OUCH: Outing the Unfortunate Characteristics of HMMs

Thanks to Keigh Rim for most of these slid

ntroducing Cortana

Cortana is an intelligent personal assistant **created by Microsoft** for Windows Mobile, Windows 10, Xbox One, Microsoft Band.

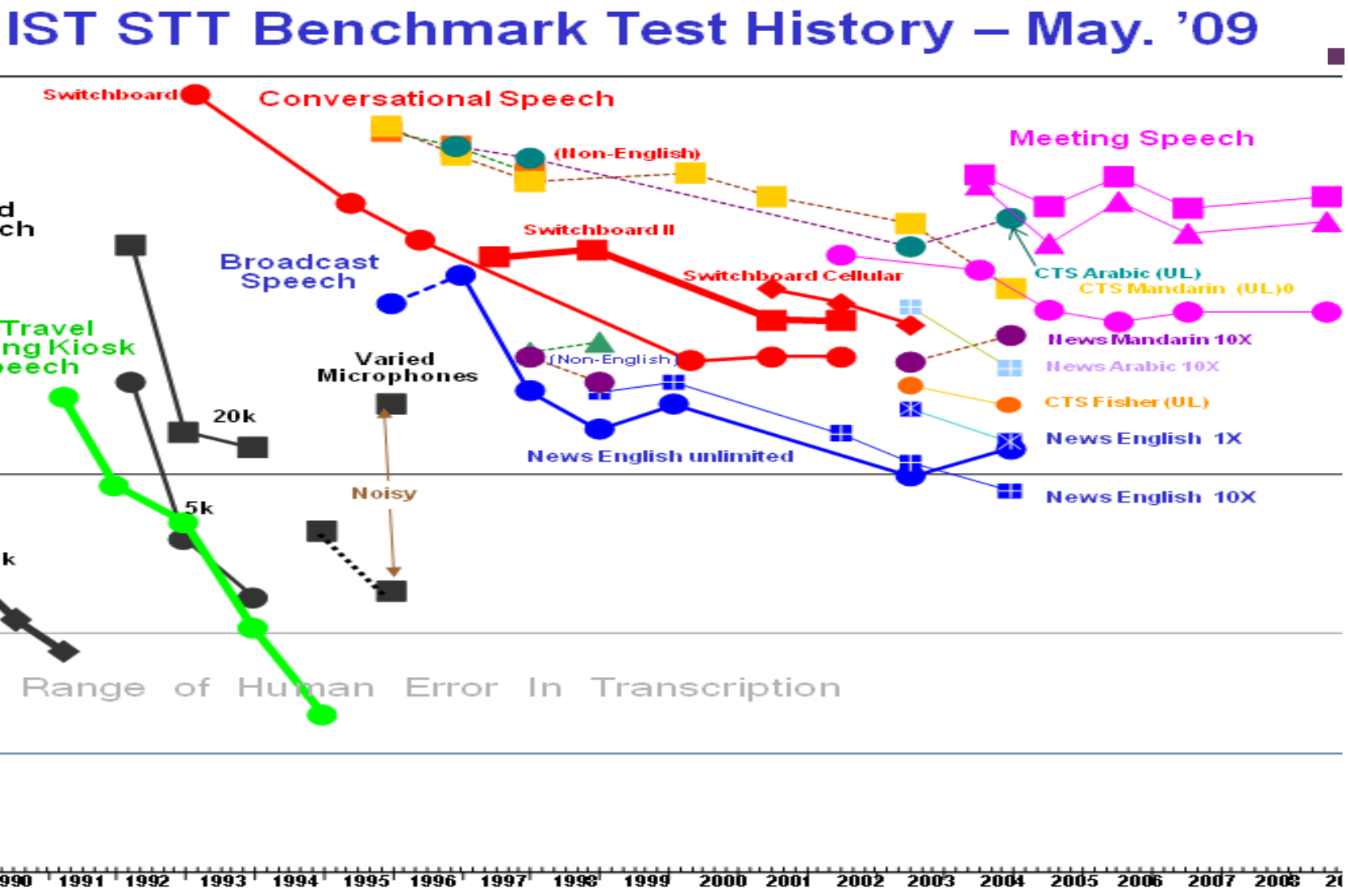
- and coming to iOS, Android.

https://en.wikipedia.org/wiki/Cortana_(software)

Cortana Demo at Sep 2015

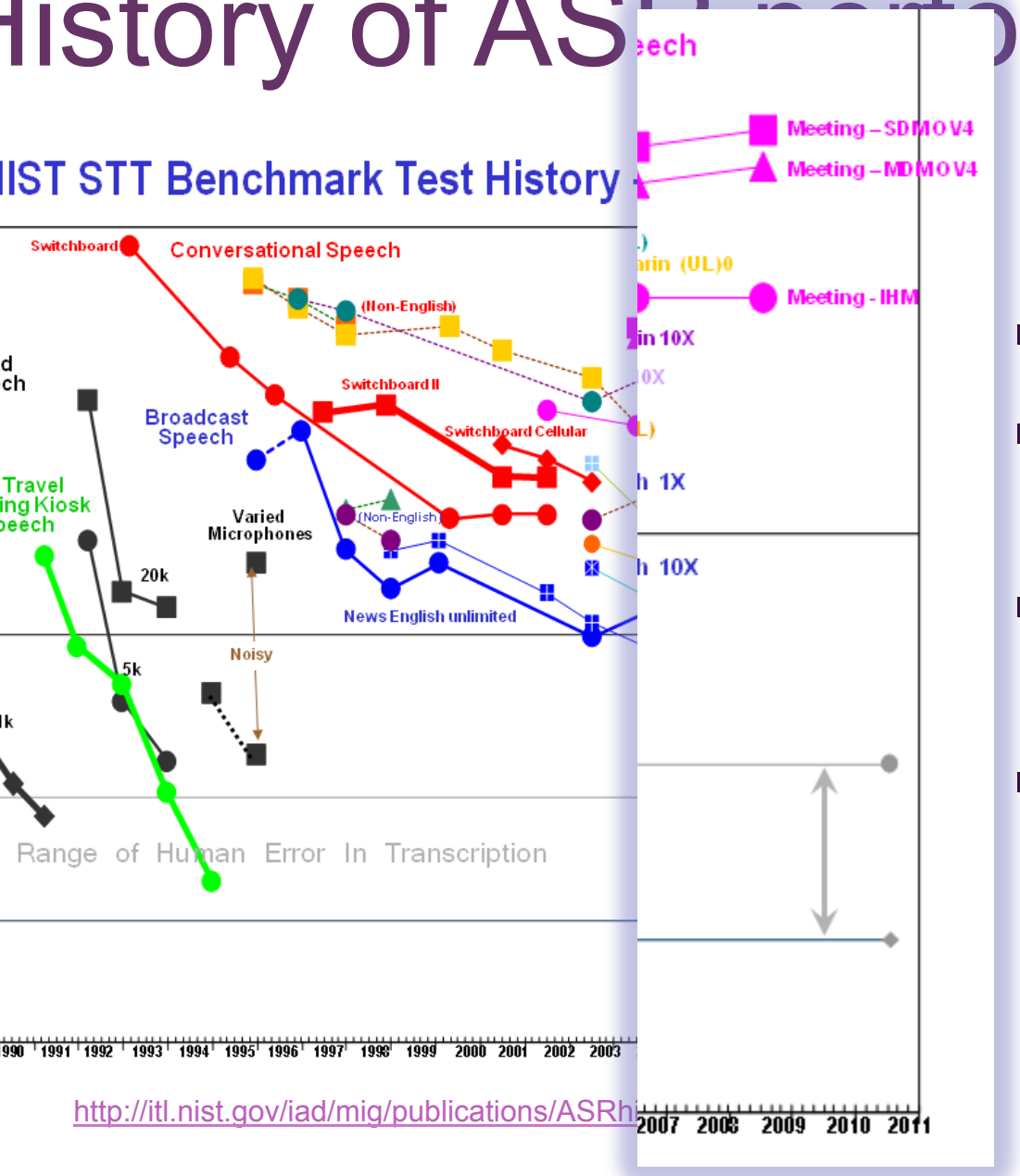
So, you tell me, what is the most at risk opportunities of Mr. Nadella?

