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 submission 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.

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 an implementation of the acoustic-phonetic approach proposed in [1]. 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 [2]. 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.

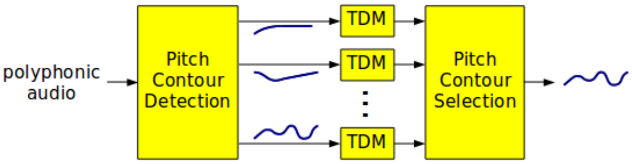

Figure 1. Short-term processing for vocal melody extraction. The goal is to extract a vocal pitch contour around time point t from the polyphonic audio. TDM stands for timbral distance measurement.

2. OVERVIEW OF THE METHOD

In this section, we first 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]:

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.

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 [1]. 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. 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.

3. 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.

4. REFERENCES

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

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