

A grouping approach to harmonic complexes

J.D. Krijnders, M.E. Niessen and T. Andringa

{j.d.krijnders,m.e.niessen,t.andringa}@ai.rug.nl Artificial Intelligence, University of Groningen



RUG

Introduction

An harmonic complex is defined as a combination of simultaneous occurring sinusoids with a frequency of n times a fundamental frequency contour ($f_0(t)$):

$$x(t) = \sum_{n=1}^N \sin(2\pi n f_0(t)t) \quad (1)$$

In computational auditory scene analysis (ASA), pitch is often used as the starting point for harmonic grouping[2]. If the pitch estimation fails, so does the grouping. We present a method for grouping without the use of pitch in which a pitch estimate is produced as side effect. The grouping algorithm works on tones extracted from a cochleogram.

Pre processing

The proposed method works in the time-frequency plane; the transformation is performed by a model of the human cochlea. The output of this model is the amplitude of the basilar membrane. The basilar membrane model used is a standard gammachirp:

$$g_{gc}(t) = at^{N-1}e^{-2\pi bB(f_r)t}e^{j(2\pi f_r t + c \log(t))} \quad (2)$$

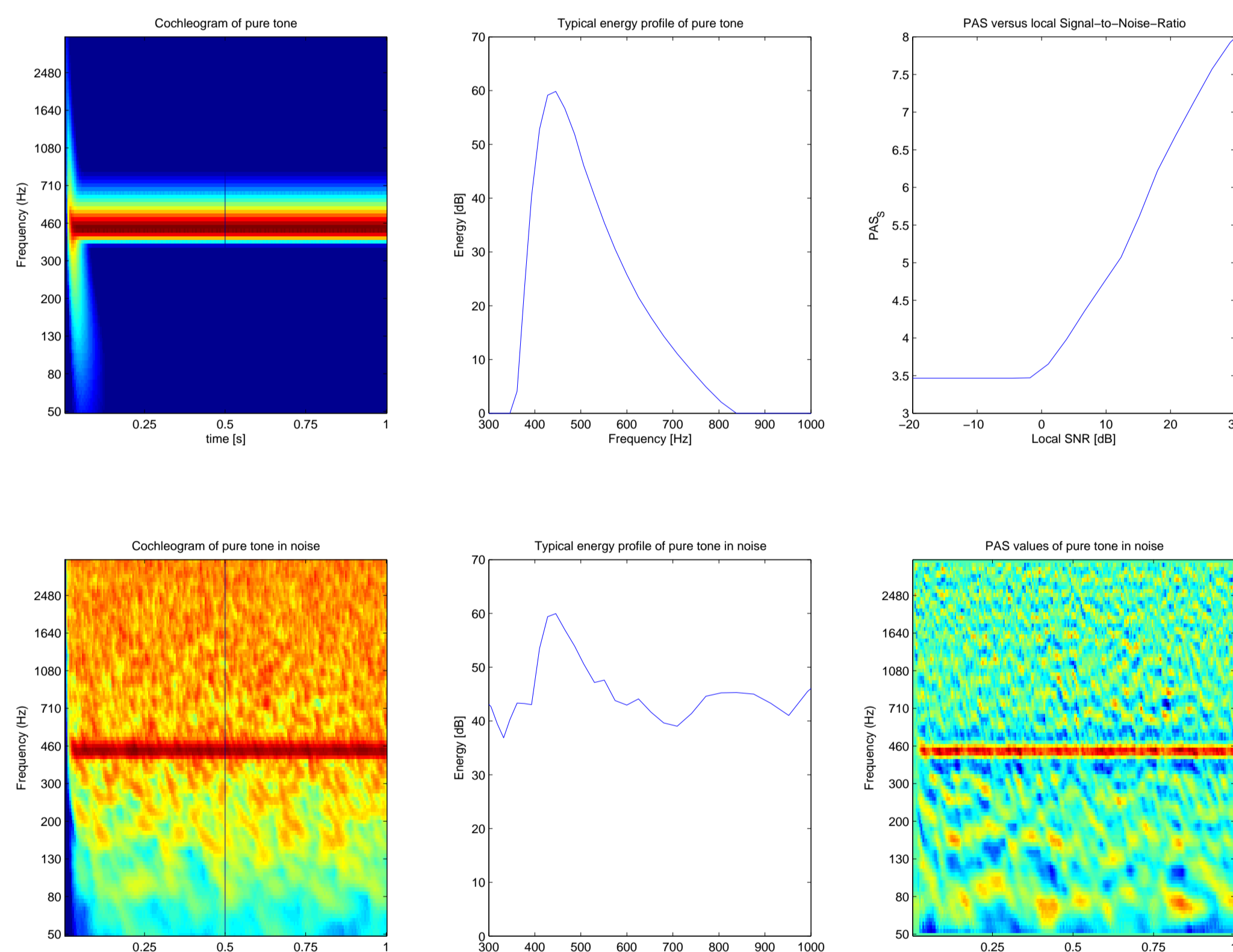
The basilar membrane output is leaky-integrated and downsampled. To compress the dynamic range of the energy spectrum, the energy is scaled to a decibel scale. This time-frequency-energy representation is called a cochleogram and is continuous in the time-frequency plane.