history of Performance of Speech Recognizers



- http://itl.nist.gov/iad/publications/ASRhistory/index.html

History of ASR performance

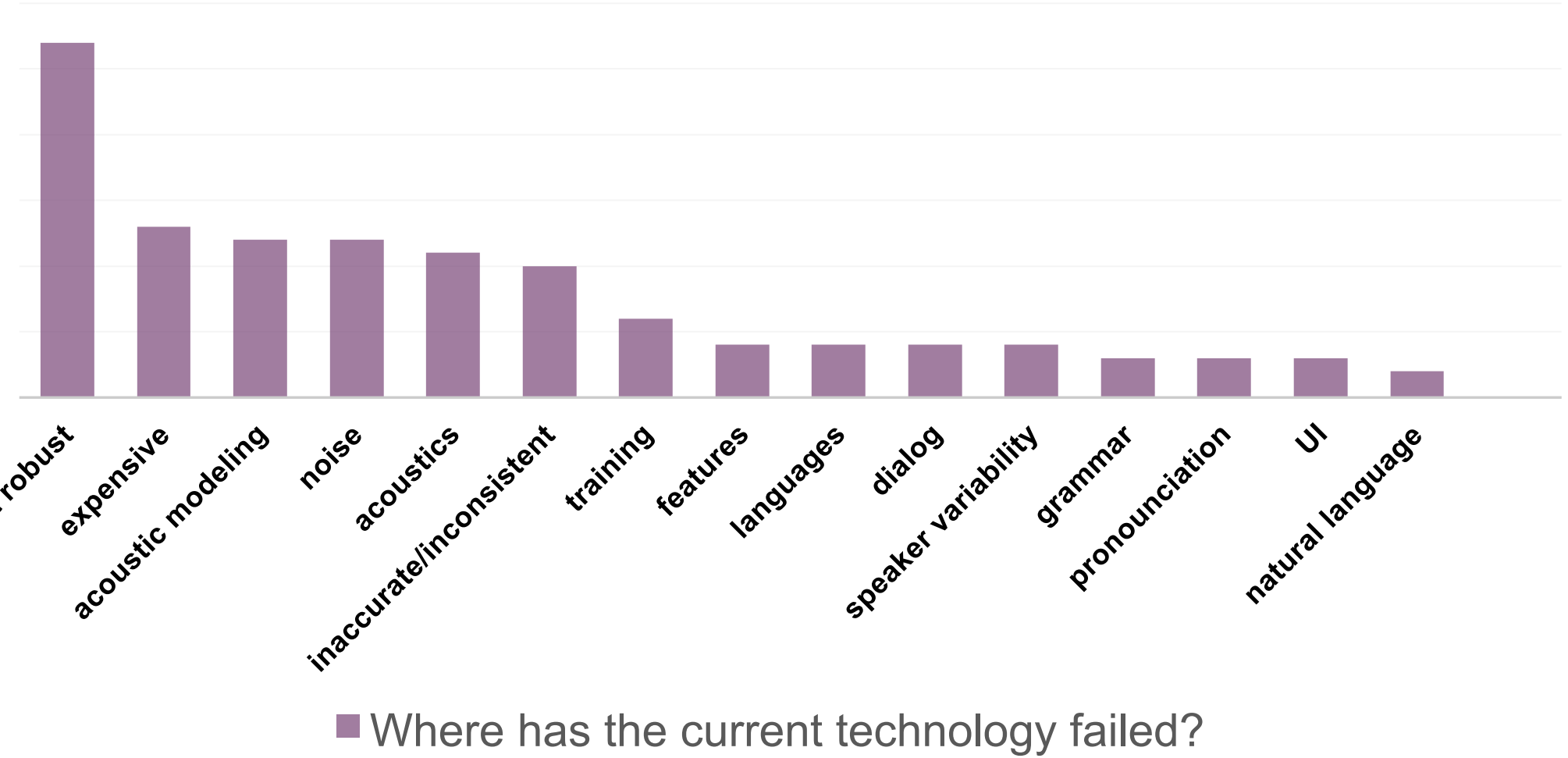


http://itl.nist.gov/iad/mig/publications/ASRh

- No Graph after 2009, why?
- Because NIST stopped evaluating, why?
- Because US gov stopped funding ASR projects, why?
- Because ASR has failed!

Where has ASR failed?

Community Survey (Morgan et al 2013)



