Experimental tests of Features and Partial Specification

Greg Kochanski
John Coleman
Christina Orphanidou
(Phonetics Laboratory, Oxford)

Christopher Alvey
A. McIntyre
Stephen Golding
(Nuffield Dept. of Surgery, Oxford)

Funded by the UK ESRC via grant RES-000-23-1094.
Talk presented by G. Kochanski, 17 December 2010,
at the Laboratoire Parole et Langage,
Université de Provence, Aix-en-Provence, France.
This can be downloaded from http://kochanski.org/gpk/papers/2010/Aix_underspecification.pdf,
http://www.phon.ox.ac.uk/releases/pdfs/Aix_underspecification.pdf, or the University of Oxford Library Service.

Philosophy

Suppose you're the first to understand X.

How can you prove that you really understand it?

- Tell a story that is consistent with my (partial) understanding of X, but goes beyond it. Add detail and implications.
- Demonstrate that you can do X.

Philosophy – why is this question important?

Pathological Science:

- Cold Fusion (c. 1989)
- Polywater (c. 1975)
- Allison Effect (c. 1940)
- N-Rays (c. 1910)

- Exciting ideas; plausible or not obviously impossible.
- Collect many supporters.
- Supporters collect much supporting evidence.
- Supporters don't look for contradictory evidence.
- Eventually fails too many experimental tests.
- Down the drain – nothing saved.

Philosophy

Suppose you're the first to understand X.

How can you prove that you really understand it?

- Tell a story that is consistent with my (partial) understanding of X, but goes beyond it. Add detail and implications.
- Demonstrate that you can do X.

Stories that start with the truth, then step outside it

- Malaria is common near swamps; you can smell swamps; therefore the disease is caused by the smell.

- "What we have found over the years in the marketplace is that derivatives have been an extraordinarily useful vehicle to transfer risk from those who shouldn't be taking it to those who are willing to and are capable of doing so." - Alan Greenspan (US Federal Reserve Chairman) 2003

"Not only have individual financial institutions become less vulnerable to shocks from underlying risk factors, but also the financial system as a whole has become more resilient." - Alan Greenspan in 2004

Stories that start with the truth, then step outside it

- “To sum up: the more productive capital grows, the more it extends the division of labor and the application of machinery; the more the division of labor and the application of machinery extend, the more does competition extend among the workers, the more do their wages shrink together.” (K. Marx, *Wage Labour and Capital*, 1847)

Philosophy

Suppose you're the first to understand X.

How can you prove that you really understand?

- Tell a story that is consistent with my (partial) understanding of X, but goes beyond it. Add detail and implications.
- Demonstrate that you can do X.
  - Predict something that hasn't yet been seen.
  - Control something that has been uncontrollable.
  - Make something that has never been made before.

Strategy

- Linguistic theorists typically provide an *account* of something.
  - Predictions are scarce.
  - An *account* explains:
    - What can happen.
    - How something might happen.
    - (generally) not what must happen.

  - Accounts are not complete descriptions:
    - They focus on one aspect.
    - They focus on a small set of examples.

- So, accounts need to be extended before they can make predictions that can be tested.
  - Add auxiliary hypotheses (assumptions).

Strategy

- Linguistic theorists typically provide an *account* of something.
  - Accounts are typically not complete descriptions.

- So, accounts need to be extended before they can be tested
  - This is related to the Quine-Duhem idea that one can never test a theory in isolation

  - Add auxiliary hypotheses (assumptions):
    - What do the subjects notice during the experiment?
    - What should we measure?
    - How do we interpret the measurements?

Strategy

- Linguistic theory provides an *account*
- Accounts need auxiliary hypotheses to allow them to be tested

- If an extended account makes the wrong prediction,
  - Is the account wrong?
  - Are the added assumptions wrong?
  - L. Laudan (1990) provides philosophical support for the Brute Force solution (below).

- Brute Force solution:
  - Test the account with many different assumptions.
    (Make many extended accounts from the same account.)
  - If it always succeeds: strong support for the account.
  - If it always fails: the account was probably not valuable.
  - In between: look for patterns of success and failure.

What will we test?

- Are distinctive features good at representing articulatory positions?

- How important is context?

- Are all distinctive features specified, or are some filled in from the neighbouring segments?

What are we testing?

- What features do we use?

  - Based on *Sound Patterns of English*
    - G. N. Clements 1991 (optional mod)
    - Schwa largely unspecified (optional mod)
    - Silence unspecified (optional mod)

  - Three complexities:
    - 5 features: HIGH, LOW, CORONAL, ANTERIOR, BACK
    - 8 features: ***drop*** ANTERIOR; ***add*** VOCALIC, SONORANT, TENSE, ATR
    - 13 features: ***add*** VOICED, ROUNDED, CONTINUANT, NASAL, ANTERIOR

- How do we get the feature representation?
  - Take the text, look it up in a dictionary to get phones.
  - Look up the phones to get the corresponding feature representation.

Methods

- •Our data are mid-sagittal magnetic resonance images.
- •We process the images to find the tongue's edges.
- •We wrote and used noise reduction software to subtract much of the noise from the MRI machine.
- •We segmented the speech to find the centres and/or edges of the phones.
  - (Because of the MRI noise, certain boundaries could not be precisely located.)

MRI Videos

Image Processing Strategy

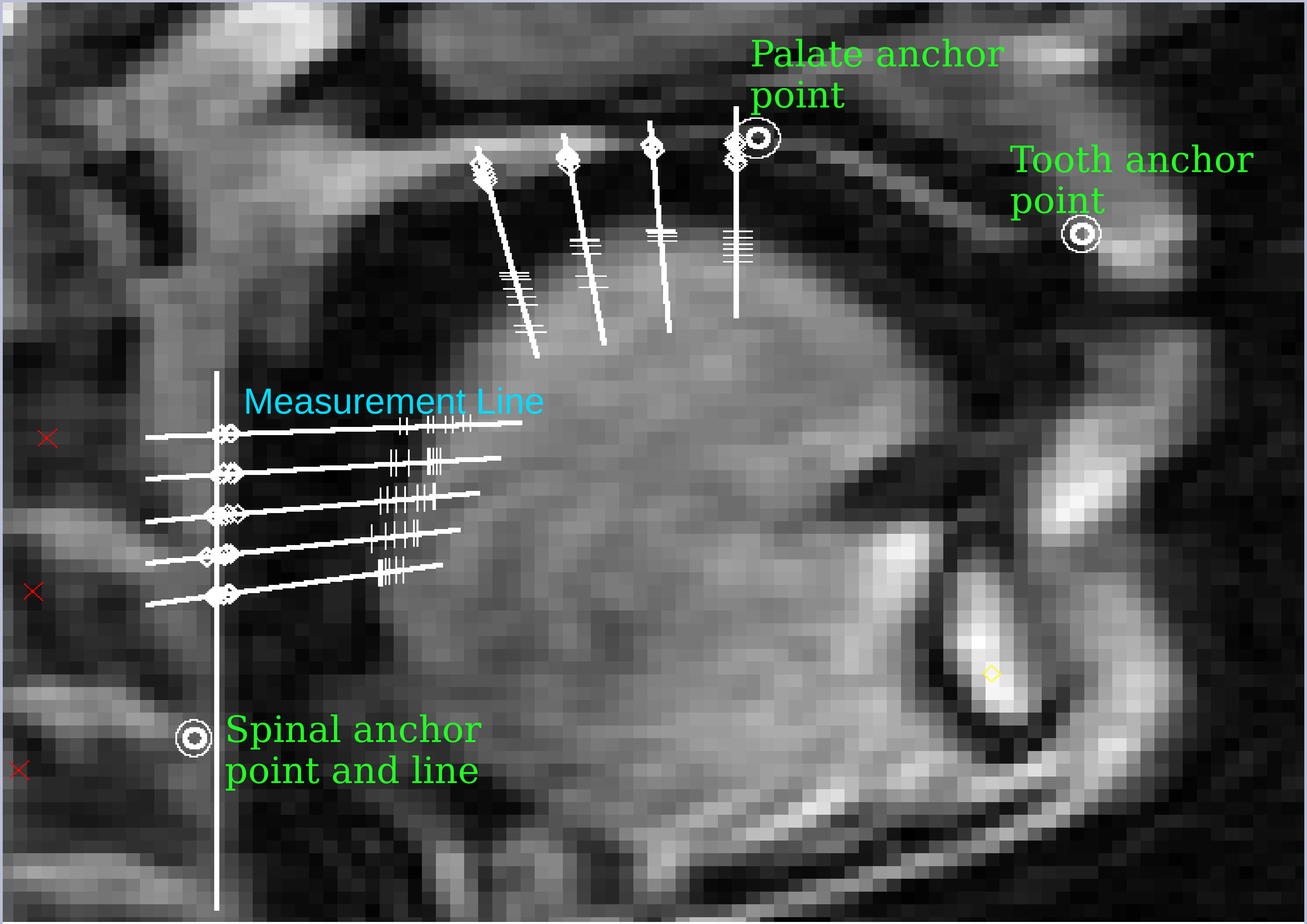
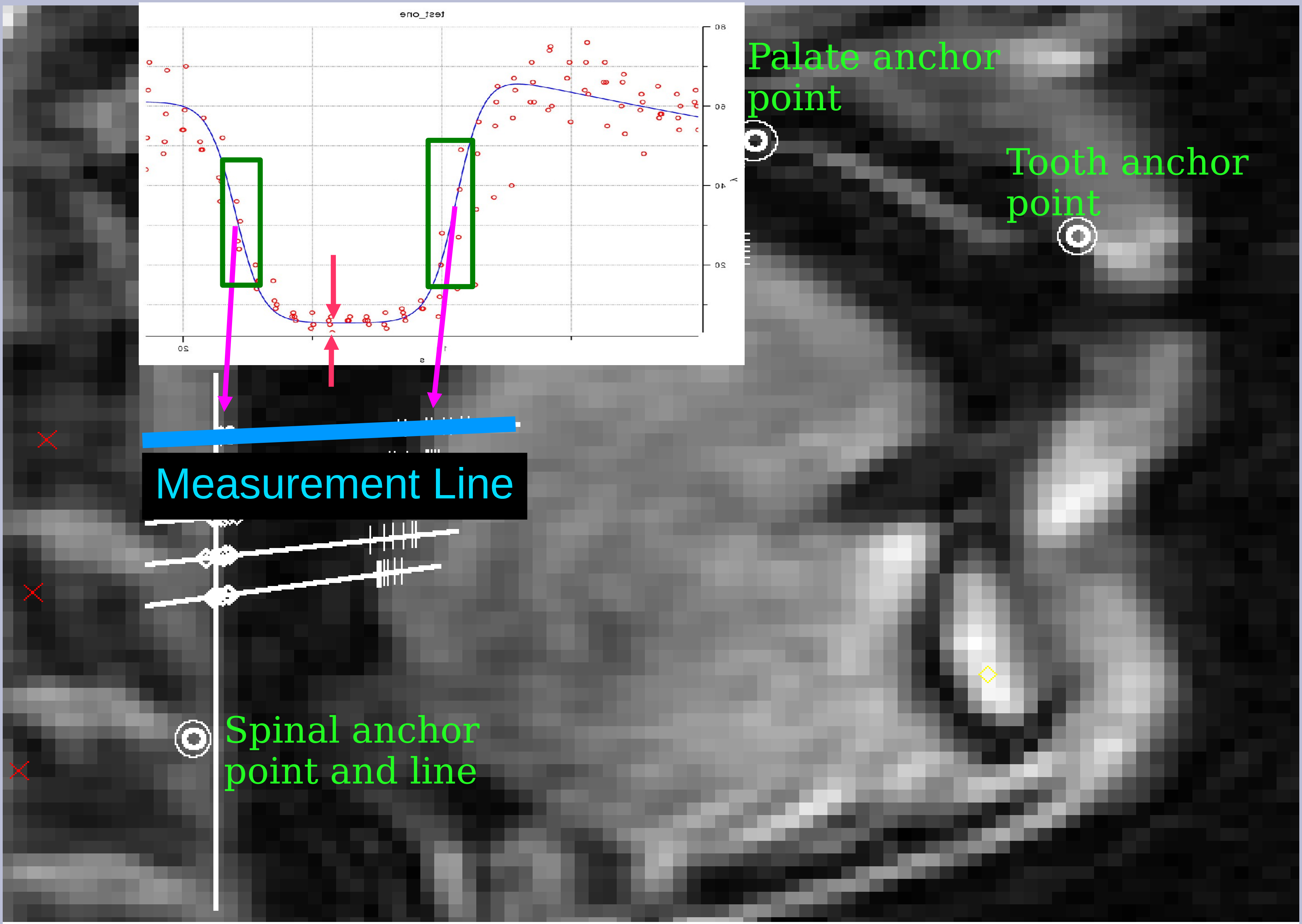


Image Processing Strategy



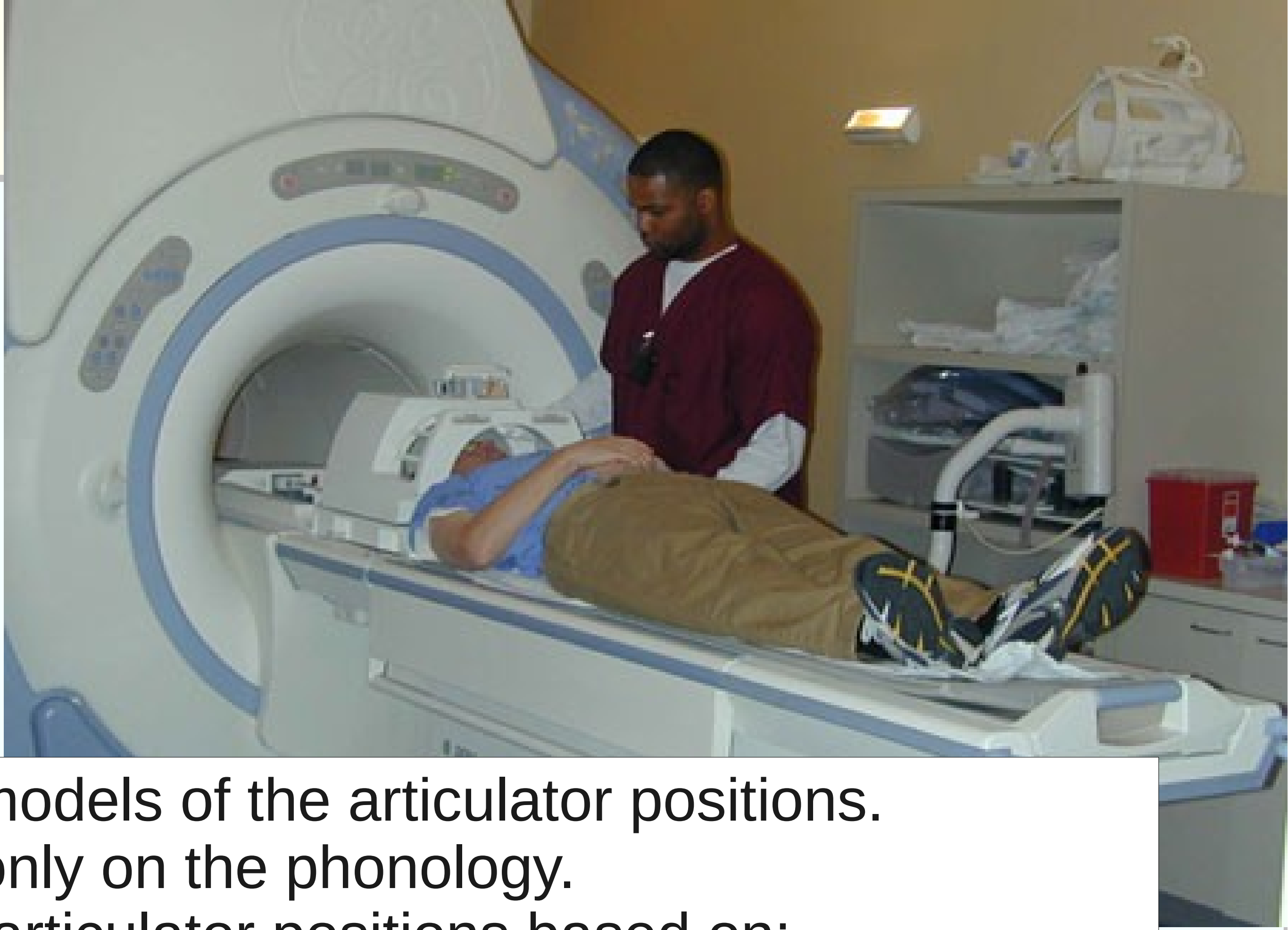
Palate anchor point

Tooth anchor point

Measurement Line

Spinal anchor point and line

Methods



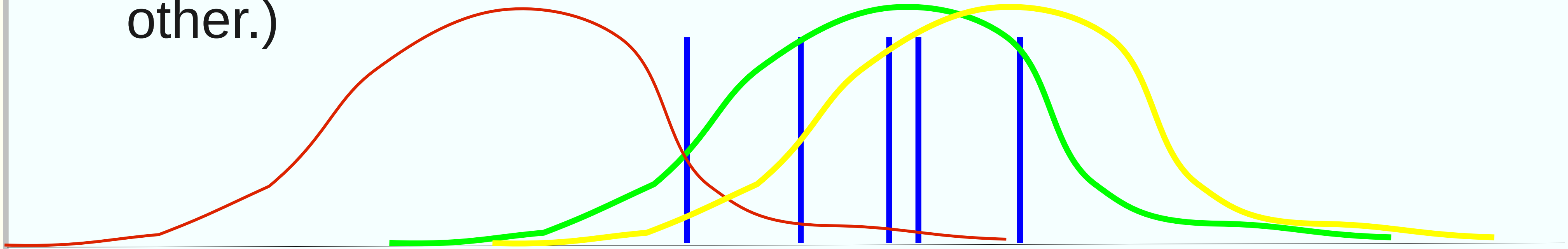
- We built mathematical models of the articulator positions.
  - The models depend only on the phonology.
  - The model predicted articulator positions based on:
    - A linear combination of the distinctive features, or
    - A postion for each phoneme, or
    - The Harshman, Ladgefoged and Goldstein model.
  - These were full Bayesian models, so the model also predicted the variance of each measurement.
  - Bayesian Model Comparison:
    - Acts as a more general F-test.

Bayesian model comparisons

- A Bayesian model tells you the probability that the model would produce an observation.
  - You tell it what you observed.
  - It tells you how likely the observation was.

- A good model:
  - Assigns high probabilities to things that are actually observed.
  - Low probabilities to things that haven't been seen.

- This can be formalized as Bayesian Model Comparison.
  - It is analogous to the classical F-test and ANOVA.
  - It can compare models that aren't related.
    - (The F-test can only compare two models if one includes the other.)

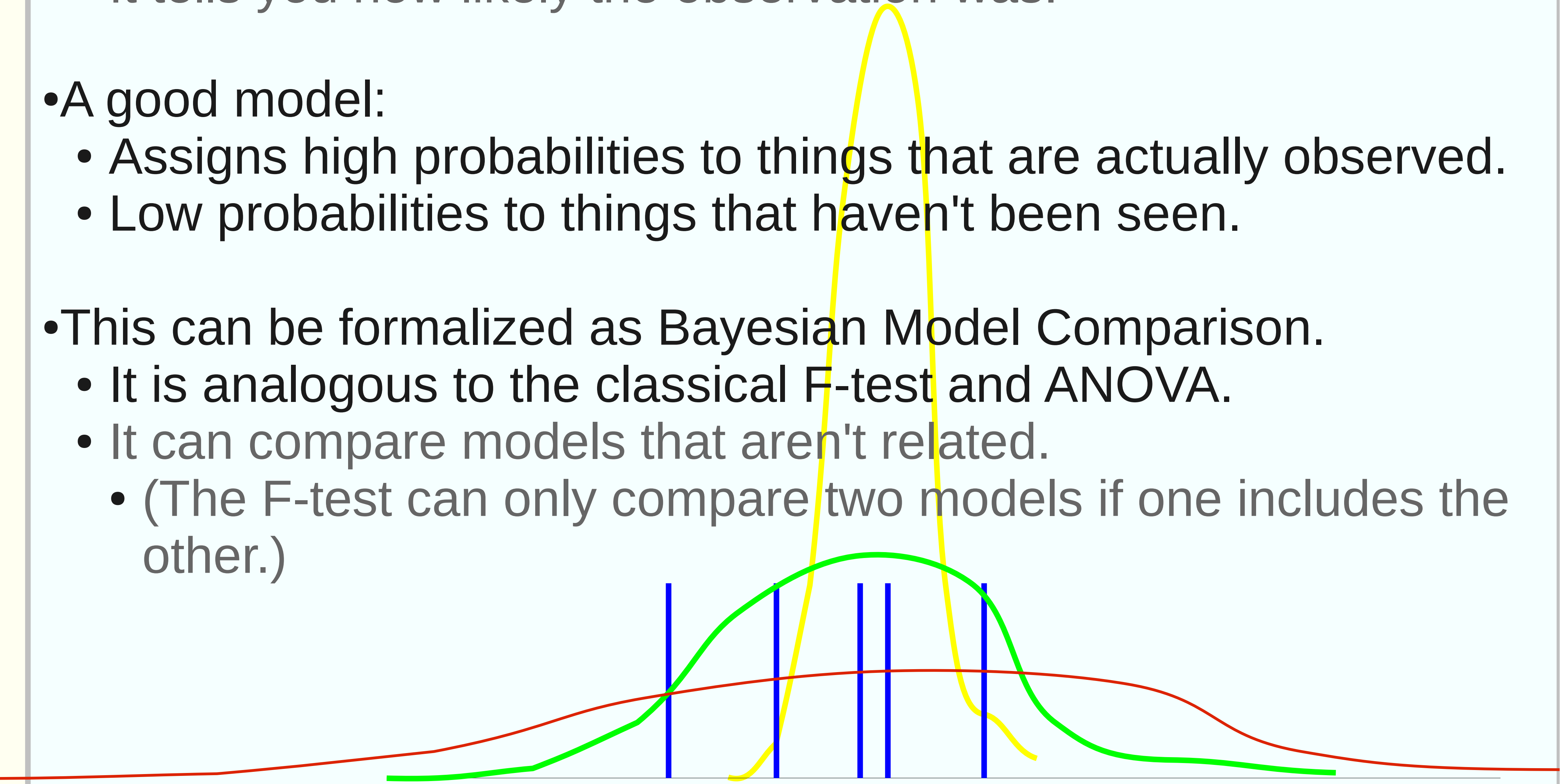


Models of Articulation

- A Bayesian model tells you the probability that the model would produce an observation.
  - You tell it what you observed.
  - It tells you how likely the observation was.

