MUMT621 Summary on Presentation 4: Singing Transcription

Ryan Groves
ryan.groves@mail.mcgill.ca

11 March 2013

Contents

1 Introduction

For many years, research has been done on the automatic transcription of musical audio recordings. The usefulness of such a technology is clear and wide-ranging. Given an effective means of transcribing audio, every musical audio recording could be turned into a score, for example, which would provide a massive dataset for automatic symbolic music analysis. However, musical transcription is a difficult and multi-faceted task. Even with current research, the automatic transformation of musical recordings into a symbolic musical representation has not been successfully completed. The transcription of the human singing voice, specifically, provides further challenges. In order to understand the task of singing transcription better, one must understand the process. The first step is to consider the types of sources of audio that are appropriate for singing transcription.

2 Sources

There are many sources that could contain a musical vocal audio performance. These sources are reduced and consolidated into two main categories- polyphonic and monophonic audio. Polyphonic audio refers to any audio source which contains not only a single vocal performance, but also other instrumental or vocal performances simultaneously being sounded. Monophonic audio is just the opposite–audio which contains only a single vocal performance. There is also a qualification on the term *vocal performance*; for singing transcription purposes, a vocal performance must only a single note at any given time.

3 Process

In general, the task of singing transcription refers to the extraction of melodic and rhythmic content from a vocal audio performance. The tasks can be categorized into pre-processing, fundamental frequency estimation, note segmentation, note labelling, with an optional transcription category.

3.1 Pre-processing

The main task for pre-processing the audio is to remove the noise in the audio recording. The noise is considered to be the sound that is not part of the vocal performance. In order to perform this, algorithms are designed to first identify the component of the audio that represents the noise, and then to attenuate it. This method was derived from earlier research done in speech recognition. Some algorithms perform better if a spectral whitening is also performed as a pre-processing step, but this step is optional.

3.2 Fundamental Frequency Estimation

Once the audio is cleaned and the vocal performance is isolated, there are a few different types of audio analysis that can be used to estimate the fundamental frequency of the

audio, over time.

3.2.1 Time-domain Methods

One of the most popular methods for fundamental frequency estimation is the Autocorrelation method. The autocorrelation method compares an audio signal with a time-shifted version of itself. Because of the periodicity of fundamental frequencies, when a signal is autocorrelated there will be peaks in the resulting signal when the time variable τ is equal to a multiple of the fundamental period. Since the fundamental period is inversely related to the fundamental frequency, the process simply requires the location of the τ value that best represents the fundamental period.

The YIN algorithm (Kawahara and de Cheveigne 2002) extended the autocorrelation method, and made it more robust against octave errors. The algorithm calculates the squared difference function to find the same τ variable. The squared difference value is also normalized, which makes the algorithm more robust against the absolute level of the signal, as well.

3.2.2 Frequency-domain Methods

Similar to autocorrelation, the Cepstrum method can be used to identify the fundamental period of an audio signal. The cepstrum method was used in speech recognition algorithms as well, however it not very robust against noise. Thus, other methods were developed.

A more intuitive approach was devised, based on a two-way mismatch procedure (Maher and J. Beauchamp 1994). This method compares the short-time spectrum of the given audio signal to that of a set of trial fundamental frequencies. Each trial fundamental frequency is assigned an error based on that comparison, and the fundamental frequency with the lowest error is chosen. Some other algorithms focused instead on the magnitude spectrum. These algorithms search the audio for peaks in the magnitude spectrum, and correlate specific fundamental frequencies to the groups of peaks.

Yet another way to approach the task is to model the human auditory system, and process the audio in the same way. Clarisse et al. modelled the human cochlea in their approach, although it was computationally expensive(Clarisse, Martens, Lesaffre, Baets, Meyer, and Leman 2002).

3.3 Note Segmentation

Note segmentation is the most challenging part of the singing transcription task. In this step of the process, note onset and offset timing information is determined by extracting a set of audio features related to the singing voice, and assembling the information they provide to calculate rhythmic information.

3.3.1 Features

The most straightforward feature for onset detection is the energy of the signal, as it is considered to reflect the loudness of the vocal performance. Often it is computed using the root-mean-square method (RMS). A simple method for exploiting this feature is to simply threshold the energy level, so that a note starts when the energy level crosses from below to above the threshold, and the note ends when its energy returns below that threshold (McNab, Smith, and Witten 1995).

