Modeling phonetic category learning from natural acoustic data

Stephanie Antetomaso[1], Kouki Miyazawa[2,4], Naomi Feldman[3], Micha Elsner[1], Kasia Hitczenko[3], Reiko Mazuka[4]

[1]The Ohio State University
[2]Fairy Devices Inc.
[3]University of Maryland
[4]RIKEN Brain Science Institute

Input to infants contains variability

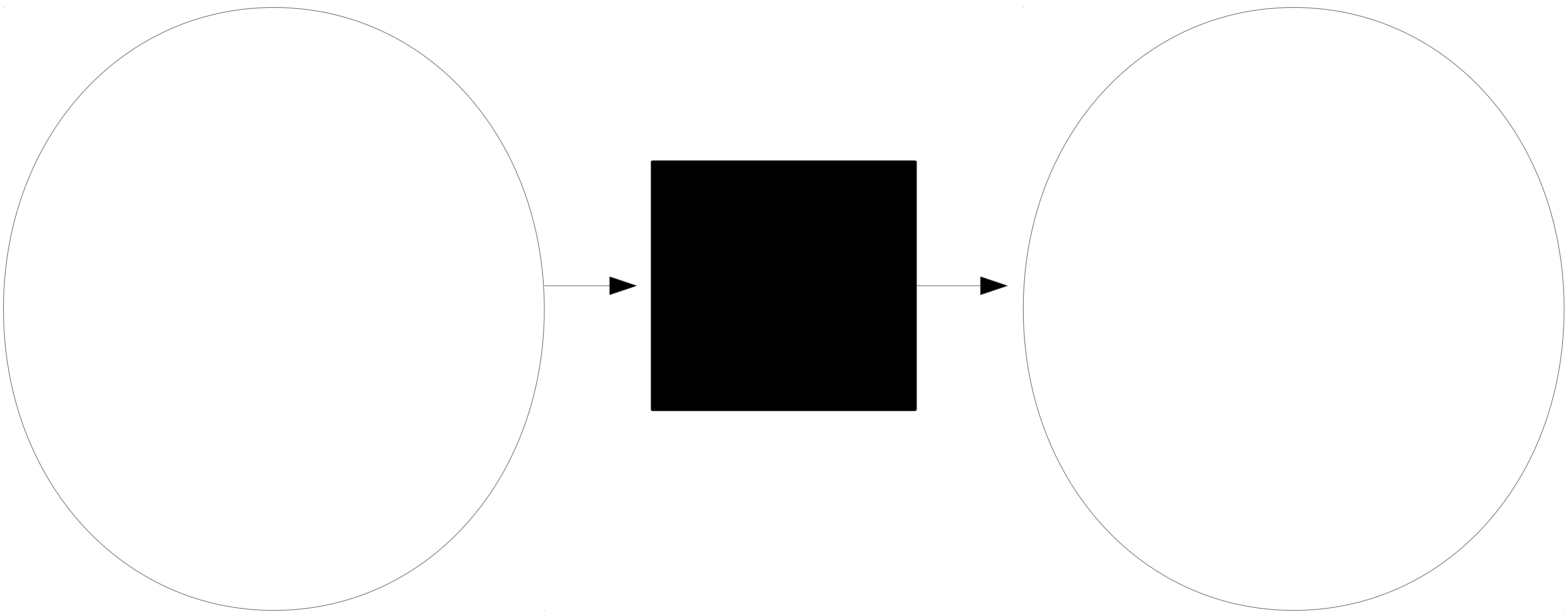
Input to infants contains variability

- Yet children acquire phonetic categories of their native language within the first year (Werker & Tees 1984, Polka & Werker 1994)

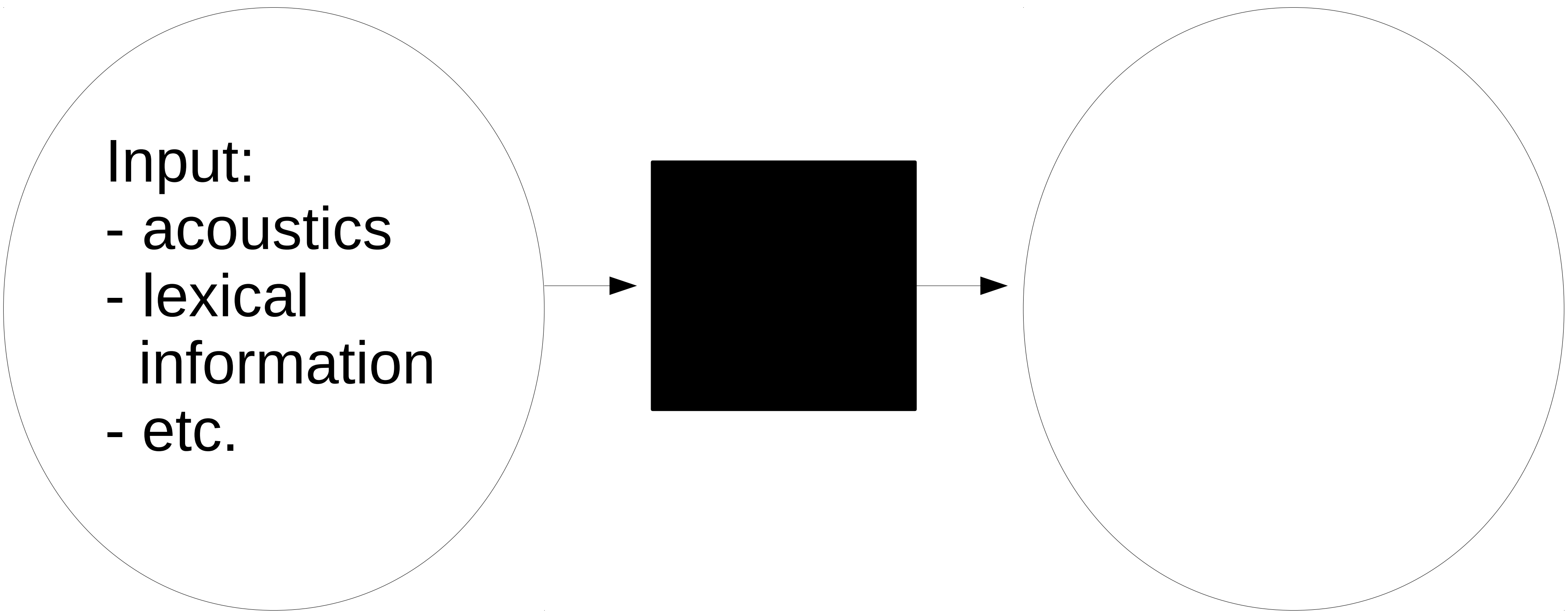
Input to infants contains variability

- Yet children acquire phonetic categories of their native language within the first year (Werker & Tees 1984, Polka & Werker 1994)
- Variability is critical for certain types of language learning (Gomez 2002, Rost & McMurray 2009)