- A good model:
  - Assigns high probabilities to things that are actually observed.
  - Low probabilities to things that haven't been seen.

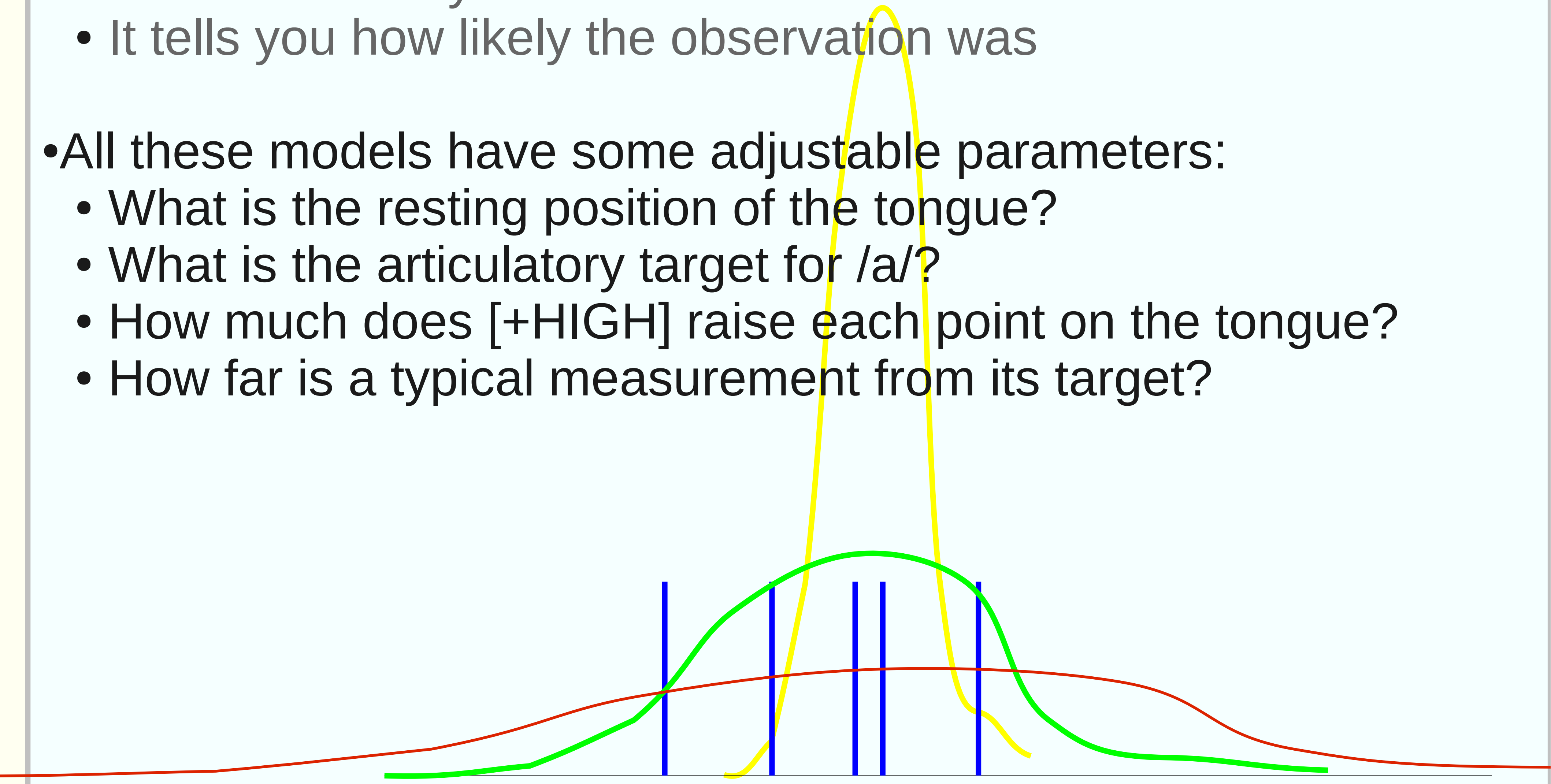
- This can be formalized as Bayesian Model Comparison.
  - It is analogous to the classical F-test and ANOVA.
  - It can compare models that aren't related.
    - (The F-test can only compare two models if one includes the other.)



Models of Articulation

- A Bayesian model tells you the probability that the model would produce an observation.
  - You tell it what you observed.
  - It tells you how likely the observation was

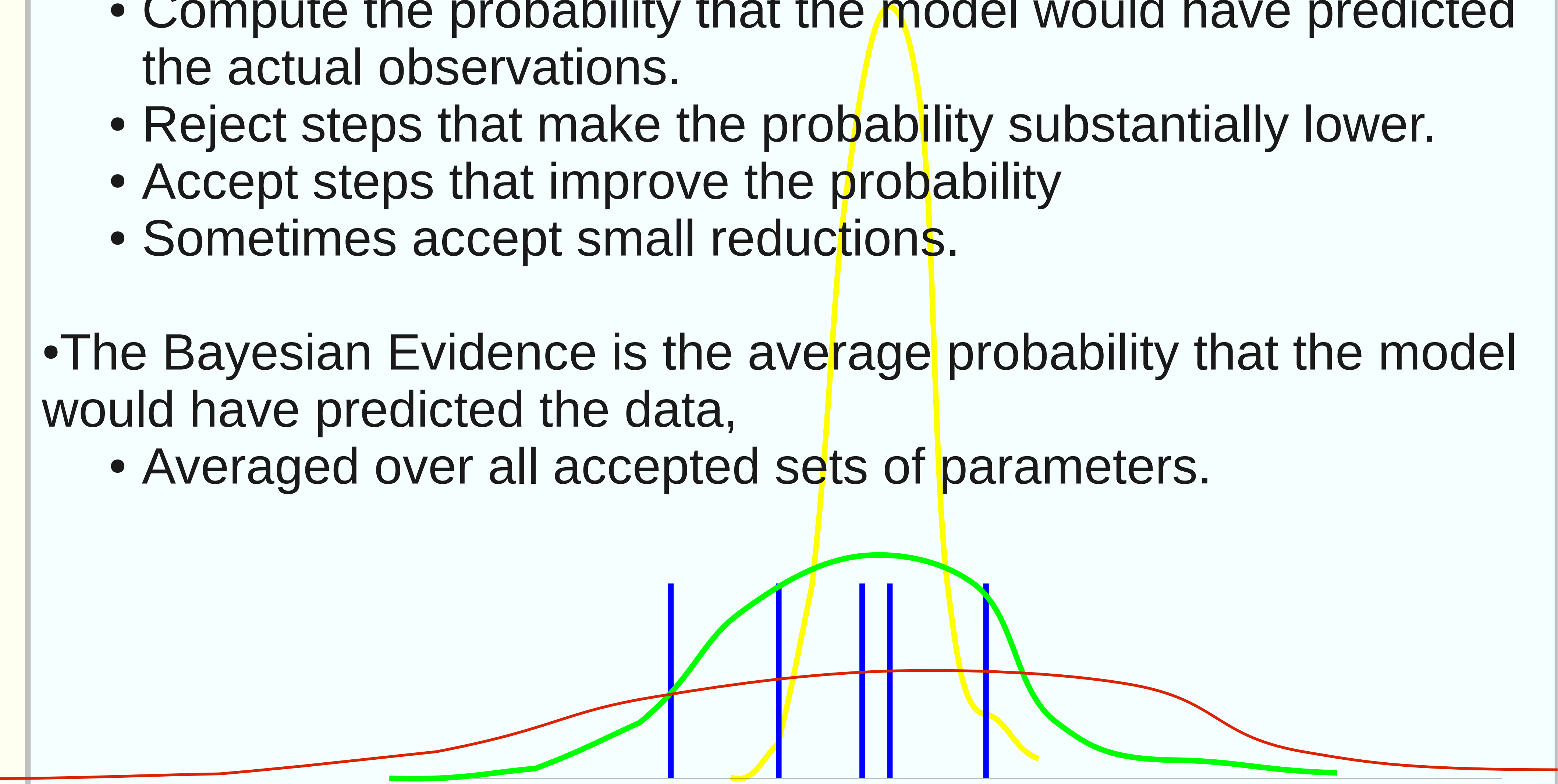
- All these models have some adjustable parameters:
  - What is the resting position of the tongue?
  - What is the articulatory target for /a/?
  - How much does [+HIGH] raise each point on the tongue?
  - How far is a typical measurement from its target?



The Model Comparison Algorithm

- Adjust the model parameters, step-by-step:
  - Use a Markov Chain Monte Carlo technique:
    - Compute the probability that the model would have predicted the actual observations.
    - Reject steps that make the probability substantially lower.
    - Accept steps that improve the probability
    - Sometimes accept small reductions.

- The Bayesian Evidence is the average probability that the model would have predicted the data,
    - Averaged over all accepted sets of parameters.



What are we testing?

- Are features good at representing articulatory positions?
- Are features better than other available models?

We compared features to:

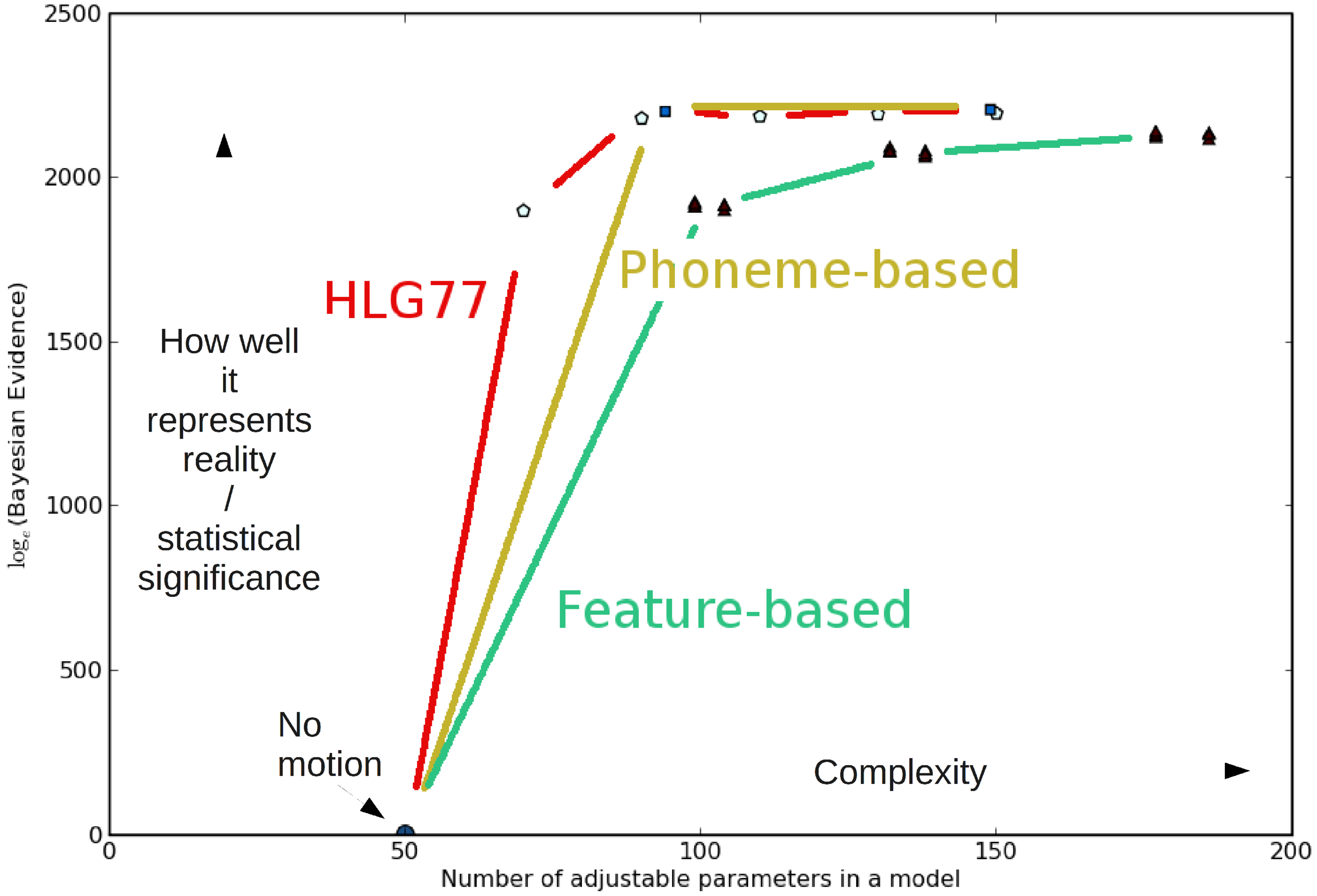
- Phonemes:
  - Two variants (more vs. less complexity in tongue shape).

- Harshman, Ladefoged, and Goldstein 1977:
  - Two factor model.
  - Factors are gradient, not binary.
  - 5 variants of the model, ranging from 1 to 5 factors.

The models in this first comparison are strictly local, with no effect of context. There are 31 models in total.

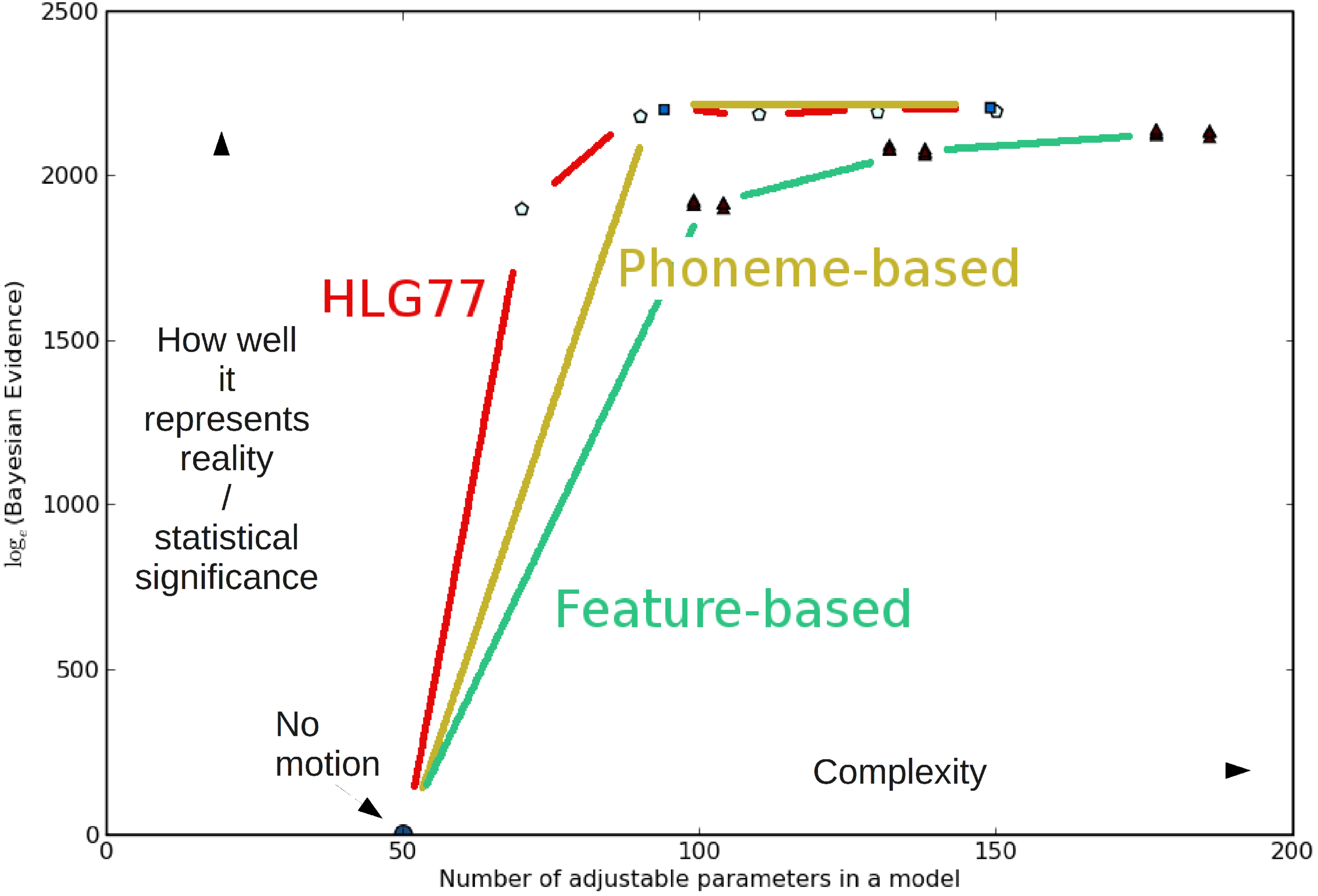
Features are significantly worse at describing articulation.

Local models: no context.



Features are significantly worse at describing articulation.

Local models: no context.



What does this imply?

- Features (absent context) are not special.
  - HLG models are better.
  - Phoneme models are better.
- Simple models capture most articulatory motion.
  - A simplified HLG model with a single(!) factor.
  - (We aren't looking at nasality, lips, or features of the larynx.)

What's a HLG model?

Harshman, Ladefoged, Goldstein 1977

•First model to quantitatively predict tongue shape.

•Fills a major hole in feature representations:

- Its features approximate the position and size of the airway constriction.
- They do not specify the rest of the tongue shape.

•Treats tongue shape as controlled by two factors that best explain tongue shapes

- One is overall upwards/backwards.
- One is root forward/front raising.
- Each vowel has a particular combination of these two motions.

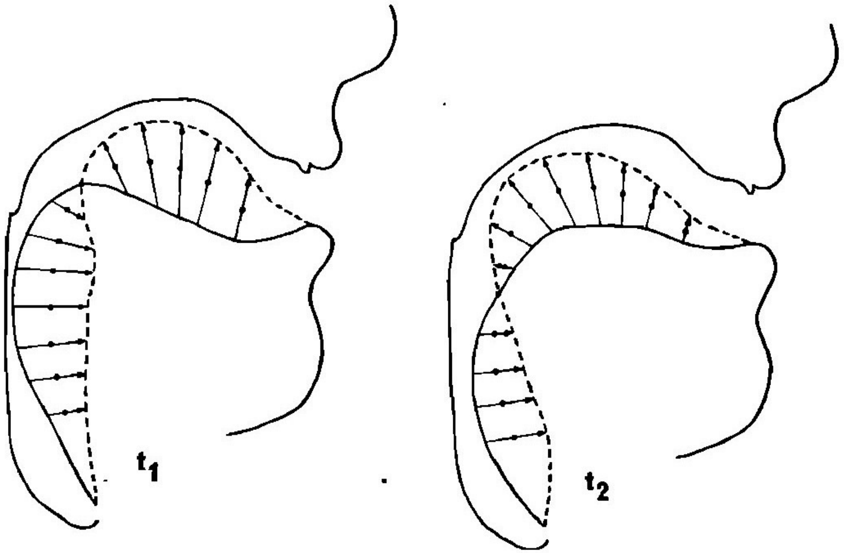


FIG. 3. A graphical representation of the two sets of constants, t_{i1} and t_{i2}, proportions of which can be combined to form different vowels. The solid points represent a neutral position of the tongue. The solid lines indicate tongue positions corresponding to large negative values of each factor. The dashed lines indicate large positive values.

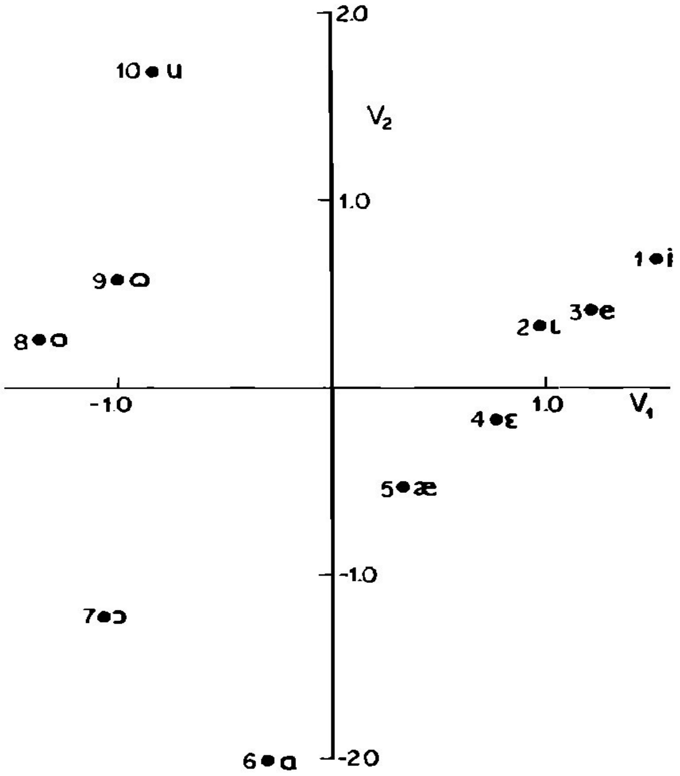


FIG. 7. The values of v_{j1} and v_{j2} for the 10 vowels in Table III.

