tone duration observing the rules of notation. In this paper, tone height within an octave is quantized into 12 equal-tempered discrete heights neglecting the absolute frequency, admitting relative names. Duration of a tone or a sile nce is quantized in unit of 2’s power, or one of full/half/quarter /eighth/16th/32th notes or rests, respectively, with a single dot placed after a note or a rest to add it one-half of its time value.

That means: we accept

- both note names and syllable names
- enharmonics, or different labeling for the same tone height on equally-tempered chromatic scale, such as Sol$^{\#}$ and La$^{\flat}$,

we neglect

- tonality, consequently we allow enharmonics as mentioned above
- metrical structure, or bars in other expression

at evaluating answers of human scorers.

So, subjects can write their answers either in absolute tone names or in syllabic names assuming any arbitrary tone as the tonic using # and ♭ to signify notes outside a scale defined by the key note, without marking bars. That means subjects are allowed to write their answers in text form but not conventional musical notations.

Furthermore, we exclude

- double-dots that add a note one-half and one fourth of its value, and
- multiplets, such as triplets, quintuplets, septuplets etc.

by designing test melodies or tone series in the current study.

EVALUATING TRANSCRIPTION PERFORMANCE

The purpose of performance comparison

The objective of the current study is to compare transcription performance by human scorers and that by the current system we are developing. Detailed investigation of tone sequences for which the system fails but most or some human scorers succeed will notice us what function is missing in our system, and we can expect it will provide us useful hints to improve the system.

Considering human cases

Possibility of memory-dependent transcription

In case a human scorer is told to transcribe a sung melody into staff notation, he/she will show good transcription performance in case he/she could recognize a similar melody in his/her memory even if the input melody is sung out of pitch. Abilities involved in that situation, however, are not only the ability of transcription, but is compound abilities of melody memorization, similarity detection, melody extraction from harmonic music, coding melody lines in mind and so on. These abilities are different from the ability of our concern. So, we have to use melodies or tone sequence that are not known to scorers in order to remove effects of marginal abilities.

Effects of tonality

Usually human scorers show better performance for highly tonal melodies rather than for melodies of poor tonality. Atonal tone sequences are extremely difficult to transcribe even for professionals having absolute pitch. So, transcription performance should be mentioned in connection with tonality level of the melodies employed.

Required time for pruning impossible keys

In order to take tonality into account, definite number of tones or definite length of melody would be necessary to establish tonal sensation or discarding impossible keys from 24 possible keys based only on what the scorer has perceived upto the current time point. So, test melodies consisting of different number of notes or different time duration are required to be used to be compared from a point of number of notes or time duration.

Definition of correct answers

In case we use sung voices as test samples, we have to prepare those that are sung properly in pitch or those that are synthesized properly for evaluating performance of a transcription system, otherwise we cannot evaluate the transcription performance not knowing the correct answers. If you want to know how human scorers perceive the height of tones in test songs regardless of actual f_0, you can evaluate the performance assuming that the majority answer of good scorers would be the correct answer. In case an atonal sequence is used as a test melody, however, answers of scorers easily diverse and you often cannot find any majority decision among answers even if you use professional musicians.

DESCRIPTION OF THE SYSTEM

Brief description

A flow-diagram of our system [27] is depicted in Fig.1 with sample signals sent on the down-arrows. First, a waveform of singing, such as depicted in the left-top in Fig.1, is read-in from a file or from a microphone via A/D converter, with sampling rate 44.1 ksamles/s, through a 2048 point (46.4ms) Hamming window by 441 point (10ms) frame shifting.

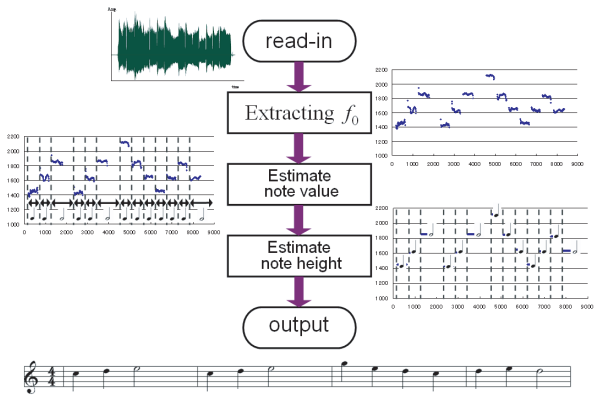

Figure 1: Flow of the system [27]

Raw fundamental frequency f_0 of each analysis frame is extracted using conventional auto-correlation function, modified correlation function (between the original signal and the prediction residual), and the cepstrum. Final f_0 for each analysis frame is determined by majority decision among the three raw f_0s. Examples of sequence of final f_0s are depicted in the right-top in Fig.1, in which abscissa is the time axis in ms and the ordinate is pitch in cents assuming A_1(=55Hz) as 0 cent.