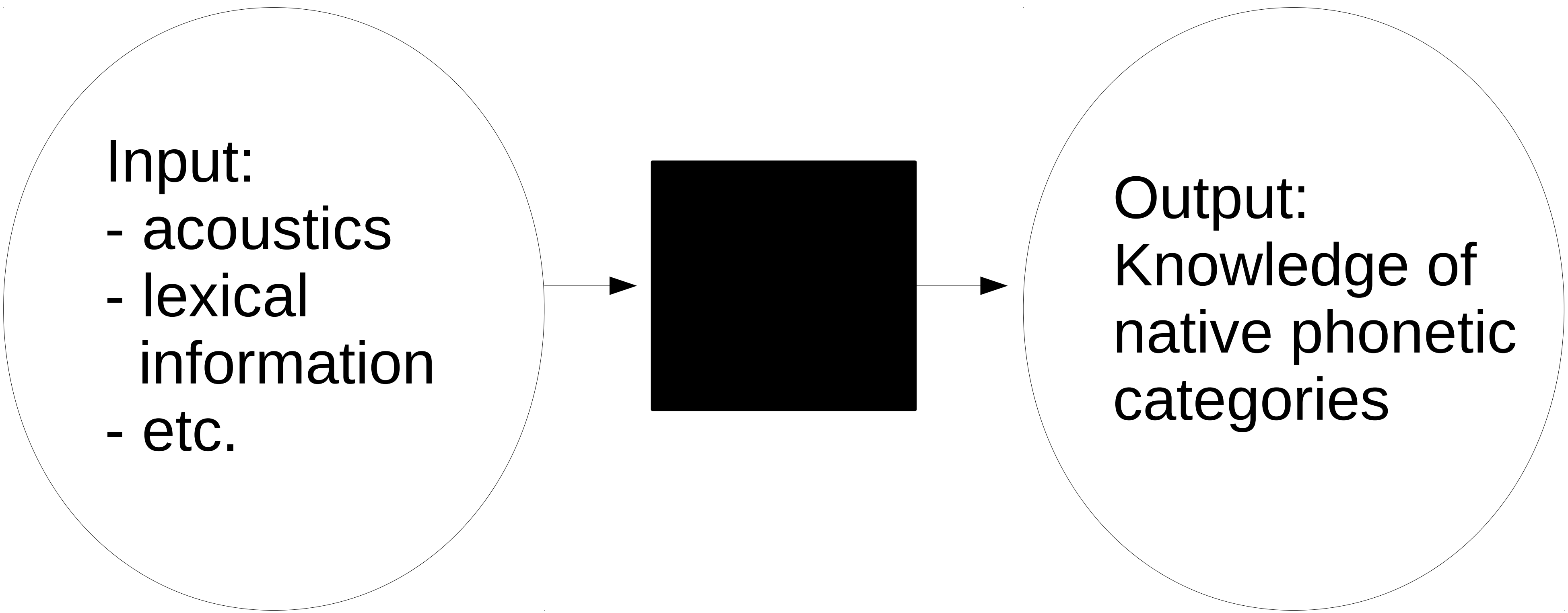
Learning problem



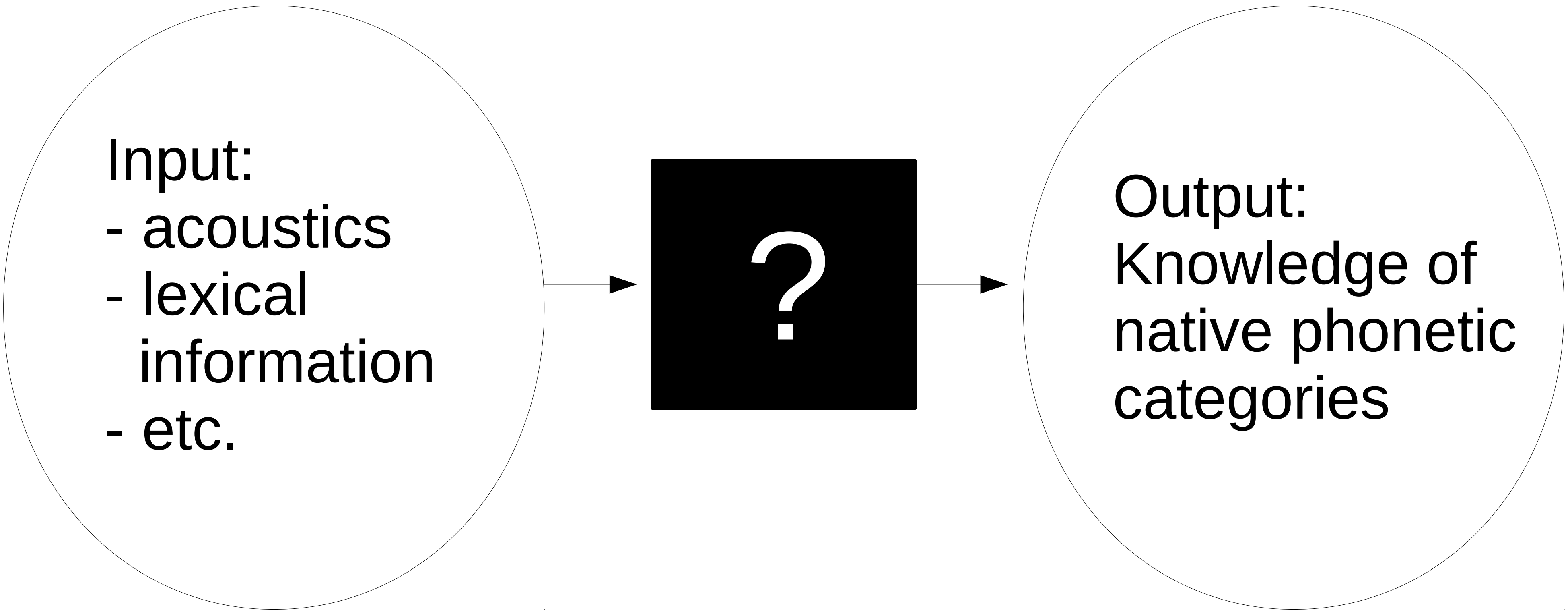
Learning problem



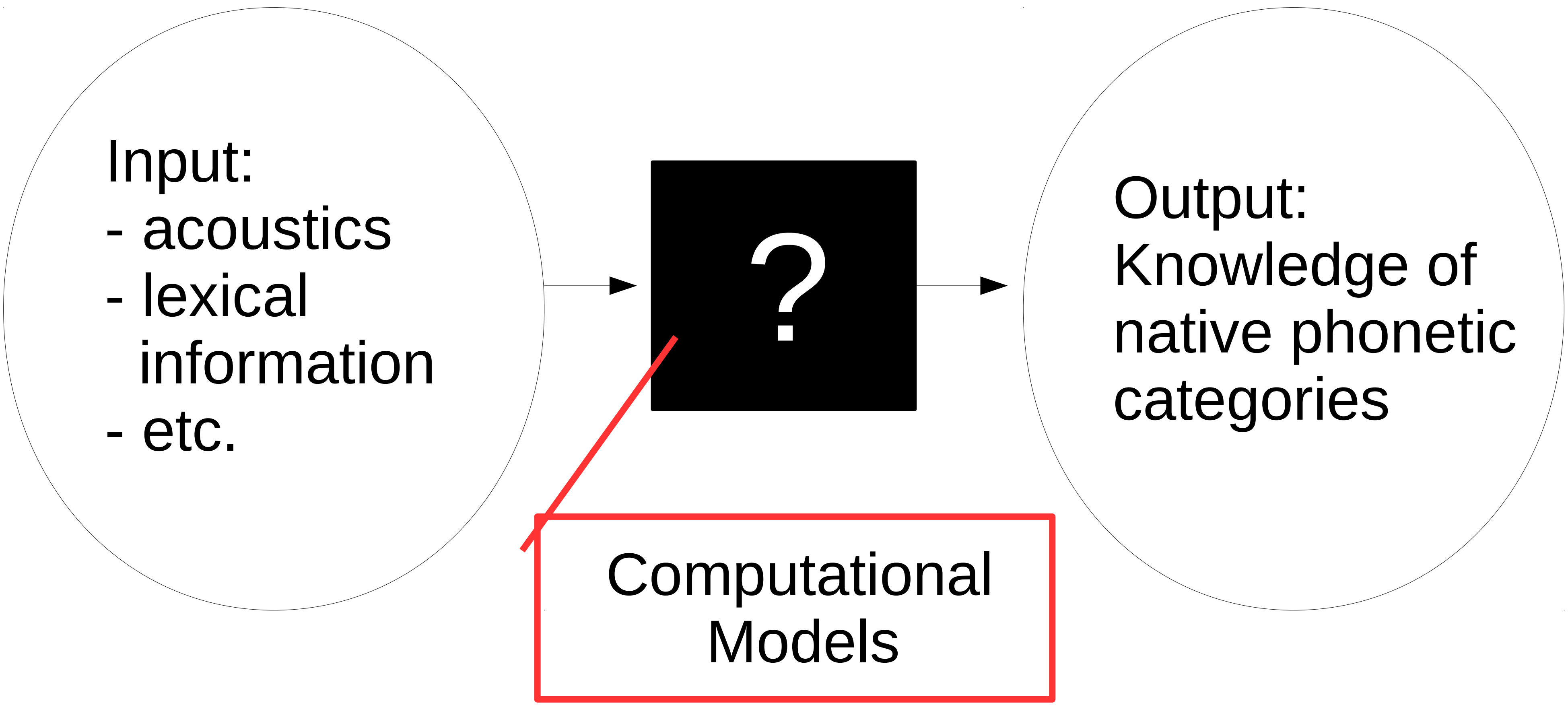
Learning problem



