# Towards a spoken dialog system capable of acoustic-prosodic entrainment

ANDREAS WEISE

6/12/2017

# Terminology

## Entrainment:

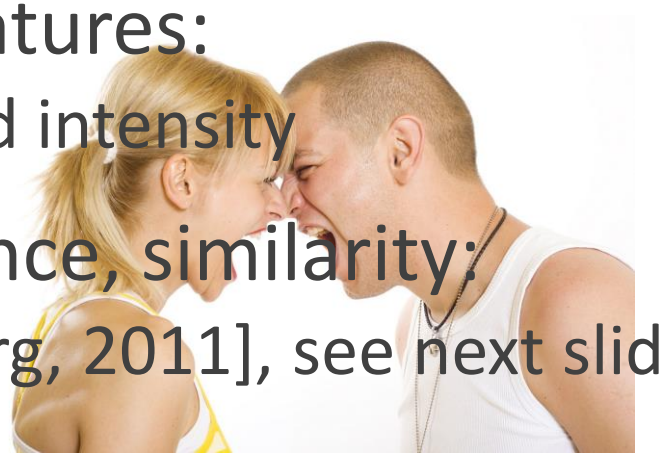
- human tendency to adapt to their interlocutor in conversation
- a.k.a. convergence, alignment, adaptation, accommodation,…