Then, Inter-Onset-Interval (IOI in short, here after) between adjacent tones is detected as the interval between beginning points of adjacent f_0 sequences as the left-bottom in Fig.1. The time duration of a unit note, a quarter note in this case, is determined based on occurence frequency distribution of IOI as explained later. Note-value of a tone is determined based on duration of the corresponding IOI divided by the duration of the unit note.

Finally, pitch or note name of each tone is determined based on the average f_0 of the middle parts in each IOI section partitioned by broken lines representing positions of note heads. The absolute frequency of the reference pitch, “Do” on a “movable Do” scale in our case, is determined based on occurence frequency distribution of frame-wise f_0 as explained later in detail. The final output of the system is staff notation, such as depicted in the bottom in Fig.1, where bars are put as results of investigating repetition period of note-value sequences.

Template Mtching on Occurrence Frequency Distribution

Detecting note boundaries

In case a song is hummed by a single singer, simultaneous multi-melody is impossible except extraordinary case such as Mongolian homey. So, a time portion corresponding to a note on a melody in staff notation is to be partitioned. Then, it is desirable to extract segments in which trajectory of f_0 is stable. However f_0 may not be stable during whole the time duration assigned as its time value. Moreover, the duration of each tone is not proportional to its note-value because each tone is generated within the duty ratio of about 80% of the nominal note-value in ordinary (*non-legato*) cases, though it can be less than 50% in *staccato* case while it may be almost 100% for *legato* case. Then it can be stated that a note-value should be determined based on the time duration of the IOI of the portion considering local tempo.

Employed here to determine note-value is IOI, though it has two defects: first, it inevitably includes the note value of the following rest if any, second, the method cannot give the note-value of the last note.

It is difficult to extract IOIs from temporal change of short-term power or amplitude, because amplitude envelope is not stable during phonation for a note, and sometimes we cannnot recognize any dip between notes. A method adopted in our system is to determine IOI by detecting points of change on trajectories of f_0, but not on amplitude envelope, as depicted in Fig. 2.

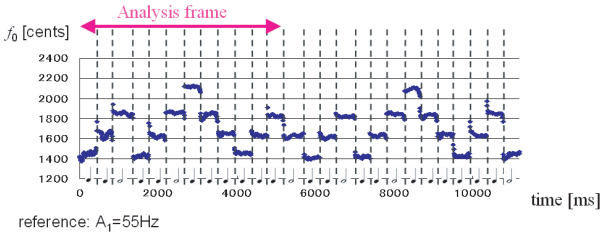

Figure 2: Detecting the head position of each note.

Extracting voiced parts having stable f_0

As mentioned above, majority decision among auto-correlation function, modified correlation function and cepstrum is employed to determine f_0 of each analysys frame, if the frame is classified to be a voiced part. Each original value for f_0 is obtained as the center frequency of a quadratic curve passing consecutive three sampling points near the principal peak on each evaluation function.

We call the trajectory obtained by connecting the results of majority decision a “sequence” though sometimes it is cut by unvoiced or silent parts or it shows very irregular movements. A sequence corresponding to a note are expected to be stable and has a definite length.

In one sequence, the occurrence frequency of erroneous peaks is thought to be fewer than that of peaks corresponding to correct pitch on a scale that the singer intends. So, sequences shorter than 100ms or those having irregular f_0 movenents are removed as false sequences of errors in f_0 extraction. Finally, sequences corresponding to notes are expected to be correctly extracted.

Note-value template

This system is designed to be able to transcribe songs sung at arbitrary tempo. As singer's range of tempo is so wide if tempo is not specified, it is required to estimate IOI corresponding to time duration of the note value of a reference note such as a quarter note or an eighth note. Then note value of each note can be assigned by evaluating which note value is the nearest to the observed IOI in concern. A note value basically takes a value in a 2's power system, so all note-values are thought to successfully correspond to either of 2's power component of a reference time duration as far as the singing tempo remain constant.

So it is easy to assign note-value of each note if the IOI corresponding to the shortest note-value in the song is correctly determined using a 2's power template on occurence frequency distribution of IOI using logarithmic time scale. In this paper, IOI corresponding to the shortest note-value is called "basic IOI" and is expressed by τ_b. Other note-values correspond to time duration about 2's power times of τ_b.

By observing occurrence distribution of IOIs of a song, distribution peaks are expected to be found in 2's power intervals. The peak corresponding to the shortest IOI in 2's power series is regarded as τ_b. So, a function $T_\tau(\tau, \tau_b)$, showing a large value at 2's power multiple over τ_b on a logorithmic time axis τ, depicted in Fig. 3 is adopted as the "note-value template" for estimating τ_b.

