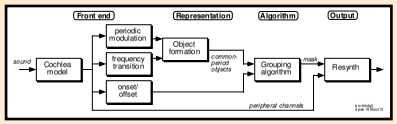
Prediction-driven Computational Auditory Scene Analysis

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

Introduction and background

Auditory Scene Analysis is the process by which listeners infer useful world information from the sound mixture at their ears – the range and weight of an approaching creature, or the words of your companion at a cocktail party. Computer modeling of this function is known as **Computational Auditory Scene Analysis**.

Several successful models based on a traditional signal-processing approach have transformed sound through a series of representations (time-frequency, discrete regions, specific features) to a domain where 'grouping principles' from experimental psychoacoustics [Breg90] may be applied to separate sound mixtures into individual sources.



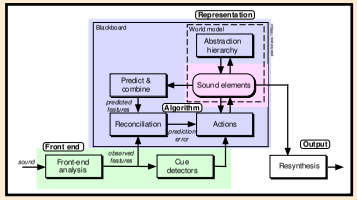
These models have a left-to-right **data-driven architecture**, as in the system of [Brown92] illustrated above. Typically, such systems will produce high-level, abstract output elements only in situations where direct evidence is observed in the input.

By contrast, human listeners can experience **illusions**, where certain scene elements may be perceived without direct evidence, but based on contextual implication alone. A well-known example is the 'continuity illusion' in which a tone is heard to continue through a noise burst that hides its presence or absence.

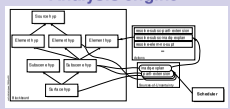
This work was motivated by the need for a model of auditory perception that can *experience illusions*.

The prediction-driven approach

The architecture operates by making **predictions** based on an **internal world-model** which must account for the **complete acoustic scene.** The predictions are **reconciled** with the actual cues observed, leading to modifications of the **generic sound elements** and **higher-level abstractions** of the model. In ambiguous situations, **multiple world-model hypotheses** will be developed in competition to determine which provides the best overall explanation.



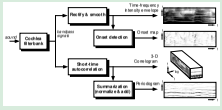
Analysis engine



The prediction-reconciliation process is accomplished with a **blackboard** system based on [CarvL91]. Each hypothesis in a hierarchy is tagged with **sources-of-uncertainty**. The scheduler invokes **actions** to resolve these uncertainties, resulting in the creation, removal or modification of competing hypotheses.

The front-end

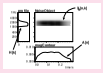
Predictions are reconciled with observations in the domains of **time-frequency intensity envelope** (smoothed at 10 ms) and the **periodogram** (the collapsed within-channel short-time autocorrelation). The full 3-dimensional correlogram [SlanL93] and a simple onset map are also used for updating the sound elements.



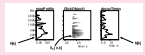
Sound elements

The world-model bottoms-out in a vocabulary of 3 sound element types:

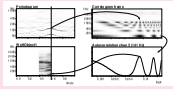
Noise clouds represent regions of unstructured energy by spectral and temporal envelopes.



Transient clicks have an onset time, initial spectrum and per-channel decay rate.

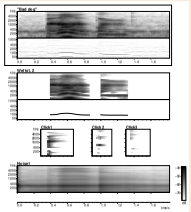


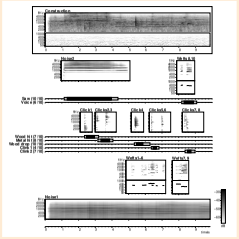
Wefts [Ellis96] represent wide-band periodic energy as a periodogram track and a smooth spectrum.



Results and evaluation

Speech in a noisy background illustrates the kind of analyses produced by the system. Six elements are created; two wefts for the periodic portions of two voiced syllables, three transient clicks at stop-releases, and a noise cloud to explain the background noise. Energy in the observed signal is apportioned between the elements where they overlap according to prediction uncertainty.





The "Construction-site ambience" sound example shows how the approach can find events in very dense sounds, just as listeners do. The noise clouds, clicks and wefts extracted by the system were compared to perceptual events gathered in subjective listening tests (horizontal bars "Saw", "Voice" etc.). There is good agreement in both quality and number of events.

References

[Breg90] Bregman, A. S. (1990). *Auditory scene analysis*, MIT Press.

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

[CarvL91] Carver, N., Lesser, V. (1991). "A new framework for sensor interpretation: Planning to resolve sources of uncertainty," Proc. Nat. Conf. on Artif. Intel.

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

[Slaney93] 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.