Singing Transcription Summary

Stephen Sinclair, MUMT 611

16 February 2006

1 Introduction

A reliable singing transcription system is desireable for a number of possible applications. These include notation of sung performances, query-by-humming systems (QBH), and tools for computer-based musical composition. Other possible extensions include transcription of lyrics and recognition of language, however these are considered outside the scope of singing transcription. QBH is the oldest of these goals, explored as early as 1994 by Japanese researchers (Kageyama & Takashima 1994), and 1995 in the United States (Ghias et al. 1995). Since then, there have been many more papers written on QBH (Jang et al. 2000; Kosugi et al. 2000; Nishimura et al. 2001).

In this summary I will outline the necessary steps for a singing transcription system. This will be followed by a list of some possible sources of error commonly encountered in these systems and a description of methods for evaluating them.

2 Steps

Singing transcription, like other kinds of transcription systems, can be broken down into a series of three main steps: audio segmentation, pitch detection, and assignment of note values. However, there are several steps or aspects of these steps, outlined below, which are particular to singing transcription. Additionally, steps such as voice separation are not always considered part of the transcription system.

2.1 Voice Separation

Depending on the source material, it may be necessary to perform separation of the voice of interest from other instruments or background noise. Transcription papers generally do not go into detail on this subject, prefering to concentrate on other aspects of the process (Weihs & Ligges 2003), however voice separation is potentially an important step if the source material is a database of popular music or opera music, for example. Additionally, voice separation may also refer to handling of multiple sung voices, and transcription of polyphony. Most papers in the literature assume pre-separated vocal tracks. No papers refering to automatic transcription of vocal polyphony were found.

2.2 Audio Segmentation

Audio segmentation is necessary in order to detect the onsets and length of each sung note. Segmentation often uses a signal energy thresholding technique (Clarisse et al. 2002), though detection of transient regions in the spectral information can also be successful. Similarly, pitch detection routines can be used for segmentation by analysing differentials of detected pitches (Weihs & Ligges 2003).

2.3 Pitch Detection

Pitch detection is usually performed using a sliding window approach. A window size is chosen, and fundamental frequency is estimated based on the spectral content of that window. The window is moved along the signal by a timestep, often equal to the size of the window. Approaches usually use autocorrelation or Fourier transform analysis to estimate fundamental frequency (Haus & Pollastri 2001). Other filtering might take place, such as in the case of Clarisse et al., who used a model of the human auditory system to filter input and perform pitch detection based on a model of the cochlea (Clarisse et al. 2002).

2.4 Island Building

Particularly with vocal music, it is necessary, after performing pitch detection, to eliminate sections which likely caused the algorithm to perform poorly. These include regions of silence as well as regions with higher noise content, such as sung consonants. Vowels do not present the same problems. Thus the elimination of these regions causes the pitch curve to resemble islands where each island represents a different note. Islands also make it easier to determine regions which must be smoothed to eliminate note assignment problems associated with vibrato. (Wang et al. 2003)

2.5 Smoothing

Since a single note value will be assigned to a section where the estimated pitch may have varied either slightly, or in the case of vibrato, noticeably, it is necessary to perform smoothing of the frequency curve (Clarisse et al. 2002). When islands are used, as described above, it is often the case of simply averaging the plateau region (Wang et al. 2003). Other smoothing functions may be used to varying effect.

2.6 Note Assignment

Once frequency has been reliably estimated, it necessary to determine which note on the musical staff is represented by the note's pitch. Additionally, the duration of this note must be detected and converted to a note type, such as quarter, eighth, half, or whole note.

Note assignment presents several difficulties particular to singing. For instance, it is often the case, especially when the singer is not accompanied, that the sung pitch is some cents off the exact frequency of the intended note. Very few singers have the ability to sing exactly the pitch associated with the notation, called "absolute pitch" (Haus & Pollastri 2001), and it is often the case that their "internal" tuning maybe somewhat stretched or offset simply because this is more pleasing to the ear. Additionally, in certain cases, such as QBH, it is very likely that a system may need to deal with untrained and poor singers.

The simplest method for note assignment, called Round MIDI, simply scales the frequency to the MIDI note number scale, and rounds to the nearest value. McNab et al. proposed a technique which takes into account the singer's "internal scale", by keeping track of relative differences between previous notes (McNab et al. 1995). Following this, Haus and Pollastri proposed an improvement which instead utilizes a constant offset for frequency values (Haus & Pollastri 2001). Wang et al. later proposed an autoregressive algorithm called Adaptive Round Semitones (ARS) for dynamically tracking relative differences between notes, providing greater accuracy for poor singers (Wang et al. 2003). In the same year, Timo Viitaniemi proposed another probabilistic pitch tracking model utilising the parallel combination of a Hidden Markov Model (HMM) and a musicological model (Viitaniemi 2003). This was followed by Ryynanen and Kapluri, who created a similar system utilizing additional features such as voicing and accent in the probabilistic model (Ryynänen & Klapur 2004) (Ryynänen 2004). They claim a greater than 90% success rate, which is a marked improvement over the 60% success rates previously reported (Clarisse et al. 2002).

