IEICE TRANS. INF. & SYST., VOL.E100–D, NO.5 MAY 2017

PAPER

Robust Singing Transcription System Using Local Homogeneity in the Harmonic Structure

Hoon HEO[†a)], ***Student Member and*** **Kyogu LEE**[†,††b)], ***Nonmember***

SUMMARY Automatic music transcription from audio has long been one of the most intriguing problems and a challenge in the field of music information retrieval, because it requires a series of low-level tasks such as onset/offset detection and F0 estimation, followed by high-level post-processing for symbolic representation. In this paper, a comprehensive transcription system for monophonic singing voice based on harmonic structure analysis is proposed. Given a precise tracking of the fundamental frequency, a novel acoustic feature is derived to signify the harmonic structure in singing voice signals, regardless of the loudness and pitch. It is then used to generate a parametric mixture model based on the von Mises–Fisher distribution, so that the model represents the intrinsic harmonic structures within a region of smoothly connected notes. To identify the note boundaries, the local homogeneity in the harmonic structure is exploited by two different methods: the self-similarity analysis and hidden Markov model. The proposed system identifies the note attributes including the onset time, duration and note pitch. Evaluations are conducted from various aspects to verify the performance improvement of the proposed system and its robustness, using the latest evaluation methodology for singing transcription. The results show that the proposed system significantly outperforms other systems including the state-of-the-art systems.

key words: *automatic music transcription, harmonic structure, music information retrieval, singing voice*

1. Introduction

Automatic music transcription, which is one of the most traditional topics in the music information retrieval (MIR) field, refers to the task of extracting a musical notation in the form of symbolic data from audio recordings. It encompasses a broad range of tasks in music signal processing, such as note onset detection, pitch estimation, and multi-instrument separation. Automatic music transcription is applicable in various fields. With the advances in mobile technologies, music education on mobile platforms are becoming popular for learning and training. Many mobile apps provide a real-time tutoring service for beginners who want to learn to play instruments such as guitar, piano, or violin. In such applications, users' performance recorded through a built-in microphone is transcribed into note-level data, and users are guided to play the given music score correctly. Further, music transcription can give useful information for higher-level MIR tasks, such as query-by-humming and melodic similarity analysis.

In most related studies, a musical note is defined by three components: onset, duration and pitch. Since the late 1990s, many approaches have been proposed to detect the onset time, defined by the exact time when a note starts [1]. In general, onsets can be categorized as hard and soft onsets depending on the attack time, which is the time taken for initial run-up of the amplitude envelope. Soft onsets that commonly appear in singing voices or in sustained string instruments such as the violin, are usually more difficult to detect, because the changes in acoustic features such as the energy envelope is very gradual and insignificant. Duration refers to the time for which the note is played; therefore, it is equal to the offset time minus the onset time of a note. Pitch is a quantitative value representing how high or low a sound is. Pitch detection algorithms estimate a sequence of successive pitch values at the frame level, which are typically defined by the fundamental frequency (F0 henceforth) in Hz or are given by MIDI note numbers. For monophonic music signals, the accuracy of pitch estimation algorithms has already reached a high level. One of the most popular pitch trackers called YIN [2] achieved an average gross error rate of 1.03%, which is still competitive today. When the input signal is a human voice such as speech or singing, it can be more reliable by using the bone-conducted signal [3].

Although the human voice is a type of musical instrument that can "perform" in the easiest way, automatic transcription for the singing voice still needs improvement. According to the Music Information Retrieval Evaluation eXchange (MIREX), the F-measures in the singing voice onset detection for the last five years have been around 0.6, which is 30% less than the results of other solo instruments. Compared to typical solo instruments, some difficulties in note detection are commonly found in singing voice signals. Note events often arise in very unpredictable ways, and it is difficult to define a single acoustic pattern. From various singing voice signals, it is observed that this unpredictability is mostly caused by two factors: loudness inconsistency and spectral heterogeneity. In singing, the dynamic range of loudness is not stable; rather, it varies among singers and their singing styles. In addition, the spectral distribution in singing depends on the pronunciation, whereas other instruments have their own timbral characteristics.

Despite all these difficulties, singing voice signals have a clear benefit for transcription. F0 estimation for the singing voice has reached a reliable level because it is ba-

Manuscript received September 8, 2016.
Manuscript revised December 20, 2016.
Manuscript publicized February 18, 2017.

[†]The authors are with the Music and Audio Research Group, Graduate School of Convergence Science and Technology, Seoul National University, Seoul 08826, Korea.
[††]The author is also with the Advanced Institutes of Convergence Technology, Suwon 16229, Korea.
a) E-mail: cubist04@snu.ac.kr
b) E-mail: kglee@snu.ac.kr

DOI: 10.1587/transinf.2016EDP7387

sically monophonic. A precise tracking of the F0 sequence can give useful information to identify not only the pitch but also important temporal attributes such as the onset and offset. Since McNab introduced a simple segmentation method for singing transcription using the pitch and amplitude [4], many approaches have been based mostly on the discontinuity in the F0 sequence. An auditory-model-based method uses the pitch continuity, together with the loudness and voicing patterns [5]. Rynnänen combined two probabilistic models to detect natural notes in a musicological sense [6]. More recently, Gómez and Bonada proposed an iterative note-consolidation technique using low-level features related to the pitch, duration, voicing and stability [7]. Molina presented a note segmentation method based on pitch-time hysteresis, making use of the dynamic average of the pitch curve [8].

However, the pitch-based approach has a problem that it cannot detect smoothly continued notes with the same pitch. Frequently observed in singing and humming, these notes can be detected using instantaneous changes in other acoustic properties. In this respect, this study begins with a hypothesis that both the beginning and the end of a note are recognized by the local homogeneity in the harmonic structure. The basic idea of using temporal changes in the harmonic structure was first attempted by using the regularity of the harmonic-related cepstrum [9]. We extend a similar approach here to make full use of the harmonic structure as an important cue for detection of note boundaries. The goal of this work is to propose a comprehensive transcription system that converts a singing voice recording into a western music score. The proposed system is presented in a unified framework, which includes extraction of a novel acoustic feature reflecting the harmonic structure, a probabilistic model for classifying the intrinsic harmonic structure, and transcription schemes for identifying the musical attributes.

The rest of this paper is organized as follows. Section 2 explains a front-end stage for F0 tracking, and describes the extraction of an acoustic feature to signify the harmonic structure. A probabilistic model to characterize the feature is also presented, followed by a transcription of note attributes such as the onset, offset, and note pitch. Section 3 presents the evaluation methodology to assess the proposed system, and the experimental results including the comparison with other systems are shown in Sect. 4. Finally, the conclusions of this paper are drawn in Sect. 5.