Learning problem

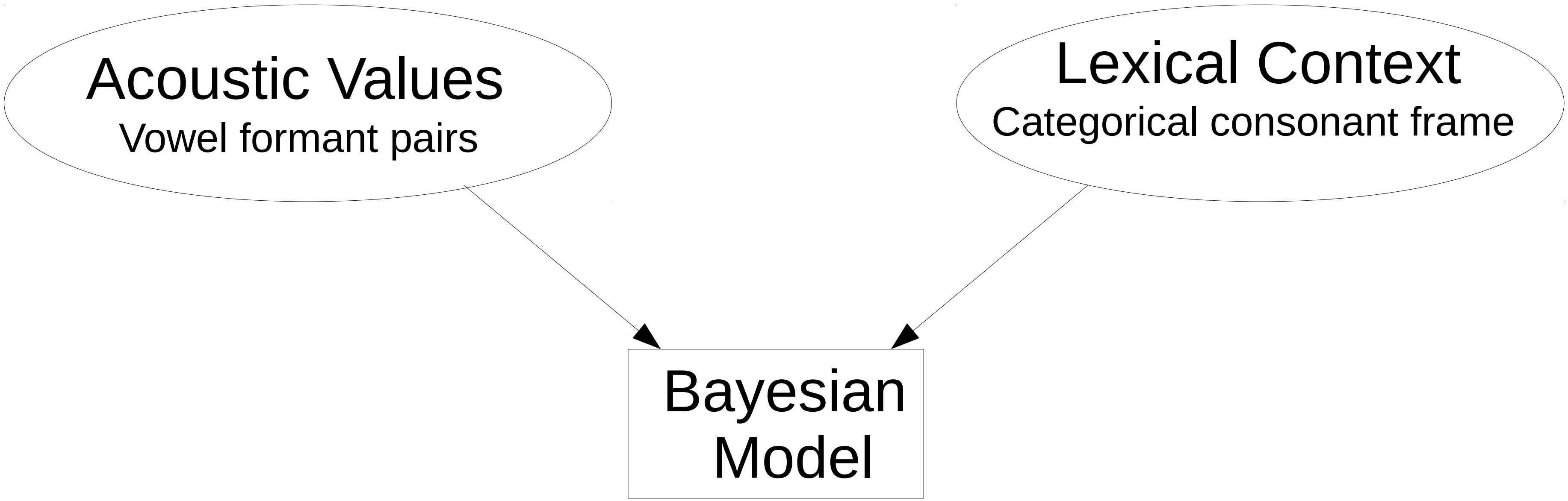


Learning problem



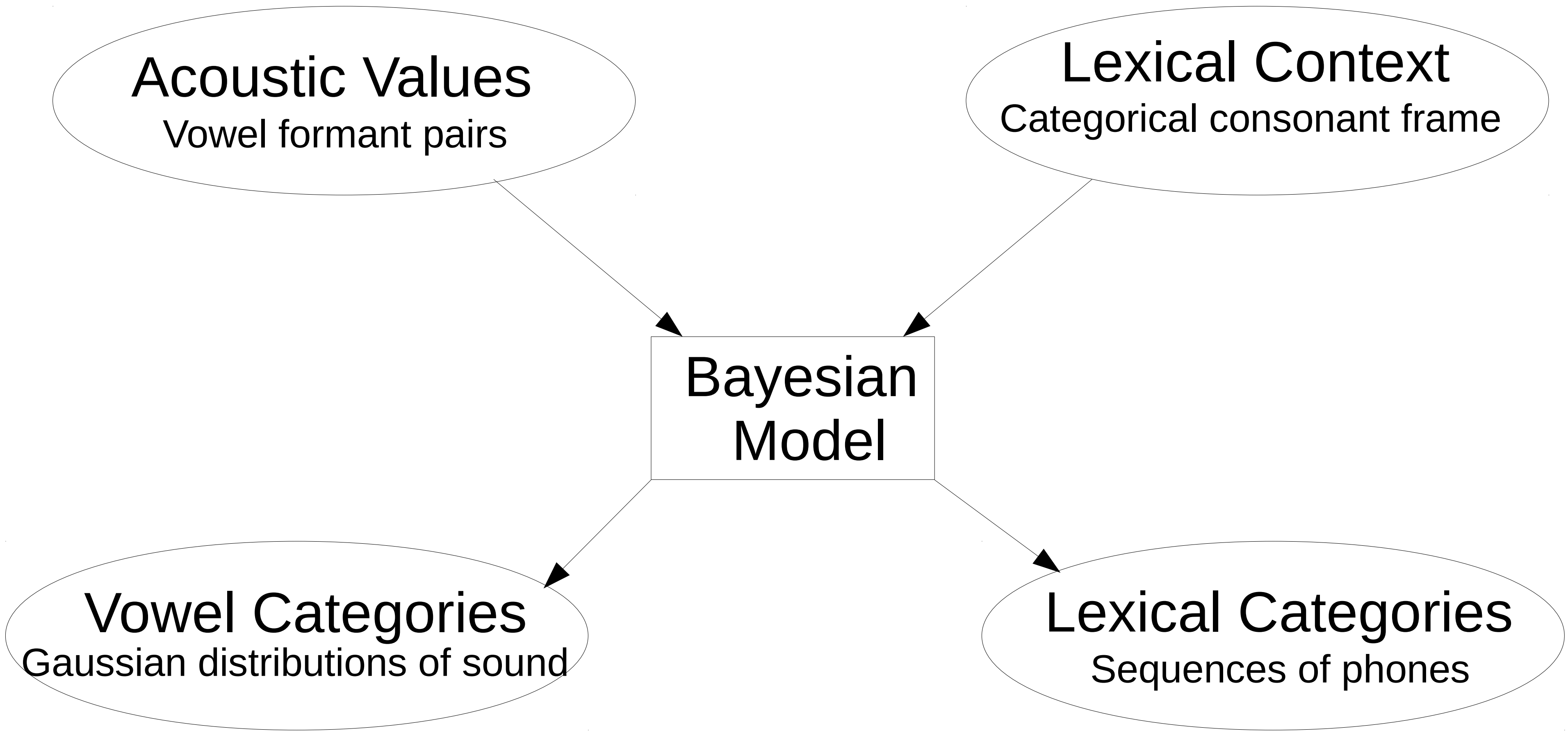
e.g. Vallabha et al. (2007), McMurray et al. (2009), Feldman et al. (2013)

