On the Automatic Transcription System of Kashino and Tanaka

Doug Van Nort
Signal Processing and Control Laboratory
Faculty of Music - McGill University
555 Sherbrooke St. West
Montreal, QC Canada H3A 1E3
doug@music.mcgill.ca

ABSTRACT

The goal of this short paper is to present the polyphonic automatic transcription system of Kunio Kashino and Hidehiko Tanaka of Tokyo University, as presented in the proceedings of the 1993 International Computer Music Conference.

1. INTRODUCTION

The ability to automatically transcribe polyphonic music has long been a goal for many researchers in the field of computer music. The process of separating an audio signal into constituent parts (unique sound sources) and then mapping the frequency content of each into symbolic data has proven to be quite difficult, and as of yet is an unsolved problem.

There have been several advances in regards to this problem however, with one of those being the transcription system of [Kashino and Tanaka 1993][1]. This particular system was groundbreaking in that it was the first such approach to utilize rules that were based on human perception. In particular, the notion of segmentation and fusion of auditory events was introduced, ideas that were informed by the work in *auditory scene analysis*. This area, simply put, deals with the human processing of auditory streams within the scope of one's sound environment. The pioneering work was done by Al Bregman, and is well documented in [Bregman 1990].

The transcription system described herein uses perceptual based rules to group frequency components that belong to the same note via bottom-up processing. This is achieved after an analysis and extraction of frequency components. After these components have been grouped at the note level, a higher level processing occurs that is successful at assembling these notes into chordal arrangements. We first describe the formal approach to sound seperation taken by the authors, and then describe the aspects of the system that we have just briefly introduced.

2. PERCEPTUAL SOURCE SEPERATION

Within the context of computational auditory scene analysis, Kashino introduces the problem of *Perceptual Sound Source Seperation*:

[1]Another important advancement that utilizes Bayesian networks can be found in [Kashino et al. 1995], but will not be discussed in this short paper.

Suppose we have mono sound signal $S(t)$ that is a mixture of m signals $\{S_1(t), \ldots, S_m\}$ and which can be represented by some parameter set $P = \{p_1(t), \ldots, p_n(t)\}$. Find some collection of parameters $P_i \subset P$ that represent signal S_i for each $i = 1 \ldots m$.

In the case of this trasciption system our set of parameters are frequency components

$$F(t) = \{F_1(t), \ldots, F_L(t)\}$$

where

$$F_j(t) = \{p_j(t),_j(t), \psi_j(t)\}$$

and $p_j(t),_j(t), \psi_j(t)$ represent power frequency and bandwidth of a spectral peak, respectively. Given this, the problem for this particular system then becomes twofold:

1. Extract frequency components $F_j(t)$ from a spectrographic analysis.

2. Cluster these components into note and finally chord structures based on perceptual rules. These rules were determined by work in psychoacoustics as presented in [Bregman 1990] and [Moore and Glasberg 1986], among others.

We now describe the different parts of the system, beginning with the extraction of frequency components.

3. FREQUENCY COMPONENT EXTRACTION

Prior to extraction, the mono input signal is sampled at 16bit/48khz. The analysis is performed by a bank of 2nd order IIR bandpass filters with log frequency scale. After this, spectral peaks are found by a method devised by the authors and described as "pinching planes." This approach is a regression plane analysis calculated by a least squares fitting. In other words two planes are fixed in the x and y (time and frequency) direction on either side of a peak, and the z (power) parameter is varied so as to minimize the sum of least squares distance between the planes and the signal. From the general expression of the planes $z = ax + by + c$ one can find the spectral bandwidth of a peak by computing the cross product of each plane's normal vector, as well as the frequency and power of the peak. The author's introduce an effective peak threshold parameter θ_e that determines the minimum power by which the system accepts/rejects potential spectral peaks, with the "winners"

going through the subsequent extraction via the described method. This threshold, along with the x and y "window size" of the pinching planes, are free parameters. As we will see there are several others that are introduced and which may greatly determine the effectiveness of the system.

4. CLUSTERING OF COMPONENTS FOR SOURCE IDENTIFICATION

4.1 Grouping Strategy

The next aspect of the system deals with the clustering of frequency components into groups that humans tend to hear as one. As such, perceptual rules based on human auditory functioning are introduced. Specifically, components are clustered based on:

1. Harmonic Mistuning
2. Onset Asynchrony.

The evaluation is in terms of probability of seperation. The probability functions used were approximations of psychoacoustic experiments conducted by the authors. The experimental results provided thresholds by which to measure similarity. For example, the threshold determined for harmonic mistuning was 2.6%, meaning that if two frequencies were inharmonic by more than 2.6% then they have probability 1 of belonging to different notes. Given these probability functions c_1 and c_2, a distance measure was created for the space of frequency components as follows:

$$m = 1 - (1 - c_1)(1 - c_2)$$

This measure integrates both probabilities, and is based on Dempster's law of combination. The actual process of clustering is achieved by pairwise assessment of the distance (relative m values) between frequency components. If m is found to be greater than some threshold parameter θ_m, then components are said to belong to different clusters. The threshold parameter is chosen by the user.

