AN ACOUSTIC-PHONETIC APPROACH TO VOCAL MELODY EXTRACTION

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

ABSTRACT

This paper addresses the problem of extracting vocal melodies from polyphonic audio. In short-term processing, a timbral distance between each pitch contour and the space of human voice is measured, so as to isolate any vocal pitch contour. Computation of the timbral distance is based on an acoustic-phonetic parametrization of human voiced sound. Long-term processing organizes short-term procedures in such a manner that relatively reliable melody segments are determined first. Tested on vocal excerpts from the ADC 2004 dataset, the proposed system achieves an overall transcription accuracy of 77%.

1. INTRODUCTION

Music lovers have always been faced with a large collection of music recordings or concert performances for them to choose from. While successful choices are possible with a small set of metadata, disappointment still recurs because the metadata only provides limited information about the musical contents. This has motivated researchers to work on systems that extract essential musical information from audio recordings. Hopefully, such systems will enable personalized recommendations for music purchase decisions.

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

The specific problem outlined above is challenging because melody extraction is prone to interference from the accompaniment unless a mechanism is in place for distinguishing human voice from instrumental sound. [6], [13], and [9] determined the predominant pitch as it accounts for the most of the signal power among all the simultaneous pitches. The concept of pitch predominance is also presented in [12] and [2], which defined the predominance in terms of harmonicity. For these methods, the problem proves difficult whenever the signal is dominated by a harmonic musical instrument rather than by the singing voice. [3] and [5] realized the timbre recognition mechanism by classification techniques; on the other hand, pitch classification entails quantization of pitch, which in turn causes loss of such musical information as vibrato, portamento, and non-standard tuning.

The contribution of this paper is an acoustic-phonetic approach to vocal melody extraction. To make judgments about whether or not each particular pitch contour detected in the polyphonic audio is vocal, we measure a timbral distance between the pitch contour and a *space of human voiced sound* derived from acoustic phonetics [4]. In this space, human voiced sound is parameterized by a small number of acoustic phonetic variables, and the timbral distance from the space to any harmonic sound can be efficiently estimated by a coordinate descent search that finds the minimum distance between a point in the space and the point representing the harmonic sound.

The proposed method offers practical advantages over previous approaches to vocal melody extraction. By imposing acoustic-phonetic constraints on the extraction, the proposed method can better distinguish human voice from

instrumental sound than the predominant pitch estimators in [2, 6, 9, 12, 13]. Furthermore, with pitch contours composed of continuous sinusoidal frequency estimates taken from interpolated spectra, the proposed method is free from the quantization errors in pitch estimation that are commonly encountered by classification-based systems [3, 5].

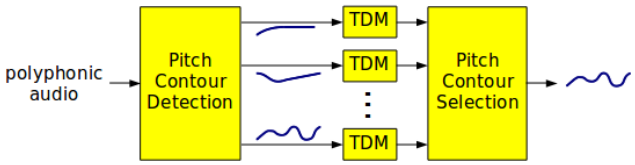

Figure 1. Short-term processing for vocal melody extraction. The goal is to extract a vocal pitch contour around time point t from the polyphonic audio. TDM stands for timbral distance measurement.

2. OVERVIEW OF SHORT-TERM PROCESSING

In this section, we consider the problem of extracting a vocal pitch contour around time point t from the polyphonic audio, provided that a singing voice exists at t. As shown in Figure 1, the extraction proceeds in three steps: 1) detecting pitch contours that each start before and end after t, 2) measuring the timbral distance between each of the detected contours and the space of human voiced sound, and 3) extracting the most salient pitch contour among any detected contours that lie in the space of human voiced sound.

In particular, the pitch contours simultaneously detected in Step 1 form a set of candidates for the vocal pitch contour. If exactly one vocal exists at this moment, then the vocal contour may be identified by timbre. Timbral distance measurement is intended here to provide the timbral information essential to the identification. In contrast to frame-based processing, here the duration of processing depends on how far pitches can actually be tracked continuously away from t in the analyzed audio. At the frame rate of 100 frames per second, it is observed that most pitch contours last for more than 10 frames; obviously, one would expect more reliable timbral judgments from contour-based processing than from frame-based processing.

3. PITCH CONTOUR DETECTION

In this section, we describe the procedure for detecting pitch contours around time point t from the polyphonic audio. It starts by detecting multiple pitches from the audio frame at t. Next, pitch tracking is performed separately for each detected pitch, from t forwards, and then also from t backwards, as depicted in Figure 2. Consequently, this procedure gives as many pitch contours as pitches are detected at t.