Bayesian lexical-distributional clustering model

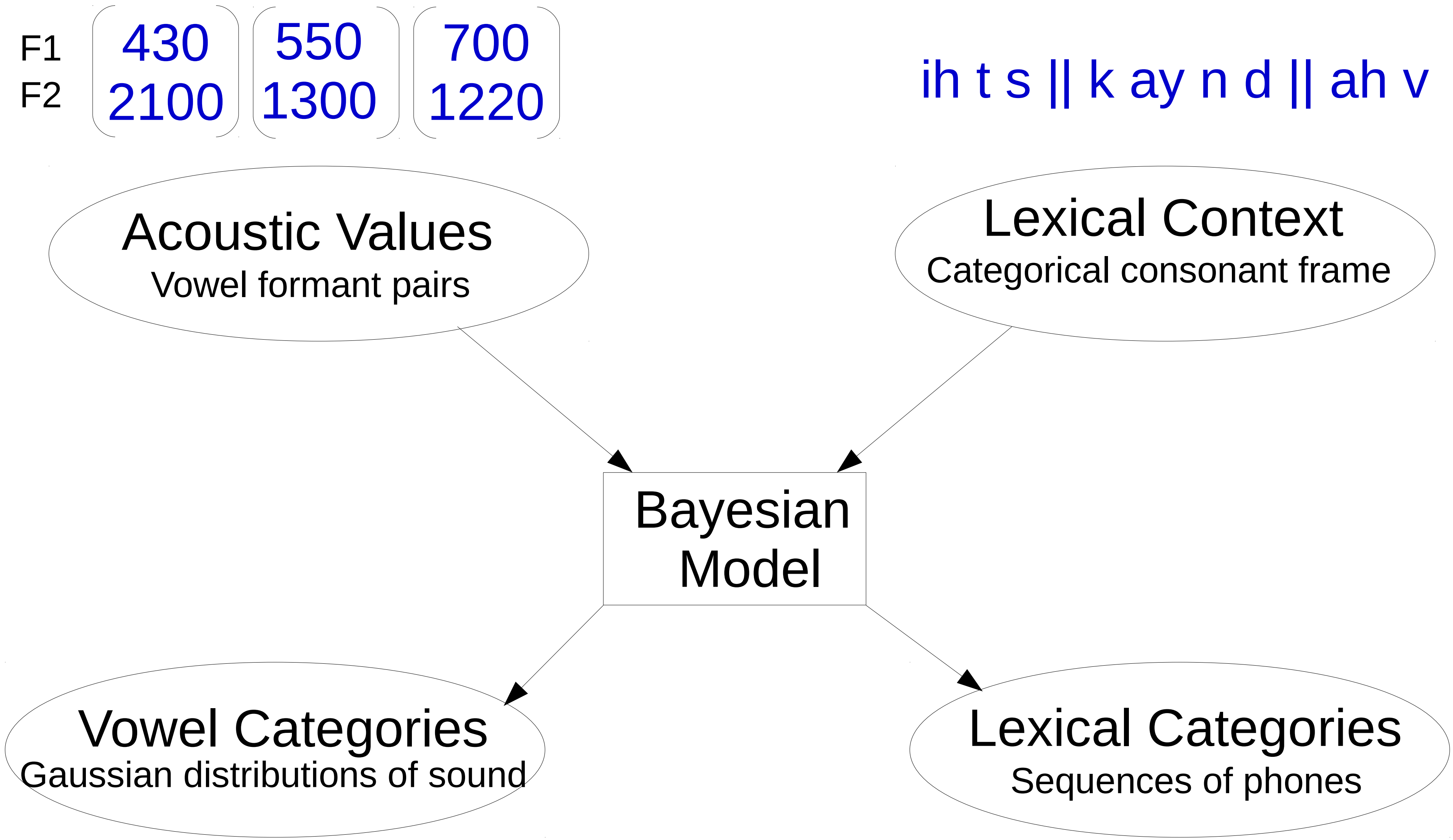


Feldman et al. 2013

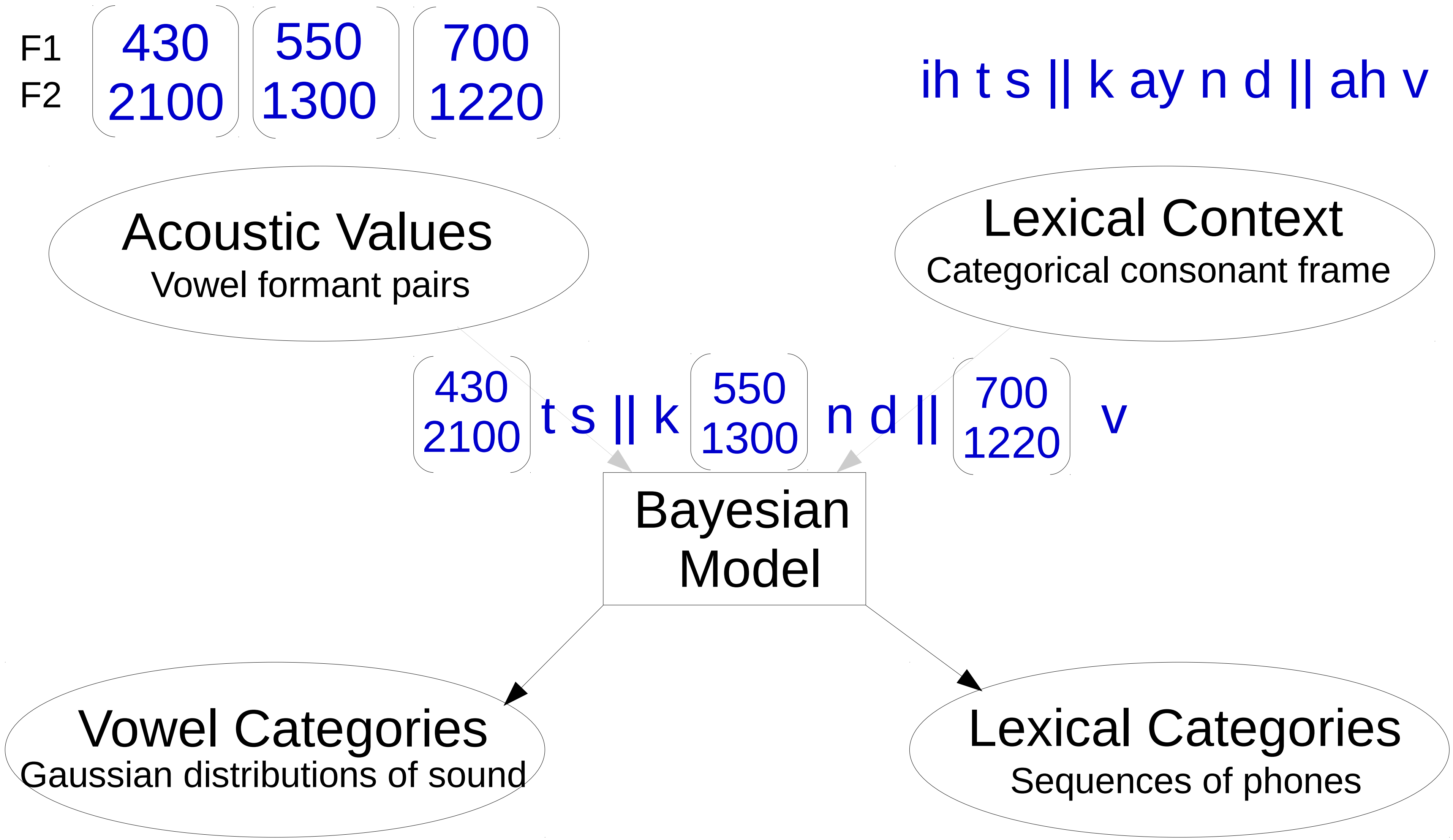
Bayesian lexical-distributional clustering model



