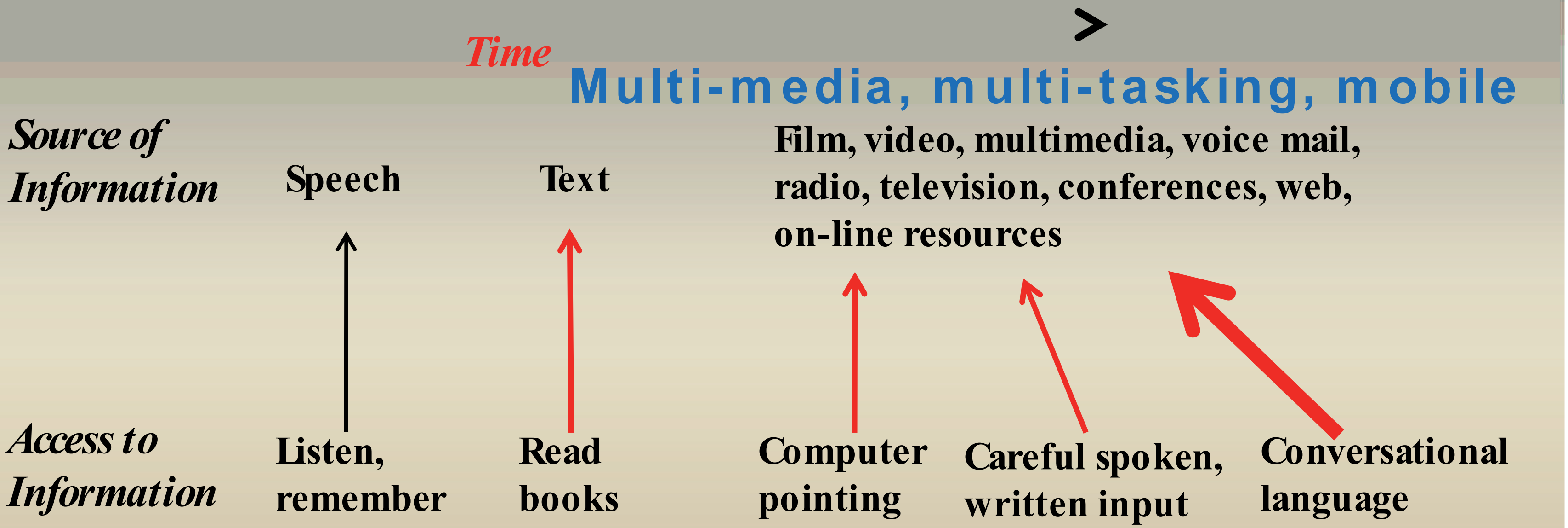
The Growing Impact of Speech Technology on Society

Patti Price, PPRICE Speech and Language Technology

- **Intro: a few million years in about a minute**
- Short review of speech recognition technology
- Speech recognition performance
- Social impact 1 (effect of society on speech)
- Social impact 2 (people vs. technology)
- Progress and challenges

Speech in the Information Age

- Speech & text were revolutionary because of information access
- New media and connectivity yield information overload
- Can speech technology help?



Time >

Multi-media, multi-tasking, mobile

Source of Information

Speech

Text

Film, video, multimedia, voice mail, radio, television, conferences, web, on-line resources

Access to Information

Listen, remember

Read books

Computer pointing

Careful spoken, written input

Conversational language

Speech is social in ways our technology is not. Can it become a complementary partner in with humans?

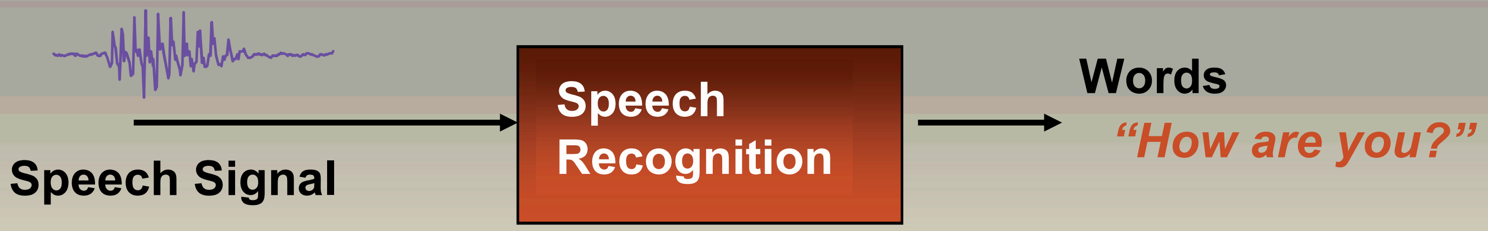
The Growing Impact of Speech Technology on Society

Patti Price, PPRICE Speech and Language Technology

- Introduction: a few million years in about a minute
- **Short review of speech recognition technology**
- Speech recognition performance
- Social impact 1 (effect of society on speech)
- Social impact 2 (people vs. technology)
- Progress and challenges

What is Speech Recognition?

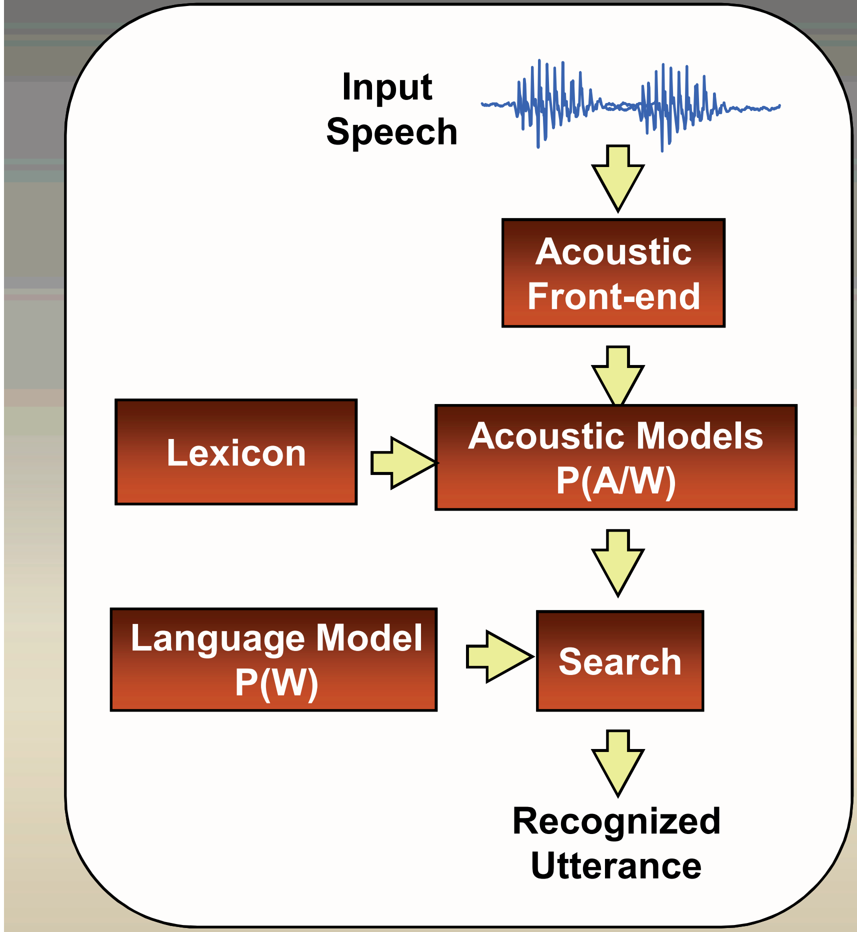
Goal: Automatically extract the string of words spoken from the speech signal



- Speech recognition does NOT determine
  - Who is talker (speaker recognition)
  - Speech output (speech synthesis or speech generation)
  - What the words mean (next two talks will address that)

Updated from 2000 AAAS talk with Joe Picone

Recognition Architectures



- The signal is converted to a sequence of feature vectors based on spectral and temporal measurements.
- Acoustic models represent sub-word units, such as phonemes, as a finite-state machine in which states model spectral structure and transitions model temporal structure.
- The language model predicts the next set of words, and controls which models are hypothesized.
- Search is crucial to the system, since many combinations of words must be investigated to find the most probable word sequence.

Probabilistic modeling requires training data, and match between test and training.

Updated from 2000 AAAS talk with Joe Picone

The Growing Impact of Speech Technology on Society

Patti Price, PPRICE Speech and Language Technology

- Intro: a few million years in about a minute
- Short review of speech recognition technology
- **Speech recognition performance**
- Social impact 1 (effect of society on speech)
- Social impact 2 (people vs. technology)
- Progress and challenges