2.7 Beats and Bars Detection

In order to correctly represent musical notation, the duration of notes in combination with the detected tempo can be used to estimate the correct locations of beats and bars. Beat detection can be quite a difficult topic in its own right and is mostly out of the scope of this summary. However, if note durations are correctly determined, bar locations can be automatically generated using notation engraving tools such as LilyPond (Nienhuys et al. 2006).

3 Sources of Error

Errors in transcription may have one of several sources. The pitch detection may be erroneous due to various reasons, such as low signal-to-noise ratio, the presence of polyphony, or the failure to correctly remove unvoiced sounds. Vibrato may cause many pitch detection problems, though it can usually be solved by using an appropriate amount of smoothing. However, this increases the minimum note duration. Glissando and other dynamics may cause problems as well, since pitch is changing while the segmentation may not be properly detected.

Segmentation errors may arise due to background noise or low onset attack. The most common cause of error, however, is failure to correct for relatively-pitched singing scales (Haus & Pollastri 2001).

4 Evaluation Methods

Since various algorithms exist for many of the steps mentioned, it is important to be able to compare them to determine which methods provide the lowest error rates. This usually involves comparing the output of a transcription system with the manual transcription done by a human expert (Weihs & Ligges 2003). Sometimes the system output is compared to the music's original score, however it can be argued that this method would convolute the singer's skills at following the score with the system's ability to transcribe what was actually sung. On the other hand, one might equally argue that a system performs better if it is able to make up for poor singing skills and deduce the original score from what was sung. The ARS system, for example, attempts to do just that.

Transcription systems are generally rated by the number of missed or inserted notes, as well as the number of wrongly classified notes (Clarisse et al. 2002). Correct detection of note duration and rests is also important, but not always included.

Often, evaluations will also compare the system's results against the same song sung with words and with single syllables (McNab et al. 1995). This can help to separate errors due to segmentation from errors due to pitch detection and note assignment.

References

Clarisse, L., J. Martens, M. Lesaffre, B. Baets, H. Meyer, and M. Leman. 2002. An auditory model based transcriber of singing sequences. In M. Fingerhut (Ed.), *Proceedings of the Third International Conference on Music Information Retrieval: ISMIR 2002*, Paris, France. 116–23. IRCAM - Centre Pompidou.

Ghias, A., J. Logan, D. Chamberlin, and B. Smith. 1995. Query by humming: musical information retrieval in an audio database. In *MULTIMEDIA '95: Proceedings of the third ACM international conference on Multimedia*, New York, NY, USA. 231–6. ACM Press.

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)*.

Jang, J., J.-C. Chen, and M.-Y. Gao. 2000. A query-by-singing system based on dynamic programming. In *International Workshop on Intelligent Systems Resolutions (8th Bellman Continuum)*.

Kageyama, T. and Y. Takashima. 1994. A melody retrieval method with hummed melody (language: Japanese). In *Transactions of the Institute of Electronics, Information and Communication Engineers*, D-II, J77D-II(8). 1543–51.

Kosugi, N., Y. Nishihara, T. Sakata, M. Yamamuro., and K. Kushima. 2000. A practical query-by-humming system for a large music database. *ACM Multimedia*.

McNab, R., L. Smith, and I. Witten. 1995. Signal processing for melody transcription. *Working paper 95/22, Dept. of Computer Science, University of Waikota*.

Nienhuys et al. 2006. LilyPond (Software). Available online: `http://lilypond.org`

Nishimura, T. et al. 2001. Music signal spotting retrieval by a humming query using start frame feature dependent continuous dynamic programming. In *Proc. ISMIR*.

Ryynänen, M. 2004. Probabilistic modelling of note events in the transcription of monophonic melodies. Master's thesis, Tampere University of Technology.

Ryynänen, M. and A. Klapur. 2004. Modelling of note events for singing transcription. In *Workshop on Statistcal and Perceptual Audio Processing (SAPA)*.

Viitaniemi, T. 2003. Probabilistic models for the transcription of single-voice melodies. Master's thesis, Tampere University of Technology.

Wang, C.-K., R.-Y. Lyu, and Y.-C. Chiang. 2003. A robust singing melody tracker using adaptive round semitones (ARS). In *Proceedings of 3rd International Symposium on Image and Signal Processing and Analysis (ISPA03)*. 18–20.

Weihs, C. and U. Ligges. 2003. Automatic transcription of singing performances. In *Bulletin of the International Statistical Institute, 54th Session*, Volume LX. 507–10.