國立臺灣大學電機資訊學院電信工程學研究所

博士論文

Graduate Institute of Communication Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Doctoral Dissertation

源自聲學語音學、用於有伴奏歌唱音訊分析之概似模型

Acoustic-Phonetic Likelihood Models for Analysis of Accompanied Singing Audio

簡御仁

Yu-Ren Chien

指導教授：鄭士康 博士、王新民 博士

Advisors: Shyh-Kang Jeng, Ph.D. and Hsin-Min Wang, Ph.D.

中華民國 105 年 1 月

January, 2016

Acknowledgments

I am indebted to my advisors, Dr. Hsin-Min Wang, Academia Sinica and Prof. Shyh-Kang Jeng, National Taiwan University, for letting me join their research groups as a Ph.D. student, where I have been able to develop and pursue my research interests. I have always been grateful to my advisors and all their group members, for listening to and commenting numerous presentations of mine in regular group meetings.

For my work on lyrics alignment, I would like to thank Dr. Annamaria Mesaros, Tampere University of Technology and Lien-Chiao Lin, National Taiwan University for sharing their ground-truth annotations for the data sets `poly_100` and `slam`. I would also like to thank Dr. Kuan-Yu Chen, Academia Sinica for his implementation of Chinese word segmentation.

Many designs in my melody extraction and lyrics alignment systems are inspired by the vocal techniques practiced in the Academia Sinica Chorus, which was founded in 2011 by Dr. Hei-Yuan Chiu and Ms. Shi-Qing Xu. I would like to thank them for my four years of musical life at Academia Sinica.

My doctoral study includes a seven-month research visit at IRCAM, France, during which I worked on an approach to source-filter analysis of human voice based on phase spectrum processing. Here I would like to thank Dr. Axel Röbel for hosting the visit, in that some programming skills helpful for experimentation, demonstration, and lyrics processing in this thesis work were learned at IRCAM.

摘要

本論文所探討的主題是有伴奏歌唱錄音的旋律分析以及歌詞分析。爲了有效進行此分析，本論文提出一種獨特的概似模型作爲方法的核心，它巧妙地結合了聲學語音學的知識以及實際蒐集而得的資料。此模型的基本要素是一套音色吻合度以及發聲狀態吻合度的量化評估方式，可爲任一候選基本頻率（基頻）或者候選母音／發聲狀態進行評分。音色吻合度意指某個基頻值的諧波振幅序列所呈現之音色與參考音色之間的相似程度，而參考音色的定義則來自一小組歌聲音色範例。爲特定基頻估算音色吻合度時，需要對所有音色範例進行基頻的修改，本論文提出的修改方式利用聲學語音學的模型，將修改前的聲帶波形以及共振峰頻率予以保留。此一概似模型在發聲狀態的部份，對弦波進行偵測、追蹤以及刪減的處理，以便在估計歌聲音量的同時，將伴奏的干擾減至最低。最後基頻或音節的估計值，是由概似模型與事前的順序模型共同決定。在使用多個資料集進行系統測試之前，此方法所涉及的所有數值參數均已完成最佳化，且使用的是數個不與測試資料有任何重複的發展資料集。對照實驗顯示，音色吻合度的使用與否，會在整體旋律正確率上面造成 13% 的差距，同時也會在平均標準化歌詞對齊誤差上面，造成 7% 的差距。

關鍵詞：旋律抽取、歌詞對齊、歌聲、聲學語音學、基頻修改、聲帶波形、共振峰頻率。

Abstract

This dissertation addresses melodic and lyrics analysis of accompanied singing recordings. Central to my approach are likelihood models that integrate acoustic-phonetic knowledge and real-world data. These models are based on a timbral fitness score and a voicing fitness score evaluated for each fundamental frequency (F0) or vowel/voicing candidate. Timbral fitness is measured for the partial amplitudes of an F0 value, with respect to a small set of vocal timbre examples. This F0-specific measurement of timbral fitness depends on an acoustic-phonetic F0 modification of each timbre example, which preserves glottal pulse shape and formant frequencies. In the voicing part of the likelihood models, sinusoids are detected, tracked, and pruned to give loudness values that minimize interference from the accompaniment. A final F0 or syllable estimate is determined by a prior sequential model in addition to the likelihood model. The numerical parameters involved in my approach were optimized on several development sets from different sources before the system was evaluated on multiple test sets separate from these development sets. Controlled experiments show that use of the timbral fitness score accounts for a 13% difference in overall melodic accuracy, and a 7% difference in average normalized lyrics alignment error.

Keywords: melody extraction, lyrics alignment, singing voice, acoustic phonetics, F0 modification, glottal pulse shape, formant frequency.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

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 content. This has motivated researchers to work on systems that index music databases by extracting such essential musical features as melodies and chords from audio recordings. Arguably, processing such as this should mimic human music listening and could thus enable machines to make personalized music purchase decisions on behalf of humans.

Songs typically come with words. The lyrics of a song determine how the song is performed in terms of phonetic articulation, and shape the timbral variations perceived by those who listen to the performance. The rhythm in which words and syllables in the lyrics are sung is highly variable, both within a song and between different songs—Some syllables are short because they are assigned to a short musical note; others are long, associated with a long note or multiple notes; and the tempo adds to the uncertainty in timing. This variability in rhythm makes it appealing to display lyric syllables synchronously in karaoke applications, or synchronized lyric lines or words as a visual augmentation to song playback. Furthermore, rhythm of lyrics could be used as a feature in music information retrieval in place of the ordinary rhythm of musical notes. Nevertheless, rhythmic information such as this is typically lacking in commercial distributions of lyrics.

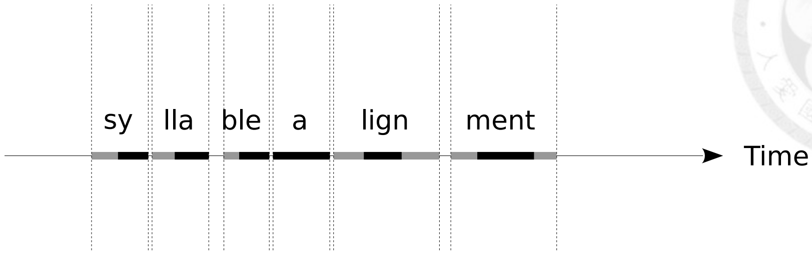


Figure 1.1: Intervals plotted on the time axis for the syllables in the phrase "syllable alignment." Vertical dotted lines mark the boundaries of these syllables. Each syllable is composed of a nucleus in black and possibly a consonant in grey preceding or following the nucleus.

1.2 Objective

In this dissertation, we focus on two specific tasks in the analysis of accompanied singing audio. One task is the extraction of *vocal melodies* from polyphonic audio signals. A melody is defined as a temporal process of variations in fundamental frequency (F0) that realizes motions from one musical pitch to another; as one might expect, melodies represent one of the most significant features that can be identified by listeners from musical pieces. Vocal melodies are of particular interest owing to the importance of vocal music in various musical cultures. By limiting the considered musical form to the common one of a solo singing voice accompanied by musical instruments, I propose a method for finding the F0 of the singing voice as a function of time while taking advantage of the timbral distinction between vocals and instruments.

The other task is automatic extraction of *lyrics rhythm* from accompanied singing audio, i.e., alignment of lyrics text with the audio. The aligned textual units in the lyrics can be syllables, words, phrases, or lines. The desired results of alignment consist of onset and offset time positions of each aligned unit. This is illustrated in Fig. 1.1 for alignment of lyric syllables.

1.3 Previous Approaches

1.3.1 Knowledge-Based Approaches

Results in psychoacoustics and musical acoustics inspired some researchers to propose methods that to some extent either imitate human auditory processing for melodies, or take advantage of acoustic characteristics of musical instruments and singing voice. Goto [10] measured the dominance of each F0 candidate by its harmonic strength in mid- and high-frequency regions and enforced temporal continuity in the F0 variations; technically, the harmonic strength derives from fitting a mixture of harmonic-structure tone models to the short-time audio spectrum, and the continuous F0 contours are tracked by multiple agents. Sutton [29] exploited F0 instability and high-frequency dominance to identify vocal F0s. Durrieu *et al.* [5] let the leading vocal part be represented by a source/filter model. Dressler [4] devised an approach where the predominant melody is tracked by an auditory streaming model that favors unstable, high-magnitude F0 contours. Hsu *et al.* [13] extracted the relative extents of vibrato and tremolo from each partial as features for classifying vocal and instrumental partials, and implemented F0 continuity by determining a sequence of tight ranges for the vocal F0. Tachibana *et al.* [30] used instability in F0 and intensity as well as shortness in duration to enhance melodic components. Salamon and Gómez [27] identified melodic F0s by vibrato or high magnitude. Joo *et al.* [17] considered vibrato extent in constructing melody lines.

Some approaches in the lyrics alignment literature make use of musical or linguistic knowledge. Iskandar *et al.* [14] set an alignment time unit to a minimum note duration determined from the tempo of a song. Wong *et al.* [31] aligned tonal contours and non-uniform rhythms in the lyrics with melodies and onsets in the audio by dynamic time warping. Kan *et al.* [18] performed beat tracking and set their alignment time unit to one beat. To locate chorus sections, they detected repeated sections with chroma features. From a structural audio segmentation of a popular song, Lee and Cremer [20] identified the chorus and verse sections by measuring an audio similarity between instances of the same section type, and aligned them with manually labeled lyric sections by dynamic programming. Fujihara *et al.* [9] attempted to enhance the alignment of unvoiced consonants in the lyrics with the audio by detecting fricative sounds in their specific frequency bands. Mauch *et al.* [22] addressed a closely related task, where timings in the audio signal are estimated for paired chords and lyric words from a song book.

1.3.2 Data-Driven Approaches

Relatively few approaches in the melody extraction literature take advantage of data in acoustic modeling. Ellis and Poliner [6] took a purely data-driven approach, letting the difference between melodic and non-melodic F0s be learned by support vector machine from a labeled set of training data. Hsu *et al.* [13] collected data for vocal and instrumental partials and trained 2 Gaussian mixture models for the 2-way classification. Joo *et al.* [17] modeled the timbre of melodic source by a set of partial amplitude examples, which is derived from *k*-means clustering of partial amplitude data points extracted from monophonic audio data.

The apparent diversity in vocal timbre among singers makes it a natural choice to take advantage of human voice data in singing voice modeling. Iskandar *et al.* [14] used speech data and pop song data to build their phoneme likelihood model, and estimated the probability distribution of syllable durations from pop song data to constrain a Viterbi alignment of lyrics. In the lyrics alignment system of Wong *et al.* [31], singing voice detection was performed by a multi-layer perceptron trained with vocal and non-vocal onset data. To train hidden Markov models for vocal segment classification in their lyrics alignment system, Kan *et al.* [18] used vocal and non-vocal segment data. They also used lyric-line duration data to estimate the prior-mean durations of phonemes in singing, which were then used to calculate a duration estimate for each input lyric line. Mesaros and Virtanen [24] built phoneme likelihood models with speech and unaccompanied singing data. In the lyrics alignment system of Fujihara *et al.* [9], predominant melodic source data was extracted from popular songs to train Gaussian mixture models for vocal activity detection. They also used speech, singing, and separated singing data to build their phoneme likelihood models. In the aforementioned approaches where both speech and singing data is used for phoneme likelihood modeling, speech data is used in training Gaussian mixture models before the models are adapted to a small amount of singing data.

1.3.3 Source Separation Techniques

Some existing approaches to melody extraction make use of techniques from single-channel source separation to isolate the desired sound source from the analyzed audio. Durrieu *et al.* [5] adapted non-negative matrix factorization to the decomposition of audio spectra into a predominant pitched source and an accompaniment. In melodic component enhancement, Tachibana *et al.* [30] processed instability and shortness by applying a harmonic/percussive separation algorithm to the audio in 2 passes with different window lengths. The first pass of this enhancement technique was adopted as a preprocessing step in [13].

To isolate the singing voice from the analyzed audio, use of techniques from audio source separation is also made by some approaches to lyrics alignment. Wong *et al.* [31] used the central panning of singing voice in a stereo recording to enhance the vocal signal before alignment. Mesaros and Virtanen [24] aligned lyrics with separated vocal signals, where the separation was based on a reconstruction of accompaniment. They performed the reconstruction by applying non-negative matrix factorization to vocal-free time-frequency regions determined from an estimated vocal melody. The voice signal with which Fujihara *et al.* [9] aligned lyrics was resynthesized from partial frequencies and amplitudes extracted from accompanied singing according to an estimated predominant melody.

1.3.4 Comparison

For both of the tasks addressed in this dissertation, there has been no research in the literature that explores the use of physical models from acoustic phonetics. Physical models of voice production are closely related to singing: A formant filter model of the vocal tract can represent various vowels sung by a singer [7], and glottal pulse shape models are relevant both to falsetto singing [28] and to personal voice quality. Although the source/filter model is an essential element in most acoustic-phonetic models, the filter model used in [5] is not constrained around a formant structure or specific vowel types, and their source model represents a single fixed voice quality.

Dependence of an approach to melody extraction on acoustic phonetics would most likely preclude its potential for instrumental melody extraction; still, I show in this dissertation how acoustic-phonetic models can provide a mechanism for timbre-preserving F0 modification of singing voice, thereby alleviating the problem of sparse pitch coverage commonly encountered in data-driven approaches to singing voice modeling. In my implementation, acoustic-phonetic F0 modification expands a small 84-sample set of vocal timbre examples by a factor of 88. In contrast, the existing data-driven approaches [6, 13, 17] apply a supervised or unsupervised learning procedure to a large audio data set composed of vocal and instrumental sound sources.

In each of the above-mentioned approaches that fit Gaussian-mixture phoneme models to human voice data, one attempts to eliminate the effect that an unknown vocal fundamental frequency (F0) has on timbre modeling, by extracting Mel-frequency cepstral coefficients (MFCCs) as a representation of the vocal spectrum envelope. The actual quality of this representation depends on the specific distribution of vocal F0 in the analyzed human voice. Since singing presents a much larger vocal pitch range than does speech, it could be difficult for the success of MFCCs in automatic speech recognition to generalize to lyrics alignment. In contrast, the wide range of vocal F0 in singing can be handled by a basic F0 parameter in a physical model of voice production. In this dissertation, I use a physical model to simulate a vowel timbre at any vocal F0 estimated from the analyzed audio, thereby circumventing pitch-blind representation of spectrum envelope.

1.4 Contribution

In the present contribution, a vocal F0 is identified at each analysis time position by its *timbral fitness*, which is a partial-amplitude similarity between the F0 and a set of vocal timbre examples. This similarity is based on a Euclidean distance between two partial-amplitude vectors representing the observed timbre and a vocal timbre example. The vocal timbre examples compactly represent vocal timbres of different genders, genres, voice types, and vowel types, but do not sample the vocal pitch range whatsoever. Since partial amplitudes are specific to an F0, each timbre example needs F0-modifying to the F0 candidate before being compared to the observation. We perform the F0 modification by estimating a set of acoustic-phonetic parameters (glottal breathiness, formant frequencies, and distortion) from the example and subsequently resynthesizing the example from the parameters and a new F0 value set to the candidate. Indeed, this similarity calculation procedure effectively checks the observed partial amplitudes against the acoustic-phonetic models used in the F0 modification. This similarity model is complemented by loudness evaluation for F0 candidates in the complete F0 likelihood model. The evaluation of loudness is based on detection, tracking, and pruning of sinusoids.

For lyrics alignment, a vocal component in the polyphonic audio is isolated according to a vocal F0 sequence estimated from the audio. At each analysis time position, a lyric vowel can be identified for the vocal component by a timbral fitness of the component with respect to the vowel, which is a partial-amplitude similarity between the vocal component and a set of timbre examples for the vowel. This similarity is based on a Euclidean distance between two partial-amplitude vectors representing the observed timbre and a *vowel* timbre example. For each lyric vowel, we have a set of timbre examples that compactly represent timbres of different genders, genres, and voice types, again without sampling the vocal pitch range whatsoever. As with vocal melody extraction, each vowel timbre example is F0-modified to the F0 estimate before being compared to the observation. The complete vowel likelihood model includes a loudness-based component for voicing detection apart from the similarity model.

In my melody extraction experiments, I verified the effectiveness of this timbral fitness measure by fitting values of all numerical parameters to 3 labeled development sets, and evaluating the melody extraction performance on 10 other test sets. Similarly for lyrics alignment, I used 2 development sets and 2 test sets.

1.5 Structure of the Document

The rest of this dissertation is organized as follows. An F0 modification procedure that underlies the proposed likelihood models is presented in Chapter 2. Algorithms (including the likelihood models) and experiments are detailed in Chapter 3 for vocal melody extraction and in Chapter 4 for lyrics alignment. Chapter 5 concludes the dissertation.

Chapter 2

Acoustic-Phonetic F0 Modification of Vocal Sinusoids

In this chapter, we consider a short-time sinusoidal representation of voiced vocal timbre that consists of a set of sinusoids, one for each vocal partial. The objective of processing considered here is to change the F0 of this representation while preserving the underlying glottal pulse shape and vocal tract transfer function. To this end, we estimate the pulse shape, formant frequencies, and distortion from the representation, and subsequently synthesize a modified sinusoidal representation according to these estimates and a target F0 value, as shown in Fig. 2.1. When applied to vocal timbre examples for melody extraction or lyrics alignment, this procedure is performed offline, with a fixed set of vocal timbre examples modified to a fixed set of F0 values that accommodates all possible results of vocal melody extraction. Before being used repeatedly in evaluation of timbral fitness in my experiments, the results of offline F0 modification were human-verified in terms of fidelity in vocal timbre.

2.1 Model of Human Voice Production

Consider the production of a voiced sound, i.e., a vowel, a nasal, an approximant, etc. In a quasi-periodic signal representation of the sound, it is approximated by a periodic signal within each short window, with fundamental frequency f_0 (hertz) and partial amplitudes $\{a_l\}_l$ (dB). Its production involves a glottal airflow filtered by the vocal tract and radiated from the lips, which can be modeled as follows [7]:

$$a_l = 20\log_{10}\left|au_l K(lf_0)\prod_{n=1}^{5} H_n(lf_0)\right| + D(lf_0), \tag{2.1}$$

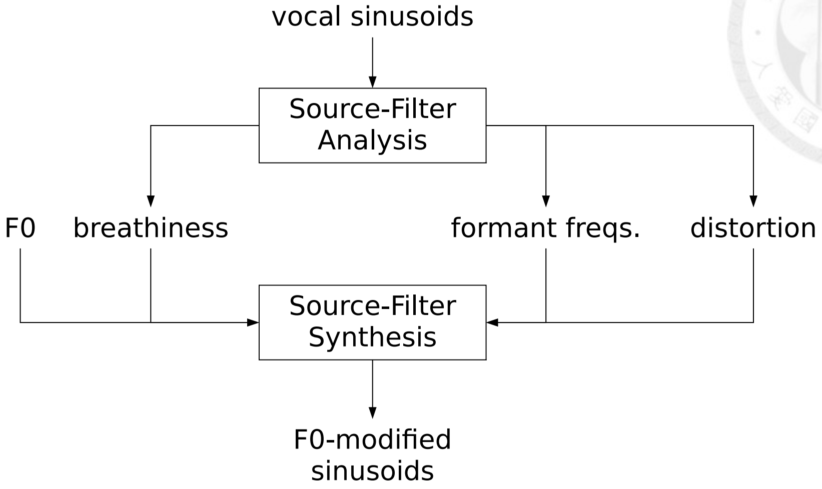


Figure 2.1: Block diagram for the F0 modification procedure.

$$l = 1, ..., \lfloor 5000/f_0 \rfloor, \tag{2.2}$$

where u_l denotes the Fourier coefficient of a derivative glottal airflow model $g(\cdot)$ for the lth partial, $a \geq 0$ scales the glottal pulse amplitude, and $K(\cdot)$, $H_n(\cdot)$, and $D(\cdot)$ are defined in Section 2.1.2. A block diagram for this model is shown in Fig. 2.2.

2.1.1 Glottal Excitation

Derivative of the glottal airflow signal, which represents the radiated glottal airflow, can be approximated by the transformed Liljencrants-Fant model [8, 21]: For $0 \leq t < T_e$,

$$g(t; T_0, E_e, R_d) = E_0 e^{\alpha t} \sin(\omega_g t); \tag{2.3}$$

for $T_e \leq t \leq T_0$,

$$g(t; T_0, E_e, R_d) = -\frac{E_e}{\varepsilon T_a}[e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_0-T_e)}], \tag{2.4}$$

where t denotes the time in seconds, T_0 denotes the fundamental period, E_e denotes the closure excitation magnitude, R_d denotes the pulse shape parameter, and values of the dependent variables E_0, α, ω_g, T_e, ε, and T_a can be fully determined by the value of (T_0, E_e, R_d). Since T_0 and E_e have no influence on the pulse shape, we define a space of glottal pulse shapes by $\mathbb{S} = \{R_d : 0.3 \leq R_d \leq 2.7\}$. As shown in Fig. 2.3, a greater value of R_d corresponds to a slower exponential decay in (2.4) [21], and thus to a shorter closed phase ($\{t : g(t) \approx 0\}$) of the glottal pulse and a breathier voice quality. Female voice and

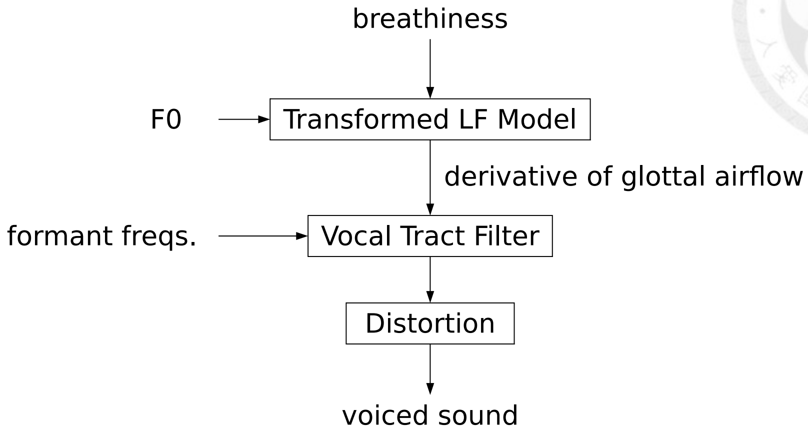


Figure 2.2: Model of voice production.

falsetto voice tend to be breathier or more akin to a single-sinusoid pure tone than do male voice and modal voice [28].