State of the Art

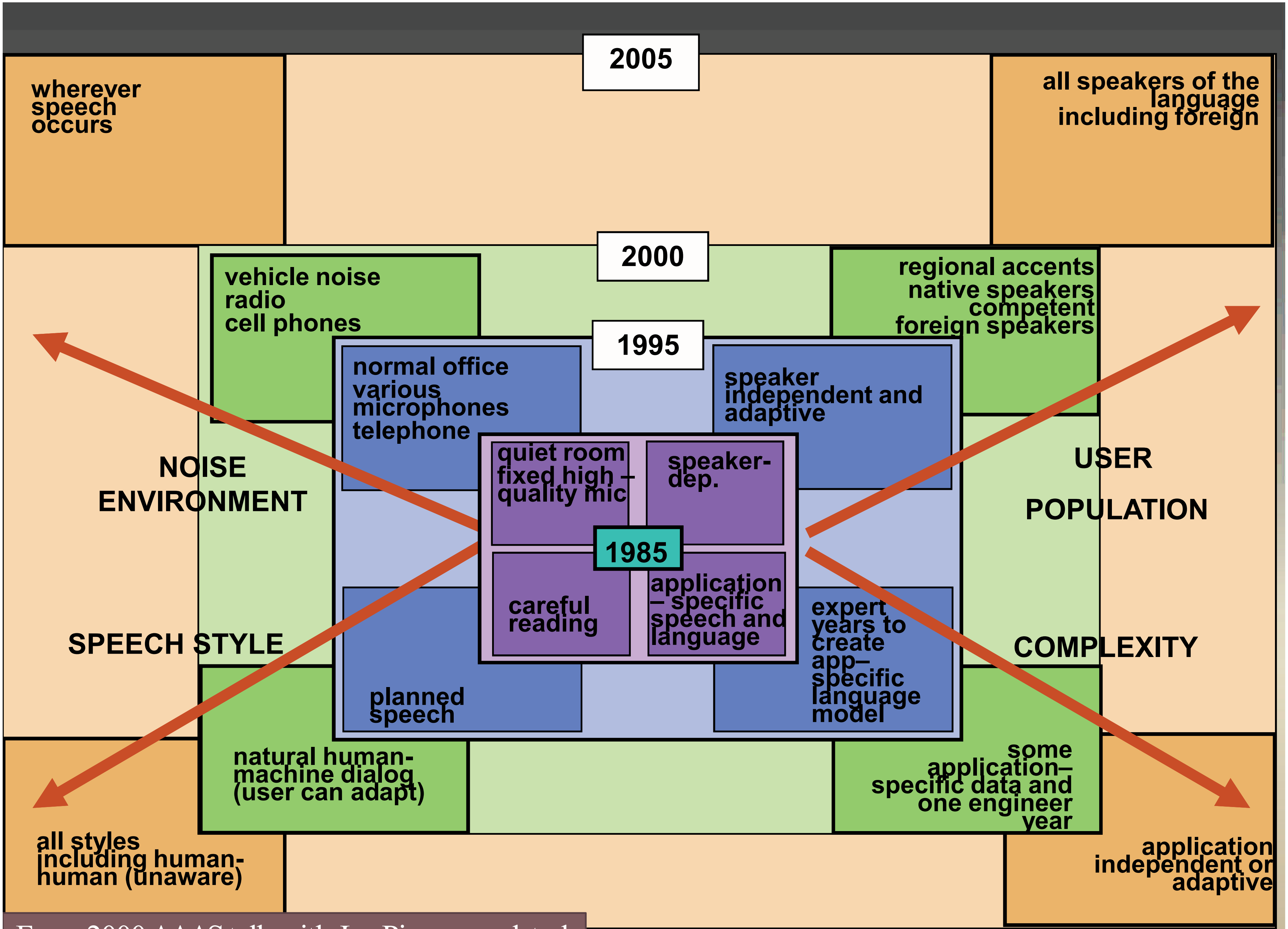
It's easy to get 99% accuracy…
What that means depends on many factors…

- Common evaluations important
- Tasks become more challenging
- Word Error Rate (WER) < 10% is 'acceptable'
- Performance in field ~2x to 4x worse

- What was training set?
- What was test set?
- Were training and test independent?
- Have other systems used same benchmark?

- How large was the vocabulary and the sample size?
- What speakers?
- What style speech?
- What kind of noise?

Updated from 2000 AAAS talk with Joe Picone



2005

2000

1995

1985

NOISE ENVIRONMENT

USER POPULATION

SPEECH STYLE

COMPLEXITY

quiet room
fixed high – quality mic

speaker-dep.

careful reading

application – specific speech and language

normal office
various microphones
telephone

speaker independent and adaptive

planned speech

expert years to create app–specific language model

vehicle noise
radio
cell phones

regional accents
native speakers
competent foreign speakers

natural human-machine dialog (user can adapt)

some application–specific data and one engineer year

wherever speech occurs

all speakers of the language including foreign

all styles including human-human (unaware)

application independent or adaptive

From 2000 AAAS talk with Joe Picone, updated

The Growing Impact of Speech Technology on Society

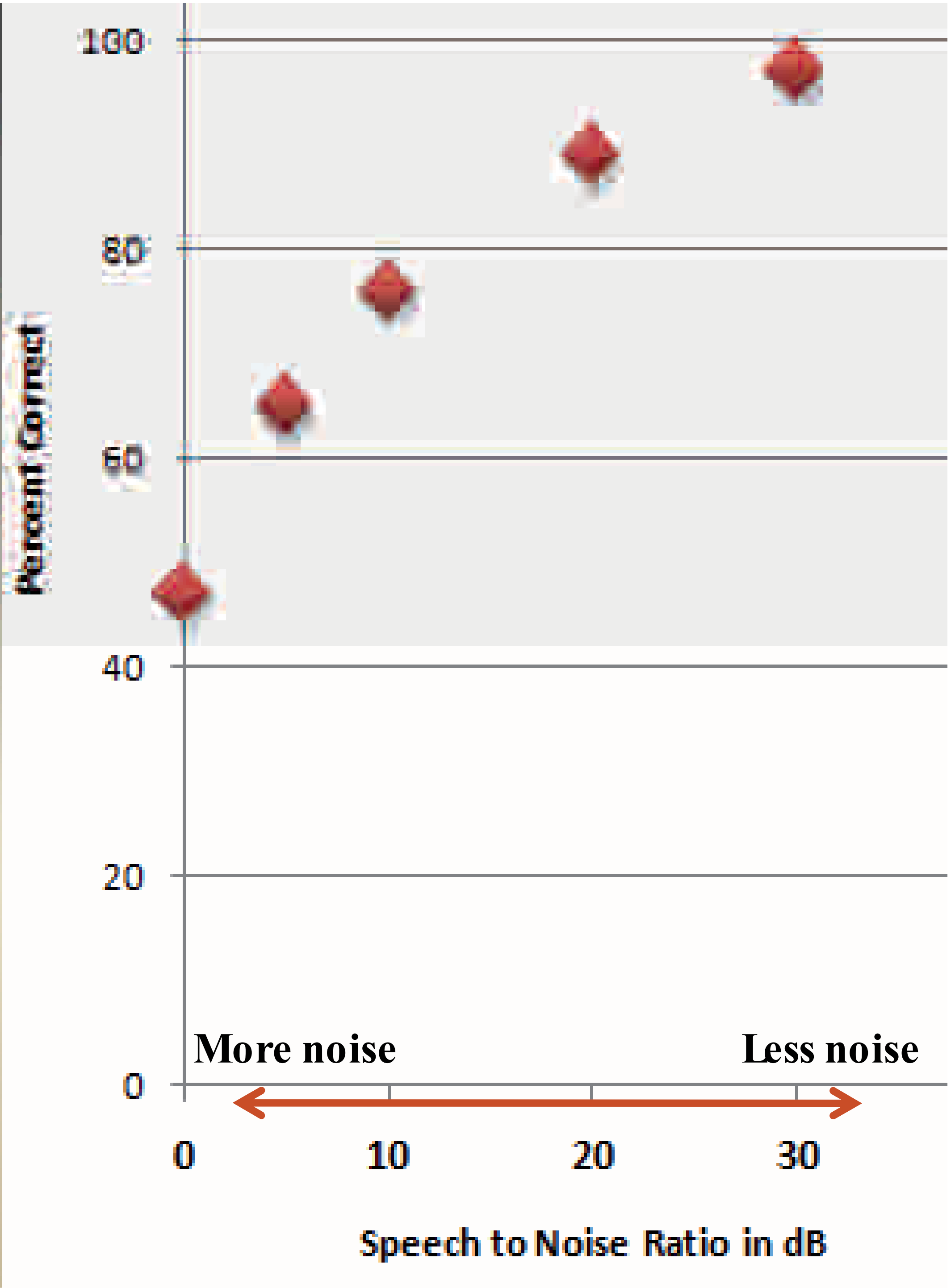
Patti Price, PPRICE Speech and Language Technology

- Intro: a few million years in about a minute
- Short review of speech recognition technology
- Speech recognition performance
- **Social impact 1 (effect of society on speech)**
- Social impact 2 (people vs. technology)
- Progress and challenges

Noise

- We talk wherever we are, but noise degrades speech recognition, **especially speech-like noise**
- Recognition of vowel-consonant-vowel consonants
- Additive speech-shaped noise
- Other noise also degrades speech recognition (speech, telephone channel, etc.)
- The world is getting noisier

