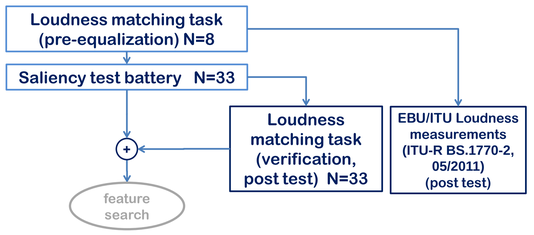
Is there more to saliency than loudness?

Francesco Tordini, *PhD candidate*
{tord@cim.mcgill.ca}
Shared Reality Lab, McGill University and
Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)

Abstract

The ability of human listeners to hear out a sound event from a complex auditory scene is well known. Salience can be defined as the property of a sound to jump to the foreground with respect to other sounds or background noise. Here we present a behavioural test battery aiming to capture the perceived salience of natural sounds in a binaural setting using two competitive sound streams.
Our results demonstrate that perceived loudness effects, although prominent, cannot completely explain foreground/background selection, raising a question as to the degree of overlap between the definitions of loudness and salience in real world scenarios. We also discuss the agreement between the recent ITU-R BS.1770x/EBU-tech3343 broadcasting standards and our subjective loudness ratings.

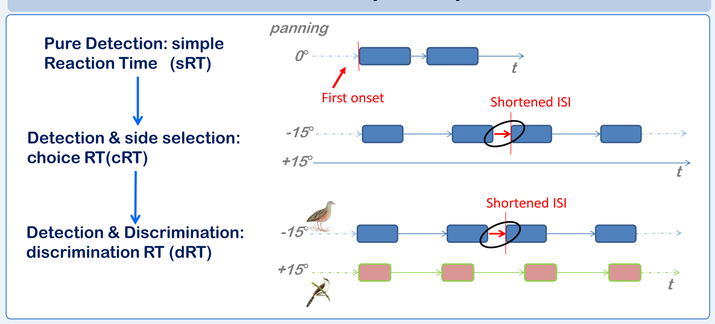
The tests pipeline: saliency vs. loudness



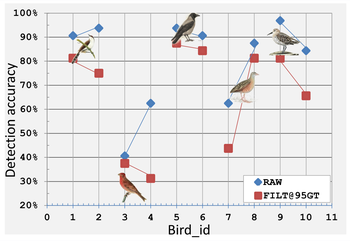
Capturing auditory saliency: our working definition

A sound is "salient" when selecting it (among others) **is "as easy" as detecting it** (in isolation)

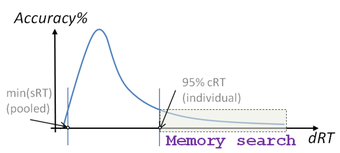
The saliency battery



Saliency, memory and accuracy



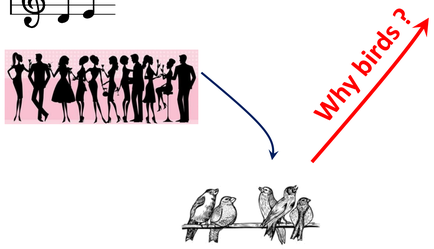
Acceptance window



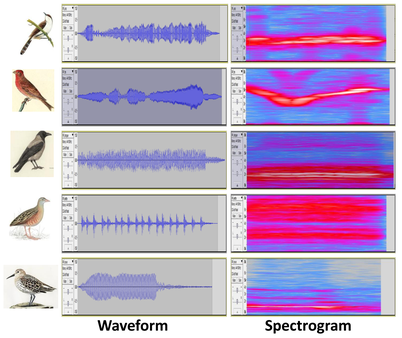
Correlation between Accuracy and dRT improves *from* R^2=0.22 *to* R^2=0.54

Corpus of everyday sounds

Listening strategies are domain/context dependent (Gaver, 1993), e.g., for music, speech, and environmental sounds

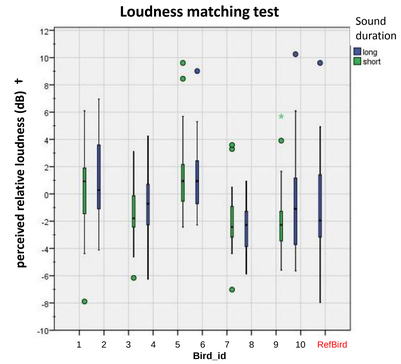


We use a corpus of bird sounds as these are less semantically loaded

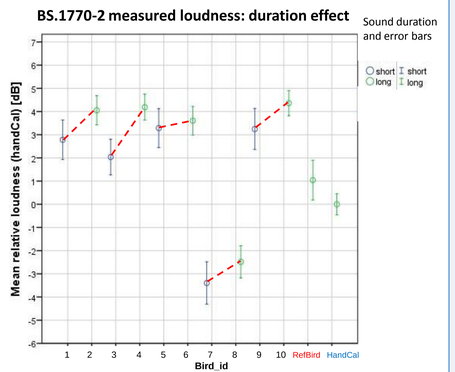


Each bird chirp comes with two sample durations: short (185 ms) or long (250 ms)