To show the invariance of this pulse shape model to F0, consider the following transformation of time variable and parameters:

$$\tilde{t} = \frac{t}{T_0}, \tag{2.5}$$

$$\tilde{T}_e = \frac{T_e}{T_0}, \tag{2.6}$$

$$\tilde{\alpha} = T_0\alpha, \tag{2.7}$$

$$\tilde{\omega}_g = T_0\omega_g, \tag{2.8}$$

$$\tilde{\varepsilon} = T_0\varepsilon, \tag{2.9}$$

$$\tilde{T}_a = \frac{T_a}{T_0}. \tag{2.10}$$

This gives

$$\tilde{g}(\tilde{t}) = E_0 e^{\tilde{\alpha}\tilde{t}} \sin(\tilde{\omega}_g \tilde{t}), \tag{2.11}$$

for $0 \leq \tilde{t} < \tilde{T}_e$, and

$$\tilde{g}(\tilde{t}) = -\frac{E_e}{\tilde{\varepsilon}\tilde{T}_a}[e^{-\tilde{\varepsilon}(\tilde{t}-\tilde{T}_e)} - e^{-\tilde{\varepsilon}(1-\tilde{T}_e)}], \tag{2.12}$$

for $\tilde{T}_e \leq \tilde{t} \leq 1$, which describe a pulse shape invariant to T_0.

Three regression formulas are essential to derivation of the dependent variables E_0, $\tilde{\alpha}$, $\tilde{\omega}_g$, $\tilde{T}_e$, $\tilde{\varepsilon}$, and $\tilde{T}_a$ as functions of (E_e, R_d):

$$R_d = (1/0.11)[0.5 + 1.2 \cdot (\tilde{\omega}_g \tilde{T}_e/\pi - 1)][\pi(\tilde{\omega}_g \tilde{T}_e/\pi - 1)/2\tilde{\omega}_g + \tilde{T}_a], \tag{2.13}$$

$$\tilde{T}_a = (-1 + 4.8R_d)/100, \tag{2.14}$$

$$\tilde{\omega}_g \tilde{T}_e/\pi - 1 = (22.4 + 11.8R_d)/100. \tag{2.15}$$

Formulas (2.14) and (2.15) can be substituted into (2.13) to give an expression for $\tilde{\omega}_g$, which in turn gives an expression for $\tilde{T}_e$ through (2.15). From $\tilde{g}(\tilde{T}_e) = -E_e$, we have

$$\frac{1 - e^{-\tilde{\varepsilon}(1-\tilde{T}_e)}}{\tilde{\varepsilon}\tilde{T}_a} = 1, \tag{2.16}$$

or equivalently, $\phi(\tilde{\varepsilon}) = 0$, where

$$\phi(\tilde{\varepsilon}) = \tilde{\varepsilon}\tilde{T}_a - 1 + e^{-\tilde{\varepsilon}(1-\tilde{T}_e)}. \tag{2.17}$$

Let $\tilde{\varepsilon}_+ = 2/\tilde{T}_a$ and $\tilde{\varepsilon}_- = \arg\min_{\tilde{\varepsilon}} \phi(\tilde{\varepsilon})$. Since $\phi(\tilde{\varepsilon}_+) > 0$ and $\phi(\tilde{\varepsilon}_-) < 0$, the zero of $\phi(\cdot)$ can be found by the bisection method. Continuity of the derivative glottal airflow at $\tilde{t} = \tilde{T}_e$ is ensured by

$$E_0 = \frac{-E_e}{e^{\tilde{\alpha}\tilde{T}_e}\sin(\tilde{\omega}_g \tilde{T}_e)}. \tag{2.18}$$

Finally, periodicity of the glottal airflow can be enforced by requiring the definite integral of derivative airflow to vanish across a fundamental period:

$$\int_0^1 \tilde{g}(\tilde{t})d\tilde{t} = 0. \tag{2.19}$$

This leads to an equation in the unknown $\tilde{\alpha}$:

$$\tilde{\varepsilon}\tilde{T}_a[\tilde{\omega}_g \cot(\tilde{\omega}_g \tilde{T}_e) - \tilde{\omega}_g e^{-\tilde{\alpha}\tilde{T}_e} \csc(\tilde{\omega}_g \tilde{T}_e) - \tilde{\alpha}] = (\tilde{\alpha}^2 + \tilde{\omega}_g^2)[\tilde{\varepsilon}^{-1} + (\tilde{T}_e - \tilde{\varepsilon}^{-1} - 1)(1 - \tilde{\varepsilon}\tilde{T}_a)], \tag{2.20}$$

which can be solved with a numerical equation solver.

2.1.2 Vocal Tract Filter

The vocal tract filter is made up of an infinite number of formants. The collective frequency response of formants beyond order five can be approximated by the following frequency response function [7]:

$$20\log_{10} K(f^h) = 0.43\left(\frac{f^h}{500}\right)^2 + 7.1 \cdot 10^{-4}\left(\frac{f^h}{500}\right)^4, \tag{2.21}$$

$$f^h \leq 5000, \tag{2.22}$$

where f^h denotes frequency in hertz. Each of the lower formants can be modeled by a continuous-time filter with a complex-conjugate pair of poles [7]:

$$H_n(f^h) = \frac{1}{\left(1 - \frac{j \cdot 2\pi f^h}{\sigma_n + j\omega_n}\right)\left(1 - \frac{j \cdot 2\pi f^h}{\sigma_n - j\omega_n}\right)}, \tag{2.23}$$

$$n = 1, ..., 5, \tag{2.24}$$

where ω_n denotes the frequency of formant n in rad/s, and $\sigma_n < 0$ has magnitude equal to half the bandwidth of formant n in rad/s. A vocal tract filter for the vowel /ɔ/ is depicted in Fig. 2.4.

The nth formant bandwidth can be approximated as a function of the corresponding formant frequency $f_n = \frac{\omega_n}{2\pi}$ by a polynomial regression model [11]:

$$\sigma_n = -\pi(k_b + \sum_{i=1}^{5} x_i f_n^i)[1 + 0.25 \cdot (f_0 - 132)/88], \tag{2.25}$$

where

$$k_b = \begin{cases} 15.8, & \text{if } f_0 > 500; \\ 165, & \text{otherwise;} \end{cases} \tag{2.26}$$

$$x_1 = \begin{cases} 8.10 \cdot 10^{-2}, & \text{if } f_0 > 500; \\ -0.674, & \text{otherwise;} \end{cases} \tag{2.27}$$

$$x_2 = \begin{cases} -9.80 \cdot 10^{-5}, & \text{if } f_0 > 500; \\ 1.81 \cdot 10^{-3}, & \text{otherwise;} \end{cases} \tag{2.28}$$

$$x_3 = \begin{cases} 5.29 \cdot 10^{-8}, & \text{if } f_0 > 500; \\ -4.52 \cdot 10^{-6}, & \text{otherwise;} \end{cases} \tag{2.29}$$

$$x_4 = \begin{cases} -1.07 \cdot 10^{-11}, & \text{if } f_0 > 500; \\ 7.50 \cdot 10^{-9}, & \text{otherwise;} \end{cases} \tag{2.30}$$

$$x_5 = \begin{cases} 7.92 \cdot 10^{-16}, & \text{if } f_0 > 500; \\ -4.70 \cdot 10^{-12}, & \text{otherwise.} \end{cases} \tag{2.31}$$

In addition to models for the glottal excitation and for the oral branch of the vocal tract, we need a mechanism for modeling the nasal branch of the vocal tract, the room resonances in recording, and any intended or unintended distortion introduced by the recording equipment. Our solution is an additional frequency response function, i.e., $D(\cdot)$ in (2.1), that models the combination of these sources of spectral distortion. Due to the inherent uncertainty and complexity in the distortion, we define this frequency response function numerically as a piecewise-linear function, which is intended as a residual in source-filter analysis that fills the gap between the human voice recording and the analytic part of the model.

2.2 Source-Filter Analysis

2.2.1 Formulation

The purpose of this analysis is to estimate from a set of vocal sinusoids (with fundamental frequency f_0 and partial amplitudes $\{a_l\}_l$) the glottal pulse shape parameter R_d, formant frequencies $\mathbf{f} = (f_1, ..., f_5)^T = (\frac{\omega_1}{2\pi}, ..., \frac{\omega_5}{2\pi})^T$ (hertz), and distortion $D(\cdot)$ defined in Section 2.1. Estimation of R_d and $\mathbf{f}$ is formulated as minimization of the sum of squared distortion values at the partials:

$$(\hat{R}_d, \hat{a}, \hat{\mathbf{f}}) = \underset{(R_d, a, \mathbf{f}) \in \mathbb{S} \times \mathbb{R}_+ \times \mathbb{V}}{\arg\min} \sum_{l=1}^{L} d_l^2(R_d, a, \mathbf{f}), \tag{2.32}$$

$$d_l(R_d, a, \mathbf{f}) = a_l - 20 \log_{10} \left| a u_l K(l f_0) \prod_{n=1}^{5} H_n(l f_0) \right|, \tag{2.33}$$

where $L = \lfloor 5000/f_0 \rfloor$, and $\mathbb{V}$ specifies constraints for the formant frequencies:

$$\mathbb{V} = \left\{ \mathbf{f} \in \mathbb{R}^5 \;\middle|\; \begin{array}{l} 250 \le f_1 \le 1000 \\ 600 \le f_2 \le 3000 \\ 1700 \le f_3 \le 4100 \\ 2500 \le f_4 \le 4500 \\ 3000 \le f_5 \le 5500 \\ f_0 \le f_1 \le f_2 \le f_3 \le f_4 \le f_5 \end{array} \right\}. \tag{2.34}$$

Here the constraint $f_0 \le f_1$ simulates singers' formant tuning behavior at high pitches [16].

The estimate defined in (2.32) can be approximated with one of the following discrete pulse shape values:

$$r_k = 0.1 \cdot 3^{(k+11)/12} \in \mathbb{S},\; k = 1, ..., 25. \tag{2.35}$$

For each value of k, we calculate the formant frequencies that minimize the objective in (2.32):

$$(a^{(k)}, \mathbf{f}^{(k)}) = \underset{(a, \mathbf{f}) \in \mathbb{R}_+ \times \mathbb{V}}{\arg\min} \sum_{l=1}^{L} d_l^2(r_k, a, \mathbf{f}). \tag{2.36}$$

The final estimate is given by

$$(\hat{R}_d, \hat{a}, \hat{\mathbf{f}}) \approx (r_{k^*}, a^{(k^*)}, \mathbf{f}^{(k^*)}), \tag{2.37}$$

$$k^* = \arg\min_k \sum_{l=1}^{L} d_l^2(r_k, a^{(k)}, \mathbf{f}^{(k)}). \tag{2.38}$$

The distortion $D(\cdot)$ is constructed as a piecewise-linear function of frequency that interpolates the minimized distortion values $\{d_l(r_{k^*}, a^{(k^*)}, \mathbf{f}^{(k^*)})\}_{l=1}^{L}$. Since the latter distortion values occur at partial frequencies $f_0, 2f_0, ..., Lf_0$, linear interpolation only defines $D(\cdot)$ at frequencies between f_0 and Lf_0. For the frequency intervals $[0, f_0)$ and $(Lf_0, 5000]$, $D(\cdot)$ is defined by a constant value set to $d_1(r_{k^*}, a^{(k^*)}, \mathbf{f}^{(k^*)})$ and $d_L(r_{k^*}, a^{(k^*)}, \mathbf{f}^{(k^*)})$, respectively.

2.2.2 Optimization

The accuracy in determining the source and filter parameters depends on how well the objective in (2.36) is numerically minimized. To be specific, if the discrete pulse shape value r_k is closest to the truth shape value and the minimum of $\sum_{l=1}^{L} d_l^2(r_k, a, \mathbf{f})$ is overestimated due to minimization being trapped in a local minimum with respect to a and $\mathbf{f}$, then r_k may very likely turn out not to be selected in (2.38). My numerical experience revealed that the best of twenty local searches for the minimum defined in (2.36), which are initialized respectively with twenty different reference points, shows great consistency in preserving vocal timbre after F0 modification. These reference points differ only in the oral formant frequencies f_1, f_2, and f_3, with numerical values taken from gender-specific averages for ten vowels of American English [19]: i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, and ɝ. Although each individual search is local by nature and can only be expected to give a local minimum in some neighborhood of the corresponding starting point, the global minimum can be found as long as it can be reached from one of the twenty initial points.

Local search for the minimum defined in (2.36) may be achieved with any local optimization technique. Here we use a simple coordinate descent algorithm, as represented in Figure 2.5, where each (all-variable) update consists of a series of one-variable updates. Each one-variable update minimizes the objective with respect to the updated variable alone while fixing all the other variables. For instance, the update of formant frequency f_2 in the jth all-variable update operates on the current point

$$(a^j, f_1^j, f_2^{j-1}, f_3^{j-1}, f_4^{j-1}, f_5^{j-1})^T \tag{2.39}$$

by computing

$$f_2^j = \mathop{\arg\min}_{f_2 \in I^{(2)}} \sum_{l=1}^{L} d_l^2 \left(r_k, a^j, (f_1^j, f_2, f_3^{j-1}, f_4^{j-1}, f_5^{j-1})^T \right), \tag{2.40}$$

$$I^{(2)} = \{ f_2 \in \mathbb{R} \mid 600 \leq f_2 \leq 3000, f_1^j \leq f_2 \leq f_3^{j-1} \}. \tag{2.41}$$

In my implementation, the subproblem (2.40) is solved by finding a local minimum over a 100-hertz-spaced sampling of f_2 around f_2^{j-1}. The subproblem for updating the amplitude a can be solved analytically, as it is equivalent to minimizing a quadratic function of a. The final numerical solution to the problem (2.36) is refined by continuing the local search with a 10-hertz spacing of formant frequency sampling.

2.3 Source-Filter Synthesis

With the glottal pulse shape parameter, formant frequencies, and distortion estimated from the vocal sinusoids as $\hat{R}_d$, $\hat{\mathbf{f}}$, and $D(\cdot)$, respectively, a new set of vocal sinusoids can now be synthesized with a target fundamental frequency f_0'. Since the pulse shape and the formant frequencies represent the specific timbre of the original vocal sinusoids, reusing them in the synthesis would serve the purpose of preserving the vocal timbre. The synthesis is a straightforward application of the model of human voice production (2.1) to the following parameter settings:

$$f_0 = f_0', \tag{2.42}$$

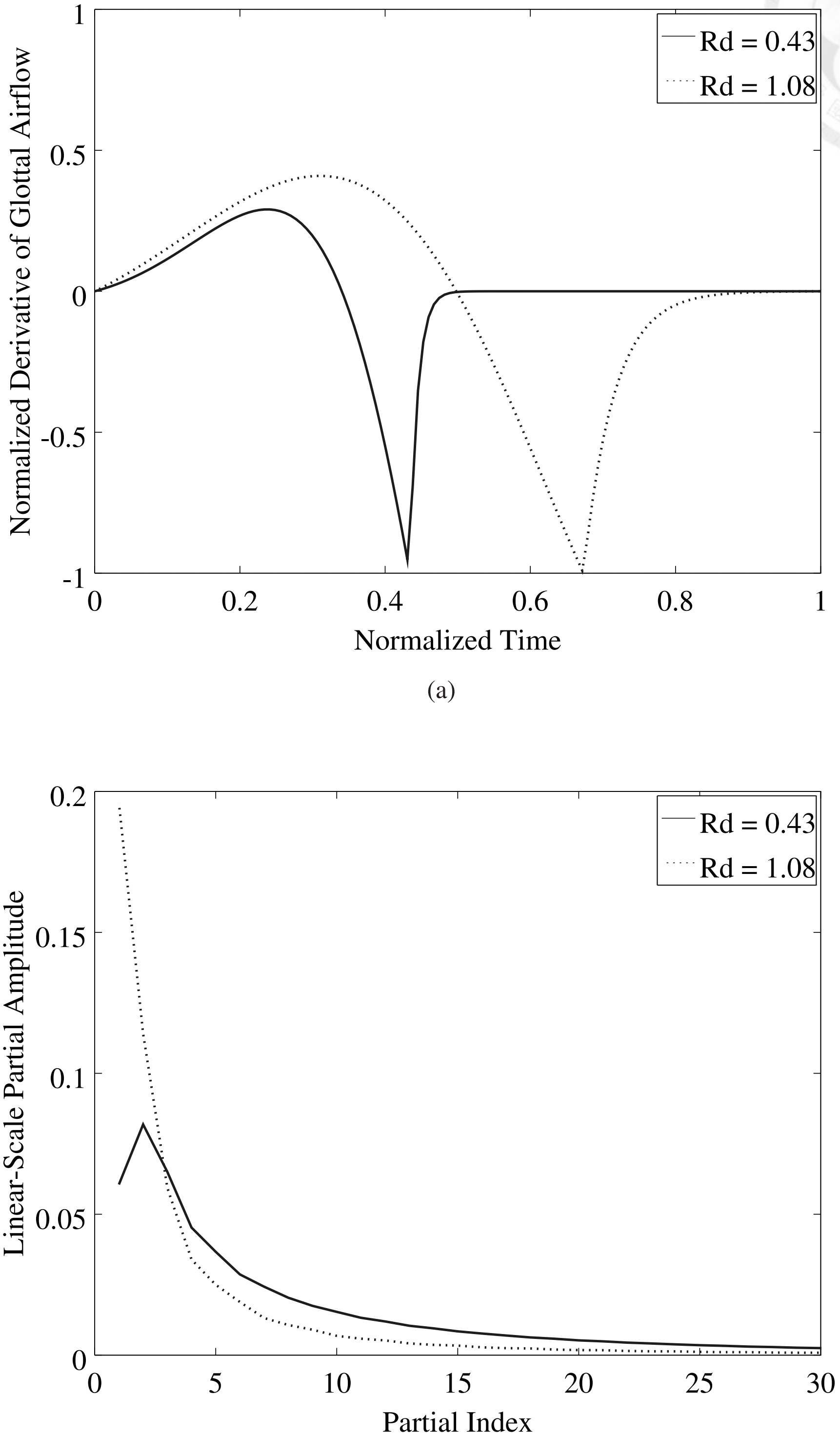
$$R_d = \hat{R}_d, \tag{2.43}$$

$$a = 1, \tag{2.44}$$

$$f_1 = \begin{cases} \hat{f}_1, & \text{if } \hat{f}_1 \geq f_0'; \\ f_0', & \text{otherwise,} \end{cases} \tag{2.45}$$

$$f_n = \hat{f}_n, \; n = 2, \ldots, 5, \tag{2.46}$$

where (2.45) simulates formant tuning [16].



(a)

(b)

Figure 2.3: Two glottal pulse shapes in the transformed Liljencrants-Fant model that are specified by shape parameter values $R_d = 0.43$ and $R_d = 1.08$. (a) In the time domain. (b) In the frequency domain.

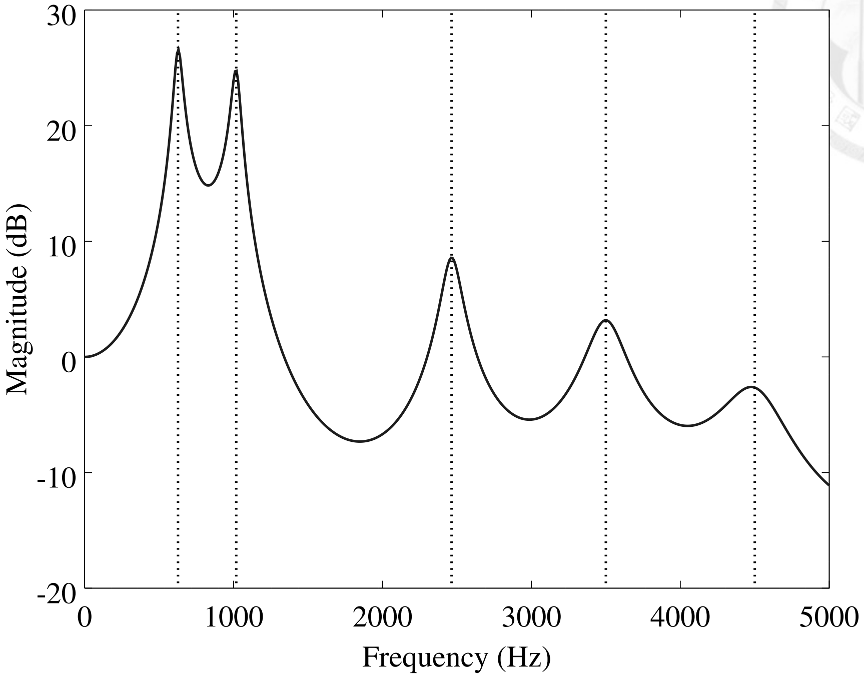


Figure 2.4: Frequency response of a vocal tract filter whose first five formant frequencies are marked by dotted lines.

