PROBABILISTIC TRANSCRIPTION OF SUNG MELODY USING A PITCH DYNAMIC MODEL

Luwei Yang[⋆], *Akira Maezawa*[†], *Jordan B. L. Smith*[‡], *Elaine Chew*[⋆]

[⋆]Queen Mary University of London, UK
[†]Yamaha Corporation, Japan
[‡]National Institute of Advanced Industrial Science and Technology, Japan

ABSTRACT

Transcribing the singing voice into music notes is challenging due to pitch fluctuations such as portamenti and vibratos. This paper presents a probabilistic transcription method for monophonic sung melodies that explicitly accounts for these local pitch fluctuations. In the hierarchical Hidden Markov Model (HMM), an upper-level ergodic HMM handles the transitions between notes, and a lower-level left-to-right HMM handles the intra- and inter-note pitch fluctuations. The lower-level HMM employs the pitch dynamic model, which explicitly expresses the pitch curve characteristics as the observation likelihood over f_0 and Δf_0 using a compact parametric distribution. A histogram-based tuning frequency estimation method, and some post-processing heuristics to separate merged notes and to allocate spuriously detected short notes, improve the note recognition performance. With model parameters that support intuitions about singing behavior, the proposed method obtained encouraging results when evaluated on a published monophonic sung melody dataset, and compared with state-of-the-art methods.

Index Terms— Singing transcription, pitch dynamic model, music processing

1. INTRODUCTION

We propose a solution to the melody transcription problem of converting an audio recording of a sung melody (usually monophonic) to symbolic note representation[1]. Transcribing some instruments is easier than others. When dealing with sung melodies, transcribing audio into an equal-tempered symbolic representation [1], is made challenging due to three issues [2]. First, the tuning frequency may vary with each singer. This deviation from the standard tuning frequency may cause the entire transcription to be a semitone above or below the ground truth. Furthermore, untrained singers tend to change the tuning frequency midway through singing. Second, the singing voice can introduce fluctuations of pitch within a note, making it difficult to segment and identify the sung pitch. For example, a singer may sing with a rich vibrato (wide frequency modulation), causing the system to transcribe this as two alternating MIDI pitches. Third, the singing voice often contains portamenti (smooth transitions between adjacent pitches) or pitch bends. These two features are often used by singers as a means of being expressive [3]. Fig.1 shows a real example of a sung melody with intra- and inter-note fluctuations.

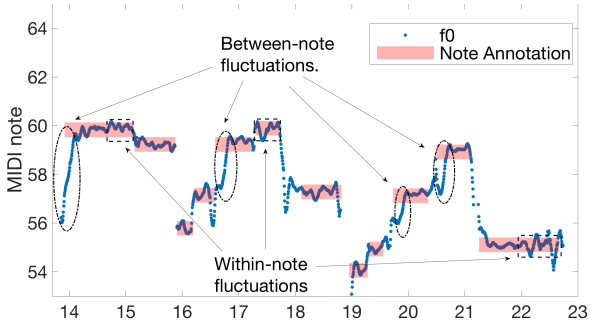

Fig. 1. Example of melody produced by a female singer with intra- and inter-pitch fluctuations.

Our method employs a hierarchical Hidden Markov Model (HMM) and a pitch dynamic model that is a two-dimensional distribution over the f_0-Δf_0 plane. An HMM-based singing transcription method was recently proposed in [4]. Another probabilistic method for finding the optimum note sequence was proposed in [5]. Dynamic averaging and hysteresis of the pitch (f_0) curve was employed in [1] to deal with the pitch fluctuations. The pitch dynamic model has been used to improve query-by-humming [6], to model singing style [7], and to synthesize singing [8]. To the best of the authors' knowledge, the pitch dynamic model has not been used in singing transcription. This paper proposes a hierarchical HMM-based sung melody transcription method

This research was supported in part by Joint Program of Queen Mary University of London and Beijing University of Posts and Telecommunications.

[1]Music excerpts of interest are assumed to be map-able to the equal-tempered chromatic scale, i.e. MIDI scale.

