HARMONIC GROUPING FOR COMPUTATIONAL AUDITORY SCENE ANALYSIS

Kathy Melih and Ruben Gonzalez

Griffith University

ABSTRACT

Blind source separation (BSS) have found considerable interest in diverse applications. However, there are some conditions where a more perceptually motivated approach is required. In these cases, computational auditory scene analysis (CASA) provides the solution. Harmonic relationships are particularly important for forming acoustical source separation and are hence an important area of research. This paper presents a novel method for performing harmonic analysis and grouping.

1 INTRODUCTION

Blind source separation (BSS) techniques have been applied in fields as diverse as medicine and radar signal processing. Under controlled conditions, BSS has proven useful to improve the performance of automatic speech recognition systems [1] to separate percussive musical sounds [2]. However, it has been shown that in unconstrained auditory environments, computational auditory scene analysis (CASA) systems offer superior performance.

CASA is the endeavour to produce computational models of the perceptual phenomenon labelled acoustical scene analysis (ASA) by Al Bregman [3]. ASA is the perceptual process by which we to make sense of the cacophonous auditory world given the noisy signal that enters our ears. The basic principle is described by **Figure 1**.

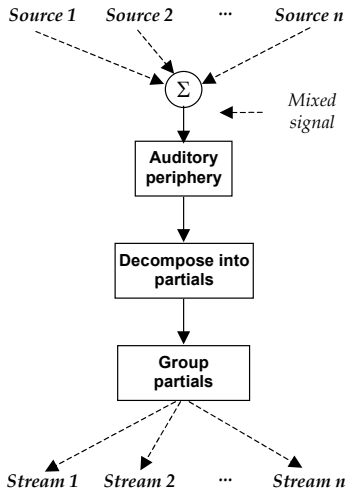

Figure 1: The basic process of ASA

Thus it is apparent that the aims of CASA are somewhat similar to those of BSS. However, there are also a number of distinctions. Firstly, CASA systems approach the problem from the point of view of modelling the human perceptual process to varying degrees of accuracy. In contrast, BSS is a purely analytical process in which the problem is posed as one of deconvolution. Secondly, and most importantly, CASA systems assume no *a priori* knowledge of the nature of the sources nor their number. Finally, BSS systems are generally interested in accurately recovering only one of the sources, as a front end for a speech recognition system, for example, while CASA systems generally aim to accurately describe the entire acoustic 'scene.'

The work reported here contributes to a larger project that aims to decompose audio signals for the purpose of audio information management tasks such as content-based retrieval and browsing. In order to maximise the utility of the system no *a priori* knowledge of the data nor of the nature of queries can be assumed. Thus, the complete picture that CASA systems offer is required.

2 BACKGROUND

CASA systems typically consist of several hierarchical processing stages that model, to varying degrees of biologic accuracy, the processes of the human perceptual system. Roughly speaking all CASA systems consist of four principal stages:

- *Peripheral processing*
  Performs a time-frequency analysis of the incoming audio signal. Typically the transform will have non-uniform quantisation matching that of the low-level perceptual system which behaves somewhat like a constant-Q filter bank. Only the amplitude *peaks* are retained.
- *Low-level feature extraction*
  Organises the peaks into continuous *tracks* (or partials) through the time-frequency-amplitude space. Tracks are monotonic along the time axis.
- *Mid-level grouping*
  The tracks are organised into *groups* that form the building blocks of higher level streams. The tracks in these groups will generally have a harmonic relationship and similar contours.

- *High-level streaming*
  Application dependent organisation of groups into *streams*. Generally a stream is what would be perceived as a single sound 'object'. In the perceptual system, stream formation is highly context dependent and this is also reflected in CASA systems.

At the mid-level grouping stage, it is known that the human perceptual system considers a number of partial characteristics in parallel to perform the grouping [3]. Harmonicity is held to be the most important characteristic for grouping in the perceptual system [4] and is not surprisingly most often employed in CASA systems. This paper presents a new algorithm to perform harmonic group formation.

The utility of harmonic grouping of time-frequency trajectories goes beyond CASA systems. Sinusoidal coding schemes also stand to benefit. Forming harmonic track groupings is fundamental to the coding gain of codecs such as the harmonic and individual lines plus noise (HILN) system found in the MPEG4 standard [5].

3 PREVIOUS WORK

Two basic approaches for harmonicity-based grouping exist. The first involves either explicitly or implicitly estimating the fundamental frequency of each of the sources present in the signal using various standard and non-standard pitch estimation techniques and using these to generate a "harmonic sieve" [4] for each source. A simplified harmonic sieve is illustrated in **Figure 2**.