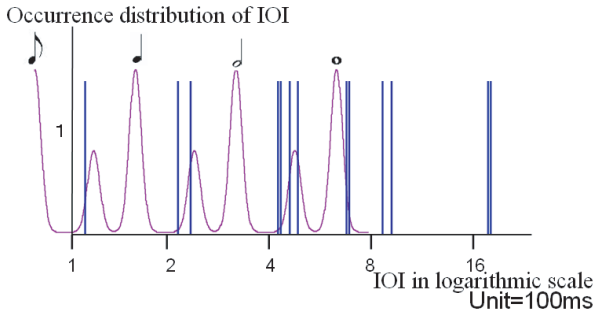

Figure 3: Occurrence frequency distribution $F_{\mathrm{IOI}}(\tau, k)$ of IOI in region k on the logarithmic time axis τ and the Note-value Template $T_\tau(\tau, \tau_b)$ located at an arbitrary posision where occasionaly τ_b for a eighth note is around 0.5.

$T_\tau(\tau, \tau_b)$ is a function having two variables τ and τ_b, where τ_b denotes the basic IOI and τ, the physical time on the logarithmic scale. Basic $T_\tau(\tau, \tau_b)$ takes the maximum value 1 at 2's power on the time axis, or each constant interval on the logarithmic time scale, while we can make it take the half value 0.5 at 50% larger time point representing time values of dotted notes. As described above, basic note-values such as a eighth note, a quarter note, a half note, and a full note take time values of 2' power multiples.

Coincidence between occurrence distribution of IOI and the note-value template becomes large if the note-value template well matches occurrence distribution of IOI. So, τ_b is estimated by finding τ that gives the maximum coincidence between occurrence distribution of IOI and the note-value template. τ_b is defined as τ that gives the maximum value of the following index representing the degree of coincidence of the note-value template to the occurence distribution of IOI, on a logarithmic scale, in evaluating region k in the input song:

$$C_\tau(\tau_b, k) = \int_{-\infty}^{\infty} T_\tau(\tau, \tau_b)\, F_{\mathrm{IOI}}(\tau, k) d\tau \qquad (1)$$

Once τ_b is determined, assigning note-value to each tone is straight forward judging the logarithmic distance to the standard value of each note-value. Results are obtained as depicted in Fig 4.

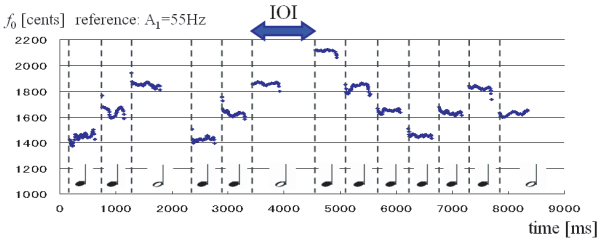

Figure 4: Assigning note-value to each note.

Scale Template

Without being given the pitch of the beginning tone, it is difficult for those who have no sense of absolute pitch to start singing at correct pitch. Even in such a case, however, systems are expected to correctly judge the height as far as the singer sings with relatively correct pitch. Generally speaking, musical intervals among notes are kept nearly correct even by amateur singers.

In this research, syllable names are identified by finding f_0 of the tonic note on a musical scale particular to the singers in concern. f_0 of the tonic is estimated by fitting a template representing a set of f_0s corresponding to 7 diatonic syllable names or 12 chromotic tones within one octave on a musical scale with an adjuster factor to occurrence distribution of observed f_0.

In case a singer sings with correct musical intervals, peaks on occurrence frequency distribution of f_0 appear on diatonic or chromotic scales. So, a function $T_\nu(\nu, \nu_{\mathrm{Do}\nu})$ having multi-peaks as depicted in the lower figure in Fig. 5 is introduced to estimate f_0 of the tonic for the singer. We call it "scale template". $T_\nu(\nu, \nu_{\mathrm{Do}\nu})$ is a function having two variables, ν and $\nu_{\mathrm{Do}\nu}$, where ν represents frequency, and $\nu_{\mathrm{Do}\nu}$ represents the frequency of "Do" on the "movable Do" for singing in relaive pitch.

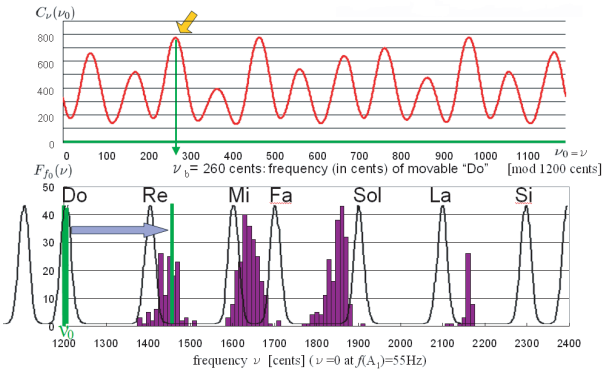

Figure 5: Searching the best position of the Scale template fitting to the Ocurrence frequency distribution of f_0