Data from Sroka & Braida, Sp. Comm., 2005

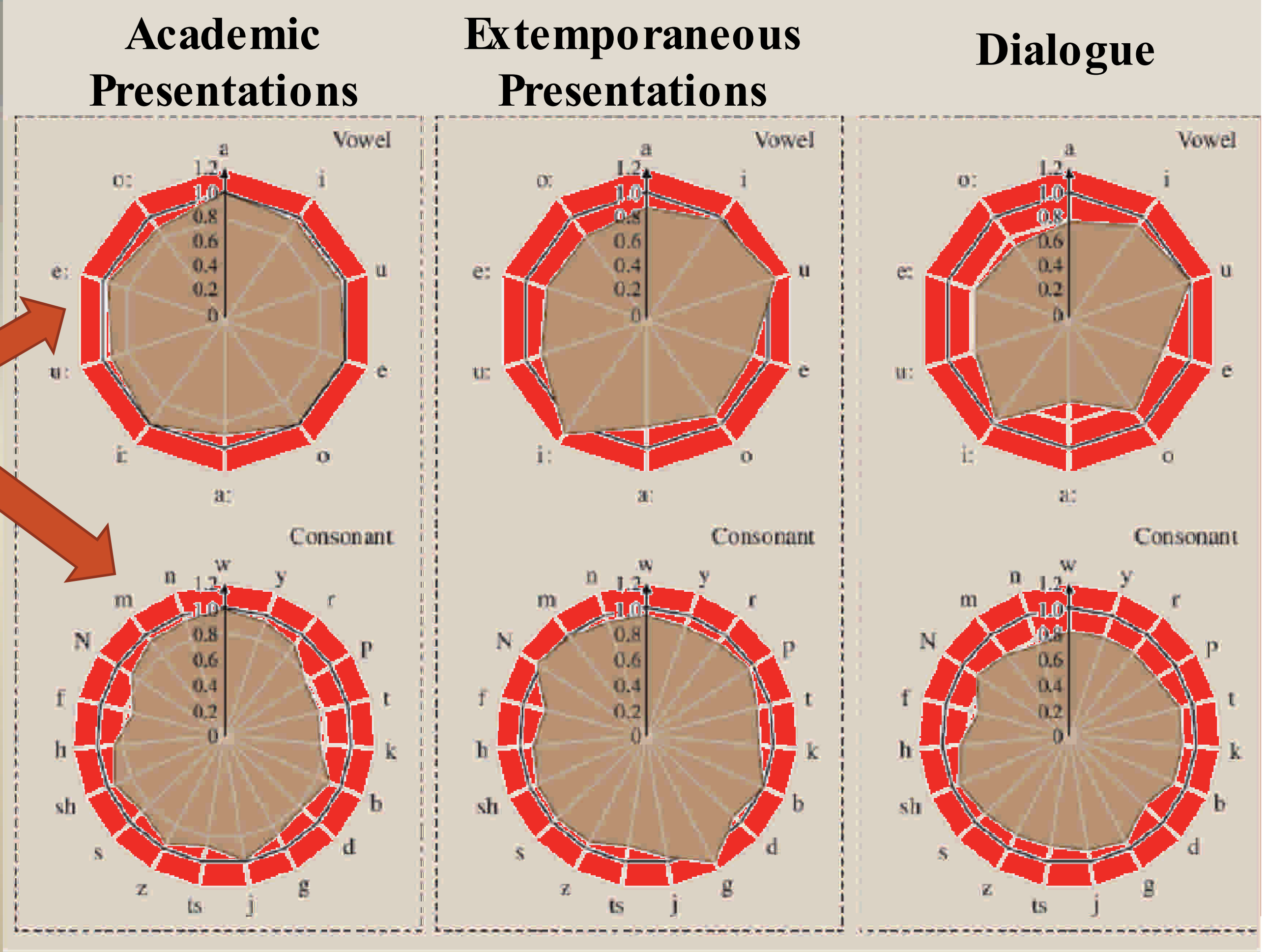


from Furui & Kawahara, Springer Handbook, 2008

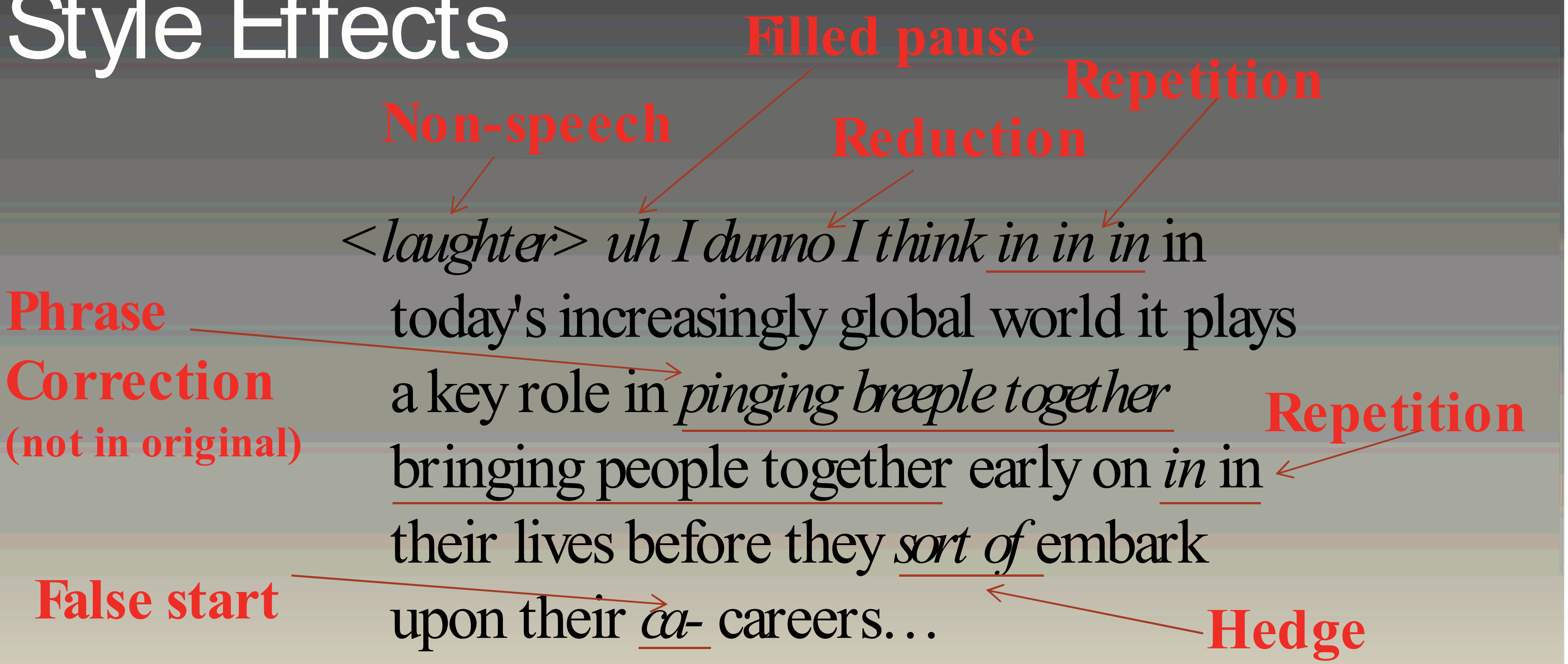
Speech Style

- **We don’t always speak carefully…**
- Orange outlines are read speech versions of originals
- 10 Japanese speakers
- Beige inside portion is relative reduction for vowels and consonants

Accuracy shrinks as reduction increases



Style Effects



Filled pause

Repetition

Non-speech

Reduction

<laughter> uh I dunno I think in in in in today's increasingly global world it plays a key role in *pinging breeple together* bringing people together early on *in* in their lives before they *sort of* embark upon their *ca-* careers…

Phrase Correction (not in original)

Repetition

False start

Hedge

In today's increasingly global world it plays a key role in bringing people together early on in their lives before they embark upon their careers.

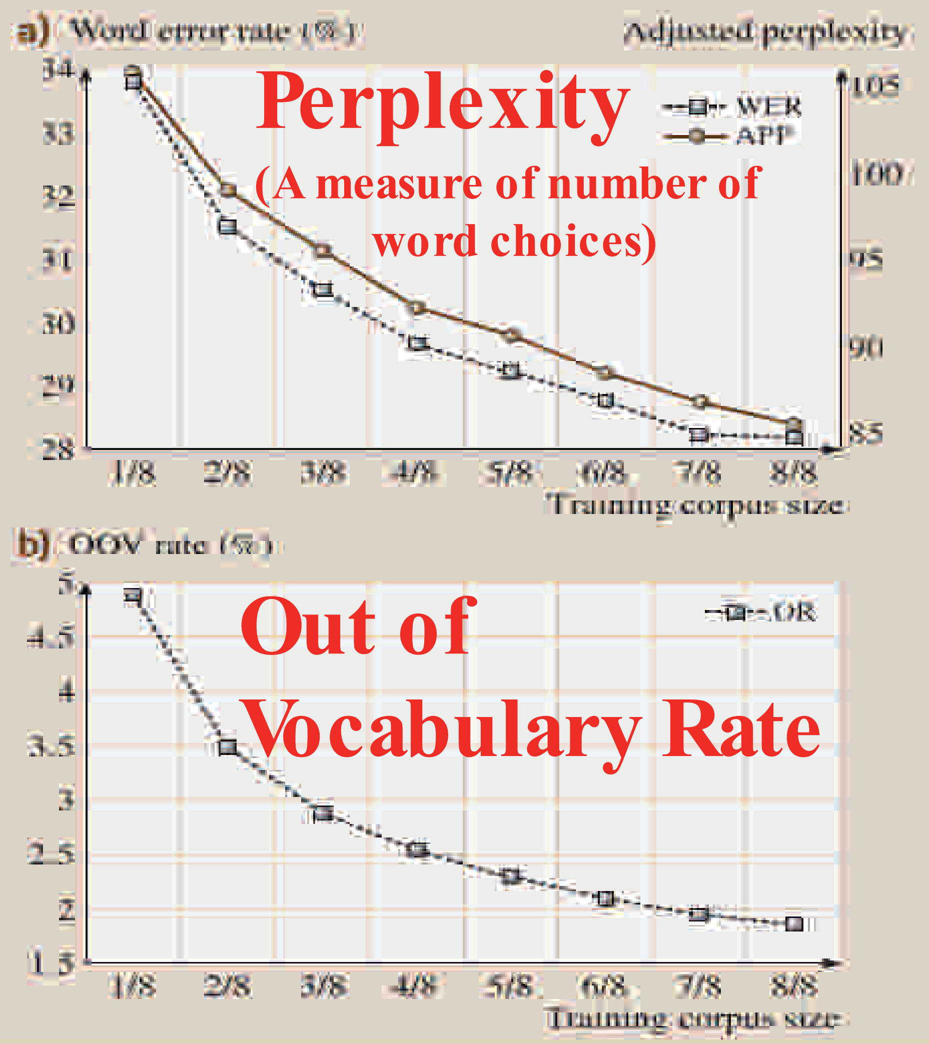
Overlapping speech not illustrated

When disfluencies are removed, spontaneous speech had same recognition error rates as read speech.
Butzberger et al. 1992

from http:/ / videolectures.net/ eae08_barlett_interview/

Some Complexity Factors

from Furui and Kawahara, Springer Handbook, 2008



Perplexity

(A measure of number of word choices)

Out of Vocabulary Rate

Error rates rise with

a) Less constraint on word choices
b) More unknown words

We wanted the knowledge navigator

We need to do better

- Instead… we got gethuman.com
- And … Julie

From Apple 1987 visionary video

From Saturday Night Live, April 2006

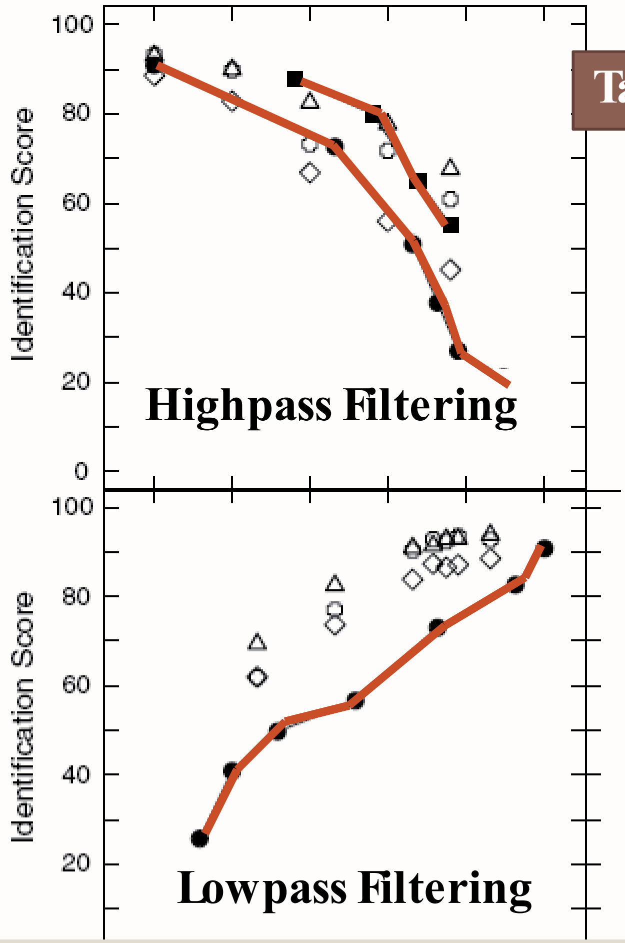
The Growing Impact of Speech Technology on Society

