SIMULATED FORMANT MODELING OF ACCOMPANIED SINGING SIGNALS FOR VOCAL MELODY EXTRACTION

Yu-Ren Chien,[1,2] **Hsin-Min Wang,**[2] **Shyh-Kang Jeng**[1,3]
[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Department of Electrical Engineering, National Taiwan University, Taiwan
yrchien@ntu.edu.tw, whm@iis.sinica.edu.tw, skjeng@ew.ee.ntu.edu.tw

ABSTRACT

This paper deals with the task of extracting vocal melodies from accompanied singing recordings. The challenging aspect of this task consists in the tendency for instrumental sounds to interfere with the extraction of the desired vocal melodies, especially when the singing voice is not necessarily predominant among other sound sources. Existing methods in the literature are either rule-based or statistical. It is difficult for rule-based methods to adequately take advantage of human voice characteristics, whereas statistical approaches typically require large-scale data collection and labeling efforts. In this work, the extraction is based on a model of the input signals that integrates acoustic-phonetic knowledge and real-world data under a probabilistic framework. The resulting vocal pitch estimator is simple, determined by a small set of parameters and a small set of data. Tested on a publicly available dataset, the proposed method achieves a transcription accuracy of 76%.

1. INTRODUCTION

Music lovers have always been faced with a large collection of music recordings or concert performances for them to choose from. Whereas successful choices are possible with a small set of metadata, disappointment recurs because the metadata only provides limited information about the musical contents. This has motivated researchers to work on systems that extract musically relevant features from audio recordings. One potential benefit of such processing would be the possibility that machines will be able to make personalized music purchase decisions on behalf of humans.

In this paper, we focus on the extraction of *vocal melodies* from polyphonic audio signals. A melody is defined as a succession of pitches and durations; as one might expect, melodies represent one of the most significant features that can be identified by listeners from musical pieces. In various musical cultures including popular music in particular, predominant melodies are commonly carried by singing voices. In view of this, this work aims at analyzing a singing voice accompanied by musical instruments. Instrumental accompaniment is common in vocal music, where the main melodies are exclusively carried by a solo singing voice, with the musical instruments providing harmony. In brief, the goal of the analysis considered in this work is finding the fundamental frequency of the singing voice as a function of time.

The specific task outlined above is challenging because melody extraction is prone to interference from the accompaniment unless a mechanism is in place for distinguishing human voice from instrumental sound. [1], [2], and [3] determined the predominant pitch as it accounts for the most of the signal power among all the simultaneous pitches. The concept of pitch predominance is also presented in [5] and [6], which defined the predominance in terms of harmonicity. For these methods, the problem proves difficult whenever the signal is dominated by a harmonic musical instrument rather than by the singing voice. [7] and [8] realized the timbre recognition mechanism by classification techniques; on the other hand, pitch classification entails quantization of pitch, which in turn causes loss of such musical information as vibrato, portamento, and non-standard tuning.

The singing voice is probably the oldest mechanism in human history for music performance. It shares considerable acoustic characteristics with speech, which have been formulated analytically in acoustic phonetics [9]. However, a typical acoustic-phonetic model involves some free parameters, i.e., the formant frequencies, which are highly variable across vowels or singers. In view of this, we take a probabilistic approach to vocal melody extraction, by which acoustic knowledge and real-world data can be integrated in a unified manner.

With an accompanied singing signal observed, estimation of the vocal pitch is based on the *pitch likelihood* (likelihood function of the pitch), which is in turn based on the *voice likelihood* (likelihood function of the singing voice). By simulating the singing voice signal, the pitch likelihood can be approximated by an average of values of the voice likelihood evaluated at the simulated set of voice examples. The simulation is realized by synthesizing voice signals of various timbres in advance according to formant frequencies extracted from a wide variety of (possibly accompanied) singing recordings. Since formant frequencies represent spectrum envelopes of the human voice, their extraction does not require the sampled singing recordings to

densely cover various pitch values, nor is it impaired by accompaniment of modest loudness in the recordings.

The proposed method offers several potential advantages over previous approaches to vocal melody extraction. First of all, imposing acoustic-phonetic constraints on the extraction enables the proposed method to better distinguish human voice from instrumental sound than the predominant pitch estimators in [1–3, 5, 6]. Secondly, the acoustic-phonetic constraints save the proposed method from large-scale data collection and labeling efforts that are common for purely data-driven systems [7]. Third, some systems [4, 10] depend on pitch instability in identifying the vocal pitch; in contrast, without discriminating between stable and unstable pitches, the proposed method allows for such cases as an unstable instrumental pitch (e.g., violin) or a stably sung vocal pitch. Fourth, although our earlier approach in [15] was also based on acoustic-phonetic knowledge, it did not statistically model the joint distribution of formant frequencies, nor did it model the accompaniment signal whatsoever. The signal model proposed here for accompanied singing promises to better represent vocal characteristics and handle interference from the accompaniment. Lastly, we highlight the advantage of the proposed method over the method in [8]. These two methods are interestingly related to each other, both adopting spectrum envelope modeling and the Viterbi algorithm. In spectrum envelope modeling, [8] extracts linear-predictive and cepstral features from sinusoidally resynthesized vocal or instrumental sounds, while our approach models vocal spectrum envelopes by formant-synthesizing voice examples. The proposed signal model turns out 1) to be applicable to both vocal pitch estimation and voicing detection, and 2) not to depend on any sound samples of musical instruments.

