multiple instance learning for classification of human behavior observations

Nassos Katsamanis, James Gibson,
Matthew Black, Shrikanth Narayanan
University of Southern California

human behavior observations

"You work too much…"

"Wanted to talk about talking…"

"Topic is really household chores stuff… "

"Temper and patience…"

Christensen et al, Journal of Consulting and Clinical Psychology, 2004

human behavior coding

10-minutes long problem solving interaction

acceptance: high
blame: low
humor: low
sadness: low

coding is only performed at the session-level

Is the husband showing acceptance?

From the manual:
“Indicates understanding and acceptance of partner’s views, feelings, and behaviors. Listens to partner with an open mind and positive attitude. … ”

Jones and Christensen, Social Support Interaction Rating System, UCLA, 1998

… but what happens at the speaker-turn level?

husband speaking turns:

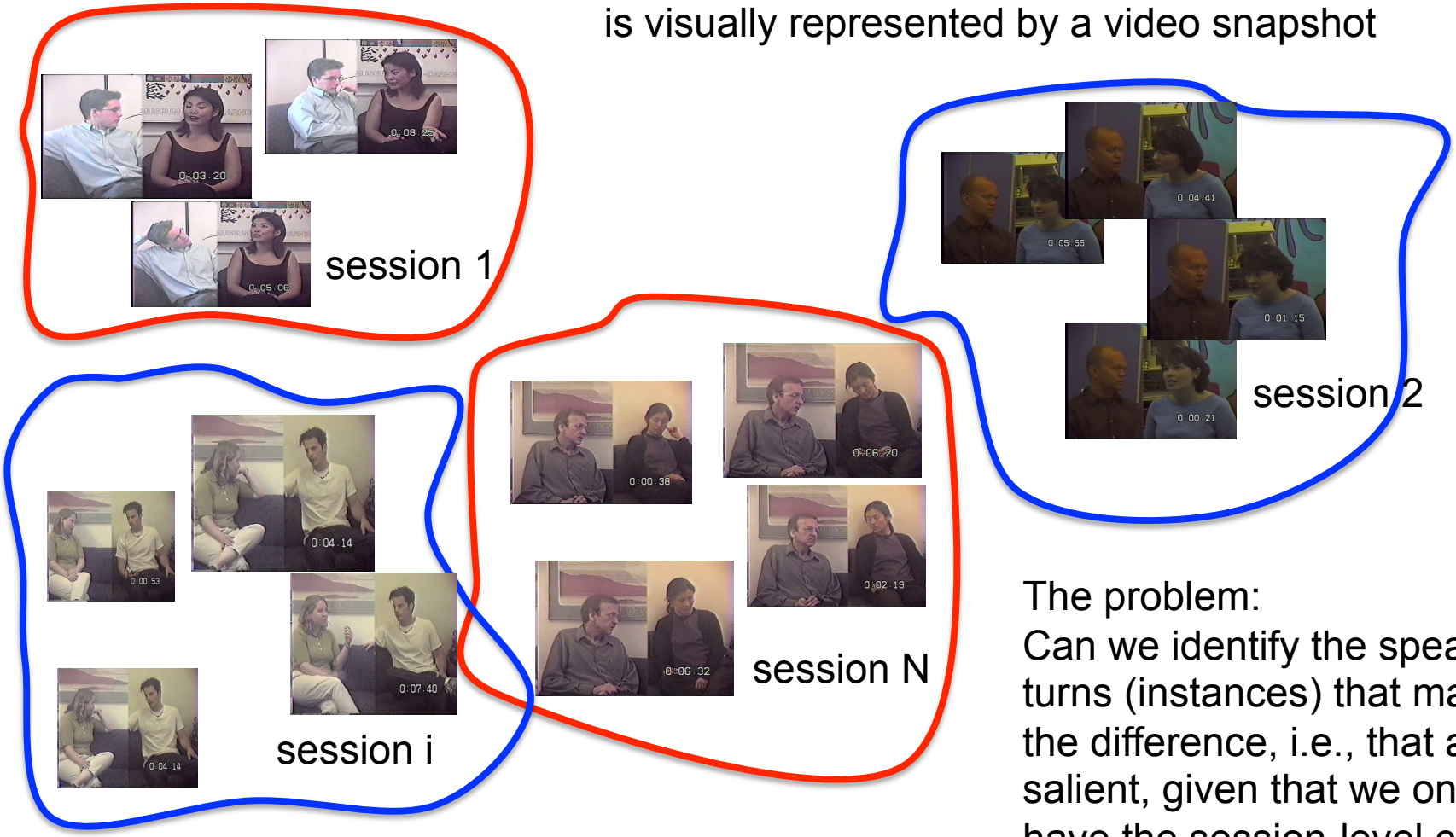
The problem:
Can we identify the speaker turns (instances) that make the difference, i.e., that are salient, given that we only have the session-level codes?