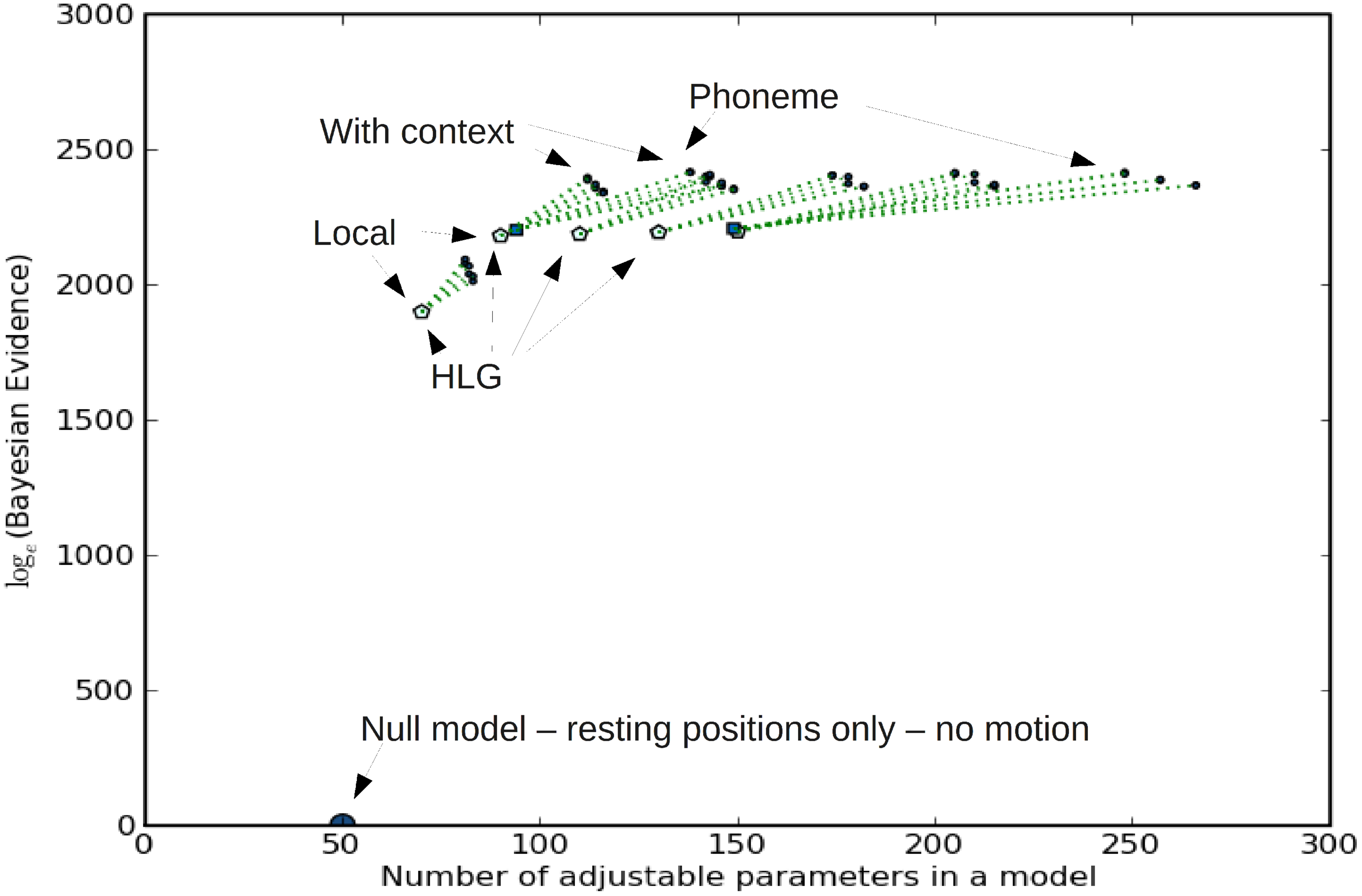
What are we testing?

- Does feature spreading improve the representation of articulatory positions?
- Do we get an improvement by adding context to phoneme or HLG models?
  - We pick two consonants from the set { /n/, /f/, /s/, /d/ }
  - We assume those two phones have different allophones, depending on:
    - Forward context (vowel, silence, or consonant), and
    - Backwards context (vowel, silence, or consonant).
    - E.g.: $/_v n_c/$ is treated as a different sound from $/_c n_s/$

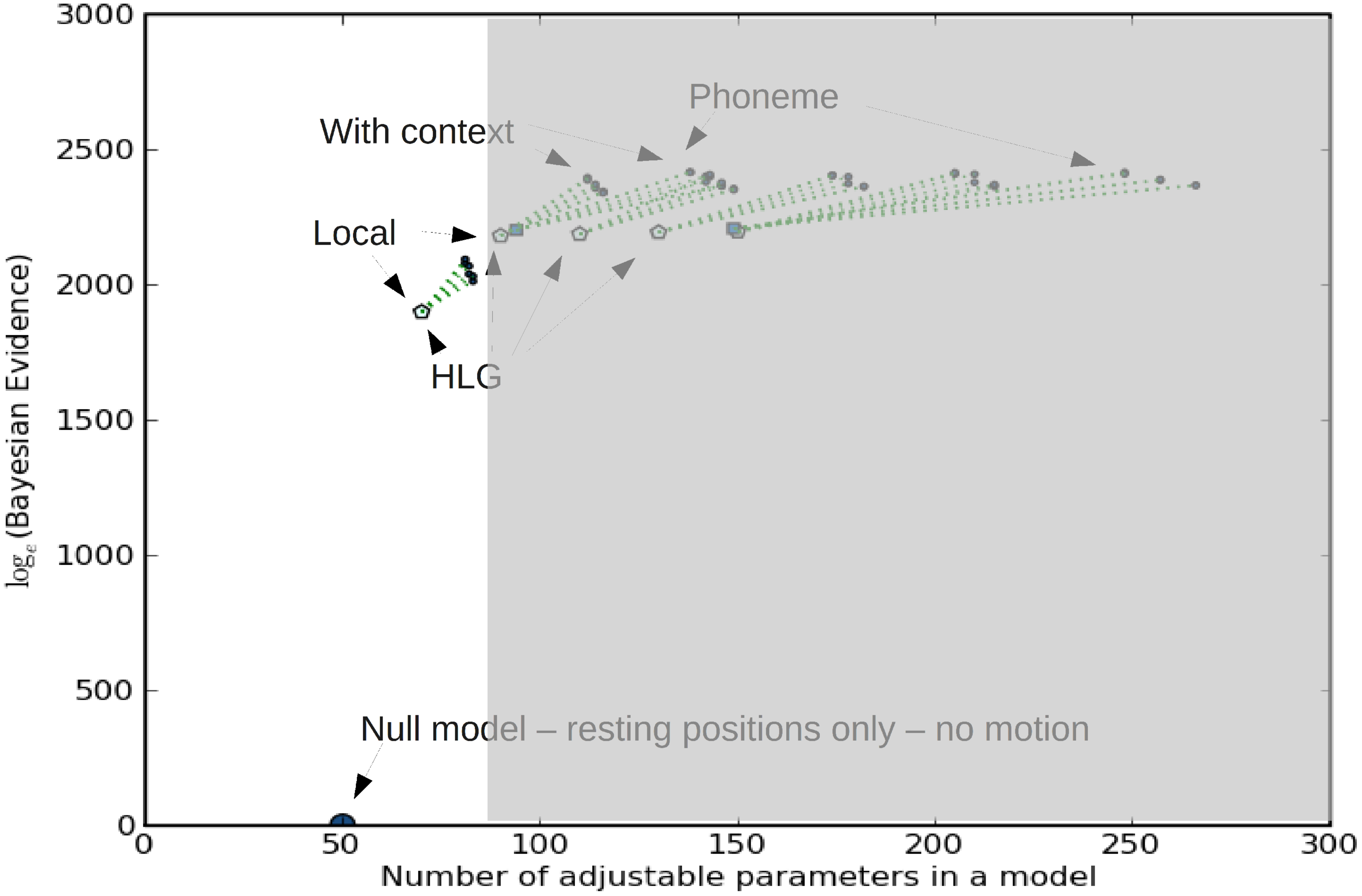
Compared to:

- Otherwise identical local models.

Adding context improves estimates of articulatory positions.



Complexity is better spent on context effects than local precision.



Phoneme

With context

Local

HLG

Null model – resting positions only – no motion

What does this imply?

- Features (absent context) are not special.
- Simple models capture most articulatory motion.
  - Five features.
  - A simplified HLG model with a single(!) factor.
  - (We aren't looking at nasality, lips, or features of the larynx.)
- It is better to add context effects than elaborate a description that doesn't use context.
  - (Except for the simplest descriptions.)

What are we testing?

- Does feature spreading improve the representation of articulatory positions?

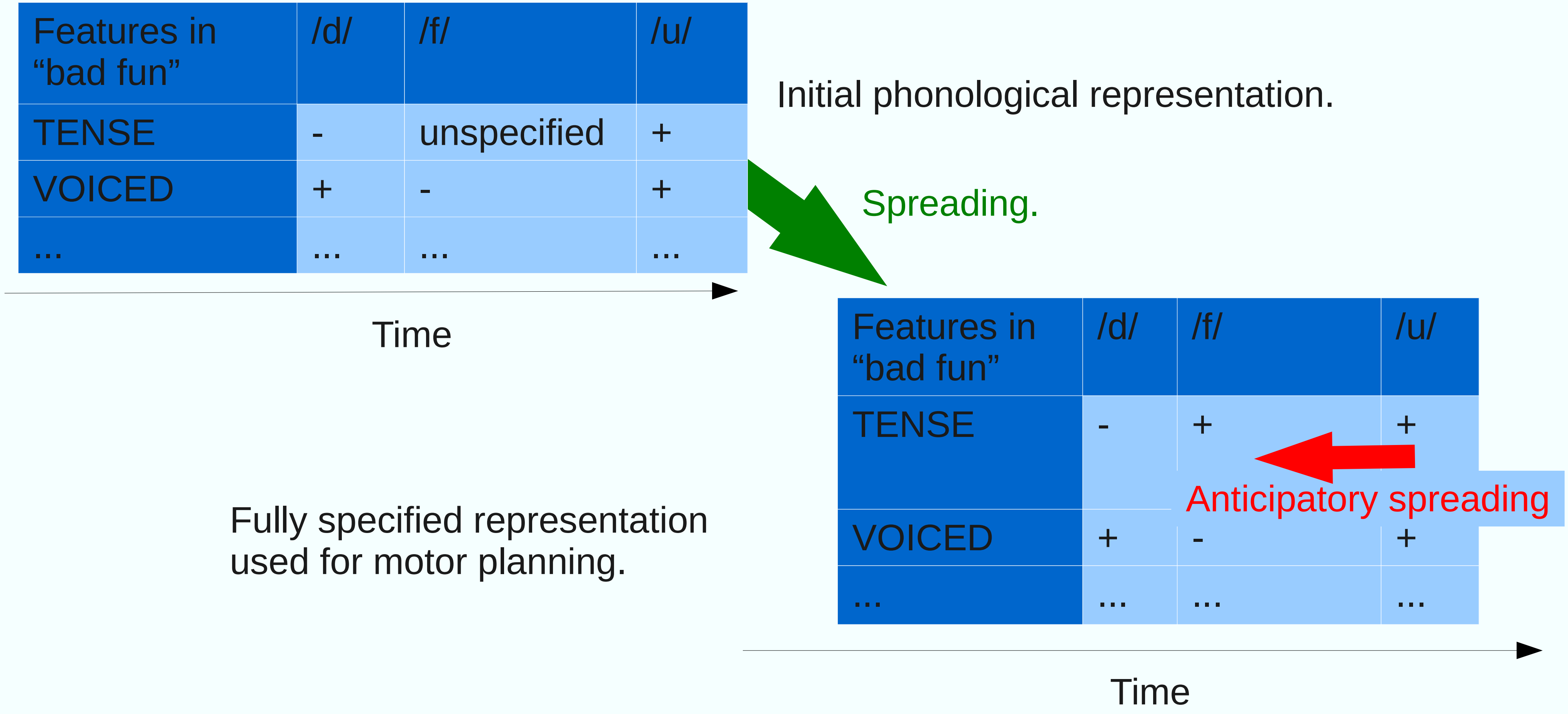
Compared to:

- The best local feature-based models.
- Phoneme models with context.
- Harshman, Ladefoged, and Goldstein models with context.

What are we testing?

- Does feature spreading improve the representation of articulatory positions?

What do we mean by “feature spreading”?



Features in “bad fun”	/d/	/f/	/u/
TENSE	-	unspecified	+
VOICED	+	-	+
...	...	...	...

Time

Initial phonological representation.

Spreading.

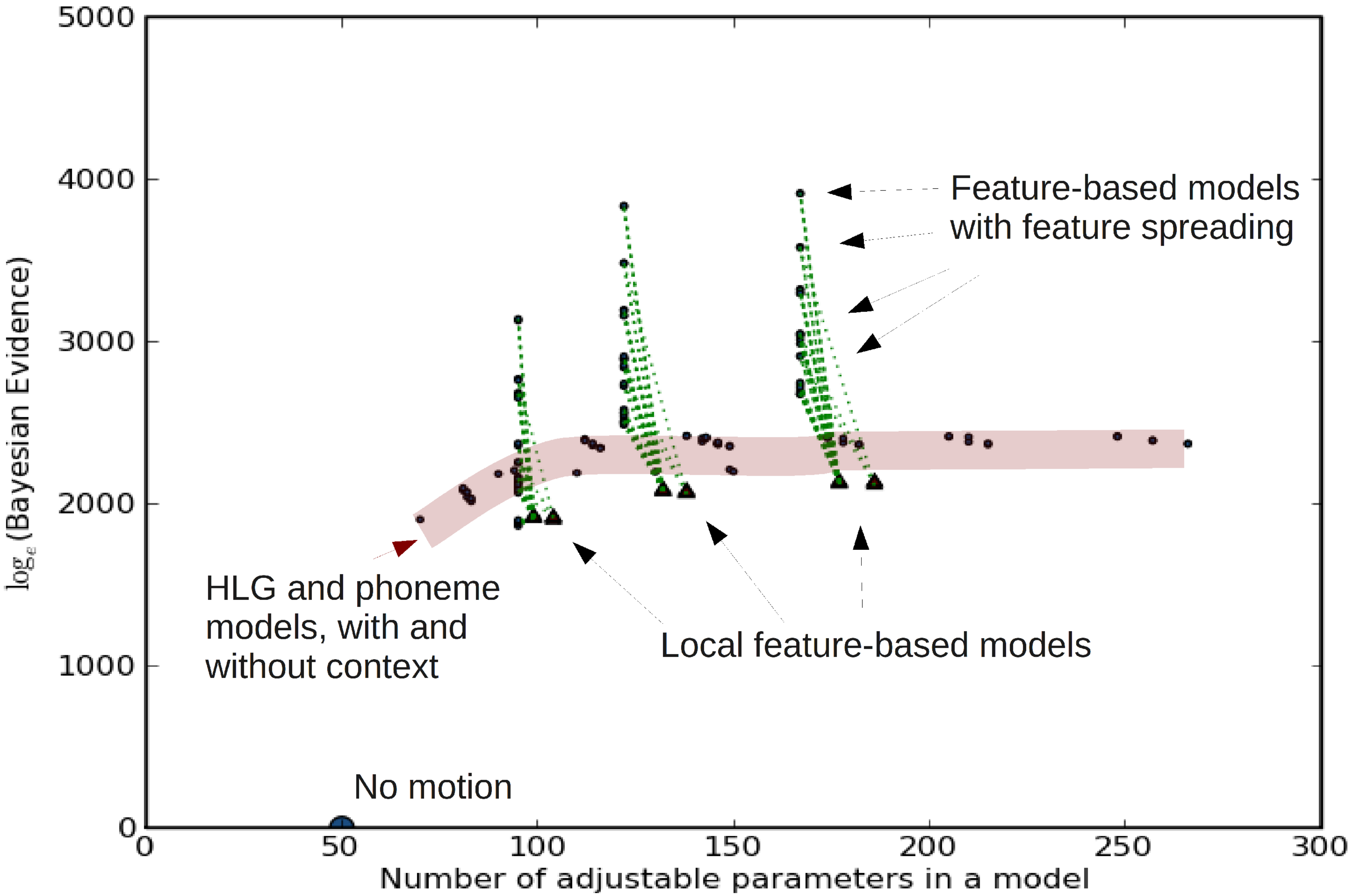
Features in “bad fun”	/d/	/f/	/u/
TENSE	-	+	+
VOICED	+	-	+
...	...	...	...

Anticipatory spreading

Time

Fully specified representation used for motor planning.

Context via feature spreading: dramatic improvement.



Feature-based models with feature spreading

HLG and phoneme models, with and without context

Local feature-based models

No motion

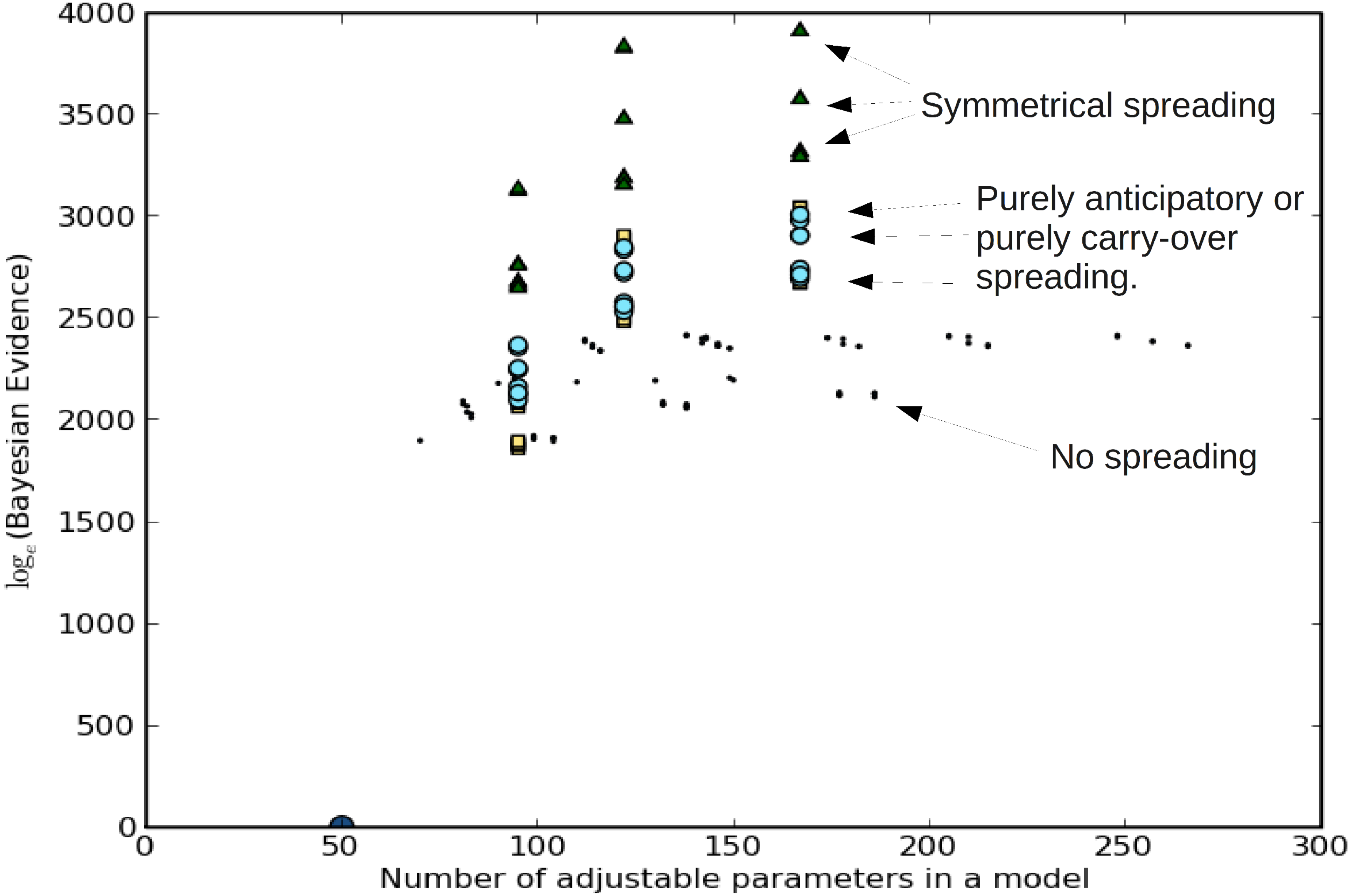
What does this imply?

