Learning spoken concepts from unlabeled audio-visual data

Karen Livescu, TTI-Chicago

Speech technologies, such as automatic speech recognition, typically address the mapping between acoustic signals and the corresponding written words. Given sufficient labeled speech data, modern speech technologies have achieved remarkable performance. However, for many languages, obtaining labeled data is expensive or even impossible. In addition, modern speech technologies typically do not consider the *semantic* content in the speech signal.

This talk will consider whether we can address both of these shortcomings by learning from unlabeled speech paired with images. I will present our recent work using images paired with spoken captions to learn speech concepts. Without access to any transcribed speech, the resulting model can spot keywords in new speech input, and can also be used to search for spoken content that is semantically relevant to a query.