## Acoustic-prosodic features:

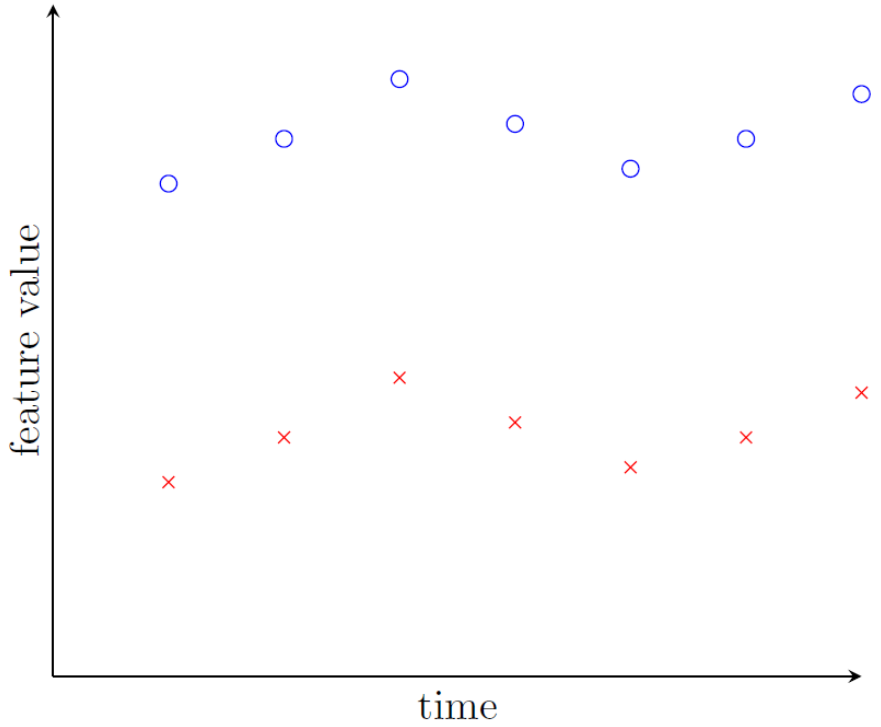
- pitch, speech rate, and intensity

## Synchrony, convergence, similarity:

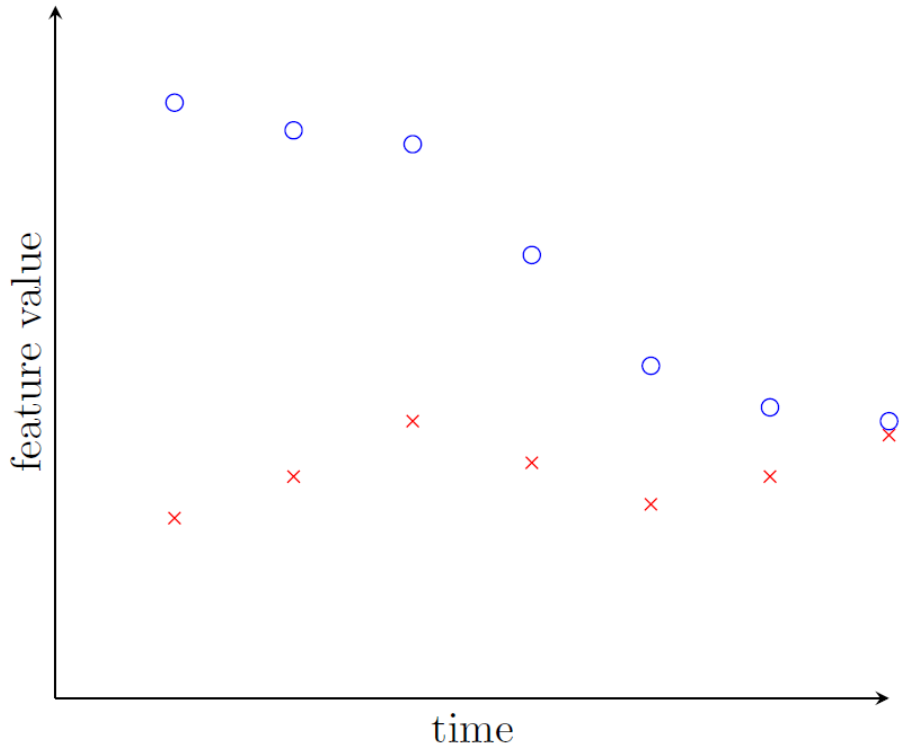
- [Levitan and Hirschberg, 2011], see next slides



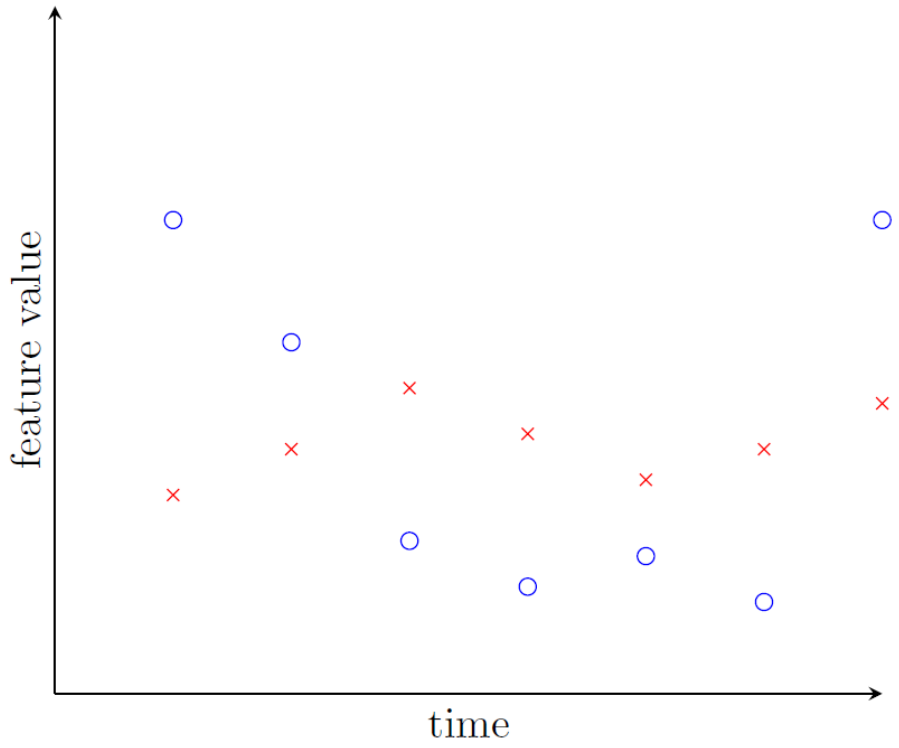
# Synchrony



# Convergence



# Similarity



# Entrainment on pitch, rate, and intensity

[Levitan and Hirschberg, 2011]:

| Feature | Similarity (g) | Similarity (l) | Convergence (g) | Convergence (l) | Synchrony |
|---|---|---|---|---|---|
| Intensity (mean) | ✔ ✔ | ✔ | ✔ | ( ✔ ) | ✔ |
| Intensity (max) | ✔ ✔ | ✔ | | ( ✔ ) | ✔ |
| Pitch (mean) | | ✔ | | ✔ | ✔ |
| Pitch (max) | ✔ | ✔ | | ✔ | ✔ |
| Rate (syll./sec) | ✔ | ✔ | | ( ✔ ) | ✔ |

[Levitan et al., 2012]:

- global similarity for intensity and rate w.r.t. outliers is even greater

# Entrainment on phonetics, lexical choices

[Pardo, 2006]:

- speakers entrain phonetically

[Brennan and Clark, 1996]:

- speakers use common referring expressions
- “recency” and “frequency” of use matter

[Niederhoffer and Pennebaker, 2002]:

- speakers use similar distributions of function words and other categories (“Linguistic Style Matching”, LSM)

# Entrainment on syntax

[Bock, 1986]:

- prepositional / double object structures are reused after “priming”

[Branigan et al., 1999]:

- syntactic priming effect decays very quickly

[Reitter et al., 2006]:

- arbitrary syntactic rules are reused after priming

# Perception-behavior link

[Chartrand and Bargh, 1999]:

- people mimic their interlocutor's mannerisms (shaking leg, rubbing face)
- perceiving a behavior increases the likelihood of engaging in it "perception-behavior link"

→ completely automatic and unconscious

(nonetheless, cognition plays an important role)

# Communication accommodation theory

[Giles et al., 1991]:

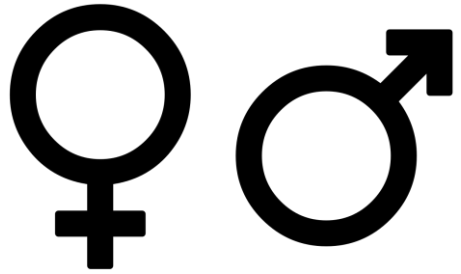
- entrainment is meant to decrease “interpersonal differences”, disentrainment is meant to increase them
- relative social status and power play and important role
- disentrainment is often “intergroup in nature”

→ partially strategic and semi-conscious

# Universality of entrainment

Entrainment has been found in a wide variety of settings:

- in Slovak, Spanish, English, Mandarin [Levitan et al., 2015]
- face to face and remotely [Lubold and Pon-Barry, 2014]
- for both genders [Levitan et al., 2012]

# Effects linked to entrainment

[Nenkova et al., 2008]:

- similarity of high-frequency word use predicts naturalness

[Reitter and Moore, 2007]:

- long-term syntactic entrainment predicts task success, short-term does not

[Lee et al., 2010]:

- similarity of pitch and intensity at turn exchanges correlates with positive affect in “seriously and chronically distressed married couples”

# How can computer science use entrainment?

1. Evaluate human-human conversation
2. Guide users' behavior
3. Entrain to user

# Why would users entrain to systems?

[Natale, 1975]:

- to increase their intelligibility

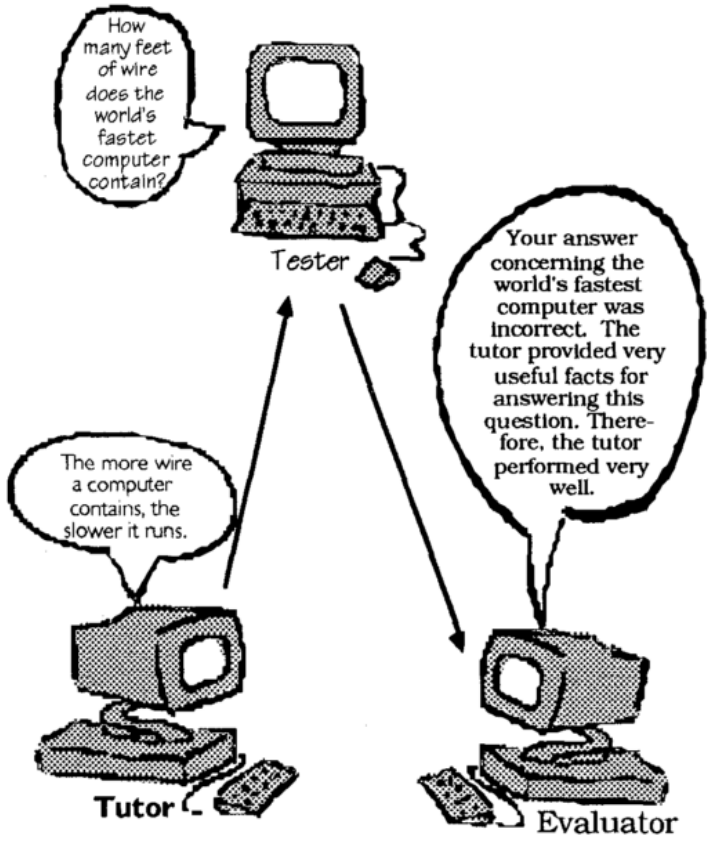
[Chartrand and Bargh, 1999]:

- because they cannot help it

[Nass et al., 1994]:

- because they treat computers as “social actors”