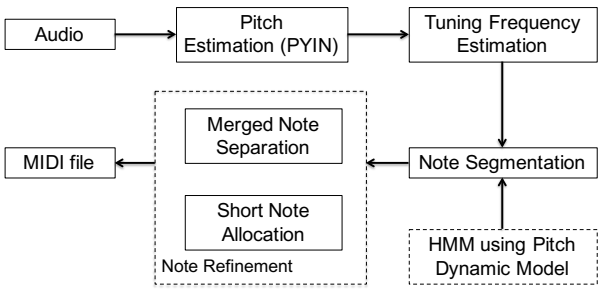

Fig. 2. A flowchart of the proposed method.

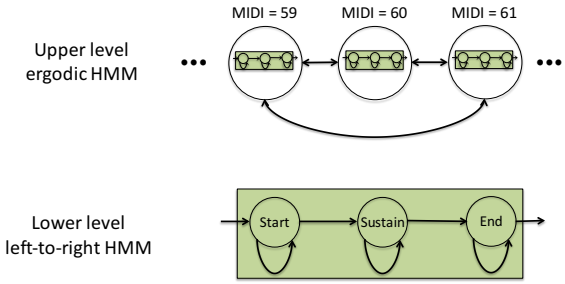

Fig. 3. Hierarchical hidden Markov model.

that incorporates the pitch dynamic model. This model explicitly serves as the observation likelihood distribution to accommodate the pitch fluctuations.

The structure of the paper is as follows: Section 2 presents the proposed method; Section 3 provides a comparison with state-of-the-art methods and discussions; and, Section 4 gives the conclusions and future work.

2. PROPOSED METHOD

The system flowchart of the proposed method is shown in Fig.2. The tuning frequency is obtained from the f_0 output using pYIN [9], and all pitches adjusted accordingly. This refined f_0 is fed into a hierarchical HMM together with the pitch dynamic model to segment and identify the notes. Following Viterbi decoding, a spectral flux-based method separates merged notes and a short-note reallocation method eliminates extraneous intermediate notes.

2.1. Tuning frequency estimation

As in [5], the method estimates a single tuning frequency for each excerpt. This is important because singers frequently sing with different tuning frequencies, more so for untrained musicians. The method we use simply shifts a melody by a constant less than 0.5 to the nearest whole number semitone.

To estimate the frequency shift required, we use the normalized histogram of f_0, given at one-cent resolution. Peaks exceeding an arbitrary threshold of 0.3 are selected, starting with the highest ones. Consecutive peaks added to the list must be at least 0.75 semitones apart. For each selected peak, we calculate two tuning frequency deviation values. The first is the *floor deviation*, which is the gap from the selected peak to the largest integer semitone smaller than itself. The second is the *ceiling deviation*, which is the gap from the peak to the smallest integer larger than itself.

The average of the floor and ceiling deviations among all peaks is used to create two adjusted f_0 series: the floor adjusted f_0, f_0^{floor}, subtracting the floor deviation from the original f_0; and the ceiling adjusted f_0, $f_0^{ceiling}$, adding the ceiling deviation to the original f_0. Note that we only consider deviations under one semitone; any gap more than one semitone will be transcribed to another MIDI note number. Three inputs, the original f_0, f_0^{floor}, and $f_0^{ceiling}$ are sent as input to the rest of the system. The one with the highest transcription performance is chosen.

2.2. Pitch dynamic model for note segmentation

Fig.3 shows the hierarchical HMM. The upper level ergodic HMM models the note transitions. The lower level left-to-right HMM models the note's start, sustain, and end states. These three states each emit the f_0 and its corresponding first-order derivative Δf_0. The model is inspired by [4] and [10]; unlike in existing studies, here the model is intended to explicitly model the dynamic characteristics of the pitch curve.

Fig.4 illustrates the basic idea behind the pitch dynamic model used to describe the sung melody dynamic characteristics. This model expresses the idea that each sung note is characterized by three distinct tendencies—start, hold steady (with or without vibrato), and end—each with specific f_0-Δf_0 behaviors.