About non-robustness

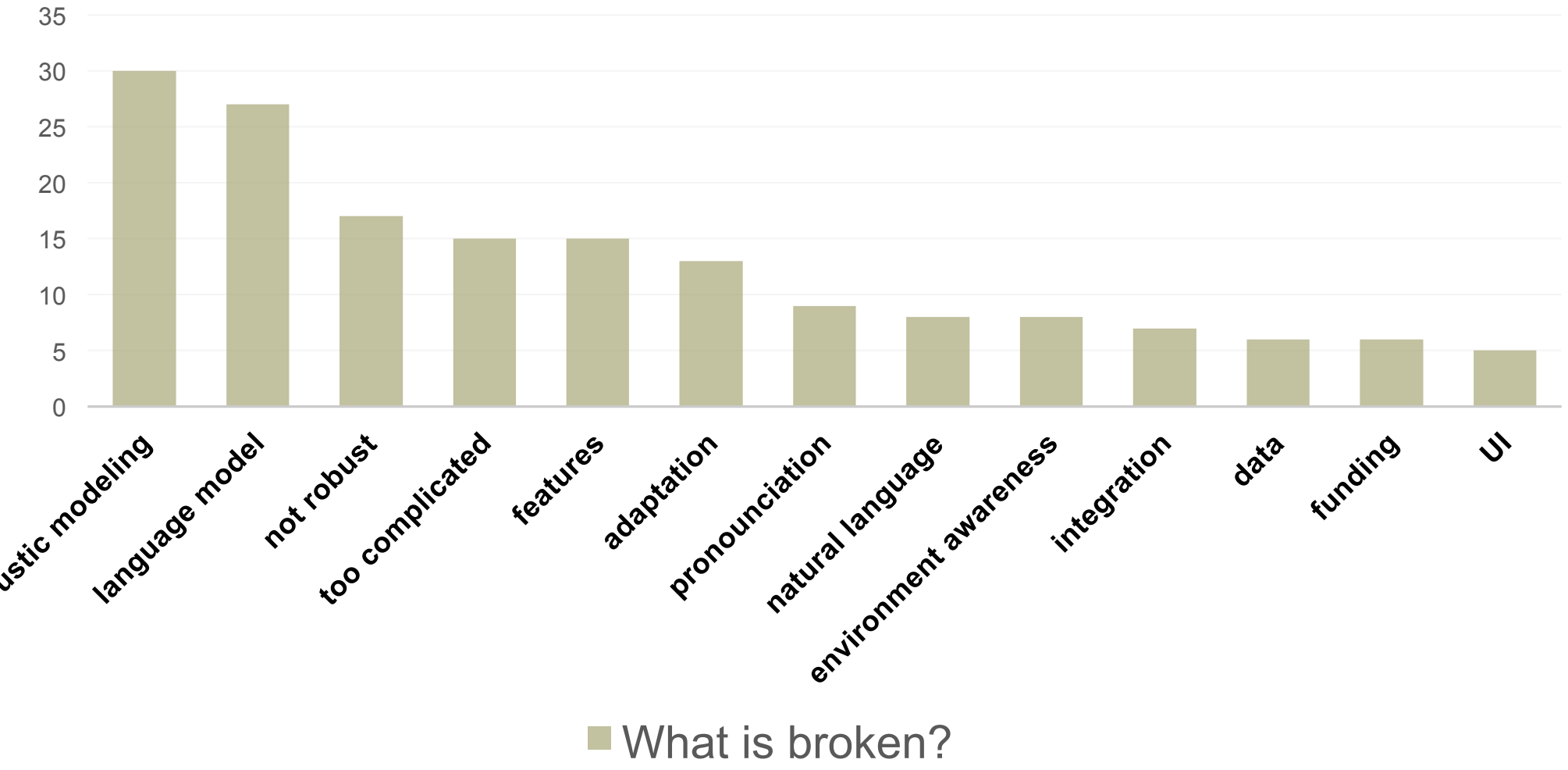
"… not robust to acoustic environments …"

"… fails for any conditions not seen in training …"

"… tuned too finely …"

"… ill equipped to handle data outside the training scenario …"

Community Survey (Morgan et al 2013)



About broken-ness

"… AMs don't communicate well with LMs …"

"… old models with new computational abilities …"

"… signal processing development was done in the 80's … no new models. HMMs and Cepstral analysis are still here …"

"… it doesn't … modeling human conversation. It assumes regimented turn-taking ..."

"… we tweak as many parameters as we can, but the caller is an unwilling participant ..."

What's wrong with Acoustic Models?

Building Acoustic Models for ASR

HMM

- Statistical model over transitions of hidden states and emissions of observations

GMM

- In ASR, probability density functions of emissions of observations

Discriminant training on previous models

- Max Mutual Information
- Max Likelihood Logit
- Min Phone Error
- …

Review on HMM

HMM is 5-tuple (Q, O, A, π, B), where

- Q is a set of (hidden) states
  $= \{q_1, q_2, \ldots, q_i, q_j\}$
- O is a set of all observations over t time
  $= \{o_1, o_2, \ldots, o_t\}$
- A is a set of transition probabilities
  $= \{\alpha_{1->1}, \ldots, \alpha_{1->j}, \ldots \alpha_{i->j}, \ldots, \alpha_{j->j}\}$ ($j \times j$ matrix)
- π is a set of special starting probabilities
  $= \{\alpha_{0->1}, \alpha_{0->2}, \ldots, \alpha_{0->j}\}$ ($1 \times j$ matrix)
- B is a set of emission probabilities
  $= \{\beta_{q1}(o_1), \beta_{q1}(o_2), \ldots, \beta_{qj}(o_t)\}$ ($t \times j$ matrix)

Review on HMM in ASR

HMM is 5-tuple (Q, O, A, π, B), where

- Q is a set of (hidden) states
  $= \{q_1, q_2, \ldots, q_i, q_j\}$ ← 10 ms sub-phone
- O is a set of all observtions over t time
  $= \{o_1, o_2, \ldots, o_t\}$ ← 39 dimensional MFCC vectors
- A is a set of transition probabilities
  $= \{\alpha_{1->1}, \ldots, \alpha_{1->j}, \ldots \alpha_{i->j}, \ldots, \alpha_{j->j}\}$
- π is a set of special starting probabilities
  $= \{\alpha_{0->1}, \alpha_{0->2}, \ldots, \alpha_{0->j}\}$ ← add ending probabilities
- B is a set of emission probabilities
  $= \{\beta_{q1}(o_1), \beta_{q1}(o_2), \ldots , \beta_{qj}(o_t)\}$

Independence” Assumptions in AM

ansition probabilities are independent from each other

- Hidden under Markov blanket.

mission probabilities are independent from each other

- Each observation is conditioned on only one state.

and B are conditionally independent

- Stationarity, at transition from $q_{i,t}$ to $q_{j,t+1}$, its probability $\alpha_{i->j}$ is independent no matter what observation, o_t is conditioned on $q_{i,t}$.

bservations are in multivariate normal distribution with diagonal covariance

- Remember that if Cov(x,y) == 0: $x \perp y$, thus, by ignoring non-diagonals , we treat all features as independent from each other.

OUCH

Outing Unfortunate Characteristics of HMMs

ndependence “Assumptions” in AM

We don’t know these conditional independences hold in real speech data, *we just assume*.

What if we have a dataset that satisfies, *for 100% sure*, the independences?

- **If HMM works differently (presumably better) with that data than real speech data, it proves that these independence assumptions on real speech are wrong.**
  (Classic form of proof by contradiction)

How can we get this particular data?

➔ We use artificial data stochastically simulated.

Sources for Data Simulation

After normally trained a acoustic model, we have

- Transition probabilities
- Emission probabilities
- Original real data♪
- Original transcript♪
- Pronunciation dictionary

ata Simulation and/or Resampling
McAllaster et al 1998, Wegmann et al 2010)

Forced alignment between real data frames and model, using transcript

Pick up pronunciations and silence from (1), generate tri-phone sequence

Recover underlying state sequence

Create pseudo utterance

- Full simulation: From a state, simulate an emission using output distribution
- Resampling: For a state, collect all associated real frames, then pick one randomly (sampling from samples, see Efron 1979)

Pseudo speech data

This reconstructed pseudo data has exactly the same length in frames with exactly the same state sequence and alignment.

Each frame is generated/picked-up from only one of mutually independent states, based on independent multivariate distributions.

That is, this data will completely satisfy the suspicious assumptions, except for that resampled data ignores the diagonal normal output distribution.