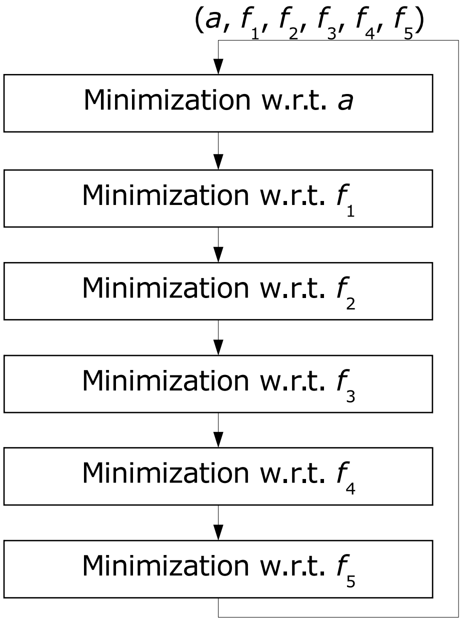

$(a, f_1, f_2, f_3, f_4, f_5)$
Minimization w.r.t. a
Minimization w.r.t. f_1
Minimization w.r.t. f_2
Minimization w.r.t. f_3
Minimization w.r.t. f_4
Minimization w.r.t. f_5

Figure 2.5: Each update in the local search for the minimum consists of a series of one-variable subproblems.

Chapter 3

Vocal Melody Extraction Based on an F0 Likelihood Model

The outline of this chapter is as follows. The complete system is summarized in Section 3.1. Section 3.2 presents an acoustic-phonetic model of F0 likelihood, which defines how likelihood scores of F0 candidates are evaluated for each analysis time position. A procedure is given in Section 3.3 for constructing the vocal melody from the temporal evolution of these likelihood scores. Experiments and results are documented in Section 3.4.

3.1 System Overview

A block diagram of the complete system is shown in Fig. 3.1. A vocal melody is extracted from the accompanied singing signal as a vocal F0 sequence and a set of vocal rests. The vocal F0 sequence specifies an F0 value for each and every analysis time position, where a vocal may or may not exist. This sequence is generated by an estimation procedure that depends on the proposed F0 likelihood model. Vocal rests are detected from the input signal by locating particular time positions where an F0 estimate implies a low vocal loudness level.

3.2 Acoustic-Phonetic Model of F0 Likelihood

Let X be an observed continuous random variable, Θ be an unobserved continuous random variable, and $p_{X|\Theta}(\cdot|\theta)$ denote the conditional density function of X given $\Theta = \theta$. Then

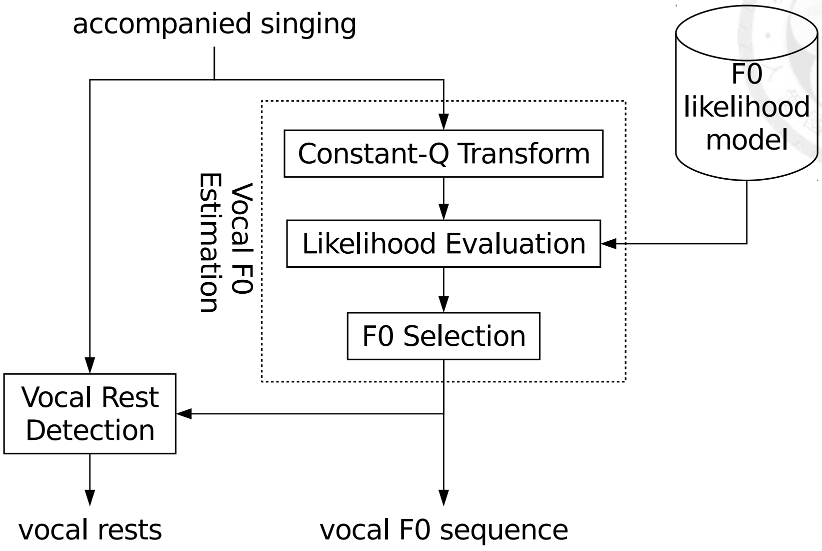


Figure 3.1: Block diagram of the complete system.

the function $\mathscr{L}_{\Theta|X}(\cdot|x)$, defined by

$$\mathscr{L}_{\Theta|X}(\theta|x) = p_{X|\Theta}(x|\theta) \tag{3.1}$$

and considered as a function of θ, is called the *likelihood function* of Θ given $X = x$. In this section, I describe a model of the likelihood function of vocal F0. In this model, each particular F0 value is associated with a sequence of partial amplitudes extracted from the spectrum of input signal. These partial amplitudes exhibit a specific timbral quality that, when checked against some vocal timbre examples, indicates how likely the F0 value gives the true vocal F0. In addition, overall loudness of the partials is also taken into account in likelihood evaluation so that low-loudness F0 values can be rejected.

Let the N-sample windowed input signal centered at time m be denoted by random N-vector $\mathbf{z}_m$, and let the corresponding quantized vocal F0 be denoted by an integer random variable w_m that measures the distance of the F0 from a reference low frequency in quarter tones. In quantizing a continuous F0 value to a discrete quarter-tone value, one produces a quantization error of 25 cents at the maximum, which will be tolerated (with a 25-cent margin) by an F0 accuracy measure that ignores all F0 errors below 50 cents. We model the likelihood function of w_m with a *timbral fitness measure* $F_h(\cdot)$ and a *loudness measure* $F_e(\cdot)$:

$$\mathscr{L}_{w_m|\mathbf{z}_m}(w|\mathbf{z}) \propto F_h(\mathbf{z}, w) \cdot F_e(\mathbf{z}, w). \tag{3.2}$$

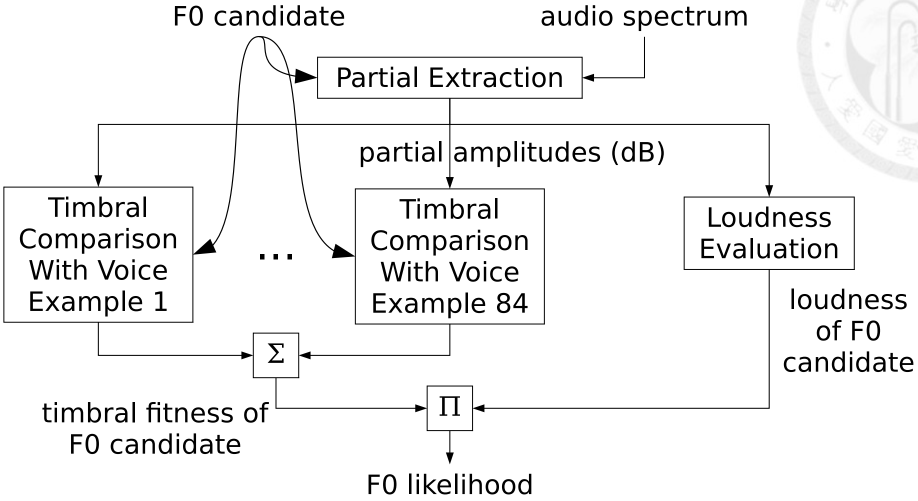


Figure 3.2: Architecture of the F0 likelihood model.

This frame-wise model will give emission probabilities in a hidden Markov model [26]. Since Viterbi search is invariant to an arbitrary scaling factor applied to all the emission probabilities, here I omit any scaling constant that would be necessary for the likelihood model to conform with the definition of probability density function. Architecture of this model is represented in Fig. 3.2.

3.2.1 Timbral Fitness Measure

To define the timbral fitness of F0 value w with respect to input signal $\mathbf{z}$, we calculate from $\mathbf{z}$ a constant-Q magnitude spectrum with quarter-tone-spaced frequency bins [2]. In an effort to simulate the dependency of human loudness perception on frequency, we correct the magnitude spectrum according to trends in the 40-phon equal-loudness contour (ELC) [15]:

$$A_f^z = |[\mathrm{CQT}\{\mathbf{z}\}]_f| \cdot 10^{(40-\kappa_f)/20},\ f \in I, \tag{3.3}$$

where $\mathrm{CQT}\{\cdot\}$ denotes the constant-Q transform, I denotes the set of frequency bins, f denotes the frequency index, and κ_f denotes the 40-phon ELC. Among all the frequency bins, we focus on a set of N_h partial frequencies[1] constructed from w:

$$I_w = \{\mathrm{round}(w + 24\log_2 l)\}_{l=1}^{N_h} \subset I. \tag{3.4}$$

[1] For values of numerical parameters, see Table 3.2.

The timbral quality exhibited by these partials is compared to N_s vocal timbre examples to determine its similarity to vocal timbre:

$$\bar{F}_h(\mathbf{z},w) = \frac{1}{N_s}\sum_{i=1}^{N_s}\exp\left\{-\frac{c_h}{N_h}\sum_{f\in I_w}\Delta_{i,f}^2\right\}, \tag{3.5}$$

$$\Delta_{i,f} = 20\log_{10}\frac{A_f^z}{\tilde{A}_f^{(i,w)}}, \tag{3.6}$$

where c_h is a nonnegative parameter that scales the effect of magnitude deviation on the likelihood score, and $\tilde{A}_f^{(i,w)}$ denotes a scaled spectrum of the ith vocal timbre example whose mean magnitude over the partials is aligned with the input signal to factor out volume level of the input signal:

$$\log_{10}\tilde{A}_f^{(i,w)} = \log_{10}A_f^{(i,w)} + \frac{1}{N_h}\sum_{f'\in I_w}\log_{10}\frac{A_{f'}^z}{A_{f'}^{(i,w)}}. \tag{3.7}$$

Here $A_f^{(i,w)}$ denotes the ELC-corrected constant-Q magnitude spectrum of the ith vocal timbre example that has been F0-modified to w with the procedure presented in Chapter 2. Note that $\sqrt{\sum_{f\in I_w}\Delta_{i,f}^2}$ represents the Euclidean distance between the observed partial-amplitude vector and the ith example partial-amplitude vector, which is converted to a similarity score by (the monotonically decreasing portion of) the Gaussian function in (3.5). The mechanism for comparing the observed timbre to an example is depicted in Fig. 3.3. A high value of $\bar{F}_h(\mathbf{z},w)$ would always imply that w is the true vocal F0 because its partials exhibit vocal timbre. Nevertheless, a low value of $\bar{F}_h(\mathbf{z},w)$ would imply a false F0 only when few of the partials suffer severe interference from the accompaniment. In case that some vocal partials should be far from any formant and an unpitched percussion instrument in the accompaniment should produce strong wide-band noise that interferes with these partials, the resulting similarity score would still be low. Pitched instruments in the accompaniment would do far less harm than the unpitched percussion because human voice is highly unstable in F0 and infrequently forms a nearly harmonic interval with a pitched instrument that cannot be resolved by my implementation of time-frequency analysis.

We make two refinements to the basic form $\bar{F}_h(\cdot)$:

- Since the F0 one octave above the true vocal F0 could have a high timbral fitness score calculated from the even partials, I attempt to identify the true F0 with a threshold parameter θ_h and reduce the score for the wrong F0 with a discount parameter $\gamma_h < 1$: If $\bar{F}_h(\mathbf{z},w-24) > \theta_h$,

$$\tilde{F}_h(\mathbf{z},w) = \gamma_h \cdot \bar{F}_h(\mathbf{z},w); \tag{3.8}$$

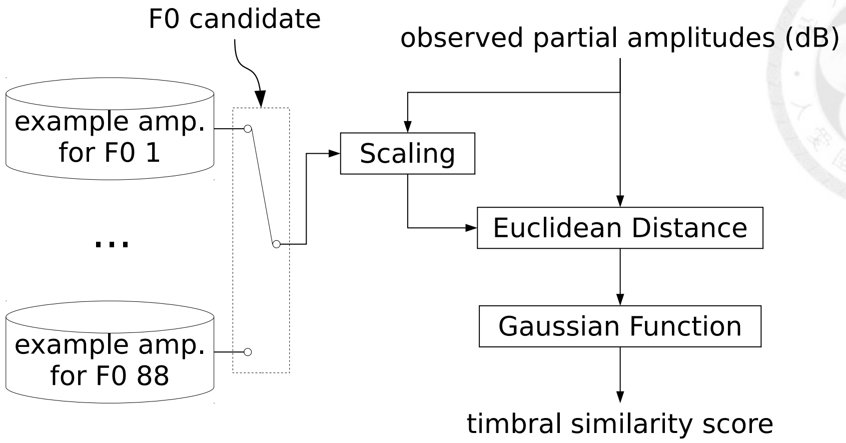


Figure 3.3: Timbral comparison to a vocal timbre example.

otherwise,

$$\tilde{F}_h(\mathbf{z},w) = \bar{F}_h(\mathbf{z},w). \tag{3.9}$$

- To prevent the comparison in timbral fitness from being dominated by any extremely high or low fitness value, we define the timbral fitness measure[2] $F_h(\cdot)$ by subjecting $\tilde{F}_h(\cdot)$ to an upper limit U_h and a lower limit L_h.

Vocal Timbre Examples

I construct the collection of N_s vocal timbre examples ($N_s = 84$) from 14 recordings of about 1 minute each. The 14 recordings (from 14 singers) 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 from the MIR-1K data set [12]. From each recording, 6 time positions are selected such that the person sings 6 vowel types respectively at these time positions, and that no partial of the vocal suffers significant interference from the accompaniment. For each of the 84 time positions, I conduct sinusoidal analysis, manually identify the fundamental of the vocal, and extract the frequencies and amplitudes of the vocal partials up to 5 kHz as one of the vocal timbre examples. Although selecting qualified time positions from accompanied singing recordings is time-consuming, these realistic recordings facilitate a collection that ensures high performance quality and high specificity in voice type and genre.

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

[2] For an example of timbral fitness scores calculated for F0 candidates, see Fig. 3.14.

(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" artists 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, possibly 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.

3.2.2 Loudness Measure

In an attempt to evaluate the loudness of F0 value w with respect to input signal $\mathbf{z}$, we conduct sinusoid tracking on the spectrum $\{A^{\mathbf{z}}_f\}_{f\in I}$, thereby generating a binary partition of the set of frequency bins I. A frequency bin where a sinusoid is detected is called a *sinusoidal frequency bin*, while one where no sinusoid is detected is called a *non-sinusoidal frequency bin*. With this partition, we can calculate a *sinusoidal power spectrum* from $\mathbf{z}$:

$$P^{\mathbf{z}}_f = (A^{\mathbf{z}}_f)^2 \cdot S_{\mathbf{z}}(f),\ f \in I, \tag{3.10}$$

where $S_{\mathbf{z}}(\cdot)$ maps sinusoidal frequency bins to unity and non-sinusoidal frequency bins to zero. Among all the frequency bins, we focus on a set of N_e partial frequencies constructed from w:

$$J_w = \{\text{round}(w + 24\log_2 l)\}_{l=1}^{N_e} \subset I. \tag{3.11}$$

The loudness is evaluated by summing sinusoidal power over these partials:

$$\bar{F}_e(\mathbf{z}, w) = \sum_{f\in J_w} P^{\mathbf{z}}_f. \tag{3.12}$$

We make two refinements to the basic form $\bar{F}_e(\cdot)$:

- To prevent the comparison in loudness from being dominated by any extremely high or low loudness value, we define function $\tilde{F}_e(\cdot)$ by subjecting $\bar{F}_e(\cdot)$ to an upper limit U_e and a lower limit L_e.
- To scale the effect of loudness on the likelihood score, we define the loudness measure[3] $F_e(\cdot)$ by raising $\tilde{F}_e(\cdot)$ to the power of $c_e \geq 0$.

[3] For an example of loudness values estimated for F0 candidates, see Fig. 3.15.

Sinusoid Detection

Here I describe a procedure for detecting sinusoids from a constant-Q magnitude spectrum $\{A^z_f\}_{f \in I}$. A sinusoid is detected at frequency $f \in I$ if the spectrum has a local maximum at f that is sufficiently prominent:

$$20\log_{10} A^z_f > \theta_p + \frac{1}{2N_a+1} \sum_{f'=f-N_a}^{f+N_a} 20\log_{10} A^z_{f'}, \tag{3.13}$$

where θ_p is a nonnegative parameter that specifies a minimum degree of prominence, and N_a controls the number of frequency bins around f over which an average magnitude level is calculated.

Sinusoid Pruning

Pruning of detected sinusoids is based on a set of sinusoidal contours on the time-frequency plane that are produced by tracking the sinusoids over a period of time [23]. In tracking a sinusoid from one time position to the next, we require that the frequency of sinusoid should not change by an amount (in quarter tones) exceeding θ_j, where the frequency is estimated by quadratically interpolating the three magnitude values at and next to the local maximum. In addition, to limit the difference in amplitude of sinusoid between the two time positions, we require that the ratio between the two amplitude values (the lower value divided by the higher value, on the linear scale) should exceed a parameter $\theta_g < 1$.

A sinusoid is removed if the sinusoidal contour to which it belongs has not only a duration (in units of 10 ms) exceeding θ_d but also a width of frequency range (the difference between the highest and lowest frequencies in quarter tones) below θ_r. By pruning sinusoids in this fashion, the procedure effectively removes some instrumental components from the input signal in that human voice does not typically maintain a constant F0 over a long period of time.

3.3 Vocal Melody Extraction

To estimate the vocal F0 at various time positions in the accompanied singing signal, we consider an equally spaced sequence of M time positions at intervals of 10 milliseconds. We assume that the vocal F0 varies continuously with time throughout the duration of signal, and that vocal rests are short time periods during which the vocal F0 variations continue with an extremely low loudness. This assumption could be explained by the fact that the tension in the vocal folds varies continuously with time, even in vocal rests, and

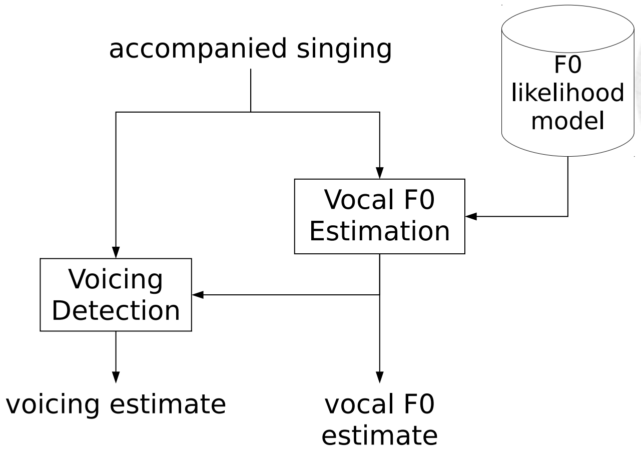


Figure 3.4: Block diagram for vocal melody extraction.

determines the F0 of voiced sound [32]. Even in the case of a disjunct melodic motion across a vocal rest, vocal folds with continuous tension would require a sufficiently long vocal rest to make preparation for the jump. With this assumption, the procedure estimates an uninterrupted sequence of vocal F0s from the sound mixture. After the estimation, vocal rests will be detected so as to remove the low-loudness F0s from the final vocal melody extract, as shown in Fig. 3.4.

3.3.1 Vocal F0 Estimation

In the estimation, vocal F0 is measured as a distance in quarter tones from the frequency of 21.205 hertz. For each time position m, vocal F0 is represented by a discrete random variable with 88 possible values, which sample the frequency range of 80–988 hertz[4] at quarter-tone intervals: $w_m \in \{46, 47, ..., 133\}$. This quantization makes it possible for us to describe the joint distribution of the vocal F0 sequence and the accompanied singing signal with an 88-state hidden Markov model (HMM) [26]. Once the likelihood scores of all the F0 candidates are calculated for all the time positions according to the F0 likelihood model presented in Section 3.2, the vocal F0 sequence can be estimated by maximizing the posterior probability of state sequence with the Viterbi algorithm, as shown in Fig. 3.5.

The F0 sequence given by the Viterbi search can be refined by considering the pitch range width of a singer, as shown in Fig. 3.6. We assume that this width (in quarter tones) never exceeds a parameter $W_r < 88$, and construct an F0 range from the unrefined F0

[4] According to the New Harvard Dictionary of Music, the range from the lowest pitch of bass to the highest pitch of soprano is E2–A5, or 82.4–880.0 hertz.