2. OVERVIEW OF VOCAL PITCH ESTIMATION

To facilitate the estimation, we quantize the vocal pitch into a discrete variable with 88 possible values. The pitch at k quarter tones ($k = 1, 2, ..., 88$) is associated with a fundamental frequency of $440 \cdot 2^{(k-60)/24}$ hertz. Therefore, the 88 pitch values are quarter-tone-spaced samples of the fundamental frequency in the vocal range from 80 hertz to 1,000 hertz.

The fact that we are now estimating a discrete-valued signal (i.e., the vocal pitch sequence) from an observed signal (i.e., the accompanied singing) makes it possible for us to characterize the pair of signals with a hidden Markov model (HMM) and find the best pitch sequence by the Viterbi algorithm [11]. Here, the accompanied singing signal is represented by a vector-valued observation sequence, which consists of 100 N-vector observations per second. Each N-vector observation is made up of N consecutive time samples of the signal. Obviously, there are 88 states in the HMM. As one might expect, the HMM is defined by two probabilistic models: the *observation model* and the *prior model*. The observation model describes the probability distribution of an observation given a particular state, while the prior model comprises the state transition probability and initial state distributions.

3. OBSERVATION MODEL

Let the accompanied singing signal, the (unobserved) singing voice, and the vocal pitch at a particular time point be denoted by the random N-vector $\mathbf{z}$, the random N-vector $\mathbf{x}$, and the random variable w, respectively. Then, the likelihood function of w, i.e., the pitch likelihood, can be expanded as a marginalizing integral:

$$p_{\mathbf{z}|w}(\mathbf{z}|w) = \int p_{\mathbf{z}|w,\mathbf{x}}(\mathbf{z}|w,\mathbf{x}) p_{\mathbf{x}|w}(\mathbf{x}|w) \mathbf{dx}. \quad (1)$$

With the term $p_{\mathbf{z}|w,\mathbf{x}}(\mathbf{z}|w,\mathbf{x})$ taken as a function of $\mathbf{x}$, this integral can be thought of as the expectation of $p_{\mathbf{z}|w,\mathbf{x}}(\mathbf{z}|w,\mathbf{x})$ and approximated by the corresponding sample mean:

$$p_{\mathbf{z}|w}(\mathbf{z}|w) \approx \frac{1}{N_e} \sum_{i=1}^{N_e} p_{\mathbf{z}|w,\mathbf{x}}(\mathbf{z}|w, \mathbf{x}^{(i,w)}), \quad (2)$$

where N_e is the number of voice examples available for each of the 88 pitch values, and $\mathbf{x}^{(i,k)}$ denotes the ith voice example for pitch k. Here, the voice examples $\{\mathbf{x}^{(i,w)}\}_{i=1}^{N_e}$ simulate the random experiment underlying the probability distribution described by the density $p_{\mathbf{x}|w}(\cdot|w)$. Given the singing voice $\mathbf{x}$, the vocal pitch w can be regarded as a constant, which is independent of any other random quantity; as a result, w can be dropped from the right-hand-side condition in (2):

$$p_{\mathbf{z}|w}(\mathbf{z}|w) \approx \frac{1}{N_e} \sum_{i=1}^{N_e} p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}^{(i,w)}), \quad (3)$$

which is an average of the values of the voice likelihood $p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\cdot)$ as evaluated at the voice examples $\{\mathbf{x}^{(i,w)}\}_{i=1}^{N_e}$.

The preparation of the voice examples will be presented in Section 3.1, which is an offline procedure performed well in advance of melody extraction. After that, we will describe the evaluation of the likelihood of each voice example, i.e., $p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}^{(i,k)})$, in Section 3.2.

3.1 Synthesizing Voice Examples

Comprehensive collection of real-world singing voice data is difficult, as results from several facts about the singing voice. In the first place, most vocal performances are accompanied, which renders unaccompanied singing voice recordings extremely scarce. Although non-professional unaccompanied singing data can be collected with less difficulty, untrained singing voice is typically of less practical relevance as compared with professional singing. Secondly, the pitches most often used in a song are confined 1) on the scale of its key and 2) within the registers of the singer; consequently, it would take a huge number of songs collected to have the entire vocal pitch range covered densely. Finally, to provide timbral variety, the collection must include various singers and various voiced sounds (vowels, nasal consonants, etc.).

To circumvent the difficulty in collecting singing voice data, we 1) collect accompanied singing data, 2) extract vocal spectrum envelopes from the data, and 3) synthesize voice examples of various pitches from the extracted envelopes. (These will be described in Sections 6.1, 3.1.1,

and 3.1.2, respectively.) Since vocal spectrum envelopes follow a well-defined formant structure, they can be extracted reliably in the presence of instrumental sounds, as long as the singing voice is sufficiently loud in comparison with the instruments. Moreover, by giving a pitch-independent description of timbre, the vocal spectrum envelopes eliminate the need for covering various pitches in data collection. In this way, sufficient data can be collected, for the sole purpose of representing the timbral diversity in singing voice.

