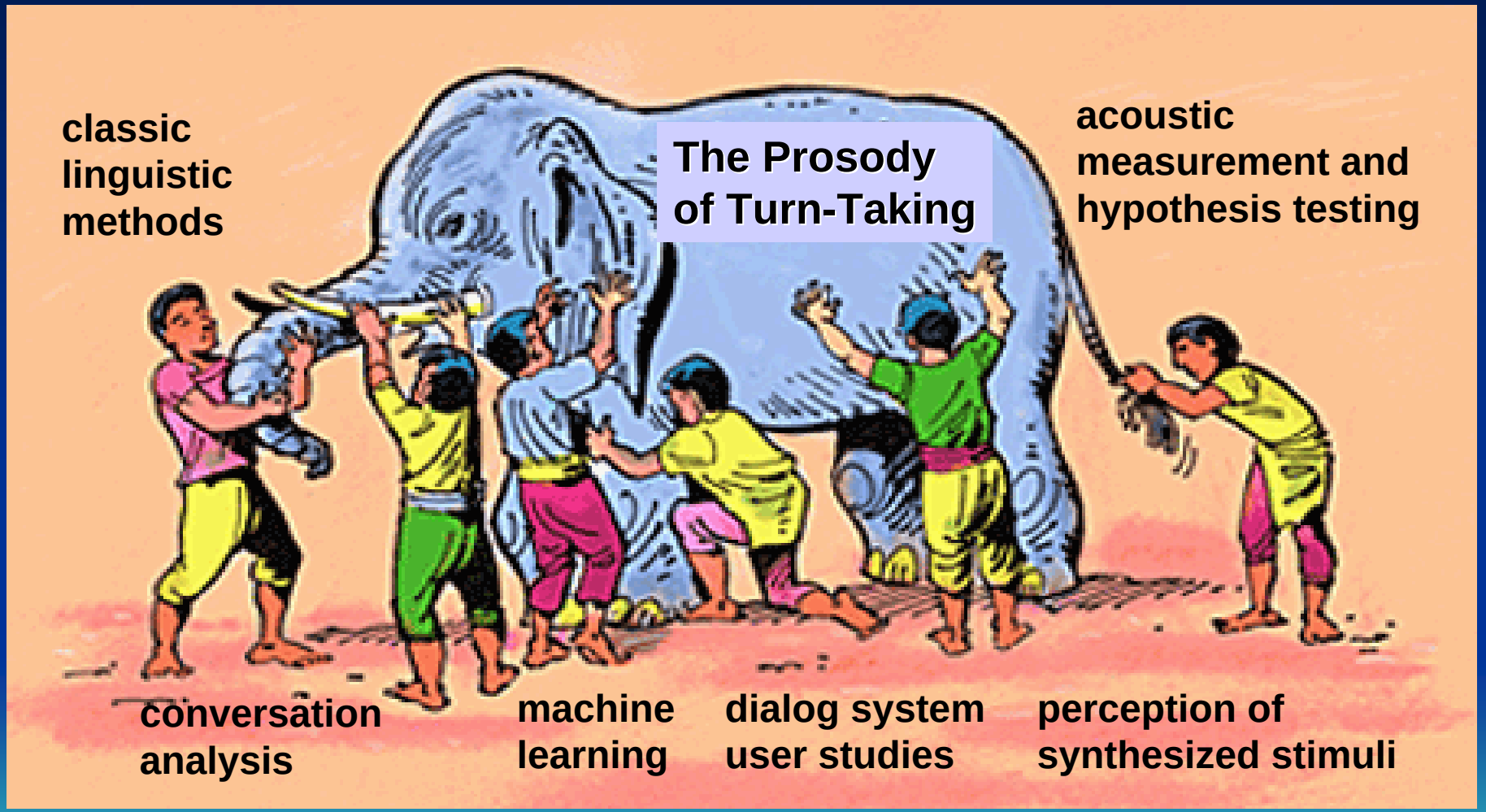
Various Approaches



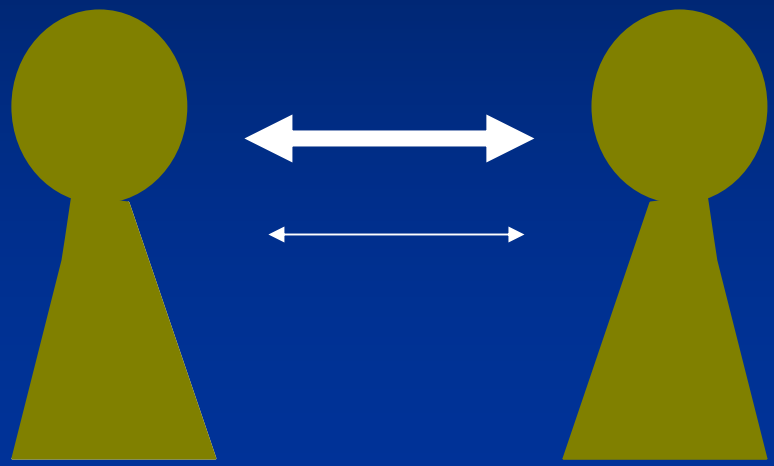
A Case Study in the Identification of Prosodic Cues to Turn-Taking

- Back-Channeling in Arabic -

Nigel Ward and Yaffa Al Bayyari

University of Texas at El Paso

The Second Channel



Form
- gesture