Our approach:
multiple instance learning

To validate:
We use the saliency based representation to classify the whole session: low vs. high acceptance

multiple instance learning

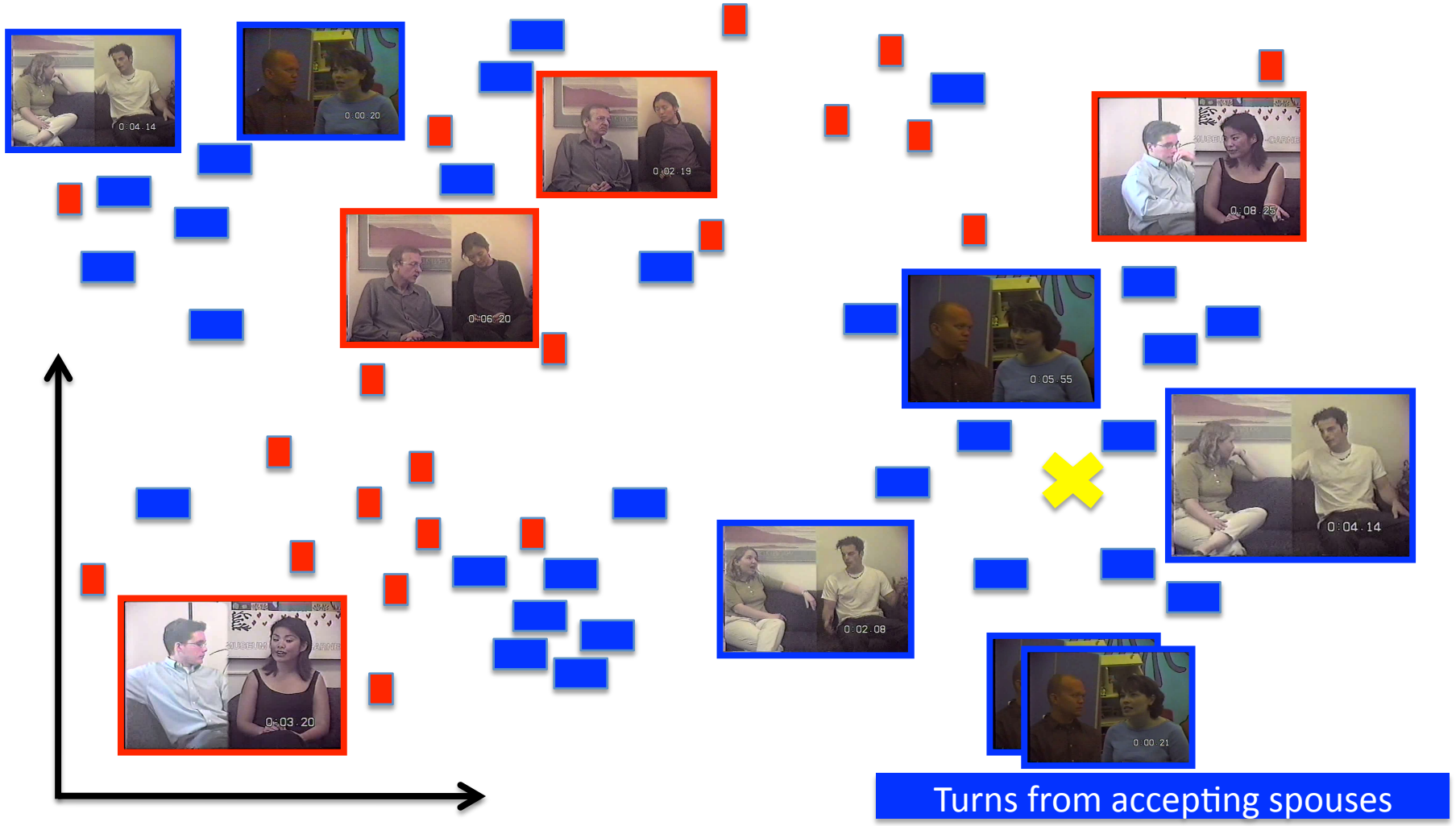
Each speaker turn is an instant (of behavior) and is visually represented by a video snapshot