- Features (absent context) are not special.
- Simple models capture most articulatory motion.
  - Five features.
  - A simplified HLG model with a single(!) factor.
  - (We aren't looking at nasality, lips, or features of the larynx.)
- It is better to add context effects than elaborate a description that doesn't use context.
  - (Except for the simplest descriptions.)
  - 5 features may be too many without context.
- Features with partial specification and spreading work well.

What are we testing?

- What kind of feature spreading works best?
  - Anticipatory
  - Carry-over
  - Symmetrical

Symmetrical Feature Spreading is Best



Symmetrical spreading

Purely anticipatory or purely carry-over spreading.

No spreading

What does this imply?

- Features (absent context) are not special.
- Simple models capture most articulatory motion.
  - Five features.
  - A simplified HLG model with a single(!) factor.
  - (We aren't looking at nasality, lips, or features of the larynx.)
- It is better to add context effects than elaborate a description that doesn't use context.
  - (Except for the simplest descriptions.)
  - 5 features may be too many without context.
- Features with partial specification and spreading work well.
- Spreading is symmetrical
  - We fill unspecified features with [-,0,+]
  - Non-binary feature filling can work.

Conclusions

If you're committed to features:

- Recognize that there's a chance something better might come along.
- Partial specification / spreading is important.
  - It may work better with symmetrical, 3-state filling.
- This group of features is important:
- HIGH, LOW, CORONAL, BACK
- Some seem unimportant compared to context:
- CONTINUANT, ANTERIOR, TENSE, ATR
- Specifying a few features with context is better than all without.

If you're not committed to features:

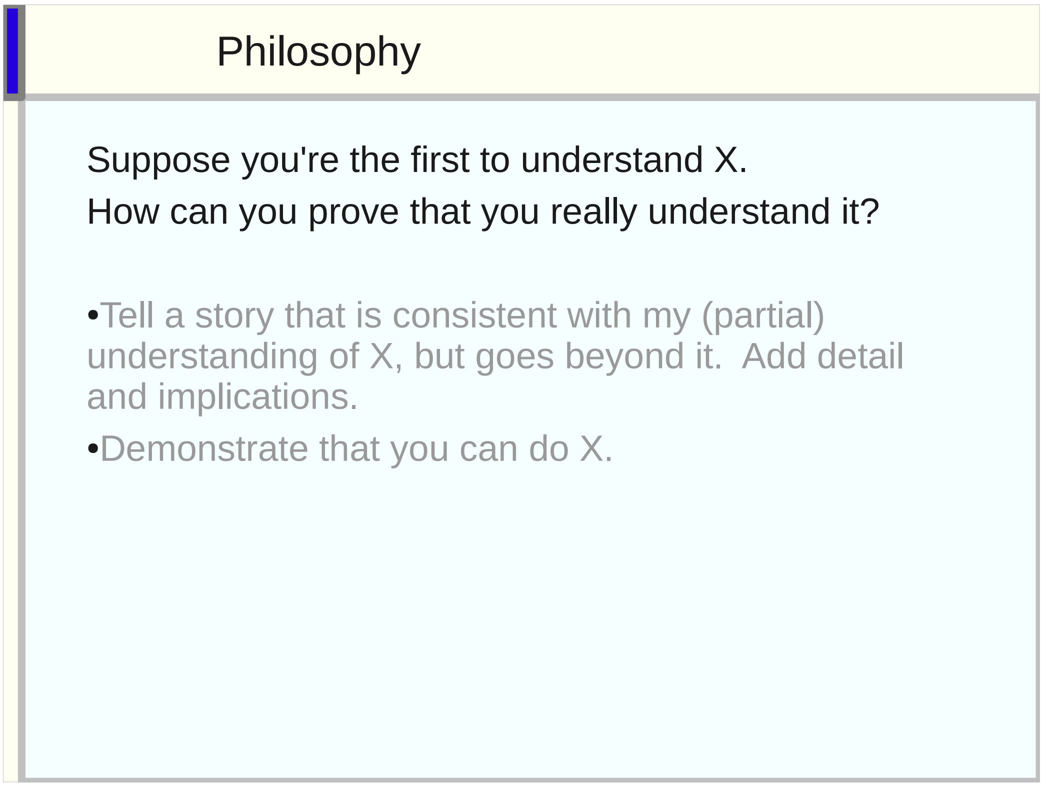
- Think of a way to add context effects to HLG models
- Think of a way to do phonology with HLG-like models.

Experimental tests of Features and Partial Specification

Greg Kochanski
John Coleman
Christina Orphanidou
(Phonetics Laboratory, Oxford)

Christopher Alvey
A. McIntyre
Stephen Golding
(Nuffield Dept. of Surgery, Oxford)

Funded by the UK ESRC via grant RES-000-23-1094.
Talk presented by G. Kochanski, 17 December 2010,
at the Laboratoire Parole et Langage,
Université de Provence, Aix-en-Provence, France.
This can be downloaded from http://kochanski.org/gpk/papers/2010/Aix_underspecification.pdf,
http://www.phon.ox.ac.uk/releases/pdfs/Aix_underspecification.pdf, or the University of Oxford Library Service.



Philosophy

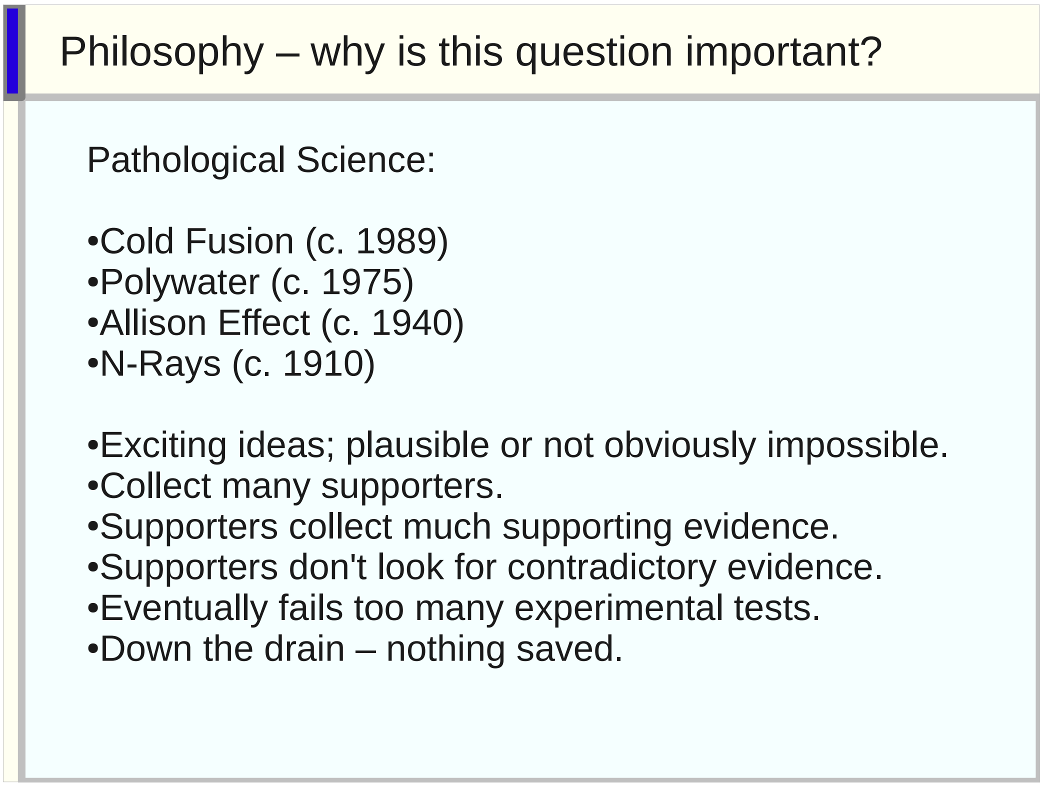
Suppose you're the first to understand X.
How can you prove that you really understand it?

- Tell a story that is consistent with my (partial) understanding of X, but goes beyond it. Add detail and implications.
- Demonstrate that you can do X.

But before I talk about the experiments, I'd like to discuss how one can test an idea that is abstract as distinctive features.

This is really the central question of science – if there is no good way to show that you really understand something, then how can the next generation of scholars tell who they should listen to? Absent some way of separating correct beliefs from false beliefs, transmission will be random, controlled by sociology, and knowledge will not progress.

Of course,the dichotomy of correct and false is an exaggeration. It would normally be better to say “a better/worse description of nature.” Likewise, “proof” is too strong a term. Science generally does not depend on proof. Rather, is is pragmatic: one uses an idea when it is the best available, with no known serious flaws. One discards an idea when it something better comes along, or when it is known to fail in some important way.



Philosophy – why is this question important?

Pathological Science:

- Cold Fusion (c. 1989)
- Polywater (c. 1975)
- Allison Effect (c. 1940)
- N-Rays (c. 1910)

- Exciting ideas; plausible or not obviously impossible.
- Collect many supporters.
- Supporters collect much supporting evidence.
- Supporters don't look for contradictory evidence.
- Eventually fails too many experimental tests.
- Down the drain – nothing saved.

Science has a failure every few decades, where people go off into a dead end. Most recently, it was Cold Fusion. It was an exciting idea, but it rapidly failed experimental tests and within weeks, most scientists had decided that it wasn't worth attention. But not all: a group of committed people hung on for months and even years, in the face of mounting evidence. See http://kochanski.org/gpk/teaching/0601Oxford/doubt.pdf .

Philosophy

Suppose you're the first to understand X.
How can you prove that you really understand it?

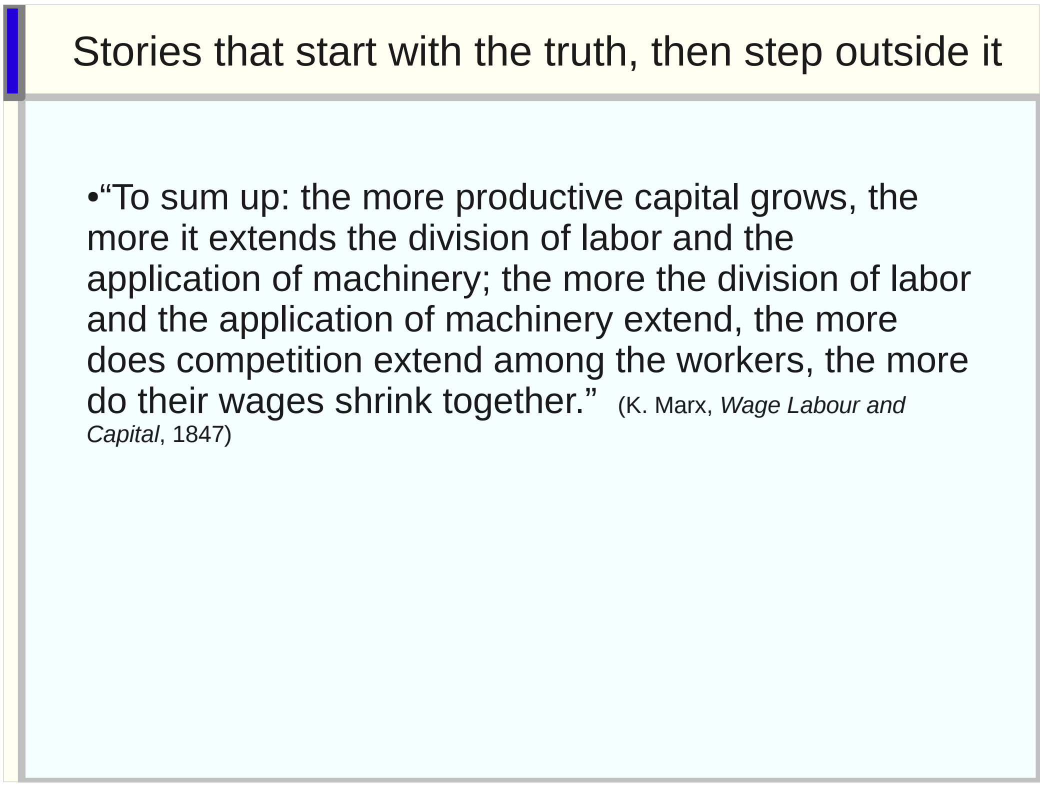
- Tell a story that is consistent with my (partial) understanding of X, but goes beyond it. Add detail and implications.
- Demonstrate that you can do X.

Stories that start with the truth, then step outside it

- Malaria is common near swamps; you can smell swamps; therefore the disease is caused by the smell.

- "What we have found over the years in the marketplace is that derivatives have been an extraordinarily useful vehicle to transfer risk from those who shouldn't be taking it to those who are willing to and are capable of doing so." - Alan Greenspan (US Federal Reserve Chairman) 2003

"Not only have individual financial institutions become less vulnerable to shocks from underlying risk factors, but also the financial system as a whole has become more resilient." - Alan Greenspan in 2004

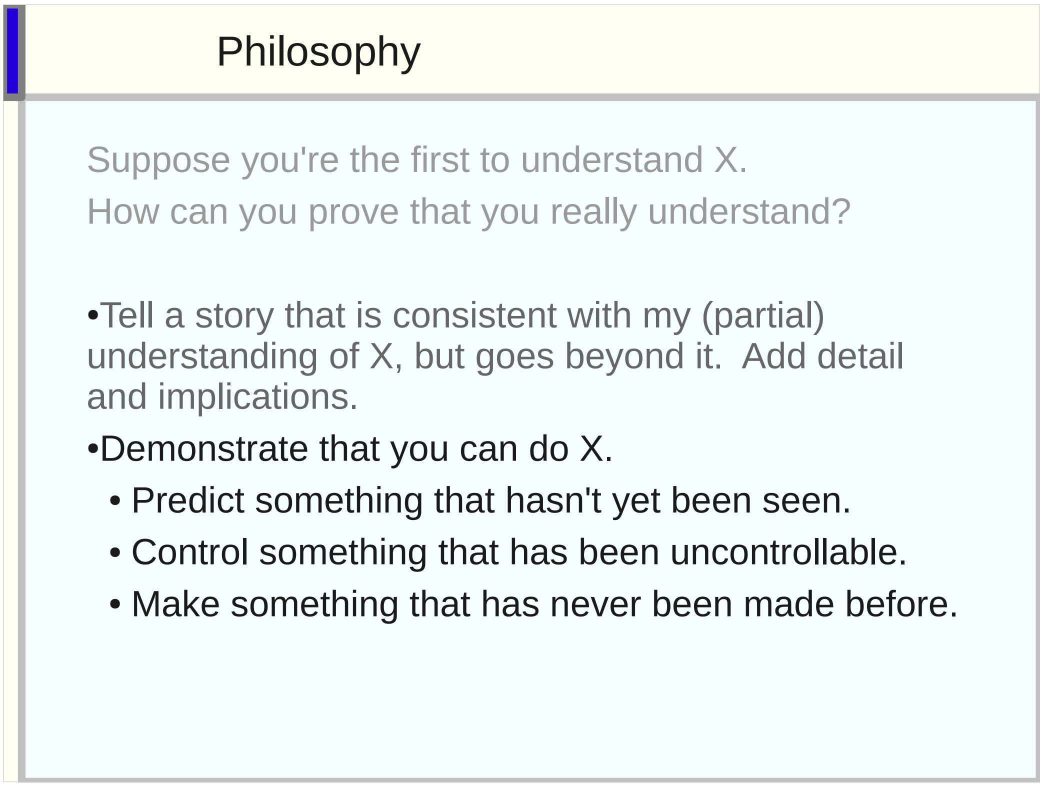


Stories that start with the truth, then step outside it

- "To sum up: the more productive capital grows, the more it extends the division of labor and the application of machinery; the more the division of labor and the application of machinery extend, the more does competition extend among the workers, the more do their wages shrink together." (K. Marx, *Wage Labour and Capital*, 1847)

Karl Marx is being very perceptive and logical here, but he is only thinking about jobs where a human is in direct competition with the machinery. For those jobs, his prediction was accurate (except for a handful of artisans who survive by producing luxury goods).

However, what he didn't imagine was that new products and industries would grow, that products would become more complicated, and that there were many things to do around the machines. As a result, workers have done fairly well over the last century, and his conclusion turned out to be false.



Philosophy

Suppose you're the first to understand X.
How can you prove that you really understand?

•Tell a story that is consistent with my (partial) understanding of X, but goes beyond it. Add detail and implications.

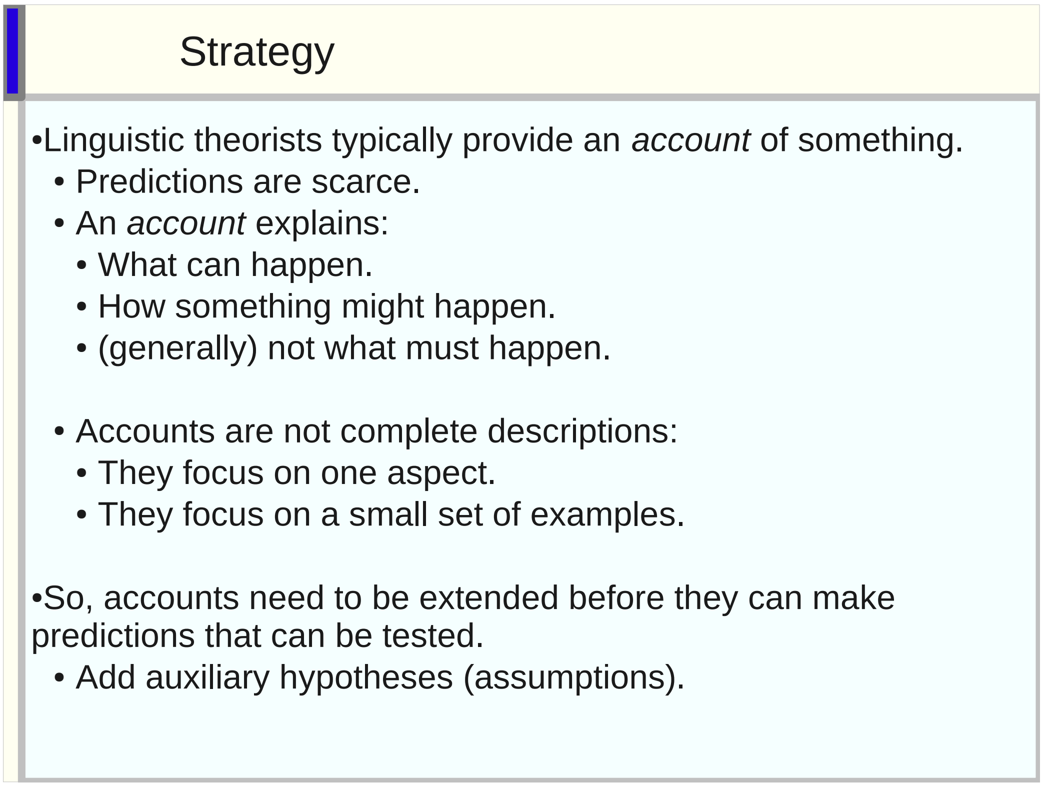
•Demonstrate that you can do X.

- Predict something that hasn't yet been seen.
- Control something that has been uncontrollable.
- Make something that has never been made before.

So, telling a story is not enough. A theory or story that is consistent with everything that we know could still be wrong once it steps away from the known facts.

A mathematical analogy can be made if we ask “What is the next number after 1, 3, 5, 7?”   The simple, tempting answer is “9”, assuming the numbers follow an arithmetical progression, adding 2 each time.   But, another perfectly good answer is to assume we have the sequence of the odd numbers that do not have integer factors – i.e. prime numbers.   Then, 9 isn't a prime, so the answer is the next prime, 11. Generally, there are many ways to continue any sequence, and therefore no argument based on the known facts is guaranteed to yield the next unknown fact.

So, if pure logic and mathematics cannot help, we need to turn to experiment of some kind.



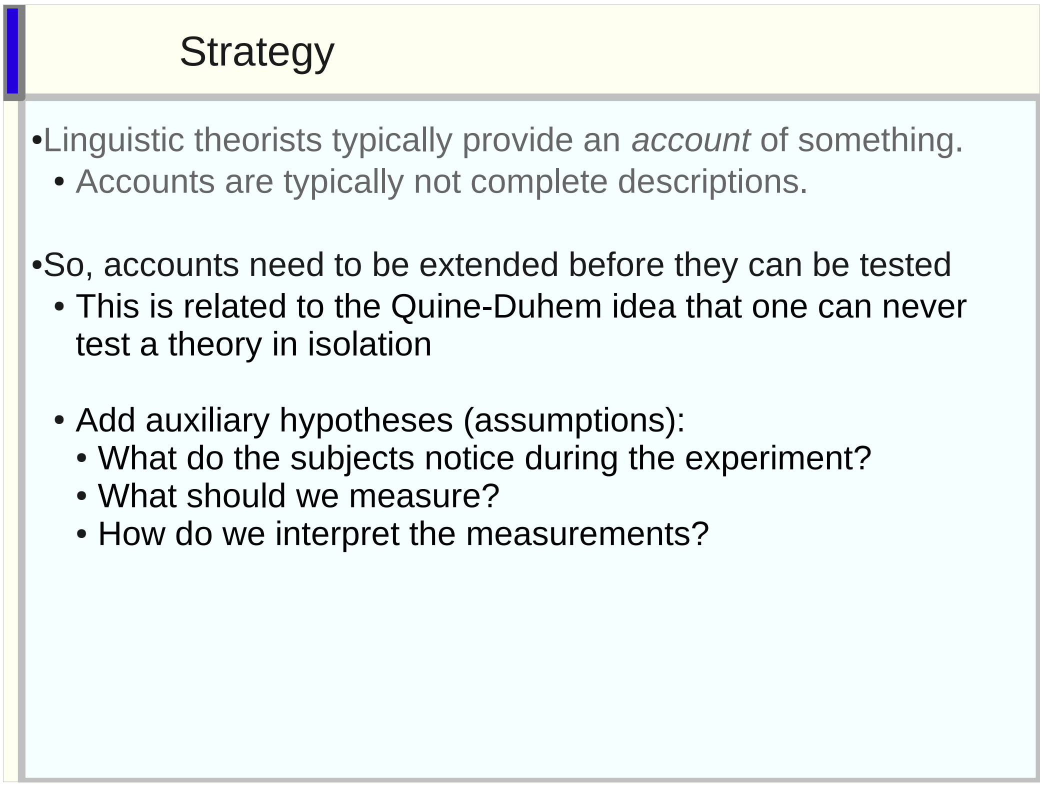
Strategy

- Linguistic theorists typically provide an *account* of something.
  - Predictions are scarce.
  - An *account* explains:
    - What can happen.
    - How something might happen.
    - (generally) not what must happen.
  - Accounts are not complete descriptions:
    - They focus on one aspect.
    - They focus on a small set of examples.
- So, accounts need to be extended before they can make predictions that can be tested.
  - Add auxiliary hypotheses (assumptions).