Most often, a sung voice begins on a note other than the target pitch. This start state forms the beginning of the transition into the target pitch. When the sung voice enters from a lower pitch, Δf_0 is positive; when it enters from a higher pitch, Δf_0 is negative. Hence, the start state always sits in the second or fourth quadrants[2] of the f_0-Δf_0 plane. Upon arrival at the target pitch, the pitch curve tends to remain near the target or oscillate around it. This sustain state models variations within the target pitch. Oscillations are represented by circles around the origin in a clockwise direction. When the singer moves to the next target pitch, the end state models the outward transition. Since transitions out and in are complementary, the end state sits in the first or third quadrant (when moving to a higher or lower pitch, respectively) of the f_0-Δf_0 plane. Note that the singing can transition in and out either way.

To mathematically model the pitch dynamics, we create a novel two-dimensional distribution, which is a combination of the Gamma distribution and a variant of the von Mises

[2]The quadrants are numbered anti-clockwise starting from the top right.

distribution. Given a pair of f_0 and Δf_0, the probability of emitting from MIDI state m is

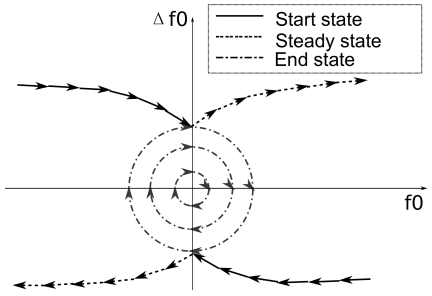

Fig. 4. Illustration of pitch dynamic model of f_0 and Δf_0 in sung melody.

$$p(r_m, \theta_m) = \frac{e^{(\kappa \cos(2(\theta_m - \mu)) - a r_m)} {r_m}^{b-1}}{2\pi I_0(\kappa)\Gamma(1+b)a^{-b}}, \quad (1)$$

where $I_0(\kappa)$ is the modified Bessel function of order 0, and r_m and θ_m are the corresponding polar coordinates,

$$r_m = \sqrt{[c(f_0 - m)]^2 + (d\Delta f_0)^2}, \quad (2)$$

$$\cos(\theta_m) = \frac{c(f_0 - m)}{r_m}, \quad (3)$$

where f_0 is the fundamental frequency on the MIDI scale, Δf_0 is the corresponding first-order difference, μ is the rotation angle during the start and end states, c and d are scaling parameters for f_0 and Δf_0, κ is a measure of the distribution concentrated around the rotating angle μ, a is the measure of the distribution concentrated around the origin, and b controls the distances between the peaks and the origin.

2.3. Note refinement

Merged note separation. Since we only use f_0 and Δf_0 for note segmentation, the above model sometimes classifies consecutive same-pitched notes (having no or only a small gap between them) as one single note. A peak in the spectral change corresponding to an onset serves as the cue to separate same-pitched consecutive notes. We use the spectral flux onset function in [11], which measures the spectral change over time summed across all frequency bins. When a spectral flux peaks in the middle of a note, that note is separated into two notes if each component note is at least 100ms long.

Short note allocation. Compared to melodies played on other instruments, sung melodies may have more portamenti or pitch bends at the start and end of each note. If the HMM enters a start state at the pitch bend, this will result in the addition of short-length spurious notes. Such spurious notes are typically identified as notes having durations less than a given threshold, and pruned [10]. Instead of deleting these notes, which could introduce onset and offset detection errors, we re-allocate very short notes to their previous or subsequent notes by comparing the weights given to these notes as follows: a neighboring note has a higher weight if it is closer in time and pitch to the very short note; two standard Gaussian functions are employed to give weights to these differences in time and pitch, and the values summed; the short note is then appended to the previous or subsequent note having the greater weight.

3. EVALUATION

We evaluated our proposed method using the sung melody dataset published in [12]. The dataset has 38 sung monophonic music pieces: 11 by adult females, 13 by adult males, and 14 by children.

3.1. State-of-the-art Methods

We compare our proposed method to four state-of-the-art methods: *Ryynänen* [4], *Gómez & Bonada* [5], *SiPTH* [1], and *Tony* [10]. We use the results for the first three methods as reported in [12]. *Tony*'s sensitivity parameter, s, was set to 0.8 and its note pruning threshold was set to 150ms. In addition, a *baseline* method—rounding the f_0 to the nearest MIDI note number and using any pitch changes as note boundaries, with note pruning using a 100ms threshold—provides additional comparison.