2. Proposed System

In the proposed system, a *stream* is defined by a region with continuous voiced F0s, which is divided by unvoiced frames. A stream may contain several notes smoothly continued, or may consist of only one note. There are two strategic benefits when the transcription process is allocated for each stream. It enables an efficient mixture model (described in Sect. 2.3) as it does not necessarily consider the whole range of an input audio. In addition, the system can be composed in a clear and unified framework because it does not need any exceptional treatments for unvoiced regions. The overall workflow of the entire proposed system is shown in Fig. 1.

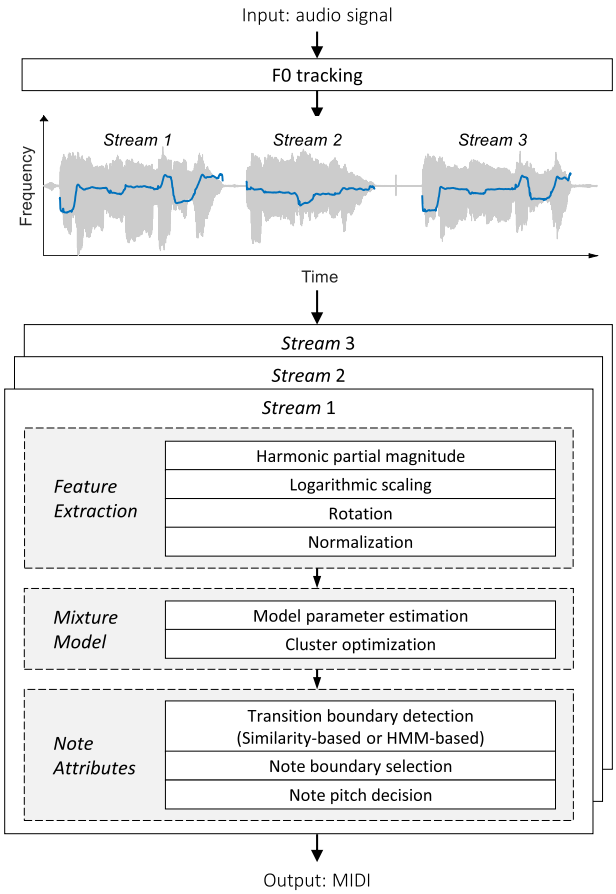

Fig. 1 Schematic flow underlying the proposed transcription system.

2.1 F0 Tracking

Before the local homogeneity in the harmonic structure is directly mentioned, a precise F0 tracking should precede it to identify the harmonic partials. In this work, it is implemented by a robust pitch tracker called PYIN [10]. This algorithm is chosen as a front-end F0 tracker of the entire transcription system due to its strength against "octave errors," which means that estimates are sometimes doubled (or half) frequencies. In order to enhance the original YIN algorithm, PYIN selects a few F0 candidates by taking valleys in the difference function of the input signal. After that, the probability of each candidate is calculated by observations in a hidden Markov model (HMM) for temporal smoothing of the F0 track, which is determined by the optimal path of pitch state decoded by the Viterbi algorithm.

The pitch space was defined from 65 Hz (C2) to 830 Hz (G#5) to cover the vocal pitch range of non-professional singers. It was divided in a step of 1/4 semitones, yielding 140 voiced pitch states in total. The same number of

unvoiced pitch states were concatenated with these voiced pitch states to construct the HMM. In the tracking result, some frames could be labeled as unvoiced if their corresponding path indicated an unvoiced state (weak probabilities of F0 candidates) or if the root-mean-square value was less than 0.1 (weak signal energy). Observation probabilities were calculated using a parameter prior modeled by the beta distribution with means 0.25, which is slightly greater than the parameter configuration that the original authors used. This is because the priority of the proposed system is a high recall, which means it aims to estimate as many frames as possible of the voiced F0s.

2.2 Feature Extraction

Extraction of an acoustic feature that reflects the harmonic structure begins with the magnitude of the harmonic partials. The use of harmonic partials has been introduced in many previous works for different tasks, such as music source separation [11] and vocal activity recognition [12]. In this work, we focus on the point that the relative ratio between the harmonic energies remains constant, regardless of the external factors including the pitch and loudness. The feature extraction process consists of the two following steps: (1) Extraction of harmonic partial magnitude and (2) Vector transformation such as scaling, rotation and normalization.

The first step of feature extraction is a time-frequency representation of an input signal using the short time Fourier transform. The input signal is downsampled to 22.05 kHz for a better computation time, and a Blackman window of 32 ms is used to split the signal into frames. Only the magnitude spectrum is considered, and the phase information ignored.

Instead of taking the magnitude at particular harmonic frequency bins, the tracking of harmonic partials is realized by a dynamic filter bank, whose frequency response is dynamically characterized by the estimated F0. The used filter bank is a series of overlapping triangular band-pass filters, so that the center frequency of one filter is equal to the lower boundary of the next filter. The center frequency of each filter is obtained from the multiple integers of the estimated F0. All the filters show a maximum response of unity at their center frequency.

The use of a filter bank offers advantages in two aspects. First, it compensates the errors arising from insufficient frequency resolution. Some algorithms [13], [14] use a multi-resolution FFT to enhance both the time and frequency resolution. However, a recent study has shown there is no significant benefit in locating the spectral peak frequency [15]. Second, frequencies slightly deviating from the exact integer multiples of the F0 can be considered. Inharmonic partials are rarely ever observed in cases of singing, but the spectral peak width can be relatively wide when the pitch is sharply changing within a frame.

The harmonic partial magnitude is not refined enough to be used as a feature vector for the harmonic structure in two respects: energy dynamics and imbalance in dimensions. The deviation in energy is too large to be characterized, and most of the spectral energy is concentrated in the first few harmonic partials. Therefore, the harmonic partial magnitude is transformed into a more refined form of feature called the Harmonic Structure Coefficient (HSC), by the three following steps of scaling, rotation and normalization.

Let a column vector $\mathbf{u} = [u_1\ u_2\ \cdots\ u_h]^T$ denotes the magnitude for up to the h-th harmonic partial at a time instance. The logarithmic scaling

$$\mathbf{x} = \log_{10}(\mathbf{u} + 1) \tag{1}$$

converts the magnitudes into non-negative values in a limited range, thereby making the data more stable for abrupt events. One example of this is the mel-scale filterbank cepstral coefficient (MFCC), which is the most popular acoustic feature that represents the timbral texture.