Frame level resampling

k of one “urn” for each state that holds
rvations

ll observations from the training data
re in that state into the urn

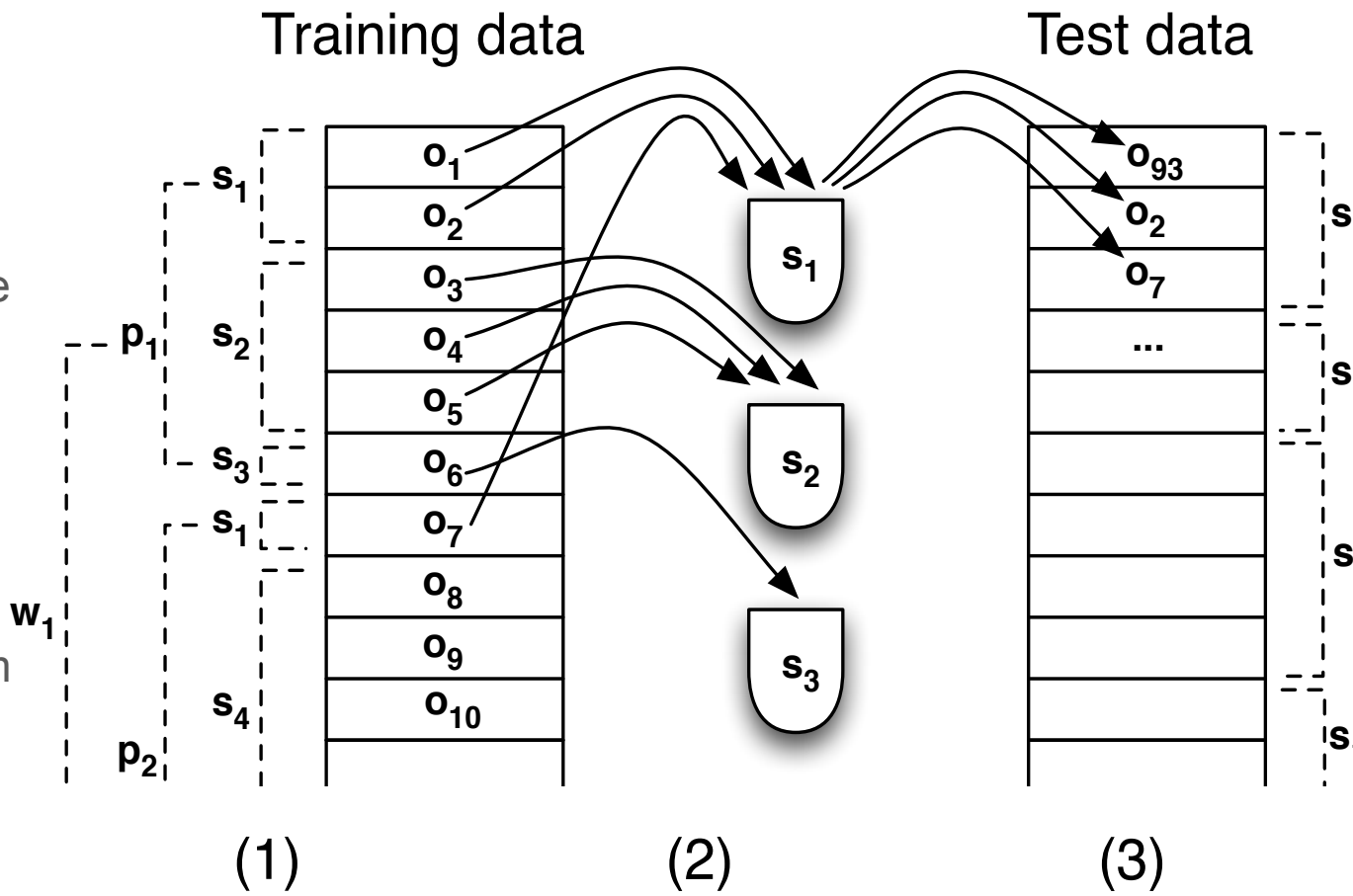
te new test utterances by Creating the
state sequence and selecting
rvations for each state randomly from
rn

observations are really independent
t shouldn’t matter

hat instance of a state they came from

hat order they are in within the state

hich speaker they are from

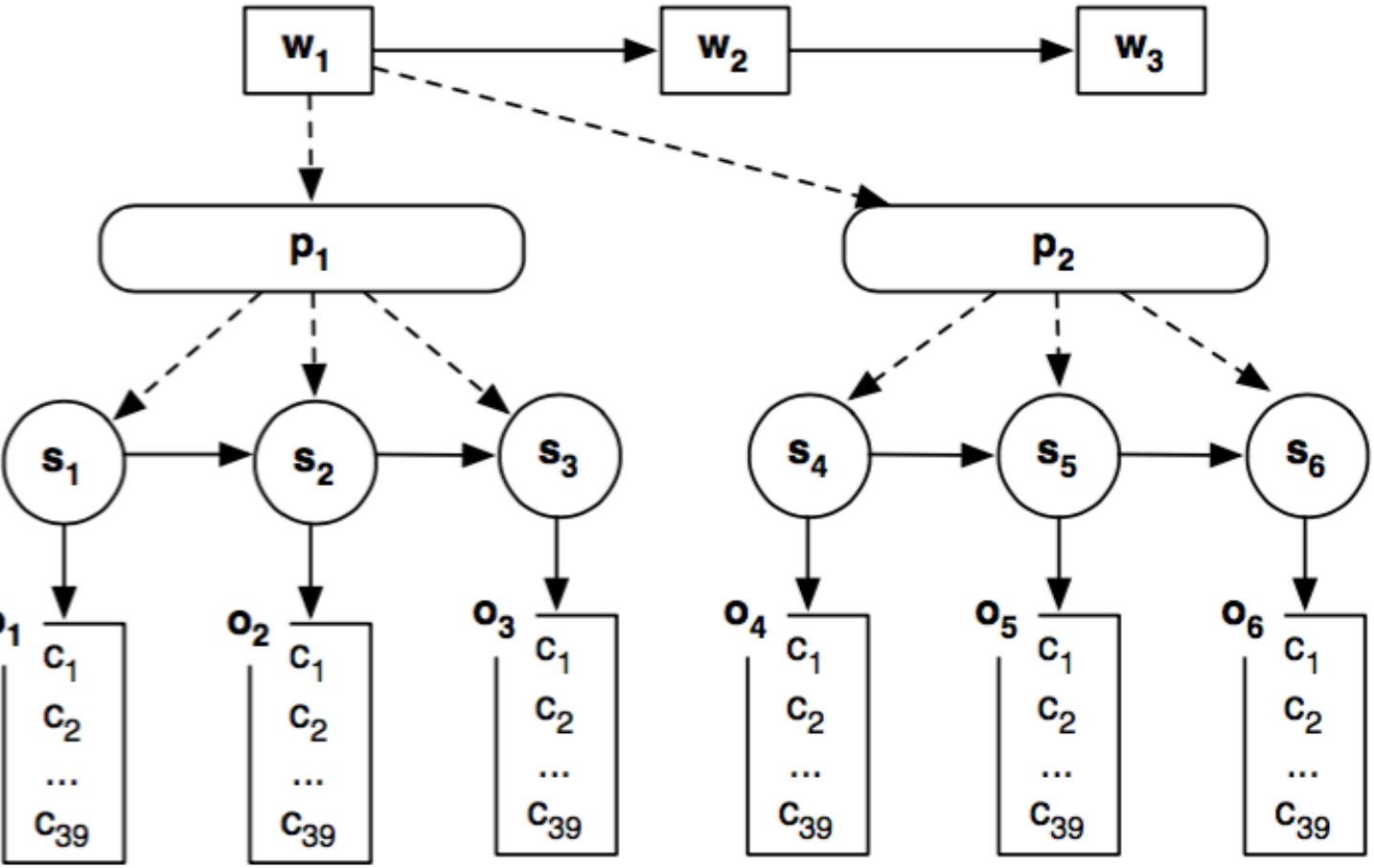


Results from Wegmann et al 2010

Dataset	WER
Original REAL speech data	.1
simulated	.0
resampled	.0
simulated using full cov matrix	.0