$$E(n, t_A) = E(n, t_A - dt_A)e^{-dt_A/\tau_n} + A(n, t_A) \quad (3)$$

$$E_{dB}(n, t) = 10 \log_{10}(E_A(n, t)) \quad (4)$$

$$(5)$$

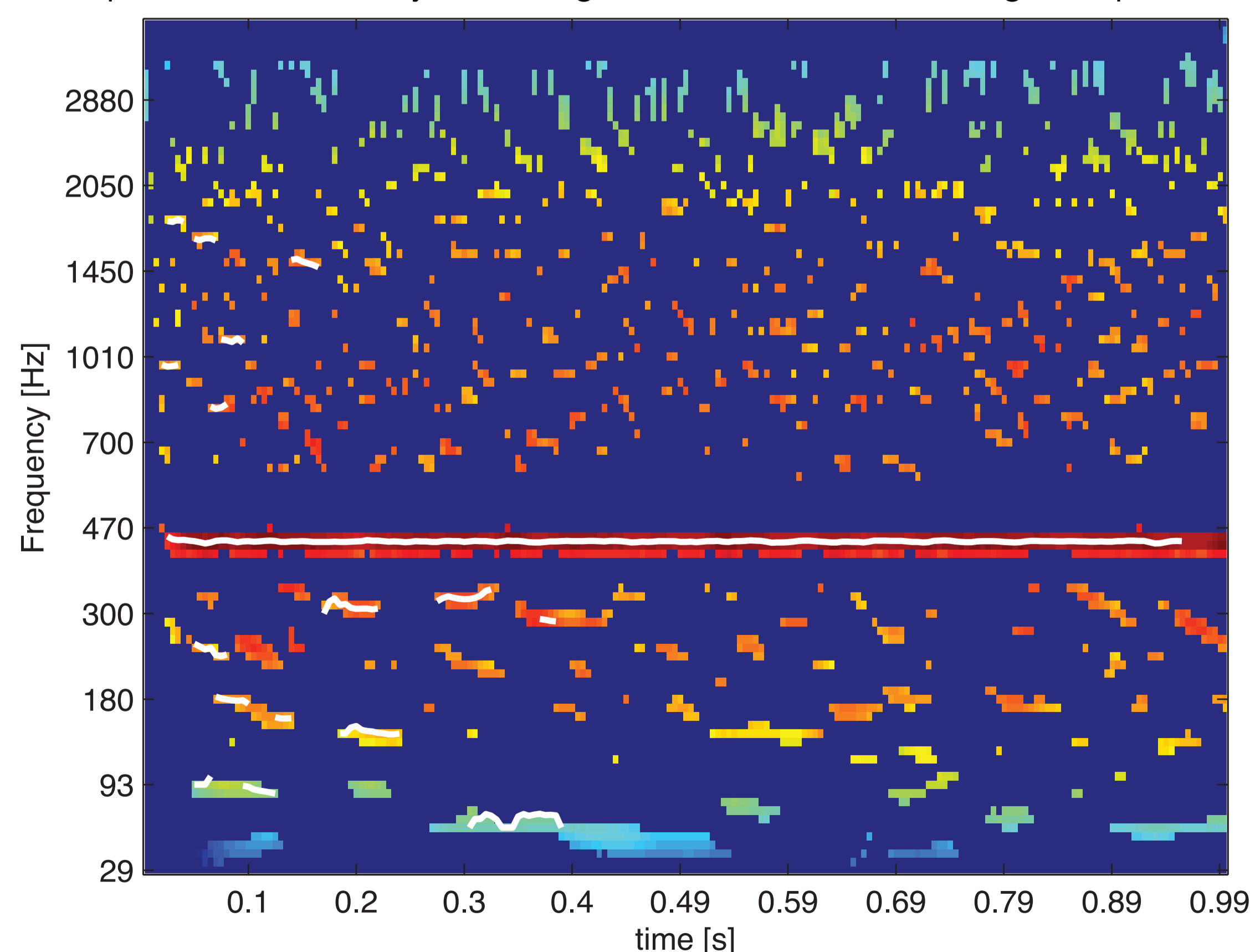
Tones are extracted by filtering the cochleogram with a matching filter based on the ideal tone response of the cochlea. The output of this filter (PAS) is correlates strongly with the local SNR when the signal is sinusoidal.