# Why would users entrain to systems?




[Nass et al., 1994] experiment setup

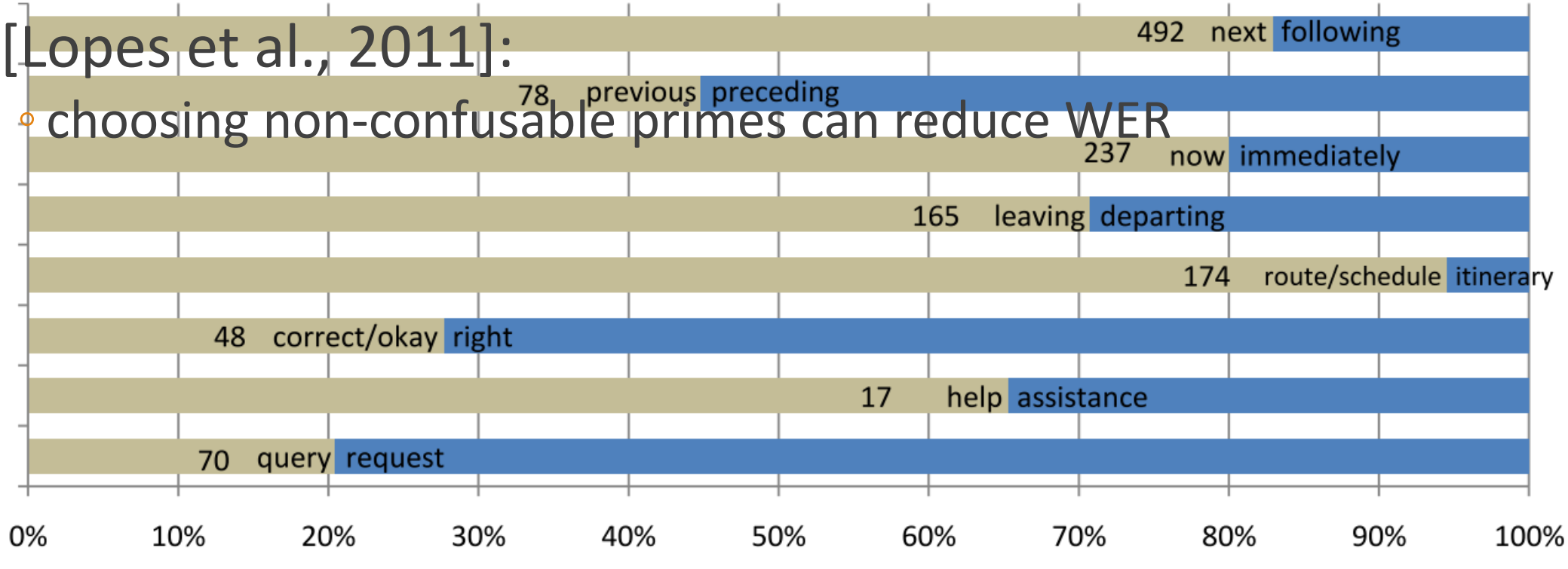
# Users entraining to systems lexically

[Parent and Eskenazi, 2010]:

- Live system: users quickly adopt new vocabulary (especially common words over rare ones)

[Lopes et al., 2011]:

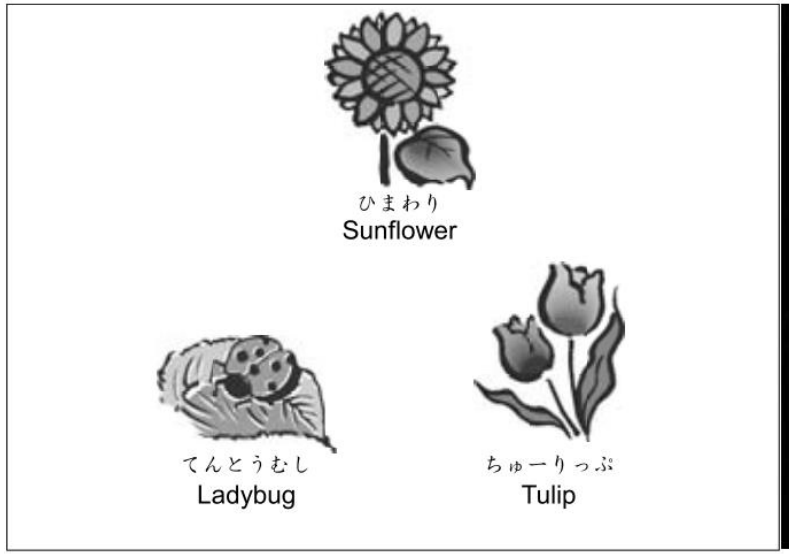
- choosing non-confusable primes can reduce WER



# Users entraining to systems prosodically

[Suzuki and Katagiri, 2007]:

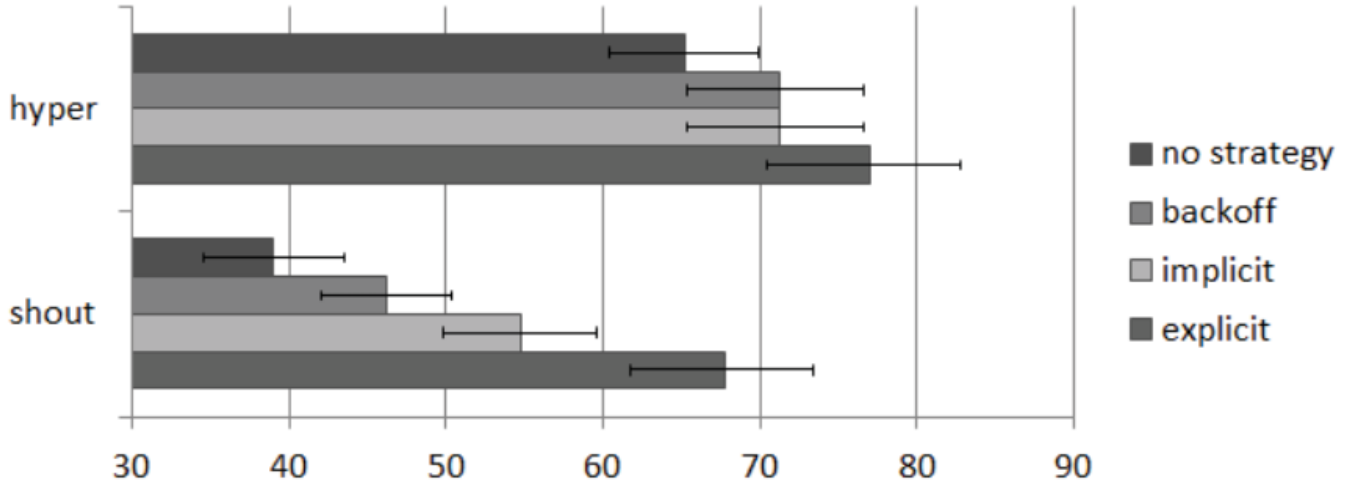
- shorter system response latency elicits shorter user latency; louder system output elicits louder user response