The horizontal axis of Fig. 5 is frequency on the logarithmic scale. $T_\nu(\nu, \nu_{\mathrm{Do}\nu})$ slides along the horizontal axis according to $\nu_{\mathrm{Do}\nu}$. Function $T_\nu(\nu, \nu_{\mathrm{Do}\nu})$ takes the value 1.0 at mod(i,12)= 0,2,4,5,7,9,11 that corresponds to the tone heights on the diatonic scale, or white keys of the piano, and 0.5 at mod(i,12)= 1,3,6,8,10, that corresponds to other tone heights, or black keys, and a small value, otherwise. $\nu_{\mathrm{Do}\nu}$ represents the frequency of

the tonic note in cents on the musical scale for the singing in concern.

The standard frequency ν_{Dov} for the singer is determined as ν that gives the maximum value for coincidence C_ν between $F_{f_0}(\nu,k)$, the occurrence distribution of ν and the "scale template", represented by function $T_\nu(\nu,\nu_{\mathrm{Dov}})$ defined as

$$T_\nu(\nu,\nu_{\mathrm{Dov}}) = \sum_{i=0}^{11} w_i \exp\left(-\frac{(\mathrm{mod}(\nu-\nu_{\mathrm{Dov}},1200)-100i)^2}{2\sigma^2}\right) \quad (2)$$

where

$$w_i = \begin{cases} 1.0 & \text{for } \mathrm{mod}(i,12) = 0,2,4,5,7,9,11 \\ w & \text{otherwise} \end{cases} \quad (3)$$

and

$$C_\nu(\nu_b,k) = \int_{-\infty}^{\infty} T_\nu(\nu,\nu_{\mathrm{Dov}})\, F_{f_0}(\nu,k)\, d\nu \quad (4)$$

which expresses the degree of coincidence of the scale template to the occurence distribution of f_o in cents of evaluating region k in the input song:

Assuming the equal temperament, all the expected fundamental frequencies of the keys on the musical scale can be easily calculated for judging syllable names of input tones. Syllable names of tones can be identified by taking the local average of the trajectory of f_0 within stable parts supposed to be corresponding to a note, evaluating which syllable name gives the nearest f_0 to the average.

Mechanism for attaining flexibility for temporal shifting in singer's standard frequency and local tempo

In order to attain both robustness and flexibility for temporal shifting of singer's standard pitch and local tempo, the currennt system adjusts the range of taking occurrence distributions of IOI and f_0. The default range of investigating occurrence frequency distribution of IOI is several seconds, and that for f_0 is 10 to 12 tones if available. The range of investigating occurrence frequency distribution is desirable to be made controlled by the stability of input singing. Detailed discussion is left for future investigation.

DESCRIPTION OF EXPERIMENTS FOR EVALUATING TRANSCRIPTION PERFORMANCE

Melodies used in the experiments

Melodies or tone sequences designed for this experiment were 60 in total as listed in Table 1. These are divided into two sets: one for human singing and the other for synthetic singing. As the subjects or human scorers of experiments for human singing participated also in the experiments using synthetic singing, the melodies cannot be common for human and synthetic singing to make all melodies unknown to human scorers. Melodies were classified by three tonality levels (high/low/atonal) and by two levels in length (8 or 2 measures long). High tonality melodies consist of only notes on a diatonic scale, while low tonality melodies contain one or two notes included in back chord of dominant chord. Atonal tone sequences are designed so as to make listeners not feel any specific key by equally allocating 12 notes in an octave.

Singers

Human Singers

Singers employed in this experiment are six female undergraduate students in vocal course of music university. Singers were asked to sing given melodies or tone sequences presented in staff notation in a sound-proof room with "ta, ta, ta....". They were allowed to check tone height by piano just before entering into the recording room. They could confirm the melody or note sequence note by note playing the piano before recording. They were allowed to make re-recording as they wanted but they had to sing looking at the score without any instrument. So, they had to memorize the note sequence so as they can sing just looking the score even if the note sequence is hard to remember. Some of atonal sequences seemed hard to sing for some singers.

Table 1: Number of melodies used in the experiment

		Human Singing		Synthetic Singing	
# of measures		2	8	2	8
tonality	high	8	2	8	2
	low	8	2	8	2
	atonal	8	2	8	2

The Synthetic Singer

We used software singer "Vocaloid 2" or so-called "Miku Hatsune" as the singer for synthetic singing. It sings with voice of a young girl according to input score without accompnying instrumental sound. We can seletively add vibrato among several modes, speed and depth. The synthetic singing attains naturalness by putting build-up process of f_0 in voicing initial. Synthetic singing by "ta, ta, ta...." was recorded on-line directly from PC.