Patti Price, PPRICE Speech and Language Technology

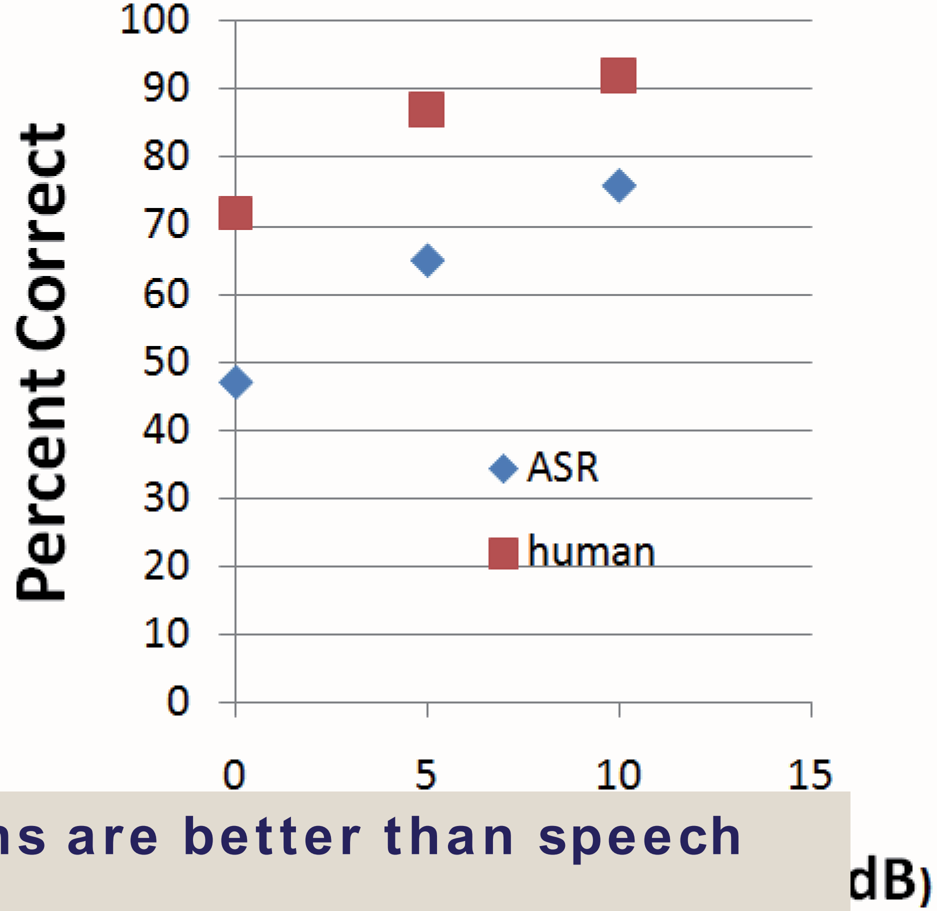
- Intro: a few million years in about a minute
- Short review of speech recognition technology
- Speech recognition performance
- Social impact 1 (effect of society on speech)
- **Social impact 2 (people vs. technology)**
- Progress and challenges

Human Recognition vs. ASR

From Sroka and Braida 2005



Task: Label consonants in CVC and CV syllables

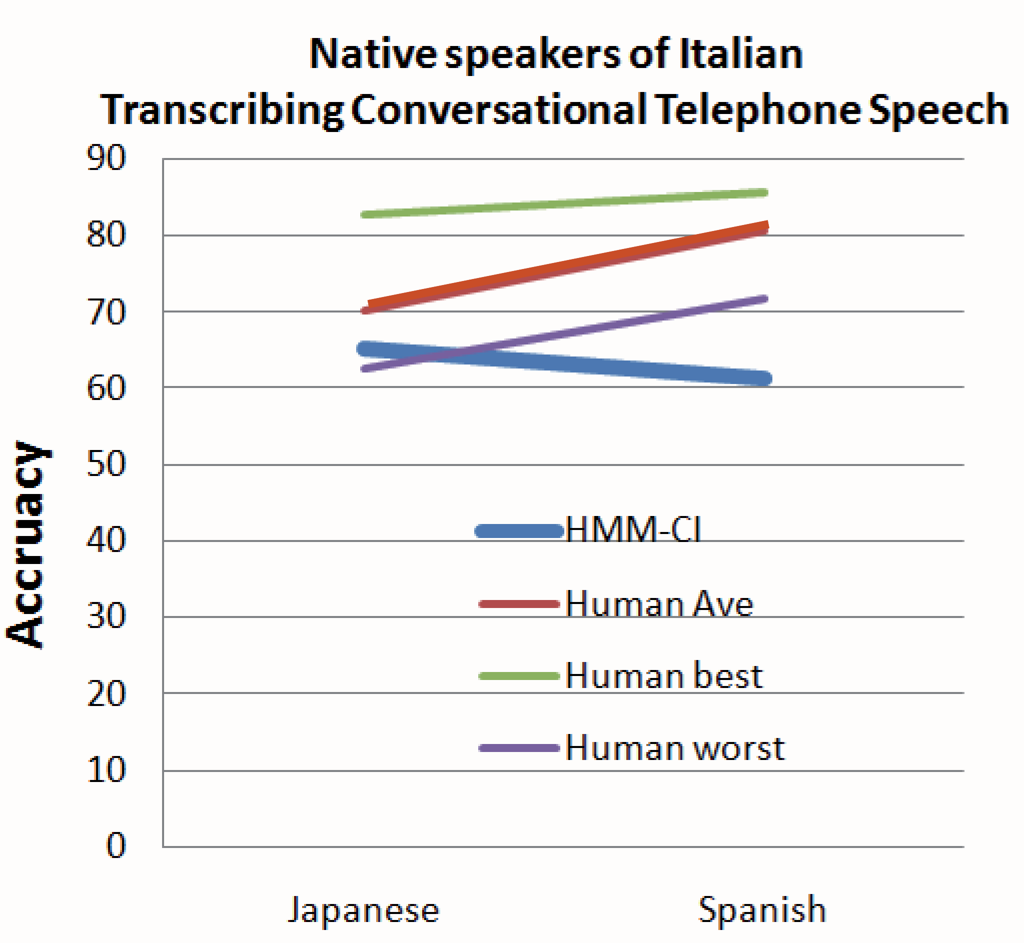


In most noise situations, humans are better than speech recognition but:

- About the same with high pass filtering
- ASR seems better with low pass filtering (data not same)

Human Recognition vs. ASR

- Remove 'language model' but still use natural speech
- Phonetic inventories are similar for Italian, Japanese and Spanish
- Simple ASR is about the same as the worst of the 15 Italian transcribers
- (Spanish and Italian are close in phonotactics)



Data from Shen et al., Interspeech 2008

ASR is sometimes better than people…

Large vocabulary tasks where people may not know the vocabulary
(e.g., thousands of names of companies in stock trading)

Small vocabulary tasks where memory plays a role
(e.g., transcribing sequence of 12 digits tracking numbers)

Artifacts of poorly designed experiments
(e.g., testing on training data, correlational data that helps…)

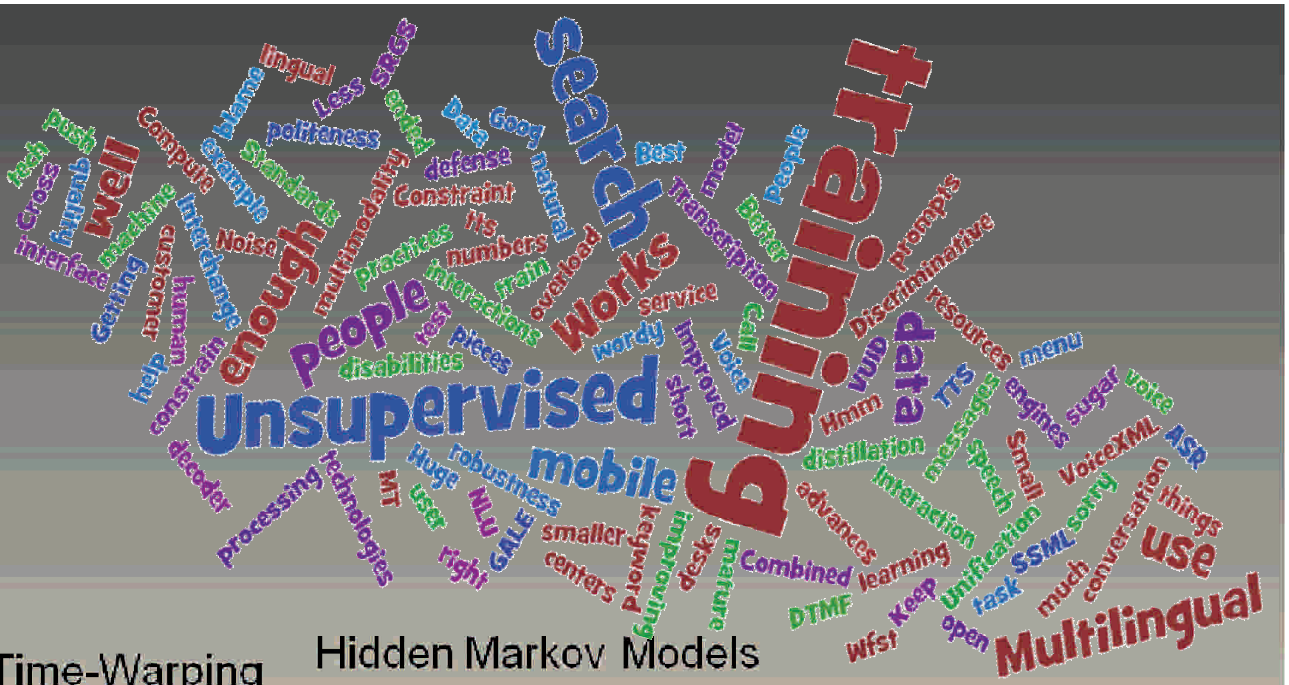
But generally people are more robust, flexible, adaptable… to situations that are normal human variability