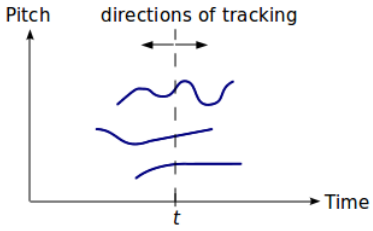

Figure 2. Bi-directional multi-pitch tracking around time point t.

3.1 Pitch Detection

In order to detect pitches at the time point t, we apply sinusoidal analysis to the short-time spectrum of the polyphonic audio signal at t. The analysis extracts (quadratically interpolated) frequencies of the loudest three peaks in the first-formant section (200–1000 hertz) of the magnitude spectrum. The loudness of a sinusoid is computed by correcting its amplitude according to the trends in the 40-phon equal-loudness contour (ELC) [8], which quantifies the dependency of human loudness perception on frequency. For each extracted sinusoidal frequency $\tilde{f}$ (hertz), the procedure "detects" up to three pitches in the 80–1000 hertz vocal pitch range, at $\tilde{f}$, $\tilde{f}/2$, and $\tilde{f}/3$, regarding the sinusoid as the fundamental, the second partial, or the third partial of a pitch. As a result, the pitch detector gives nine pitches at the most for the time point t. The ambiguity among the first three partials will not be resolved until a selection is made among pitch contours.

3.2 Pitch Tracking

Suppose that we are now appending a new pitch to the end of a growing pitch contour. Calculation of the new pitch proceeds in three steps: 1) finding in the new spectrum a set of sinusoids around (within one half tone of) the first three partials of the last pitch in the contour, 2) finding among the sinusoids the one with the highest amplitude, and 3) dividing the frequency (hertz) of this sinusoid by the corresponding harmonic multiple (1, 2, or 3). In other words, the pitch contour is guided by nearby high-energy pitch candidates. The growth of a pitch contour stops once the amplitude of the loudest partial drops (cumulatively) from a peak value by more than 9 dB, i.e., a specific form of onset or offset is detected, with the loudness of each partial evaluated over the entire contour as a time average.

4. TIMBRAL DISTANCE MEASUREMENT

In this section, we develop a method for measuring the timbral deviation of a pitch contour $\mathcal{C}$ from human voiced sound, which is based on an acoustic-phonetic parameterization of

human voiced sound, and finding within the space of human voiced sound the minimum distance from $\mathcal{C}$, as illustrated in Figure 3.

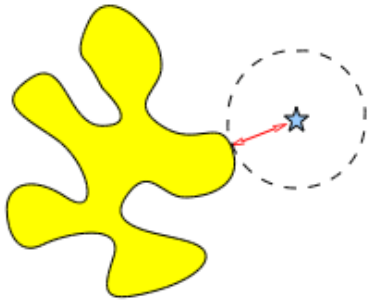

Figure 3. Measuring the timbral distance between a pitch contour (star) and the space of human voiced sound.

4.1 Parameterization of Human Voiced Sound

In order to model the space of human voiced sound, it is desirable to identify every point in the space with a set of acoustic-phonetic parameters. To this end, we let each short-time magnitude spectrum of human voiced sound be represented by seven parameters: the amplitude, the fundamental frequency, the first three formant frequencies, and the nasal formant and anti-formant frequencies [11]. Such a parameterization is appropriate for specifying human voiced sound in that sinusoidal parameters of the voice can be obtained from the acoustic-phonetic parameters through well-defined procedures. Obviously, partial frequencies of the human voiced sound can be derived as integer multiples of the fundamental frequency. On the other hand, partial amplitudes of the human voiced sound can be derived on the basis of formant synthesis [4], which has been applied to synthesizing a wide range of realistic singing voice [15].

Consider a point in the space of human voiced sound

$$\mathbf{s} = (a, f_0, f_1, f_2, f_3, f_p, f_z)^T, \tag{1}$$

where a is the amplitude (in dB), f_0 is the fundamental frequency (in quarter tones), f_1, f_2, and f_3 are the first three formant frequencies (in hertz), and f_p and f_z are the nasal formant and anti-formant frequencies (in hertz). Amplitude of partials can be calculated from $\mathbf{s}$ by [4]

$$a_i^p = a + 20\log_{10}\left|U_R(if_0^h)K_R(if_0^h)\prod_{n\in I_f} H_n(2\pi\cdot if_0^h)\right|, \tag{2}$$

where a_i^p is the amplitude of the ith partial in dB, $i \le 10$, f_0^h denotes the fundamental frequency in hertz:

$$f_0^h = 440\cdot 2^{(f_0-105)/24}, \tag{3}$$

$U_R(\cdot)$ represents the (radiated) spectrum envelope of the glottal excitation [4]:

$$U_R(f) = \frac{f/100}{1+(f/100)^2}, \tag{4}$$

$K_R(\cdot)$ represents all formants of order four and above [4]:

$$20\log_{10} K_R(f) \approx 0.72\left(\tfrac{f}{500}\right)^2 + 0.0033\left(\tfrac{f}{500}\right)^4, \quad f \le 3000, \tag{5}$$