Log-scaled magnitudes are then rotated in such a way that the eigenvector with the minimum eigenvalue is parallel to the mean vector of a stream. This vector rotation allows the data distribution to be grouped easily when it is projected onto a unit hypersphere. Given a sequence of log-magnitude vectors $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_N]$ with a stream length of N, its distribution can be expressed by the mean vector $\mu_\mathbf{x}$ and the covariance matrix $\mathbf{C}_\mathbf{x} = \mathrm{Cov}(\mathbf{X})$, reflecting the center point and the spreadness in the h-dimensional Euclidean space, respectively. Since $\mathbf{C}_\mathbf{x}$ is a $h \times h$ square matrix, eigen decomposition $\mathbf{C}_\mathbf{x} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ can be applied to find the eigenvectors and eigenvalues of $\mathbf{C}_\mathbf{x}$. Then, the eigenvector $\mathbf{q}_{\min}$ with the minimum eigenvalue is chosen to determine the rotation angle.

The generalized form of the rotation matrix between two arbitrary vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as [16]

$$\mathbf{R} = \mathbf{I} - \mathbf{u}\mathbf{u}^T - \mathbf{v}\mathbf{v}^T + \begin{bmatrix}\mathbf{u} & \mathbf{v}\end{bmatrix}\begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\mathbf{u} & \mathbf{v}\end{bmatrix}^T \tag{2}$$

where

$$\mathbf{u} = \frac{\mathbf{a}}{\|\mathbf{a}\|},$$
$$\mathbf{v} = \frac{\mathbf{b} - (\mathbf{u}\cdot\mathbf{b})\,\mathbf{u}}{\|\mathbf{b} - (\mathbf{u}\cdot\mathbf{b})\,\mathbf{u}\|},$$
$$\theta = \arccos\frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}.$$

The first three terms in Eq. (2) find a projection onto the rotation subspace using the orthonormal basis $\mathbf{u}$ and $\mathbf{v}$. The last term performs a two-dimensional rotation on a plane generated by two vectors $\mathbf{a}$ and $\mathbf{b}$, and maps it back to the original dimension. By substituting with $\mathbf{a} = \mathbf{q}_{\min}$ and $\mathbf{b} = \mu_\mathbf{x}$, the rotation is fixed with the angle between $\mathbf{q}_{\min}$ and the mean vector $\mu_\mathbf{x}$. This allows that the feature vectors are widely dispersed when projected onto the hypersphere, by keeping the basis with the lowest spreadness parallel to the mean vector. Figure 2 illustrates a graphical example of the two-dimensional vector rotation, showing two distinct groups in

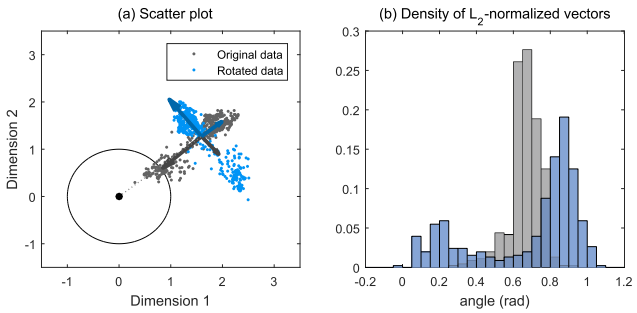

Fig. 2 A two-dimensional example of the vector rotation. (a) A scatter plot of the original and the rotated data. Eigenvectors and eigenvalues are depicted by the direction and the length of arrows. (b) A density plot of angles for both data when normalized onto the unit circle.

the rotated data when the normalization is applied.

As the final step, the HSC is defined by the rotation around the mean vector followed by normalization.

$$\mathbf{y} = \mathbf{R}\left(\mathbf{x} - \mu_{\mathbf{x}}\right) + \mu_{\mathbf{x}} \tag{3}$$

$$\mathrm{HSC} = \frac{\mathbf{y}}{\|\mathbf{y}\|} \tag{4}$$

The rotation enables to find the best perspective to interpret the clustered data, while preserving the relative information between dimensions. Besides, the normalization removes absolute information about the energy, thus the HSC only includes the relative information between the harmonic partials. In other words, the HSC eventually contains only the essential information to represent the harmonic structure, regardless of other acoustic properties such as pitch and loudness.

2.3 Parametric Mixture Model

As mentioned in the previous section, it is assumed that perception of a note boundary is closely related to a significant transition of the harmonic structure. If a stream contains several notes with different pronunciations that can be clearly distinguished, the HSCs would form several clusters on the surface of the unit hypersphere. Ideally, the number of clusters would be equal to the number of vowel pronunciations. Unsupervised classification is known as a standard solution for identifying these clusters; however, clustering methods such as the K-means or Gaussian mixture model are not suitable for the data in this particular distribution. Alternatively, a mixture model based on the von Mises–Fisher distribution is used here.

The von Mises–Fisher (vMF) distribution provides a suitable model to fit the data on the surface of a multidimensional unit sphere. The vMF distribution is applied in recent topics of information retrieval such as text mining, allowing it not to have a huge bias towards only a few words with highly frequent occurrence [17]. It is parametrized by the mean direction μ and the concentration parameter κ, which refers to the spread of the distribution around the mean. Its probability density function (pdf) for the h-dimensional unit vector x is defined by

$$p\left(x|\mu,\kappa\right) = \frac{\kappa^{h/2-1}}{(2\pi)^{h/2}\, I_{h/2-1}\left(\kappa\right)} e^{\kappa x^T \mu} \tag{5}$$

where $I_r\left(\kappa\right)$ is the modified Bessel function of the first kind at order r.

In the mixture model, the Expectation-Maximization (EM) algorithm is used to estimate the mean and concentration parameters of each vMF distribution as formulated by Banerjee [18]. In a general EM framework, the model may converge to a local maximum of the likelihood function depending on setting the initial point, and it does not guarantee that the model is correctly converged to the global maximum. To avoid this, all the steps of parameter estimation are repeated 10 times with different initial points, and the iteration is selected for which the log-likelihood sum is maximized. The mean vector of randomly selected samples, for which the mixing proportions are uniform, gives the initial point.