Figure 6 shows variation of vibratos by the singer software, where abscissa is time axis of 2 s long and the ordinate is f_0 in cents, among them we used normal vibrato in which f_0 fluctuates both sides of nominal pitch with depth about 130 cents, and speed about 4.5 Hz. Figure 7 depicts frequencyy spectra of vowel /a/ generated in A_2(=110Hz), A_3(=220Hz), A_4(=440Hz) and A_5(=880Hz).

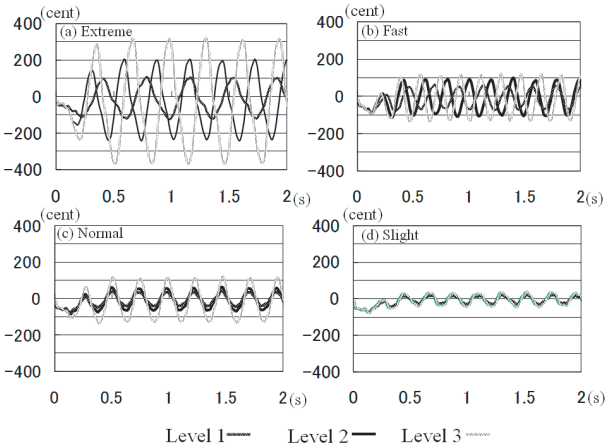

Figure 6: Four different modes of f_0 patterns
Each mode has 3 levels in vibrato depth with fixed/controlled vibrato rate.
abscissa: time, ordinate: f_0 in cents, 0=note height, 100cents=semitone

Subjects as Human Scorers

Table 2 is a list of human scorers employed for transcription of human singing. Each scorer was asked to transcribe six 8-measure melodies and 18 2-measure melodies sung by different combination of singers following two-set 6×6 Latin square. Scorers A, C, E, F, H and K in Table 2 were employed also for transcription of synthetic singing.

Instruction to the Subjects

We gave the following instructions to human scorers.

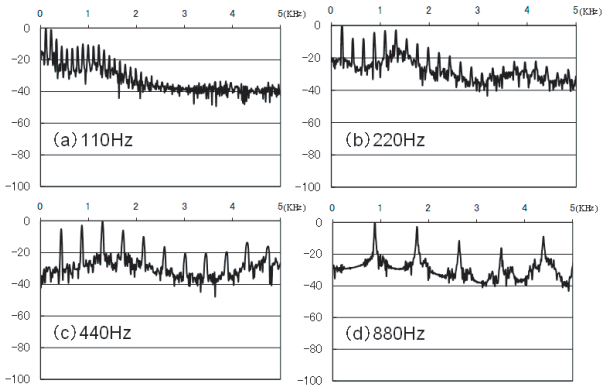

Figure 7: Frequency spectra of synthetic /a/ in A_2(=110Hz), A_3(=220Hz), A_4(=440Hz), and A_5(=880Hz)

Table 2: Scorers for human singing

scorer ID	sex	specialty
A	M	Prof., Composition
B	M	Prof., Composition
C	M	Researcher, Musical Inf. Proc.
D	F	Prof., Composition
E	F	Prof., Music Therapy
F	F	Prof., Infant Education
G	M	Graduate student, Pf.
H	M	Graduate student, Guitar
I	M	Graduate student, Cond.
J	F	Graduate student, Pf.
K	F	Graduate student, Oboe
L	F	Graduate student, Pf.

- Transcribe the melodies recorded in the CD. You can play-back any times as you want in your own order, You can rewrite your answers. You can use musical instruments if you want.
- You can use either of note names for enharmonics, for example C# can be noted also as D♭.
- You can use syllable names in “movable Do” system instead of note names, as our current primary interest is perception of relative intervals on a scale system.
- You are not asked to put bars in answer sheet, as note-value is not target of our interest in this experiment.
- In case you feel the singer made mistakes, you can modify the height of your answer from what you perceive.

Evaluation policy

Correct recognition rate defined as follows is our principal index for evaluating transcription performance.

$$R = \frac{\text{Correct Answers} - \text{Exess Answers} - \text{Missing Answers}}{\text{Number of notes in the task score}} \times 100(\%) \tag{5}$$

However, “Correct Answer” is not clear in case of human singing as the song might not be faithful to the given score. So, some feasible method to determine the correct answer is to be assumed. One possibility is majority answer of human professionals. Two kinds of answers are presumed to be possible correct answers: one is the original score presented to singers assuming that they sang with considerable accuracy, and the other is majority answer by human scorers assuming that singer might have failed in singing and most of scorers perceived as their majority answer. In case synthtic singing is employed, the output singing is thought to be reliable in pitch and note-value though it lacks naturalness.