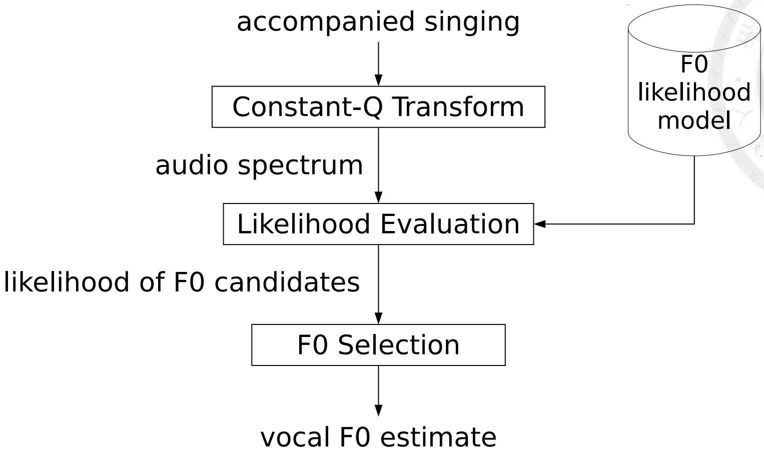


Figure 3.5: Block diagram for vocal F0 estimation.

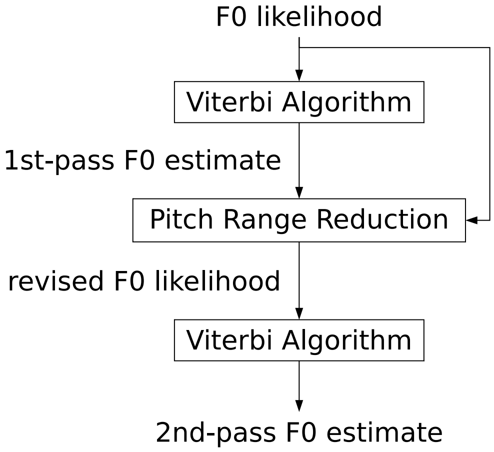


Figure 3.6: Block diagram for F0 selection.

sequence. The range (with width W_r) is constructed by placing its center at the median of the unrefined F0 sequence. With the F0 range constructed, the F0 sequence is refined by repeating the Viterbi search with all likelihood scores for F0s outside this range set to zero.

We use a Markov chain $\{w_m\}_{m=1}^{M}$ to model the vocal F0 sequence:

$$\begin{aligned} P(w_1,...,w_M) = {} & P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)\cdots \\ & P(w_m|w_1,...,w_{m-1})\cdots \\ & P(w_M|w_1,...,w_{M-1}) \qquad (3.14) \\ = {} & P(w_1)\prod_{m=2}^{M} P(w_m|w_{m-1}). \qquad (3.15) \end{aligned}$$

The equality in (3.15) results from the Markovianity that given the previous F0 w_{m-1}, the current F0 w_m is independent of all the earlier F0s $w_{m-2}, w_{m-3}, ..., w_1$. The initial state distribution is assumed to be uniform over all possible F0 values:

$$P(w_1 = w) = \frac{1}{88}, \ \forall w \in \{46, 47, ..., 133\}. \tag{3.16}$$

As shown in Fig. 3.7, the state transition probability distribution constrains each vocal F0 to stay within 1 quarter tone of the previous F0:

$$P(w_m = w | w_{m-1} = w') = p_{w'-45, w-45}, \tag{3.17}$$

$$p_{i,j} = \begin{cases} \frac{1}{v_p+1}, & \text{if } (i,j) = (1,1), (88,88) \\ \frac{v_p}{v_p+1}, & \text{if } (i,j) = (1,2), (88,87) \\ \frac{1}{2v_p+1}, & \text{if } 2 \leq i = j \leq 87 \\ \frac{v_p}{2v_p+1}, & \text{if } 2 \leq i \leq 87 \text{ and } j = i \pm 1 \\ 0, & \text{otherwise,} \end{cases} \tag{3.18}$$

where $v_p > 0$ controls the tendency for the vocal F0 to deviate from the previous F0. When the previous F0 is not at the boundary of the vocal pitch range, there are 3 F0 values around the previous F0 that are assigned a nonzero probability for the current F0. When the previous F0 is at the boundary, only the same F0 value and the neighboring value inside the boundary are possible for the current F0. Note that this transition model simulates the physical continuity in the temporal variations of the F0 of vocal-fold vibrations, rather than any symbolic or music-theoretic melodic motion from one member pitch of a scale to another. This simulation is justified by the relatively high analysis frame rate, i.e., 100 analysis time positions per second.

3.3.2 Vocal Rest Detection

We detect vocal rests as time positions where the loudness of singing voice is low. With vocal F0 at time position m estimated as $\hat{w}_m$, the vocal loudness at time position m can readily be estimated as $\bar{F}_e(\mathbf{z}_m, \hat{w}_m)$. To derive an adaptive loudness threshold, a representative loudness value is calculated by median-filtering (with filter length N_m) the sequence of vocal loudness values and taking the maximum of the filter output. The loudness threshold is calculated by multiplying the representative loudness and a parameter $\theta_v \in (0,1)$. In addition, we apply an upper limit U_v to the threshold.

Within true vocal rests, the estimated vocal loudness is typically low and allows effective detection of these rests. This is illustrated in Fig. 3.8, where 4 partials of the vocal

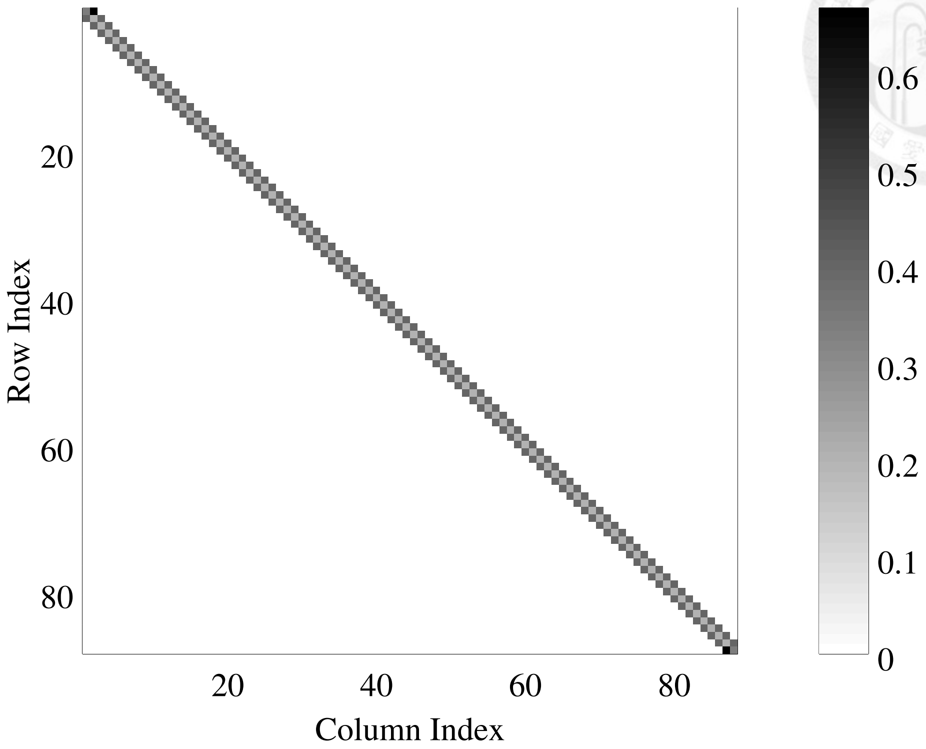


Figure 3.7: Matrix of transition probabilities ($[p_{i,j}]_{i,j}$ with $v_p = 2.0$) used in Viterbi search.

F0 estimate in a vocal rest are plotted on top of the spectrogram of analyzed signal, and the loudness estimated from these partials is also plotted. As we can see in the figure, the sequence of vocal F0 estimates are constrained to form a continuous contour that connects the true vocal F0s before and after the rest. This constraint prevents the F0 estimate from freely reaching the loudest instrumental F0 in the vocal rest, thereby giving typically low loudness values that characterize the vocal rests.

3.4 Experiments

3.4.1 Data Sets

The methods presented in Sections 3.2 and 3.3 are reliant on a set of numerical parameters. To ascertain reasonable values for these parameters without the risk of overfitting, I used 3 *development sets*: `adc2004`, `labrosa`, and `mir1k_dev`. Data set `adc2004` is a subset of the data used in the ISMIR 2004 Audio Description Contest (ADC 2004). ADC 2004 in its entirety consists of 20 audio recordings, among which 8 recordings have instrumental melodies, and the other 12 have vocal melodies. Since this work does not

Table 3.1: Summaries for the data sets.

Data Set	#Excerpts	Excerpt Length (s)
adc2004	10	~20
labrosa	9	~30
mir1k_dev	379	4–13
mirex05	16	10–40
mirex08	4	120
mirex09	374	4–13
mir1k	253	4–13
medley1	61	13–514
medley2	44	13–385 (vocal sections only)

consider instrumental melodies, only vocal recordings are included in adc2004, including 2 pop song excerpts, 4 song excerpts with synthesized vocal, and 4 opera excerpts. The other 2 vocal excerpts are not included here because they contain an ensemble of vocals. Data set labrosa consists of English popular song excerpts from the data set prepared for polyphonic melody extraction by LabROSA, Columbia University. Data set mir1k_dev consists of Chinese popular song excerpts with non-professional vocals and synthetic accompaniment from the MIR-1K data set [12], covering 4 male subjects (fdps, geniusturtle, jmzen, and Kenshin) and 4 female subjects (heycat, tammy, titon, and yifen). Each entry in mir1k_dev is produced by adding the vocal and accompaniment channels in the corresponding MIR-1K audio file without balance modification. Numbers of excerpts and lengths are listed in Table 3.1 for these data sets.

To assess the generalization performance of parameter optimization, I used 10 test sets without any overlap with the development sets: mirex05, mirex08, mirex09, mirex09+, mirex09-, mir1k, mir1k+, mir1k-, medley1, and medley2. Each data set with the prefix "mirex" refers to a full data set used in the Music Information Retrieval Evaluation eXchange (MIREX) Audio Melody Extraction task, to the 2014 edition of which I submitted the algorithm presented in this dissertation. An exception to this is mirex05, which is a vocal subset of the MIREX05 data set. Data set mirex08 is composed of Indian classical song excerpts. Data set mirex09 is composed of Chinese popular song excerpts with non-professional vocals and synthetic accompaniment. Data set mirex09+ is a variant of `mirex09` with accompaniment attenuated by 5.0 dB (MIREX09 +5dB), and for `mirex09-` the accompaniment is amplified by 5.0 dB (MIREX09 -5dB). Data set `mir1k` is a test counterpart of `mir1k_dev` that covers 2 male subjects (`abjones` and `bug`) and 2 female subjects (`amy` and `ariel`). Similarly to the MIREX09 sets, data sets `mir1k+` and `mir1k-` are balance-modified variants of `mir1k`. Data set `medley1` is an unaltered vocal subset of the MedleyDB data set [1], covering the genres of classical music, rock, pop, musical theatre, singer/songwriter, and jazz. Note that excerpts in `medley1` present specific variations in the existence of vocals that are realistic on the one hand, and on the other, deviate from what I believe would favor the use of timbral distinction between vocals and instruments, i.e., the existence of exactly one vocal mentioned in Section 1.2. For an opportunity to work with MedleyDB while factoring out this deviation, I derived another data set from MedleyDB that conforms with our assumption about vocal existence, which we refer to as `medley2`. Not all vocal excerpts in MedleyDB are included in `medley2`: Of the 61 vocal excerpts in MedleyDB, 17 are excluded because of their use of multiple vocals. Many vocal excerpts in MedleyDB are actually full songs and contain instrumental sections that last for tens of seconds. To ensure a roughly sustained presence of a vocal in the machine-analyzed audio of `medley2`, I manually split each such excerpt into an alternating sequence of vocal and instrumental sections. By ignoring short instrumental sections of length below 7.5 seconds, I identified from an excerpt 6 instrumental sections at the most, and one or more vocal sections where no vocal rest is longer than 7.5 seconds. For each split excerpt, only its vocal sections are used for the `medley2` evaluation, with a separate melody extracted from each vocal section. Stereo audio in MedleyDB is converted to monaural audio by averaging the left- and right-channel signals. Numbers of excerpts and lengths are listed in Table 3.1 for these data sets.

3.4.2 Performance Measures

In the experiments documented here, the tested system gives vocal melodies in the format of a voicing/F0 value for each analysis time position (100 equally spaced time positions per second). If voice is detected at a time position, the output specifies the F0 estimate for the time position; otherwise, the output specifies that the time position is in a vocal rest.

The MIREX evaluation adopts several measures for evaluating the performance of a melody extraction system [25]. In the first place, to determine how well the system performs voicing detection, we use the *voicing detection rate* and the *voicing false alarm rate*. The voicing detection rate is computed as the proportion of time positions that are both labeled and estimated to be voiced, among all the time positions that are labeled voiced. The voicing false alarm rate is computed as the proportion of time positions that are estimated to be voiced but are actually in a vocal rest, among all the time positions that are in a vocal rest according to the reference transcription.

Second, to determine how well the system performs F0 estimation, we use the *raw pitch accuracy* and the *raw chroma accuracy*. The raw pitch accuracy is computed as the proportion of time positions that are labeled voiced and have F0 estimated within one quarter tone of the true F0, among all the time positions that are labeled voiced.[5] 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 F0 is here measured in terms of chroma, or pitch class, a quantity derived from the F0 by wrapping the F0 into one octave.

Finally, the performance of voicing detection and F0 estimation can be measured jointly by the *overall accuracy*, defined as the proportion of time positions that receive correct voicing classification and, if voiced, an F0 estimate within one quarter tone of the true F0, among all the time positions.

3.4.3 Results on the Development Sets

The search for optimal parameter values was performed on the development sets in 2 stages. In the first stage, errors in extracted vocal melodies were analyzed to indicate particular parameter values that were responsible for the errors. Each error analysis was expected to suggest an adjustment in parameter value that would eliminate the analyzed error. Parameters were repeatedly and selectively adjusted according to such error analyses until the remaining errors could no longer be explained by an inadequate parameter value. In the second stage, parameters were further optimized by coordinate ascent in 4 cycles, with respect to a roughly 10-point sampling of each one-dimensional parameter space. In the first cycle, one-parameter line searches were performed, each as a one-dimensional grid search, for all the parameters in the order shown in Table 3.2. In each line search, an overall accuracy is evaluated on the development sets for each sample point. The sample point that gives the highest accuracy is assigned to the parameter as an update, which may or may not equal the old parameter value. The overall accuracy is averaged over all excerpts in each development set before being averaged among the 3 development sets without weighting. In the second and third cycles, the same one-parameter line searches were repeated, with increasingly more parameters receiving a stationary up-

[5] My implementation produces F0 estimates in estimated vocal rests to factor out voicing false negative errors in the raw pitch accuracy.

date. Finally in the fourth cycle with the same line searches, all the initial parameter values, as given by the third cycle, were found to be optimal with respect to the sampling of parameter space.

Table 3.2 shows the overall accuracy evaluated on the development sets for various parameter settings. Each row contains 3 parameter settings derived from the optimal setting by assigning 3 different values to the designated parameter: The first value is the smallest sample point, the second value is the optimal value, and the third value is the largest sample point. As shown in Fig. 3.9, for the number of partials N_h, the accuracy well over 0.7 for $N_h = 2$ suggests that the first 2 partials could play an important role in timbral comparison, and the optimal value 6 suggests that amplitude observations for partials above the sixth may be relatively noisy due to low partial amplitude. As shown in Fig. 3.10, for the lower limit of loudness L_e, the accuracy is close to 0.78 for all values of L_e between $4 \cdot 10^{-6}$ and $4 \cdot 10^{-5}$, but is close to zero for $L_e = 0$. A nonzero lower limit of loudness is crucial in that the posterior probability of a desired F0 sequence is zero if it has a zero loudness value in a vocal rest. As shown in Fig. 3.11, the optimal width of pitch range $W_r = 44$ is close to the typical 2-octave pitch range of a singer; however, the accuracy for this width is only marginally higher than that for the much larger width of 60 quarter tones. This suggests that the median of the unrefined F0 sequence might not be an adequate estimate of the true center of pitch range, and that a non-Markovian prior model of the F0 sequence that incorporates this pitch range characteristic will be desirable if a computationally efficient search algorithm is developed for the model.

3.4.4 Results on the Test Sets

Performance of the proposed approach is evaluated on the test data sets, as presented in Table 3.3. For a comparative basis for performance assessment, note that many existing approaches gave overall accuracies between 0.6 and 0.8 for `mirex05` and between 0.7 and 0.8 for `mirex08` and `mirex09`, as shown in Table 3.4. Accordingly, we are now interested in whether the proposed approach can give overall accuracies all above 0.7 across various data sets. It is encouraging that the overall accuracies turned out to be each above 0.7, except for `medley1`, `medley2`, and the two strong-accompaniment data sets `mirex09-` and `mir1k-`. Multiple vocals and long instrumental sections, which are present in `medley1` but not in `medley2`, resulted in a 0.041 drop in raw chroma accuracy and smaller differences in the other measures, all suggesting a performance degradation. The tiny difference 0.006 in overall accuracy suggests a mixed effect of the inclusion of long instrumental sections—Vocal F0 estimation suffered slightly while vocal rests, which are detected at a rate around 80% according to voicing false alarm rate, constituted a larger proportion in the analyzed audio. Each of the raw chroma accuracies is only slightly higher than the corresponding raw pitch accuracy, which reveals that my method rarely gives octave errors. The average overall accuracy over the 5 data sets `mirex05`, `mirex08`, `mirex09`, `mir1k`, and `medley1` is 0.715, which is 0.065 below the development-set average 0.780 and demonstrates a reasonable generalization performance for the determination of parameter values driven by the development sets.

As shown in Table 3.4, my approach gave overall accuracies within 0.07 of those given by the best-performing approaches, i.e., [4], [27], and [13], for all the MIREX data sets except `mirex08`. Unfortunately, given the rather limited sizes (16 and 4 excerpts) of `mirex05` and `mirex08`, one would not consider the comparison statistically significant for these two sets. For each of the `mirex09` variants, I calculated a p-value by the two-sided Wilcoxon signed-rank test on the 374 pairs of overall accuracies given by an existing approach and my approach, as listed in Table 3.5. If we consider a significance level of 5%, 10 of the 15 `mirex09` (variant) comparisons are shown to have statistical significance. Note that the major difference between the proposed approach and the best-performing approaches [4], [27], and [13] would be the attempt to detect the spectral timbre of singing voice. Although the detection can be helpful for vocal melody extraction, it requires that most partials of the vocal should not suffer significant interference from the accompaniment, so that observed vocal partial amplitudes can truthfully represent the vocal timbre. Even if the vocal loudness is high, for certain vowels such as /i/ the partials between the first two formants can still be rather weak and susceptible to such interference. As a result, one would not strongly expect the proposed approach to outperform other mechanisms for vocal detection, vibrato detection in particular, which underlies the best-performing approaches [4], [27], and [13].

3.4.5 Examples

A numerically represented vocal melody extracted from a song signal can be auralized as a simple quasi-periodic signal that realizes the sequence of F0s and rests in the melody. By listening to the song signal and the auralized melody, one could make a personal assessment of the performance of extraction. By auralization, I present 25 melody extraction examples, which are made up of the 25 excerpts with overall accuracies ranking between the 115th and the 139th in `mir1k`. The examples are located at `http://homepage.ntu.edu.tw/~d98942017/english/melody.htm`.

Here I present analysis of the excerpt that gave the median overall accuracy 0.747

in mir1k, whose spectrogram is shown in Fig. 3.12. The likelihood scores of the 88 F0 candidates are displayed in gray scale in Fig. 3.13 for a sequence of time positions, where we can see a high-likelihood path around $10^{2.6} \approx 398$ hertz that is consistent with the F0 contour noticeable from Fig. 3.12. The timbral fitness scores are plotted in Fig. 3.14, where we can see some unwanted low-magnitude, high-fitness regions apart from the expected high-fitness path. These regions could result from leakage of vocal energy from the spectral main lobes of harmonically modulated Hamming windows to a number of distant side lobes, which gives a spectral trend resembling vocal timbre to non-sinusoidal frequency bins. Sinusoidal F0 candidates are clearly identified in the loudness image of Fig. 3.15. Vocal F0 estimation gave the F0 sequence plotted in Fig. 3.16, where F0 estimates are identical to ground-truth values for the most part. The result of vocal rest detection is depicted in Fig. 3.17, where two short bursts of voicing errors are noticeable around $t = 1.9$ and $t = 3.8$.

F0 and voicing errors in results for the median-accuracy excerpt shed some light on the limitations of my approach. The F0 error around $t = 2.8$ results from pruning of the vocal fundamental sinusoid due to its constant F0—Although vocal F0 is typically unstable, stability can be observed on some occasions. The short ground-truth F0 contour around $t = 3.8$, which the algorithm failed to track, is a rapid down-chirp that the sinusoid detection procedure cannot handle. This F0 error also led to the false vocal rest at the same time position. The ground-truth vocal rest around $t = 1.9$, which the algorithm failed to detect, contains a strong bass sound whose 3rd partial is coincidentally located one half tone below the immediately preceding vocal F0. This led the F0 estimation procedure to track the bass partial, leaving no low-loudness interval at the vocal rest. In brief, this analysis demonstrates that, essentially based on the typical characteristics of accompanied singing signals, my approach is slightly sensitive to extreme cases.