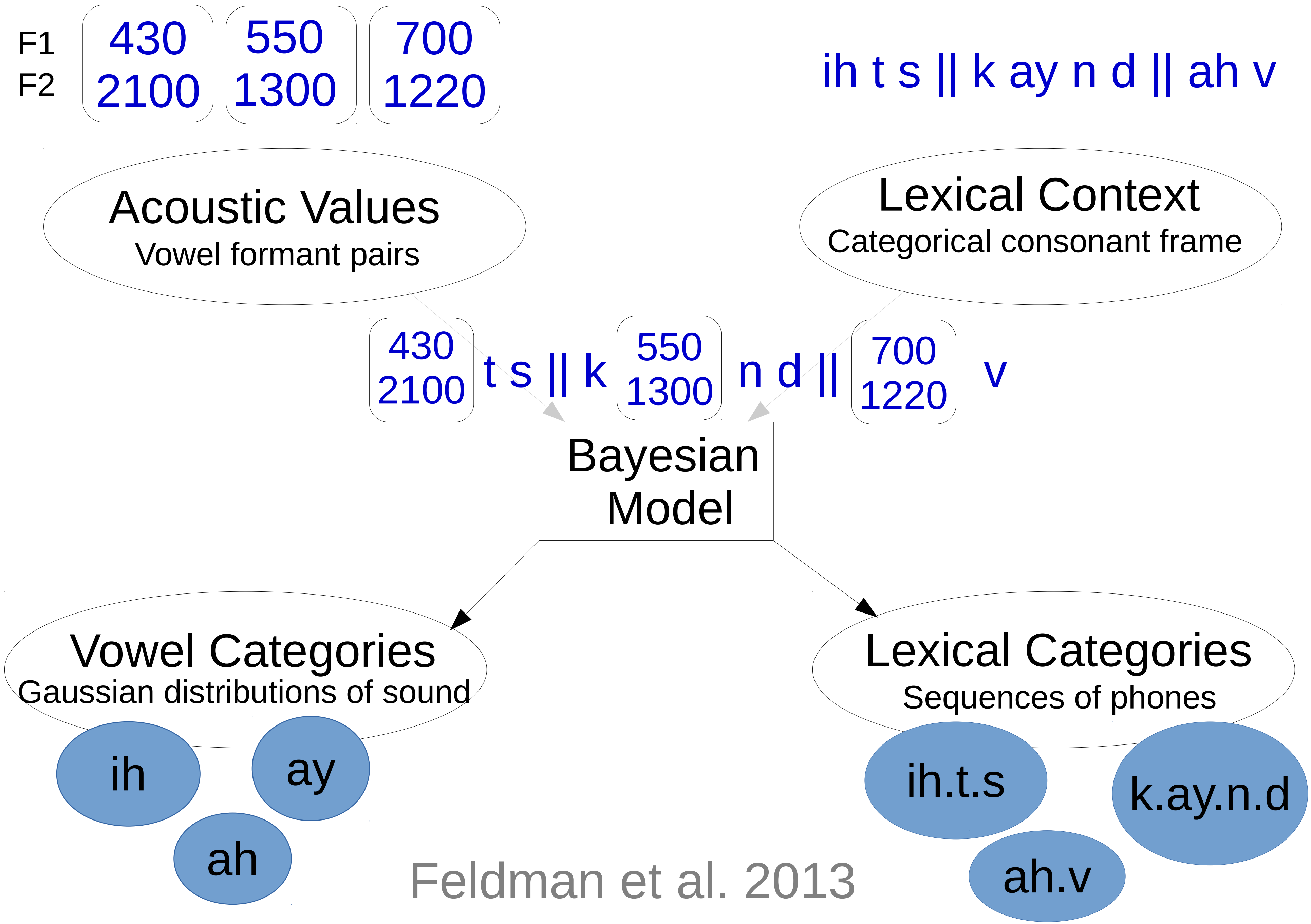
Feldman et al. 2013



Feldman et al. 2013

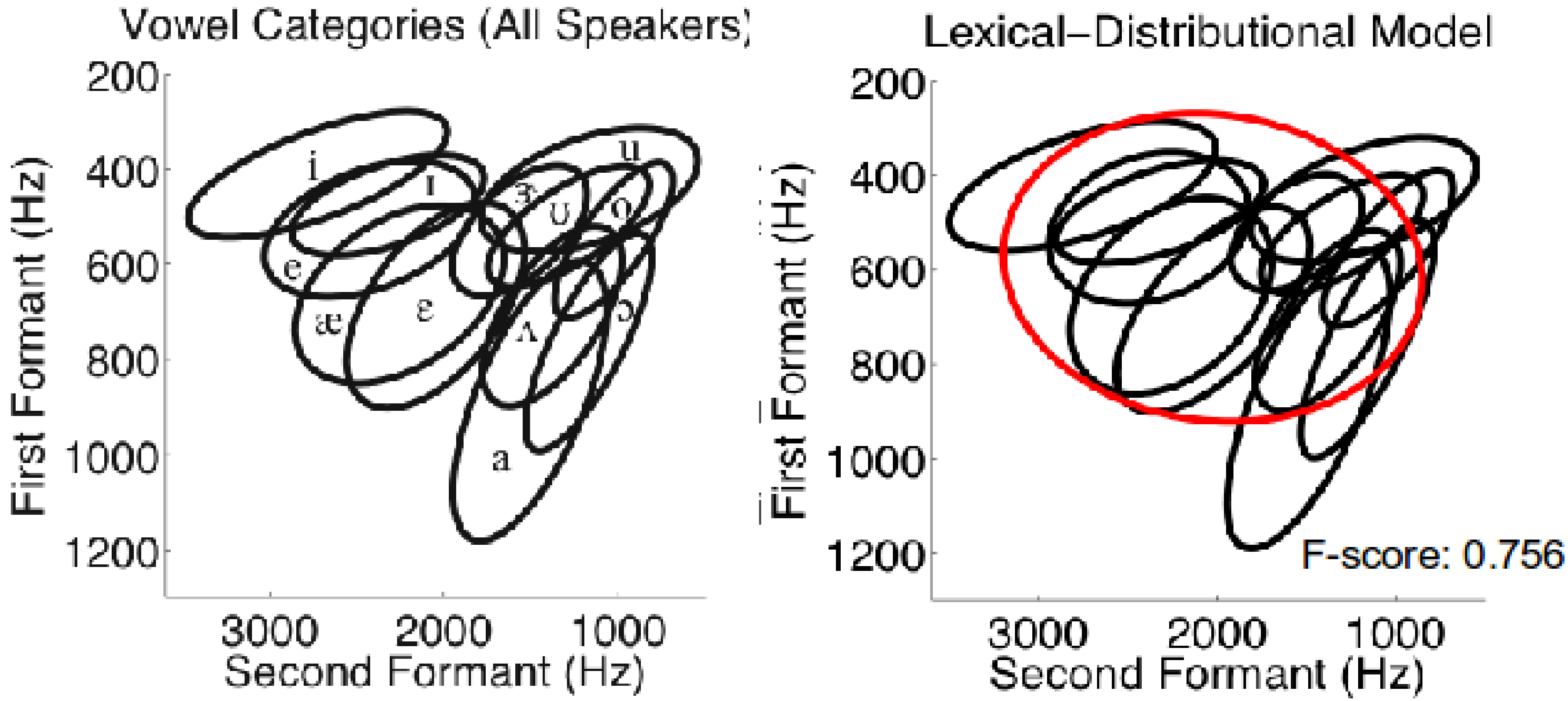


Feldman et al. 2013



F1
F2
430
2100
550
1300
700
1220
ih t s || k ay n d || ah v
Acoustic Values
Vowel formant pairs
Lexical Context
Categorical consonant frame
430
2100
t s || k
550
1300
n d ||
700
1220
v
Bayesian Model
Vowel Categories
Gaussian distributions of sound
ih
ay
ah
Lexical Categories
Sequences of phones
ih.t.s
k.ay.n.d
ah.v
Feldman et al. 2013

Feldman et al. 2013 results



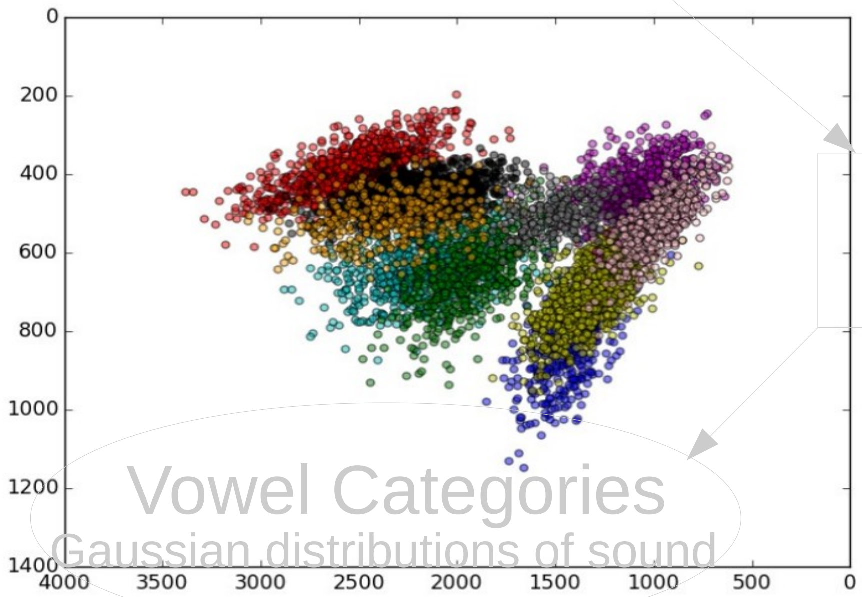
Distributional	Lexical-Distributional
0.45	0.76