Accounts are not theories, in that they do not (typically) claim that their internal mechanisms reflect the brain's mechanisms. Nor do they usually allow for prediction. An account for X can typically be interpreted as “Here's one way to explain X, but there might be other ways.”

Overall, accounts are unsatisfactory because they are (usually) so cautiously stated that they cannot be used to make any predictions, so one can never disprove them.

But, accounts can be made more complete, by adding extra assumptions. You can change “...one way...” to “...the way...”, and assume that the account generalizes across a whole class of examples. Likewise, you can add plausible ties between the internal mechanisms of the account and observable effects. Given enough additional assumptions, you can create a predictive theory that is consistent with the linguistic account.

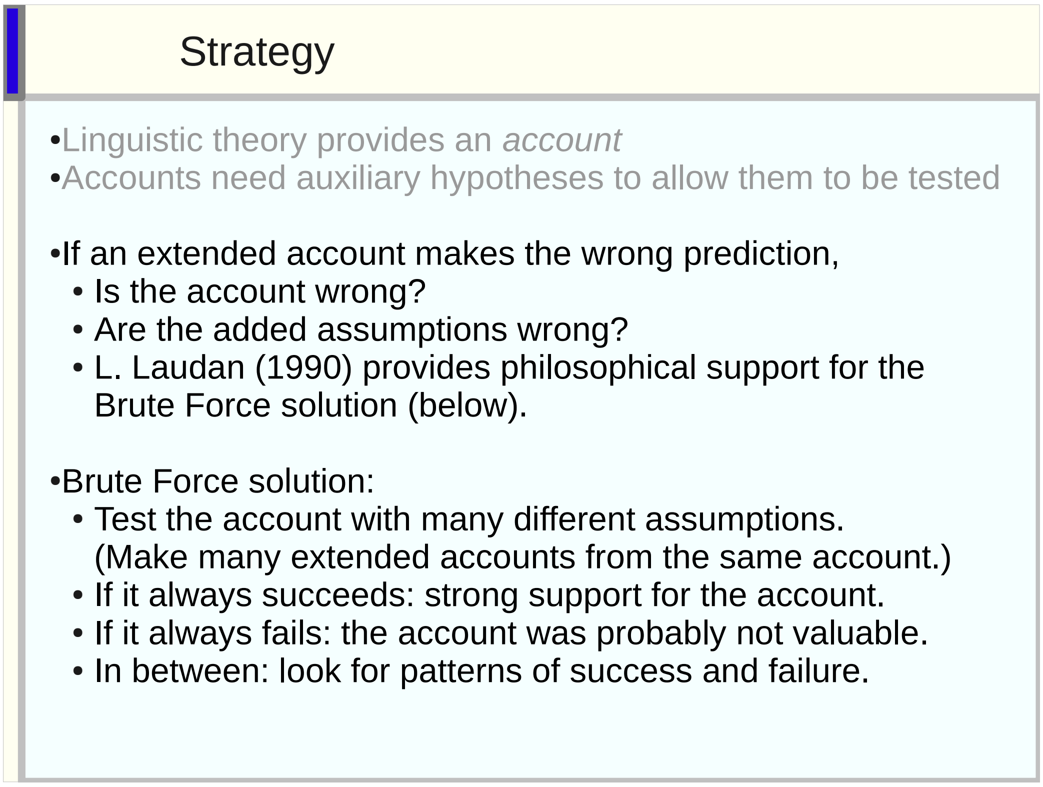


Strategy

- Linguistic theorists typically provide an *account* of something.
  - Accounts are typically not complete descriptions.
- So, accounts need to be extended before they can be tested
  - This is related to the Quine-Duhem idea that one can never test a theory in isolation
  - Add auxiliary hypotheses (assumptions):
    - What do the subjects notice during the experiment?
    - What should we measure?
    - How do we interpret the measurements?

The D. V. Quine - Pierre Duhem thesis states that one cannot test any theory in isolation. Even in the hard sciences, where one can (for example) predict a magnetic field strength, one still needs auxiliary hypotheses to specify thow one's measuring instruments respond to the field.

In linguistic experiments, we likewise need to specify how the experimental conditions affect the system under test (i.e. our experimental subject) and how we will measure and interpret his/her responses. These are in addition to any auxiliary hypotheses that we need to convert an account into a predictive theory.



Strategy

- Linguistic theory provides an *account*
- Accounts need auxiliary hypotheses to allow them to be tested

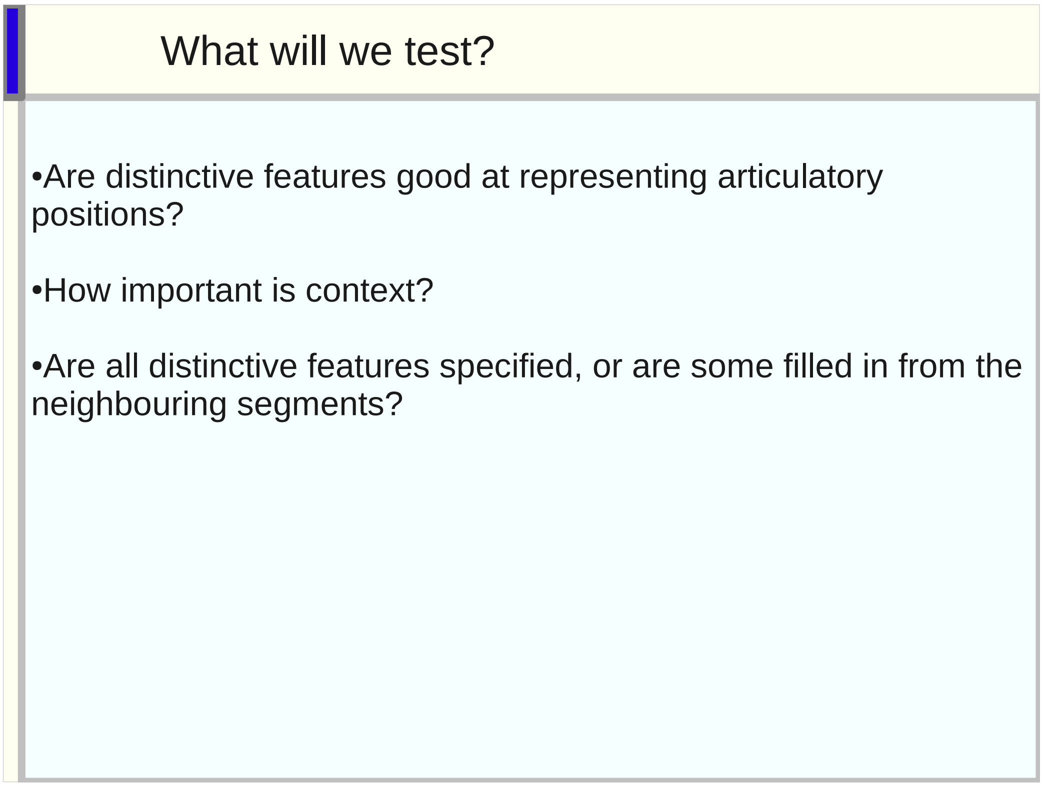
- If an extended account makes the wrong prediction,
  - Is the account wrong?
  - Are the added assumptions wrong?
  - L. Laudan (1990) provides philosophical support for the Brute Force solution (below).

- Brute Force solution:
  - Test the account with many different assumptions. (Make many extended accounts from the same account.)
  - If it always succeeds: strong support for the account.
  - If it always fails: the account was probably not valuable.
  - In between: look for patterns of success and failure.

The problem is that there is no certainty that we have added correct auxiliary hypotheses to our original account. So, if the extended account gives bad predictions, we do not know where the error lies: is it in the original account, or does it lie in the assumptions that we have added?

D.V. Quine sees this as a serious problem: that we cannot know what change to make in our web of beliefs as the result of an experiment. We take a more pragmatic approach similar to Laudan (Larry Laudan, University of Chicago Press, 1990 *Science and relativism: some key controversies in the philosophy of science*, isbn:9780226469492.) Some changes are more sensible than others.

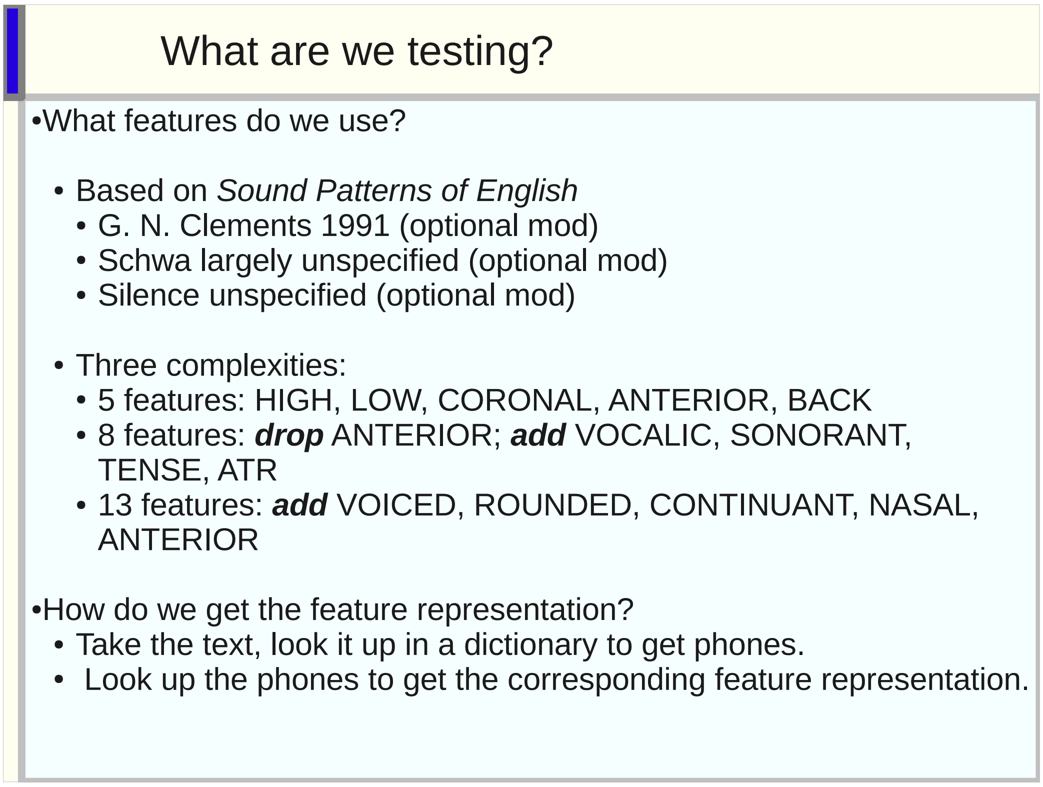
Then, we make use of this by our brute-force approach. We build many, different testable hypotheses from one linguistic account, and then we test each one. Te account is (ideally) the one shared link in all the tests, so if there a strong pattern of success or failure, Occam's razor tells us to put the praise or blame on the shared link.



What will we test?

- Are distinctive features good at representing articulatory positions?
- How important is context?
- Are all distinctive features specified, or are some filled in from the neighbouring segments?

These are our three basic research questions. We will answer them with a series of comparisons between testable hypotheses derive from different linguistic accounts.



What are we testing?

•What features do we use?

- Based on *Sound Patterns of English*
  - G. N. Clements 1991 (optional mod)
  - Schwa largely unspecified (optional mod)
  - Silence unspecified (optional mod)

- Three complexities:
  - 5 features: HIGH, LOW, CORONAL, ANTERIOR, BACK
  - 8 features: ***drop*** ANTERIOR; ***add*** VOCALIC, SONORANT, TENSE, ATR
  - 13 features: ***add*** VOICED, ROUNDED, CONTINUANT, NASAL, ANTERIOR

•How do we get the feature representation?

- Take the text, look it up in a dictionary to get phones.
- Look up the phones to get the corresponding feature representation.

Our core research question will be: are features useful. We will use standard Chomsky feature representations of phones, plus a few minor variants.

We take the words, look them up in a dictionary to get a phone sequence, then look up the features corresponding to each phone.

Methods

- •Our data are mid-sagittal magnetic resonance images.
- •We process the images to find the tongue's edges.
- •We wrote and used noise reduction software to subtract much of the noise from the MRI machine.
- •We segmented the speech to find the centres and/or edges of the phones.
  - (Because of the MRI noise, certain boundaries could not be precisely located.)

MRI Video:

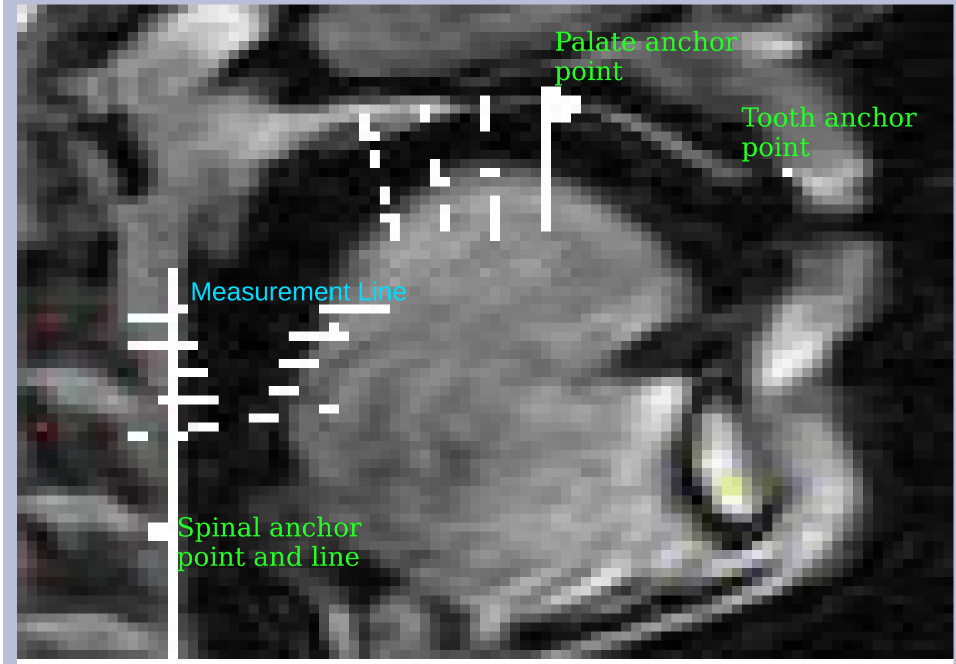
http://kochanski.org/gpk/papers/2010/Aix_underspecification/hr3.mpg

Here is some sample MRI data we used in this experiment.

The data is collected in “gated” mode: i.e. stroboscopically. The volunteer repeats a phrase about 18 times to the click of a metronome (“under the desk. Under the desk. Under the desk....) and from that, the MRI system reconstructs 60 time slices. So, effectively, these are 60 independent time slices at about 15 millisecond intervals. It is a very good imaging technique for phrases that are short enough to be precisely repeated.

The overlay shows the first step of the data analysis: finding the tongue edges.

Image Processing Strategy



We measure the proton density along nine lines that are two pixels wide. The lines are scaled to different people's anatomy by anchoring them to:

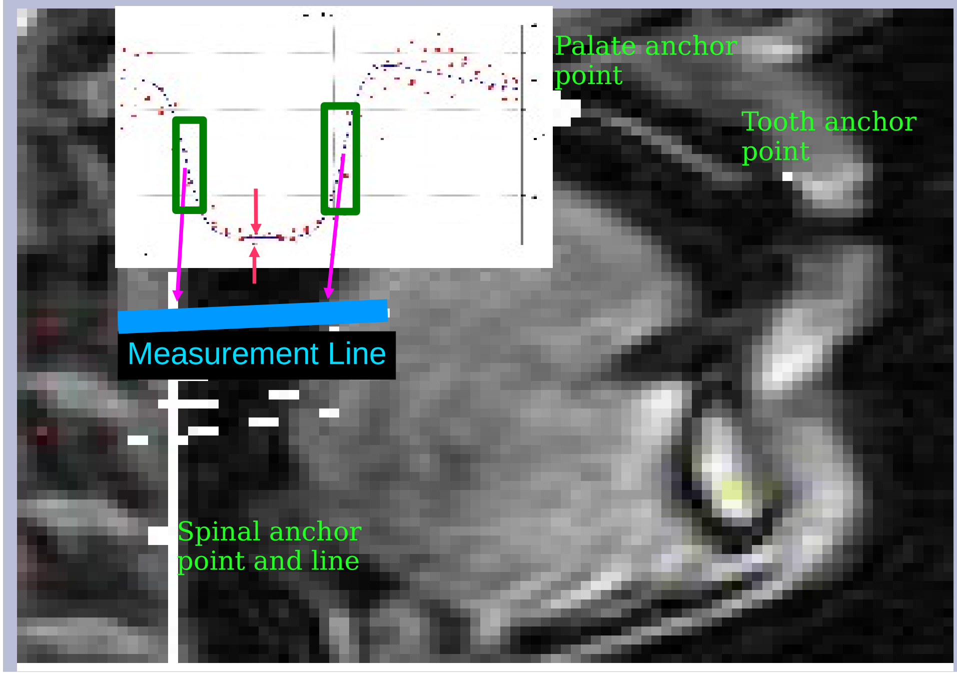
* a line along the back of the throat,

* the point where that line crosses exits the second cervical vertebra,

* the front centre tooth, and

* the top of the arch of palate.

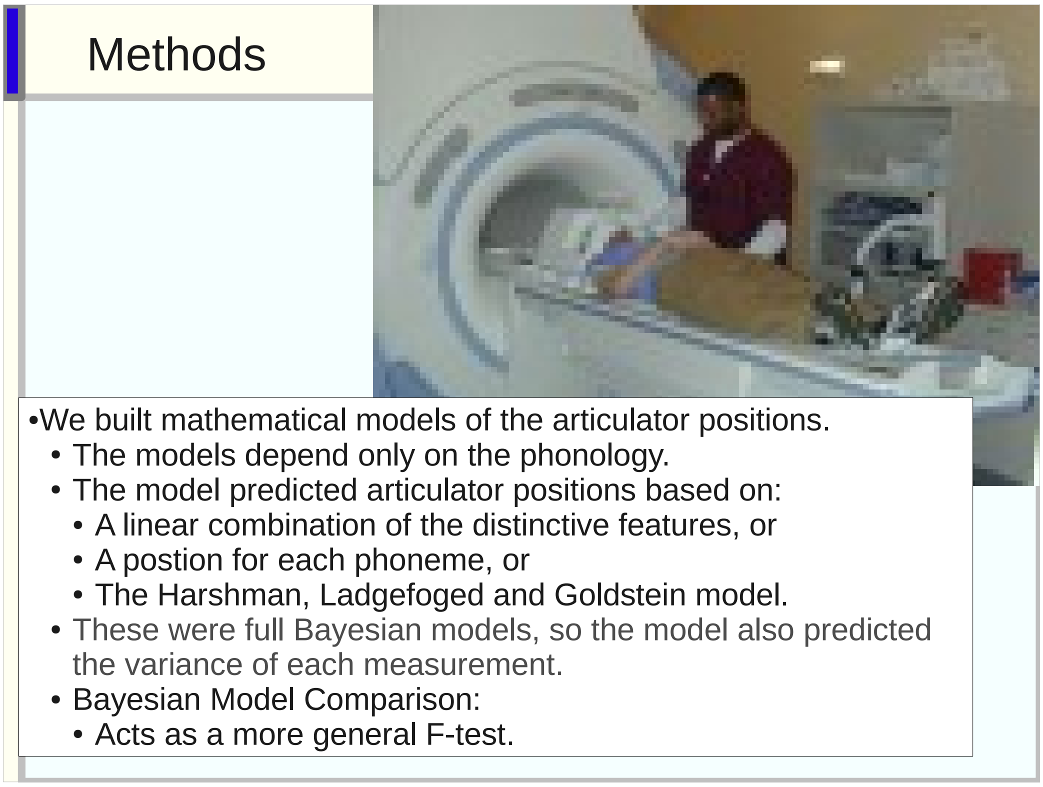
Image Processing Strategy



Along each line, we tale the pixels within about a pixel of the center-line and fit them to a function that has a high proton density at each edge, corresponding to the tongue or the tissue at the back of the throat, and a low density in the middle, correponding to the empty airway. The edges are blurred to represent the MRI system's imaging resolution.