3.2. Parameter Settings

To match the human singing voice, the pitch range of interest was set between MIDI note number 35 (61.74Hz) and 80 (830.61Hz), with integer resolution. As a result, there are 46 states in the upper-level HMM, with a MIDI number for each state. Without any prior information such as the score or key, the transition probabilities were set as Gaussian distributions with the MIDI note's pitch as mean and 4 MIDI numbers (semitones) as standard deviation.

To simplify the problem, we make all upper-level HMM states identical. Each upper-level HMM state is associated with one lower-level HMM. Each lower-level HMM has an observation likelihood distributed represented by the pitch dynamic model.

For the lower-level HMM state transition probabilities, we select 0.1, 0.9, and 0.4 as the probabilities for the start, sustain and end state self-transitions, values found through a grid search for the best parameters matching the training dataset. Note that the upper- and lower-level HMMs could also be better tailored to specific styles.

The parameters for the pitch dynamic model are set as shown in Fig.5. The start state observation distribution only exists in the second and fourth quadrants, with the rotating angle at 135 degrees to the horizontal axis, which corresponds to the parameter settings: $\kappa = 0.5, a = 0.4, b = 1, c = 1, d =$

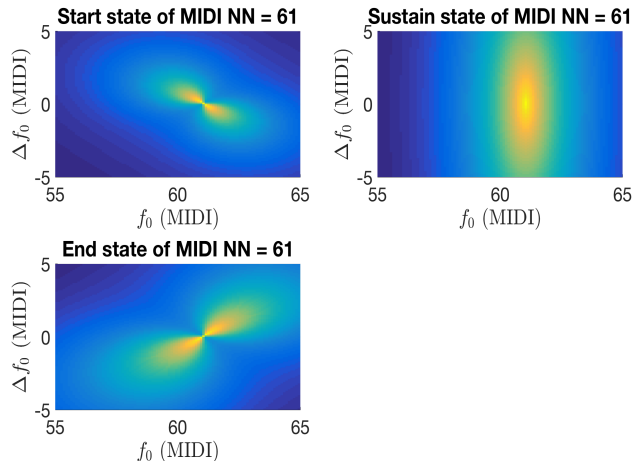

Fig. 5. The observation likelihood distribution using the proposed pitch dynamic model for MIDI state 61.

Method	Precision	Recall	F-measure
Baseline	0.254	0.223	0.234
Ryynänen [4]	0.304	0.315	0.308
Gómez & Bonada [5]	0.430	0.373	0.398
SiPTH [1]	0.397	0.440	0.415
Tony [10]	**0.510**	**0.534**	**0.520**
Proposed	0.409	0.436	0.421

Table 1. Note-level evaluation.

$1, \mu = \frac{3}{4}\pi$. The end state has the same parameters except $\mu = \frac{1}{4}\pi$ for a rotating angle of 45 degrees for the first and third quadrants; $a = 0.2$ in order to make the distribution more diffused in order to force the HMM to stay on the upper-level (MIDI) state. To make the sustain state open to sharp pitch bends (high Δf_0) and vibratos (low f_0 range), its parameters are set to $\kappa = 0, a = 0.1, b = 1, c = 5, d = 1, \mu = 0$.

3.3. Results and Discussions

Table 1 presents the precision, recall and F-measure of the note-level evaluation. Based on [12], we require the onset, offset, and pitch to be within a narrow threshold in order to consider a transcription to be successful. More precisely, a note is considered to be correct if the transcribed onset is within $\pm$50ms of the ground truth, pitch within $\pm$0.5 semitones of the ground truth, and offset within $\pm$50ms of the ground truth or result in a duration within $\pm 20\%$ of the ground truth, whichever is larger.

According to this measure, all methods performed better than the baseline approach. Tony achieves the best performance. The proposed method obtains the second highest F-measure and is comparable to SiPTH and Gómez & Bonada.

Table 2 shows the error rate for three different error types: erroneous splits, meaning the ground truth note is split into a number of separate transcribed notes; merge error refers to the case where a number of consecutive ground truth notes are merged into one transcribed note; and, spurious errors refer to cases where a transcribed note does not overlap with any ground truth note.