- gaze
- prosody …

Content
- uncertainty,
- novelty,
- dialog control …

Value
- efficiency
- satisfaction
- + (Shriberg 2005)
- …

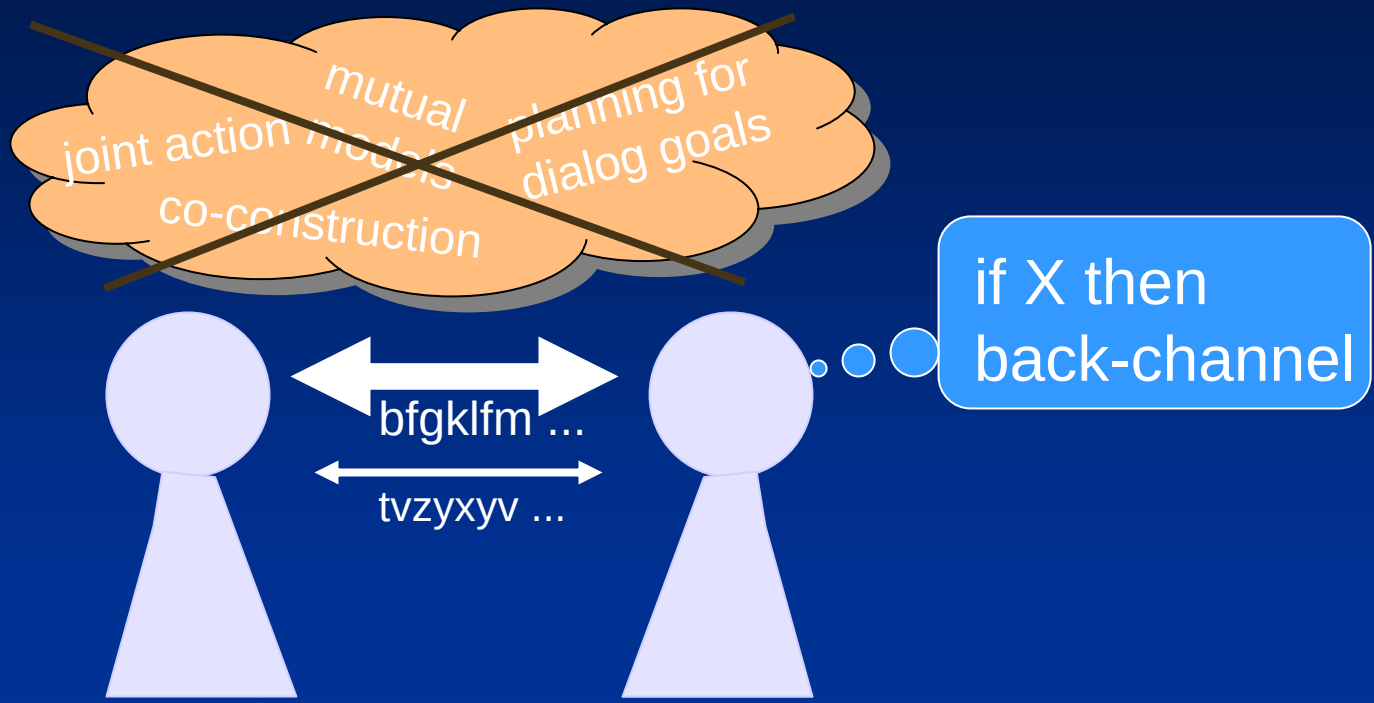
1. Project Aims

Discover the rules governing back-channeling in Arabic

to teach soldiers how to "show you're listening"

