A Dynamic Modelling Approach to Music Recognition

Simon Dixon
Department of Computer Science
Flinders University of South Australia
GPO Box 2100, Adelaide SA 5001
email: dixon@cs.flinders.edu.au

Abstract

This paper focusses on some of the more difficult issues involved in creating an automatic transcription system. The initial stages of the project follow traditional approaches based on Fourier analysis, but as these methods are not sufficiently robust to process arbitrary musical data correctly, they are augmented by models of auditory perception derived from auditory scene analysis and dynamic models of the sources. We argue that by using dynamic modelling, it is possible to solve many of the constituent problems of automatic transcription.

Keywords: automatic transcription, music recognition

1 Introduction

Most approaches to machine transcription suffer from brittleness — they have a steep degradation in performance under non-ideal conditions, such as the minor variations and imperfections in tempo, rhythm and frequency which commonly occur in musical performances. This problem has been noticed in computer implementations of both frequency tracking and beat tracking algorithms, and is cited as one of the major weaknesses of beat tracking programs [Dannenberg, 1991]. For example, if at some stage of the music the program gets out of synchronisation, it is difficult to rectify the problem and bring the program back into step with the music. A similar problem was found in frequency tracking [Dixon, 1996], where music played at about half a semitone sharper than concert pitch caused poor note identification. In this case the frequency analyser was out of synchronisation with the music.

In contrast, the human perceptual system is robust; it is capable of recognising musical information under extremely poor conditions. Although it is quite easy to obtain better performance from computer implementations by carefully setting and adjusting parameters, this defeats the goal of automatic transcription. The dynamic modelling approach emulates human perception in the sense that it is self-adjusting; that is, it "tunes in" to the features of the music it is analysing. This approach extends ideas appearing in previous note identification systems [Chafe et al., 1985] of using context and the expectations derived from the context to disambiguate difficult passages of music.

After developing a dynamic model of the musical sources, the system uses the model to perform its analysis. This methodology can be applied multiple times, extracting different features of the music each time. In this way, the system automatically adapts to different types of instruments, instead of requiring manual setting of parameters.

This approach has been applied successfully in the frequency domain to provide automatic tuning of the system. In current work, the same approach is being used in the time domain to perform beat tracking. The next stage of the system will apply dynamic modelling to the amplitude envelopes and spectral envelopes of instruments, to achieve a greater degree of discrimination of fundamentals from harmonics and then also to perform part separation.

In this paper, we discuss dynamic modelling as it applies to automatic tuning, tempo recognition and synchronisation, analysis of spectral envelopes (formants) and analysis of amplitude envelopes. For a fuller description of the underlying system, readers are referred to [Dixon, 1996].

2 Automatic Tuning

For the vast majority of people, music recognition relies almost entirely on the relationships between frequencies of notes rather than the precise frequencies themselves. For example, a shift of several percent in all of the frequencies in a piece of music does not produce a significant change in the perception of the music, whereas a similar change in a single note will be immediately noticed as being out of tune. Even for those who have absolute pitch, the frequency offset may be noticed, but the relationships between notes are still recognised as being the same as they were without the offset.

Computer implementations of frequency tracking do not necessarily share the same robustness. In a prototype system [Dixon, 1996], it was found that the spectral peak detection algorithm suffered a considerable degradation in performance as the relative tuning shifted by up to half a semitone. This was due to the fact that the system was tuned to recognise the even-tempered semitones of western music, so that frequencies were quantized to the nearest semitone. The system lacked the human ability to "tune in" to the initial frequencies.

To create a model of the tuning of the source, the first step was to select an initial reference frequency ($A_4 = 440\,\mathrm{Hz}$), and construct an even-tempered chromatic scale from this reference. The next step was to measure the deviation of the frequency of each note from its expected value (the nearest note in the chromatic scale), and then adjust the reference frequency towards reducing the error, and reconstruct the chromatic scale. These steps were iterated until the reference frequency converged on a reasonably stable value.

More precisely, starting from the given reference frequency, we construct an even-tempered chromatic scale. Then for each spectral peak, the precise frequency of the peak is estimated by calculating the phase change in the frequency bin over a single sample. (This is the frequency estimation method of the phase vocoder [Flanagan and Golden, 1966, Dolson, 1986].) The frequency is then matched to the nearest note in the chromatic scale, and the deviation from the expected value is calculated.

The average deviation could be used to generate a new reference frequency, but due to the large variability in frequencies detected from natural sounds, and the presence of many unwanted artefacts in the Fourier transform, the algorithm is not stable if the new value is computed in this way. Similarly, using the average deviation suffers too much influence from outlying values, and the average does not converge to a stable value. Thus any error close to half a semitone is discarded, as it is not possible to determine the direction of the error. Then, to reach a stable estimate for the reference frequency, the previous reference frequency is adjusted slightly by the error term, by calculating a geometric mean of the previous and present estimates of the reference frequency, heavily weighted towards the previous value. In this way, the effect of erroneous values is greatly reduced, and the corrections gradually bring the tuning in line with the correct value. Although the convergence is not absolute (it continues to vary within a range of about 0.5%, which is less than the variance of the frequency estimates of individual notes), it provides a sufficiently accurate reference for spectral peak detection.