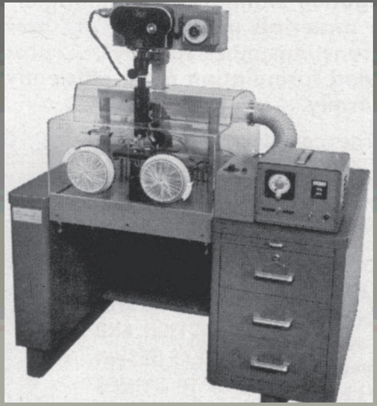
The Growing Impact of Speech Technology on Society

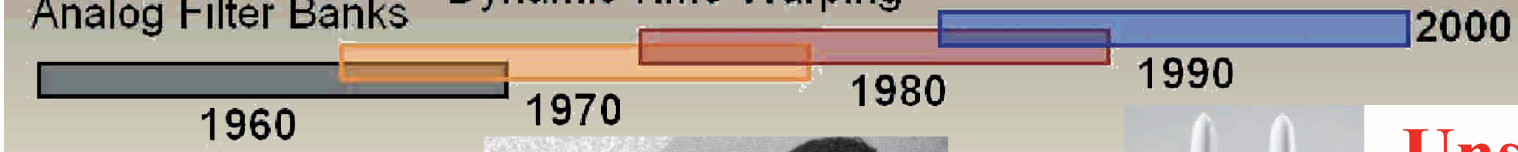
Patti Price, PPRICE Speech and Language Technology

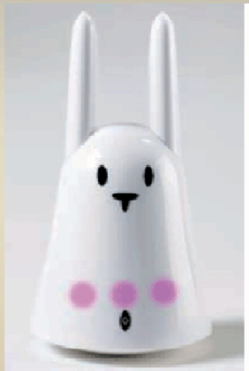
- Intro: a few million years in about a minute
- Short review of speech recognition technology
- Speech recognition performance
- Social impact 1 (effect of society on speech)
- Social impact 2 (people vs. technology)
- **Progress and challenges**

Progress









Unsupervised training

New languages

User Interface

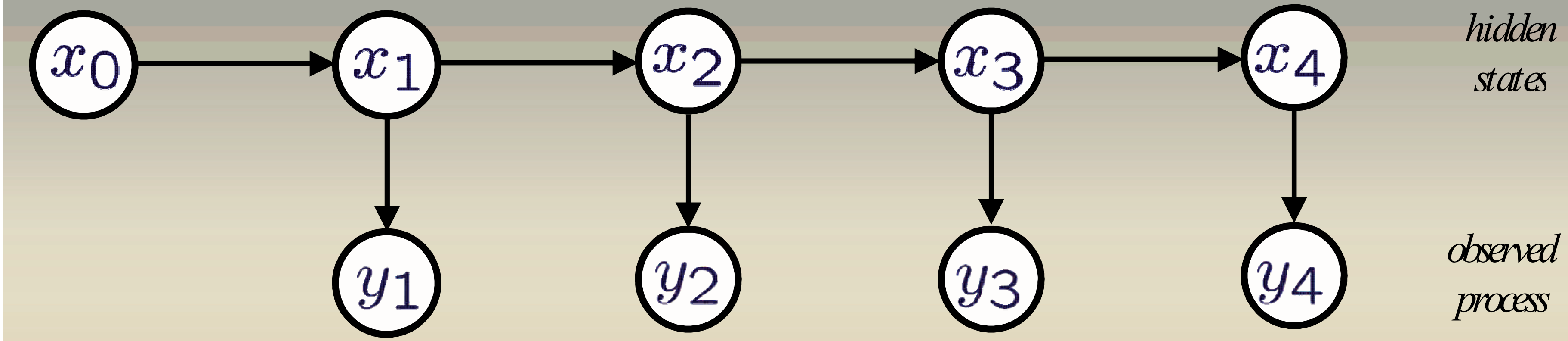
Distillation, NL

From 2000 AAAS talk with Joe Picone, **updated, images from** ***Science***

The Challenge of Hidden Markov Models

- Few realistic time series directly satisfy the assumptions of Markov processes:

"Conditioned on the present, the past & future are independent"



From Vas Digalakis

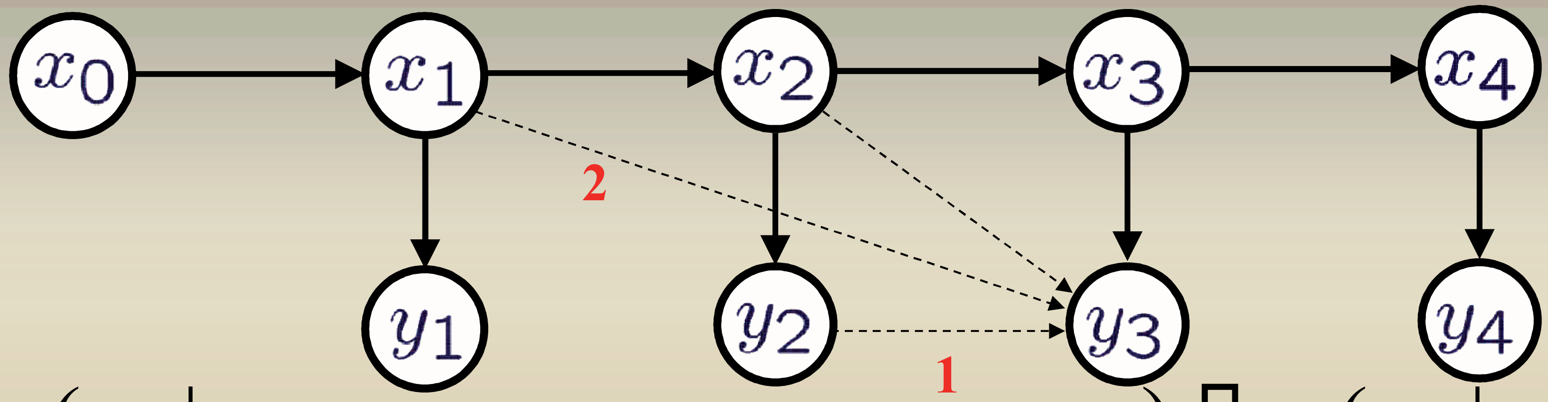
Conditional Independence Assumptions

1. Observations are conditionally independent given the HMM state

   We know this is false. It is a poor model.

2. Observations are conditionally independent of past states given the current state

   This is also false. We make up for it with trigrams, etc.



$$p(y_t \mid x_1, x_2, \ldots, x_t, y_1, y_2, \ldots, y_{t-1}) \approx p(y_t \mid x_t)$$

But that's just adjusting the boundary and not fixing the model…

From Vas Digalakis

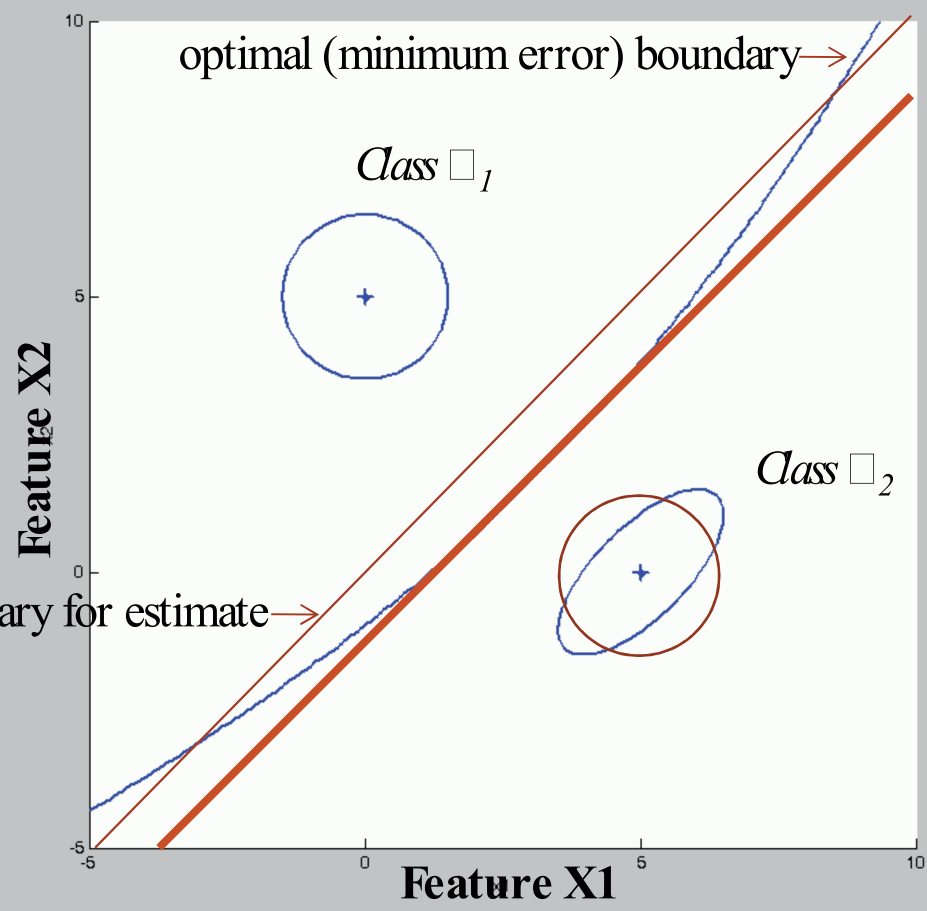
Challenge: A Simple 2-Class Problem

- Simple two-class, two-feature problem, with priors

$$p(\square_1) = 1/3 \quad p(\square_2) = 2/3$$

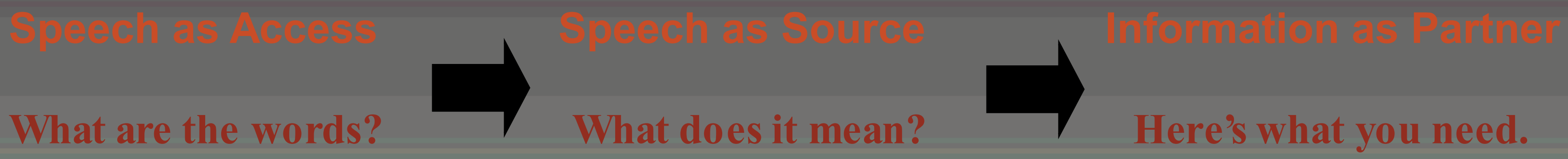
Using an exponential weight for the priors:

$[p(\square_1)]^a \quad [p(\square_2)]^a$



From Vas Digalakis

Trends

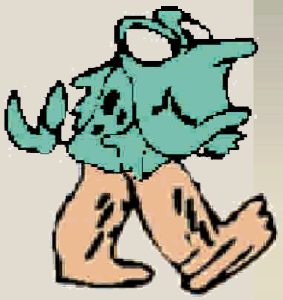

Speech as Access → Speech as Source → Information as Partner

What are the words? What does it mean? Here's what you need.