The edges of the airway are defined to be the points where the density is half-way between the airway and the edge.

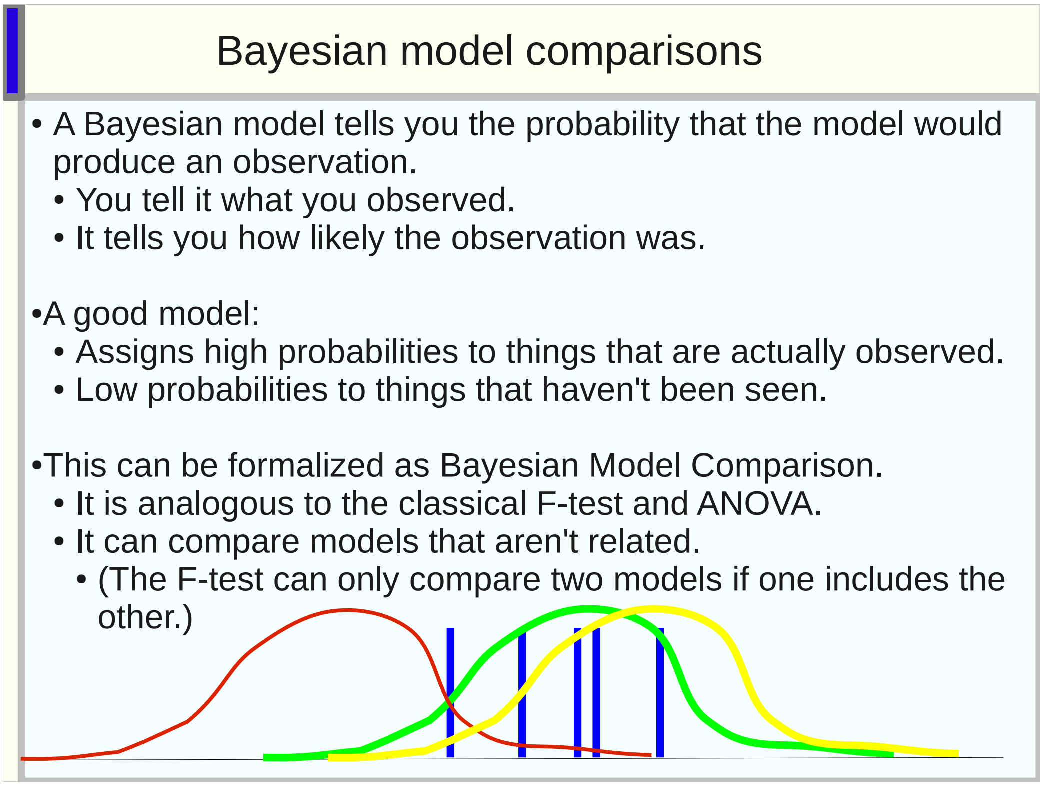
To estimate the uncertainties in the tongue edge, we use a bootstrap resampling technique. This involves taking different collections of pixels, re-running the analysis, and looking at the scatter of the results. In this image, the different estimates of the tongue edge are given by the tick-marks perpendicular to the measurement line.



Methods

- We built mathematical models of the articulator positions.
  - The models depend only on the phonology.
  - The model predicted articulator positions based on:
    - A linear combination of the distinctive features, or
    - A postion for each phoneme, or
    - The Harshman, Ladgefoged and Goldstein model.
  - These were full Bayesian models, so the model also predicted the variance of each measurement.
  - Bayesian Model Comparison:
    - Acts as a more general F-test.

A Bayesian model predicts the most likely outcome of an experiment, but also the full probability distribution of results. Typically, if one operates with a normal distribution, that means needs to predict the variance as well as the mean.



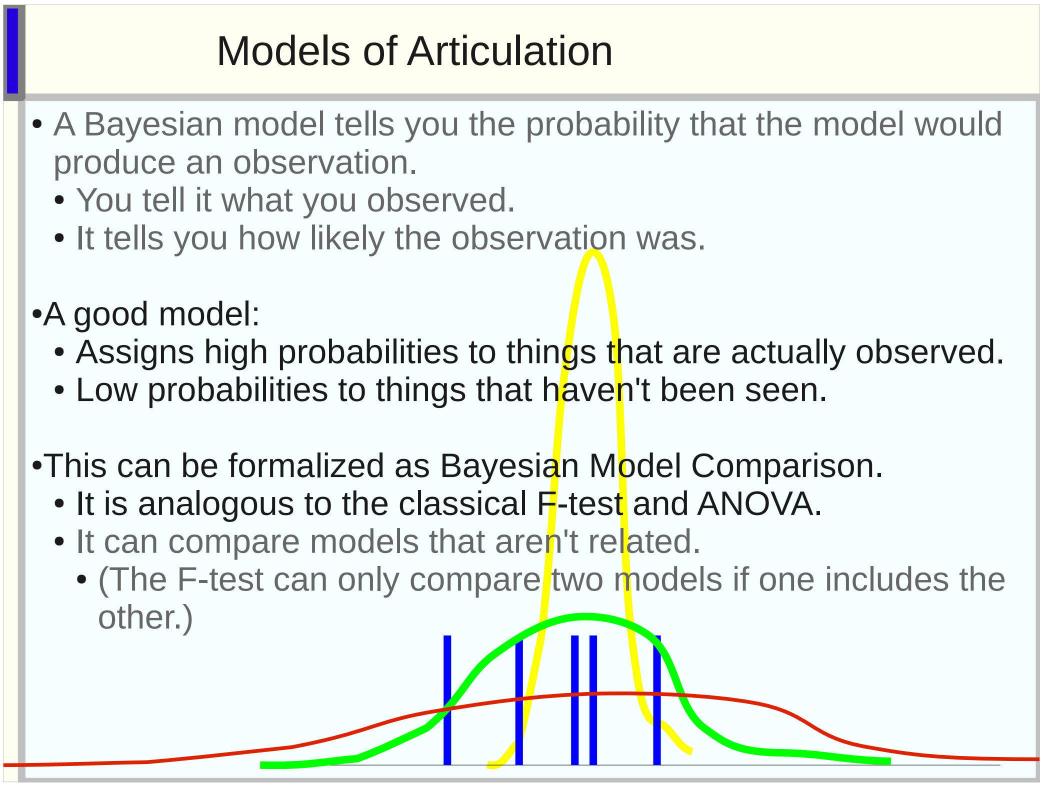
Bayesian model comparisons

- A Bayesian model tells you the probability that the model would produce an observation.
  - You tell it what you observed.
  - It tells you how likely the observation was.
- A good model:
  - Assigns high probabilities to things that are actually observed.
  - Low probabilities to things that haven't been seen.
- This can be formalized as Bayesian Model Comparison.
  - It is analogous to the classical F-test and ANOVA.
  - It can compare models that aren't related.
    - (The F-test can only compare two models if one includes the other.)

In the drawing at the bottom, the green curve is the probability distribution that is the most likely to lead to the observed data (blue lines). The red curve assigns very low probabilities to the three right-most observations, and the yellow curve would be unlikely to generate the left-most observation.

As a result, the Bayesian posterior probability of the green probability distribution is higher, and it would win in a Bayesian model comparison.

Each of these probability distributions corresponds to a different mathematical model, of course.



Models of Articulation

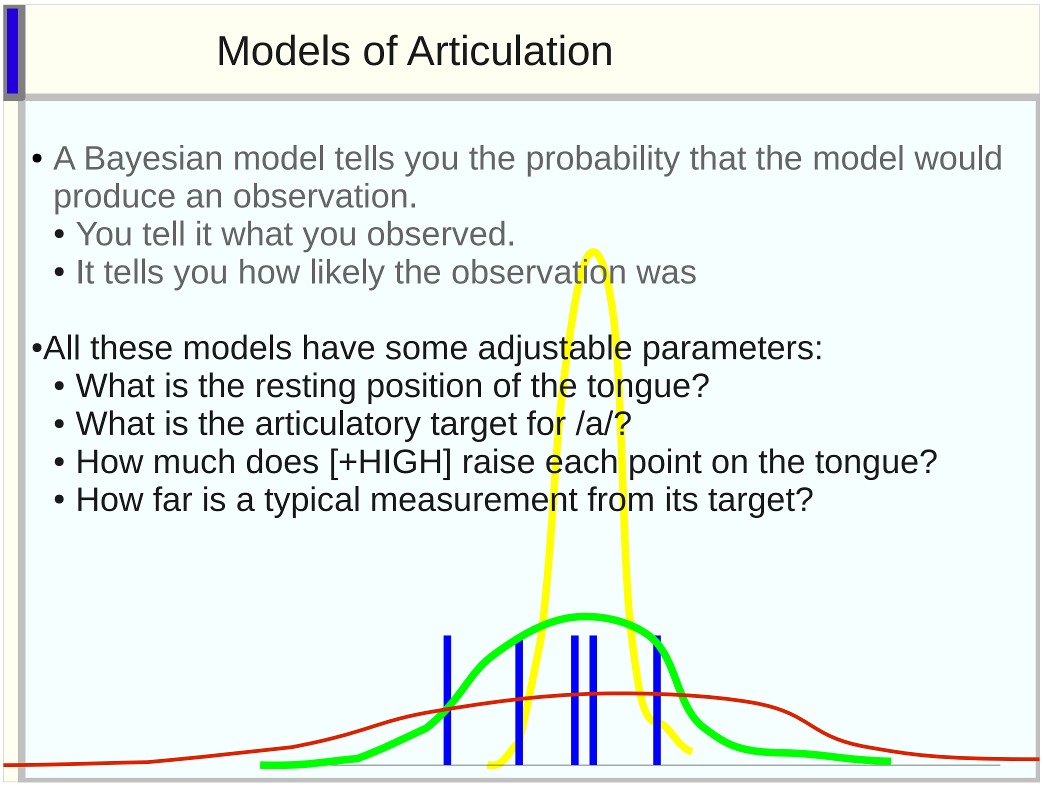
- A Bayesian model tells you the probability that the model would produce an observation.
  - You tell it what you observed.
  - It tells you how likely the observation was.

- A good model:
  - Assigns high probabilities to things that are actually observed.
  - Low probabilities to things that haven't been seen.

- This can be formalized as Bayesian Model Comparison.
  - It is analogous to the classical F-test and ANOVA.
  - It can compare models that aren't related.
    - (The F-test can only compare two models if one includes the other.)

Bayesian model fitting also matches the width of the probability distribution predicted by the model to the width of the data distribution. In the figure at the bottom, the green distribution has the highest overall probability of generating the data. While the yellow is very likely to predict some of the data, it will not generate the leftmost one.

On the other hand, the red distribution *can* generate the observed data, but since it is so broad, the probability of generating the actual (blue lines) data are somewhat smaller, compared to the green distribution. Recall that probability distributions integrate to one, so that a broader distribution is lower: because it allows the data to be generated in more places, they are less likely to occur in any particular place.



Models of Articulation

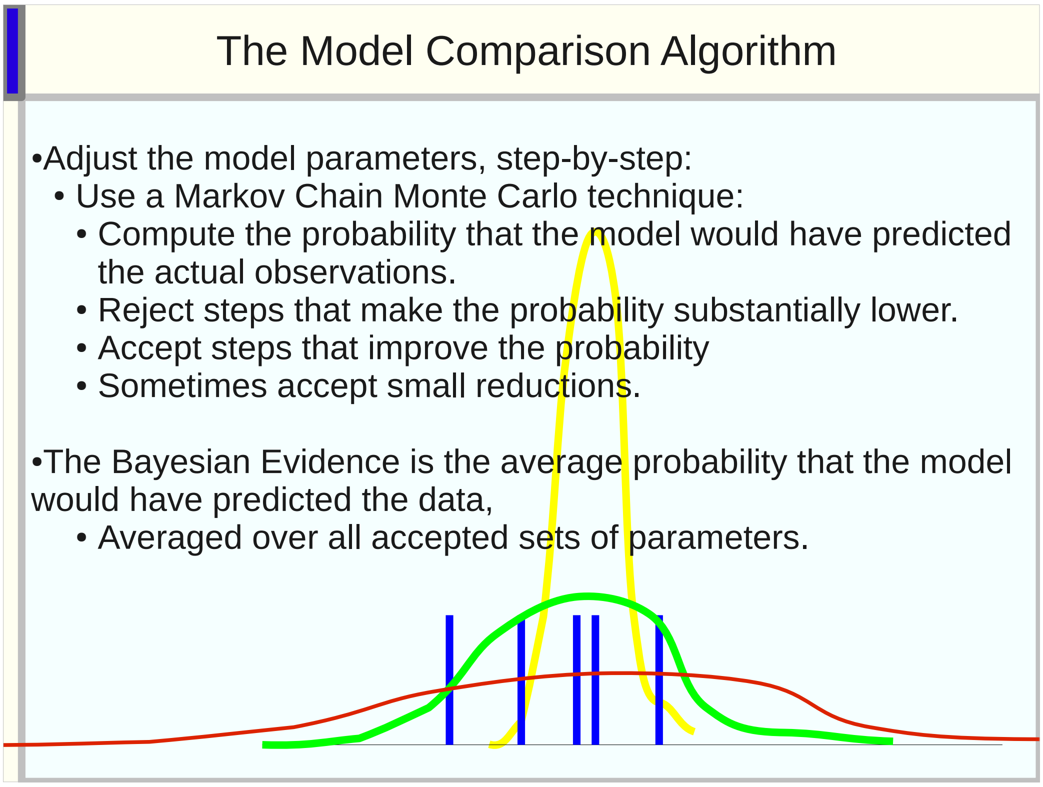
- A Bayesian model tells you the probability that the model would produce an observation.
  - You tell it what you observed.
  - It tells you how likely the observation was

•All these models have some adjustable parameters:
  - What is the resting position of the tongue?
  - What is the articulatory target for /a/?
  - How much does [+HIGH] raise each point on the tongue?
  - How far is a typical measurement from its target?

When we say “model” here, we really mean a set of equations (or, equivalently, an algorithm) that predicts a probability distribution for the data. A model is a more general thing than a hypothesis. A hypothesis is a particular probaability distribution for the data; a model can generate a family of hypotheses as one adjusts some parameters.

For instance:
•“The table is 1.35 meters long with an uncertainty of 0.02 meters” is a hypothesis.
•“The table is *X* meters long with an uncertainty of 0.02 meters” is a model, and *X* is a parameter of the model.

We call these “adjustable parameters” because one can adjust them to find which model(s) in the family are a good match to the available data. Adjusting a model's parameters is an example of Bayesian model comparison within a family of hypotheses (within a model), and one can also do similar comparisons between families.

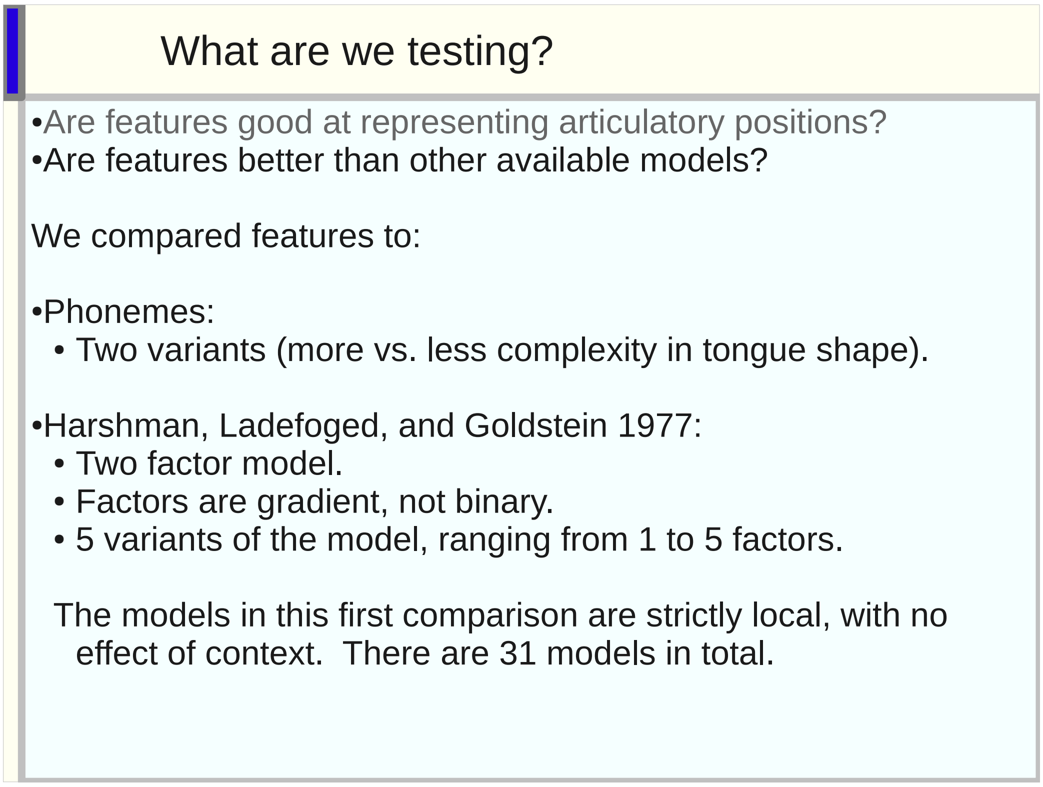


The Model Comparison Algorithm

- Adjust the model parameters, step-by-step:
  - Use a Markov Chain Monte Carlo technique:
    - Compute the probability that the model would have predicted the actual observations.
    - Reject steps that make the probability substantially lower.
    - Accept steps that improve the probability
    - Sometimes accept small reductions.

- The Bayesian Evidence is the average probability that the model would have predicted the data,
    - Averaged over all accepted sets of parameters.

If the modified parameters give a probability p for the model generating the data, and the last accepted probability is P, then accept the step with a probability A=min(1, p/P). This is the more precise description of the algorithm on the slide. This Markov Chain Monte Carlo algorithm will explore all the combinations of parameters that are reasonably consistent with the data. It will visit the most consistent ones most often.



What are we testing?

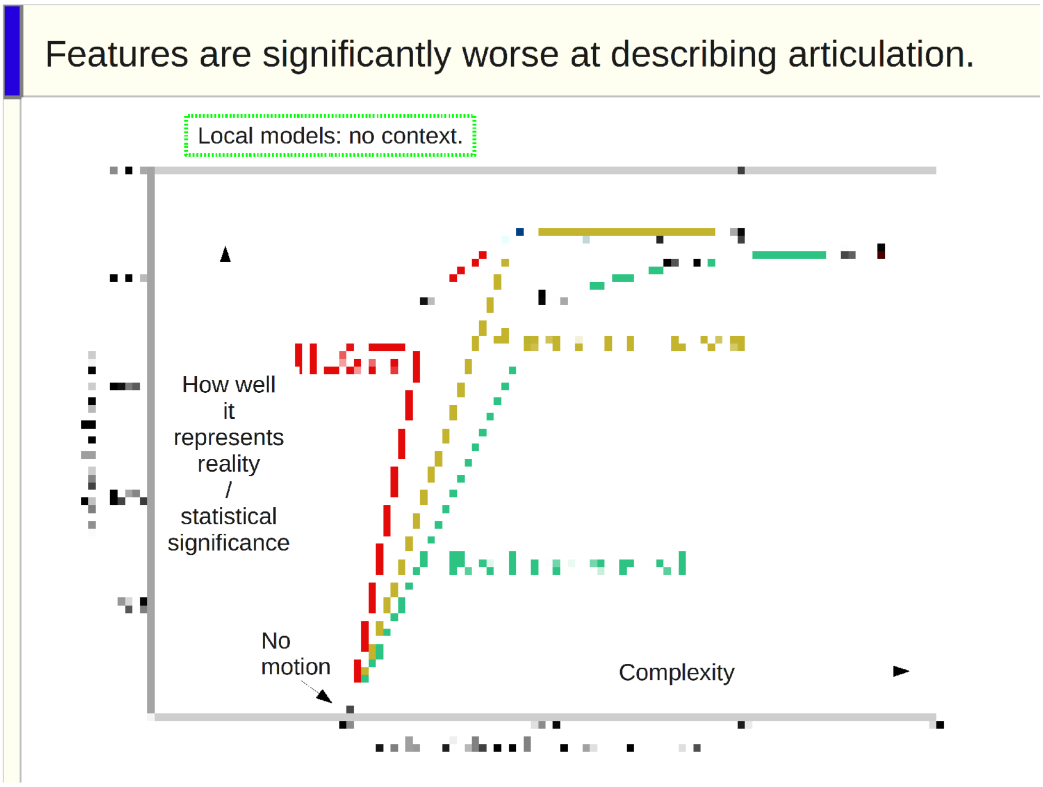
- Are features good at representing articulatory positions?
- Are features better than other available models?

We compared features to:

- Phonemes:
  - Two variants (more vs. less complexity in tongue shape).
- Harshman, Ladefoged, and Goldstein 1977:
  - Two factor model.
  - Factors are gradient, not binary.
  - 5 variants of the model, ranging from 1 to 5 factors.

The models in this first comparison are strictly local, with no effect of context. There are 31 models in total.

