ELEN 4810 PROJECT PROPOSAL

Yi Luo yl3364

Xiaowen Zhang xz2461

1. PROJECT DESCRIPTION

For the course project, we aim at creating a system for singing transcription. The main idea of singing transcription is to take raw monophonic singing signal as the input, using signal processing and machine learning technique to generate a series of symbolic representations (e.g. the score) of the original signal. This technique is widely needed in query-by-humming/singing systems, Karaoke scoring systems and pedagogical systems.

There's a widely accepted framework for singing transcription. First, the raw singing voice is pre-processed to remove the noise or strengthen the frequencies of human voice. Next, low-level features such as fundamental frequency (F0), voiced/unvoiced regions, etc. are extracted. Following that is the process of segmenting the notes (which contains detection of onset and offset of a note) and assigning symbolic representations to the notes, and these two jobs can be done separately or jointly. Finally, some post-processing techniques are carried through to generate the final symbolic representation of the whole signal. We'll follow this framework.

2. PROJECT PLAN

Since I've done some work on singing transcription, I hope that this time I could design a system that performs much better than the previous one, so first I need to introduce what I have done in my previous system.

In my previous system, first I used the pYIN [1] algorithm to generate the F0 curve and the voiced/unvoiced regions of the singing voice jointly. Then I used a two-step thresholding segmentation method to segment the F0 curves into notes (a note consists of several consecutive frames)). Specifically speaking, I first adopted a large threshold (170 cents) to find the large deviations in pitch which may represent note changing or the end of a phrase. After segmentation in this step, I calculate the note pitch of each note (assigning a pitch value for the whole note) by using "dominant pitch" of the note, which is calculated by calculating the mean of the bar with the most quantity of pitch values in the pitch histogram of all the frames inside the note. Next I adopted a small threshold (65 cent) to find the transitions of notes that have a relatively small pitch deviation, but before doing that, I utilized the autocorrelation method on the F0 curve to find the possible vibrato notes incase that the vibrato range is larger than 65 cents hence be falsely segmented by the small threshold. Finally, I adjust the onset and offset positions of the notes and recalculate the note pitch, and take the result as the output of this system.

What I want to do now is much different from the system above. Although I said at the office hour that I only hope to make use of HMM and chromagram to enhance the performance of the system, now I want to do more than that after reading papers and books there days. I'll list the things that I want to modify below.

(1) As said at the office hour, I hope to add denoising in the pre-processing process. I think Bayesian methods could help if I assume that the noise is Gaussian.

(2) Since the F0 curve of the signal may contain multiple outliers, I think a smoothing process is necessary.

(3) I used pYIN algorithm before to detect voiced/unvoiced regions and calculate the pitch for each frame. However, pYIN requires accurate training sets and its performance varies if we haven't chosen a training set that is good enough. Hence I want to substitute it with YIN algorithm and find another way to detect the voiced/unvoiced regions. The things that I could make use of includes zero crossing rates, aperiodicity, STFT, bandpass filter, etc.

(4) I think only by doing thresholding without some assumption about the F0 curve is too ad hoc. Maybe I should not use this naïve thresholding method. What I'm thinking about now is: since the main point of thresholding is to find the onset and offset of the notes, maybe I could make use of the property of neurons – that is, try to encode the onsets of notes to the spikes of the neuron? This idea is motivated by the computational neuroscience course I take this semester, and I think it deserves trying. If it would not work, maybe I need to try something like prominent functions or moving average of pitch curve to detect the onset and offset.

(5) There is another concern about the onset and offset: since the metric of the singing voice should be constant, the onset and offset should only happen on specific times. Hence maybe we could first calculate the metric information of the singing voice and use it to support our detection of onset and offset. The ways to calculate metric information includes transient event detection using STFT, ICA on STFT, oscillating filter, etc. I'll try to find one that is most proper.

(6) The usage of HMM and N-gram models in note-pitch assigning have been studied by many researchers. I also want to make use of them to assign correct pitch value to the notes. Maybe I could also use something else like CRF.

(7) In the previous system I detect vibrato by calculating autocorrelation function in pitch-curve. This method sometimes falsely notates some unintended pitch deviations as vibrato regions. This time I hope to use STFT or to design a specific filter to detect the vibrato regions more robustly.

(8) It's very important that we consider the musical context while doing transcription. Information such as music key and music scale may help when we calculate note pitch. I haven't thought of that before, so this time I'll make use of them

That's what I hop to do in this project. Please let me know if there are anything that needs to change.