3.1.1 Extracting Vocal Spectrum Envelopes

A vocal spectrum envelope is an amplitude function of frequency that models the spectrum envelope of a particular voiced sound (a vowel, a nasal consonant, etc.). By giving partial amplitudes as its samples at partial frequencies, it provides a pitch-independent description of the specific timbre of the voiced sound. In our implementation, it is determined by seven parameters: the first five oral formant frequencies $f_1, f_2, ..., f_5$ (hertz), a nasal formant frequency f_p (hertz), and a nasal anti-formant frequency f_z (hertz) [12]. To be more specific, it is defined by (see [9])

$$A(f^h) = 20\log_{10}\left|U_R(f^h)K_R(f^h)\prod_{n\in I_f} H_n(2\pi f^h)\right|, \tag{4}$$

where $A(\cdot)$ is the amplitude function in dB, f^h denotes the frequency in hertz, $U_R(\cdot)$ represents the (radiated) spectrum envelope of the glottal excitation [9]:

$$U_R(f^h) = \frac{f^h/100}{1+(f^h/100)^2}, \tag{5}$$

$K_R(\cdot)$ represents all formants of order six and above [9]:

$$20\log_{10} K_R(f^h) \approx 0.43\left(\tfrac{f^h}{500}\right)^2 + 7.1\cdot 10^{-4}\left(\tfrac{f^h}{500}\right)^4, \quad f^h \le 5000, \tag{6}$$

$I_f = \{1,2,3,4,5,p,z\}$, and $H_n(\cdot)$ represents frequency response of formant n [9]:

$$H_n(\omega) = \frac{1}{\left(1-\frac{j\omega}{\sigma_n+j\omega_n}\right)\left(1-\frac{j\omega}{\sigma_n-j\omega_n}\right)}, \quad n = 1,2,3,4,5,p, \tag{7}$$

$$H_z(\omega) = \left(1-\frac{j\omega}{\sigma_z+j\omega_z}\right)\left(1-\frac{j\omega}{\sigma_z-j\omega_z}\right). \tag{8}$$

In (7), ω_n is the frequency of formant n in rad/s, i.e., $\omega_n = 2\pi f_n$, and σ_n is half the bandwidth of formant n in rad/s, which can be approximated as a function of ω_n by a polynomial regression model [13]. As an example, a vocal spectrum envelope is plotted in Figure 1, which was extracted by the following procedure from a recording of Dietrich Fischer-Dieskau's performance.

In a short-time spectrum (computed by the constant-Q transform [14]) of accompanied singing, amplitudes at the partial frequencies of a (manually identified) vocal pitch constitute a noisy observation for estimating the underlying vocal spectrum envelope. As a consequence, the vocal

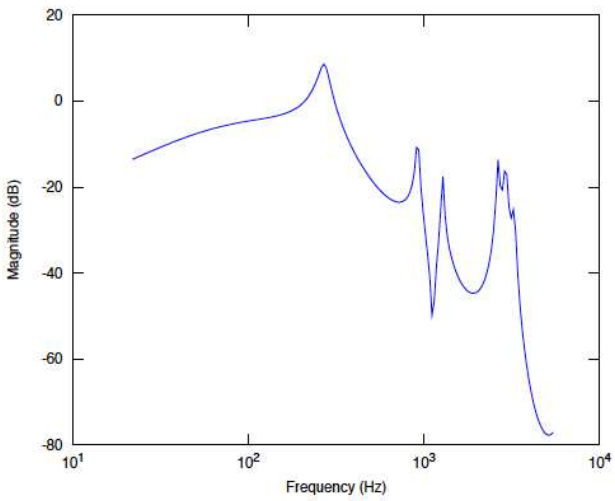

Figure 1. A vocal spectrum envelope with formant frequencies (in hertz) of $f_1 = 270$, $f_2 = 1274$, $f_3 = 2630$, $f_4 = 2920$, $f_5 = 3270$, $f_p = 920$, and $f_z = 1120$.

spectrum envelope can be estimated by fitting its spectral samples to the observed amplitudes:

$$\hat{\mathbf{v}} = \arg\min_{\mathbf{v}\in\mathcal{V}} \sum_{l=1}^{40} \left(a_l^q - a - A(lf_0^h)\right)^2, \tag{9}$$

where a is an amplitude (in dB) variable that modifies the overall magnitude of the spectrum envelope,

$$\mathbf{v} = (a, f_1, f_2, f_3, f_4, f_5, f_p, f_z)^T, \tag{10}$$

$\mathcal{V}$ describes constraints imposed on the formant frequencies:

$$\mathcal{V} = \left\{ \mathbf{v}\in R^8 \;\middle|\; \begin{array}{l} 250 \le f_1 \le 1000 \\ 600 \le f_2 \le 3000 \\ 1700 \le f_3 \le 4100 \\ 2500 \le f_4 \le 4500 \\ 3000 \le f_5 \le 5500 \\ 200 \le f_p \le 4000 \\ 200 \le f_z \le 4000 \\ f_1 \le f_2 \le f_3 \le f_4 \le f_5 \\ f_p \le f_z \end{array} \right\}, \tag{11}$$

a_l^q denotes the amplitude (in dB) observed at the lth partial, and f_0^h denotes the vocal pitch in hertz. The constrained optimization problem in (9) is solved by the multi-start coordinate-descent distance minimization procedure described in [15].