The problem:
Can we identify the speaker turns (instances) that make the difference, i.e., that are salient, given that we only have the session-level codes?

red sessions: non-accepting spouse
blue sessions: accepting spouse

diverse density (Maron et al., NIPS 1998)



Points of locally maximum diverse density
Speaker turns close to these points are salient

instance (speaker turn) representation - text

Bag-of-words representation:
"term frequency times inverse document frequency" (tfidf) values of a selected set of words

- information gain to select the words that are most informative for discriminating, e.g., low vs. high acceptance

Behavior	Informative words
(high vs. low) acceptance	um, told, nothing, mm, yes, everything, ask, more, (laugh), can't
(high vs. low) blame	nothing, everything, your, no, said, always, can't, never, mm, told
(high vs. low) humor	(laugh), topic, good, missing, cool, treat, seemed, truly, accept, case
(high vs. low) positivity	um, kind, nothing, mm, good, (laugh), told, can't, mean, why

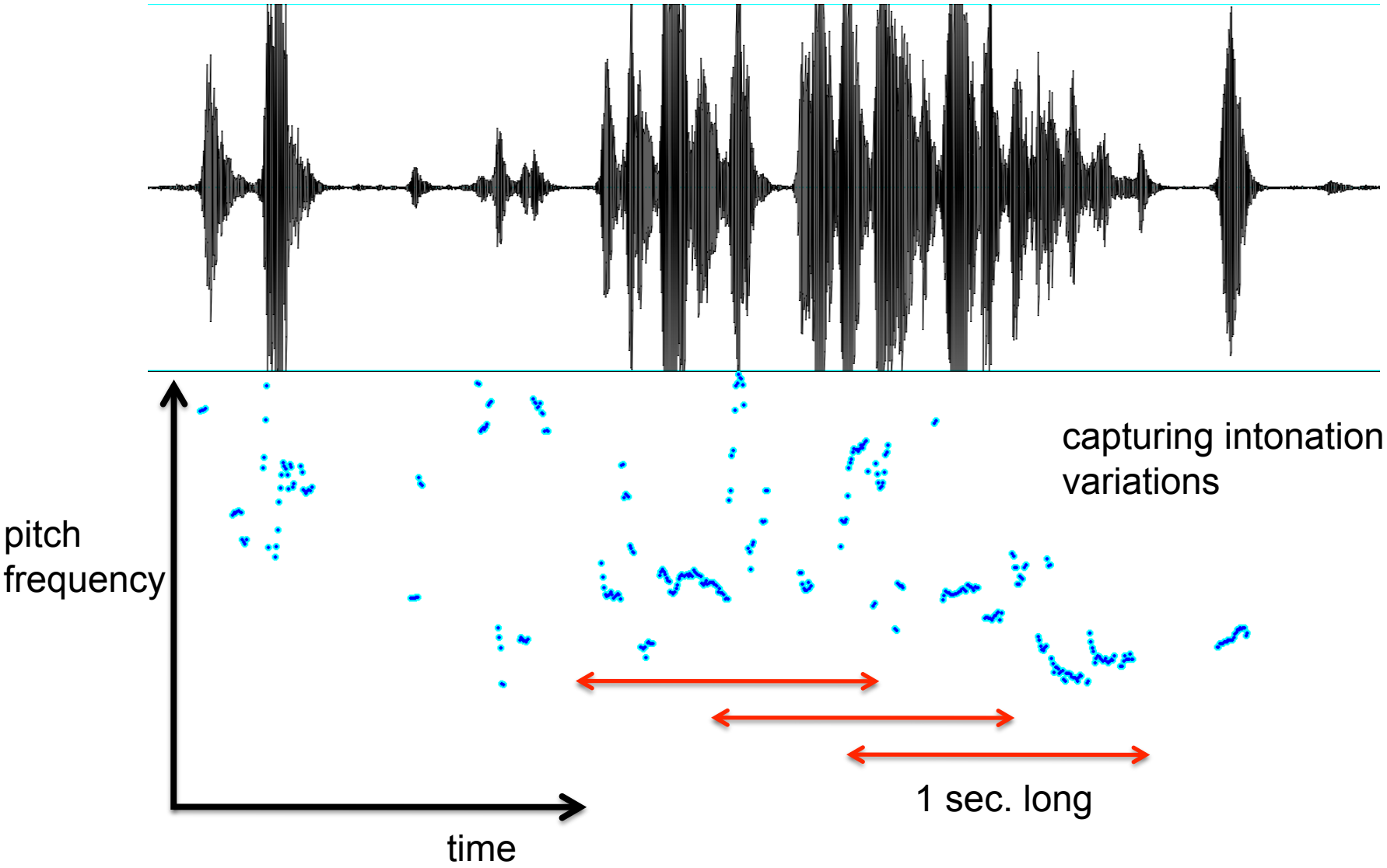
instance (speaker turn) representation - text