SNR is defined as the maximum power of the signal minus the mean power of the noise at the same frequency, both in dB.

Signal components

Tonal signal components are formed by connecting all local maxima within a region of positive local SNR:



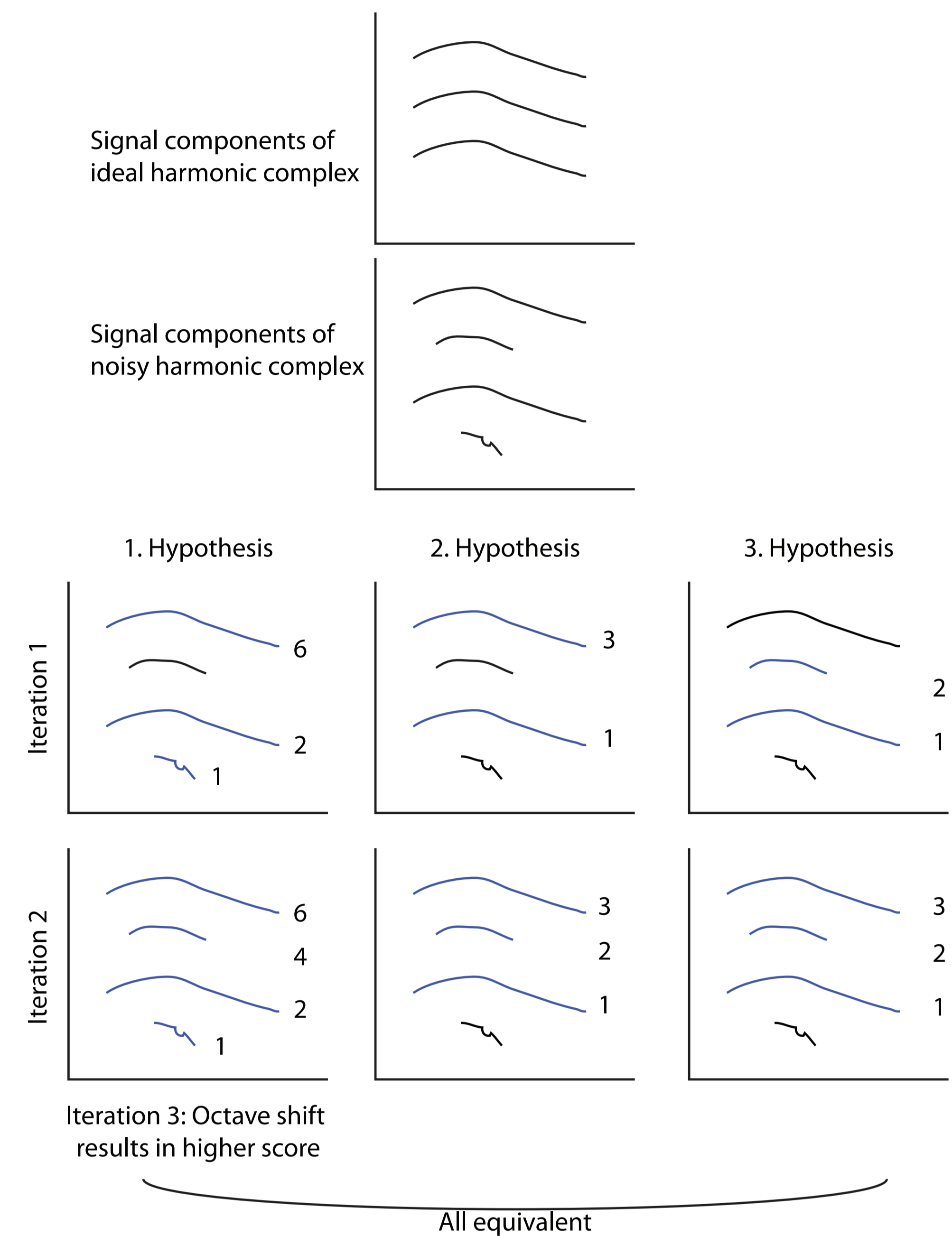
Grouping

First task is to map which signal components occur simultaneously, this step is expensive in off-line processing, but cheap for on-line processing. This step's implementation is non crucial, it should only ensure that enough grouping hypotheses are generated so no complexes will be missed.

Second, an iterative process is used to extend the hypotheses, merge duplicates and score each hypothesis. In each step the following occurs, until all hypotheses are stable.