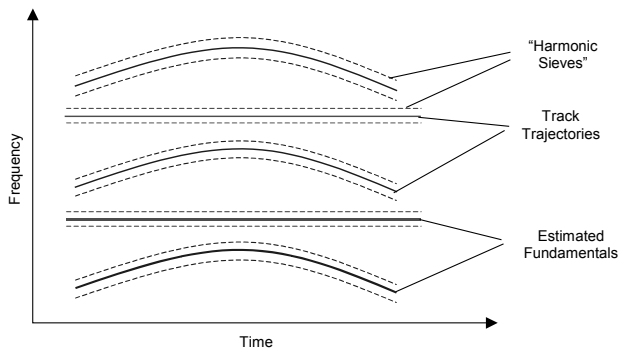

Figure 2: Simplified harmonic sieve

The chief problem associated with the existing harmonic sieve approaches is that the pitch estimation techniques employed generally assume that only one source exists in the given time-frame. Hence, accurate pitch estimates can be difficult to obtain. The method proposed in this paper applies a theory of pitch perception to determine all likely fundamental frequencies in a given analysis frame.

The second method exploits the fact that if two frequencies are harmonically related, their ratio will be equivalent to the ratio of two small positive integers, (a, b):

$$\frac{f_i}{f_j} = \frac{a}{b} \qquad (1)$$

Determining whether two tracks are harmonically related is then simply a matter of determining whether their frequencies satisfy the above condition within acceptable error bounds.

One example of the second technique was proposed by Virtanen and Klapuri [6] who used a look-up table to find integer ratio that was closest to the ratio of frequencies and then calculating the error between the two. The range of allowable values for the integers a and b were restricted such that the fundamental frequency could not be below the minimum frequency in the data set. Parenthetically, there is a disadvantage in this choice of minimum fundamental frequency as it is possible that the perceived fundamental frequency may not actually be present in the data [7].

4 BREGMAN'S PITCH PERCEPTION THEORY

Bregman [3] observed that the differences between frequencies play an important role in pitch perception. In the simplest case, we have a single harmonic series of frequencies:

$$f_1 = f_0,\ f_2 = 2f_0,\ f_3 = 3f_0,\ f_4 = 4f_0,\ \ldots \qquad (2)$$

It should be obvious that the difference between each pair of adjacent frequencies is equal to the fundamental frequency, f_0. That is, given H harmonics in the series:

$$(f_2 - f_1) = (f_3 - f_2) = \ldots\ (f_H - f_{H-1}) = f_0 \qquad (3)$$

Further, the difference between any two non-adjacent frequencies will be some integer multiple of the fundamental. For example,

$$f_3 - f_1 = 3f_0 - f_0 = 2f_0. \qquad (4)$$

If the data were perfectly noise free, and we only ever dealt with a single source, determining the fundamental frequency would thus be a very simple matter of determining the difference between any two adjacent frequencies in the frame. In practice, however, to deal with both noisy data and multiple source separation, we must use a histogram to record the difference between all pair-wise combinations of peak frequencies available. This will obviously lead to a peak at each of the fundamental frequencies present. Equation 4, shows that more accuracy can be achieved recording fractional differences as well.

5 HARMONIC GROUPING

5.1 Group Formation

Harmonic grouping involves determining the fundamental track (or tracks in the case of a mixed source) and then using this as the basis for a harmonic sieve as illustrated in **Figure 2**. It may be noted that Bregman's pitch estimation procedure outlined in the previous section assumes that the signals (tracks) represent steady state tones of a single

frequency. In practice, this will rarely be the case as even the simplified example in **Figure 2** illustrates.

One possibility is to pick a single representative frequency for each track. The major disadvantage of this approach is that no matter how this representative frequency is selected, it is highly likely that two totally unrelated tracks will appear the same as shown in **Figure 3**.

