UNSUPERVISED LEARNING OF OVERLAPPED SPEECH MODEL PARAMETERS FOR MULTICHANNEL SPEECH ACTIVITY DETECTION IN MEETINGS

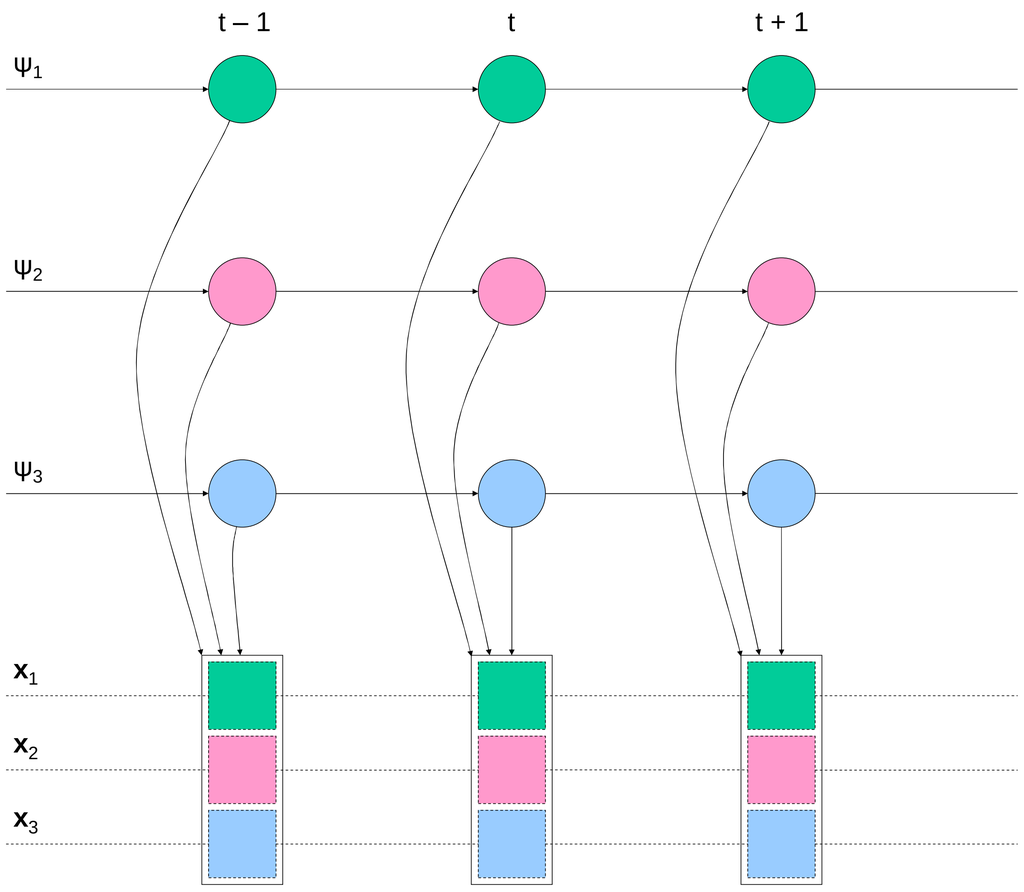
Kornel Laskowski and Tanja Schultz

interACT, Carnegie Mellon University, Pittsburgh, USA

kornel@cs.uka.de **tanja@cs.cmu.edu**

A. The Problem

1. Current speech activity detection systems treat participants independently; **they ignore the fact that participants are speaking together.**
2. This is clearly wrong, since regardless of the number of participants, about ~80% of the time is spent in one-speaker-at-a-time talk.



$$P(\mathbf{\Psi} \mid \mathbf{X}) = \prod P(\boldsymbol{\psi}_k \mid \mathbf{X})$$

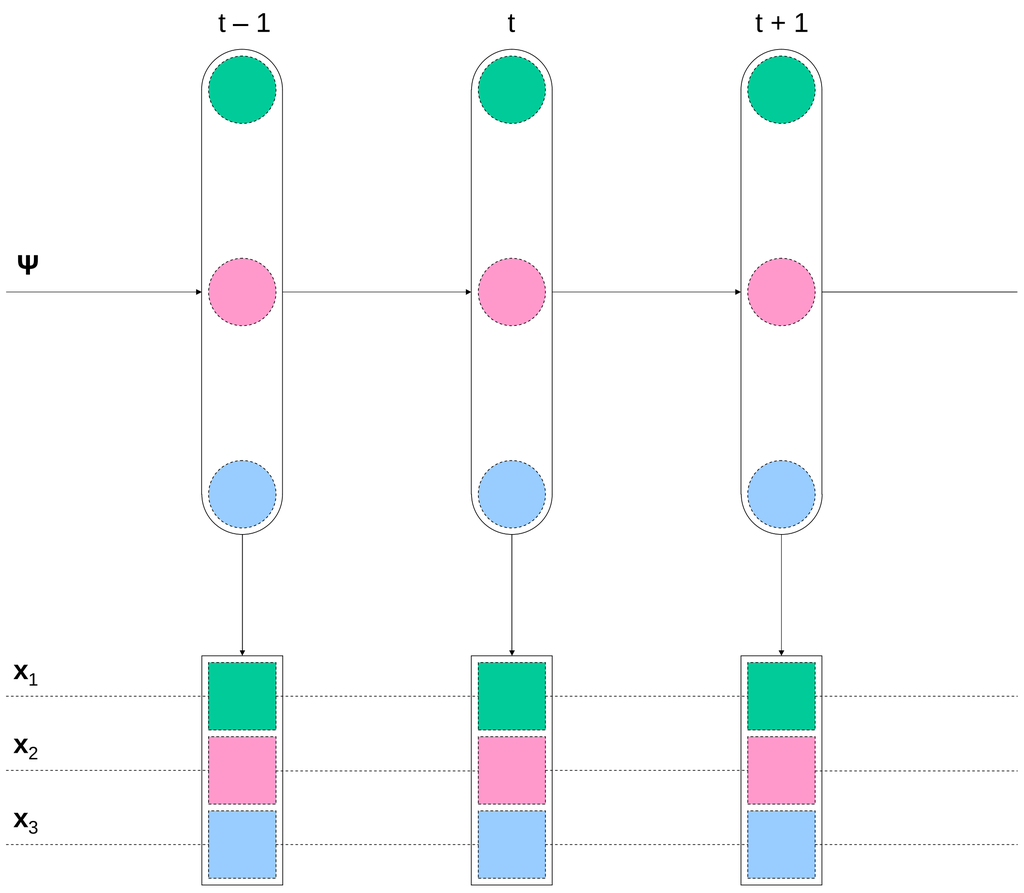
$$\boldsymbol{\psi}_k^* = \text{argmax } P(\boldsymbol{\psi}_k \mid \mathbf{X})$$

$$1 \le k \le K$$

3. Such systems treat overlap and interruption (and crosstalk!) as purely coincidental, emergent phenomena, obtained by combining the independently decoded channels.

B. Proposed Solution

1. Instead of decoding in each participant's **vocal activity state** space (2 states), decode in the joint **vocal interaction state** space (2^K states).
2. For transition probabilities, **partition the state space by the number of currently active speakers**, ignoring their identities.



$$\mathbf{\Psi}^* = \text{argmax } P(\mathbf{\Psi} \mid \mathbf{X})$$

$$\boldsymbol{\psi}_k^* = (\mathbf{\Psi})_k$$

$$1 \le k \le K$$

3. Acoustic training data will be extremely rare for the majority of events. Train models of overlapped speech using the test data, with **labels supplied by an initial pass system which is high-precision in the single-speaker intervals.**