Bag-of-words representation:
"term frequency times inverse document frequency" (tfidf) values of a selected set of words

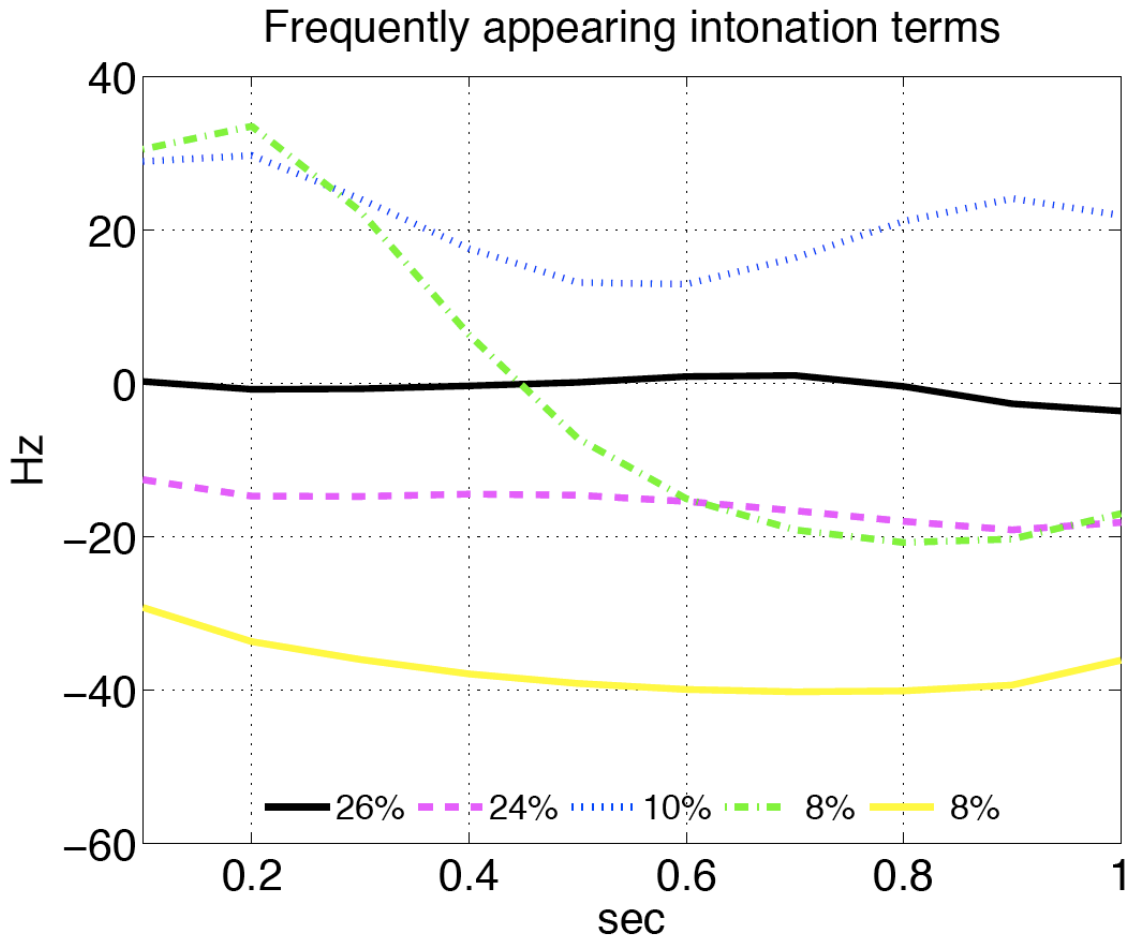
- information gain to select the words that are most informative for discriminating, e.g., low vs. high acceptance