3.1.2 Synthesis From a Spectrum Envelope

Let $A^{(i)}(\cdot)$ denote the ith vocal spectrum envelope extracted from accompanied singing data ($i = 1, ..., N_e$). To synthesize the ith voice example for pitch k (i.e., $\mathbf{x}^{(i,k)}$), we compute its partial amplitudes according to the envelope $A^{(i)}(\cdot)$:

$$a_l^{(i)} = A^{(i)}(l\cdot 440\cdot 2^{(k-60)/24}), l = 1, ..., L, \quad L = \left\lfloor \frac{5000}{440\cdot 2^{(k-60)/24}} \right\rfloor, \tag{12}$$

where $a_l^{(i)}$ denotes the amplitude (in dB) of the lth partial. Then, the voice example can be synthesized as

$$x_t^{(i,k)} = \sum_{l=1}^{L} 10^{\frac{a_l^{(i)}}{20}} \cos\left(2\pi l \cdot 440 \cdot 2^{\frac{k-60}{24}} \cdot \frac{t}{11025}\right), \quad t = 1, ..., N. \tag{13}$$

3.2 Likelihood of a Voice Example

To evaluate the likelihood of the voice example $\mathbf{x}^{(i,k)}$, we take advantage of the fact that the accompanied singing signal is the sum of the singing voice signal and the accompaniment signal:

$$\mathbf{z} = \mathbf{x} + \mathbf{y}, \tag{14}$$

where $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are random N-vectors representing the singing voice, the accompaniment, and the accompanied singing, respectively. By taking (14) as a transformation of $\mathbf{y}$ into $\mathbf{z}$, the likelihood can be evaluated as

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}^{(i,k)}) = p_{\mathbf{y}|\mathbf{x}}(\mathbf{z} - \mathbf{x}^{(i,k)}|\mathbf{x}^{(i,k)}). \tag{15}$$

Approximate independence can be assumed between the vectors $\mathbf{x}$ and $\mathbf{y}$, in that they represent separate sound sources with independent phases; hence, we have

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}^{(i,k)}) \approx p_{\mathbf{y}}(\mathbf{z} - \mathbf{x}^{(i,k)}). \tag{16}$$

The dependence and trend among the time samples in $\mathbf{y}$ represent the specific timbre or polyphony of the accompaniment, of which, however, we do not have any knowledge at the time of melody extraction. In consequence, approximate i.i.d. is assumed among the time samples:

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}^{(i,k)}) \approx \prod_{t=1}^{N} p_y(z_t - x_t^{(i,k)}). \tag{17}$$

To determine the probability distribution of each time sample in $\mathbf{y}$, we collected 256,000 time sample values by randomly sampling the accompaniment data in the MIR-1K dataset [16]. The histogram of these values, as plotted in Figure 2, suggests that the probability distribution can be approximated by a zero-mean Gaussian distribution. The Q-Q plot of these values against the standard normal distribution, as shown in Figure 3, confirms the approximation by presenting a curve that resembles a straight line. With this approximation, we have

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}^{(i,k)}) \propto \exp\left\{-\frac{||\mathbf{z} - \mathbf{x}^{(i,k)}||^2}{2\sigma_y^2}\right\}, \tag{18}$$

where σ_y denotes the standard deviation of the Gaussian distribution.

The voice examples $\{\mathbf{x}^{(i,k)}\}_{i=1}^{N_e}$ are initially intended for simulating the random experiment underlying the probability distribution described by the density $p_{\mathbf{x}|w}(\cdot|k)$; even so, we cannot afford to synthesize a huge number of voice examples that collectively represent the diversity in such trivial signal specifications as various loudness levels and

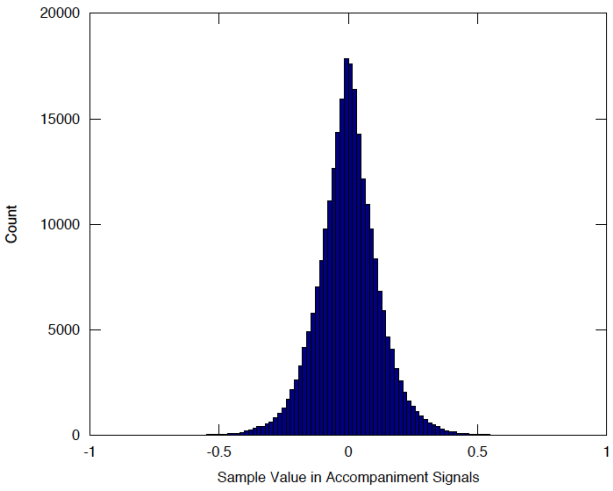

Figure 2. Histogram of sample values in accompaniment signals.

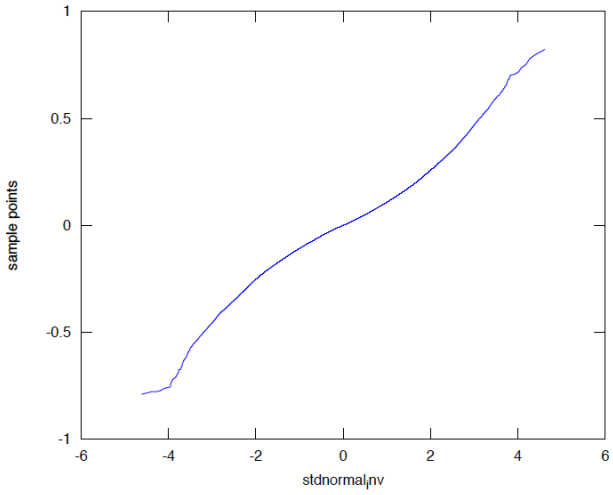

Figure 3. Q-Q plot of sample values in accompaniment signals against the standard normal distribution.

various sinusoidal phase angles. Therefore, as described in Section 3.1, the N_e examples serve only to represent the timbral variety in singing voice; meanwhile, each voice example needs to be matched against the accompanied singing in a phase- and loudness-insensitive fashion.