This first experiment intends to find out if there are any intrinsic advantages to using features for representing a single phoneme.



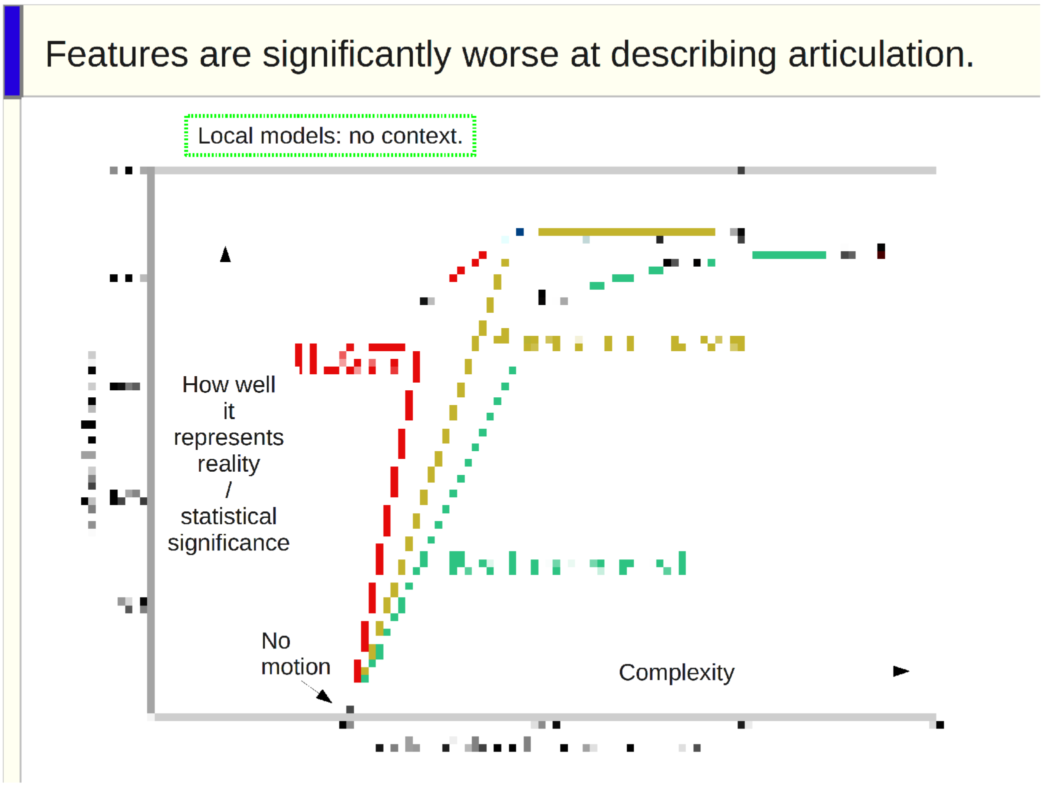
Features are significantly worse at describing articulation.

Local models: no context.

The vertical axis is the log of the Bayesian evidence: it is a measure of how effective each model can be at predicting the experimental data.

One can think of it as a measure of goodness of fit, but one that is averaged over the hypotheses generated by the model. So, a model that *can* make a good prediction but has many ways to make a mediocre prediction will not be rated as highly as a model that makes only good predictions.

The horizontal axis is the complexity of the model, measured in terms of the number of adjustable parameters it takes. Models with more parameters typically are more flexible and can be adjusted to match almost any data set. As a result, they tend to fit well, but they have very little predictive power. So, a simpler model is preferred over a complex one, if they have equal Bayesian evidence.



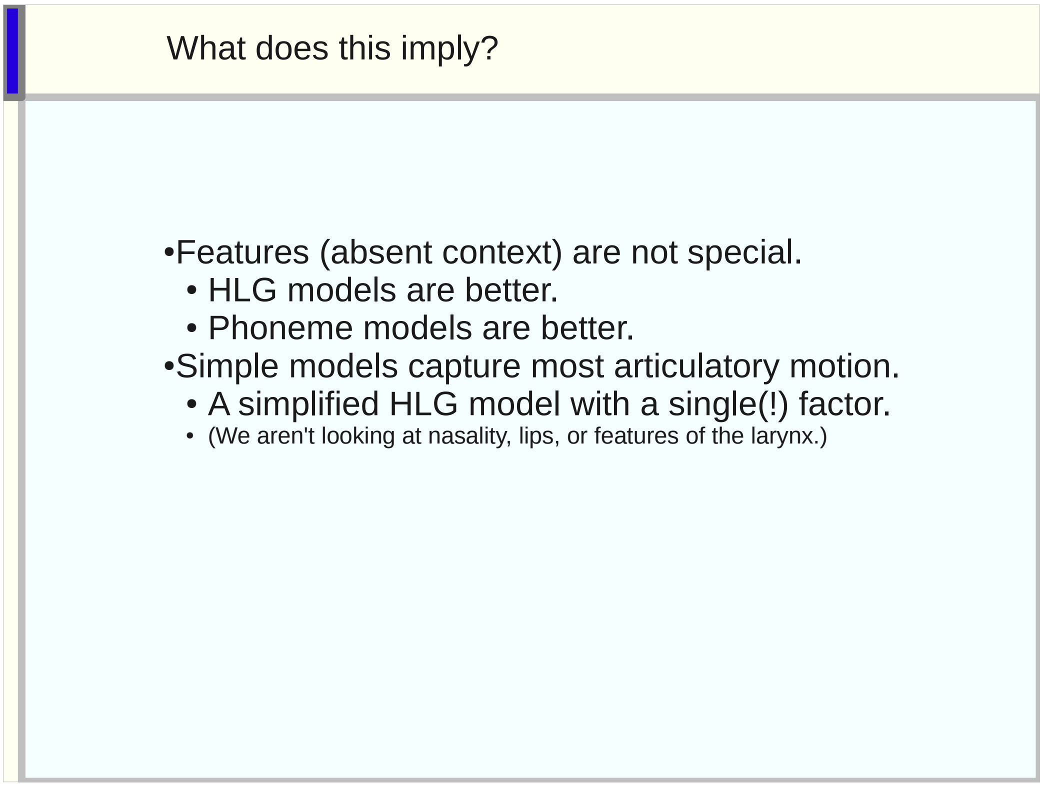
Features are significantly worse at describing articulation.

Local models: no context.

Here, we compare the three classes of models (HLG, Phoneme, and Feature) against each other. Note that there are several versions of each, of different complexity.

Each model is referenced against a model where the tongue doesn't move. In this minimal model, there is no difference between one speech sound and another. That minimal model has 50 parameters which describe the inter-subject differences in tongue position, along with the measurement errors (variances).

You can see that the feature-based models (green) have a substantially lower Bayesian evidence for equal complexities. The phoneme-based (brown) and HLG models (red) are comparably good, at least for the more-complex models. Note that the HLG models perform fairly well, even with a one-feature HLG model, a simplified version of HLG's idea.



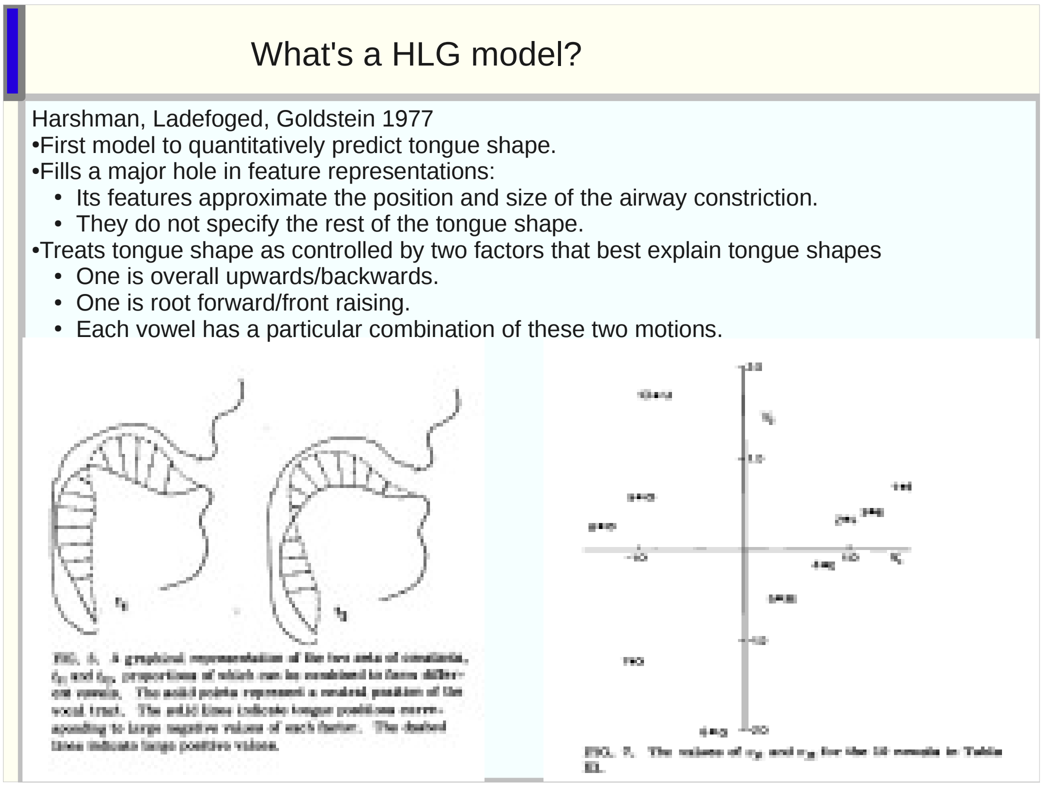
What does this imply?

- Features (absent context) are not special.
  - HLG models are better.
  - Phoneme models are better.
- Simple models capture most articulatory motion.
  - A simplified HLG model with a single(!) factor.
  - (We aren't looking at nasality, lips, or features of the larynx.)

It should be emphasized that we are not looking at all the articulators. We see the mid-section of the tongue and the back/root.   We don't see the tongue tip or the velum, which is perhaps why a single-factor HLG models does fairly well, and why the two-factor HLG model does as well as anything.

Features are not particularly good at describing articulatory positions. That is because these feature models assume that there is no interaction between features: for instance, we assume that +HIGH has the same effect whether or not +BACK is active. Presumably, this is not the case.

One would be tempted to argue that “But of course.   A true feature model would involve complicated interactions between features.” and that might be true, but such a model would no longer be simple once all the interactions are specified. For instance, adding interactions between 11 features adds 495 adjustable parameters. All the fully interacting feature models are far off the right-hand edge of the previous plot.



What's a HLG model?

Harshman, Ladefoged, Goldstein 1977

•First model to quantitatively predict tongue shape.

•Fills a major hole in feature representations:

- Its features approximate the position and size of the airway constriction.
- They do not specify the rest of the tongue shape.

•Treats tongue shape as controlled by two factors that best explain tongue shapes

- One is overall upwards/backwards.
- One is root forward/front raising.
- Each vowel has a particular combination of these two motions.

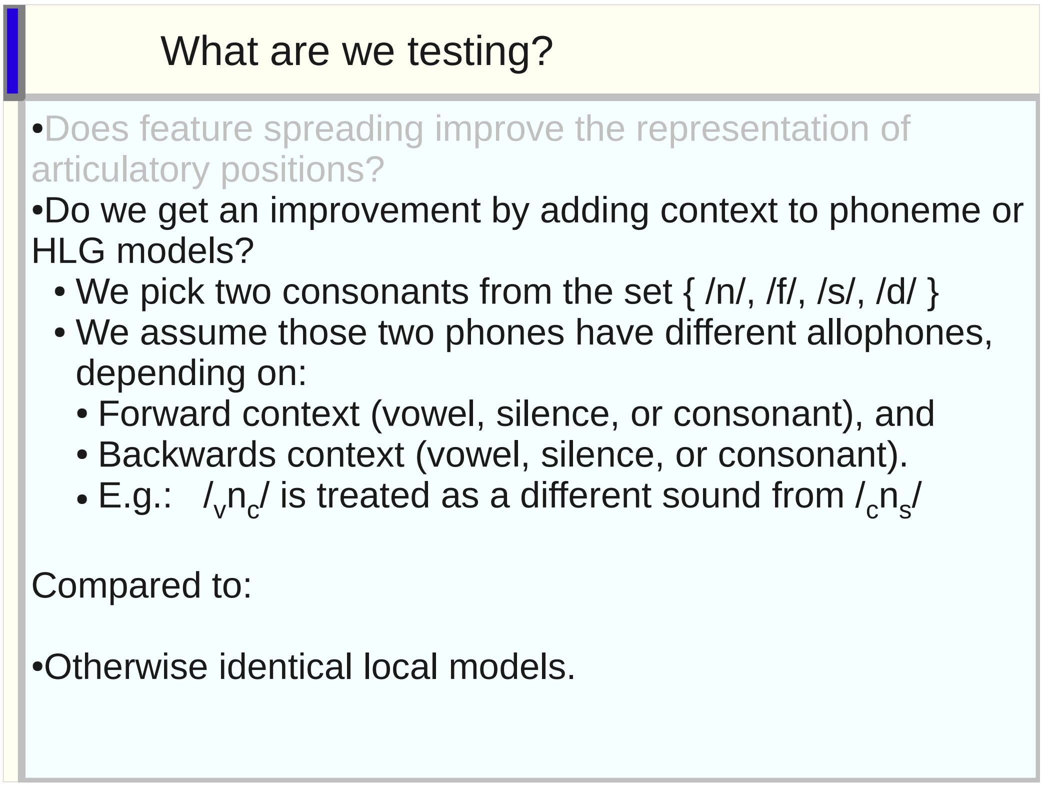
FIG. 5. A graphical representation of the two axes of constriction, t_I and t_{II}, proportions of which can be combined to form different vowels. The solid points represent a neutral position of the vocal tract. The solid lines indicate tongue positions corresponding to large negative values of each factor. The dashed lines indicate large positive values.

FIG. 7. The values of c_I and c_{II} for the 10 vowels in Table III.

The original HLG model, by definition, is the best two-factor additive model of the airway width. The factors, though, are not the binary values that one expects in linguistics: they can take any real-number value. The plot on the right shows the value of the features in the original HLG paper.

Because HLG is defined procedurally, the factor values and the tongue shapes will vary from data set to data set. Indeed ours are somewhat different from those of HLG. Note that HLG defined their model over all English vowels, and we have defined ours to cover the set of phonemes used in out data (5 vowels and 5 consonants).

We generated the 1-, 3-, 4-, and 5-factor equivalents of the HLG model.



What are we testing?

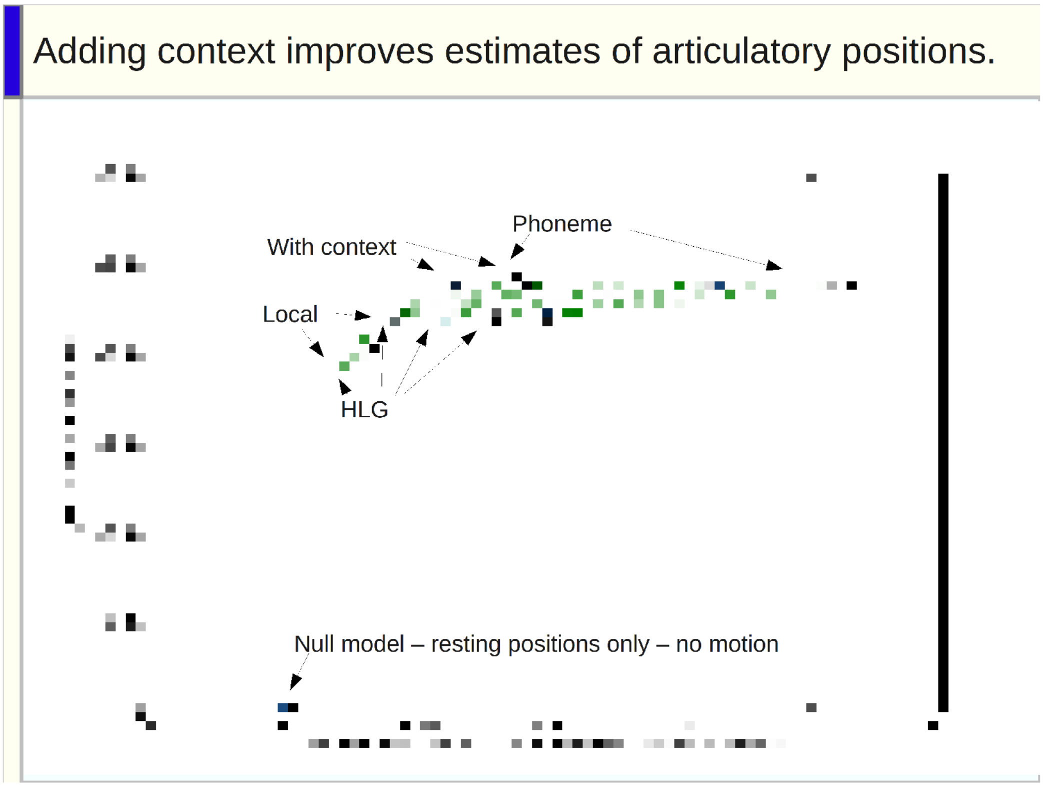
- Does feature spreading improve the representation of articulatory positions?
- Do we get an improvement by adding context to phoneme or HLG models?
  - We pick two consonants from the set { /n/, /f/, /s/, /d/ }
  - We assume those two phones have different allophones, depending on:
    - Forward context (vowel, silence, or consonant), and
    - Backwards context (vowel, silence, or consonant).
    - E.g.: $/_{v}n_{c}/$ is treated as a different sound from $/_{c}n_{s}/$

Compared to:

- Otherwise identical local models.

This second experiment is designed to find out for how much improvement could be expected by introducing some context into the models. This will be a “default”, relatively uninspired way to include context. Then, in the next experiment, we shall see if adding context to features (via partial specification) is more or less effective

So, we will add context to two consonants at a time, and see how much that improves the ability of the model to predict the data. The context breaks each of the consonants into six allophones that differ in terms of what phone is to their left or their right.

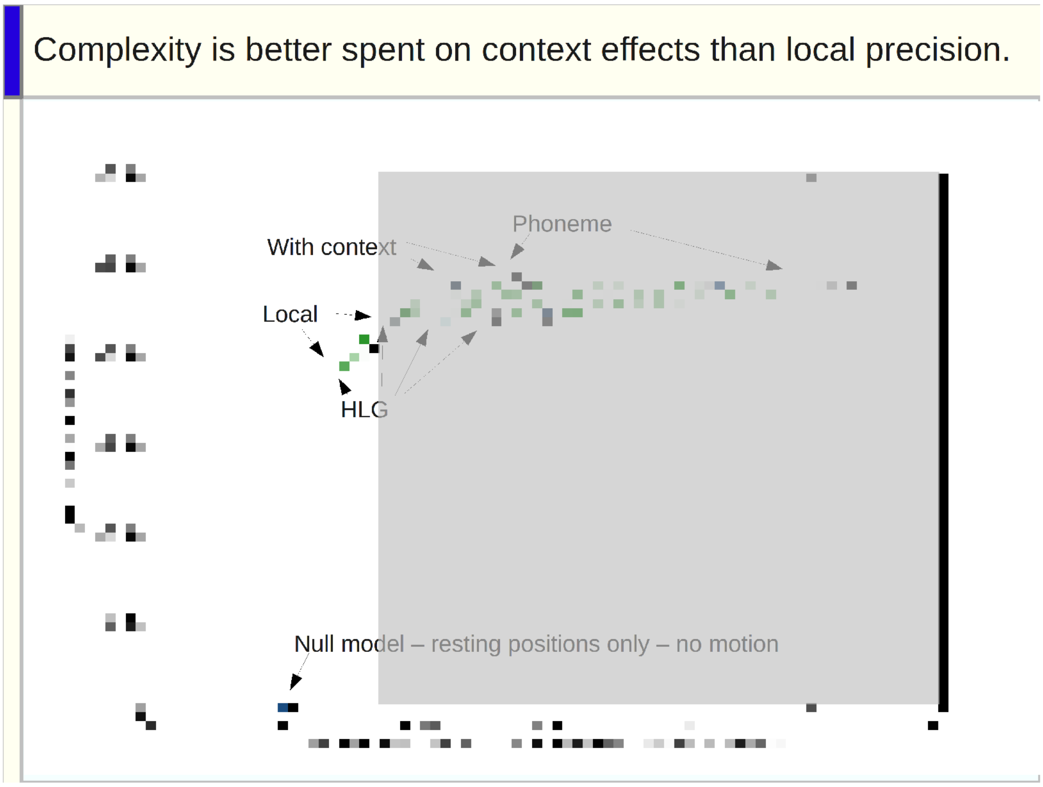


Adding context improves estimates of articulatory positions.

This figure shows the changes in the Bayesian evidence (vertical) and changes the complexity of the models (horizontal) as we add context. Plotted as before, except that corresponding models with and without context are joined by dashed green lines.

Each local model (down and left) has six variants: we add context to two consonants out of four. (The six variants differ in their complexity because some consonants are found in more different contexts than others.)

One can see some modest improvements to the Bayesian evidence by adding context.