Can our technology to help with information overload?

150 years after Darwin's famous British A. A. S. debate

- How do we get ourselves out of a niche as we learn more?
- Do we wait for a mutation? Wait for the old models to die?
- It's tough to start completely from scratch, but:
  - We can borrow useful mutations from others...
  - We can partner with our technology



Limitations to speech technology arise from the evolution of speech as a social construct

- Constrained by evolutionary history, production, perception, cognition
- Balancing needs of both speaker and hearer
- At the least, an existence proof, at best, a model we can improve on
- Speech technology lacks social skills; what do we do?

2009 Informal Survey

Sondra Ahlen

Fil Aleva

Francoise Beaufays

Joe Campbell

Rolf Carlson

Gerard Chollet

Mike Cohen

Martin Cooke

Deborah Dahl

Vas Digalakis

Farzad Ehsani

Sadaoki Furui

Juan Gilbert

John Makhoul

Bill Meisel

Roger K. Moore

Ariane Nabeth

Joe Picone

Alex Rudnicky

Paul Sawyer

Malcolm Slaney

Michel Stella

Gary Strong

Orith Toledo

Carl Turner

Fuliang Weng

Steve Young

THANKS!!!

Predictions Survey:

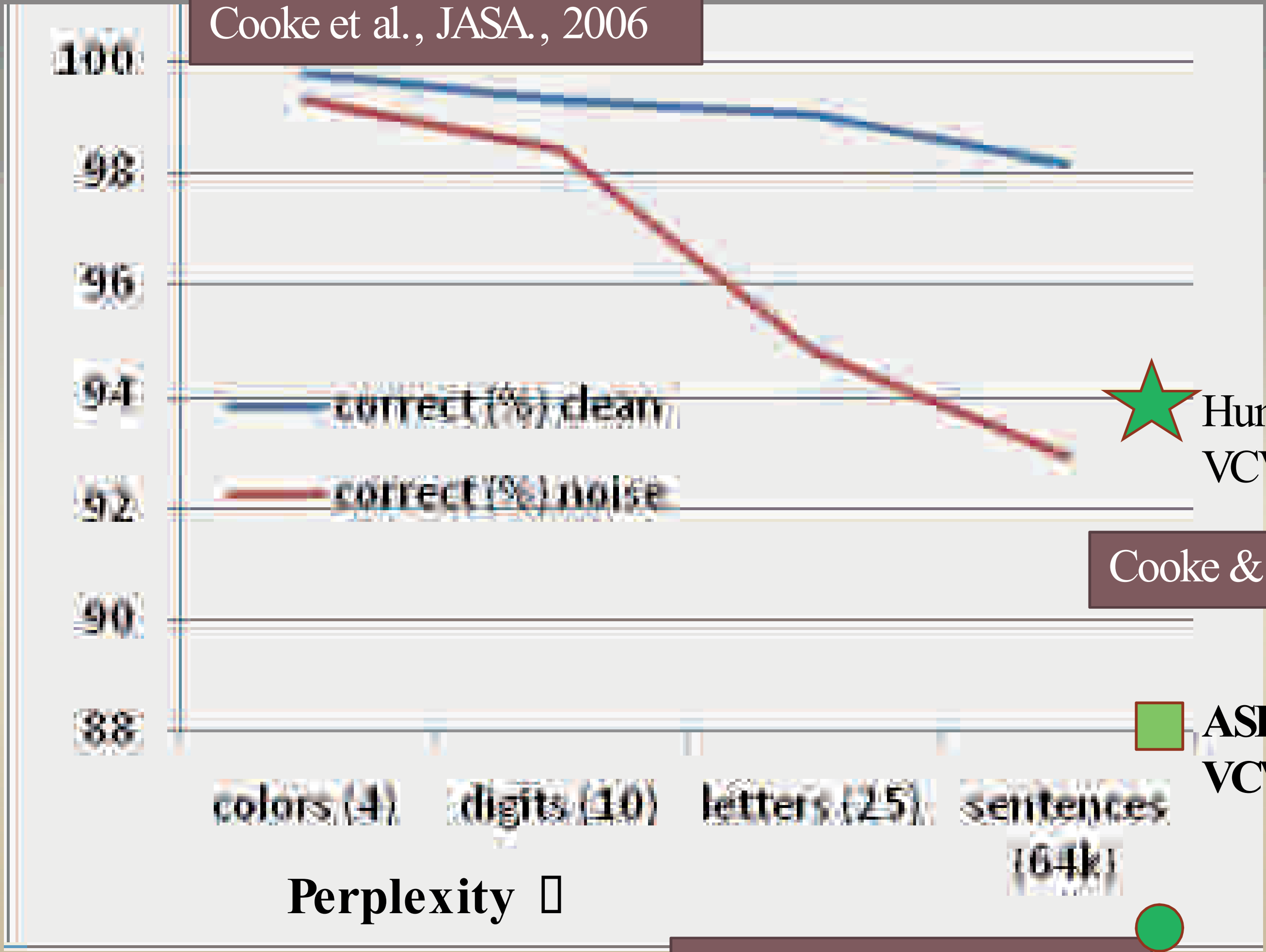
- 1997, 2003, 2009 speech conference attendees
- Participants suggest year (or “never”) when each statement might become true
- Example: 1998, Kurzweil “By 2009 most routine business transactions take place between a human and a virtual personality (including an animated visual presence that looks like a human face).”
- **In fact, in each sample the future gets farther away!**
- Except for ‘year when no speech research needed’ (which is always “never”)