As all vectors belong to the $(h-1)$-sphere, the mean vector should be calculated in the $h-1$ dimensional angular coordinate, instead of the Euclidean space. The angular coordinates ϕ_i can be converted from the Cartesian coordinates $x_1, \ldots, x_h$ as

$$\phi_i = \arccos \frac{x_i}{\sqrt{x_h^2 + x_{h-1}^2 + \cdots + x_i^2}} \tag{6}$$

where $i = 1, 2, \ldots, h-1$. For a special case of $x_h < 0$, $\phi_{h-1} = 2\pi - \arccos \frac{x_{h-1}}{\sqrt{x_h^2 + x_{h-1}^2}}$. Given N sample vectors, the mean angle of each coordinate $\bar{\phi}_i$ is calculated by

$$\bar{\phi}_i = \mathrm{atan2}\left(\mathrm{Im}\left(\bar{z}_i\right), \mathrm{Re}\left(\bar{z}_i\right)\right) \tag{7}$$

$$\text{where } \bar{z}_i = \frac{1}{N}\sum_{n=1}^{N} e^{j\phi_i}. \tag{8}$$

The mean vector $\bar{\mathbf{x}} = [\bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_h]^T$ is finally obtained by the inverse transformation from the angular coordinates as follows:

$$\bar{x}_i = \begin{cases} \sin(\bar{\phi}_1)\cdots\sin(\bar{\phi}_{h-2})\cos(\bar{\phi}_{h-1}) & \text{if } i < h \\ \sin(\bar{\phi}_1)\cdots\sin(\bar{\phi}_{h-2})\sin(\bar{\phi}_{h-1}) & \text{if } i = h \end{cases} \tag{9}$$

Meanwhile, estimating the optimal number of mixture components (i.e. clusters) is not a simple issue, especially when the statistical characteristic of the data is not specified. In this work, fortunately, it is possible to assume roughly that the number of notes is proportional to the length of the stream. A heuristic regression approximated the correlation between the stream length and the note count. Using the ground truth in the dataset (see details in Sect. 3.1), streams were first segmented so that each stream was divided by a short interval (> 0.1 s). By counting the notes for each stream, it was noticed that the maximum note count could be roughly approximated to five times the stream length in seconds. To contain unnecessary clusters for a short transition, the maximum number of clusters is limited to five so that clusters are generated for only significant harmonic

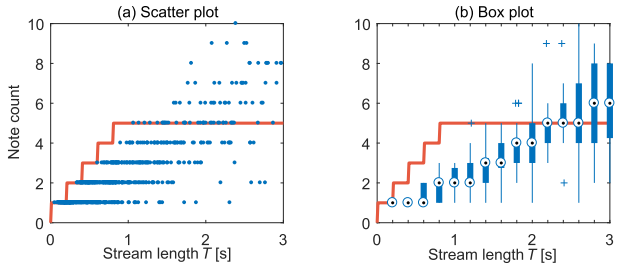

Fig. 3 Note counts by different stream lengths and the heuristic regression of the maximum number of clusters. Each dot in the scatter plot represents a stream. Variances in the box plot are shown with stream groups divided in a step of 0.2 s. The regression function $g(T) = \min(\lceil 5T \rceil, 5)$ is depicted by the red line.

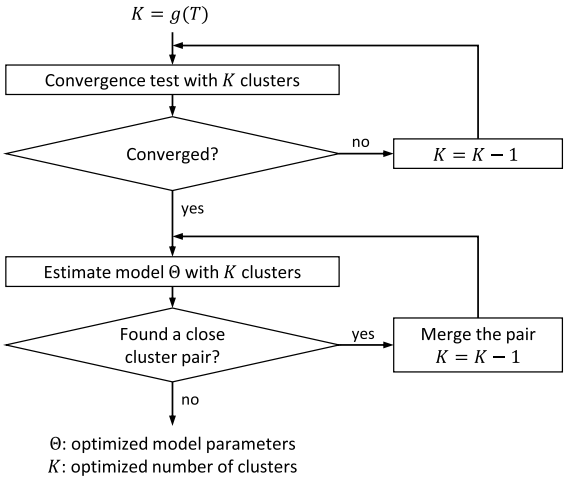

Fig. 4 Flowchart on the cluster optimization.

structures. Figure 3 shows the approximation of the initial number of clusters using the actual note counts.

In practical cases, streams may contain less notes than the maximum number. Moreover, the number of intrinsic harmonic structures can be even lower when some notes have the same vowel pronunciation. To this end, an iterative method is developed to optimize the number of clusters as shown in Fig. 4, using the regression function of the maximum number of clusters.

Once the maximum number of clusters is initially determined by the stream length, the largest number of clusters that the mixture model converges within 100 EM iterations is found first. Next, by decreasing the number of clusters K, the EM algorithm is repeated to estimate the model parameters $\Theta = \{\mu_{1\cdots K}, \kappa_{1\cdots K}\}$, as long as the distance between the means of two clusters is shorter than a threshold $d_{\min}$. Since all the cluster means are located on the $(h-1)$-sphere, the distance is defined by the arc length between two points on the unit hypersphere,

$$d = \arccos \boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_j, \qquad 0 < d \leq \pi. \tag{10}$$

A close pair of clusters is merged by taking the mean vector of the two cluster means, and the initial points of the next vMF model are determined by the mean vector and all the other cluster means. This method is based on the agglomerative clustering, a bottom-up approach to merge pairs that are closely formed. It is advantageous to make the final clusters as distant to each other as possible.

2.4 Note Attributes

This sub-section describes the methods for determining the three basic attributes of a note: the onset, offset, and note pitch. Significant transitions in the harmonic structure are primarily detected to identify note boundaries. Then, the actual onsets and offsets are selected from the harmonic structure transitions, and a single pitch that represents a note will be finally decided.

2.4.1 Transition Boundary Detection

Detection of the harmonic structure transition is achieved in two different methods. The first builds a detection function representing the degree of local changes in the feature, using the self-similarity (or self-distance) analysis. The self-similarity analysis has been used mainly for music segmentation since early studies [19], [20]. The purpose of these works is to automatically find some points of significant structural transitions in music, such as a chorus after verses. In this work, a similar technique is applied at the note level to detect onsets instead of segments. A self-similarity matrix is obtained by subtracting from one the cosine distance between two HSC vectors, i.e.,

$$S_{i,j} = 1 - \mathrm{HSC}_i \cdot \mathrm{HSC}_j \tag{11}$$

where HSC_n denotes a row vector of the harmonic structure coefficient at the n-th frame. Note that the denominator of the cosine distance formula is removed since the L_2 norm of the HSC is unity. The novelty function is determined by

$$Novelty(n) = \sum_{i=-N/2}^{N/2} \sum_{j=-N/2}^{N/2} W_{i,j} \cdot S_{n+i,n+j} \tag{12}$$

where W is a Gaussian-tapered checkerboard kernel [21] sliding alongside the diagonal elements of the self-similarity matrix. A small kernel allows the detection of short notes but increases the chance of false positives. Conversely, a large kernel can be considered when the transcription system should avoid detecting spurious notes. In order to locate the transition boundaries, all the peaks (i.e., local maxima) in the novelty function are found first, and only the peaks higher than a peak-picking threshold δ_{peak} are chosen. Note that this similarity-based method does not use the mixture model.