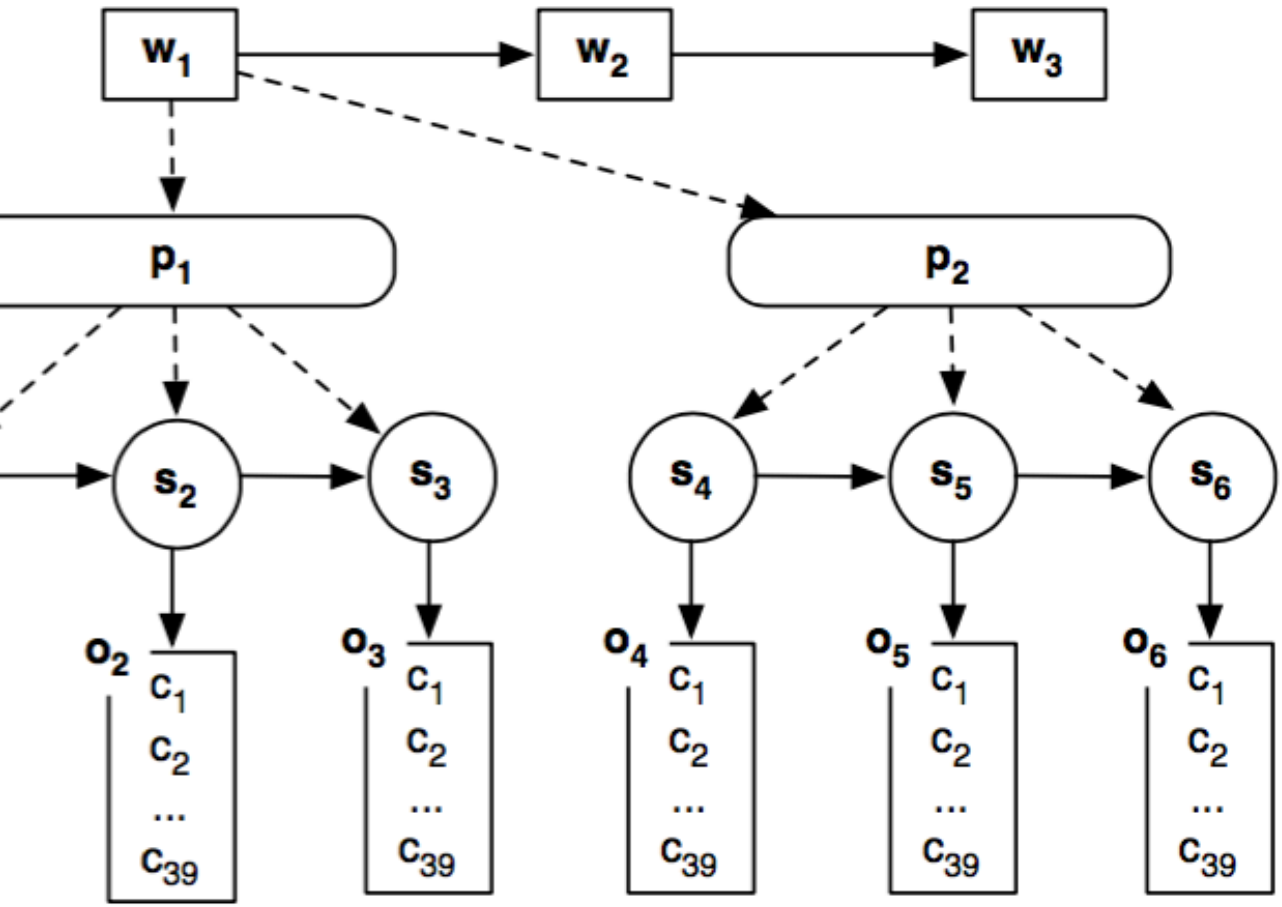
Conclusion: We have a serious problem in our model assumptions, and diagonal simplification is definitely not the problem.

Multi-level resampling (Gillick et al 2011)



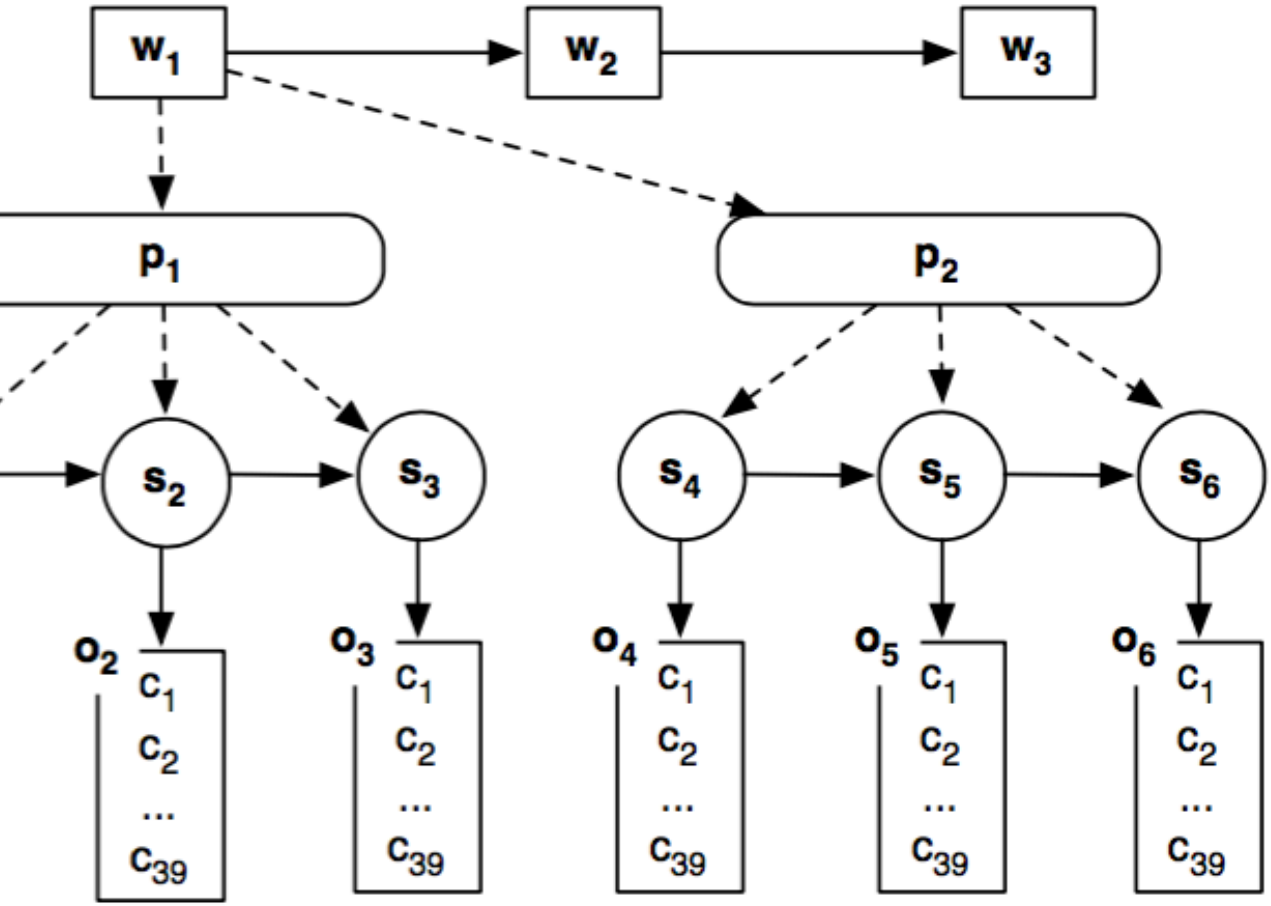
- Same idea, simila
  procedure but on
  - state level
  - phone level
  - word level