To achieve the phase-insensitivity, we substitute a scaled frequency-domain total power for the N-sample signal energy in (18):

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x}^{(i,k)}) \propto \exp\{-c \sum_{f=1}^{192} |A_f^z e^{j\phi_f^z} - A_f^{(i,k)} e^{j\phi_f^{(i,k)}}|^2\}, \tag{19}$$

where c is a manually specified scaling constant ($c = 2 \cdot 10^5$), f is a frequency index to a constant-Q spectrum [14] with 192 quarter-tone-spaced bins, A_f^z and ϕ_f^z denote the constant-Q magnitude and phase spectra of the accompanied singing signal, and $A_f^{(i,k)}$ and $\phi_f^{(i,k)}$ denote those of the voice example. Now, for the phase-insensitivity, we

relax the phase of the voice example and maximize the likelihood with respect to the relaxed phase, thereby creating a modified voice example $\bar{\mathbf{x}}^{(i,k)}$ with phase spectrum $\{\phi_f^z\}_{f=1}^{192}$:

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\bar{\mathbf{x}}^{(i,k)}) \propto \exp\{-c\sum_{f=1}^{192}(A_f^z - A_f^{(i,k)})^2\}, \quad (20)$$

$$CQT\{\bar{\mathbf{x}}^{(i,k)}\} = \{A_f^{(i,k)}e^{j\phi_f^z}\}_{f=1}^{192}, \quad (21)$$

where $CQT\{\cdot\}$ denotes the constant-Q transform.

Next, to achieve the insensitivity to loudness, we relax the loudness of the voice example and maximize the likelihood with respect to the relaxed loudness, thereby creating an amplified or attenuated voice example $\tilde{\mathbf{x}}^{(i,k)}$:

$$\begin{aligned} & p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\tilde{\mathbf{x}}^{(i,k)}) \\ \propto \quad & \exp\left\{-c\left[\sum_{f=1}^{192}(A_f^z)^2 - \frac{(\sum_{f=1}^{192}A_f^zA_f^{(i,k)})^2}{\sum_{f=1}^{192}(A_f^{(i,k)})^2}\right]\right\}, \end{aligned} \quad (22)$$

$$\tilde{\mathbf{x}}^{(i,k)} = \frac{\sum_{f=1}^{192}A_f^zA_f^{(i,k)}}{\sum_{f=1}^{192}(A_f^{(i,k)})^2}\bar{\mathbf{x}}^{(i,k)}, \quad (23)$$

which orthogonally projects the accompanied singing onto the subspace of amplified or attenuated versions of the voice example.

In the end, to evaluate the likelihood of pitch w, we substitute the modified voice examples $\{\tilde{\mathbf{x}}^{(i,w)}\}_{i=1}^{N_e}$ for the voice examples $\{\mathbf{x}^{(i,w)}\}_{i=1}^{N_e}$ in (3):

$$p_{\mathbf{z}|w}(\mathbf{z}|w) \approx \frac{1}{N_e}\sum_{i=1}^{N_e}p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\tilde{\mathbf{x}}^{(i,w)}). \quad (24)$$

4. PRIOR MODEL

We use a Markov chain $\{w_m\}_{m=1}^{M}$ to model the vocal pitch sequence, which consists of 100 pitch values per second:

$$\begin{aligned} & P(w_1, ..., w_M) \\ = \quad & P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\cdots \\ & P(w_m|w_1, ..., w_{m-1})\cdots P(w_M|w_1, ..., w_{M-1}) \\ = \quad & P(w_1)\prod_{m=2}^{M}P(w_m|w_{m-1}), \end{aligned} \quad (25)$$

where random variable $w_m \in \{1, ..., 88\}$ represents the mth element in the vocal pitch sequence. The second equality in (25) results from the Markovianity that given the previoius pitch w_{m-1}, the current pitch w_m is independent of all the earlier pitches $w_{m-2}, w_{m-3}, ..., w_1$. The initial state distribution is assumed to be uniform over all possible pitch values:

$$P(w_1 = k) = \frac{1}{88}, \forall k \in \{1, ..., 88\}. \quad (26)$$

The state transition probability distribution is also assumed to be uniform, but only over pitch values within 2 quarter tones of the previous pitch:

$$\begin{aligned} & P(w_m = k_2|w_{m-1} = k_1) \\ = \quad & \begin{cases} \frac{1}{3} & \text{if} \quad k_1 \in \{1, 88\}, |k_1 - k_2| \le 2; \\ \frac{1}{4} & \text{if} \quad k_1 \in \{2, 87\}, |k_1 - k_2| \le 2; \\ \frac{1}{5} & \text{if} \quad 3 \le k_1 \le 86, |k_1 - k_2| \le 2; \\ 0 & \text{if} \quad |k_1 - k_2| > 2. \end{cases} \end{aligned} \quad (27)$$

In almost all cases, there are five pitch values around the previous pitch that are assigned a nonzero probability ($\frac{1}{5}$) for the current pitch. Other cases are associated with only 3 or 4 pitch values. For example, when the previous pitch is 88, the only possible values for the current pitch are 86, 87, and 88.

5. VOICING DETECTION

In addition to estimating the vocal pitch sequence from accompanied singing, vocal melody extraction finds particular time points at which no singing voice is actually sounding. Such time points may be found during vocal rests, at plosives, etc. For each of these time points, the pitch estimate should be overriden by a state indicating the absence of singing voice. In other words, we need a mechanism for detecting the singing voice for each time point.