Although this approach is quite simple and easy to understand, choosing a proper peak-picking threshold heavily affects the transcription performance. Thus, another approach based on the hidden Markov model (HMM) is proposed as well, applying the parametric mixture model. The proposed HMM consists of a transient state and the same

number of sub-HMMs as clusters from the mixture model. Each sub-HMM contains several one-way states to model a harmonic structure with a minimum duration constraint. This constraint prevents the state path from fluctuating instantaneously, as the state path is forced to stay in a cluster for $T_{\min}$ seconds at least. All transition probabilities are determined by an input parameter, which decides the probability of staying in the current cluster or the transient state. This "self-transition probability" parameter controls the sensitiveness of the note event detection. If they become closer to unity, the transition is less likely to occur, thus less number of notes will be detected.

Observation probabilities are given by a function of the likelihood $p(x|\mu,\kappa)$ of each cluster as defined in Eq. (5). Since the pdf can be greater than unity by its definition, the pdf is so normalized that the probabilities sum to unity at every instance of time. Given the normalized pdf $p_{k,n}$ for all K clusters, the observation probabilities are calculated in the range between 0 and 1 as

$$b_{k,n} = \begin{cases} p_{k,n} \cdot \exp(p_{k,n} - 1) & \text{(sustain state)} \\ \sum_{k=1}^{K} \dfrac{\Delta p_{k,n+1} + \Delta p_{k,n}}{2} & \text{(transient state)} \end{cases} \tag{13}$$

where $\Delta p_{k,n} = \left| p_{k,n} - p_{k,n-1} \right|$. The observation probabilities of the transient state are determined by changes in the pdf of the clusters. At the end, the prior probability is uniformly given to all clusters and the transient state. After the three HMM parameters are determined for all the states, the optimal state path $v = \{v_1, \ldots, v_N\}$ is decoded by the Viterbi algorithm. Accordingly, transition boundaries, at which the state path changes from the transient state to a sustain state, are simply detected.

2.4.2 Note Boundary Selection

Arguably, the transition boundaries indicate the points at where the harmonic structure significantly changes. However, not all transitions are directly converted into the note onset, because some voiced consonants such as [l], [m] and [ŋ] can be included. These voiced consonants, commonly observed in humming, may cause low detection accuracy, if they are detected as independent notes. Therefore, it is necessary to exclude the voiced consonants from the note boundary, using their distinguishing spectral characteristic due to the nasal sound.

Let $x_{i,t}$ denotes the log-magnitude of the i-th harmonic partial at a time instance t. Mean height $\bar{\delta}_\tau$ at a transition boundary time τ is defined by

$$\bar{\delta}_\tau = \frac{1}{h} \sum_{i=1}^{h} \left(\max_{t \in (\tau, \tau+T)} x_{i,t} - \min_{t \in (\tau-T, \tau)} x_{i,t} \right) \tag{14}$$

where $T = T_{\min}/2$. When a voiced consonant is followed by a normal vowel, the harmonic partial magnitude decreases except in the first few partials. A note boundary is selected at τ only if $\bar{\delta}_\tau > \delta_{\text{note}}$, and determines *onset* and *offset*.

2.4.3 Note Pitch Decision

When the note boundary and F0s are given, the simplest way to decide the note pitch would be by taking a mean or median value of the F0s between the onset and the offset. In singing, however, it is sometimes difficult to specify a single value of the F0s within a note. Singing voices often

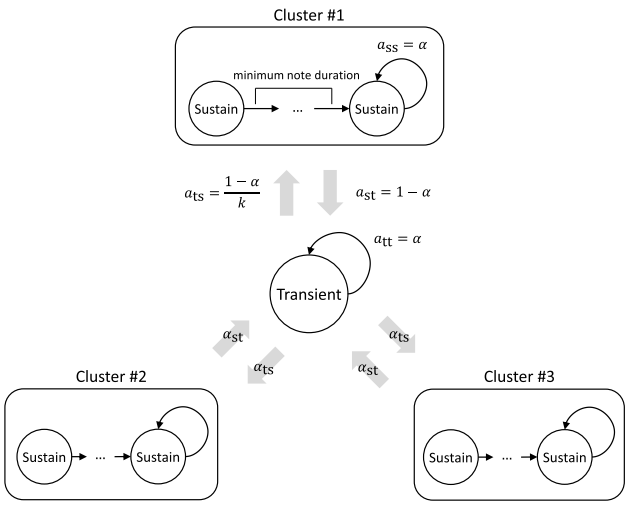

Fig. 5 Transitions in the HMM.

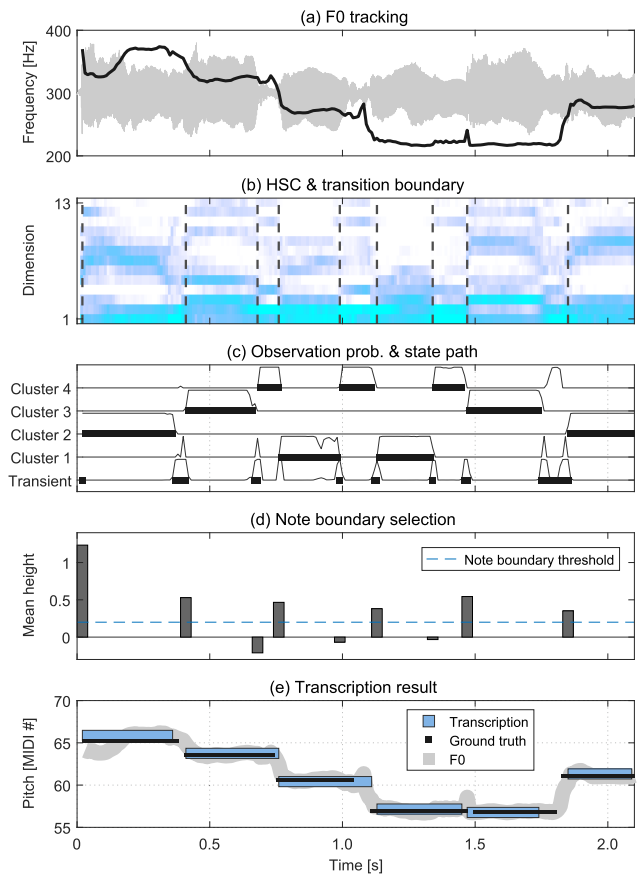

Fig. 6 Transcription result from an excerpt of `afemale10.wav` in the dataset.

include musical expressions and ornaments such as a grace note, which is a separate pitch prefixed to a principal note. The longest region for which the pitch deviations are kept below a tolerance of 50 cents (100 cents = 1 semitone) is selected, and the pitch at the beginning of the region decides the note pitch. In doing so, a note pitch that is most likely to be perceived can be chosen.