3.4.6 Results of Controlled Experiments

Timbre Model

In order to assess the effect of timbral fitness measure, I repeated the medley2 experiments with F0 likelihood defined by the loudness measure alone:

$$\mathscr{L}_{w_m|\mathbf{z}_m}(w|\mathbf{z}) \propto F_e(\mathbf{z}, w). \tag{3.19}$$

This timbre-insensitive F0 likelihood model gave results listed in row “Timbre” in Table 3.6. For lack of timbre modeling, the performance of vocal F0 estimation is significantly reduced with a decrease of 0.33 in raw pitch accuracy, which in turn leads to a decrease of 0.13 in overall accuracy. This confirms the effectiveness of the timbral fitness measure in guiding the identification of vocal F0s.

Improvement in F0 Likelihood Model

A predecessor of the current F0 likelihood model is derived from a white Gaussian model of accompaniment signal [3], with the likelihood defined by spectral comparison of the accompanied singing signal with vocal timbre examples for the F0 candidate. The full-spectrum comparison performed by the previous model virtually combines timbral fitness evaluation and loudness evaluation without a weighting. In contrast, the current model carries out timbral comparison with respect to partial amplitudes and scales the effect of timbre and loudness with the parameters c_h and c_e. With a proper weighting, selection of high-loudness instrumental F0s can be avoided in vocal F0 estimation.

In an attempt to determine the practical effect of this model improvement, I repeated the `medley2` experiments with F0 likelihood evaluated by the previous model, giving results listed in row “New Likelihood” in Table 3.6. The improvement is confirmed by a difference of 0.11 in overall accuracy.

Transformed Liljencrants-Fant Model

In the F0 modification of vocal timbre examples, we implement an invariance in the glottal pulse shape, which is represented by the transformed Liljencrants-Fant model with a parameter (R_d) value estimated from the original example. To evaluate the effect of this specific model of glottal excitation, I repeated the `medley2` experiments with the (radiated) glottal excitation modeled in the conventional way by a fixed spectrum envelope [7]:

$$U_R(f^h) = \frac{f^h/100}{1+(f^h/100)^2}. \tag{3.20}$$

Note that an F0-invariant pulse shape implies F0-invariant amplitude ratios among the partials and thus a spectrum envelope that stretches along the frequency axis as the F0 increases. The repeated experiments gave results listed in row “LF Model” in Table 3.6. The difference of 0.13 in raw pitch accuracy confirms the advantage of transformed Liljencrants-Fant model in modeling vocal timbre.

Diversity Among Vocal Timbre Examples

The vocal timbre examples are collected with diversity in gender, genre, voice type, and vowel type. In an effort to evaluate the separate effect of these diversity factors, I repeated the `medley2` experiments with 4 subsets of the vocal timbre examples where one of these diversity factors has been eliminated. The 7-singer male-only (42-example) subset gave results listed in row "Gender" in Table 3.6. Labeled "Genre" are results for the 4-singer opera-only (24-example) subset. The row labeled "Voice Type" is for the 36-example high-voice-only subset, which consists of the tenor, the soprano, and the other 4 high-voice singers. Results obtained with the 14 examples of vowel /ɑ/ are found in the row titled "Vowel Type."

Comparison of these 4 rows of results to the full-feature results reveals that none of these diversity factors is associated with a drop larger than 0.05 in overall accuracy or raw pitch accuracy. This suggests that for the purpose of distinguishing vocal timbre from instrumental timbre, it is not necessary for the collection of vocal timbre examples to capture all the 4 diversity factors, and that the current 4-factor collection is sufficiently diverse for the purpose. For example, a sound of vowel /u/ can be much closer in timbre to an example of vowel /ɑ/ than is an instrumental sound to any example of vowel /ɑ/, so that the 14 examples of vowel /ɑ/ are sufficient for this application.

For the purpose of evaluating the composite effect of diversity factors, I repeated the `medley2` experiments with two subsets of the vocal timbre examples where 3 or 4 diversity factors have been eliminated. With the 6 tenor examples only, the repeated experiments gave results listed in row "Singer" in Table 3.6, which show that elimination of the 3 singer diversity factors lowers the raw pitch accuracy by 0.04. When only one vocal timbre example (tenor's /ɑ/) is used, lack of diversity results in a decrease of 0.27 in raw pitch accuracy, as shown in the row labeled "Diversity." This extremely biased representation of vocal timbre leads to accuracies slightly higher than those obtained with timbre entirely ignored. This confirms the significance of diversity in representing vocal timbre.

Loudness Model

To see the effect of loudness measure, I repeated the `medley2` experiments with F0 likelihood defined by the timbral fitness measure alone:

$$\mathscr{L}_{w_m|\mathbf{z}_m}(w|\mathbf{z}) \propto F_h(\mathbf{z},w). \tag{3.21}$$

This loudness-insensitive F0 likelihood model gave results listed in row "Loudness" in Table 3.6, where lack of loudness modeling brought about a dramatic drop in raw pitch accuracy. Many of the numerous incorrect F0 estimates are associated with low loudness and high timbral fitness that could result from leakage of vocal energy from spectral main lobes of harmonically modulated Hamming windows to a number of distant side lobes. This confirms the need for the loudness measure in guiding the identification of vocal F0s.

F0 Prior Modeling

To test the effect of prior modeling of vocal F0 sequence, I repeated the `medley2` experiments with the vocal F0 sequence estimated by maximizing the F0 likelihood separately for each time position. This frame-wise maximum-likelihood approach to vocal F0 estimation yielded results listed in row "Prior Model" in Tabel 3.6, where lack of prior modeling caused a 0.20 reduction in raw pitch accuracy. This confirms the need for smoothing likelihood-based predictions in fulfillment of the continuity constraint in vocal F0.

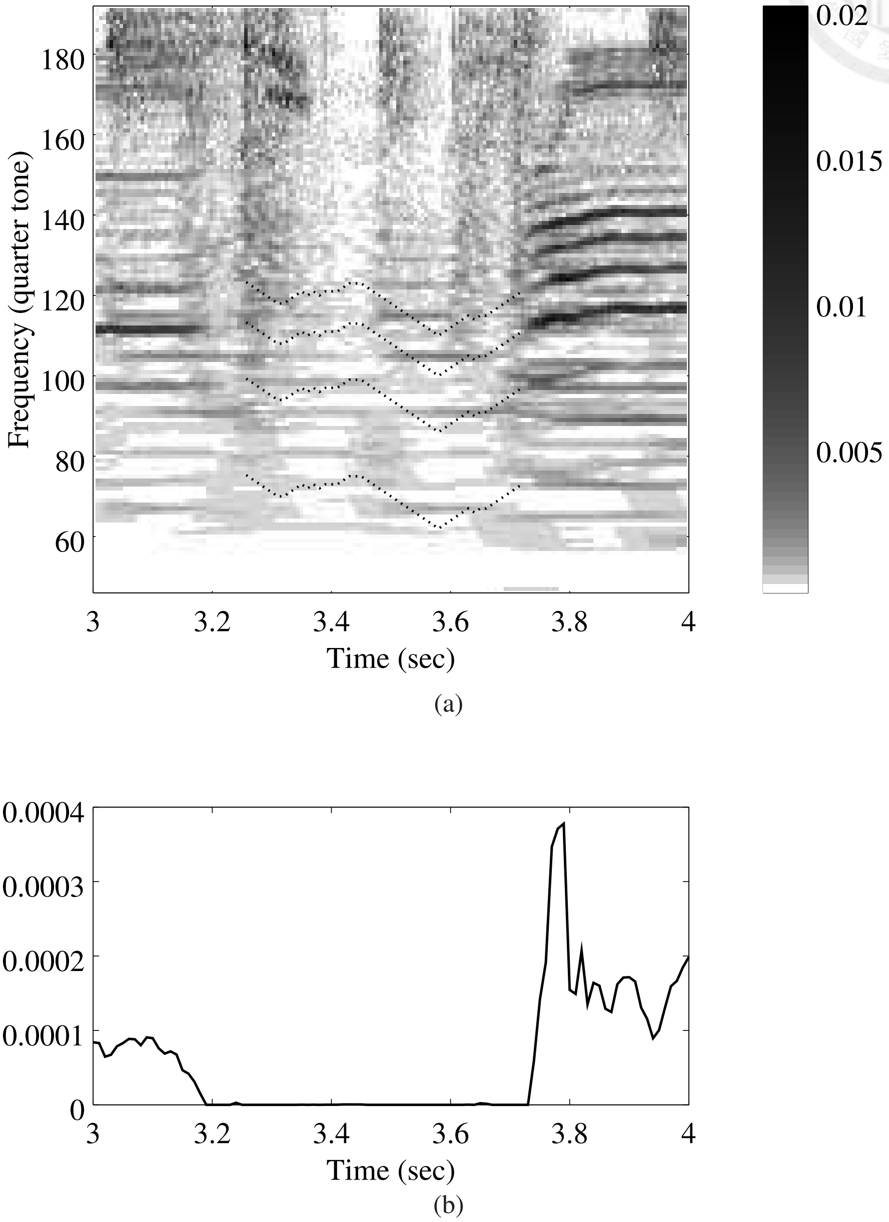


Figure 3.8: A vocal rest. (a) Spectrogram of analyzed signal overlaid with partial frequency contours (dotted lines) of the estimated vocal F0 in the vocal rest. Tick labels on the color bar are expressed in the unit of linear-scale signal magnitude. The frequency is measured as a distance from 21.205 hertz in quarter tones. (b) Estimated vocal loudness in the unit of linear-scale signal power. Some visible instrumental sinusoids crossed by the plotted partials have been pruned in loudness calculation.

Table 3.2: Optimal and extreme parameter settings and the resulting overall accuracies evaluated on the development sets. In each ordered pair, the first entry gives a parameter value and the second entry gives an accuracy. The reader can refer to Sections 3.2 and 3.3 for the units of parameter values.

Param.	**Smallest Value**	**Optimal Value**	**Largest Value**
N_h	(2, 0.750)	(6, 0.780)	(12, 0.772)
c_h	(0.005, 0.742)	(0.019, 0.780)	(0.025, 0.766)
c_e	(0.15, 0.768)	(0.55, 0.780)	(1.15, 0.776)
U_h	(0.05, 0.750)	(0.15, 0.780)	(0.5, 0.776)
L_h	(0.01, 0.776)	(0.02, 0.780)	(0.11, 0.651)
U_e	($4 \cdot 10^{-5}$, 0.734)	(0.00164, 0.780)	(0.00164, 0.780)
L_e	(0, 0.006)	($4 \cdot 10^{-5}$, 0.78)	($4 \cdot 10^{-5}$, 0.78)
v_p	(0.05, 0.730)	(2.0, 0.780)	(10, 0.571)
N_a	(1, 0.000)	(7, 0.780)	(10, 0.780)
θ_p	(0, 0.738)	(12.0, 0.780)	(20.0, 0.600)
θ_r	(0.1, 0.712)	(0.7, 0.780)	(2.1, 0.631)
θ_j	(0.5, 0.759)	(1.5, 0.780)	(1.5, 0.780)
θ_g	(0.05, 0.769)	(0.45, 0.780)	(0.75, 0.723)
θ_d	(1, 0.776)	(7, 0.780)	(21, 0.762)
N_e	(2, 0.708)	(4, 0.780)	(8, 0.772)
θ_h	(0.07, 0.774)	(0.19, 0.780)	(0.39, 0.761)
γ_h	(0.05, 0.772)	(0.35, 0.780)	(0.95, 0.764)
N_m	(1, 0.762)	(35, 0.780)	(280, 0.768)
U_v	(0, 0.631)	($1.2 \cdot 10^{-4}$, 0.78)	($2 \cdot 10^{-4}$, 0.779)
θ_v	($5 \cdot 10^{-4}$, 0.726)	(0.02, 0.780)	(1.0, 0.697)
W_r	(24, 0.750)	(44, 0.780)	(60, 0.776)

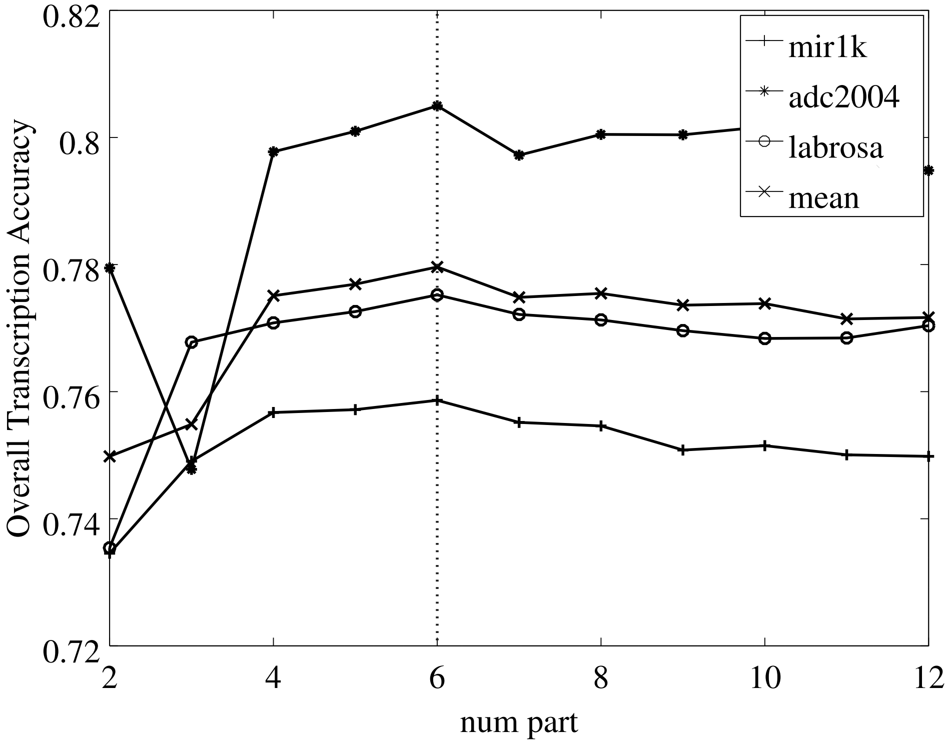


Figure 3.9: Overall accuracy evaluated on the development sets for sample points of N_h. The vertical dotted line marks the optimal parameter value.

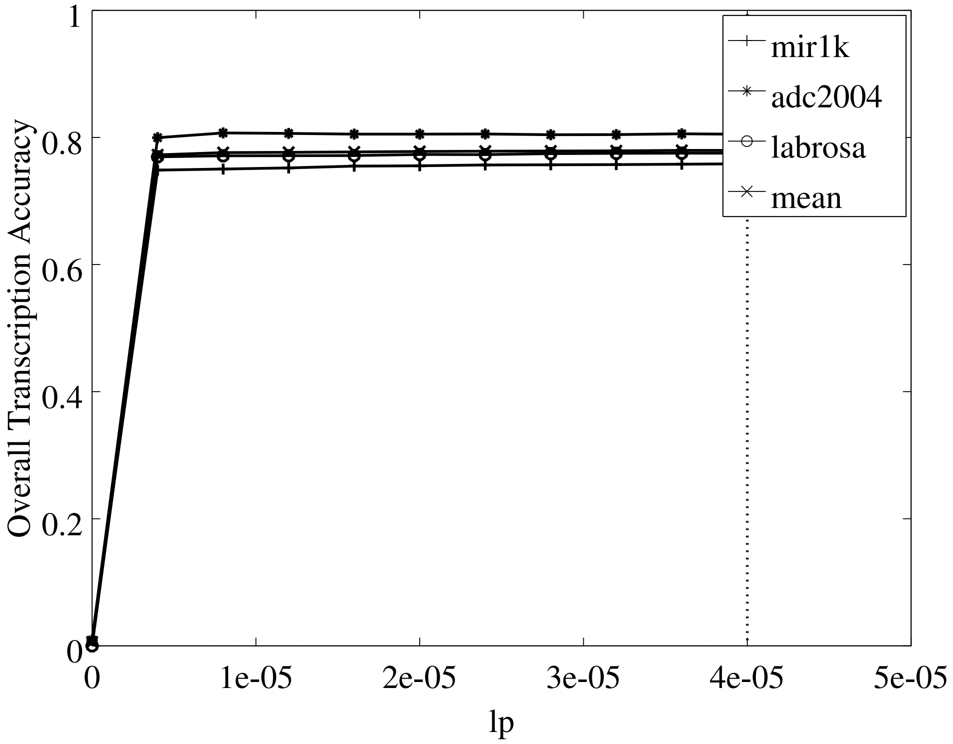


Figure 3.10: Overall accuracy evaluated on the development sets for sample points of L_e. The vertical dotted line marks the optimal parameter value.

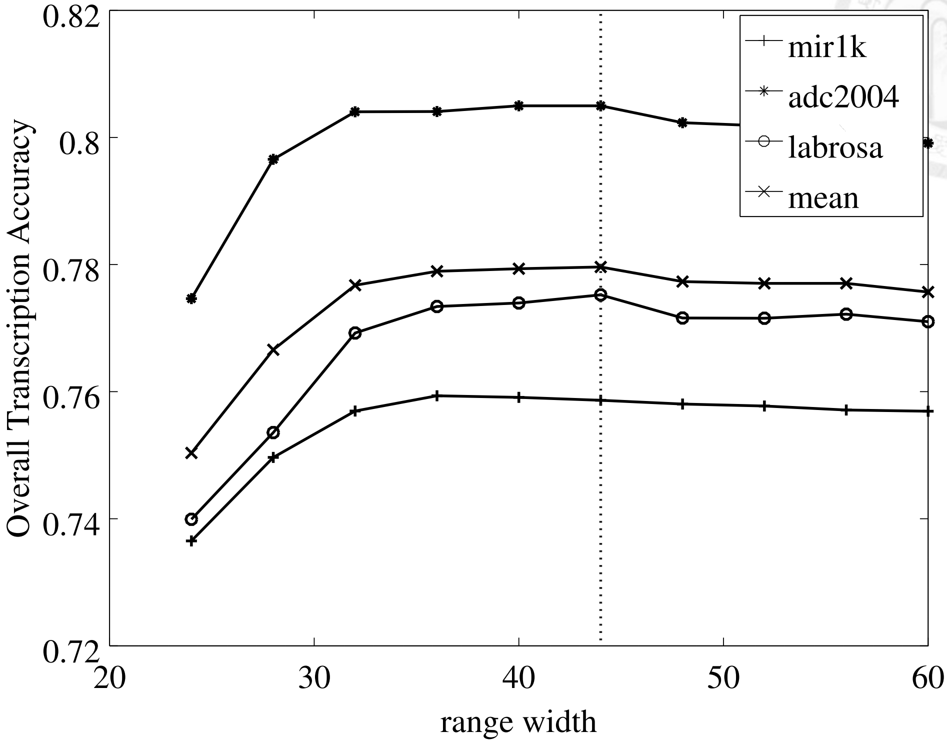


Figure 3.11: Overall accuracy evaluated on the development sets for sample points of W_r. The vertical dotted line marks the optimal parameter value.

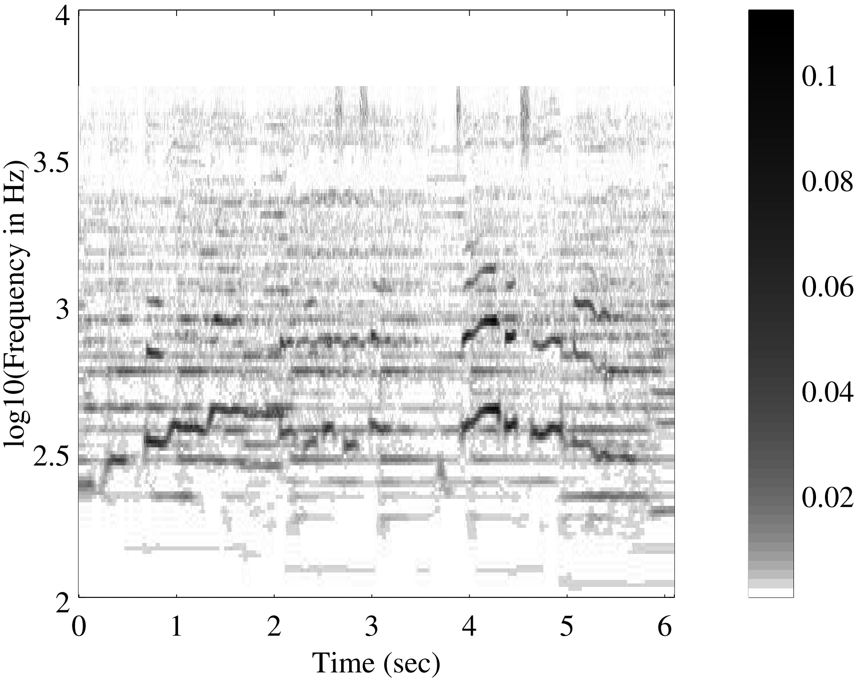


Figure 3.12: Spectrogram of the median-accuracy excerpt in `mir1k`.

Table 3.3: Results of performance evaluation on the test sets. OA = Overall Accuracy; RPA = Raw Pitch Accuracy; RCA = Raw Chroma Accuracy; VDR = Voicing Detection Rate; VFAR = Voicing False Alarm Rate.