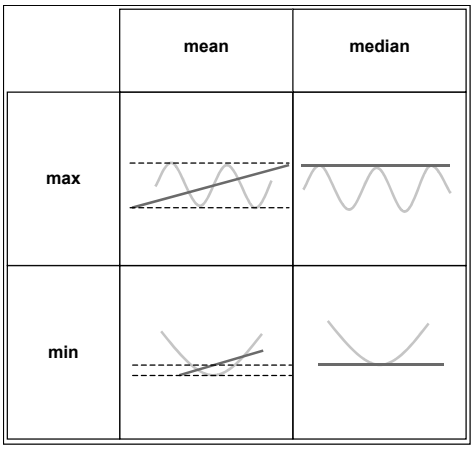

Figure 3: Examples of ambiguity arising from single frequency representation of tracks

To overcome this disadvantage the individual peak frequency values were used to populate a histogram that determines the fundamental frequency estimates as per Bregman's procedure. These estimates were then used to determine the most likely harmonic number of each peak in the corresponding time frame. Finally, the most likely harmonic number of each track was deemed to be the harmonic number assigned to the majority of its peaks.

5.2 Histogram generation

Given that the resolution of the human ear is not fixed and that there is no restriction on the fundamental frequency of real data to be an integer, a conventional fixed-width histogram was inappropriate for the task since it is difficult, if not impossible, for a set of uniformly spaced frequency bins to adequately sort the data such that the fundamental frequencies become apparent.

Hence, a proportionally spaced histogram was developed. The bin centres were derived from the actual data values and the bin width was an empirically determined proportion of the centre frequency. Figure 4 illustrates the important parameters of a histogram bin.

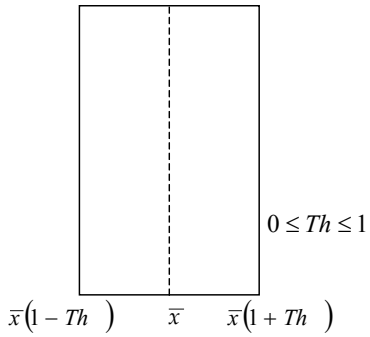

Figure 4: Histogram bin characteristics

The algorithm for generating the histogram is as follows:

1. Place each data value in its own zero-width bin.
2. Sort the bins in ascending order of bin centre.
3. Set n to 0
4. WHILE n <B ,

IF $bin[n] < bin[n+1] \times (1 - Th)$ **(5)**

$$bin[n] = \frac{bin[n] \times cnt[n] + bin[n+1]}{cnt[n] + 1} \quad \textbf{(6)}$$

increment $cnt[n]$

ELSE increment n

END WHILE

where B is the number of bins and $bin[n]$ is the centre frequency of bin n and $cnt[n]$ is the number of entries in bin n.

The peaks in the histogram will then provide the candidate fundamental frequencies, f_{0i}, for $0 \le i \le$ M where M is the number of candidate fundamental frequencies.

5.3 Harmonic Number Determination

Having obtained the fundamental frequency estimates as described above, the most likely harmonic number for each peak frequency, f_n, in the current time frame is obtained as follows:

Find e_{min} such that:

$$e_{min} = \mathrm{MIN}\left\{\left|\mathrm{ROUND}\left\{\frac{f_n}{f_{0i}}\right\} - \frac{f_n}{f_{0i}}\right|\right\} \quad \textbf{(7)}$$

for $0 \le i \le$ M

The harmonic number, h_n is then:

$$h_n = \mathrm{ROUND}\left\{\frac{f_n}{f_{0i_{min}}}\right\} \quad 0 \le n < N \quad \textbf{(8)}$$

where i_{min} is the index corresponding to e_{min} in 7.

6 RESULTS

Two versions of the algorithm were tested: the first populated the histogram with only frequency differences (BREG) while the second also included fractional differences (BREG FRAC). These were tested against an implementation the Virtanen and Klapuri algorithm mentioned previously (V&K); a slight variation, V&K (mod), and a third (original) method that used equation (1) and Euclid's method of continuing fractions to determine a harmonicity metric.

The algorithms were tested over a set 25 files containing various two source combinations of speech, music and artificially generated signals (chirps, tones and FM). To provide a reference for the tests, the files were first grouped manually. Once the algorithms were run, two

performance metrics were calculated by comparing the grouping achieved by the algorithm against the manually grouped reference. The tests were performed over a range of threshold values. This threshold represents the tolerance value in the harmonic sieve.