To this end, we estimate the short-time spectra of the singing voice on the basis of the accompanied singing. Ideally, the estimation will give a zero spectrum for each time point that is not voiced. To estimate the spectrum at a particular time point, we use its minimum mean square error (MMSE) estimator:

$$\begin{aligned} & E[CQT\{\mathbf{x}\}|\mathbf{z}] \\ = \quad & \int CQT\{\mathbf{x}\}p_{\mathbf{x}|\mathbf{z}}(\mathbf{x}|\mathbf{z})\mathbf{dx} \\ = \quad & \int CQT\{\mathbf{x}\}\frac{p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x})}{p_{\mathbf{z}}(\mathbf{z})}p_{\mathbf{x}}(\mathbf{x})\mathbf{dx}. \end{aligned} \quad (28)$$

With the term $CQT\{\mathbf{x}\}\frac{p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x})}{p_{\mathbf{z}}(\mathbf{z})}$ taken as a function of $\mathbf{x}$, this integral can be thought of as the expectation of $CQT\{\mathbf{x}\}\frac{p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x})}{p_{\mathbf{z}}(\mathbf{z})}$ and approximated by the corresponding sample mean:

$$\begin{aligned} & E[CQT\{\mathbf{x}\}|\mathbf{z}] \\ \approx \quad & \frac{1}{88N_e}\sum_{k=1}^{88}\sum_{i=1}^{N_e}CQT\{\tilde{\mathbf{x}}^{(i,k)}\}\frac{p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\tilde{\mathbf{x}}^{(i,k)})}{p_{\mathbf{z}}(\mathbf{z})}. \end{aligned} \quad (29)$$

The density $p_{\mathbf{z}}(\mathbf{z})$ can again be approximated in this fashion:

$$\begin{aligned} & p_{\mathbf{z}}(\mathbf{z}) \\ = \quad & \int p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\mathbf{x})p_{\mathbf{x}}(\mathbf{x})\mathbf{dx} \\ \approx \quad & \frac{1}{88N_e}\sum_{k=1}^{88}\sum_{i=1}^{N_e}p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\tilde{\mathbf{x}}^{(i,k)}). \end{aligned} \quad (30)$$

Since all the modified voice examples share the same phase spectrum, the magnitude of the spectrum estimate is evaluated as

$$\begin{aligned} & |E[CQT\{\mathbf{x}\}|\mathbf{z}]|_f \\ \approx \quad & \frac{1}{88N_e}\sum_{k=1}^{88}\sum_{i=1}^{N_e}\tilde{A}_f^{(i,k)}\frac{p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}|\tilde{\mathbf{x}}^{(i,k)})}{p_{\mathbf{z}}(\mathbf{z})}, \\ & f = 1, ..., 192, \end{aligned} \quad (31)$$

where $\tilde{A}_f^{(i,k)}$ denotes the constant-Q magnitude spectrum of the modified voice example $\tilde{\mathbf{x}}^{(i,k)}$. Eventually, the loudness of the singing voice can be estimated by correcting the magnitude spectrum according to the trends in the 40-phon equal-loudness contour (ELC) [17], which quantifies the dependency of human loudness perception on frequency:

$$\Lambda(\mathbf{z}) = \sum_{f=1}^{192} (|E[CQT\{\mathbf{x}\}|\mathbf{z}]|_f \cdot 10^{(40-\kappa_f)/20})^2, \quad (32)$$

where κ_f denotes the 40-phon ELC, plotted in Figure 4. If, and only if, $\Lambda(\mathbf{z})$ exceeds the empirical threshold of $2 \cdot 10^{-5}$, the time point is deemed voiced.

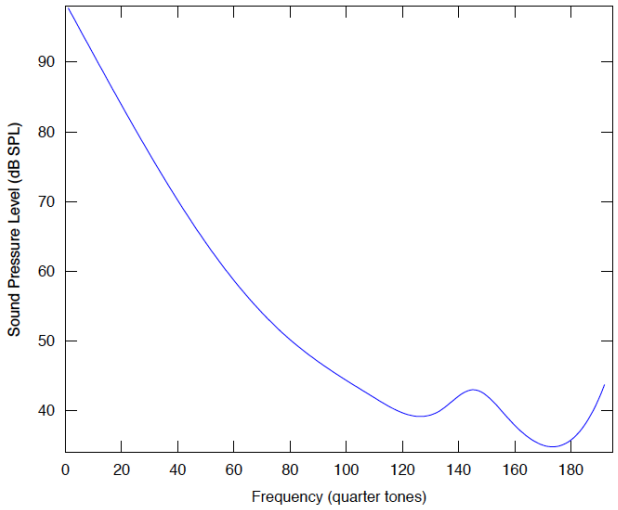

Figure 4. 40-phon equal-loudness contour.

6. EXPERIMENTS

In this section, to provide comparison of our method with some existing methods, we conduct vocal melody extraction experiments on a publicly available dataset. Since the synthesis of voice examples is based on a collection of accompanied singing data, we start by describing the collection.

6.1 Data Collection for Voice Example Synthesis

To synthesize voice examples, we extracted $N_e = 84$ vocal spectrum envelopes from 14 recordings of about 1 minute each. The 14 recordings represent 14 distinct types of singing voice, including 10 recordings of professional (accompanied) singing captured from YouTube, and 4 recordings of non-professional (unaccompanied) singing adapted from some clips in the MIR-1K dataset [16]. From each recording, 6 spectrum envelopes were extracted that represent 6 distinct types of voiced sound.