Data Set	OA	RPA	RCA	VDR	VFAR
mirex05	0.703	0.727	0.767	0.751	0.144
mirex08	0.710	0.853	0.856	0.737	0.119
mirex09	0.739	0.785	0.797	0.791	0.197
mirex09+	0.817	0.850	0.857	0.816	0.086
mirex09-	0.594	0.629	0.660	0.738	0.336
mir1k	0.741	0.816	0.825	0.813	0.246
mir1k+	0.824	0.886	0.891	0.815	0.100
mir1k-	0.584	0.648	0.670	0.801	0.425
medley1	0.684	0.689	0.715	0.715	0.223
medley2	0.690	0.728	0.756	0.742	0.206

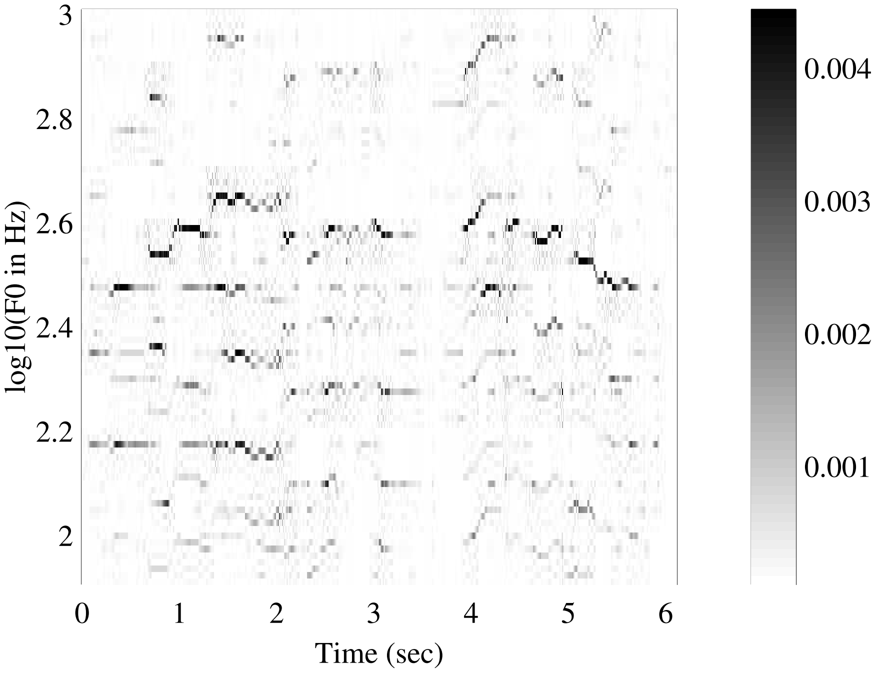


Figure 3.13: Likelihood scores of F0 candidates for the median-accuracy excerpt in mir1k.

Table 3.4: Performance comparison with several state-of-the-art approaches by overall accuracy. The approaches are sorted by the `mirex05` accuracy in descending order. Some approaches were evaluated before some data sets were created; therefore, results are not available for some approach-data pairs. M5 = `mirex05`; M8 = `mirex08`; M9 = `mirex09`; M9+ = `mirex09+`; M9- = `mirex09-`.

Work	**M5**	**M8**	**M9**	**M9+**	**M9-**
[4]	0.770	0.807	0.682	0.784	0.517
[27]	0.734	0.844	0.781	0.852	0.611
[13]	0.718	0.768	0.762	0.834	0.629
This Dissertation	0.703	0.710	0.739	0.817	0.594
[17]	0.693	0.710	0.739	0.827	0.536
[29]	0.673	–	–	–	–
[5]	0.667	0.750	–	–	–
[6]	0.650	–	–	–	–
[30]	0.628	0.715	0.742	0.817	0.623

Table 3.5: Results (*p*-values) of two-sided Wilcoxon signed-rank tests for the performance comparison with several state-of-the-art approaches with respect to the `mirex09` variants. For lack of per-track results in MIREX 2009, the *p*-values cannot be calculated for [4].

Data Set	**[27]**	**[13]**	**[30]**	**[17]**
`mirex09`	0.000	0.000	0.145	0.760
`mirex09+`	0.000	0.000	0.049	0.001
`mirex09-`	0.001	0.000	0.000	0.000

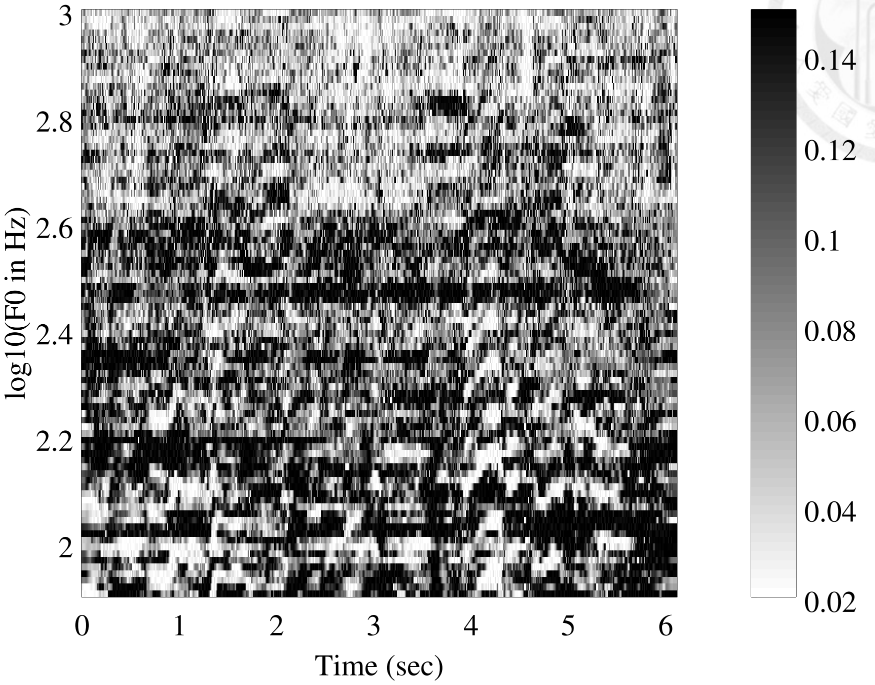


Figure 3.14: Timbral fitness scores of F0 candidates for the median-accuracy excerpt in `mir1k`.

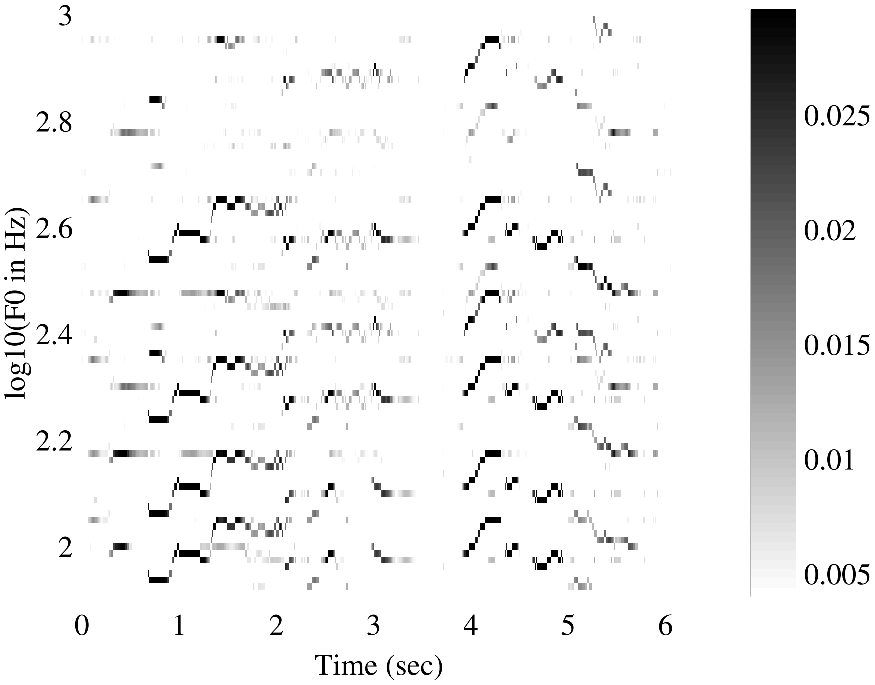


Figure 3.15: Loudness values of F0 candidates for the median-accuracy excerpt in `mir1k`.

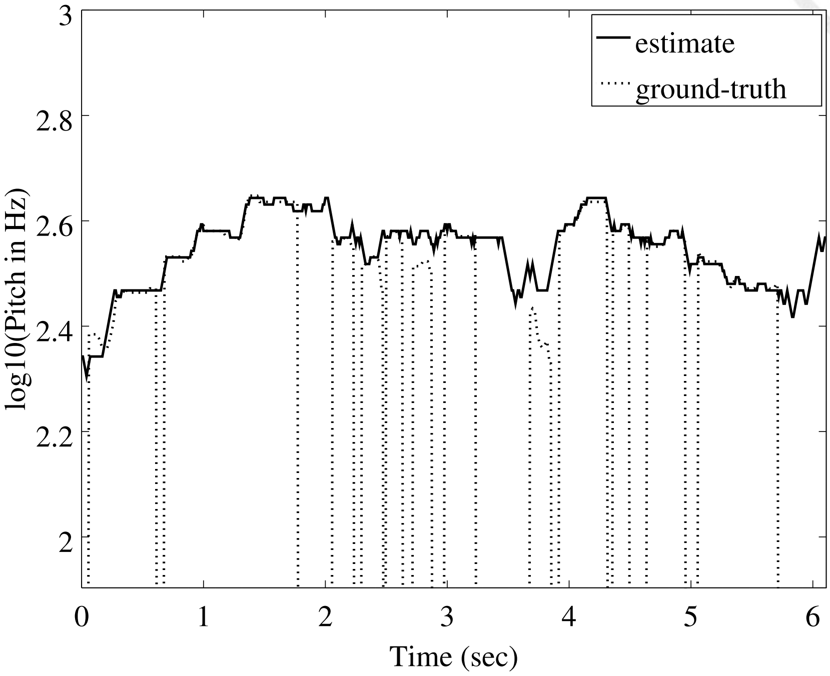


Figure 3.16: Vocal F0 sequence estimated from the median-accuracy excerpt in `mir1k`. Dotted vertical lines mark the boundaries of ground-truth vocal rests represented by 0-hertz F0 values.

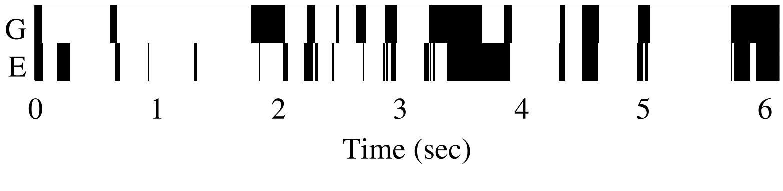


Figure 3.17: Result of vocal rest detection for the median-accuracy excerpt in `mir1k`. Vocal rests are depicted in black, both for the ground truth (G) and for the estimate (E).

Table 3.6: Results of experiments conducted on `medley2` with an algorithmic feature removed from the proposed approach. OA = Overall Accuracy; RPA = Raw Pitch Accuracy; RCA = Raw Chroma Accuracy; VDR = Voicing Detection Rate; VFAR = Voicing False Alarm Rate.

Removed Feature	**OA**	**RPA**	**RCA**	**VDR**	**VFAR**
None	0.690	0.728	0.756	0.742	0.206
Timbre	0.560	0.402	0.479	0.591	0.138
New Likelihood	0.585	0.581	0.628	0.666	0.236
LF Model	0.636	0.603	0.651	0.697	0.190
Gender	0.694	0.735	0.761	0.741	0.207
Genre	0.705	0.757	0.770	0.739	0.202
Voice Type	0.668	0.680	0.722	0.735	0.203
Vowel Type	0.701	0.746	0.765	0.738	0.203
Singer	0.667	0.688	0.716	0.730	0.204
Diversity	0.576	0.459	0.518	0.659	0.185
Loudness	0.233	0.037	0.064	0.467	0.457
Prior Model	0.578	0.530	0.614	0.802	0.288

Chapter 4

Lyrics Alignment Based on a Vowel Likelihood Model

This chapter is outlined as follows. An overview is presented in Section 4.1 for the complete alignment system. Section 4.2 is dedicated to an acoustic-phonetic model of vowel likelihood, which defines how likelihood scores of vowel candidates are evaluated for each analysis time position. Described in Section 4.3 is an algorithm for lyrics alignment based on the vowel likelihood model. Experiments and results are documented in Section 4.4.

4.1 System Overview

A block diagram of the complete system is shown in Fig. 4.1. With the lyrics regarded as an alternating sequence of syllables and vocal rests, we construct an alignment of lyrics as a *syllabic position* sequence, which specifies a syllabic position (initial rest, first syllable, rest following first syllable, second syllable, rest following second syllable, etc.) for each and every analysis time position in the accompanied singing signal. This sequence is generated by an estimation procedure that 1) extracts partial amplitudes from the audio spectrum according to an F0 estimate produced by the vocal melody extractor presented in Chapter 3, 2) evaluates likelihood scores for syllabic position candidates with the proposed vowel likelihood model, and 3) uses sequential constraints among the lyric syllables to select a syllabic position for each time position. Note that whereas vocal F0 is undefined within vocal rests, the melody extractor gives an F0 estimate for every time position.

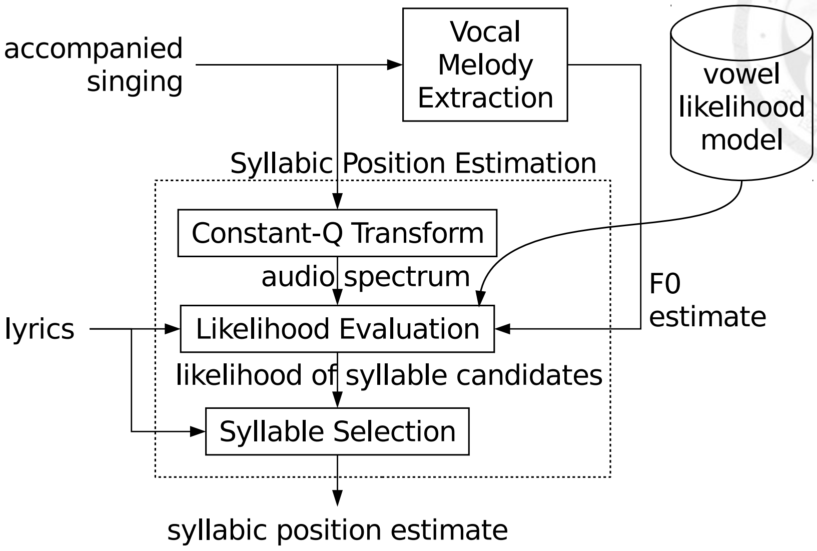


Figure 4.1: Block diagram of the complete system.

4.2 Acoustic-Phonetic Model of Vowel Likelihood

In this section, we consider the likelihood function[1] of *vowel type* given the input signal at an analysis time position, for which a computational model is presented. In this model, the observation is represented by a sequence of partial amplitudes extracted from the spectrum of input signal according to a vocal F0 estimate. These partial amplitudes exhibit a specific timbral quality that, when checked against timbre examples of a vowel hypothesis, indicates how likely the vowel hypothesis gives the true vowel type. In addition, the overall loudness of the partials is also taken into account in likelihood evaluation so that vocal rests can be distinguished from vowels. Note that many approaches to automatic speech recognition or lyrics alignment are based on phoneme likelihood, i.e., likelihood function of one of the vowels and consonants. In this work, vowel likelihood is sufficient because I let each lyric syllable be approximated by its nucleus, a decision based on the fact that lyric syllables are typically prolonged in musical notes by extending their nuclei instead of their consonant margins, leaving all lyric consonants relatively short in singing. As with a typical phoneme likelihood model where a phoneme can be a short pause, here the vowel type can be either a specific vowel or a vocal rest.

Let the N-sample windowed input signal centered at time m be denoted by random

[1] See Section 3.2 for the definition of likelihood function.

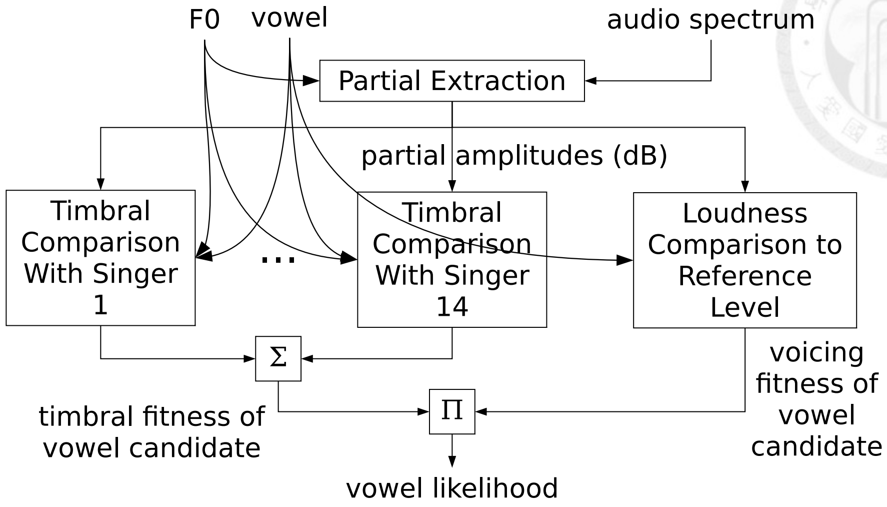

F0
vowel
audio spectrum
Partial Extraction
partial amplitudes (dB)
Timbral Comparison With Singer 1
...
Timbral Comparison With Singer 14
Loudness Comparison to Reference Level
Σ
Π
timbral fitness of vowel candidate
voicing fitness of vowel candidate
vowel likelihood

Figure 4.2: Architecture of the vowel likelihood model.

N-vector $\mathbf{z}_m$, the corresponding vocal F0 be denoted by an integer random variable $w_m \in \{46, 47, ..., 133\}$ that measures the distance of the F0 from a reference low frequency (21.205 hertz) in quarter tones, and the corresponding vowel type be denoted by a discrete random variable v_m. We model the likelihood function of v_m with a *timbral fitness measure* $F_h(\cdot)$ and a *voicing fitness measure* $F_e(\cdot)$:

$$\mathscr{L}_{v_m|\mathbf{z}_m,w_m}(v|\mathbf{z},w) \propto F_h(\mathbf{z},w,v) \cdot F_e(\mathbf{z},w,v). \tag{4.1}$$

This frame-wise model will give emission probabilities in a hidden Markov model [26]. Since Viterbi search is invariant to an arbitrary scaling factor applied to all the emission probabilities, here I omit any scaling constant that would be necessary for the likelihood model to conform with the definition of probability density function. Architecture of this model is represented in Fig. 4.2.

4.2.1 Timbral Fitness Measure

To define the timbral fitness of vowel hypothesis v with respect to input signal $\mathbf{z}$ and vocal F0 estimate w, we calculate from $\mathbf{z}$ a constant-Q magnitude spectrum with quarter-tone-spaced frequency bins [2]. In an effort to simulate the dependency of human loudness perception on frequency, we correct the magnitude spectrum according to trends in the 40-phon equal-loudness contour (ELC) [15]:

$$A_f^z = |[\text{CQT}\{\mathbf{z}\}]_f| \cdot 10^{(40-\kappa_f)/20},\ f \in I, \tag{4.2}$$

where CQT$\{\cdot\}$ denotes the constant-Q transform, I denotes the set of frequency bins, f denotes the frequency index, and κ_f denotes the 40-phon ELC. Among all the frequency bins, we focus on a set of N_h partial frequencies[2] constructed from w:

$$I_w = \{\mathrm{round}(w + 24\log_2 l)\}_{l=1}^{N_h} \subset I. \tag{4.3}$$

The timbral quality exhibited by these partials is compared to N_s timbre examples of vowel v to give a timbral fitness score:

$$F_h(\mathbf{z},w,v) = \frac{1}{N_s}\sum_{i=1}^{N_s}\exp\left\{-\frac{c_h}{N_h}\sum_{f\in I_w}\Delta_{i,f}^2\right\}, \tag{4.4}$$

$$\Delta_{i,f} = 20\log_{10}\frac{A_f^z}{\tilde{A}_f^{(v,i,w)}}, \tag{4.5}$$

where c_h is a nonnegative parameter that scales the effect that magnitude deviation has on the likelihood score, and $\tilde{A}_f^{(v,i,w)}$ denotes a scaled spectrum of the ith timbre example for vowel v whose mean magnitude over the partials is aligned with the input signal to factor out volume level of the input signal:

$$\log_{10}\frac{\tilde{A}_f^{(v,i,w)}}{A_f^{(v,i,w)}} = \frac{1}{N_h}\sum_{f'\in I_w}\log_{10}\frac{A_{f'}^z}{A_{f'}^{(v,i,w)}}. \tag{4.6}$$

Here $A_f^{(v,i,w)}$ denotes the ELC-corrected constant-Q magnitude spectrum of the ith timbre example for vowel v that has been F0-modified to w with the procedure presented in Chapter 2. The mechanism for comparing the observed timbre to an example is depicted in Fig. 4.3.

In the case where hypothesis v is a vocal rest, we need a timbral fitness score that measures how well the observed partial amplitudes fit our expectation for a vocal rest. At a vocal rest, what the melody extractor gives in addition to an expected symbol for the rest, is a frequency w that does not typically match any vocal or instrumental F0. Such an F0 estimate often results in partial frequencies at non-sinusoidal, low-magnitude positions in the spectrum. In consequence, we assume that any timbre is possible for these partials:

$$F_h(\mathbf{z},w,v) = 1 \tag{4.7}$$

whenever v is a vocal rest. Obviously, vocal rests cannot be recognized with the timbral fitness score alone; rather, detection of vocal rests relies on the voicing fitness measure.