Having obtained the corrected reference frequency, a new chromatic scale is generated, and the process is iterated for the next sample. It is important not to use the same sample for the iterations, as the notes within the sample may not be representative of the tuning of the piece. This could occur, for example, if the sample contained a transition between two stable notes, or the onset of a note from a percussive instrument such as a guitar or piano, which tends to be slightly sharper than the same note as it decays.

The automatic tuning was tested on several pieces of solo guitar music, and it was found that the system converged on a value within the first few seconds of music, even when the initial error was maximal (half a semitone). Faster convergence can be obtained if required, but slower convergence is preferred, as it is more robust in noisy data.

As it stands, the system calibrates itself at the beginning of a piece, and then matches all notes to this calibrated scale. This is sufficient for music in which the tuning is stable over the duration of the piece, but it is possible that it may be necessary to recalibrate periodically; for example, unaccompanied singing may drift in pitch by a substantial amount during a piece.

3 Spectral Envelopes and Amplitude Envelopes

There is a considerable amount of information available in musical data to disambiguate the parts of music which appear to have more than one possible transcription. For example, the problem of recognising harmonics as being part of a complex tone, which are not notated in the printed score, can be solved much more accurately if the context of the surrounding notes is taken into account. The spectral composition of other notes gives clues to the resonances of the instrument and of the room in which the recording was made, and if these factors are modelled, it becomes possible to predict the spectral composition of other notes, and thus differentiate between various possible decompositions of chords containing complex tones.

The approach suggested in [Tanguiane, 1993] which reduces spectra to boolean values (each component is either present or absent) is naive, as it requires detection thresholds to be sufficiently well-tuned to give highly accurate spectra. Experiments involving solo guitar music show that this is not possible, and a range of acoustic clues must be combined to verify or deny the presence of a component. Clearly the difficulty of this task is magnified with multiple instruments, so it does not appear to be realistic to obtain accurate boolean spectra.

Amplitude envelopes also give clues about the decomposition of multiple notes, as shown by research in auditory scene analysis [Bregman, 1990, Brown and Cooke, 1994]. This work cites the Gestalt principle of common fate to suggest that if partials have a similar amplitude envelope, such as beginning and ending together, or having a similar and parallel pattern of decay, then they may have come from the same source. But partials do not decay at the same rate, so the pattern of *change* in the spectral envelope may in fact be more useful in providing information about the source of the sound.

The current system does not yet make use of dynamic models in this area. The auditory scene analysis principles provide a static model which is used for matching of harmonics to fundamentals.

4 Tempo Recognition and Synchronisation

Another necessary constituent of an automatic transcription system is a procedure which can determine the underlying rhythmic structure of a piece of music. Unlike commercial notation and sequencing software, which requires the speed and time signature to be specified explicitly, we assume that such information is not available from the user and must be generated automatically.

A dynamic modelling approach can also be used at this level to automatically extract such features as the bar and phrase structure of musical pieces. This is the approach advocated in [Rosenthal, 1992], and is similar to the one proposed in [Tanguiane, 1993].

In the implementation, rhythm is detected by a two-stage process. A small window size is used to create a high-resolution time-line of events. The current rhythm detection stage relies on the sudden onset of percussive instruments, and is not suitable for analysing music containing other instruments. The percussive onset gives a peak across a wide range of frequencies, which is easily detected, and signals the beginning of a note, which can then be identified by examining the spectrum from a large time window, which gives high resolution in the frequency domain. This gives an accurate mapping to a MIDI-like representation, but lacks the underlying rhythmic structure. To obtain this, statistical correlation methods are used to determine the best estimation of a beat, and to synchronise the beat to the sample counts.

5 Conclusions

The philosophy behind this project is to develop an unsupervised transcription system, that is, one which requires little or no adjustment of parameters or a priori information about the piece of music being analysed. Therefore, it is desirable to make as few assumptions as possible about the nature of the music and the instruments.

To achieve this aim, it is necessary to develop models of the sound sources on the fly, and then use these models to further analyse the music. This iteration makes it difficult to use the system for real-time analysis, but provides a degree of refinement which would not otherwise be possible.

Although much of the work still remains to be done, the initial results are encouraging, and it is planned to continue this line of research to the completion of an automatic music transcription system.

Acknowledgement

The author wishes to thank David Powers for many useful ideas and discussions.

References

[Bregman, 1990] Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organisation of Sound*. Bradford, MIT Press.

[Brown and Cooke, 1994] Brown, G. and Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336.

[Chafe et al., 1985] Chafe, C., Jaffe, D., Kashima, K., Mont-Reynaud, B., and Smith, J. (1985). Techniques for note identification in polyphonic music. In *Proceedings of the International Computer Music Conference*. Computer Music Association, San Francisco CA.

[Dannenberg, 1991] Dannenberg, R. (1991). Recent work in real-time music understanding by computer. *Proceedings of the International Symposium on Music, Language, Speech and Brain*.

[Dixon, 1996] Dixon, S. (1996). Multiphonic note identification. *Australian Computer Science Communications*, 17(1).

[Dolson, 1986] Dolson, M. (1986). The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27.

[Flanagan and Golden, 1966] Flanagan, J. and Golden, R. (1966). Phase vocoder. *Bell System Technical Journal*, 45:1493–1509.

[Rosenthal, 1992] Rosenthal, D. (1992). Emulation of human rhythm perception. *Computer Music Journal*, 16(1):64–76.

[Tanguiane, 1993] Tanguiane, A. (1993). *Artificial Perception and Music Recognition*. Springer-Verlag.