We investigate the problem of audio-visual (AV) person diarization in broadcast data. That is, automatically associate the faces and voices of people and determine when they appear or speak in the video (see Fig 1).

The contributions are twofolds. First, we formulate the problem within a novel CRF framework that simultaneously performs the AV association of voices and face clusters to build AV person models, and the joint segmentation of the audio and visual streams using a set of AV cues and their association strength.
Secondly, we use for this AV association strength a score that does not only rely on lips activity, but also on contextual visual information (face size, position, number of detected faces,...) that leads to more reliable association measures. Experiments on 6 hours of broadcast data show that our framework is able to improve the AV-person diarization especially for speaker segments erroneously labeled in the mono-modal case mainly when AV association is reliable. Besides, presence of off voices and multi-head shots make still AV association challenging (see Fig.2 images b, d and f).