From Roger K. Moore, 2005

Human vs. ASR Recognition in Noise, Clean

- Noise is speech shaped noise at 6, 4 and 2 dB SNR

Cooke et al., JASA., 2006



★ Human, consonants in VCV, clean

Cooke & Schorenborg IS 2008

🟩 **ASR, consonants in VCV, clean**

● NNS, VCV clean

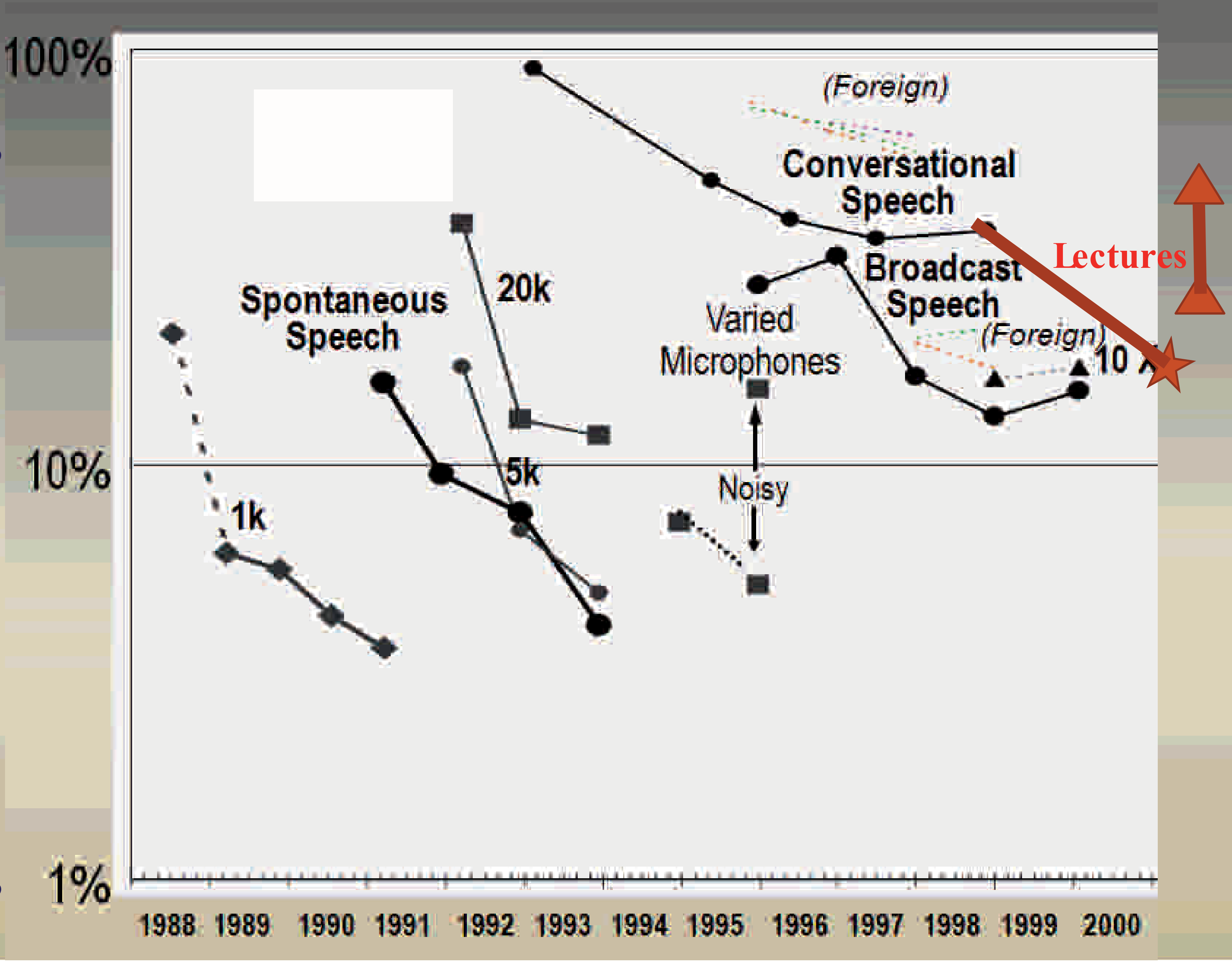
Lecumberri et al., IS 2008

Error Rates

From 2000 AAAS talk with Joe Picone, updated

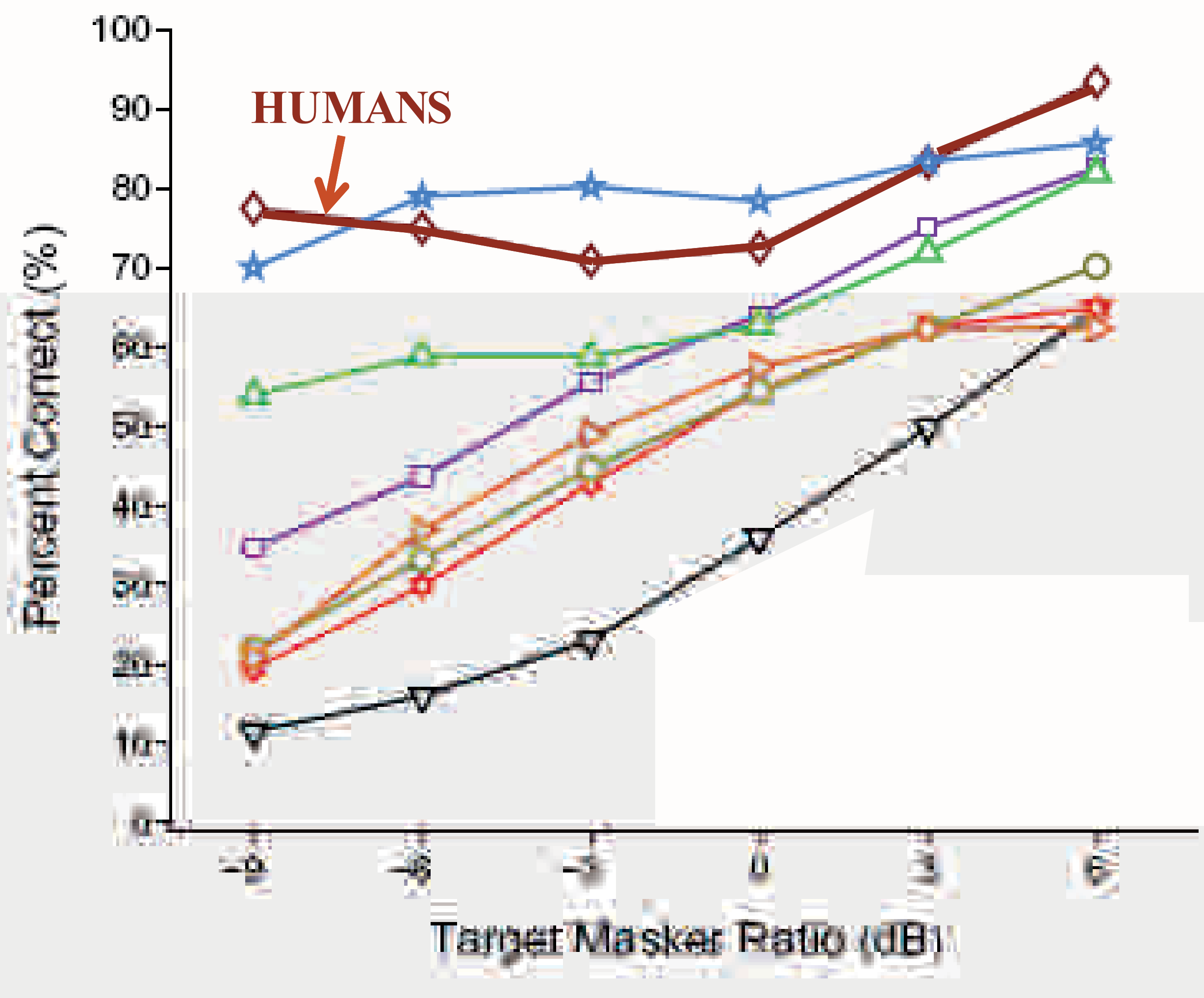
More errors

Fewer errors



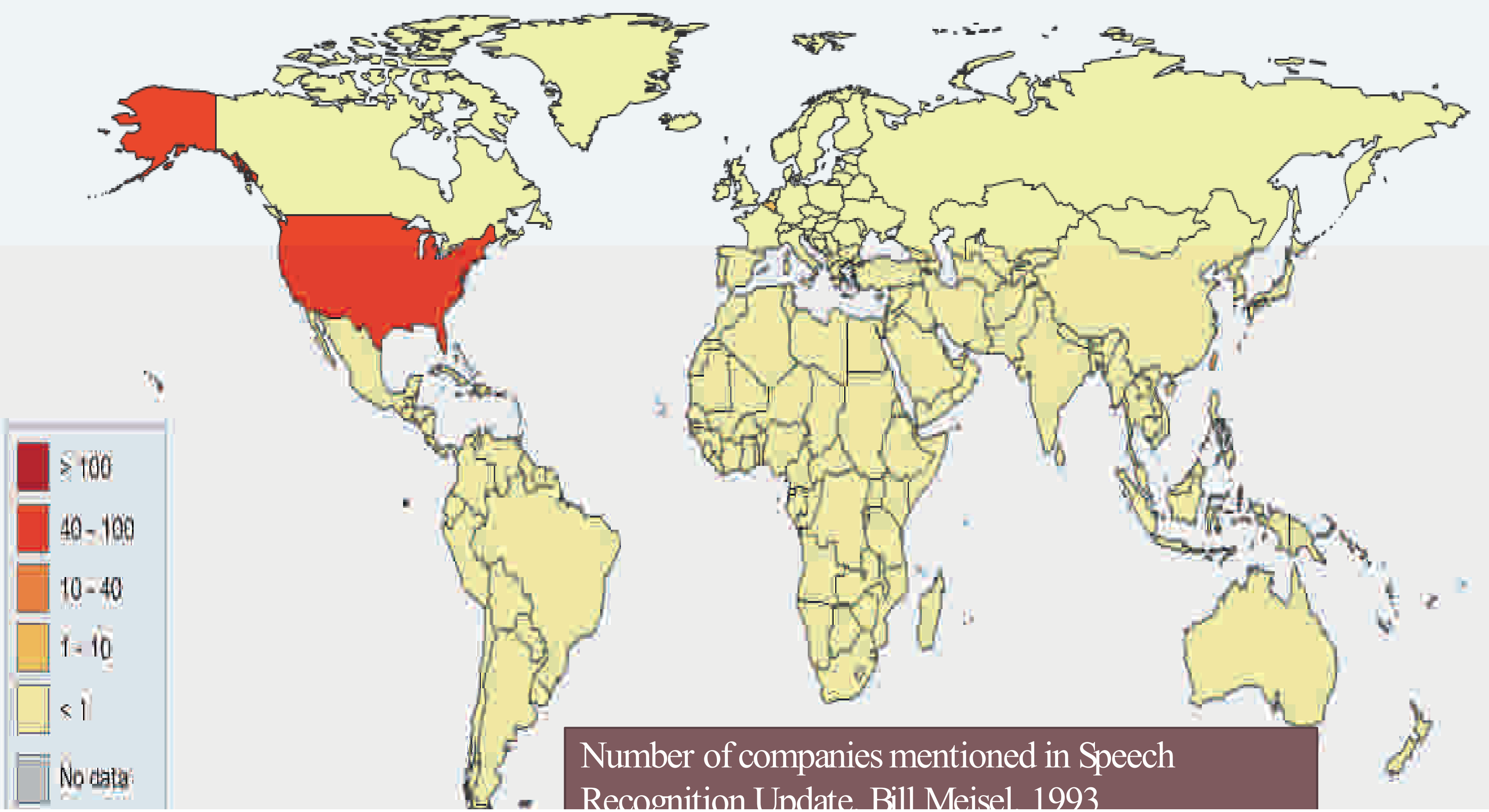
Human Recognition vs. ASR

- Task: Identify key words in sentences masked by similar sentences
- Humans do fairly well, even when masker is louder target
- Some systems took advantage of the fact that the target was always at a fixed level…



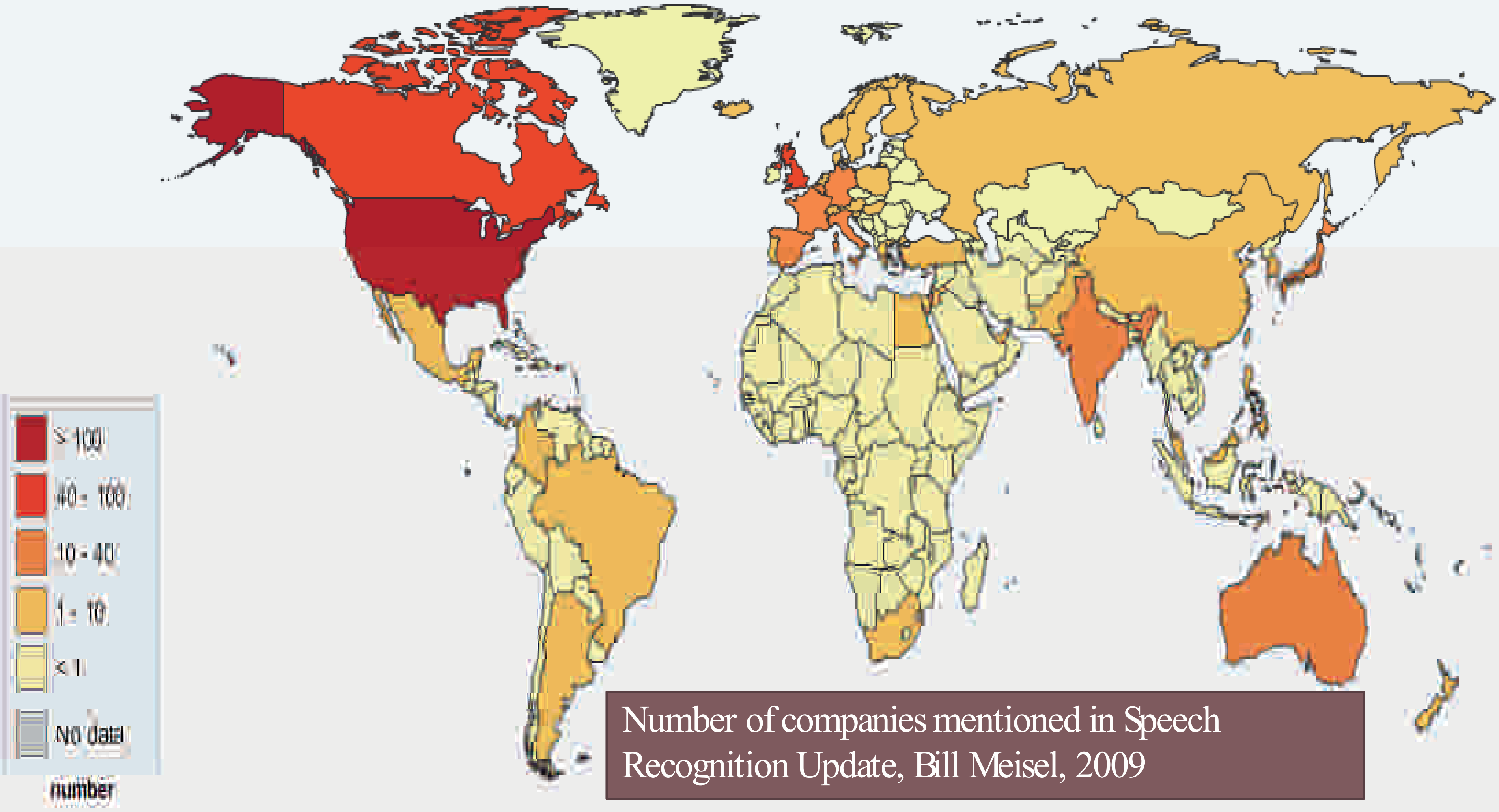
Cooke, Hershey and Rennie, Sp. Comm. 2010