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

$$H_n(\omega) = \frac{1}{\left(1-\frac{j\omega}{\sigma_n+j\omega_n}\right)\left(1-\frac{j\omega}{\sigma_n-j\omega_n}\right)}, n = 1, 2, 3, p, \tag{6}$$

$$H_z(\omega) = \left(1-\frac{j\omega}{\sigma_z+j\omega_z}\right)\left(1-\frac{j\omega}{\sigma_z-j\omega_z}\right). \tag{7}$$

In (6), ω_n is the frequency of formant n in rad/s, i.e., $\omega_n = 2\pi f_n$, and σ_n is half the bandwidth of formant n in rad/s, which can be approximated as a function of ω_n by a polynomial regression model [7].

4.2 Distance Minimization

Suppose that the instantaneous pitch values in contour $\mathcal{C}$ have mean $f_{\mathcal{C}}$. Now, let the vector

$$\mathbf{x} = (a, f_1, f_2, f_3, f_p, f_z)^T \tag{8}$$

denote any point on the hyperplane $f_0 = f_{\mathcal{C}}$ in the space of human voiced sound. Then we can define the distance between $\mathbf{x}$ and $\mathcal{C}$ as

$$D_{\mathcal{C}}(\mathbf{x}) = \sqrt{\sum_{i=1}^{10}\left(\tfrac{a_i^q - a_i^p}{\sigma_a}\right)^2}, \tag{9}$$

where a_i^q is the mean amplitude (in dB) of the ith partial of $\mathcal{C}$, a_i^p is the amplitude (computed as in (2)) of the ith partial of $\mathbf{x}$, and σ_a is an empirical constant set to 12. The timbral distance between $\mathcal{C}$ and the space of human voiced sound can now be measured as

$$\min_{\mathbf{x}\in\mathcal{X}} D_{\mathcal{C}}(\mathbf{x}), \tag{10}$$

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

$$\mathcal{X} = \left\{ \mathbf{x}\in R^6 \;\middle|\; \begin{array}{l} 250 \le f_1 \le 1000 \\ 600 \le f_2 \le 3000 \\ 1700 \le f_3 \le 4100 \\ 200 \le f_p \le 500 \\ 200 \le f_z \le 700 \\ f_p, f_z \le f_1 \le f_2 \le f_3 \end{array} \right\}. \tag{11}$$

The accuracy in determining whether or not $\mathcal{C}$ is vocal depends on how well the distance in (9) is numerically minimized. To be specific, if $\mathcal{C}$ is vocal and the timbral distance between $\mathcal{C}$ and the space of human voiced sound is overestimated due to distance minimization being trapped in a local minimum, then $\mathcal{C}$ may very likely turn out to be mistaken by the procedure for an instrumental contour. Our numerical experience revealed that the best of twenty local searches for the minimum defined in (10), which are initialized respectively with twenty different reference points, shows great consistency in associating vocal pitch contours with short timbral distances. These reference points differ only in the oral formant frequencies f_1, f_2, and f_3, with numerical values taken from the gender-specific averages for ten vowels of American English [10]: 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.

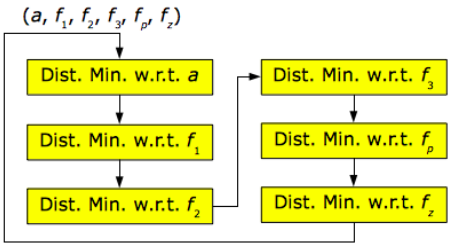

Figure 4. Each update in the local search for the minimum distance consists of a series of one-variable subproblems.