Behavior	Informative words
(high vs. low) acceptance	um, told, nothing, mm, yes, everything, ask, more, (laugh), can't
(high vs. low) blame	nothing, everything, your, no, said, always, can't, never, mm, told
(high vs. low) humor	(laugh), topic, good, missing, cool, treat, seemed, truly, accept, case
(high vs. low) positivity	um, kind, nothing, mm, good, (laugh), told, can't, mean, why

instance (speaker turn) representation - audio



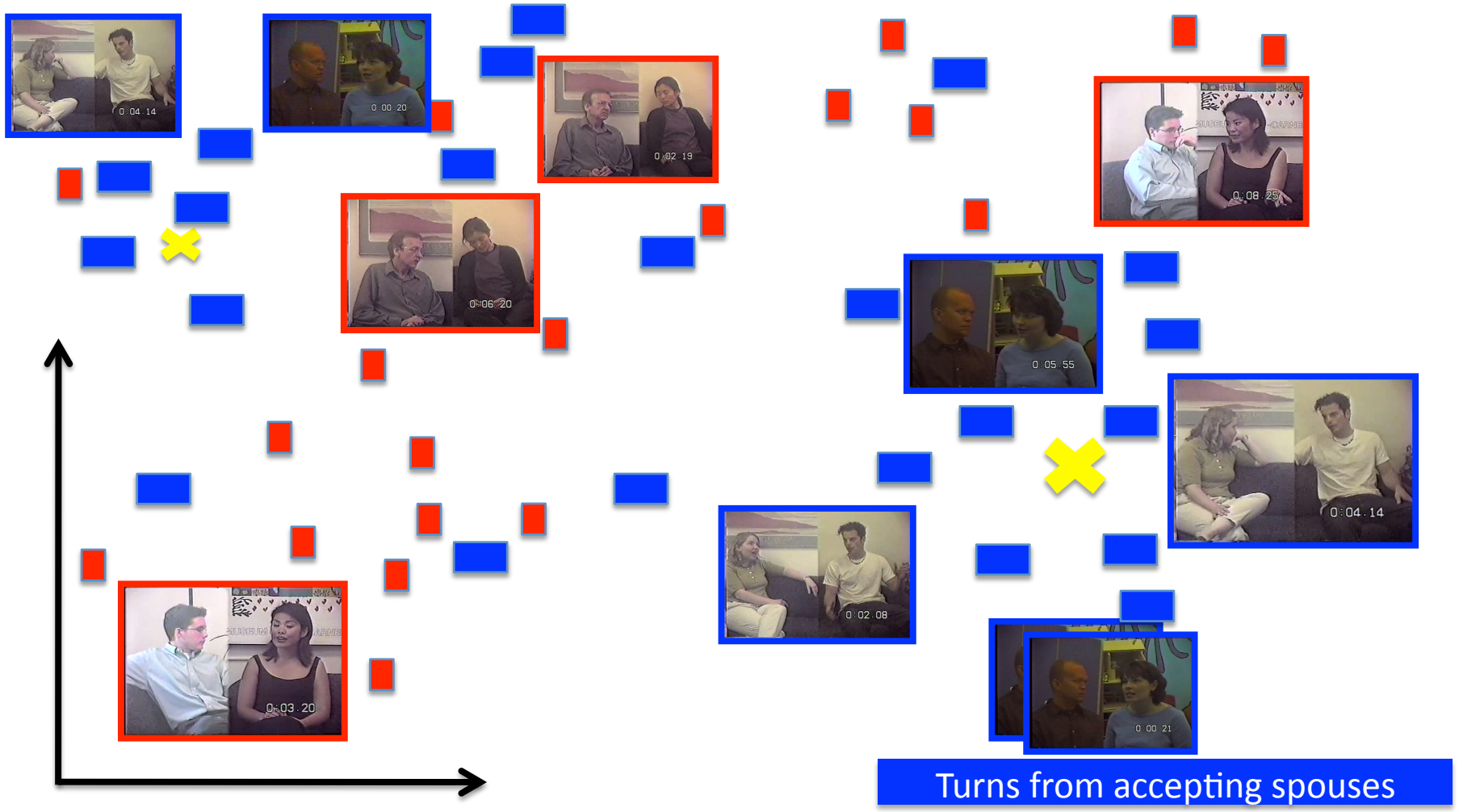
instance (speaker turn) representation - audio