Another audio useful audio feature for note segmentation is zero-crossing rate, which indicates the transient sounds that generally represent consonant enunciation in singing. The zero-crossing rate (ZCR) simply measures how many times the audio signal crosses the zero line in the sample representation. A higher ZCR suggests that there are lots of high frequencies at the current time in the audio.

Ryynanen and Klapuri suggested a novel representation of potential onsets, which he calls the accent (Ryynanen and Klapuri 2004). Their accent signal is a measure of spectral change over time, and is found using the differentiation of the power signal. They use the accent signal to find both potential note onsets and to predict the metrical accents of the audio.

3.4 Note Labelling

The human vocal instrument is unique to most other instruments (not all) in that it does not have a way to tune to absolute pitch. A singer must ground themselves to a specific frequency using their ears as well as their training, and must then perform in the tempered scale when moving from note to note. In monophonic vocal performances, it is easy for even a professional singer to sing outside of absolute pitch tuning, or even to change their tuning as they continue to sing. Different algorithms have chosen to address these issues (or chosen not to) with various methods.

Haus and Pollastri assumed that the singer would remain in the tuning in which they began throughout the vocal performance, and thus the tuning would be an offset from absolute tuning(Haus and Pollastri 2001). In order to correct for this, they tallied the most common deviation from the mean of the fundamental frequency of each note segment to the closest absolute pitch. Afterwards, they adjusted the whole set of pitches by this most frequent deviation.

Others decided this was not enough, and that the system should adapt to the current tuning of the performance as time progressed(McNab, Smith, and Witten 1995). In this system, the rounding of the previous note (to the absolute scale) is used to adjust the current note. The drawback of this approach is that it is easy to offset the tuning too much.

In general, the formula used for converting fundamental frequency estimates into MIDI notes is (Chamberlin, Ghias, Logan, and Smith 1995)

$$n_{MIDI} = 12log_2\frac{F0}{440Hz} \tag{1}$$

3.5 Machine Learning Methods

There has been much interest in the automatic determination of both the note segments and the note pitch labels using machine learning methods. Klapuri and Ryynnen, for example, create a comprehensive approach using the accent feature in conjunction with a Hidden Markov Model. In fact, each note had its own HMM, in which the state set consisted of Attack, Sustain, and Silence/Noise states. It was shown to perform with less than ten percent error for the data set they generated (Ryynanen 2006). Viitaniemi et al. created a method using an HMM where each state was instead a note, and the different note transitions were modelled (Klapuri, Viitaniemi, and Eronen 2003)

4 Conclusion

It is evident that the transcription of singing performances is a well-researched topic with a lot of progress made. The uses of such technologies could range from query-by-humming musical searches, training tools for singers, automatic transcription of melodies, and more. Methods have been developed already for the automatic transcription of the note sequences (Cemgil, Desain, and Kappen 2000), which could lead to the population of large data sets of transcribed audio. However, there still remains to be a lot of work done in the realm of transcription in general, especially when incorporating multiple instruments and voices.

References

Cemgil, T., P. Desain, and B. Kappen. 2000. Rhythm quantization for transcription. *Computer Music Journal* 24 (2): 60–76.

Chamberlin, D., A. Ghias, J. Logan, and B. C. Smith. 1995. Query by humming - musical information retrieval in an audio database. *ACM Multimedia* 95: 231–36.

Clarisse, L., J. Martens, M. Lesaffre, B. Baets, H. Meyer, and M. Leman. 2002. An auditory model based transcriber of singing sequences. In *Third International Conference on Music Information Retrieval: ISMIR 2002*, 116–23.

Haus, G., and E. Pollastri. 2001. An audio front-end for query-by-humming systems. In *Proceedings of the International Symposium on Music Information Retrieval: ISMIR*, 116–23.

Kawahara, H., and A. de Cheveigne. 2002. Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America* 111 (4): 1917–30.

Klapuri, A., T. Viitaniemi, and A. Eronen. 2003. Probabilistic models for the transcription of single-voice melodies. *Tampere University of Technology*: 59–63.

Maher, R., and J. J. Beauchamp. 1994. Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of the Acoustical Society of America* 95 (4): 2254–63.

McNab, R., L. Smith, and I. Witten. 1995. Signal processing for melody transcription. *Department of Computer Science, University of Waikato*: 4.

Ryynanen, M. 2006. Singing transcription. *Signal Processing Methods for Music Transcription* 4: 361–90.

Ryynanen, M., and A. Klapuri. 2004. Modelling of note events for singing transcription. In *Proceedings of IEEE Workshop on Statistical and Perceptual Audio Processing: SAPA*, 216–221.