[2] For values of numerical parameters, see Table 4.2.

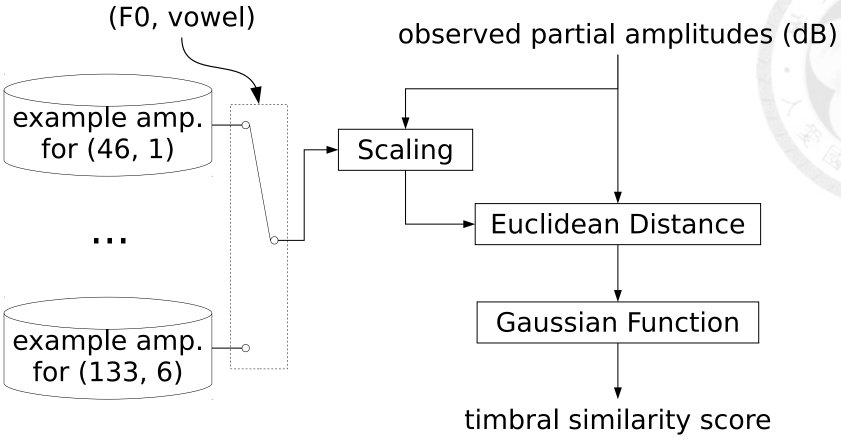


Figure 4.3: Timbral comparison to a vowel timbre example.

Vowel Timbre Examples

I collected timbre examples for 6 vowel types: /i/, /ɛ/, /ɑ/, /ɔ/, /u/, and /ə/. Since apparently many vowels are not included in this set, we are in fact dividing all possible vowels into the 6 categories with a set of rules for categorical mapping. For my experiments on songs with English lyrics, I map each lyric diphthong to its first component vowel, each lyric /æ/ to /ɛ/, and each lyric /ʌ/ to /ə/. Here the diphthong mapping is based on the singing practice of gliding toward the end of note, ignoring the typically short ending component vowel. Similarly, for Chinese lyrics, I map each lyric falling diphthong to its first component vowel, each lyric /y/ to /i/, and each lyric /ɨ/ to /ə/.

For each of the 6 vowel types, I construct a collection of N_s timbre examples ($N_s = 14$) from 14 recordings of about 1 minute each. The 14 recordings (from 14 singers and shared among the 6 vowel types) 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 from the MIR-1K data set [12]. From each recording, a time position is selected such that the person sings the vowel type at the time position, and that no partial of the vocal suffers significant interference from the accompaniment. For each of the 14 time positions, I conduct sinusoidal analysis, manually identify the fundamental of the vocal, and extract the frequencies and amplitudes of the vocal partials up to 5 kHz as one of the timbre examples for the vowel type.

The 14 types of singing voice are tenor, soprano, baritone, mezzo-soprano, pop high male voice, pop high female voice, pop low male voice, pop low female voice, pop nasal male voice, pop nasal female voice, non-professional high male voice, non-professional high female voice, non-professional low male voice, and non-professional low female voice. For further details on these timbre examples, see Section 3.2.1.

4.2.2 Voicing Fitness Measure

In an attempt to distinguish a vowel from a vocal rest, we estimate the loudness of singing voice from input signal $\mathbf{z}$ and vocal F0 estimate w. A high vocal loudness would indicate a vowel, while a low vocal loudness would indicate a vocal rest. To estimate the loudness, we conduct sinusoid tracking on the spectrum $\{A_f^z\}_{f\in I}$, thereby generating a binary partition of the set of frequency bins I. A frequency bin where a sinusoid is detected is called a *sinusoidal frequency bin*, whereas one where no sinusoid is detected is called a *non-sinusoidal frequency bin*. With this partition, we can calculate a *sinusoidal power spectrum* from $\mathbf{z}$:

$$P_f^z = (A_f^z)^2 \cdot S_z(f),\ f \in I, \tag{4.8}$$

where $S_z(\cdot)$ maps sinusoidal frequency bins to unity and non-sinusoidal frequency bins to zero. Among all the frequency bins, we focus on a set of N_e partial frequencies constructed from w:

$$J_w = \{\mathrm{round}(w + 24\log_2 l)\}_{l=1}^{N_e} \subset I. \tag{4.9}$$

The loudness is evaluated by summing sinusoidal power over these partials:

$$\Lambda_e(\mathbf{z}, w) = \sum_{f\in J_w} P_f^z. \tag{4.10}$$

Overestimation of vocal loudness can occur when a vocal F0 estimate coincides with an instrumental F0 at a vocal rest. This is alleviated by the sinusoid tracking procedure, where many instrumental sinusoid contours are removed for their almost constant frequency (see Section 3.2.2).

A *voicing fitness score* is calculated to measure the fitness of hypothesis v in terms of presence of singing voice. This score can now be defined by comparing the vocal loudness estimate $\Lambda_e(\mathbf{z}, w)$ to a reference loudness level chosen for the hypothesis: If v is a vowel,

$$F_e(\mathbf{z}, w, v) = \exp\left\{-\frac{(\lambda_e - \lambda_s)^2}{2\sigma_s^2}\right\}; \tag{4.11}$$

otherwise,

$$F_e(\mathbf{z}, w, v) = \exp\left\{-\frac{(\lambda_e - \lambda_r)^2}{2\sigma_r^2}\right\}. \tag{4.12}$$

Here, λ_s (dB) denotes an adaptive loudness level for a vowel, which is calculated by median-filtering (with filter length specified by N_m) the vocal loudness estimate $\Lambda_e(\mathbf{z}, w)$ across all analysis time positions and taking the maximum of the filter output. By median

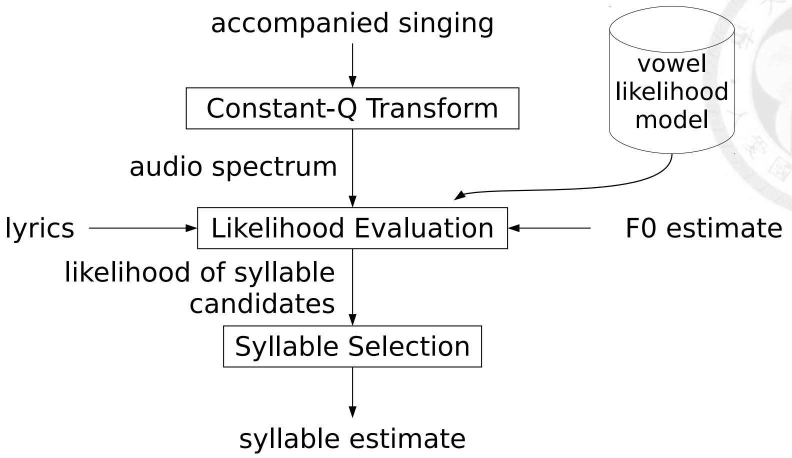


Figure 4.4: Block diagram for syllabic position estimation.

filtering, the procedure rejects spikes in the loudness variations. With the reference level for a vowel defined, that for a vocal rest, denoted by λ_r, is then determined by subtracting a dynamic range ρ_d (dB) from λ_s. In (4.11) and (4.12), λ_e (dB) denotes a smoothed, compressed vocal loudness estimate. The initial loudness estimate $\Lambda_e(\mathbf{z}, w)$ is compressed with a lower limit of λ_r and an upper limit of λ_s. After that, the loudness is median-filtered across all analysis time positions (with filter length N_v) to smooth out any isolated, brief loudness dips resulting from F0 estimation errors. The effect loudness deviation has on the likelihood score is scaled by nonnegative parameters σ_s and σ_r.

4.3 Syllabic Position Estimation

As the first step in aligning lyric syllables with audio, the lyrics text is processed to generate a list of syllabic position candidates. Estimation of syllabic position is performed on a grid of 100 time positions per second. At each time position, likelihood scores are evaluated for all the candidates with the proposed vowel likelihood model. Last, a syllabic position is selected for each time position according to the likelihood scores and the sequential relation among syllabic positions. This is depicted in Fig. 4.4.

4.3.1 Lyrics Preprocessing

To convert the sequence of words in each lyric line into a sequence of vowels, each word is looked up in a digital pronunciation dictionary, thereby converting the word sequence into a sequence of phonemes. Next, consonants are removed from the phoneme sequence, and each vowel is mapped to one of the 6 vowel types, /i/, /ɛ/, /ɑ/, /ɔ/, /u/, and /ə/. When multiple pronunciations are available in the dictionary for a syllable, all vowel variants are kept for the syllable. My implementation of the lyrics preprocessor can process North American English and Standard Chinese. The English dictionary used is the CMU Pronouncing Dictionary. For Chinese lyrics, a 72,647-word dictionary is used for maximum-matching word segmentation and pronunciation lookup.

To generate a complete set of syllabic position candidates for lyrics alignment, we create a unique candidate for each element in the vowel sequence. For a syllable with vowel variants, a unique candidate is created for each variant, with multiple variant candidates sharing the same syllabic position. Unique candidates are also declared for vocal rests between every two neighboring elements in the vowel sequence, before the first vowel, and after the last vowel. For instance, for a lyrics text file consisting of 50 syllables, at least 101 syllabic position candidates should be created. The resulting set of candidates is denoted by $C = \{1, 2, ..., N_c\}$, where each candidate is represented by a positive integer and N_c denotes the number of candidates. Each syllabic position $s \in C$ is associated with a syllable index n_s, which locates s as the n_sth syllable in the lyrics or the vocal rest following the n_sth syllable. A zero syllable index $n_s = 0$ indicates that s is the initial vocal rest. Each syllabic position $s \in C$ is also associated with a vowel type $v^{(s)} \in \{0, 1, ..., 6\}$. A zero vowel type $v^{(s)} = 0$ indicates that s is a vocal rest.

4.3.2 Likelihood Evaluation

Selection of a syllabic position for an analysis time position is based on likelihood evaluation performed over all syllabic position candidates for the same time position. The proposed vowel likelihood model yields 7 likelihood scores respectively for the 7 vowel types: /i/, /ɛ/, /ɑ/, /ɔ/, /u/, /ə/, and vocal rest. Each syllabic position candidate is assigned one of the 7 scores according to its vowel type. Notice that with this likelihood evaluation scheme, many candidates can share the same likelihood score because they have the same vowel type, a fact that reveals the importance of selecting syllables jointly over all analysis time positions with sequential relation taken into account.

In the implementation, likelihood scores for the first and last analysis time positions are defined in a way that directly ensures proper syllabic positions for these two time positions. Since the audio should start with the initial vocal rest or the first syllable, all other candidates are assigned a zero score for the first time position. Similarly, all candidates are assigned a zero score for the last time position, except for the final vocal rest and the last syllable.

4.3.3 Syllable Selection

To estimate the evolution of syllabic position in the singing voice, we let the joint probability distribution for the unobserved and observed signals, i.e., the syllabic position evolution and the accompanied singing audio, be represented by a hidden Markov model (HMM) [26]. In this HMM, likelihood scores of syllabic position candidates are defined as in Section 4.3.2, and state transition probabilities among the candidates encode the sequential relation among syllabic positions. With this probabilistic model, the estimation can be achieved by maximizing the posterior probability of state sequence with the Viterbi algorithm.

The evolution of syllabic position is modeled with a Markov chain $\{s_m\}_{m=1}^{M}$, where M denotes the number of analysis time positions:

$$\begin{aligned} P(s_1,\ldots,s_M) &= P(s_1)P(s_2|s_1)P(s_3|s_1,s_2)\cdots \\ &\quad P(s_m|s_1,\ldots,s_{m-1})\cdots \\ &\quad P(s_M|s_1,\ldots,s_{M-1}) \qquad (4.13) \\ &= P(s_1)\prod_{m=2}^{M} P(s_m|s_{m-1}). \qquad (4.14) \end{aligned}$$

The equality in (4.14) results from the Markovianity that given the previous syllabic position s_{m-1}, the current syllabic position s_m is independent of all the earlier syllabic positions $s_{m-2}, s_{m-3}, \ldots, s_1$. A uniform initial state distribution can be assumed here because a proper syllabic position at the first time position has been guaranteed by likelihood scores:

$$P(s_1 = s) = \frac{1}{N_c},\ \forall s \in C. \qquad (4.15)$$

The conditional probability $P(s_m|s_{m-1})$ considers a particular syllabic position given for the previous time position, and defines the probabilities of all possible syllabic positions for the current time position. We derive specific values for these probabilities from the fact that syllabic positions must be visited in the same order as they appear in the lyrics. Since only a continuation of the same syllabic position or a transition to a succeeding syllabic position is allowed, we assign zero probabilities to all the other syllabic positions. First, consider the case where s_{m-1} is a vowel, i.e., $v^{(s_{m-1})} \neq 0$. In this case, a probability of 0.5 is assigned to the continuation, and the remaining probability of 0.5 is distributed among several eligible successors. If $n_{s_{m-1}}$ points to the end of a lyric line, the only eligible successor is the vocal rest immediately following s_{m-1}, which is expected to last relatively long; otherwise, we make the vocal rest optional, with the rest taking 0.25 and all vowel variants for the next syllable uniformly sharing the remaining 0.25. Second,

Table 4.1: Summaries for the data sets.

Data Set	**#Excerpts**	**Excerpt Length (s)**
`adc2004`	9	~20
`labrosa`	9	~30
`poly_100`	100	9–49
`slam`	130	9–52

consider the case where s_{m-1} is a vocal rest, i.e., $v^{(s_{m-1})} = 0$. Again, if this is at a line break, we use a parameter $P_c < 0.5$ to control the tendency for this rest to last longer than vocal rests within a lyric line, assigning a probability of $1 - P_c$ to the continuation; otherwise, the continuation has probability 0.5. The remaining probability is again shared uniformly among vowel variants of the $(n_{s_{m-1}} + 1)$th syllable.

4.4 Experiments

4.4.1 Data Sets

In Sections 4.2 and 4.3, the vowel likelihood model and the procedure for syllabic position estimation have been defined with a set of numerical parameters, whose values need to be determined empirically. To select appropriate values for these parameters, I considered their effect on alignment performace by carrying out alignment experiments on 2 *development sets*: `adc2004` and `labrosa`. Data set `adc2004` is adapted from the data used for audio melody extraction in the ISMIR 2004 Audio Description Contest (ADC 2004). ADC 2004 in its entirety consists of 20 audio recordings with melody annotations, among which 8 recordings are fully instrumental, and the other 12 are accompanied singing. For lyrics alignment, only recordings with words are included in `adc2004`, including 4 pop song excerpts, 2 song excerpts with synthesized vocal, and 3 opera excerpts[3]. Melody annotations are replaced with syllable onset and offset annotations and vowel-sequence transcriptions, the latter serving as a substitute for lyrics lacking in this data set. I intend to produce syllable-level annotations to make full use of this small data set. Data set `labrosa` is similarly adapted from a data set created for polyphonic melody extraction by

[3] Two excluded vocal excerpts (`daisy3` and `daisy4`) do not have words. Another vocal excerpt (`opera_male3`) was excluded for its excessive difficulty, featuring an exaggerated loudness contrast in singing.

LabROSA, Columbia University, consisting of English popular song excerpts. Numbers of excerpts and lengths are listed in Table 4.1 for these data sets.

To assess the generalization performance of parameter optimization, I used 2 test sets without any overlap with the development sets: `poly_100` and `slam`. Data set `poly_100` was created by Mesaros and Virtanen [24], composed of excerpts extracted from 17 English popular songs (8 female artists and 9 male artists), their lyrics, and lyric-line onset and offset annotations. A total of 4–8 excerpts are extracted from each song, with each excerpt capturing a complete structural section in a song, such as a chorus and a verse. From a collection of 20 Chinese popular songs (10 female artists and 10 male artists), I derived a similar data set `slam` with 3–9 excerpts extracted from each song. For this data set, I produced onset and offset annotations for a smaller unit of alignment, a lyric phrase. Chinese lyric phrases are separated by spaces or line breaks, and each lyric line consists of one or more lyric phrases. Numbers of excerpts and lengths are listed in Table 4.1 for these data sets.

4.4.2 Performance Measures

In the experiments documented here, a textual unit for alignment is specified for each alignment task, which can be a lyric syllable, a lyric word, a lyric phrase, or a lyric line. Onset and offset time estimates are extracted for the sequence of aligned units from the evolution of syllabic position given by the tested system. The extraction proceeds by finding onsets and offsets for syllables before identifying specific syllabic onsets and offsets that correspond to boundaries of the aligned units.

To measure the overall performance of a tested system in aligning lyrics of an excerpt with its audio, we calculate the *average absolute alignment error* and the *average normalized alignment error*. The average absolute alignment error is computed by taking the average of distances (in seconds) of all the onset and offset estimates from ground-truth annotations. When each of these distances is normalized by the true duration of the aligned unit, the normalized distance represents an alignment error as a proportion in the true duration. An upper limit of unity can be further applied to the normalized distance to avoid discriminating among any excessive distances. We calculate the average of such normalized, limited distances to give the average normalized alignment error.

Alignment of a textual unit with audio attempts to determine the position and duration of the unit as measured on the time axis of audio. To focus performance measurement on one of these two factors, we calculate the *average normalized position error* and the *average relative duration error*. With the position of a unit defined by the midpoint between onset and offset, the average normalized position error is computed by taking the average of, again, normalized and limited distances of all the position estimates from ground-truth values. The same calculation applied to the duration gives the average relative duration error.

4.4.3 Results on the Development Sets

The search for optimal parameter values was performed on the development sets in 2 stages. In the first stage, errors in syllable alignments, as indicated by particular syllables aligned with excessive absolute errors, were analyzed to isolate particular parameter values that were responsible for the errors. Each error analysis was expected to suggest an adjustment in parameter value that would eliminate the analyzed error. Parameters were repeatedly and selectively adjusted according to such error analyses until the remaining errors could no longer be explained by an inadequate parameter value. In the second stage, the average absolute alignment error was further minimized by coordinate descent in 3 cycles, with respect to a roughly 10-point sampling of each one-dimensional parameter space. In the first cycle, one-parameter line searches were performed, each as a one-dimensional grid search, for all the parameters in the order shown in Table 4.2. In each line search, an average absolute alignment error is evaluated on the development sets for each sample point. The sample point that gives the lowest error is assigned to the parameter as an update, which may or may not equal the old parameter value. The average absolute alignment error is averaged over all excerpts in each development set before being averaged between the 2 development sets without weighting. In the second cycle, the same one-parameter line searches were repeated, with more parameters receiving a stationary update. Finally in the third cycle with the same line searches, all the initial parameter values, as given by the second cycle, were found to be optimal with respect to the sampling of parameter space.

The results of parameter optimization are presented in Table 4.2, where optimal settings are compared to extreme settings with average absolute alignment errors given by syllable alignment experiments conducted on the development sets. Each row contains 3 parameter settings derived from the optimal setting by assigning 3 different values to the designated parameter: The first value is the smallest sample point, the second value is the optimal value, and the third value is the largest sample point. As shown in Fig. 4.5, the limited effect that changes in P_c have on the error, suggests that the assumed relative lengthiness of vocal rests at line breaks may not be sufficiently consistent in practice to deserve a special transition probability assignment.

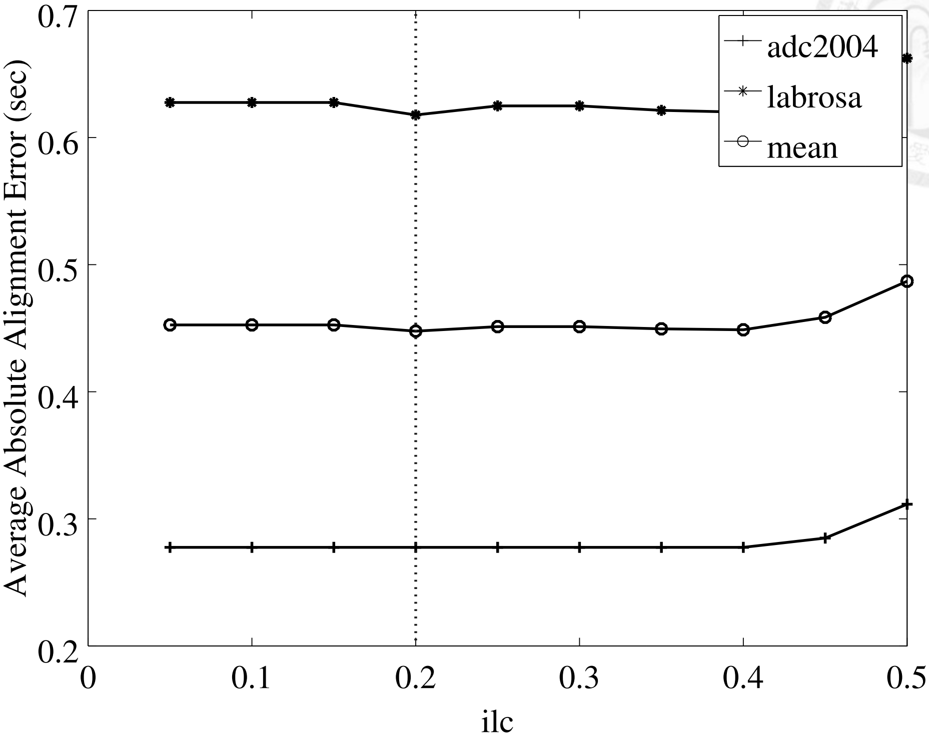


Figure 4.5: Average absolute alignment error evaluated on the development sets for sample points of P_c. The vertical dotted line marks the optimal parameter value.

