# VOCAL MELODY EXTRACTION BASED ON AN ACOUSTIC-PHONETIC MODEL OF PITCH LIKELIHOOD

**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@ntu.edu.tw

## ABSTRACT

This submission deals with the task of automatically extracting vocal melodies from accompanied singing recordings. The extraction is based on a model of vocal pitch likelihood that integrates acoustic-phonetic knowledge and real-world data. The likelihood model evaluates a timbral fitness score, as well as the loudness, of each pitch candidate. The timbral fitness is measured for the partial amplitudes of the pitch candidate, with respect to a small set of vocal timbre examples. The pitch-specific measurement of timbral fitness depends on an acoustic-phonetic pitch transformation of each timbre example. In the loudness part of the likelihood model, sinusoids are detected, tracked, and pruned to give loudness values that minimize the interference from the accompaniment. The final pitch estimate is determined by a prior model of pitch sequence in addition to the likelihood model. The extraction is completed by detecting voiced time positions according to the singing voice loudness variations given by the estimated pitch sequence.

## 1. INTRODUCTION

In this submission, 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.

This submission is a variant of the acoustic-phonetic approach proposed in [2]. To evaluate the likelihood of each pitch candidate, we calculate its *timbral fitness* and loudness. The timbral fitness measures the timbral similarity of the observed partial amplitudes to a set of vocal timbre examples that compactly represent vocal timbres of different genders, genres, voice types, and vowel types. In calculating such a pitch-specific timbral similarity score, each timbre example is pitch-transformed to the pitch candidate 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 pitch value set to the candidate.

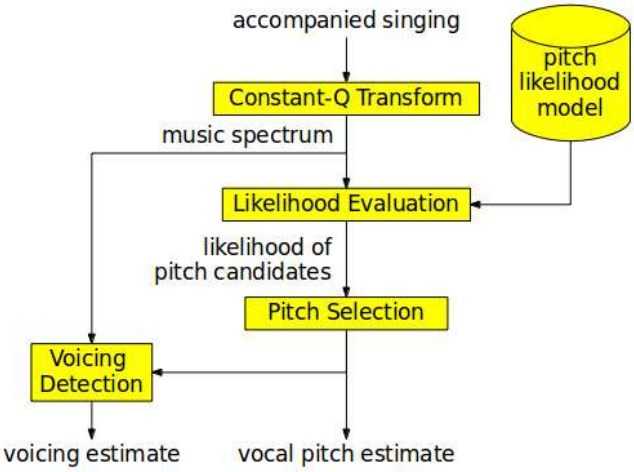


**Figure 1**. Block diagram of the submitted melody extraction system.

## 2. OVERVIEW OF THE METHOD

As shown in Figure 1, the extraction procedure starts by calculating a sequence of constant-Q spectra [1] from the input signal. After that, for each time position at which a spectrum has been calculated, the procedure evaluates the likelihood of each pitch candidate according to an acoustic-phonetic model of pitch likelihood. Next, a pitch estimate is selected for each time position, where both the pitch continuity between adjacent time positions and a reasonable global pitch range are taken into consideration in addition to the likelihood scores.

As the final step in the extraction procedure, we perform voicing detection at each time position by detecting a high power value on the vocal pitch estimate, where the power value is estimated by summing the squared sinusoidal amplitude over several partials of the vocal pitch estimate. Since the vocal pitch estimate is required to always evolve continuously over time, time positions that do not actually have a melodic pitch typically have a vocal pitch estimate around an earlier or later true pitch, and thus see a low power value.

## 3. ACKNOWLEDGMENTS

This work was supported in part by the Taiwanese Ministry of Science and Technology under Grant: NSC 102-2221-E-001-008-MY3.

## 4. REFERENCES

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

[2] Yu-Ren Chien, Hsin-Min Wang, and Shyh-Kang Jeng. Simulated formant modeling of accompanied singing signals for vocal melody extraction. In *Proc. the 9th Sound and Music Computing Conference (SMC)*, 2012.