Intonation patterns are determined
via vector quantization

Bag-of-words representation:
Normalized appearance frequency values of a set of intonation patterns

… and we estimate the diverse density in this feature space



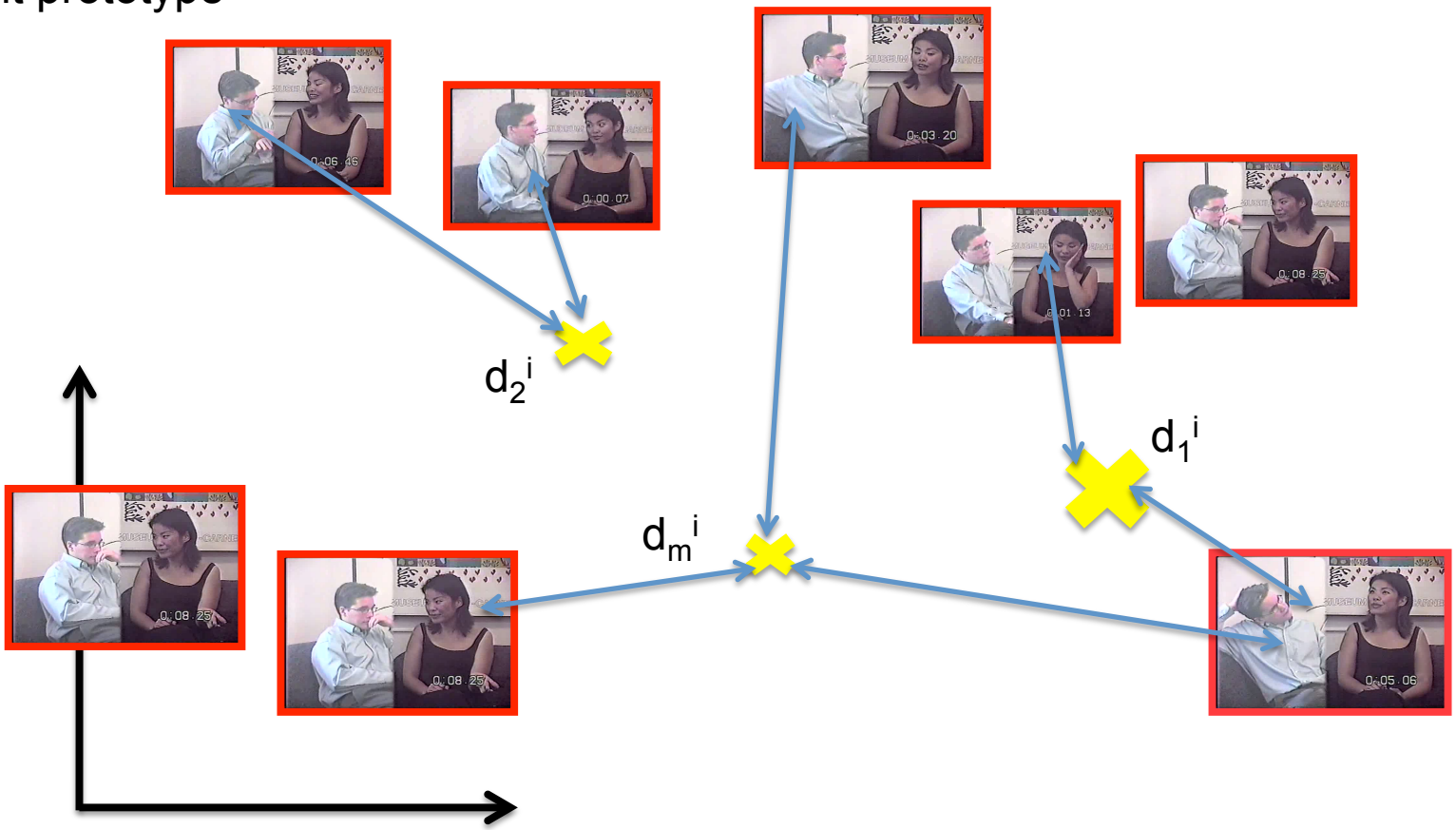
Turns from accepting spouses

Turns from non-accepting spouses

Points of locally maximum diverse density
Speaker turns close to these points are salient

saliency estimation

For each turn we can have a saliency estimate based on how far the turn is from a salient prototype



saliency estimation – session representation

session instances (speaker turns, i = 1… N)