EVALUATION OF TRANSCRIPTION PERFORMANCE

Evaluation on Human Singing

Figure 8 compares correct recognition rate of 12 human scorers and that of our system for human singing. The left figure is the results asuuming that correct answer is the orinal score presented to singers, while the right figure is the results assuming that majority answer by human scorers is the correct answer. Looking at the results arranged in the order of tonality degree, tonality dependency of correct recognition rate is obvious for both human scorers and the automatic transcription system.

The fact that “correct recognition rate assuming majority answer to be the correct answer is much better than that assuming the original score to be the correct answer” seems to declare that human singing contained many mistakes for atonal melodies in particular. The fact that “assuming majority answer to be the correct answer also increases correct recognition rate by the system” indicates that the system acts somewhat similar to human performance. However, as far as assuming majority answer to be the correct answer, correct recognition rate by the system cannot exceed that of human scorers.

As the current system does not have facilities related to tonality, except simple evaluation of diatonic nature of the scale, correct recognition rate of the sytem is expected to be almost the same despite the degree of tonality. So, evaluation using singing data having known correct answer is required to evaluate transcription performance of the system properly.

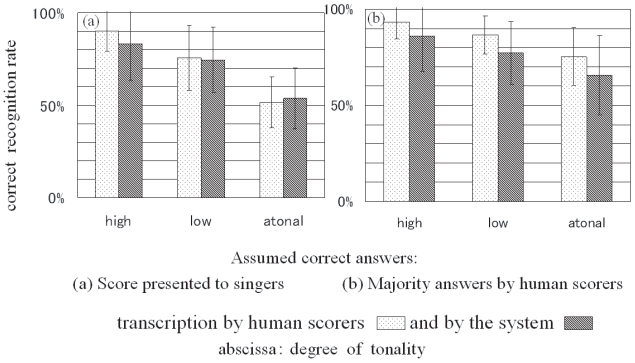

Figure 8: Comparison between correct recognition rates by human scorers and by the system for human singing from a point of tonality degree of the sung melodies employing two different criteria for correct answers

Figure 9 is the counter results for synthetic singing, for which correct answer is definitely the original score. Looking into Fig. 9 we can see that correct recognition rate of the system is almost the same regardless of tonality degree as we expected. Figure 9 says that correct recognition rate of the system is superior to that of human scorers regardless of tonality degree, though no significant difference is recognized in high tonality case.

Effect of length

Figure 10 compares correct recognition rate of human scorers and that by the transcription system for synthetic singing. There is no significant difference in results of different length of presented melodies, though clear difference is recognized between correct recognition ratio by human scorers and that by the system. The result for the human scorers means that two measure length is enough for tonality processing or eight measure length is still insufficient for tonality processing. Anyway, rssults by the system are the same regardless of tonality degree of the melody.

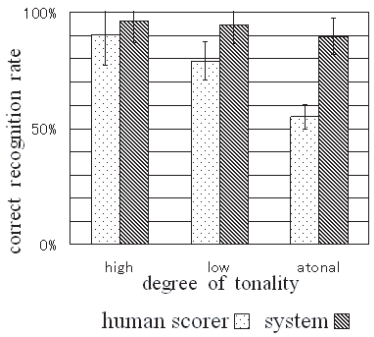

Figure 9: Tonality dependency of correct recognition rate for synthetic singing

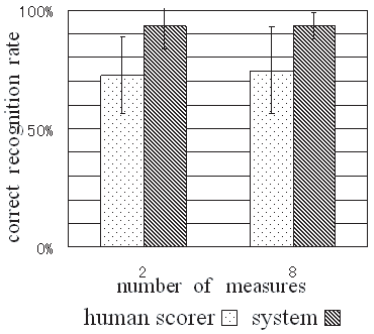

Figure 10: Dependency of correct recognition rate on length of presented singing for synthetic singing

INTERPRETATION OF THE RESULTS

Appropriateness of evaluation procedure

Looking at Fig. 8 and Fig. 9, we would dare say that we could get quite reasonable results. For human singing, performance of human scorers are better than the system for high tonality melodies in either of the correct answer cases. Assuming the majority answer of human scorers be the correct answer, both the correct recognition rate increases for less tonal melodies and atonal tone sequences, while performance for tonal melodies remain the same. The reason for that is singers' faults in singing less tonal melodies and atonal tone sequences. As far as we assume majority answer to be correct answer, performance of the system never exceed that of human scorers.

For synthetic singing, the system showed far better performance for atonal tone sequences in particular. Although no significan difference is recognized between the system and human scorers, the average score of the system is slightly better even for tonal melodies. One human scorer, however, showed slightly better performance than the system for highly tonal melodies. Fig. 11 is the result of the scorer. This scorer shows very high performance for highly tonal melodies and atonal tone sequences compared with other scorers. Transcription errors for atonal tone sequence are less than the average error of all scorers participated in this experiment. Quantitatively speaking, his error rate for atonal tone sequences is 22%, while the average errror rate is 45%. Judging from his error rate, transcribing atonal tone sequence seems to be hard even for him.