- a qualitative description to use for teaching, plus
- a quantitative description to drive the characters

2. Problem Formulation



- using only past information

(no look-ahead)

- using only features computable from the signal

(no hand-labeling)

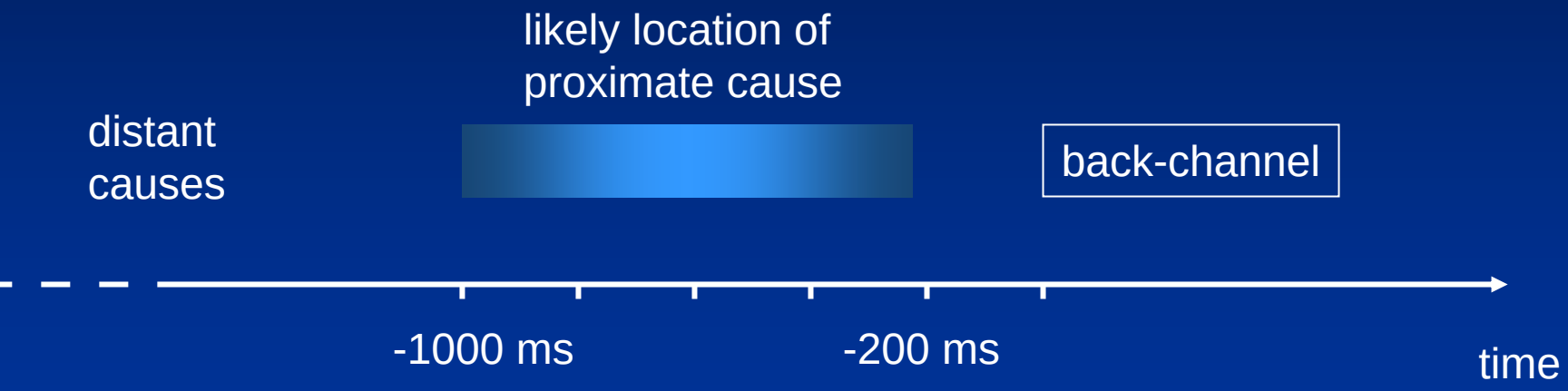
3. Corpus Preparation

All the usual issues ...

big enough to

- find a good dialog
- do proper evaluation

4a: Feature Discovery unclear where to look



Complications:

- time from cue to back-channel varies (complicates Machine Learning)
- salient events can obscure the cues (complicates perceptual analysis)

Example: what do the following have in common?

4b: Feature Discovery
the overwhelming multitude of features

computed over prosody (pitch, energy, timing), voicing ... (possibly in combination)

for example:

- height of highest pitch peak in the last 400 ms relative to the baseline over the past 2000 ms
- first coefficient of a second-order approximation to the pitch curve over the last three syllables before a pause of at least 200 milliseconds
- presence of a 150 millisecond region with the pitch consistently below the 26th percentile

...

4c: Feature Discovery: harnessing perception

audio inspection

- perceive lots of information specific places
- hard to focus on specific features
- hard to scan to

visual inspection

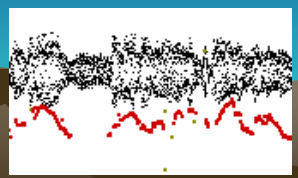
- perceive only what's graphically salient
- easy to focus on specific features
- easy to scan to specific places