Acoustic simplification: lab productions

Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Bayesian Model



Vowel Categories
Gaussian distributions of sound

Lexical Categories
Sequences of phones

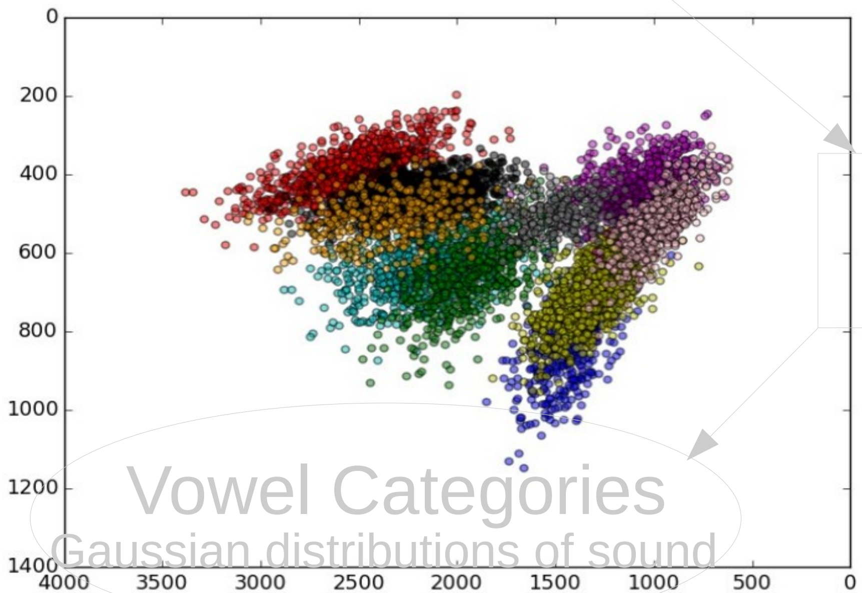
Hillenbrand et al. 1995

Feldman et al. 2013

Acoustic simplification: lab productions

Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame



Hillenbrand et al. 1995

Bayesian Model

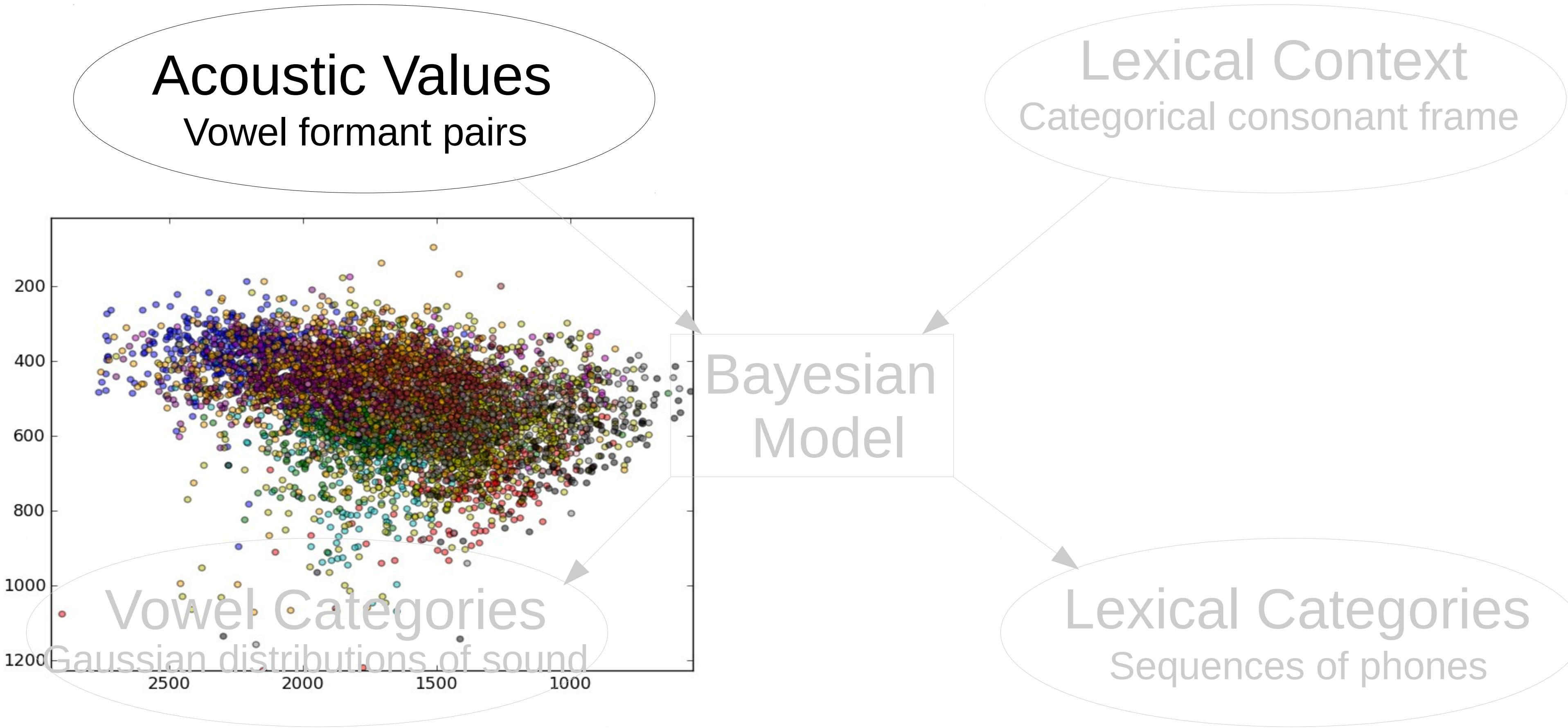
- Stressed, single syllable
- No noise, reduction, co-articulation
- No **prosodic variability** (affects vowel quality and duration)