- Find additional signal components that match the properties of the hypothesis
- See whether the score increases when the hypothesis is shifted one octave up or down
- Merge hypotheses that overlap or are equal

Cartoon example



Scoring

Scoring is done based on

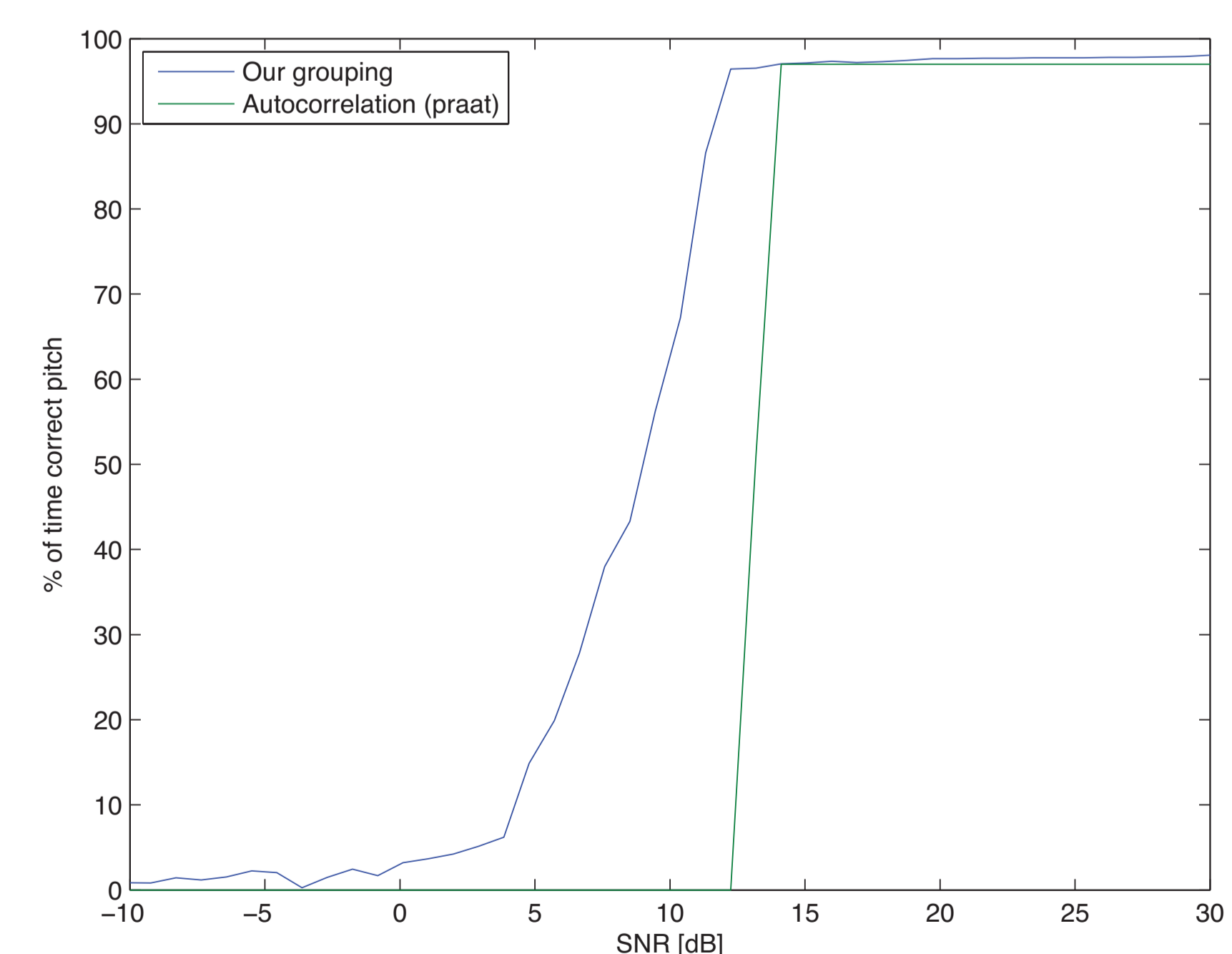
- The number of available harmonics (α)
- The number of subsequent harmonics (β)
- The existence of a signal component at fundamental frequency (δ)
- The average PAS value under the harmonics (γ)
- How well the harmonics fit the properties of the hypothesis (ϵ)

These measure are combined as follows:

$$s = (\alpha * \delta) + \beta + \gamma - \epsilon \quad (6)$$

Results and discussion

Our pitch estimates compared to a standard pitch estimation algorithm, as used by Praat[1].



Discussion

We have shown a method of grouping tonal signal components into harmonic complexes. The method produces pitch as a side product and this pitch estimate is better than the estimate as done by Praat with default settings. Further research includes comparison to other grouping algorithms, multi-pitch performance and inclusion of context information to improve the performance of the algorithm.

[1] Paul Boersma and David Weenink. Praat: doing phonetics by computer (version 4.4.32) [computer program]. Technical report, University of Amsterdam, 2007.

[2] DeLiang Wang and Guy J. Brown. *Computational Auditory Scene Analysis*. John Wiley and Sons, Holoken, NJ, 2006.