neither

- no subjectivity
- no insight

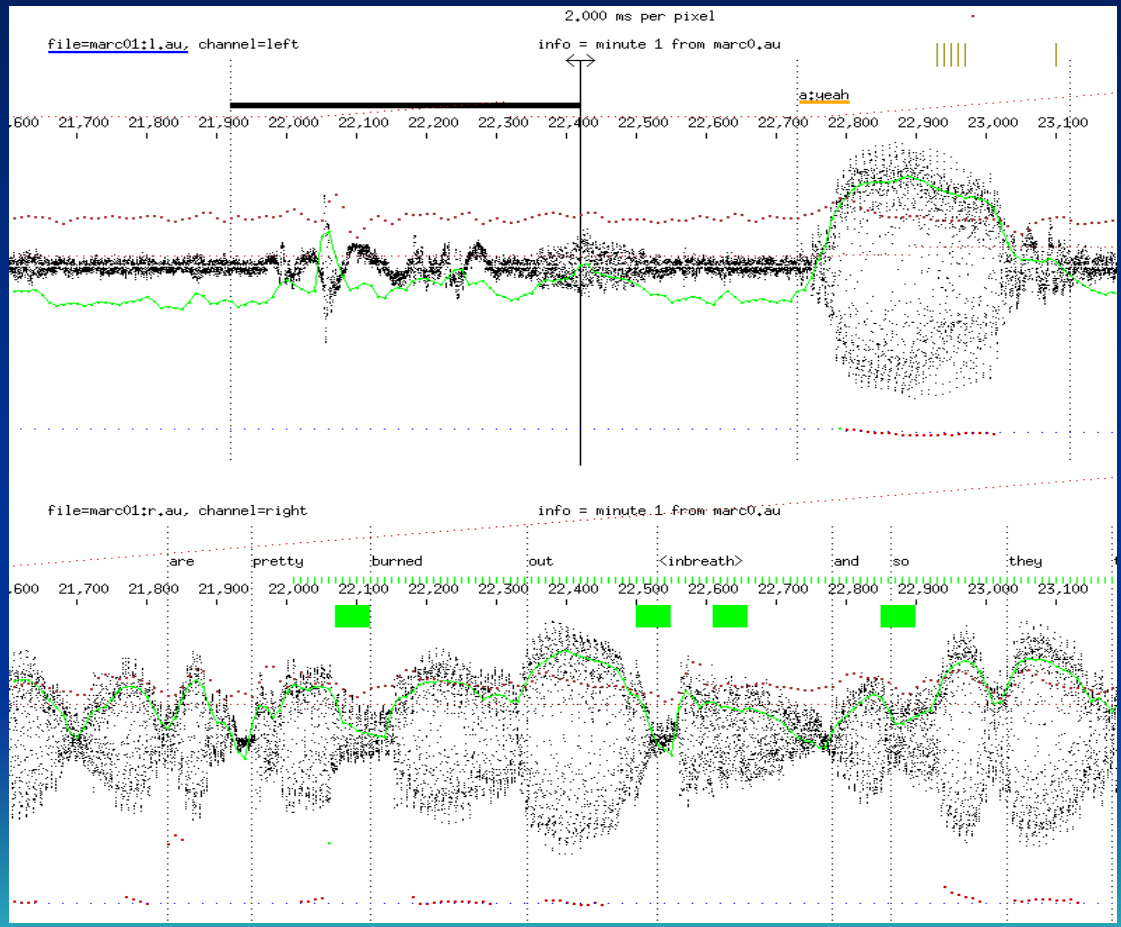
both

- perceive lots of information
- navigate quickly
- focus easily
- need tools