DISCUSSIONS

The principal objective of this research is to find how we can improve transcription performance of our current automatic scoring or transcription system by comparing system performance with human performance. Knowing that transcription performance of the current system is superior to average of

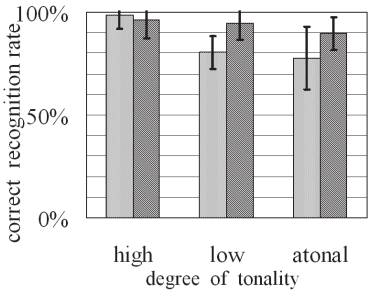

Figure 11: Results of a subject who showed the best marks beating the system for high tonality melodies

human scorers, tasks left for future are improvements in very delicate processing to make it be able to manage tough situations where the current system fails but some human scorers succeed.

Typical examples of that kind are something like shown in Fig. 12. The bottom figure shows trajectory of extracted f_0, top score is human transacription and the second top score is the system output. Three circled notes in the system output are erroneous. You may recognize that note-values are doubled in human transcription compared with the systen output, but we ignore the absolute values by extracting the time ratio. Also we neglect metrical structure, so we admit the difference that the last note in the top score is a half note but the corresponding part in the lower score is divided into two tied quarter notes. We regard these are different expressions of the same contents neglecting the bars.

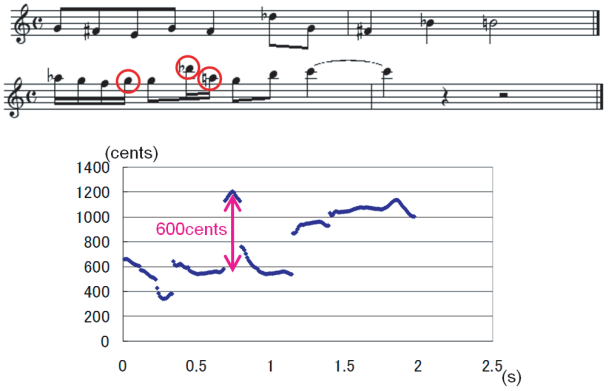

Figure 12: Examples of human modification in transcribing a melody.

Also you will find pitch difference between the top score and the second score by 11 semitones. Our evaluation ignores the absolute height but puts focus on interval progression. In that case, actual jump from 5th tone to 6th tone is about 620 ccents (though nominal interval is perfect 5th or 700 cents). In such case, human scorers often regard the interval to be perfect 5th instead of augmented 4th or diminished 5th because of their queerness, assuming that the target pitch of the extremal point should have been s little bit higher than perceived.

However the system extracts f_0 of the top position a llitle bit lower than actual instantaneous f_0 because of lowering effect of windowing with a definite length. To make the system modify to act like humans do is easy, but if we add that facility to the system, the system may produce new errors as by-productive reactions. So we should be careful to modify the system.

CONCLUSIONS

Reasonable and expected results were obtained concerning comparative study on transcription performance between our system and human scorers employing both human singing and synthetic singing. We have learned that we should use test songs whose correct answers are known to us, otherwise we cannnot evaluate the answers. Obtained results are summarized as follows:

- Transcription performance of the system is better than that of human scorers for low tonality melodies and atonal tone sequences, while there is no significant difference for highly tonal melodies, though the current system has no facility to make use of tonality for transcription.
- Transcription performance of the system is the same regardless of tonality degree of the given melody.
- Performance of human scorers gets worse drastically along with decreasing tonality.
- Correct recognition rate could not reach 100% even for synthetic singing. The reason for it would be failure in extracting f_0 in fast passages together with difficulties in determining note boundaries.
- Failure in extracting f_0 usually occurs in short notes or sharp summits or dips in f_0 trajectories.
- Possible candidate plan for improving the performance is to introduce mechanism for assuming target value of f_0 in case of fast f_0 transition.

ACKNOWLEDGEMENTS

This work has been originated in graduate course study of Mr. Masahide Onji, currently working with Mitsubishi Aero Space, through studies by Mr. Daiske Kimachi and Mr. Kohei Yasui. Authors express their thanks to the frontier works of the three graduate students.

REFERENCES

[1] John Cage: "*Notations*", Something Else Press, NY, 1969.

[2] F. Pachet and A. Zils: "Evolving Automatically High-Level Music Description from Acoustic Signals", *Proc. Intern. Symp. on Computer Music Modeling and Retrieval*, Montpellier, France, pp.42-53, May, 2003.

[3] Hugues Vinet: "The representation levels of music information", *Proc. Intern. Symp. on Computer Music Modeling and Retrieval*, Montpellier, France, pp.193-209, May, 2003.

[4] H. K. Taube : "*Notes from the Meta Level*", Taylor & Francis, London and New York, 2004.