# Users entraining to systems prosodically

[Fandrianto and Eskenazi, 2012]:

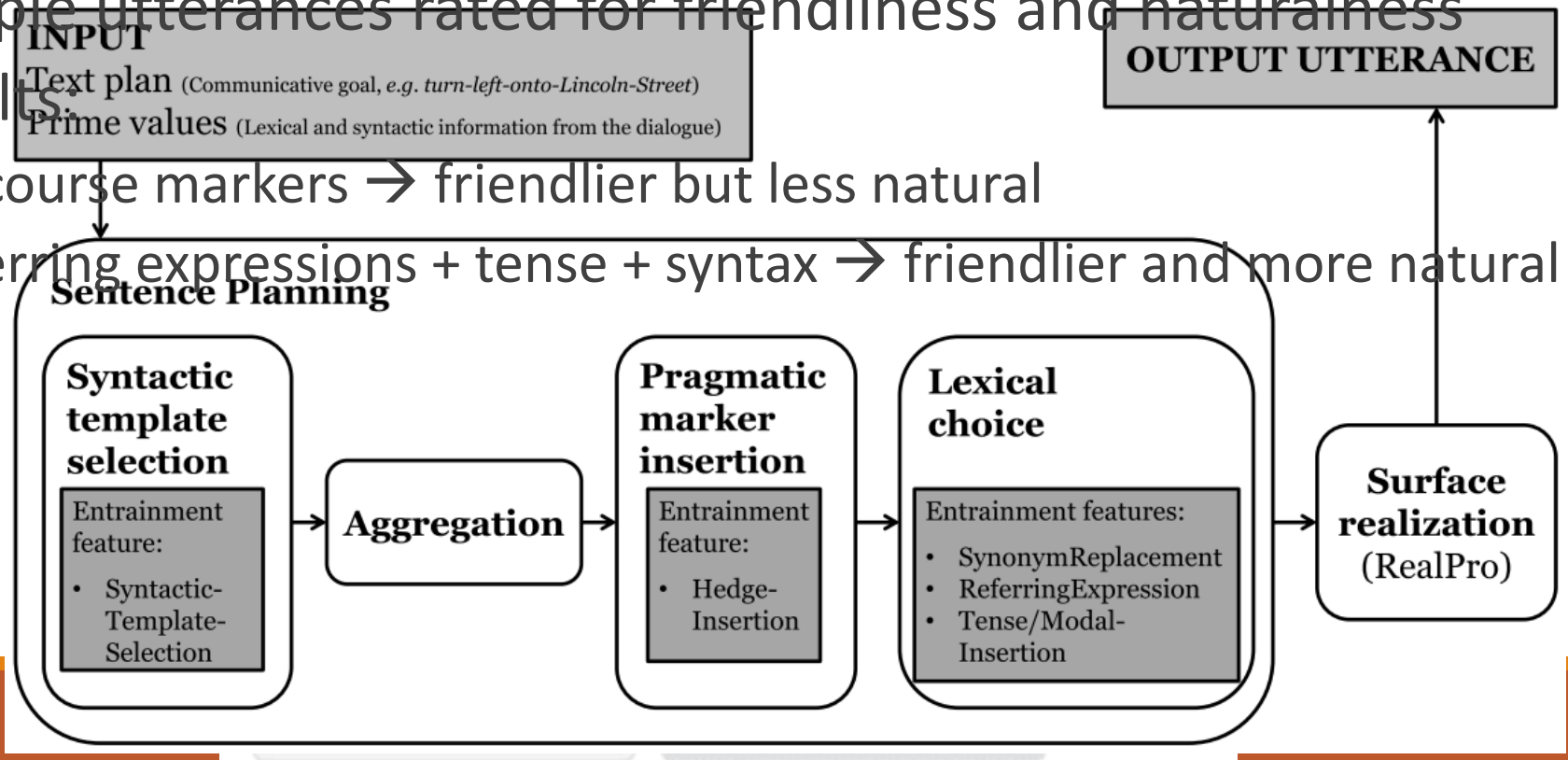
- quieter, faster system output can reduce shouting and hyperarticulation