Multi-level resampling (Gillick et al 2011)



- These higher level resamp violate model assumptions
- Note that *frames* are leave replacing them with artificia pieces (simulation) or segm from the original (resamplin doesn't percolate down♪

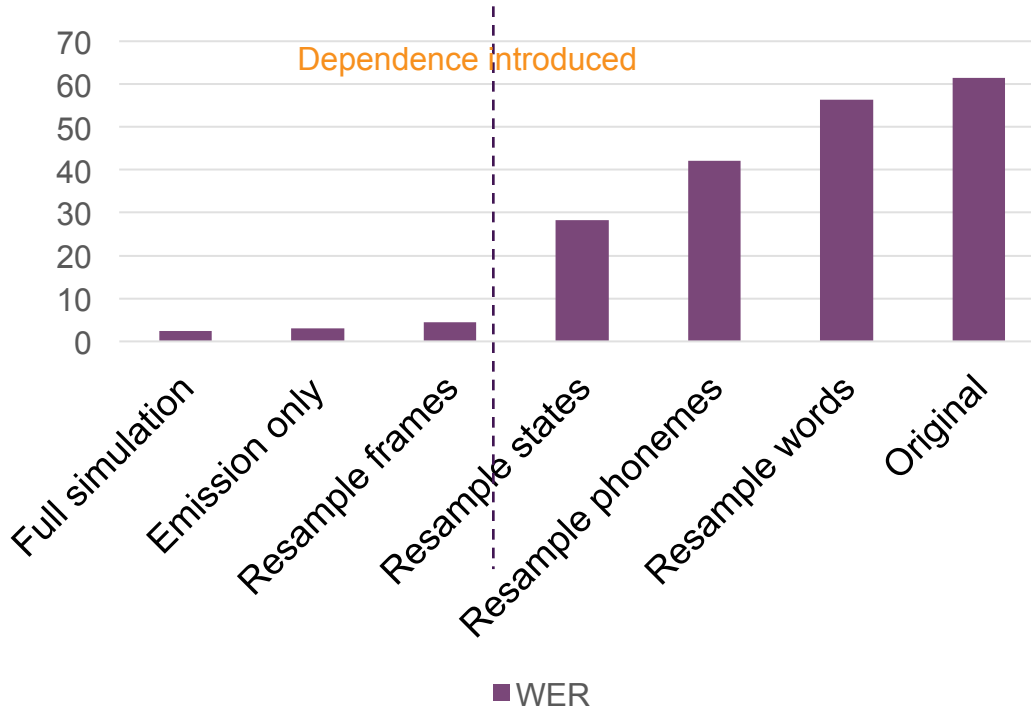
Multi-level resampling (Gillick et al 2011)



- At higher levels, replacing by resampling will guarant inter-node independence.
- But we cannot control wha inside the node, that is, dependency from the origi data is brought into pseud data.

Results on SWBD from Gillick et al 2011

Dataset	WER
Original	61.5
Full simulation	2.4
Emission only simulation	3.0
Resample frames	4.5
Resample states	28.2
Resample phonemes	42.1
Resample words	56.4



Conclusion: the largest increase in WER is observed when we move from frame resampling to state resampling ➔ this is where we first need to look at!