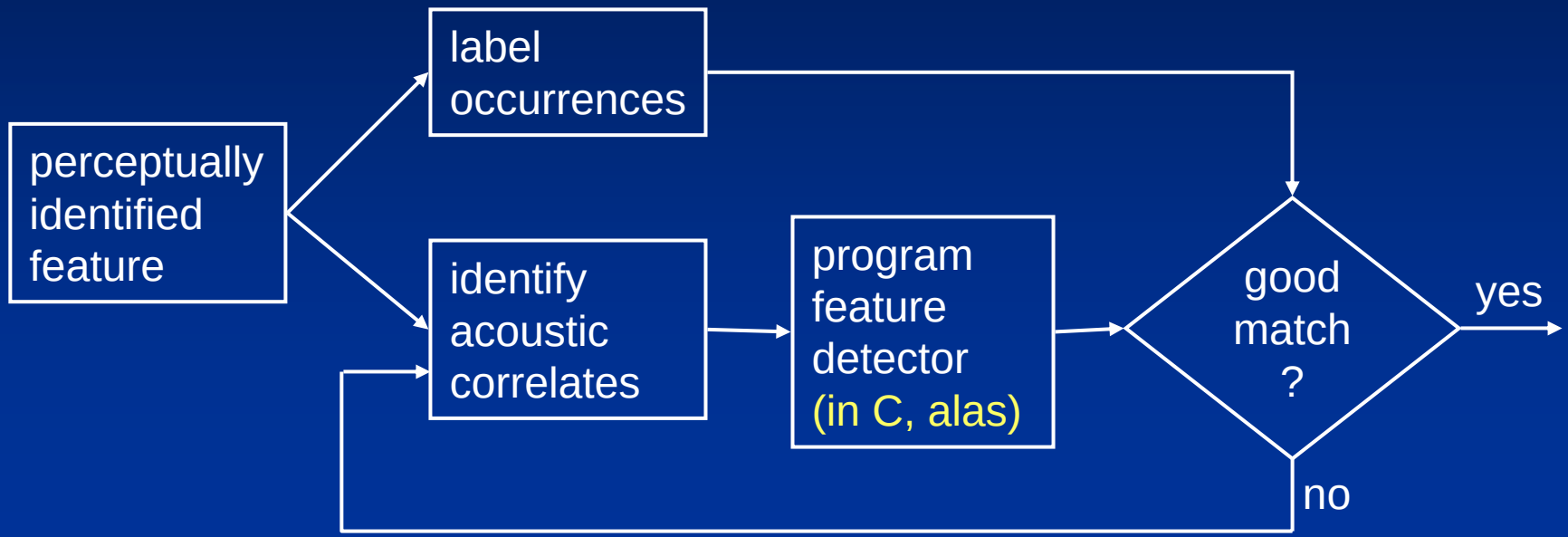


A Custom Tool for Integrated Analysis

Didi



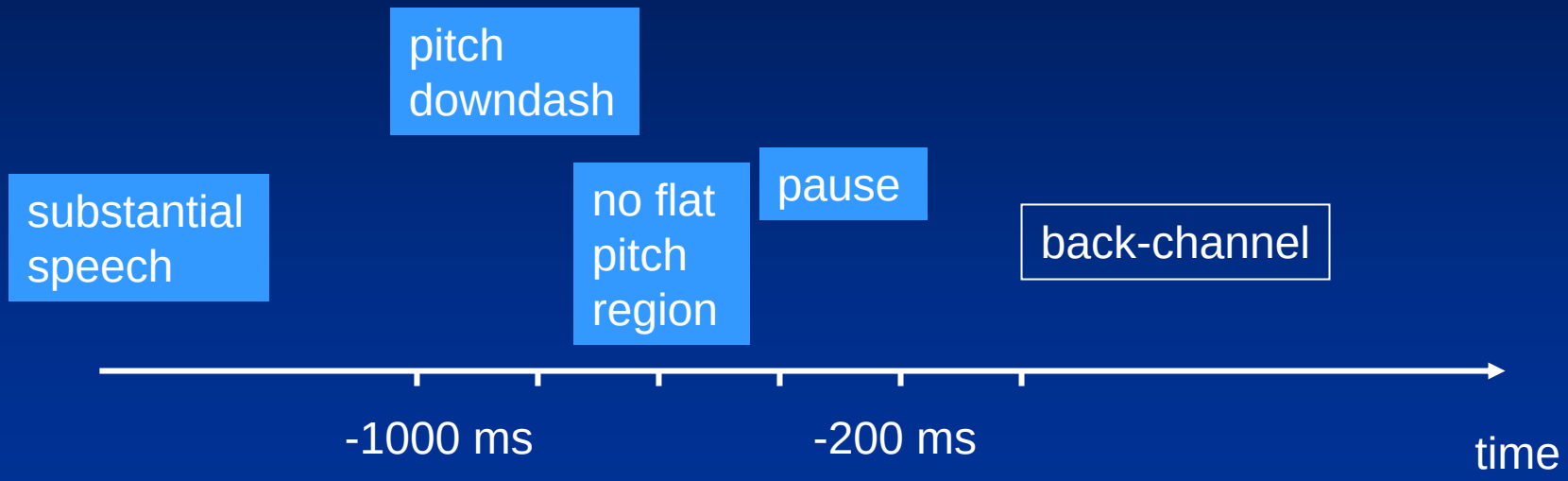
4d: Feature Discovery quantifying perceptions