The 14 types of singing voice are tenor (José Carreras), soprano (Kiri Te Kanawa), baritone (Dietrich Fischer-Dieskau), mezzo-soprano (Cecilia Bartoli), pop high male voice (Terry Lin), pop high female voice (Stella Chang), pop low male voice (Shifeng Luo), pop low female voice (Inn-Jae Chen), pop nasal male voice (Wakin Chau), pop nasal female voice (Chiou-Feng Tsai), non-professional high male voice (Bobon), non-professional high female voice (Annar), non-professional low male voice (Davidson), and non-professional low female voice (Ani). The "nasal" singers are well-known in Taiwan for nasalizing their vowels significantly.

The 6 types of voiced sound are /i/, /ɛ/, /ɑ/, /ɔ/, /u/, and a miscellaneous type defined by /ə/, /z̩/, /ʐ̩/, /m̩/, /n̩/, or /ŋ̩/. Each sound in the miscellaneous type does not occur in all recordings: /ə/ is absent in all 4 Taiwanese-language recordings, perhaps because it seldom occurs in the northern speech of the Taiwanese language; the syllabic nuclei /z̩/ and /ʐ̩/ are specific to languages such as Mandarin Chinese; and the nasal hummings, due to their low loudness, are rarely used in operatic singing. To extract vocal spectrum envelopes, the first author subjectively selected 6 short-time spectra from each recording that exemplify the 6 sound types, respectively.

6.2 Dataset Description

The dataset adopted for performance evaluation is a subset of the one built for the Melody Extraction Contest in the ISMIR 2004 Audio Description Contest (ADC 2004). The whole ADC 2004 dataset consists of 20 audio recordings, each around 20 seconds in duration, among which eight recordings have instrumental melodies, and the other twelve have vocal melodies. Since this work considers vocal melodies only, experiments are carried out exclusively on 9 of the 12 vocal recordings, including two pop song excerpts, three song excerpts with synthesized vocal, and four opera excerpts. The other three vocal excerpts are not included here because one contains falsetto singing and the other two contain an ensemble of vocals. The dataset has been in use in several Music Information Retrieval Evaluation Exchange (MIREX) contests since 2006; therefore, it affords extensive comparison among methods.

Before melody extraction, each audio file in the dataset is resampled at 11,025 hertz and constant-Q transformed [14] ($Q = 34$) into a sequence of short-time spectra. Each resulting spectrum is a quarter-tone-spaced sampling of a continuous spectrum that is capable of resolving the interference between two half-tone-spaced sinusoids from 21.827 hertz all the way to 5,428.6 hertz.

6.3 Performance Measures

In the experiments documented here, the tested system gives vocal melodies in the format of a voicing/pitch value for each frame (at the rate of 100 frames per second). If a frame is estimated to be voiced, the output specifies the pitch estimate for the frame; otherwise, the output specifies that the frame is estimated to be not voiced.

MIREX adopts several measures for evaluating the performance of a melody extraction system [18]. In the first place, to determine how well the system performs voicing detection, we use the voicing detection rate, the voicing false alarm rate, and the discriminability. The voicing detection rate is computed as the fraction of frames that are both labeled and estimated to be voiced, among all the frames that are labeled voiced. The voicing false alarm rate

is computed as the fraction of frames that are estimated to be voiced but are actually not voiced, among all the frames that are not voiced according to the reference transcription. The discriminability combines the above two measures in such a way that it can be deemed independent of the value of any threshold involved in the decision of voicing detection:

$$d' = Q^{-1}(P_F) + Q^{-1}(1 - P_D), \quad (33)$$

where $Q^{-1}(\cdot)$ denotes the inverse of the Gaussian tail function, P_F denotes the false alarm rate, and P_D denotes the detection rate.

Second, to determine how well the system performs pitch estimation, we use the raw pitch accuracy and the raw chroma accuracy. The raw pitch accuracy is computed as the fraction of frames that are labeled voiced and have pitch estimated within one quarter tone of the true pitch, among all the frames that are labeled voiced. To focus on pitch class estimation while ignoring octave errors, we compute the raw chroma accuracy, which is computed in the same way as the raw pitch accuracy, except that the pitch is here measured in terms of chroma, or pitch class, a quantity derived from the pitch by wrapping the pitch into one octave.

Finally, the performance of voicing detection and pitch estimation can be measured jointly by the overall transcription accuracy, defined as the fraction of frames that receive correct voicing classification and, if voiced, a pitch estimate within one quarter tone of the true pitch, among all the frames.

Excerpt	Accuracy (%)			PD (%)	PF (%)	d'
	All	Voiced	Chroma			
pop3	61.089	59.606	67.303	88.931	33.488	1.6494
pop4	75.623	74.985	77.387	89.924	22.004	2.0493
daisy1	76.529	97.388	97.388	99.813	90.419	1.5942
daisy2	82.636	96.417	96.417	100	88.268	-
daisy4	99.122	99.323	99.323	100	100	-
opera_fem2	68.032	71.474	72.988	88.287	44.101	1.3379
opera_fem4	81.668	84.956	84.956	93.191	50.781	1.4706
opera_male3	62.300	58.708	63.483	76.629	8.636	2.0902
opera_male5	73.727	72.643	78.259	83.668	18.705	1.8697

Table 1. Experimental results. ("pop1" and "pop2," which contain an ensemble of vocals, are not included here. "daisy3" is excluded because it contains falsetto singing.)