Loudness: perceived vs. measured (ITU-R BS 1770-2)

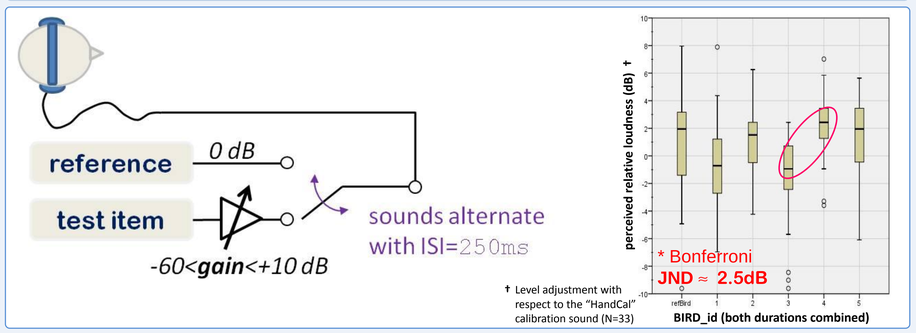


† Level adjustment with respect to the "HandCal" calibration sound (N=33)



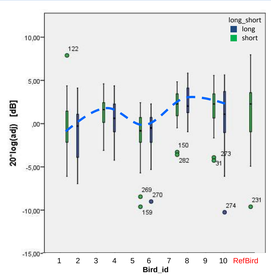
Error bars: +/- 2SD

Loudness matching task and significant perceptual separation (JND)

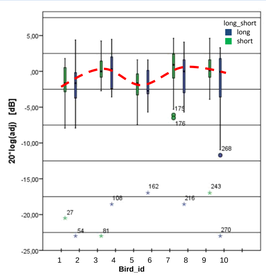


† Level adjustment with respect to the "HandCal" calibration sound (N=33)

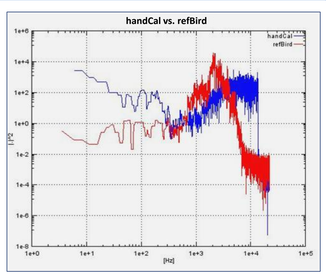
Loudness ratings using references with different power spectra



Loudness ratings using **HandCal**



Loudness ratings using **RefBird**



Different power spectra lead to similar ratings

Funding sources This research is funded through the Graphics, Animation, and New Media (GRAND) Networks of Centres of Excellence, HP Labs Innovation Research Program 2011 (HPL-IRP2011), and the Centre for Interdisciplinary Research in Music, Media and Technology (CIRMMT).

Ongoing and future work

Matching features: An initial selection includes features derived from the autocorrelation function (ACF) as the *energy decay profile* (τ_e) as defined by Ando et al. [1] and [3] but calculated over different time scales between 50 ms and 1s. Moreover, *spectral roughness and sparseness* as defined by Peeters et al. [5] are studied over different time scales.

Further work will investigate the relationship between the JNDs for the "raw" loudness (ITU-R BS.1770-3 [4] and EBU-Tech-3343 [6]) and the perceptual ratings derived from the loudness matching task.

Finally, we will validate the current results with a new behavioral test using **three** competitive streams to test the categorical nature of foreground/background organization.

Conclusions

1. Perceptual loudness can explain 38% of the trials but has little sensitivity to duration variations. It partially explains detection variance between different birds, but not within-bird differences *(i.e., loudness uses longer integration windows).*
2. ITU/EBU loudness (ITU-R BS.1770-x) is sensitive to sound duration and pattern structure.
3. Perceptual and measured loudness lead to similar rankings.
4. A conservative loudness JND is preferred.
5. Detection performance (accuracy) depends on memory. Saliency should be evaluated under conditions involving minimal access to memory (RT filtering, using a suitable baseline).
6. Saliency seems to be organized in bands: High (mid) Low. *How many objects can we attend to at the same time?*
7. Accuracy and RT are in better agreement when filtered for memory effects. Should we avoid the speed-accuracy-trade-off (SAT) to capture primitive reactions *(i.e., low perceptual load)*?

References

[1] Y. Ando, T. Okano, and Y. Takezoe, "*The running autocorrelation function of different music signals relating to preferred temporal parameters of sound fields*", JASA, vol. 86, no. 2, pp. 644–649, 1989.
[2] A. S. Bregman, Auditory Scene Analysis. The perceptual organization of sound. MIT Press, 1990.
[3] D. D'Orazio, S. De Cesaris, and M. Garai, "*A comparison of methods to compute the 'effective duration' of the autocorrelation function and an alternative proposal*", JASA, vol. 130, no. 4, p. 1954, 2011.
[4] Rec. ITU-R BS.1770.3 (Aug 2012) and ITU-R BS.1770-2 (May 2011), available at *http://www.itu.int/searchcenter/Pages/SearchRecommendations.aspx*
[5] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "*The timbre toolbox: Extracting acoustic descriptors from musical signals*", JASA, vol. 130, pp. 2902–2916, 2011
[6] EBU Tech Doc 3343-v2-2011 (Aug 2011), available at *https://tech.ebu.ch/docs/tech/tech3343.pd*