The first performance metric was a straightforward measure of the number of files for which best possible grouping was achieved. The second was the proportion of miss-grouped tracks with respect to the number of groups found. This metric was calculated as follows:

$$MT = \frac{1}{M_a} \sum_{i=0}^{M_a - 1} MT_{ai} \quad (9)$$

$$MT_{ai} = \frac{(2C_{over} + C_{under} + C_{wrong})}{N_{ai}} \quad (10)$$

$$C_{over} = \begin{cases} N_{ai} - N_{rj} & N_{ai} > N_{rj} \\ 0 & otherwise \end{cases} \quad (11)$$

$$C_{under} = \begin{cases} N_{rj} - N_{ai} & N_{rj} > N_{ai} \\ 0 & otherwise \end{cases} \quad (12)$$

$$C_{wrong} = \frac{N_{ai} - C_{over} - SIZE\{G_{ai} \cap G_{rm}\}}{N_{ai}} \quad (13)$$

$$m \ni SIZE\{G_{rm} \cap G_{ai}\} = MAX\{SIZE\{G_{rj} \cap G_{ai}\}\}$$
$$\forall\ 0 \le j \le N_{rj} \quad (14)$$

The performance of all the algorithms is shown in **Figure 5**. This figure reveals that both original methods outperform the V&K algorithms with the straightforward BREG method performing the best. Given previous discussion, it may be surprising that recording fractional differences would degrade performance. However, this may be accounted for by the fact that recording fractional differences causes peaks to form in the histogram below the actual fundamental frequencies. Because of increasing noise sensitivity at lower fundamental frequencies, this can cause ambiguity and errors in harmonic determination.

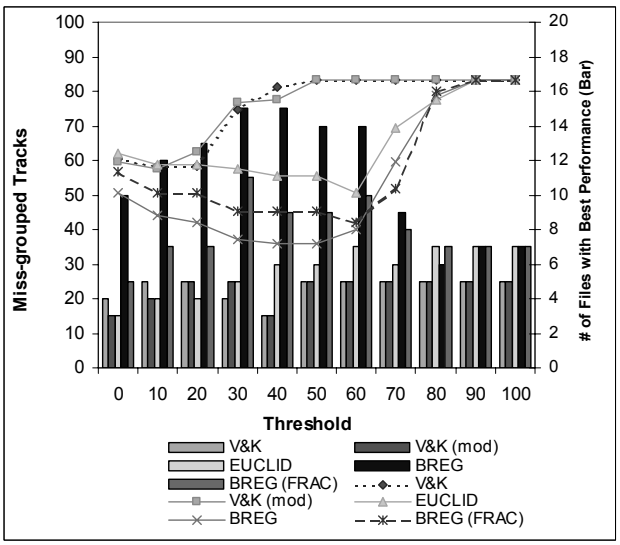

Figure 5: Miss-grouped tracks performance measure

Figure 6 shows an example of the grouping achieved using the Bregman-based algorithm recording only the raw frequency differences in the histogram.

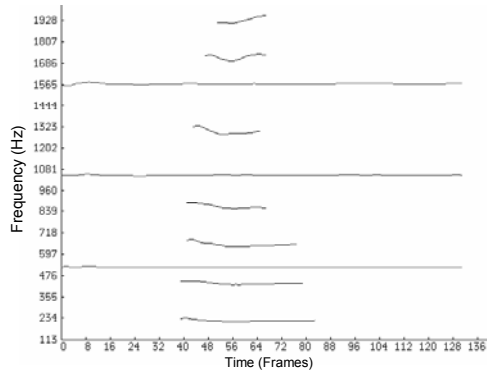

Figure 6: Example of grouping

7 CONCLUSIONS

Harmonic grouping plays an important role in computational auditory scene analysis (CASA) as well as other signal processing applications such as parametric coding. A new algorithm to perform harmonic grouping of tracks in a sinusoidal representation has been presented and has been shown to outperform another recent CASA-based approach.

8 REFERENCES

1 A. Koutras, E. Dermatas and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2000)*, 2000.

2 M. Baeck and U. Zölzer, "Performance Analysis of a Source Separation Algorithm," *Proceedings of DAFX02*, 2000.

3 A. S. Bregman, Auditory Scene Analysis: the Perceptual Organization of Sound, MIT Press, 1990.

4 S. Grossberg, "Pitch-based Streaming in Auditory Perception," Musical Networks: Parallel Distributed Perception and Performance, N. Griffith and P. Todd (eds), Cambridge, MA, MIT Press, 1996.

5 H. Purnhagen, H. Meine, "HILN – The MPEG-4 Parametric Audio Coding Tools," *IEEE International Symposium on Circuits and Systems*, ISCAS2000, Geneva, May 2000.

6 T. Virtanen, A. Klapuri, "Separation of Harmonic Sound Sources Using Sinusoidal Modeling," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* ICASSP 2000.

7 J. F. Schouten, "The Perception of Pitch," *Philips Technical Review*, vol. 5, 286-298, 1940.