4.2 Source Identification

Once components are grouped into clusters that represent singular sound events (based on human auditory functioning), the next step is to identify the source of each and to group sounds based on similar sources. This part of the system is a higher level approach that looks at global characteristics of each cluster. Based on the parameters employed in this method, it is intended for groups that contain a single note.

Once again a distance measure is constructed; in this instance the structure of each sound cluster is compared based on physical characteristics, rather than perceptual. Specifically, the constructed measure is defined as

$$D = w_1 f_p + w_2 f_q + w_3 t_a + w_4 t_s$$

where

f_p = Peak power ratio of second harmonic to fundamental
f_q = Peak power ratio of third harmonic to fundamental
t_a= attack time
t_s = sustain time.

The w_k are user determined coefficients, allowing such a user to weigh one or another of these parameters more heavily. Seemingly, this would have great effect on the algorithms ability to properly group sources. In terms of the chosen physical parameters, it seems that only certain sources would have these as salient characteristics. Perhaps this is one of the stronger factors that contribute to the systems apparent inability to recognize all types of sources. Indeed, the tests described in [Kashino et al. 1998] were restricted to several instruments.

5. TONE MODEL BASED PROCESSING

In the event that a sound cluster contains only a single note, the process described in the preceding section is sufficient for accurate transcription (varying effectiveness based on type of input). In the case of simultaneously sounding notes, further processing is required. This is accomplished in this work via tone model based processing. The initial implementation of this technique generates tone models, which are two dimensional matrices with rows corresponding to different frequency components and columns to time steps. At each matrix location is stored a two dimensional vector of normalized power and frequency. Thus, the matrix represents time-varying frequency and power values for each component. These tone models are generated and compared to the sound clusters based on "mixture hypotheses." These hypotheses set the following standardizations and rules for comparison:

•Limit the number of simultaneous notes in a model to three
•adjust amplitude of the model so that the lowest frequency component is the same as that of the sound cluster, and
•Time shift the tone model from -20 ms to 20 ms, and compare against sound clusters.

The metric for comparison is defined as

$$D_t = \sum_{i=1}^{F} \sum_{j=1}^{T} |p_{ij} - p'_{ij}| \cdot f'_{ij}$$

where F is the number of frequency components in the sound cluster, T is the number of time steps in the tone model, p_{ij} is the power component at the corresponding matrix location for the tone model, and p'_{ij}, f'_{ij} are similar power and frequency values for the sound cluster. Therefore, the size of the matrices for comparison are determined by the frequency components of a sound cluster, and the time discretization introduced by the time steps of the tone model. The two matrices (tone model and discretized sound cluster) are compared at several relative time shifts, and the model that minimizes the metric D_t is chosen as the chordal representation.

This tone based processing implementation is said to work well for chord recognition, while being restricted to three note polyphony. This method requires user registration of tone models based on knowledge and assumption of what the structure might be (as well as knowledge of existing instrument models). In order to avoid the need for such a registration process, the authors implement an automatic tone model based processing scheme. The difference between this and the above is that the metric D_t is modified such that the number of "hits" are taken into account. That is, the number of frequency indices that overlap "well" between cluster and model give a measure of correspondence

between the two which allows for the model to automatically generate new frequency components or to shift existing ones. When the metric falls below a specified threshold, the model is kept as being the correct representation.

6. CONCLUSIONS

The technique of polyphonic transcription and source seperation that has been described in this paper was an early example of transcription systems defined by perceptual rules. In this sense it can be regarded as noteworthy in a historical sense. Beyond this, however, the implementation shows promise based on the results given in the 1993 report, thereby providing validity for such perception based systems. In the reported results, both automatic and registration based tone modeling studies gave high recognition rates for two note polyphony, with registration based modeling further giving high rates for 3 note polyphony. As with all polyphonic transcription systems to date, there are tradeoffs and limitations, however. The maximum polyphony is 3, the number of free parameters means that plenty of user adjustment is required, and further the psychoacoustic parameters of section 4.1 and "global cluster" variables of section 4.2 seem to imply certain instrument types. Thus, we can conclude that this system is quite effective for particular listening contexts, but that there is still a lot of work to be done to achieve a general polyphonic seperation and symbolic transcription system.

7. REFERENCES

Bregman, A. (1990). *Auditory Scene Analysis.* Cambridge, Massachusetts: M.I.T. Press.

Kashino, K., K. Nakadai, T. Kinoshita, and H. Tanaka (1995). Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism. In *IJCAI*, pp. 158–164.

Kashino, K., K. Nakadai, T. Kinoshita, and H. Tanaka (1998). Application of the Bayesian Probability Network to Music Scene Analysis. In *Computational auditory scene analysis*, pp. 21–26. Lawrence Erlbaum Associates.

Kashino, K. and H. Tanaka (1993). A Sound Source Separation System with the Ability of Automatic Tone Modeling. In *Proc. of the 1993 ICMC*, pp. 248–255.

Moore, B. and B. Glasberg (1986). Threshold for Hearing Mistuned Partials as Separate Tones in Harmonic Complexes. *J. Acoust. Soc. Am. 80*(2), 479–483.