Understanding scenes around us is a fundamental capability in human intelligence. Similarly, designing computer algorithms that can understand scenes is a fundamental problem in artificial intelligence. Humans consciously or unconsciously use all five senses (vision, audition, taste, smell, and touch) to understand a scene, as different senses provide complimentary information. Existing machine scene understanding algorithms, however, are designed to rely on just a single modality. Take the two most commonly used senses, vision and audition, as an example, there are scene understanding algorithms designed to deal with each single modality. However, no systematic investigations have been conducted to integrate these two modalities towards more comprehensive audio-visual scene understanding. Here, I will talk about two recent works from my group. The first work addresses the cross-modal audio-visual generation problem leveraging the power of deep generative adversarial training. We show state-of-the-art performance in constrained domains such as music performance and lip reading. The second work addresses the general scene audio-visual event localization. We show using both modalities cohesively in a deep time series model outperform using visual-alone or audio-alone.