…


salient prototypes (m=1…M)



⋮


distances of session instances from salient prototypes

d_1^i

d_2^i

⋮

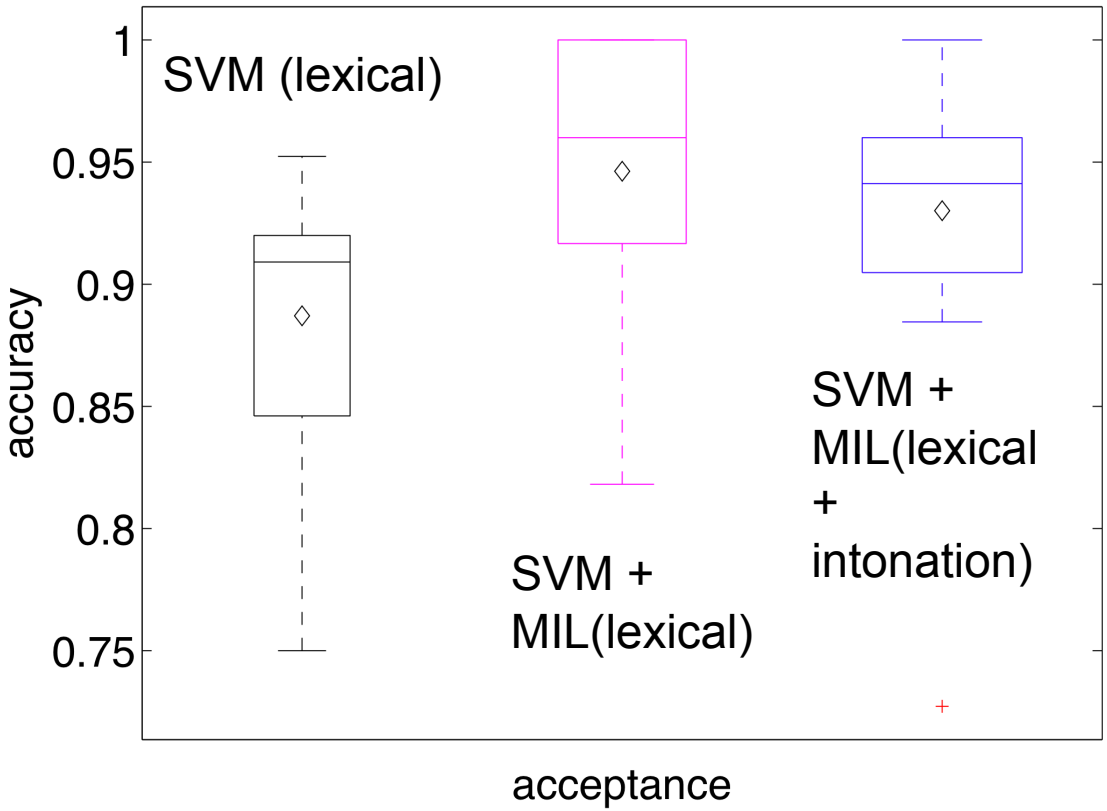
d_m^i

feature vector for session

$$\begin{bmatrix} \min_i d_1^i \\ \min_i d_2^i \\ \vdots \\ \min_i d_m^i \\ \vdots \end{bmatrix}$$

to validate the proposed representation, we run classification experiments using Support Vector Machines

classification results



10-fold Cross-validated results: high vs low acceptance

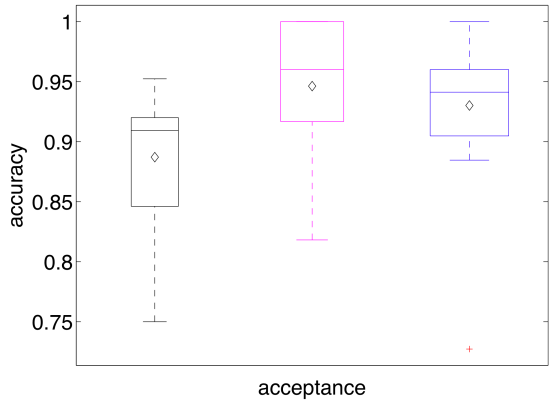
The black boxes correspond to the baseline classification results:
- Bag-of-words lexical representation of the whole session (without exploiting saliency estimates)