Since some features are pervasive, hence un-informative,

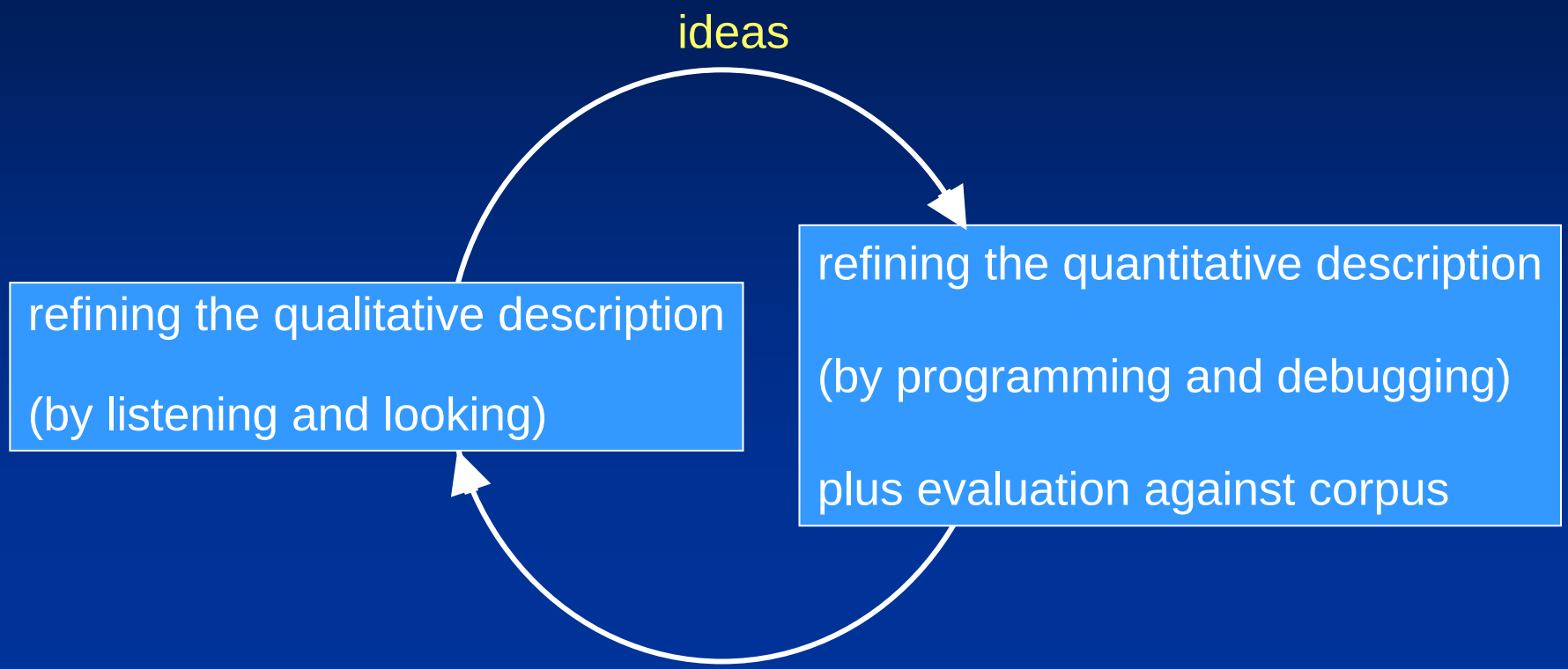
listen casually first, to get familiar with the pervasive patterns.

