The recently proposed audio-visual scene-aware dialog task paves the way to a more data-driven way of learning virtual assistants, smart speakers and car nav