6.4 Results

The results are listed in Table 1. The overall transcription accuracies listed in the column titled "All" range from 61% to 99%, with an average at 75.636%. The minimum is found at the excerpt "pop3." A significant error made in the analysis of this excerpt is depicted in Figure 5, which reveals that the system mistakenly selected the pitch (87 hertz) of a high-energy (instrumental) bass from 1.34 s to 1.77 s because of a tendency of the proposed signal model to assume a low-energy accompaniment. Still, the energy of frequency components away from the hypothesized partials is irrelevant to the timbre of the hypothesized voice. This suggests that further improvement to the accuracy may be made by leaving out the off-partial frequency components in the calculation of the voice likelihood. At the other end of the accuracies, we see that the maximum occurs at the excerpt "daisy4," which might have been particularly easy for our approach because its melodic source is a synthesized vocal. The raw pitch accuracies in the column titled "Voiced" are highly correlated with the overall transcription accuracies, which suggests that further improvement to this system should be made in pitch estimation, not in voicing detection. The column titled "Chroma" contains raw chroma accuracies similar to the raw pitch accuracies, which suggests that octave errors were successfully avoided by the system.

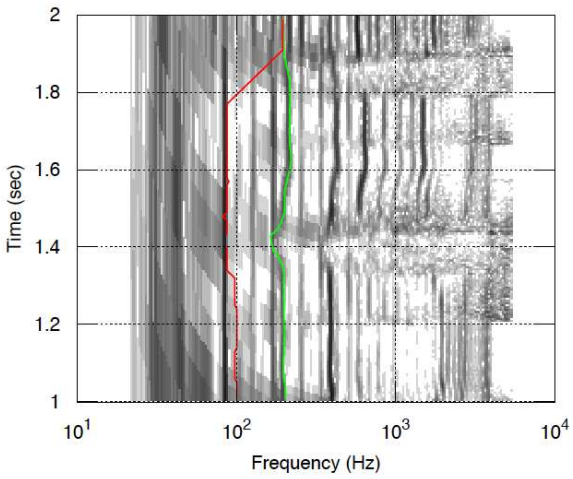

Figure 5. Spectrogram of a segment of the test excerpt "pop3," overlayed with the true melody in green and the melody estimate in red.

Shown in Table 2 is a comparison of the proposed method with the MIREX 2011 submissions in terms of the overall transcription accuracy (OTA). Notably, if the proposed method had entered the evaluation in 2011, it would have ranked 5th out of a total of 11 submissions. Moreover, the accuracy of the proposed system is within 10% of the highest accuracy in the 2011 evaluation. Compared with the method we proposed in [15], which corresponds to Method 6 in Table 2, our current method turns out to give a slighly lower accuracy. This confirms the feasibility of adopting this new approach as the foundation for our future work on vocal melody extraction.

Method	1	2	3	4	5	6	7	8	9	10	Proposed
OTA (%)	66	65	65	71	56	78	83	75	84	78	76

Table 2. Comparison with the MIREX 2011 Audio Melody Extraction results.

7. CONCLUSIONS

A novel approach to vocal melody extraction has been presented that integrates acoustic-phonetic knowledge and real-world data in estimating the vocal pitch sequence. The performance of the proposed method has been evaluated on a

publicly available dataset to be comparable to the state-of-the-art performance. In the future, we expect a minor modification to the proposed signal model that will further improve the performance in vocal pitch estimation.

Acknowledgments

This work was supported in part by the Taiwan e-Learning and Digital Archives Program (TELDAP) sponsored by the National Science Council of Taiwan under Grant: NSC 100-2631-H-001-013. Comments from the anonymous reviewers were valuable for the enhanced quality of this paper.

8. REFERENCES

[1] M. Goto and S. Hayamizu, "A real-time music scene description system: Detecting melody and bass lines in audio signals," in *IJCAI-CASA*, 1999.

[2] R. P. Paiva, T. Mendes, and A. Cardoso, "On the detection of melody notes in polyphonic audio," in *Proc. ISMIR*, 2005.

[3] S. Jo and C. D. Yoo, "Melody extraction from polyphonic audio based on particle filter," in *Proc. ISMIR*, 2010.

[4] K. Dressler, "An auditory streaming approach for melody extraction from polyphonic music," in *Proc. ISMIR*, 2011.

[5] M. Lagrange, L. Martins, J. Murdoch, and G. Tzanetakis, "Normalized cuts for predominant melodic source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 278–290, 2008.

[6] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. ICASSP*, 2008.

[7] D. P. W. Ellis and G. E. Poliner, "Classification-based melody transcription," *Mach. Learn.*, vol. 65, no. 2-3, pp. 439–456, 2006.

[8] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and Viterbi search," in *Proc. ICASSP*, 2006.

[9] G. Fant, *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*. The Hague: Mouton, 1970.

[10] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2145–2154, 2010.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, 1989.

[12] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 971–995, 1980.

[13] J. W. Hawks and J. D. Miller, "A formant bandwidth estimation procedure for vowel synthesis," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1343–1344, 1995.

[14] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Am.*, vol. 92, no. 5, pp. 2698–2701, 1992.

[15] Y.-R. Chien, H.-M. Wang, and S.-K. Jeng, "An acoustic-phonetic approach to vocal melody extraction," in *Proc. ISMIR*, 2011.

[16] 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, Lang. Process.*, 2009.

[17] ISO 226, "Acoustics—normal equal-loudness contours," 2003.

[18] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1247–1256, 2007.