5. Feature Combination



feature combination is tricky,
since features not always synchronized

6. Hypothesis Refinement



a back-channel cue in Spanish

a false alarm

7. Hypothesis Tuning

hill-climbing suffices

(iff the previous steps were done well)

Resulting rule:

If
- an utterance has lasted at least 1.2 seconds, and
- contains a pitch downdash
    - lasting at least 40 milliseconds, with
    - a pitch drop of at least 0.7% every 10 ms

...

then
- predict a back-channel in response, 300 ms later

8. Evaluation

- by native-speaker acclamation
- by interacting with it
- by correspondence to the corpus

  (51% coverage, 16% accuracy)

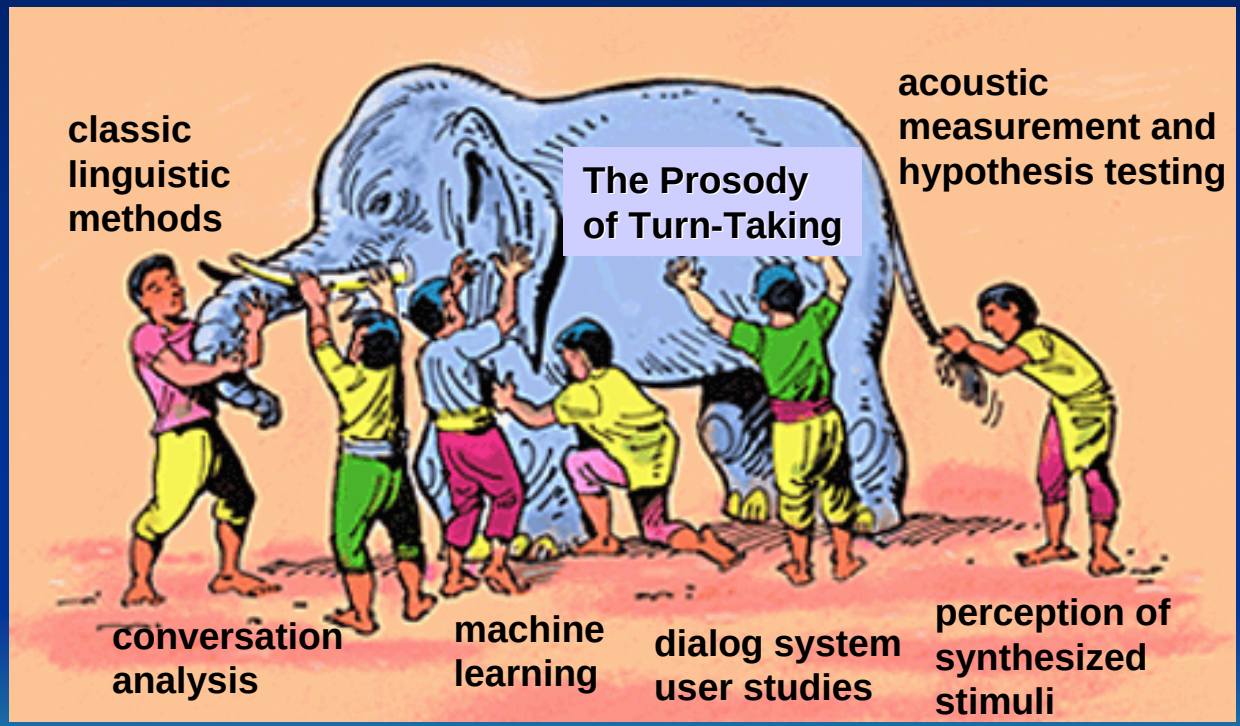
Summary

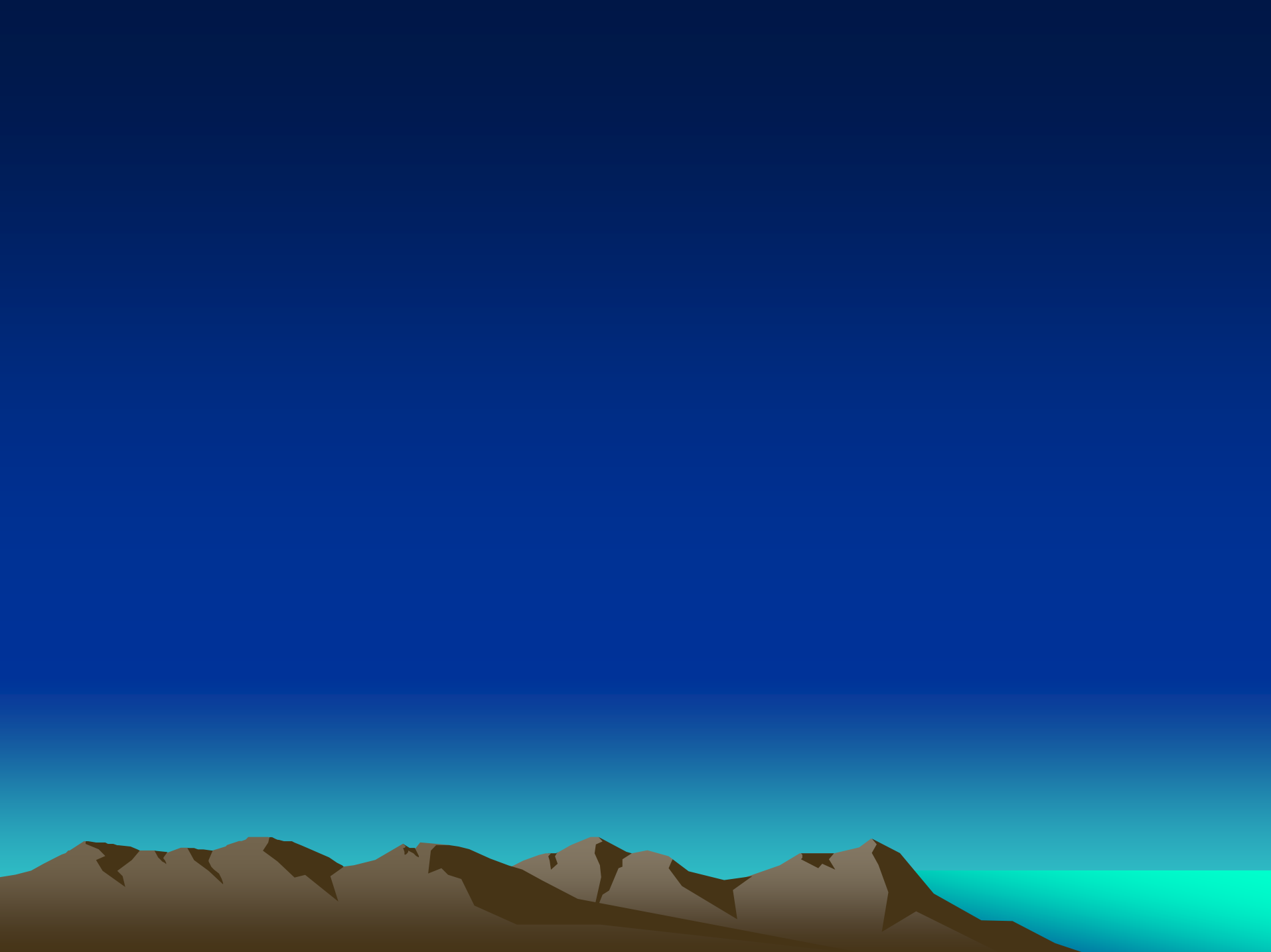
An integrated answer (qualitative + quantitative)

- achievable
- costly (~$90,000)

What Next?

- need more usable tools
- need more feature-rich tools





An Integrated Method

Eight steps to discovery of a prosodic cue

1. Project aims
2. Problem formulation
3. Corpus preparation
4. Feature discovery
5. Feature combination
6. Hypothesis refinement
7. Tuning
8. Evaluation

Fostering Progress

let’s build tools!

let’s look at the same elephants!

Why Engineers Should Care

- Spontaneous speech is different, in ways that affect recognition (Shriberg 2005)
- Dialog systems are pervasive but unnatural and disliked

- Intrinsic scientific interest
- Language teaching applications

3. Corpus Preparation

Corpus size is a Goldilocks question

- labeling too expensive
- can't listen to the data

this corpus is too big

- results not general
- can analyze too deeply

this corpus is too small

- can find a good dialog
- can evaluate properly

this corpus is just right (for us)

Applications

Making Machines more like People

- acknowledgements in tutorial systems
- adapting pace in information-delivery systems
- noticing user reactions in persuasive systems

Making People more like People

- learning to show you're listening ... actively