Method	Split	Merged	Spurious
Baseline	0.205	**0.146**	**0.057**
Ryynänen	0.105	0.248	0.116
Gómez & Bonada	0.140	0.167	0.071
SiPTH	0.074	0.309	0.157
Tony	0.079	0.230	0.112
Proposed	**0.064**	0.230	0.120

Table 2. Note-level error rate.

The proposed method obtains the lowest split error rate, which means that our method is better able to tolerate intra-note pitch curve fluctuations such as vibratos and unstable pitches. This may be due to the use of the pitch dynamic model to handle intra-note pitch fluctuations, and the short-note allocation procedure in the post-processing stage.

Note that our method, along with Ryynänen and Tony (which uses a similar HMM note model) and SiPTH, all have high merged note error rates. One possible explanation is that all these methods consider only the pitch curve. Consecutive notes—separate notes having no gap between them, or only a small one—having the same pitch are very hard to identify using only the pitch curve. Tony uses an amplitude-based onset detection to separate merged notes, while our proposed method uses the spectral-based onset detection to deal with this problem. Gómez & Bonada has fewer merged errors as it aggregates short notes to form longer ones; however, for the same reason, it has a split error rate higher than other methods. The baseline has the lowest merged and spurious error rates because the baseline method considers all pitch changes to be note boundaries; however, this strategy results in the highest split error rate.

4. CONCLUSIONS

In this paper, we have presented a pitch dynamic model-based probabilistic sung melody transcription method. We created a novel two-dimensional distribution, a combination of the Gamma distribution and a variant of the von Mises distribution, to capture the pitch dynamics in singing. This pitch dynamic model was applied to a hierarchical HMM. A spectral flux-based method was used to separate merged notes and a short note allocation method applied to enhance the results. Using model parameters that corroborated intuitions about singing behavior, the proposed model obtained encouraging results compared to state-of-the-art methods. Future work will explore the automatic derivation of model parameters for the pitch dynamic model and the hierarchical HMM.

5. REFERENCES

[1] Emilio Molina, Lorenzo J. Tardón, Ana M. Barbancho, and Isabel Barbancho, "SiPTH: Singing transcription based on hysteresis defined on the pitch-time curve," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 252–263, 2015.

[2] Matti Ryynänen, "Singing transcription," in *Signal processing methods for music transcription*, pp. 361–390. Springer, 2006.

[3] Johan Sundberg, "Expressivity in singing. A review of some recent investigations," *Logopedics Phoniatrics Vocology*, vol. 23, no. 3, pp. 121–127, 1998.

[4] Matti P Ryynänen and Anssi P Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.

[5] Emilia Gómez and Jordi 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.

[6] Yasunori Ohishi, Masataka Goto, Katunobu Itou, and Kazuya Takeda, "A stochastic representation of the dynamics of sung melody," in *Proc. of the 8th Internatioal Society for Music Information Retreval Conference*, 2007.

[7] Tatsuya Kako, Yasunori Ohishi, Hirokazu Kameoka, Kunio Kashino, and Kazuya Takeda, "Automatic identification for singing style based on sung melodic contour characterized in phase plane," in *Proc. of the 10th International Society for Music Information Retrieval Conference*, 2009.

[8] Takeshi Saitou, Masashi Unoki, and Masato Akagi, "Development of an f0 control model based on f0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, no. 3, pp. 405–417, 2005.

[9] Matthias Mauch and Simon Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 659–663.

[10] Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon, "Computer-aided melody note transcription using the Tony software: Accuracy and efficiency," in *Proc. of the 1st International Conference on Technologies for Music Notation and Representation*, 2015, pp. 23–30.

[11] Simon Dixon, "Onset detection revisited," in *Proc. of the 9th International Conference on Digital Audio Effects*, 2006, vol. 120, pp. 133–137.

[12] Emilio Molina, Ana M. Barbancho, Lorenzo J. Tardón, and Isabel Barbancho, "Evaluation framework for automatic singing transcription," in *Proc. of the 15th International Society for Music Information Retrieval Conference*, 2014, pp. 567–572.