Vowel Categories
Gaussian distributions of sound

Lexical Categories
Sequences of phones

Feldman et al. 2013

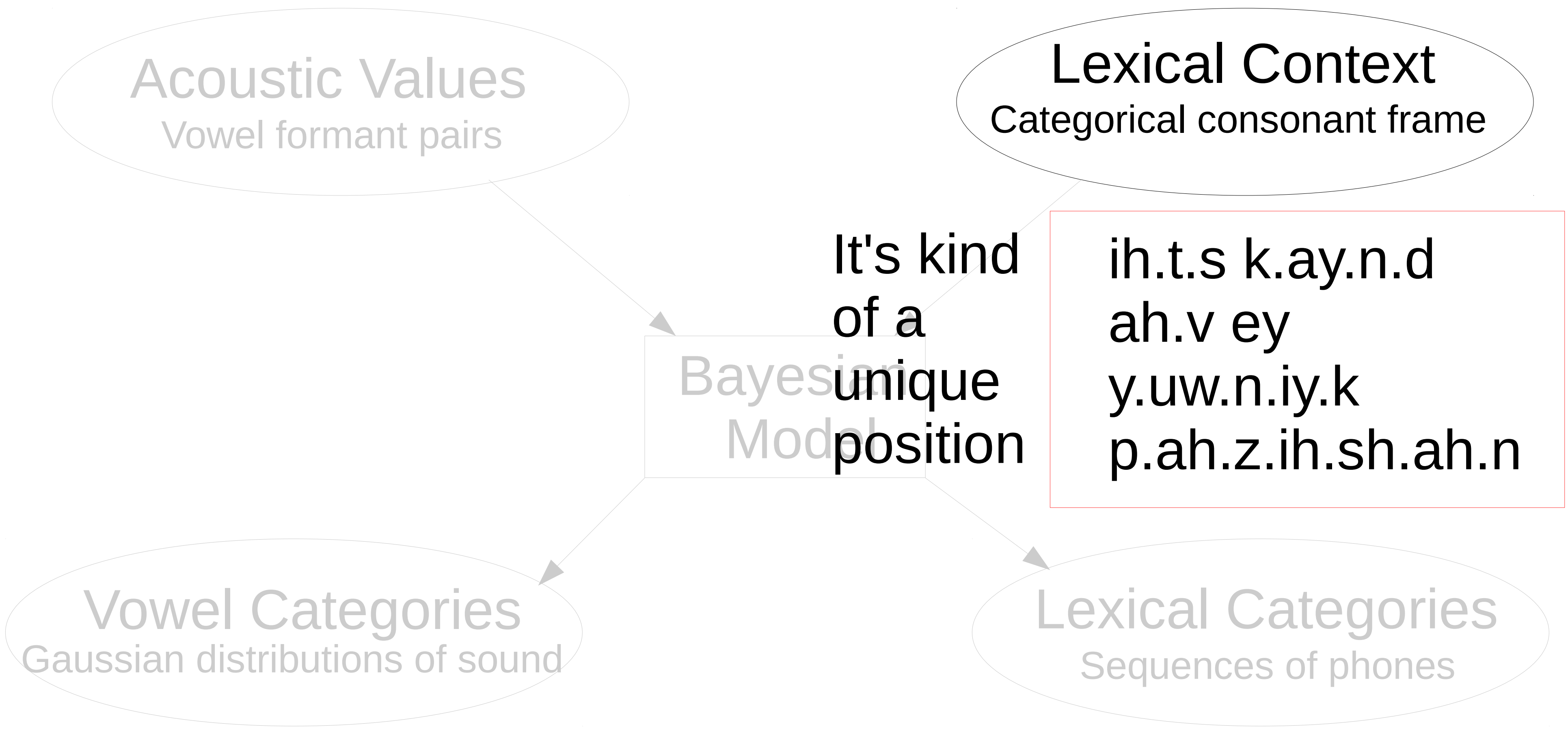
Acoustic simplification: corpus vowels



Buckeye Speech corpus (Pitt et al. 2007)