The local search for the minimum defined in (10) may be achieved with any local optimization technique. Here we use a simple coordinate descent algorithm, as represented in Figure 4, where each (all-variable) update consists of a series of one-variable updates. Each one-variable update minimizes the distance with respect to the variable alone while fixing the other variables. For instance, the update of the formant frequency f_2 in the jth all-variable update operates on the current point

$$(a^{(j)}, f_1^{(j)}, f_2^{(j-1)}, f_3^{(j-1)}, f_p^{(j-1)}, f_z^{(j-1)})^T \tag{12}$$

by computing

$$\begin{aligned} &f_2^{(j)} = \arg\min_{f_2 \in I_2} \\ &D_{\mathcal{C}}\left((a^{(j)}, f_1^{(j)}, f_2, f_3^{(j-1)}, f_p^{(j-1)}, f_z^{(j-1)})^T\right), \\ &I_2 = \{f_2 \in R \mid 600 \le f_2 \le 3000, f_1^{(j)} \le f_2 \le f_3^{(j-1)}\}. \end{aligned} \tag{13}$$

In our implementation, the subproblem (13) 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 (10) is refined by continuing the local search with a 10-hertz spacing of formant frequency sampling.

5. PITCH CONTOUR SELECTION

In this section, we present a procedure for selecting the vocal pitch contour from a set of pitch contours detected around time point t. To begin with, it prunes those pitch contours that have been associated with a long timbral distance from the space of human voiced sound. A pitch contour is accepted only if the timbral distance does not exceed the empirical threshold of $\sqrt{-2\log 0.4}$. In addition, if the mean amplitude over even partials of a pitch contour exceeds that over odd partials by more than 7 dB, the contour is rejected, taken as the octave below a true pitch contour.

Secondly, the procedure prunes some pitch contours that can be seen as an overtone as related to another pitch contour. To this end, the overlap time interval between each pair of contours is calculated, and the pitch interval between two contours is determined on the basis of the mean pitch during the overlap. The procedure rejects any pitch contour that has a mean pitch at the 2nd, 3rd, or 4th partial of another contour.

Lastly, the procedure selects the loudest pitch contour from any contours that survived the prunings, thereby providing a mechanism for identifying the predominant lead vocal out of several simultaneous singing voices. The loudness of each pitch contour is defined as the mean of its instantaneous loudness values, which are each calculated by summing the linear-scale, ELC-corrected instantaneous power over the partials.

6. LONG-TERM PROCESSING

At the excerpt level, the goal of processing is an interleaved sequence of vocal pitch contours and pauses. To this end, we maintain a list of *visited frames* throughout the segmentation process. A frame is considered visited whenever a vocal pitch contour has been extracted whose duration covers the frame.

Suppose that at this moment the procedure has extracted k vocal pitch contours from the excerpt, with the list of visited frames updated accordingly. The procedure attempts to extract the $(k+1)$th contour around time point t, which is set to the unvisited frame that has the highest signal loudness among all the unvisited frames. Here, the loudness of a frame is calculated by summing the linear-scale, ELC-corrected power over sharp peaks in the spectrum. The sharpness threshold of each spectral local maximum is set to 9

dB above the mean amplitude over the neighboring 5 frequency bins. In case that the new contour should overlap with an existing contour, the new contour would be truncated to resolve the conflict. This procedure continues until the loudness of every unvisited frame is below the excerpt-wide median. These remaining unvisited frames form the final pauses between vocal pitch contours.

7. EXPERIMENTS

In this section, to provide comparison of our method with some existing methods, we conduct vocal melody extraction experiments on a publicly available dataset.

7.1 Dataset Description

The dataset is a subset of the one built for the Melody Extraction Contest in the ISMIR2004 Audio Description Contest (ADC 2004). The whole ADC 2004 dataset consists of 20 audio recordings, each around 20 seconds in duration, among which eight recordings have instrumental melodies, and the other twelve have vocal melodies. Since this work considers vocal melodies only, experiments are carried out exclusively on the 12 vocal recordings, including four pop song excerpts, four song excerpts with synthesized vocal, and four opera excerpts. The dataset has been in use in several Music Information Retrieval Evaluation Exchange (MIREX) contests since 2006; therefore, it affords extensive comparison among methods.

Before melody extraction, each audio file in the dataset is resampled at 11,025 hertz and constant-Q transformed [1] ($Q = 34$) into a sequence of short-time spectra. Each resulting spectrum is a quarter-tone-spaced sampling of a continuous spectrum that is capable of resolving the interference between two half-tone-spaced sinusoids from 21.827 hertz all the way to 5,428.6 hertz.

