The paper “Multimodal Variational Auto-Encoder based Audio-Visual Segmentation” was interesting to read. I had a few thoughts as I read it:

* I was not clear what Audio-Visual Segmentation was and did not understand it until I was reading section 3.3.
* Section 3.4 was the most interesting to me. More on that below.
* The authors seem to focus on the highest value being best. They do not present a calculation of statistical significance in the paper. While some of the values are higher, I am not sure that they are significantly higher. Most likely they are within experimental uncertainty.

Section 3.4 is called Shared-Information Completeness. I find it interesting that while they have two separate sources for Video and Audio data, they treat them as two separate views of the same target. The latent space is fused but they have sca and scv from the latent space. They use Bayes law to come up with a shared information completeness loss function, lsic. I find this very interesting to take the two different streams and then try to find a way to maximize the shared information between them.

Obviously, this makes sense for the paper where they are trying to determine what part of the visual makes the sound. But it seems to me that this technique could be used in other areas to disentangle information while still being able to find interactions. For example, with a stream of audio from a cocktail party, could you use a technique like this to single out one specific conversation?