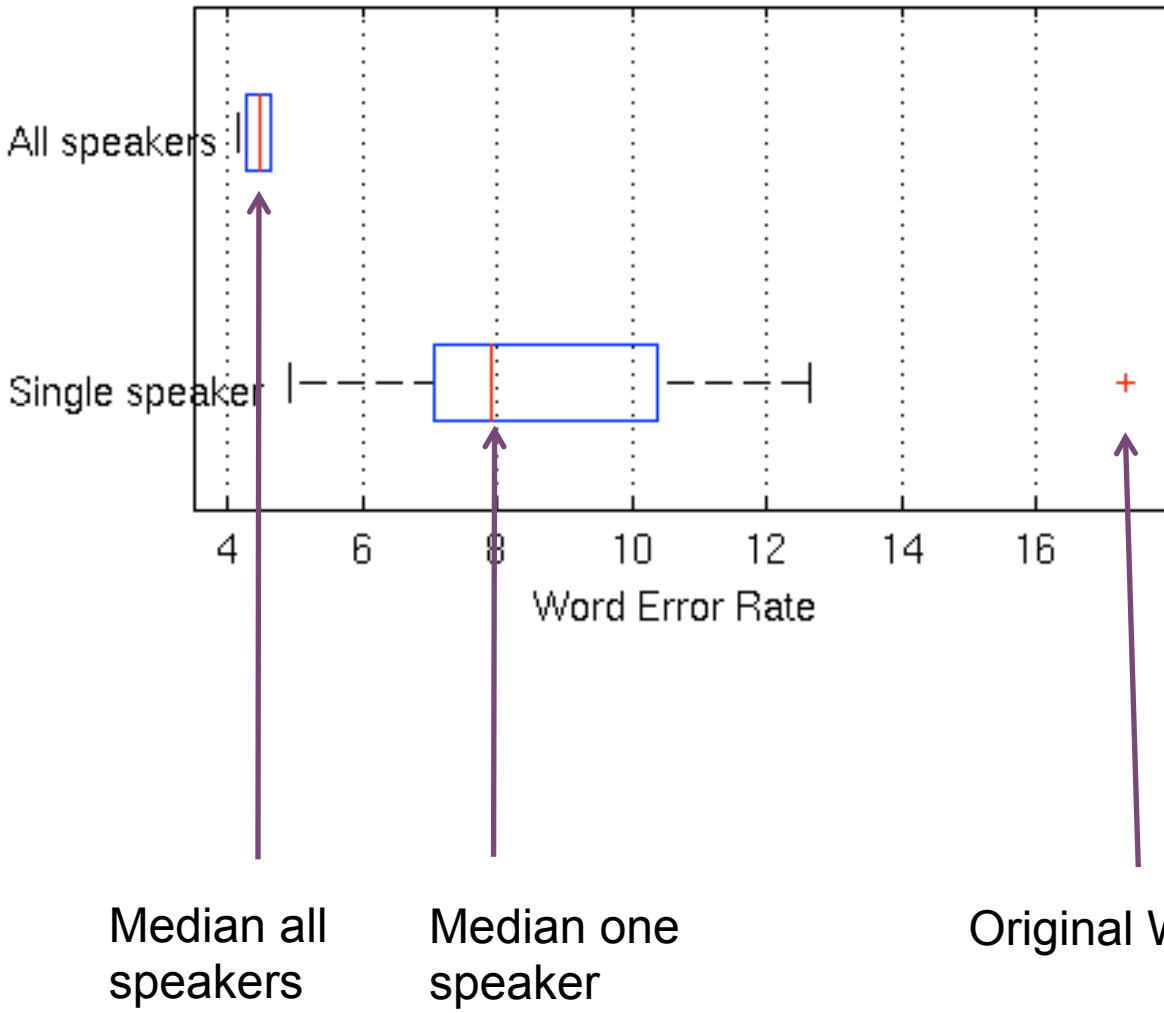
mpact of speaker dependence

All speakers: Data is sampled independent of the speakers

Single speaker: Only select samples from a single speaker

If these are close, the problem is more to do with the temporal dependence (that frames need to be sequential in time)

If the single speaker is close to the original, then the issue is capturing speaker characteristics



ntroducing Different Acoustic Conditions Parthasarathi et al 2013)

Remember from the community survey;

"… not robust to acoustic environments …"

"… fails for any conditions not seen in training …"

How bad is the effect of difficult acoustic input?

The ICSI meeting corpus: Two recordings on the same speech, close mic (Near Field) & far mic (Far Field)

Parallel data sets

Three experiments

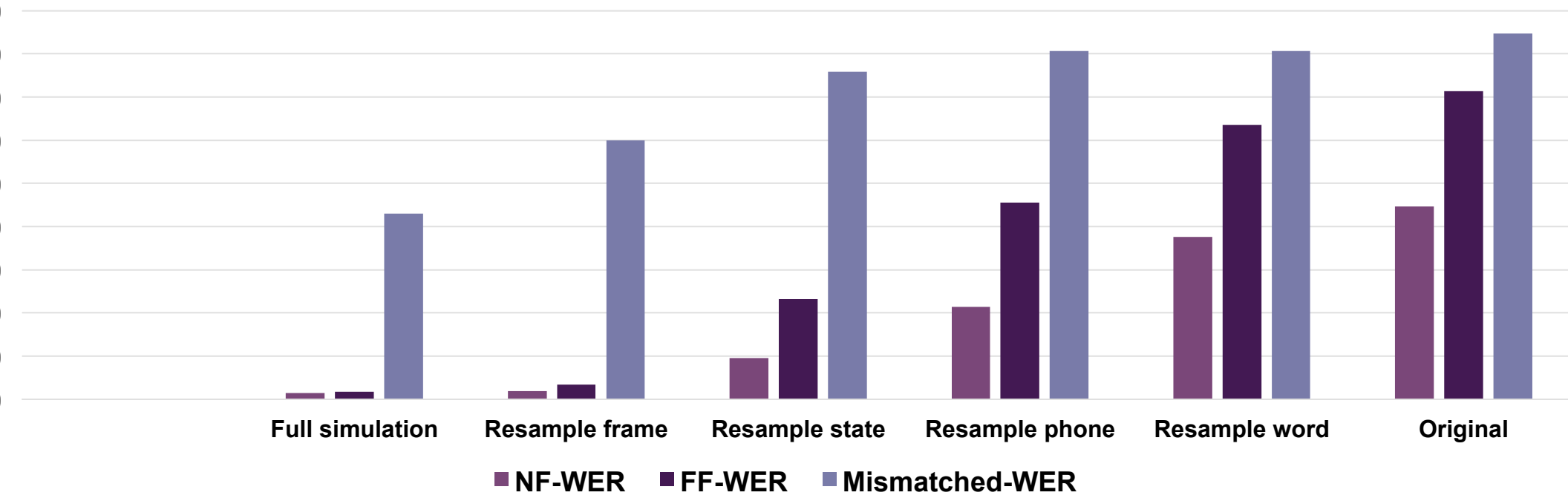
	Train on	Test against
NF set	NF recording	NF recording
FF set	FF recording	FF recording
Mismatched set	NF recording	FF recording

Conducted the same sets of simulation/resampling experiments over different datasets to find out the more dominant problem;
Bad acoustic condition vs. False model assumptions.

Results from Parthasarathi et al 2013

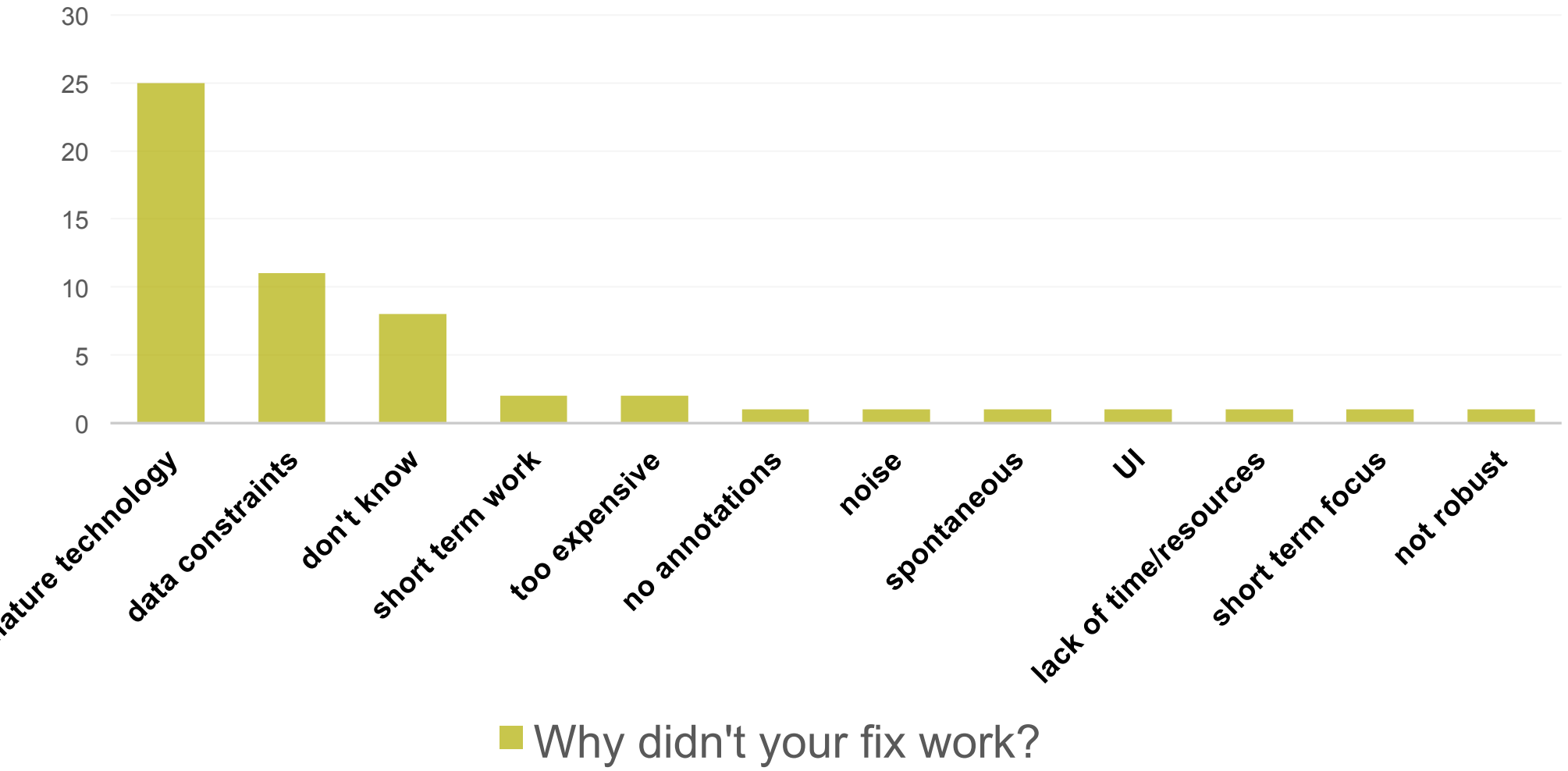
Experiment	NF-WER	FF-WER	Mismatched-WER
Original	44.7	71.4	84.7
Full simulation	1.4♪	1.8	43.0
Resample frame	1.9	3.4	59.9
Resample state	9.6	23.2	75.8
Resample phon	21.4	45.5	80.6
Resample word	37.6	63.5	80.6