7.2 Performance Measures

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

MIREX adopts several measures for evaluating the performance of a melody extraction system [14]. In the first place, to determine how well the system performs voicing detection, we use the voicing detection rate, the voicing false alarm rate, and the discriminability. The voicing detection rate is computed as the fraction of frames that are both labeled and estimated to be voiced, among all the frames that are labeled voiced. The voicing false alarm rate is computed as the fraction of frames that are estimated to be voiced but are actually not voiced, among all the frames that are not voiced according to the reference transcription. The discriminability combines the above two measures in such a way that it can be deemed independent of the value of any threshold involved in the decision of voicing detection:

$$d' = Q^{-1}(P_F) + Q^{-1}(1 - P_D), \tag{14}$$

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

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

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

Excerpt	Accuracy (%)			PD (%)	PF (%)	d'
	All	Voiced	Chroma			
pop1	61.515	60.548	62.027	88.110	34.529	1.5785
pop2	65.656	65.481	65.551	85.704	33.855	1.4835
pop3	78.422	79.008	82.634	86.196	23.721	1.8045
pop4	82.271	81.136	82.308	91.798	13.508	2.4944
daisy1	84.116	85.012	88.433	92.786	18.762	2.3467
daisy2	89.409	88.925	90.337	92.291	8.101	2.8233
daisy3	96.301	96.301	96.301	99.472	-	-
daisy4	96.486	96.479	96.682	97.833	0.000	-
opera_fem2	61.018	61.275	65.418	87.649	39.888	1.4139
opera_fem4	75.917	74.347	74.822	86.936	8.594	2.4896
opera_male3	62.000	61.798	64.326	82.921	36.364	1.2998
opera_male5	70.978	72.437	74.395	80.526	39.209	1.1344

Table 1. Experimental results.

7.3 Results

The results are listed in Table 1. The overall transcription accuracies listed in the column titled "All" range from 61% to 96% and have their average at 77.007%. The minimum is found at the excerpt "opera_fem2." A close look at a significant error made in the analysis of this excerpt revealed that the system mistakenly selected the octave below a true

vocal pitch contour because the octave had a timbral distance of $\sqrt{-2\log 0.41}$, slightly shorter than the upper limit set for a vocal contour. Still, the distance measured for the true vocal pitch contour was much shorter, at $\sqrt{-2\log 0.98}$. This suggests that a relative threshold for the timbral distance may be implemented along with the absolute threshold to further improve the accuracy. To see the effect of timbral distance measurement on the average accuracy, we repeated the experiments with the distance threshold set to infinity, so that no contour was pruned because of a large timbral deviation from human voiced sound. This turned out to reduce the mean accuracy from 77.007% to 75.233%, which verifies the benefit of timbral distance measurement. The raw pitch accuracies in the column titled "Voiced" are highly correlated with the overall transcription accuracies, which suggests that further improvement of this system should be made in pitch estimation, not in voicing detection. The column titled "Chroma" contains raw chroma accuracies similar to the raw pitch accuracies, which suggests that octave errors were successfully avoided by the system.

Shown in Table 2 is a comparison of the proposed method with the MIREX 2009 submissions in terms of the overall transcription accuracy (OTA). Notably, if the proposed method had entered the evaluation in 2009, it would have ranked 5th out of a total of 13 submissions. Moreover, the accuracy of the proposed system is within 10% of the highest accuracy in the 2009 evaluation.

Method	1	2	3	4	5	6	7	8	9	10	11	12	Proposed
OTA (%)	75	75	80	78	49	45	75	86	71	86	74	51	77

Table 2. Comparison with the MIREX 2009 Audio Melody Extraction results.

8. CONCLUSION

We have presented a novel method for vocal melody extraction which is based on an acoustic-phonetic model of human voiced sound. The performance of this method is evaluated on a publicly available dataset and proves comparable with state-of-the-art methods. [1]

9. ACKNOWLEDGMENTS

This work was supported in part by the Taiwan e-Learning and Digital Archives Program (TELDAP) sponsored by the National Science Council of Taiwan under Grant: NSC 100-2631-H-001-013.

[1] Octave code available at `http://www.iis.sinica.edu.tw/~yrchien/english/melody.htm`

10. REFERENCES

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

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

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

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

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

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

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

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

[9] S. Jo and C. D. Yoo. Melody extraction from polyphonic audio based on particle filter. In *ISMIR*, 2010.

[10] Ray D. Kent and Charles Read. *The acoustic analysis of speech*. Singular/Thomson Learning, 2002.

[11] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *JASA*, 67(3):971–995, 1980.

[12] M. Lagrange, L.G. Martins, J. Murdoch, and G. Tzanetakis. Normalized cuts for predominant melodic source separation. *IEEE Trans. on ASLP*, 16(2):278–290, 2008.

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

[14] 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. on ASLP*, 15(4):1247–1256, 2007.

[15] J. Sundberg. The KTH synthesis of singing. *Advances in Cognitive Psychology*, 2(2-3):131–143, 2006.