Companies 1993



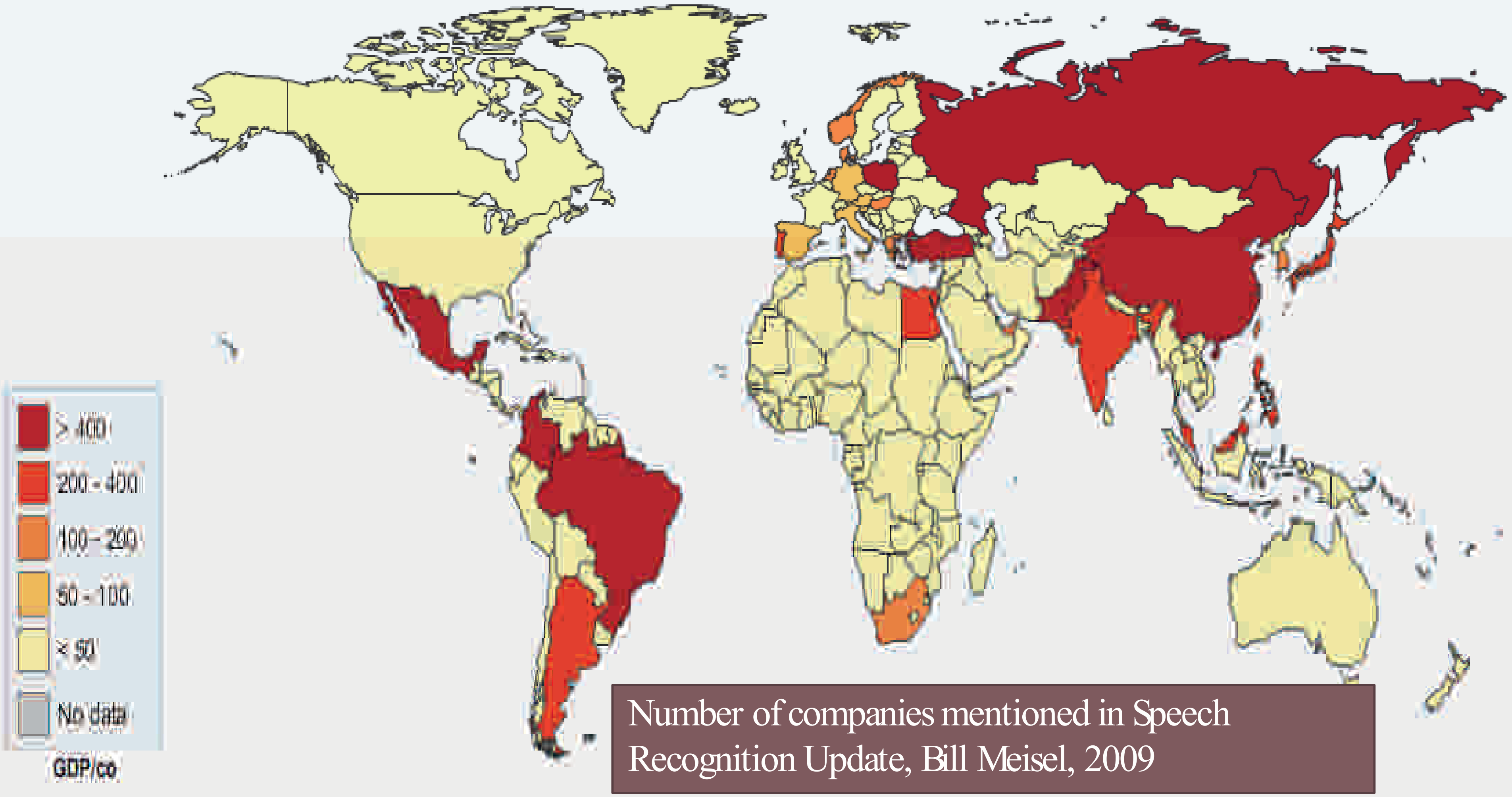
Number of companies mentioned in Speech Recognition Update, Bill Meisel, 1993

Companies 2009



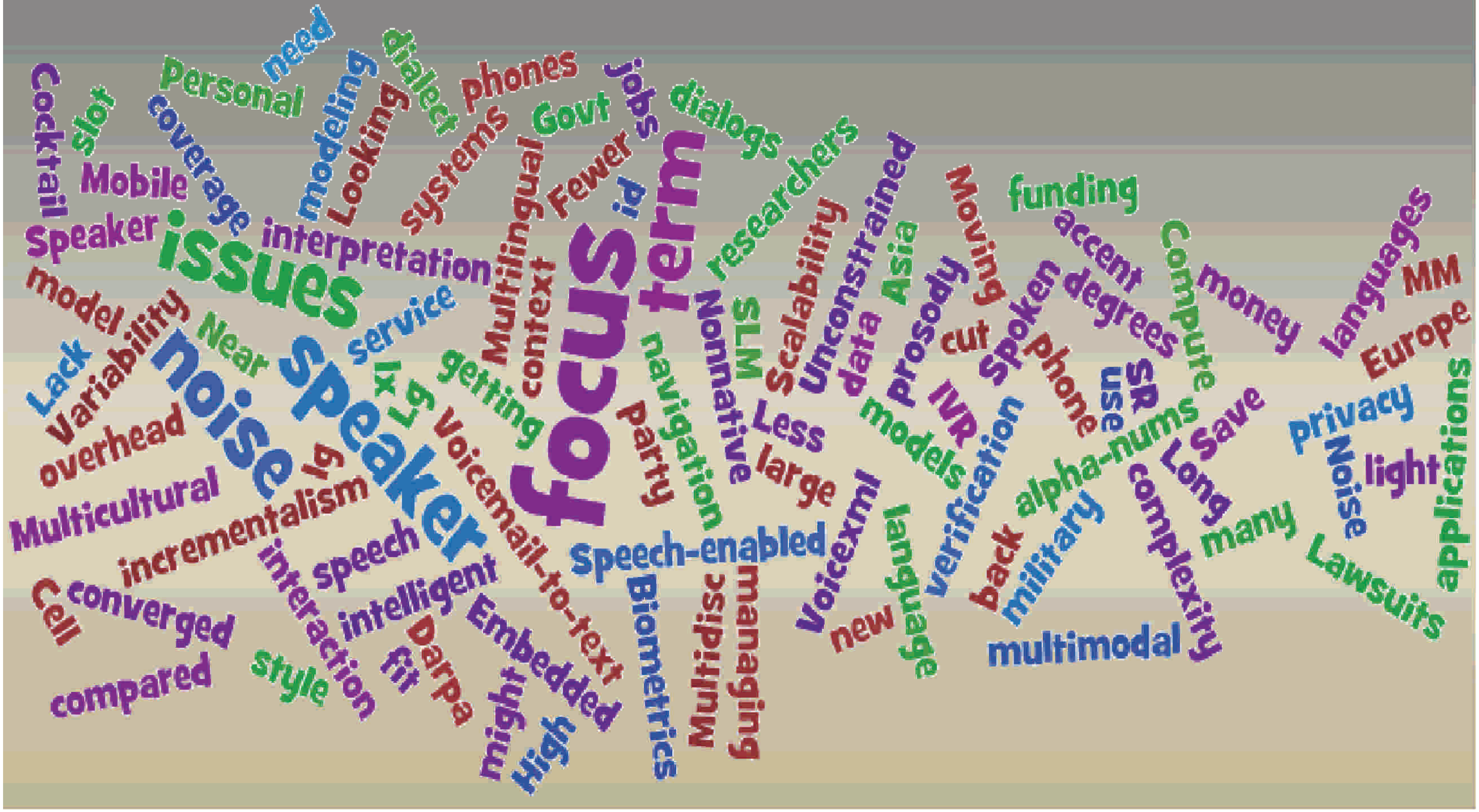
Number of companies mentioned in Speech Recognition Update, Bill Meisel, 2009

Companies 2009, GDP/ Company



Number of companies mentioned in Speech Recognition Update, Bill Meisel, 2009

Challenges


Cocktail
slot
Personal
need
coverage
modeling
Looking
dialect
phones
systems
Govt
jobs
dialogs
Mobile
Speaker
issues
interpretation
Multilingual
Fewer
id
term
focus
researchers
Scalability
Unconstrained
Moving
funding
accent
Compute
money
languages
MM
model
Variability
Near
service
context
navigation
Nonnative
SLM
data
Asia
prosody
cut
Spoken
degrees
Europe
Lack
noise
speaker
lx
Lg
getting
party
Less
large
models
IVR
phone
SR
use
Save
privacy
Noise
light
applications
overhead
Multicultural
lg
incrementalism
Voicemail-to-text
Speech-enabled
Voicexml
verification
alpha-nums
Long
complexity
many
Lawsuits
speech
intelligent
interaction
Cell
converged
compared
style
fit
Darpa
Embedded
might
High
Biometrics
Multidisc
managing
new
language
back
military
multimodal

Most Cited, Most Downloaded

understanding
communication
predominant
reconstruction
sequences
filling
form
large
presence
multimodal
review
scaling
melody
algorithm
adaptive
missing
audio
modeling
classifiers
signals
variability
emotion
estimation
highly
detection
dialogue
human
prosodic
feature
statistical
verification
like
data
systems
usability
system
tutorial
repeated
resources
using
automatic
language
long
processor
towards
noise
music
scene
parametric
algorithms
effective
room
vocal
synthesis
sources
speaker
reverberation
robust
intonation
methods
signal
features
role
research
error
paradigms
F0
decoding
margin
spectral
efficient
environments
spoken
applications
small
strategies
information
recovery
non-stationary
description
voice
time
lines
world
measures
activity
binaural
telephone
expressions
real
multi-level
detecting
envelope
designs
evaluation
term
bass

Predictions