The Weft: Auditory Scene Analysis of periodic sounds

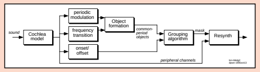
Dan Ellis • International Computer Science Institute, Berkeley CA • <dpwe@icsi.berkeley.edu>

Motivation: Auditory separation of multiple pitches

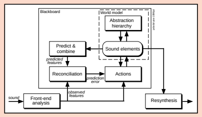
- The '**pitch cue**' (source periodicity) is central to our ability to attend to individual sounds in a mixture [AssmS90]. The signal processing behind this ability is not well understood.
- Psychoacoustic pitch phenomena (missing fundamental etc.) are well modeled by **autocorrelation** of unresolved harmonics in broadening frequency channels [MeddH91].
- As part of a **computational auditory scene analysis** system, we would like to recover the energy in different frequency bands due to each periodicity present.
- The auditory system isvery successful at this task. Hoping to uncover its secrets, we base our signal separation the **correlogram** [SlaneyL93], an auditory model including approximately-constant-Q filtering and autocorrelation.
- **The Weft** is a discrete element describing individual periodic sources, and an algorithm for extracting them from the correlogram representation.

Computational Auditory Scene Analysis (CASA)

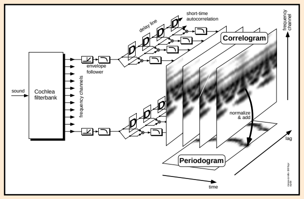
... is computer modeling of listeners' ability to **organize** sound mixtures according to their independent **sources**.



- A popular approach is to calculate a time-frequency 'mask' based on periodicity & common onset cues, and use it to filter out the energy of a single source [Brown92].
- Listeners work at a more abstract level, able to interpret energy in a single channel as several sources.



- The **prediction-driven** approach [Ellis96] looks for consistency between the input and the aggregate predictions of internal models, which can reflect high-level context.
- **Mid-level representations** form a critical bridge between observable features and abstract structure. **Wefts** fill a part of this role.

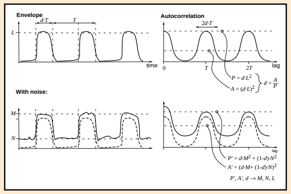


Preprocessing

- The **correlogram** represents sound as a 3-dimensional function of time, frequency and autocorrelation lag.
- In a time-slice of the correlogram, each cell is the short-time autocorrelation of the envelope of one of the cochlea filterbank frequency channels.
- The periodogram summarizes the 'prominence' of each periodicity. It is the sum across all frequency channels of normalized autocorrelations (lag vs. time).
- We use log time (= log frequency) sampling on the lag axis to match human pitch acuity.

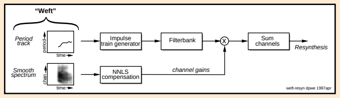
Analysis

- A prominent peak in the periodogram triggers the creation of a new weft, which will track that 'pitch' through subsequent time frames. Autocorrelation aliases of tracked periods are suppressed.
- For a channel excited only by a single periodicity, the autocorrelation maximum at that period is the average squared envelope.
- Noise adds in the power domain, but has a nonlinear effect on the autocorrelation.
- The ratio of peak-to-average autocorrelation varies from 1 for aperiodic input to a maximum specific to each channel/periodicity pair. We can work backwards from this robust feature to eliminate the effect of noise on the autocorrelation peak.
- Tracking wefts through time gives a temporal context that can be used for interpolation through masking.



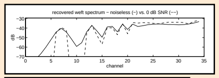
Resynthesis

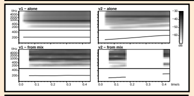
- A mid-level representation should be **perceptually adequate**, implying sufficient detail for independent resynthesis - an important goal.

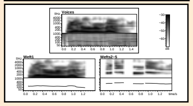


- A weft is defined by its **period track** (time-varying excitation period) and **smooth spectrum** (energy in each frequency channel for each time frame).
- Resynthesis is through a simple source-filter process.
- Overlap of the adjacent frequency bands is precompensated through Non-negative least squares inversion (NNLS).

Results and conclusions







- Extracting the spectrum of a 'flat' periodic signal with and without added noise verifies the basic algorithm.
- A mixture of two periodic signals does not completely separate their spectra. Also, only one signal is extracted when the periods collide.
- Male and female voices result in multiple wefts for each region of voicing. Octave collisions are successfully resolved.
- Wefts form a compact and plausible representation of periodic sounds as part of the vocabulary of a computational auditory scene analysis system.

Why not harmonics?

Weft extraction is complicated. Why not use a fixed narrow bandwidth analysis to track individual harmonics, as has proven successful in the past?

- The ear does not resolve harmonics above the first few. Presumably, this bias towards finer time resolution has some evolutionary benefit.
- Tracking the upper spectrum of periodic signals can be difficult and leads to bulky representations.
- Finding periodicity within broad channels gives different detectability than decisions made on individual harmonics.

References

[AssmS90] Assmann, P. F., Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acous. Soc. Am. 88(2).

[Brown92] Brown, G. J. (1992). "Computational auditory scene analysis : a representational approach," Ph.D. thesis, Sheffield University.

[Ellis96] Ellis, D. P. W. (1996). "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, M.I.T.

[MeddH91] Meddis, R., Hewitt, M. J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acous. Soc. Am. 89(6).

[SlaneyL93] Slaney, M., Lyon, R. F. (1993). "On the importance of time - a temporal representation of sound," in *Visual Representations of Speech Signals*, ed. M. Cooke et al, Wiley.