4.4.4 Results on the Test Sets

Performance of the proposed approach is evaluated on the test data sets, as presented in Table 4.3. The absolute alignment error is around one second for both sets. Since many Chinese lyric lines in `slam` contain only one phrase, line-level alignment with `poly_100` and phrase-level alignment with `slam` would not be expected to give dissimilar absolute errors. The normalized alignment error is below 0.3 for both sets. For an onset or offset, a normalized alignment error of 0.3 means an absolute error equal to 30% the true duration of aligned unit, but does not indicate a specific error in estimation of position or duration: If errors of 0.3 are in the same direction for onset and offset, the position error will be 0.3, and the duration error will be zero; otherwise, the position error will be zero, and the duration error will be 0.6. Results in normalized position and relative duration errors indicate that errors associated with these two types of estimation are in fact below or close to 0.3 on average for both data sets.

As shown in Table 4.4, my approach gave an average absolute alignment error slightly lower than those given by variant approaches of Mesaros and Virtanen [24] in experiments conducted on `poly_100`. Their approach to lyrics alignment linearly adapts several (3, 8,

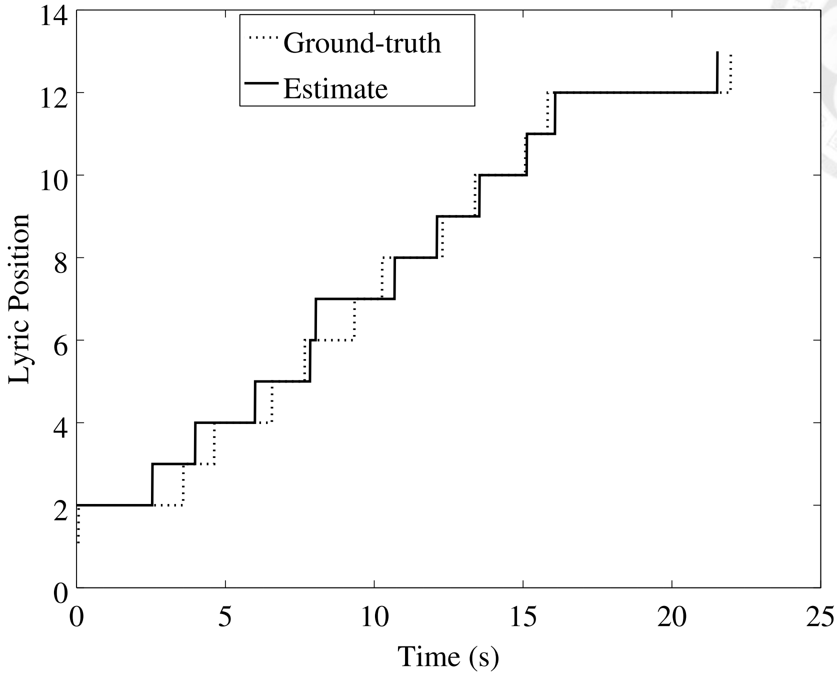


Figure 4.6: Lyric position sequence estimated from the excerpt with the 50th lowest average normalized alignment error in `poly_100`. The initial vocal rest is represented by position 1, the first line is represented by position 2, the vocal rest following the first line is represented by position 3, and so on. In other words, odd positions represent vocal rests at line breaks, whereas even positions represent lyric lines.

or 22) categories of speech phoneme likelihood models to a small amount of singing data, with all the models in a category sharing the same linear transformation. This comparison shows that, if my approach does not significantly outperform their approach, these two approaches could be considered comparable in performance.

4.4.5 Example

To gain insight into my test results, consider the excerpt in `poly_100` for which the resulting average normalized alignment error was 0.197, which is the 50th lowest error among all the 100 error values and serves as a median item that could represent the entire data set. As shown in Fig. 4.6, the estimated evolution of lyric position adequately matched the true evolution at all 6 lyric lines except the third line, where the transition to vocal rest occurred prematurely. Inspection of its spectrogram shown in Fig. 4.7 reveals that the third line is sung (from $t = 7.7$ s to $t = 9.3$ s) with a relatively low loudness, which led my algorithm to treat the segment of signal as a vocal rest. Many instances of low

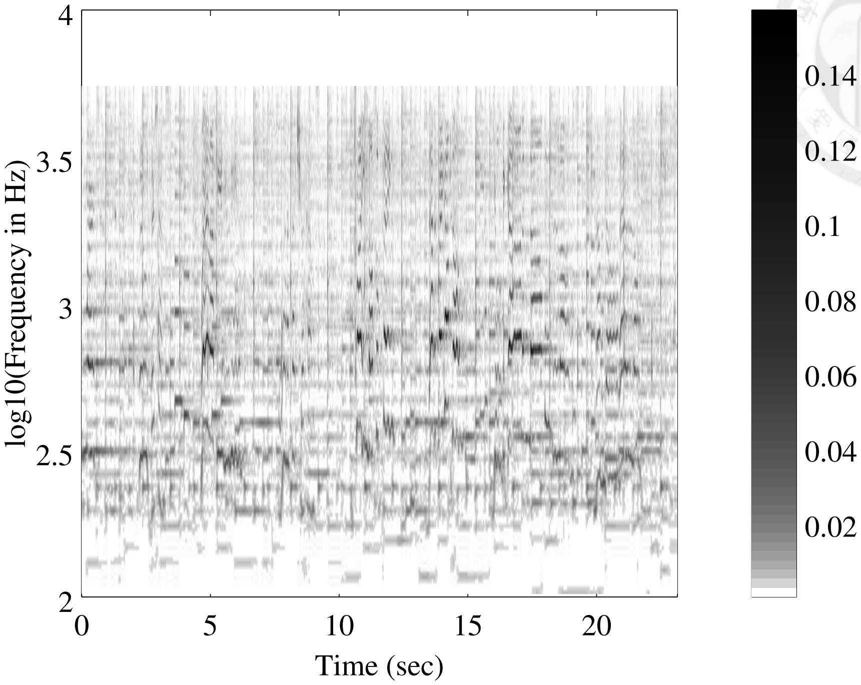


Figure 4.7: Spectrogram of the excerpt with the 50th lowest average normalized alignment error in poly_100.

estimated vocal loudness actually indicate vocal rests that separate adjacent lyric syllables or lyric lines; as a result, my approach is inevitably confused by soft singing in some circumstances.

4.4.6 Results of Controlled Experiments

Timbre Model

In order to assess the effect of timbral fitness measure, I repeated the poly_100 experiments with vowel likelihood defined by the voicing fitness measure alone:

$$\mathscr{L}_{v_m|\mathbf{z}_m,w_m}(v|\mathbf{z},w) \propto F_e(\mathbf{z},w,v). \tag{4.16}$$

This timbre-insensitive likelihood model gave results listed in Table 4.5 on row "Timbre." For lack of timbre modeling, the normalized alignment error grows by 0.07, which confirms the effectiveness of timbral fitness measure in identifying lyric vowels. Moreover, this result reveals that lyrics alignment based solely on estimating the loudness or voicing of singing voice, not performing any phonemic discrimination, could adequately estimate line-level evolution of lyric position. As an example of timbre-blind alignment, an evolution of lyric position is displayed in Fig. 4.8, which is estimated without the timbre model

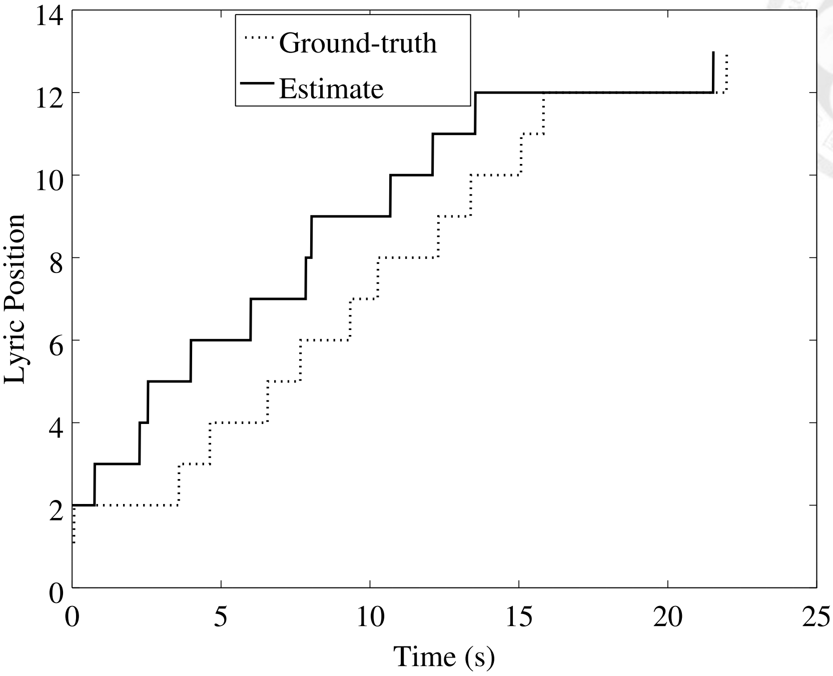


Figure 4.8: Lyric position sequence estimated without the timbre model, from the excerpt with the 50th lowest average normalized alignment error in poly_100. For the definition of lyric position, see the caption of Fig. 4.6.

from the same excerpt considered in Section 4.4.5. In this example, the impact of timbre model removal is far above the average, giving a normalized alignment error of 0.772 and a clear deviation from the true evolution in Fig. 4.8.

Transformed Liljencrants-Fant Model

In the F0 modification of vowel timbre examples, we implement an invariance in the glottal pulse shape, which is represented by the transformed Liljencrants-Fant model with a parameter (R_d) value estimated from the original example. To evaluate the effect of this specific model of glottal excitation, I repeated the poly_100 experiments with the (radiated) glottal excitation modeled in the conventional way by a fixed spectrum envelope [7]:

$$U_R(f^h) = \frac{f^h/100}{1+(f^h/100)^2}. \tag{4.17}$$

Note that an F0-invariant pulse shape implies F0-invariant amplitude ratios among the partials and thus a spectrum envelope that stretches along the frequency axis as the F0 increases. The repeated experiments gave results listed in Table 4.5 on row “LF Model.” The difference of 0.027 in average normalized alignment error confirms the advantage of transformed Liljencrants-Fant model in modeling vowel timbre.

Diversity Among Vowel Timbre Examples

In Section 4.2.1, vowel timbre examples are collected with diversity in gender, genre, and voice type. In an effort to evaluate the separate effect of these diversity factors, I repeated the `poly_100` experiments 3 times, each time using a timbre example subset where one of these diversity factors has been eliminated. The 7-singer male-only (7-example) subset gave results listed in Table 4.5 on row "Gender." Labeled "Genre" are results for the 4-singer opera-only (4-example) subset. The row labeled "Voice Type" is for the 6-example high-voice-only subset, which consists of the tenor, the soprano, and the other 4 high-voice singers. Removal of each diversity factor is found to raise the normalized alignment error by an amount between 0.015 and 0.026, which exhibits the benefit of including these factors in timbre example collection. For the purpose of evaluating the composite effect of diversity factors, I repeated the `poly_100` experiments with only one tenor example for each vowel type, giving results listed in Table 4.5 on row "Diversity." This shows not only that complete lack of diversity results in an increase of 0.046 in normalized alignment error, but that even with this single-example timbral comparison, a reduction in error still exists (0.023) from the timbre-blind approach.

Voicing Model

To see the effect of voicing fitness measure, I repeated the `poly_100` experiments with vowel likelihood defined by the timbral fitness measure alone:

$$\mathscr{L}_{v_m|\mathbf{z}_m,w_m}(v|\mathbf{z},w) \propto F_h(\mathbf{z},w,v). \tag{4.18}$$

This loudness-insensitive vowel likelihood model gave results listed in Table 4.5 on row "Voicing," where lack of a voicing model brought about a dramatic rise in average normalized alignment error. My parameter optimization efforts made on the development sets determined that, in order to minimize alignment errors, the alignment should predominantly depend on voicing clues provided by the variations in estimated vocal loudness, which is made evident by this result.

Table 4.2: Optimal and extreme parameter settings and the resulting average absolute alignment errors evaluated on the development sets. In each ordered pair, the first entry gives a parameter value and the second entry gives an error in seconds. The reader is referred to Sections 4.2 and 4.3 for the units of parameter values. Some of the listed parameters are defined in Section 3.2.2 for sinusoid tracking.

Param.	**Smallest Value**	**Optimal Value**	**Largest Value**
P_c	(0.05, 0.453)	(0.2, 0.448)	(0.5, 0.487)
ρ_d	(9.0, 2.492)	(27.0, 0.448)	(36.0, 0.965)
σ_s	(0.1, 2.304)	(0.4, 0.448)	(1.0, 1.142)
σ_r	(0.1, 1.384)	(0.4, 0.448)	(1.0, 1.329)
c_h	(0.003, 0.522)	(0.023, 0.448)	(0.043, 0.468)
N_h	(5, 0.660)	(10, 0.448)	(15, 0.497)
N_m	(1, 0.931)	(35, 0.448)	(280, 1.065)
N_v	(10, 0.699)	(22, 0.448)	(30, 0.546)
N_a	(2, 1.656)	(7, 0.448)	(10, 0.475)
θ_p	(0.0, 0.795)	(12.0, 0.448)	(20.0, 1.196)
θ_r	(0.1, 0.833)	(0.9, 0.448)	(2.1, 0.945)
θ_j	(0.5, 0.484)	(1.0, 0.448)	(1.5, 0.484)
θ_g	(0.002, 0.448)	(0.1, 0.448)	(0.5, 0.700)
θ_d	(1, 0.540)	(9, 0.448)	(21, 0.498)
N_e	(6, 0.532)	(14, 0.448)	(16, 0.474)

Table 4.3: Results of performance evaluation on `poly_100` and `slam`. AA = Average Absolute Alignment Error; NA = Average Normalized Alignment Error; NP = Average Normalized Position Error; RD = Average Relative Duration Error.

Data Set	**AA**	**NA**	**NP**	**RD**
`poly_100`	0.897 s	0.251	0.229	0.306
`slam`	1.069 s	0.295	0.260	0.327

Table 4.4: Performance comparison with variants of the approach of Mesaros and Virtanen [24] by average absolute alignment errors evaluated on `poly_100`. The approaches are sorted by error in ascending order.

Approach	**Error**
This Dissertation	0.90 s
[24], 8-Class Adaptation	0.94 s
[24], 3-Class Adaptation	0.97 s
[24], 22-Class Adaptation	1.07 s

Table 4.5: Results of experiments conducted on `poly_100` with an algorithmic feature removed from the proposed approach. AA = Average Absolute Alignment Error; NA = Average Normalized Alignment Error; NP = Average Normalized Position Error; RD = Average Relative Duration Error.

Removed Feature	**AA**	**NA**	**NP**	**RD**
None	0.897 s	0.251	0.229	0.306
Timbre	1.198 s	0.320	0.301	0.385
LF Model	1.012 s	0.278	0.253	0.328
Gender	0.979 s	0.269	0.253	0.320
Genre	0.993 s	0.266	0.244	0.334
Voice Type	1.032 s	0.277	0.256	0.338
Diversity	1.103 s	0.297	0.281	0.334
Voicing	9.929 s	0.885	0.898	0.979

Chapter 5

Conclusions

5.1 Contribution

An approach to vocal melody extraction has been presented, which is based on a novel model of F0 likelihood. The F0 likelihood model is built upon a set of vocal timbre examples for the F0 candidate that are generated by F0-modifying a small set of singing voice samples. The F0 modification is achieved by source-filter analysis and synthesis with state-of-the-art models from the field of acoustic phonetics.

The proposed approach to vocal melody extraction has been tested extensively both to evaluate its performance and to investigate the significance of various algorithmic features in the approach. My approach achieved an overall accuracy in the range between 70% and 75% for various data sets, in which range one also finds many state-of-the-art accuracies according to the annual MIREX evaluation. With a series of controlled experiments, we verified that timbral discrimination in the F0 likelihood model is effective as a mechanism for guiding F0 estimation away from instrumental F0s. Even so, discrimination in loudness and prior modeling of vocal F0 are indispensable for vocal F0 estimation.

In addition, an approach to lyrics alignment with audio has been presented, which is based on a novel model of vowel likelihood. The vowel likelihood model is built upon a set of vowel timbre examples for the F0 estimate that are generated by F0-modifying a small set of singing voice samples. The proposed method has been evaluated in multiple experiments, not only to test its efficacy, but to look into the importance of various algorithmic features in the method. For two data sets alike, which are collected from different sources and contain lyrics of different languages, my approach achieved an average normalized alignment error below 0.3 and an average absolute alignment error around one second. A state-of-the-art approach was previously evaluated on one of the two data sets, also giving an absolute error around one second. With a series of controlled experiments, we verified that timbral discrimination in the vowel likelihood model is effective as a mechanism for rendering the alignment phonetically sensitive. Still, voicing modeling based on estimation of vocal loudness is indispensable for lyrics alignment.

5.2 Further Work

This work demonstrates how acoustic-phonetic models lend themselves to distinguishing human voice from instrumental sound, and distinguishing different vowel sounds. It would be intriguing to further inquire application of acoustic-phonetic models to other aspects of singing voice in the future. An example would be the personal timbre that characterizes one's singing, which can be represented with acoustic-phonetic models in an approach to singer recognition.

The task of vocal melody extraction has been approached in this work in the single-vocal scenario. It would definitely be interesting to consider in the future the multiple-vocal case, e.g., melody extraction from a recording of a piano-accompanied mixed-voice four-part choral performance. To that end, one could perform a four-voice version of vocal F0 estimation and select the F0 sequence that exhibits the strongest melodic quality, such as conjunct melodic motion and high loudness.

Bibliography

[1] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *Proc. the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[2] J. C. Brown and M. S. Puckette. An efficient algorithm for the calculation of a constant Q transform. *J. Acoust. Soc. Am.*, 92(5):2698–2701, 1992.

[3] Y.-R. Chien, H.-M. Wang, and S.-K. Jeng. Simulated formant modeling of accompanied singing signals for vocal melody extraction. In *Proc. the 9th Sound and Music Computing Conference (SMC)*, 2012.

[4] K. Dressler. An auditory streaming approach for melody extraction from polyphonic music. In *Proc. the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[5] J. L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, March 2010.

[6] D. P. W. Ellis and G. E. Poliner. Classification-based melody transcription. *Mach. Learn.*, 65(2-3):439–456, 2006.

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

[8] G. Fant. The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR*, 1995.

[9] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno. LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1252–1261, October 2011.

[10] M. Goto. A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43:311–329, 2004.

[11] J. W. Hawks and J. D. Miller. A formant bandwidth estimation procedure for vowel synthesis. *J. Acoust. Soc. Am.*, 97(2):1343–1344, 1995.

[12] 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.*, 18(2):310–319, Feb 2009.

[13] C.-L. Hsu, D. Wang, and J.-S. Jang. A trend estimation algorithm for singing pitch detection in musical recordings. In *2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 393–396, May 2011.

[14] D. Iskandar, Y. Wang, M.-Y. Kan, and H. Li. Syllabic level automatic synchronization of music signals and text lyrics. In *Proc. the 14th Annual ACM International Conference on Multimedia*, 2006.

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

[16] E. Joliveau, J. Smith, and J. Wolfe. Vocal tract resonances in singing: The soprano voice. *Journal of the Acoustical Society of America*, 116(4):2434–2439, October 2004.

[17] S. Joo, S. Park, S. Jo, and C. D. Yoo. Melody extraction based on harmonic coded structure. In *Proc. the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[18] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy. LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):338–349, February 2008.

[19] R. D. Kent and C. Read. *The acoustic analysis of speech*. Singular/Thomson Learning, 2002.

[20] K. Lee and M. Cremer. Segmentation-based lyrics-audio alignment using dynamic programming. In *Proc. the 9th International Conference on Music Information Retrieval (ISMIR)*, 2008.

[21] H.-L. Lu. *Toward a High-Quality Singing Synthesizer with Vocal Texture Control*. PhD thesis, Stanford University, 2002.

[22] M. Mauch, H. Fujihara, and M. Goto. Integrating additional chord information into HMM-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):200–210, January 2012.

[23] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, Aug 1986.

[24] A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(546047):1–11, 2010.

[25] 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.*, 15(4):1247–1256, 2007.

[26] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[27] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, Aug 2012.

[28] G. L. Salomão and J. Sundberg. What do male singers mean by modal and falsetto register? An investigation of the glottal voice source. *Logopedics Phoniatrics Vocology*, 34:73–83, 2009.

[29] C. Sutton. Transcription of vocal melodies in popular music. Master's thesis, Queen Mary, University of London, 2006.

[30] H. Tachibana, T. Ono, N. Ono, and S. Sagayama. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. In *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 425–428, March 2010.

[31] C. H. Wong, W. M. Szeto, and K. H. Wong. Automatic lyrics alignment for Cantonese popular music. *Multimedia Systems*, 12:307–323, 2007.

[32] W. R. Zemlin. *Speech and hearing science: anatomy and physiology*. Allyn and Bacon, Boston, 4th ed. edition, 1998.