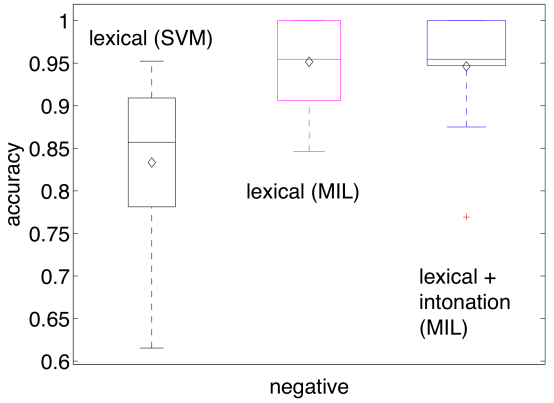
Two things can be noted:

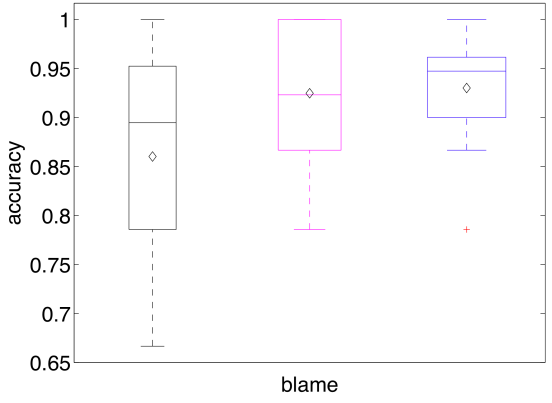
- performance is significantly improved when switching to the multiple instance learning setup
- inclusion of the intonation features does not lead to further consistent accuracy improvements.

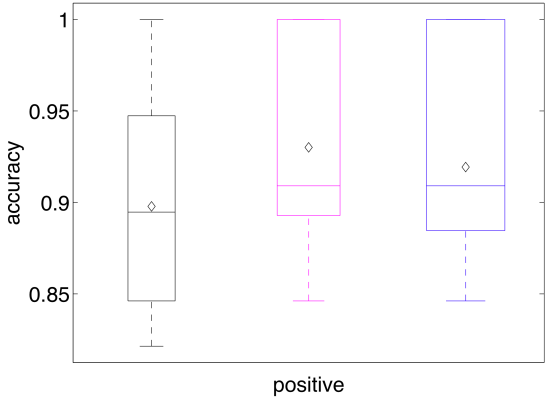
classification results

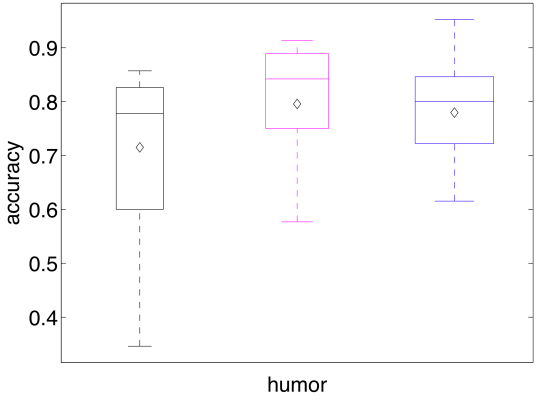
10-fold cross-validated results for six behavioral codes (high vs low)











discussion/directions

- Saliency is determined based on diverse density and is fully data-driven
  - Currently, we are trying to validate and relate the estimates to saliency annotations from experts
  - We need to provide reliability of our saliency estimates to increase their usability
- Visual information can also be exploited in this scheme, e.g., facial expressions and body language
- We have also tried to apply similar ideas to identify salient instances of entrainment during the interaction (Lee et al., poster session tomorrow)

acknowledgements - We are grateful to Brian Baucom and Andrew Christensen from the Psychology Departments of USC and UCLA respectively for giving us access to the couple therapy dataset and for many fruitful discussions. This research is supported by the National Science Foundation.