Figure 6 summarizes the whole transcription process of a female singing voice signal. Panel (b) illustrates the HSC and eight detected transition boundaries. Panel (c) shows the observation probabilities of the HMM and the corresponding state path. Mean height for each transition boundary is depicted in the panel (d), showing six of them are selected as note onset. In the last panel, the transcription result is displayed in the form of a piano-roll representation. It is notable that two connected notes with the same pitch (the fourth and the fifth note) are correctly transcribed.

3. Evaluation

3.1 Dataset

Evaluations have been conducted using a publicly available dataset [22], [23], released for the purpose of evaluation on singing transcription. The dataset consists of 38 audio recordings of monophonic singing, recorded with a sample rate of 44.1 kHz and a 16-bit resolution. The melodies in the dataset come from several excerpts of popular songs including The Beatles. Singers are categorized in three classes: adult males (13 recordings), adult females (11 recordings), and children (14 recordings). The pitch and loudness are quite unstable as the singers are untrained. The duration of the whole dataset is up to 19 minutes 15 seconds in total. All the recordings were very freely performed with musical articulations and ornaments.

The dataset also contains the note-level ground truth by manual annotations. The ground truth provides annotation of the onset, offset, and note pitch for all the 2154 notes in the dataset. The onset and offset are given by their exact time in seconds, and the note pitch is by a MIDI note number with two decimal places. The MIDI note number is converted from the frequency in Hz by $12\log_2(\text{frequency}/440) + 69$.

3.2 Criteria and Measure

Precision and recall have been commonly considered the standard measures for binary classification such as the onset detection. Combining the precision and the recall, the F-measure is the most representative measure for an overall performance. However, a note transcription system needs to adopt more extensive criteria, because it includes the overall evaluation for the three note attributes. Thus, recent criteria were extended particularly for singing transcription [8]. The qualitative meanings in the criteria are described as follows:

- COnPOff (correct onset, pitch and offset): The most restrictive criterion, meaning the correct rate of onset (±50 ms), offset (±20% of the ground-truth note duration or 50 ms, whichever is larger) and pitch (±0.5 semitones). A note is correctly transcribed only if its onset, offset, and pitch satisfy the criteria simultaneously.
- COnP (correct onset and pitch): A less restrictive criterion, accounting for both the onset and pitch, using the same size of tolerance window as above.
- COn (correct onset): Similar to the above two criteria, but only onset is considered in this case. This is equal to the traditional metric for onset detection.
- Split: The rate of ground truth notes incorrectly segmented into consecutive notes by transcription.
- Merge: The rate of ground truth notes merged as they are transcribed into the same note (complementary to Split).
- Spurious: The rate of transcribed notes not having any overlap with ground truth notes (neither in time nor pitch domain).
- Non-detected: The rate of ground truth notes not having any overlap with transcribed notes (neither in time nor pitch domain).

Note that COnPOff, COnP, and COn are chosen as major criteria for the overall performance of note transcription. Each criterion has its numerical measures such as precision, recall and F-measure. For other criteria such as Split, Merge, Spurious and Non-detected, the measures are expressed by the rate of incorrectly transcribed notes that each criterion defines, emphasizing the more specific points of wrong transcription.

3.3 Experimental Setup

Two evaluations were conducted in various aspects of singing transcription at the note level, rather than the assessment for the front-end pitch tracker at the frame level. This is because the existing algorithms for monophonic pitch tracking have already accomplished a reliable performance, and the proposed note transcription system is based on the assumption that the F0 is known.

The first evaluation assessed the transcription performances among two methods for transition boundary detection, and to examine the influence of different parameter configurations. By comparing the results, the best method and the most optimized parameter were determined. On the other hand, the second evaluation shows the improvement of the proposed system compared to other systems including the state-of-the-arts. For a fair comparison, the experiment was conducted under an identical experimental setup including dataset and metrics. The default parameter configuration in all the experiments is summarized in Table 1.

All the experiments were conducted on a personal computer with a 3.3 GHz CPU and 8 GB RAM. The computational time for the entire transcription system depends on whether the probabilistic models are included or not, and

Table 1 Parameter configuration.

Parameter	Description	Value
h	Number of harmonic partials	13
$d_{\min}$	Minimum cluster distance (rad)	0.25
δ_{peak}	Peak picking threshold	0.03
α	Self-transition probability	0.5
δ_{note}	Note boundary threshold	0.2
$T_{\min}$	Minimum note duration (s)	0.1

Table 2 Computational time of the proposed system.

	Using the given F0 track		Including F0 tracking	
	Similarity-based	HMM-based	Similarity-based	HMM-based
Time (s)	37	110	296	374

most of the time was spent on F0 tracking. The detailed time taken for all input signals with a total length of 1155 seconds is displayed in Table 2.

4. Result

As the first evaluation, the overall performance was compared by using different parameters, including the number of harmonic partials and the detection sensitivity. This experiment was conducted using the two methods for transition boundary detection, the similarity analysis and the HMM-based note event model. As shown in Fig. 7, the performance improvement was saturated in both methods with more than 11 partials, and the highest F-measure of 0.82 was achieved by the HMM-based method. As the number of partials increases, the performance of the similarity-based method slightly decreases while the HMM-based method does not change. It is also noticeable that the similarity-based method scored a very low performance when only a few partials were used. This result implies that the HMM-based method is more robust than the similarity-based method.

To verify the robustness of the HMM-based method more clearly, the precision-recall curve for both detection methods is reported in Fig. 8, showing the trade-off between precision and recall. The precision and the recall were obtained by varying the parameters δ_{peak} and α, which determine the detection sensitivity for the similarity-based and the HMM-based method, respectively. While the HMM-based method achieved a reliable performance for various detection sensitivities, the precision rapidly decreased in the similarity-based method as the peak-picking threshold increased. In most cases, it was reported that the recall tends to be greater than the precision.