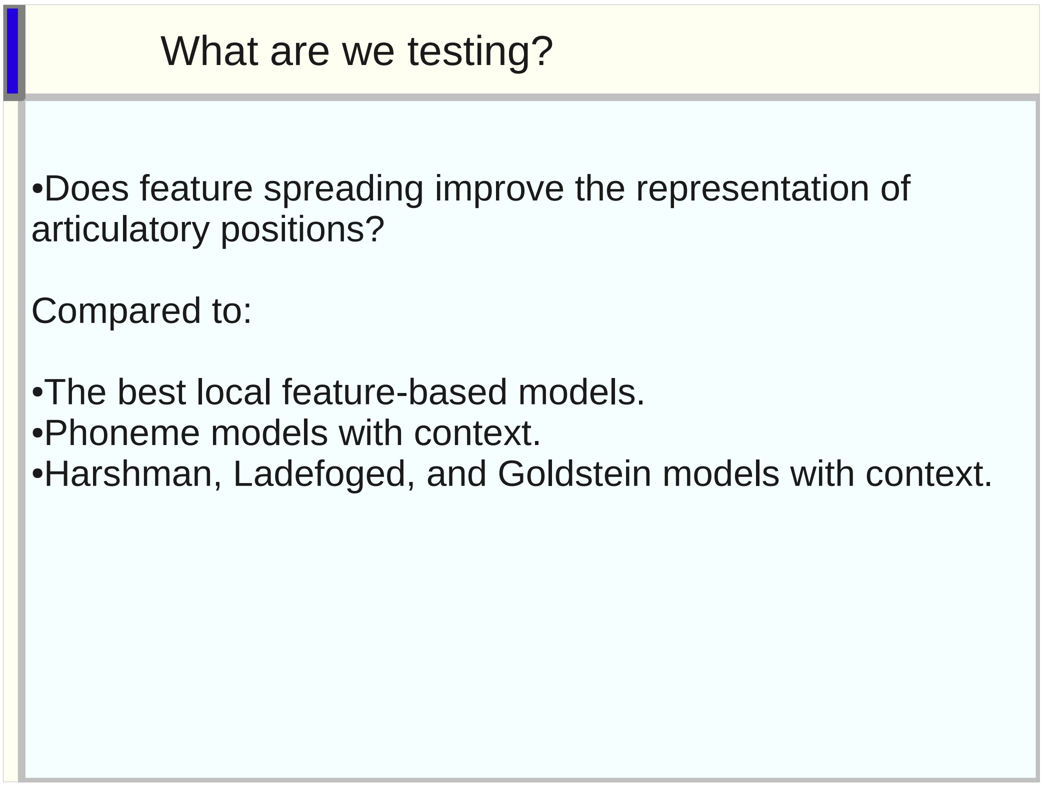


Complexity is better spent on context effects than local precision.

Now, focus on the area where the models are about all equally good (the shaded region). One can see that the improvement by context is typically much larger than the difference between phoneme models or the difference among variants of the HLG model.

What does this imply?

- Features (absent context) are not special.
- Simple models capture most articulatory motion.
  - Five features.
  - A simplified HLG model with a single(!) factor.
  - (We aren't looking at nasality, lips, or features of the larynx.)
- It is better to add context effects than elaborate a description that doesn't use context.
  - (Except for the simplest descriptions.)



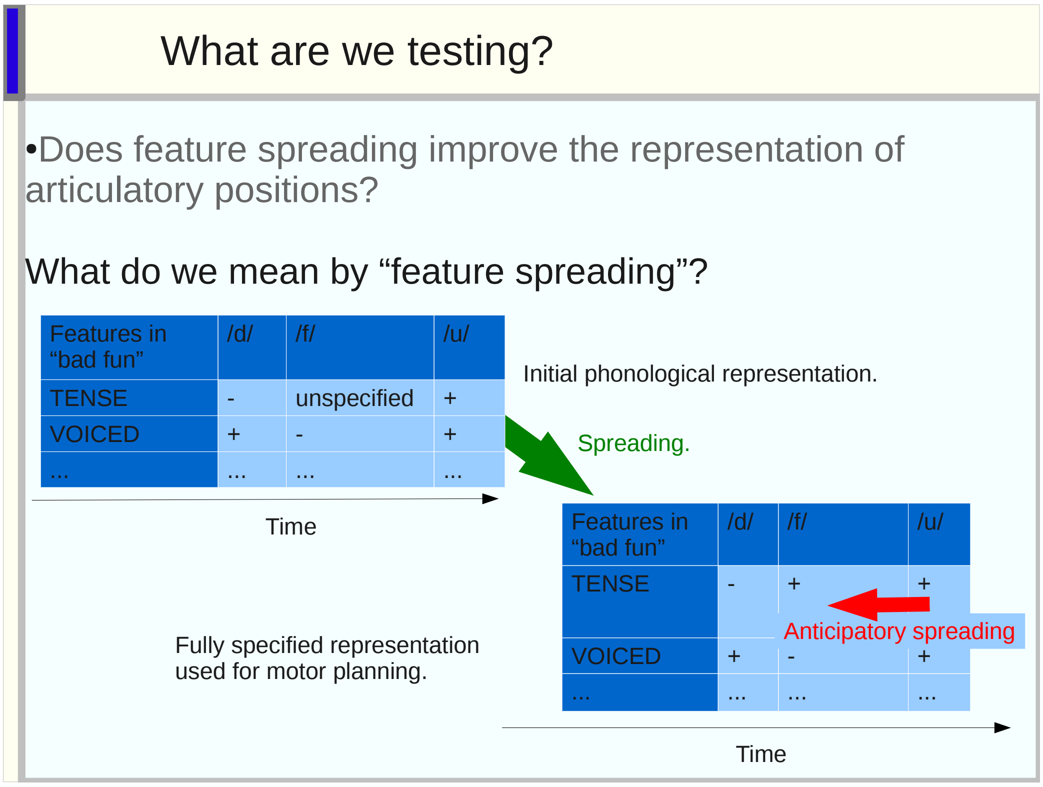
What are we testing?

- Does feature spreading improve the representation of articulatory positions?

Compared to:

- The best local feature-based models.
- Phoneme models with context.
- Harshman, Ladefoged, and Goldstein models with context.

Finally, we investigate feature spreading. First, to see if it can improve the match to the data, and if so, to see if it does a better job than the “default” context of our second experiment.



What are we testing?

- Does feature spreading improve the representation of articulatory positions?

What do we mean by “feature spreading”?

Features in “bad fun”	/d/	/f/	/u/
TENSE	-	unspecified	+
VOICED	+	-	+
...	...	...	...

Initial phonological representation.

Time

Spreading.

Fully specified representation used for motor planning.

Features in “bad fun”	/d/	/f/	/u/
TENSE	-	+	+
VOICED	+	-	+
...	...	...	...

Anticipatory spreading

Time

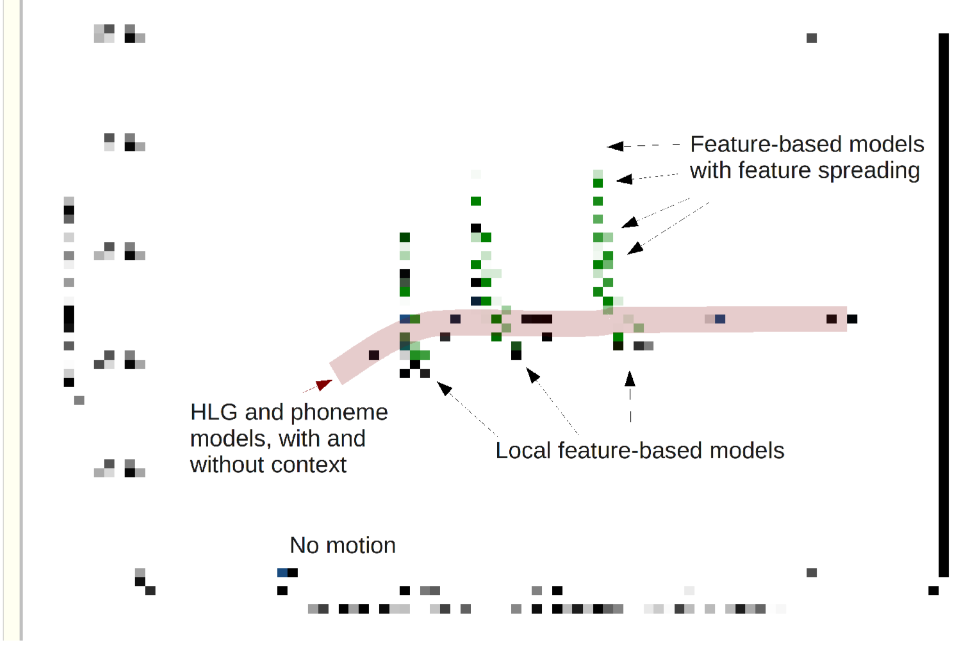
By “spreading” I mean the process by which an unspecified feature acquires a value.   For instance, the Advanced Tongue Root feature is unspecified for the /f/ sound in Chomsky's *Sound Patterns of English*. Acoustically, the tongue root position may not matter for that sound, but the tongue root is certainly somewhere at all times, and with our MRI system, we can measure where it is. (The example above shows “tense” rather than ATR to simplify the display.)

A reasonable hypothesis is that the motor control system makes use of the fact that the tongue root positions does not matter for /f/, and uses the time interval of the /f/ to move the tongue root from wherever it needed to be for the previous phone to wherever it will need to be for the next phone.

In this case, a specified feature is one that is necessary to generate the correct sound so that the listener can understand what is being said. In contrast, an unspecified feature is one that is unimportant for intelligibility and this provides freedom for the motor control system to optimize the articulatory trajectory. It might then try to minimize effort my making a relatively slow move from one position to the next (c.f. Kochanski, Shih, Jing2005).

This example shows anticipatory spreading, where the motor control system concentrates on getting ready for the next phone. Other reasonable possibilities might be to hold on to the previous position, and not move until necessary, or some mixture of the two.

Context via feature spreading: dramatic improvement.

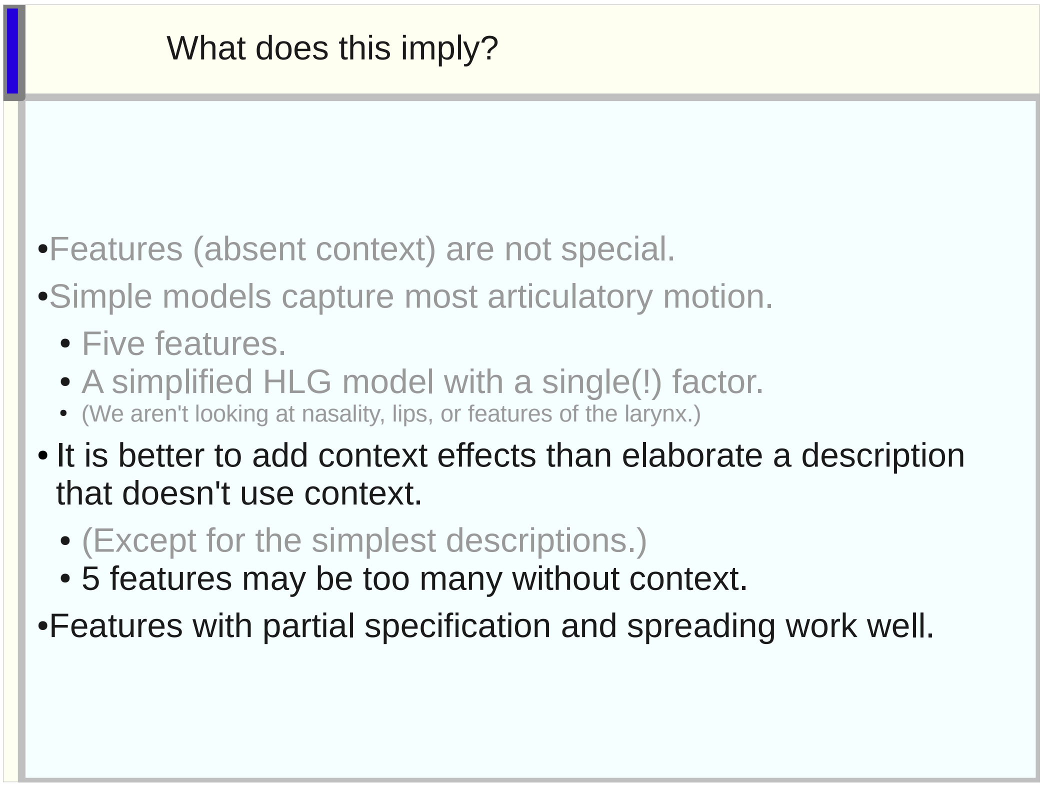


In this experiment, we introduce context via feature spreading and see how it affects the Bayesian evidence (vertical axis) and the complexity of the models (horizontal axis). The pink band contains the comparison models from the previous plot: these are all the phoneme-based and HLG variants both with and without context.

The dashed green lines join corresponding pairs of models, with and without feature spreading. The models with spreading have far higher Bayesian evidence: they do a much better job of predicting the observed data.

Interestingly, the models with spreading also have fewer parameters. That is because in the local models, without spreading, still need to fill that slot. Since they are not filled with a 0(-) or 1(+) from the phonology, we fill them with whatever value gives the best fit to the data: in other words, they are assigned an adjustable parameter. (Incidentally, this adjustable parameter can take on fractional values, so our local models will provide a better match to the data than if the unspecified slots were filled with either 0 or 1.)

However, in the models with spreading, we no longer need to allocate an adjustable parameter to filling in the slot, since it is filled in from the neighbouring phones.

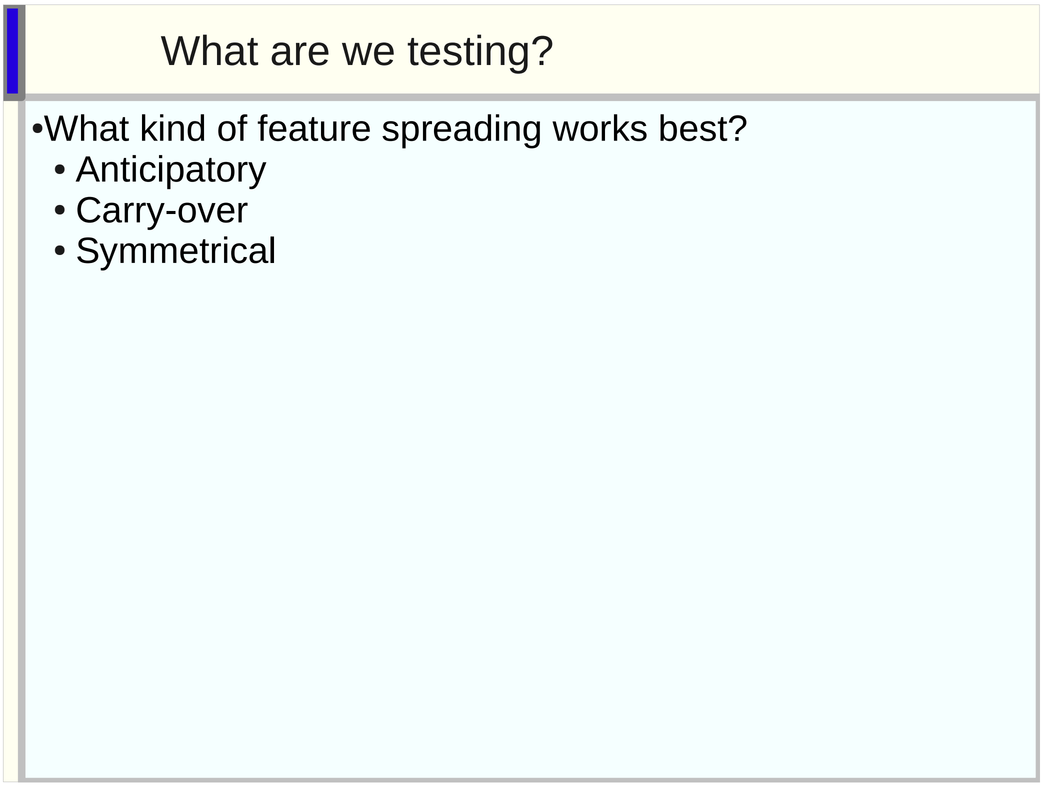


What does this imply?

- Features (absent context) are not special.
- Simple models capture most articulatory motion.
  - Five features.
  - A simplified HLG model with a single(!) factor.
  - (We aren't looking at nasality, lips, or features of the larynx.)
- It is better to add context effects than elaborate a description that doesn't use context.
  - (Except for the simplest descriptions.)
  - 5 features may be too many without context.
- Features with partial specification and spreading work well.

Although features were unimpressive in local models, doing a rather poor job, they did remarkably well when we included context effects via feature spreading.

And, once again we see that context effects are much more important than elaborating a model. The simplest feature-based model with context (5 features) does far better than the 13-feature model without.

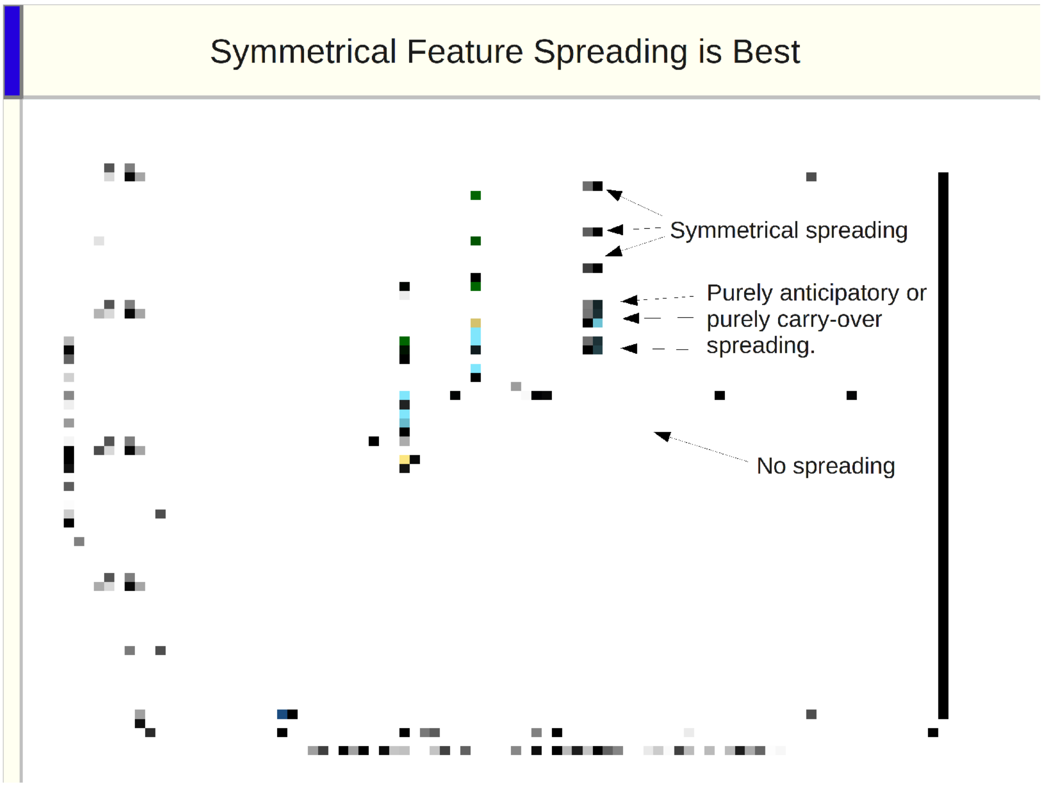


What are we testing?

- What kind of feature spreading works best?
  - Anticipatory
  - Carry-over
  - Symmetrical

Let's look in more detail to see what kind of feature spreading is most effective. The literature has arguments that anticipatory spreading is dominant, and also that carry-over is dominant. There is a long-standing disagreement that we can resolve. (Anticipatory means that the feature is filled from the next phone; carry-over means it is filled from the previous phone.)

We also tried symmetric spreading just for completeness: in symmetric spreading, a feature is filled with the average of the corresponding features on both the left and right sides.



Symmetrical Feature Spreading is Best

This plot shows the same models as the previous plot, but it labels them with the type of feature spreading.

It can be seen that the models with symmetrical feature spreading are substantially better than either anticipatory or carry-over spreading. This difference is dramatic: they are far and away the best models.

This leads to an interesting challenge to conventional phonological theory: the symmetrical spreading process fills a feature with three, not two values. These three values correspond to cases where the neighbouring features are both positive (+?+) it yields +; where both are negative (-?-) it yields -, and where they disagree (+?- or -?+) it yields half-way in between.

Incidentally, it's not shown here, but anticipatory and carry-over spreading are about equally good, which may be why there is a long-standing debate over which is better.

What does this imply?

- Features (absent context) are not special.
- Simple models capture most articulatory motion.
  - Five features.
  - A simplified HLG model with a single(!) factor.
  - (We aren't looking at nasality, lips, or features of the larynx.)
- It is better to add context effects than elaborate a description that doesn't use context.
  - (Except for the simplest descriptions.)
  - 5 features may be too many without context.
- Features with partial specification and spreading work well.
- Spreading is symmetrical
  - We fill unspecified features with [-,0,+]
  - Non-binary feature filling can work.

Conclusions

If you're committed to features:

- Recognize that there's a chance something better might come along.
- Partial specification / spreading is important.
  - It may work better with symmetrical, 3-state filling.
- This group of features is important:
- HIGH, LOW, CORONAL, BACK
- Some seem unimportant compared to context:
- CONTINUANT, ANTERIOR, TENSE, ATR
- Specifying a few features with context is better than all without.

If you're not committed to features:

- Think of a way to add context effects to HLG models
- Think of a way to do phonology with HLG-like models.