# Systems entraining to users lexically/synt.

[Hu et al., 2014]:

- system that can entrain lexically and syntactically
- sample utterances rated for friendliness and naturalness
- results:
  - discourse markers → friendlier but less natural
  - referring expressions + tense + syntax → friendlier and more natural

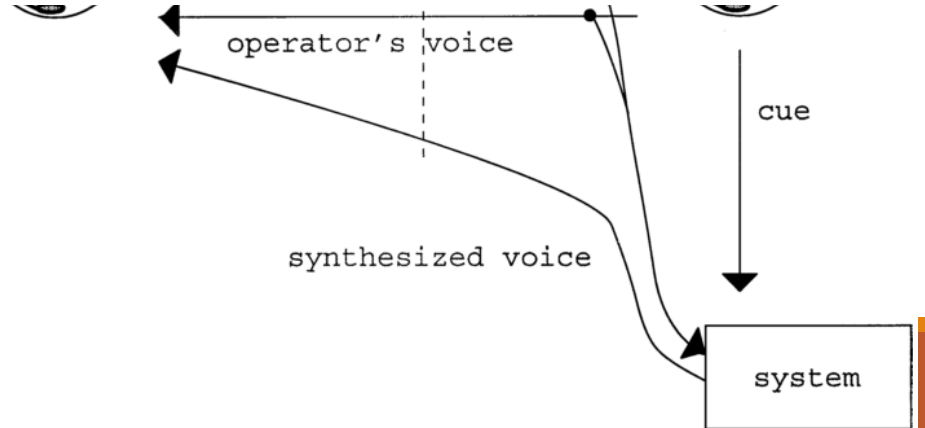


# Systems entraining to users prosodically

[Ward and Nakagawa, 2004]:

- corpus of “directory assistance dialogs”
- predict operator’s speaking rate during dictation of numbers from user’s speaking rate and initial reaction time
- correlation of 0.41 between prediction and real value achieved

| | | |
|---|---|---|
| **operator:** | Directory Assistance, Suzuki speaking. | 1 |
| **user:** | *Hello. I'd like the number for the University of Tokyo, in Bunkyo-ku. Tokyo* | 2 |
| **operator:** | University of Tokyo, Bunkyo-ku, Tokyo? | 3 |
| **user:** | *yes.* | 4 |
| **operator:** | here is the number ... | 5 |
| **operator:** | ...03 3812 2111. | 6 |



# Systems entraining to users prosodically

[Levitan et al., 2016]:

- how to integrate entrainment in an SDS?
- do subjects trust entraining avatars more?
- results:

| Language | Avatar gender | Entrainment | | Baseline | Entrainment × Advice score |
|---|---|---|---|---|---|
| | | Features | Method | | |
| English | Male | Intensity Speech rate | Absolute | Random | + ($p < 0.001$) |
| Spanish | Female | Speech rate | Relative | Disentrain | - ($p < 0.1$) |
| | | | | Constant | no effect |
| Spanish | Female | Speech rate | Relative | Disentrain | - ($p < 0.05$) |
| | | | | Constant | no effect |

# Open questions / proposed work

1. How to get from analytical model to generative model?
2. How to determine the “best” parameters?
3. How to evaluate parameters on subjects?

Thank you!

Questions?