Both experimental results show that the use of the mixture model not only improves the overall performance, but also accomplishes the robustness of the system. The similarity-based method is heavily influenced by the parameters and the characteristic of the input signal, since it is difficult to choose a proper threshold for peak picking. Whereas, the mixture model is effective for classifying the intrinsic harmonic structures in a stream, even when a limited number of partials are given. Nonetheless, the overall performance of the similarity-based method is still higher than the recent average results of the onset detection for the singing voice class in the MIREX. This infers that the HSC is a very effective feature to represent the harmonic structure, and is suitable for singing transcription even without the mixture model.

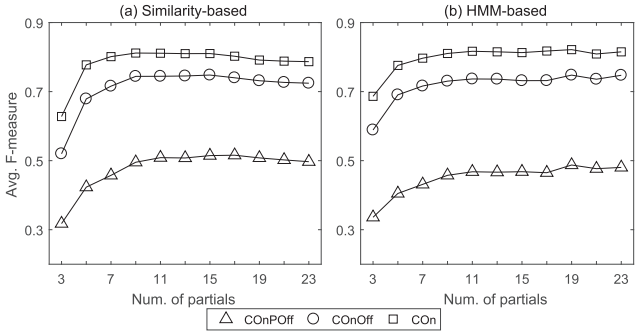

Fig. 7 Average F-measures in three criteria by different number of harmonic partials.

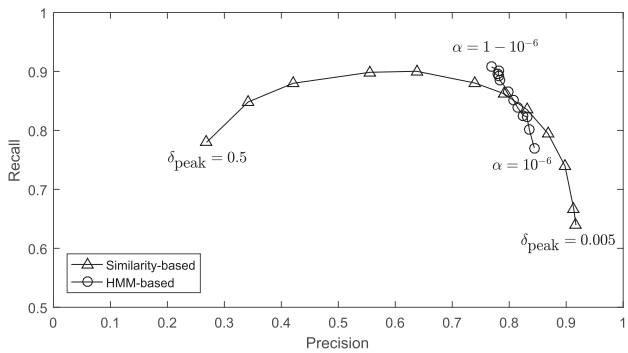

Fig. 8 Precision-recall curves in the COn criterion for two transition detection methods.

The second evaluation was conducted to compare the system performance with five other methods. All the results are excerpts from the original papers [8], [23] that use the same dataset and criteria. The results attained by Rynnänen's note event model approach [6], Gómez & Bonada's method [7], a commercial system named Melotranscript [24] were cited from Molina's evaluation framework [23]. The SiPTH system has only one overall performance about COnPOff, since the authors do not mention the result on COnP and COn in their paper [8]. In case of Tony [25], their best result was chosen (reported as pYIN s = 0.8, prn = 0.10) among different parameter configurations.

As shown in Fig. 9, the overall performance of the proposed system outperforms other systems including the state-of-the-arts. In case of COn, the best performance (average F-measure 0.82, 95% confidence interval 0.80 to 0.84) was achieved using the HMM-based method. The performance improvement on COnP becomes more significant compared to the first three systems. It implies that the local homogeneity within the harmonic structure, which is the most

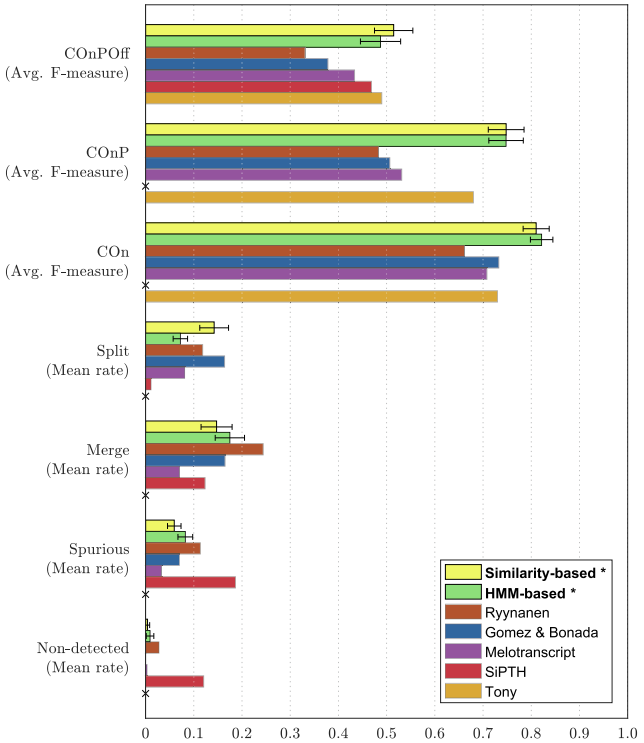

Fig. 9 Evaluation comparison of the proposed system (marked by asterisk) and other algorithms. Labels on the y-axis indicate the criteria and their numerical measure. Items marked by crosses are not publicly announced.

distinguishing point to other approaches, can be an effective feature for singing transcription, as it has an advantage that connected notes with the same pitch can be detected.

However, the proposed system did not improve much when the offset detection is included. The relatively low improvement on COnPOff can be explained by two factors. First, even with the feature normalization to remove the influence of the loudness, it cannot reflect the changes in harmonic structure as the singing becomes softer at the end of a note. Second, it may be caused by the ambiguity in the offset annotation for the singing voice.

Split and Merge are complementary to each other. As the detection sensitivity becomes higher, Merge decreases and Split increases. In the proposed system, the detection sensitivity mainly depends on the note boundary threshold δ_{note}. When it increases from 0.2 to 0.3, it was observed the system produces only Splits less than 0.05% of the entire ground truth notes, while the overall performance is still higher than others (over 80% COn). Since it cannot say that either Split or Merge is more critical, it is required to use appropriate settings depending on the purposes of transcription.

Although the proposed system accomplished the best overall performance, it is not always the best approach for all cases. One example is a stepwise pitch change with the same pronunciation, which can be easily detected by pitch-based systems. It is expected that the system can be further improved when the time-pitch curve is also considered.

5. Conclusion

A singing transcription system based on the analysis of harmonic structure was presented. Given the estimated F0 sequence, a novel acoustic feature called the harmonic structure coefficient (HSC) was derived by extracting the harmonic partial magnitude with several refinement steps of vector transformation. In doing so, the HSC is defined on the surface of a unit hypersphere, representing the relationship between harmonic partials.

