**Background & Motivation**
- Accurate and timely recognition of the trigger keyword is vital.
- There is a trade-off needed between accuracy and latency.
- Existing works focus on accuracy and computational latency.

**Proposed system:**
Unified speculation, detection, and verification model
Speculation makes an early decision, which can be used to give a head-start to downstream processes on the device.
Verification verifies previous decision after the keyword span.

**Model Architecture**
CNN encoder behaves as an efficient feature extractor to model local temporal and spectral dependencies:
- Convolutional neural network front-end has a receptive field of 34 frames and has a stride of 6 frames.
- Each layer is composed of a convolution layer, a rectified linear unit activation layer, an optional max pooling layer, a batch normalization layer and a drop out layer.
- Outputs are vectorized and fed to RNN decoder

**Model Architecture (cont.)**
CNN decoder captures dependencies among different frames:
- A long short-term memory (LSTM) layer captures dependencies using “gating” mechanism.
- A fully-connected (FC) layer is used to further transform features before Softmax output.
- Due to the similarity of speculation, detection, and verification tasks, i.e., all of them try to detect the same word from audio, we share the same convolutional front-end, LSTM, and FC layer for them to reduce model size.
- We only add three output heads with separate linear layers for dimension reduction and Softmax outputs.
- The additional computations to achieve these three tasks simultaneously are only introduced by these small output heads, hence are negligible.

**Experimental Results**
- Speculation models from both the single task baseline and USDV have the earliest detection.
- Verification models achieve the lowest FAR with more right context.
- USDV model is able to achieve three tasks with different accuracy and latency trade-off, which validates the effectiveness of the MTL training and latency-aware max-pooling loss.
- USDV model achieves same level of performance as baseline models, which shows that the CRNN architecture has enough capacity to perform all three task simultaneously.

**Conclusions**
- We propose an CRNN-based unified speculation, detection, and verification keyword detection model.
- We propose a latency-aware max-pooling loss, and show empirically that it teaches a model to maximize accuracy under the latency constraint.
- A USDV model can be trained in a MTL fashion and achieves different accuracy and latency trade-off across these three tasks.