[5] A. Ghias, J. Logan, D. Chamberlin and B. C. Smith: "Query by humming: musical information retrieval in audio database", *Proc. of 3rd ACM intern. Conf. on Multimedia*, pp.231-236, 1995.

[6] R. Typke, F. Wiering and R. Veltkamp: "A survey of music information retrieval systems", in J. D. Reiss and G. A. Wigginns(Eds.): "*Proc. 6th ICMIR*", pp.163-160, London, Queen Mary, U. of London, 2005.

[7] David Huron: "Themefinder", http://www.themefinder.org and Helmut Schaffrath: "*Essen Folksong Collection*.

[8] Denys Parsons: "*The Directory of Tunes and Musical Themes*", Spencer Brown, 1975.

[9] Y. E. Kim, W. Chai, R. Garcua and B. Vercoe: "Analysis of a contour-based representation for melody", *Proc. Intern. Symp. on Music Information Retrieval*, 2000.

[10] P. Ferraro, P. Hanna, L. Imbert and T. Izard: "Accelerating Query-by-Humming on GPU", *Proc. 10th ISMIR, Kobe*, pp.279-284, 2009.

[11] W. B. Hewlett and E. Selfridge-Field: "*Melodic Similarity*", MIT Press, 1998.

[12] A. Klapuri and M. Davy : "*Signal Processing Methods for Music Transcription*", Springer, New York, 2006.

[13] R. Typke and L. Prechelt : 'An interface for melody input", *ACN Transactions on Computer-Human Interaction*, pp.133-149, 2001.'

[14] L. Diana, G. Haus and M. Longari: "Towards a General Architecture for Musical Archive Information Systems", in U. K. Wiil(Ed.): "*Computer Music Modeling and Retrieval*", Intern. Symp. CMMR2003, pp.23-33, Springer, 2004.

[15] R. J. McNab et al.: "Toward the digital music library: tune retrieval from acoustic input", *Proc. of 1st ACM Conf. on Digital Libraries*, pp.11-18, 1996.

[16] A. L. Uidetdenbogerd and J. Zobel: "Matching techniques for large music databases", in D. Bulterman, K. Jeffay and H. J. Zhang (Eds.): "Proc. ACM Multimedia Conf.", Orlando, pp.57-66, 1999.

[17] R. Bod: "A General Parsing Moddel for Music and Language", in C. Anagnostopoulou et al.(Eds.): "*Music and Artificial Intelligence*", Proc. 2nd ICMAI, Edinburgh, 2002, pp.5-17, Springer, 2002.

[18] T. Winkler: "Composing Interactive Music", MIT Press, 1998.

[19] N. EMURA, M. MIURA and M. YANAGIDA: "A modular system generating Jazz-style arrangement for a given set of a melody and its chord name sequence", *Acoustical Science and Technology*, Vol. 29, No. 1, pp.51-57, 2008.

[20] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering and R. van Oostrum: "Using Transportation Distances for Measuring Melodic Similarity", Tech. Rep., UU-CS-2003-024, Inst. of Inf. & Computing Sci., Utrecht Univ., 2003.

[21] J. Goldstein : "An optimum processor theory for the central formation of the pitch of complex tones", *J. Acoust. Soc. Am.*, Vol.54, pp.1496-1516, 1973.

[22] H. D. Thornburg and R. J. Leistikow : "A New Probablistic Spectral Pitch Estimator: Exact and MCMC-approximate Strategies", *Proc. 2nd Intern. Symp. on Computer Music Modeling and Retrieval*, Esbjerg, Denmark, pp.41-60, May, 2004.

[23] C. Chafe, D. Jaffe, K.Kashima, B. M-Reynaud and J. Smith : "Techiniques for Note Identificatiion in Polyphonic Music", *CCRMA Report*, No. STAN-M-29, Oct., 1985.

[24] H. Kameoka, T. Nishimoto and S. Sagayama: "Multi-pitch Detection Algorithm Using Constrained Gaussian Mixture Model and Information Criterion for Simultaneous Speech", *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing*, pp.533-536, 2004.

[25] Naoko Kosugi, Y. Nishihara, T. Sakata M. Yamamuro and K. Kushima: "A practical Query-by-Humming system for a large music database", Proc. 8th ACM IC on Multimedia, Los Angels, pp.333-342, 2000.

[26] I. NAKAYAMA and M. YANAGIDA: "Introduction to database of traditional Japanese singing with examples of comparative studies on formant shifts and vibrato among genres", *Acoustical Science and Technology*, Vol. 29ĄCNo. 1, pp.58-65, 2008.

[27] M. Onji, J. Shimizu, M. Miura and M. Yanagida : "Automatic scoring of hummed songs", *Proc. APSCOM2005*, Seoul, pp.169-174, Aug. 4-9, 2005.