A parametric mixture model based on the von Mises–Fisher distribution was used to characterize the feature space. Further, an optimization technique was proposed to determine the optimal number of clusters, so that the intrinsic harmonic structure could be correctly classified.

To detect significant transition boundaries in the harmonic structure, two different methods were presented based on the self-similarity analysis and the HMM. Then, note attributes were finally determined by excluding the voiced consonant from the detected transition boundaries.

The proposed system was evaluated using the latest evaluation methodology for singing transcription. Comparing results of the two proposed methods for transition boundary detection showed that the mixture model and the note event model improve the transcription performance and robustness. When comparing with the existing systems, the evaluation results clearly indicate that the proposed transcription system significantly outperforms other systems including the state-of-the-art systems.

References

[1] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler, "A tutorial on onset detection in music signals," IEEE Trans. Audio, Speech, Language Process., vol.13, no.5, pp.1035–1047, Sept. 2005.

[2] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., vol.111, no.4, pp.1917–1930, 2002.

[3] M.S. Rahman and T. Shimamura, "Pitch determination from bone conducted speech," IEICE Trans. Inf. & Syst., vol.E99-D, no.1, pp.283–287, Jan. 2016.

[4] R.J. McNab, L.A. Smith, I.H. Witten, et al., "Signal processing for melody transcription," Australian Computer Science Communications, vol.18, pp.301–307, 1996.

[5] L. Clarisse, J.P. Martens, M. Lesaffre, B. De Baets, H. De Meyer, and M. Leman, "An auditory model based transcriber of singing sequences," Proc. Int. Symp. Music Information Retrieval (ISMIR), Citeseer, 2002.

[6] M.P. Ryynänen and A.P. Klapuri, "Modelling of note events for singing transcription," ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing, 2004.

[7] E. Gómez and J. Bonada, "Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing," Computer Music Journal, vol.37, no.2, pp.73–90, 2013.

[8] E. Molina, L.J. Tardón, A.M. Barbancho, and I. Barbancho, "Sipth: Singing transcription based on hysteresis defined on the pitch-time curve," IEEE/ACM Trans. Audio, Speech, Language Process., vol.23, no.2, pp.252–263, Feb. 2015.

[9] H. Heo, D. Sung, and K. Lee, "Note onset detection based on harmonic cepstrum regularity," in Proc. IEEE Int. Conf. Multimedia and Expo. (ICME), pp.1–6, July 2013.
[10] M. Mauch and S. Dixon, "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP), pp.659–663, May 2014.
[11] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," IEEE Trans. Audio, Speech, Language Process., vol.16, no.4, pp.766–778, May 2008.
[12] M. Mauch, H. Fujihara, K. Yoshii, and M. Goto, "Timbre and melody features for the recognition of vocal activity and instrumental solos in polyphonic music," Proc. Int. Symp. Music Information Retrieval (ISMIR), pp.233–238, 2011.
[13] C.-L. Hsu and J.-S.R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1K dataset," IEEE Trans. Audio, Speech, Language Process., vol.18, no.2, pp.310–319, Feb. 2010.
[14] K. Dressler, "Audio melody extraction for mirex 2009," 5th Music Inform. Retrieval Evaluation eXchange (MIREX), vol.79, pp.100–115, 2009.
[15] J. Salamon and E. Gomez, "Melody extraction from polyphonic music signals using pitch contour characteristics," IEEE Trans. Audio, Speech, Language Process., vol.20, no.6, pp.1759–1770, Aug. 2012.
[16] S. Montgomery-Smith, "Finding the rotation matrix in n-dimensions," Mathematics Stack Exchange, URL: http://math.stackexchange.com/q/598782 (version: 2016-06-19).
[17] J. Reisinger and R.J. Mooney, "Multi-prototype vector-space models of word meaning," Human Language Technologies: The 2010 Ann. Conf. the North American Chapter of the Association for Computational Linguistics, HLT '10, pp.109–117, Stroudsburg, PA, USA, Association for Computational Linguistics, 2010.
[18] A. Banerjee, I.S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von mises-fisher distributions," J. Machine Learning Research, vol.6, pp.1345–1382, Sept. 2005.
[19] M. Cooper and J. Foote, "Summarizing popular music via structural similarity analysis," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.127–130, Oct. 2003.
[20] J. Foote, "Visualizing music and audio using self-similarity," in Proc. 7th ACM Int. Conf. Multimedia, MULTIMEDIA '99, New York, NY, USA, pp.77–80, ACM, 1999.
[21] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in Proc. IEEE Int. Conf. Multimedia and Expo. (ICME), pp.452–455, Aug. 2000.
[22] J. Salamon, J. Serrà, and E. Gómez, "Tonal representations for music retrieval: from version identification to query-by-humming," Int. J. Multimedia Information Retrieval, vol.2, no.1, pp.45–58, 2013.
[23] E. Molina, A.M. Barbancho, L.J. Tardón, and I. Barbancho, "Evaluation framework for automatic singing transcription," Proc. Int. Symp. Music Information Retrieval (ISMIR), pp.567–572, 2014.
[24] T.D. Mulder, J.P. Martens, M. Lesaffre, M. Leman, B.D. Baets, and H.D. Meyer, "Recent improvements of an auditory model based front-end for the transcription of vocal queries," in Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP), vol.4, pp.iv-257–iv-260, May 2004.
[25] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the tony software: Accuracy and efficiency," Proc. 1st Int. Conf. Technologies for Music Notation and Representation (TENOR), pp.23–30, 2015.

Hoon Heo received the B.S. and M.S. degree in Electrical Engineering from Seoul National University, Seoul, Republic of Korea, in 2008 and 2011, respectively. Currently, he is pursuing his Ph.D. degree at the Music and Audio Research Group at the Graduate School of Convergence Science and Technology at Seoul National University, Seoul, Korea. His research interests include automatic music transcription and various applications in music information retrieval.

Kyogu Lee received the B.S. degree in Electrical Engineering from Seoul National University, Seoul, Korea, in 1996, the M.M. degree in Music Technology from New York University, New York, in 2002, and the M.S. degree in Electrical Engineering and the Ph.D. degree in Computer-based Music Theory and Acoustics from Stanford University, Stanford, CA, in 2007 and 2008, respectively. He worked as a Senior Researcher in the Media Technology Lab at Gracenote from 2007 to 2009. He is now an associate professor at the Graduate School of Convergence Science and Technology at Seoul National University, Seoul, Korea, and is leading the Music and Audio Research Group. His research focuses on signal processing and machine learning applied to music/audio.