Results from Parthasarathi et al 2013

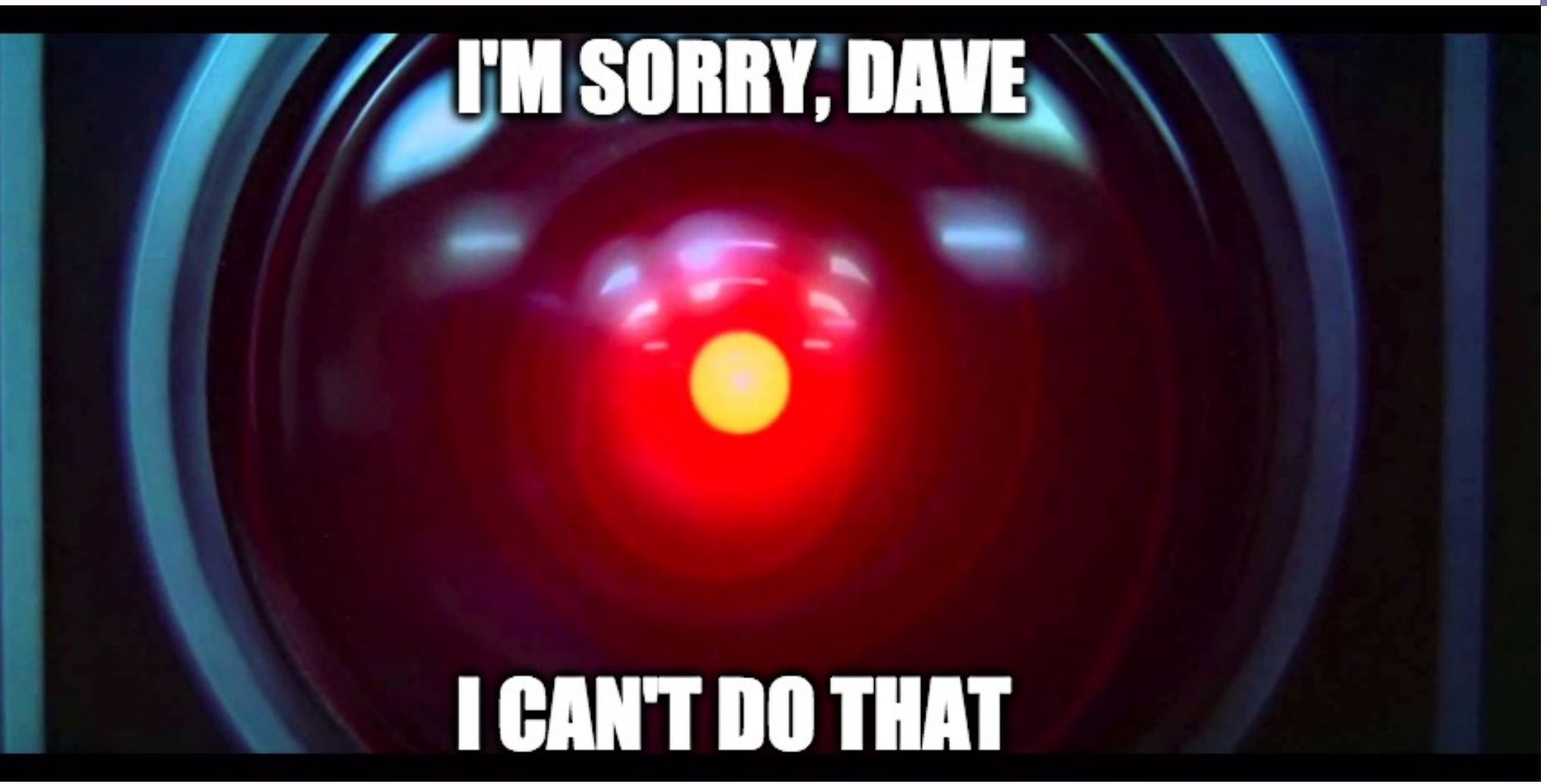


Conclusion: when the conditions are matched the model errors dominate. Mismatched conditions contribute as much to the total errors as does the model.

Community Survey (Morgan et al 2013)



Current ASR is indeed seriously broken



low can we fix this? - Some suggestions from Morgan et al 2013

Diagnose, diagnose, diagnose.

- We need diagnostic analysis.
- Not simply seeing WER/perplexity going down, we need some kind of methodology of specificity and efficiency.
- Encouraging a diagnostic spirit could have very broad effects.

low can we fix this? - More suggestions from Morgan et al 2013

We need new models, obviously!

- Most successful recognition system: human brain. Model human speech perception.
- Model low dimensional parameters of environmental variability: how to compensate accents, noise, spontaneous conversations?
- Better stochastic models from theorists.

It's time for a transition from a difficult, immature technology to a robust, mature technical system!

References

rgan, N., Cohen, J., Parthasarathi, S. H. K., Chang, S., & Wegmann, S. (2013). *Final Report : OUCH Project (Outing Unfortuna*
racteristics of HMMs).

allaster, D., Gillick, L., Scattone, F., Newman, M., & Street, N. (1998). Studies With Fabricated Switchboard Data : Exploring
rces of Model-Data Mismatch. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*.

gmann, S., & Gillick, L. (2010). Why has (reasonably accurate) Automatic Speech Recognition been so hard to achieve?

ick, D., Gillick, L., & Wegmann, S. (2011). Don't multiply lightly: Quantifying problems with the acoustic model assumptions
ech recognition. *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 71–76.

on, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, *Vol. 7, No. 1*, 1–26.

thasarathi, S., & Chang, S. (2013). The blame game in meeting room ASR: An analysis of feature versus model errors in nois
mismatched conditions. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 6758 –
2.