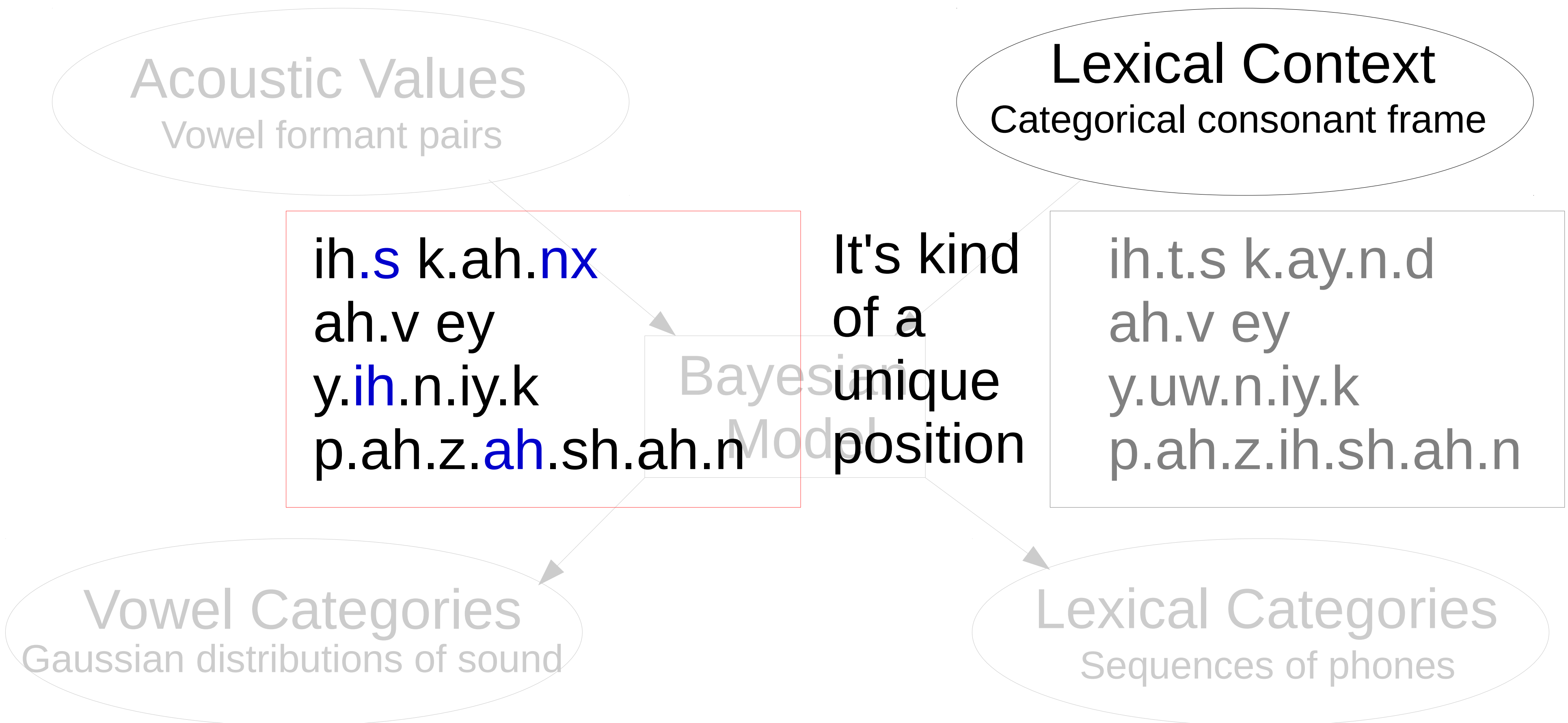
Feldman et al. 2013

Lexical simplification: phonemic transcription



Feldman et al. 2013

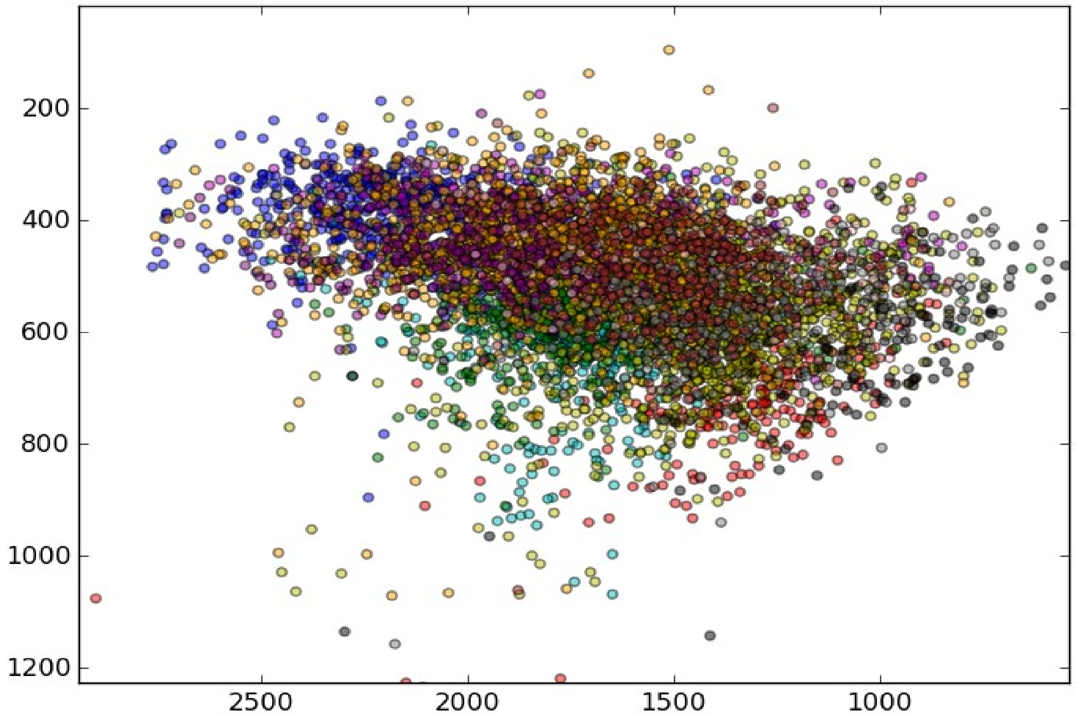
Lexical simplification: phonetic transcription



Feldman et al. 2013

What Children Actually Hear

Natural Vowels

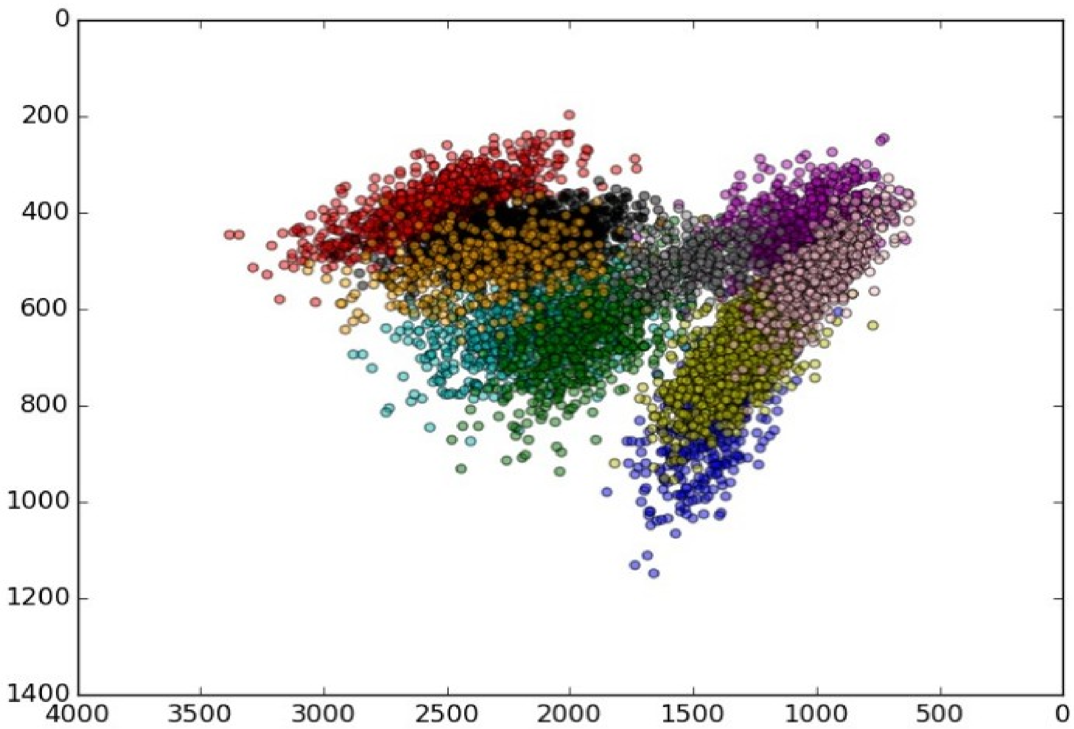


Reduced Lexical Items

ih.s k.ah.nx
ah.v ey
y.ih.n.iy.k
p.ah.z.ah.sh.ah.n

What Models Actually Receive

Lab Vowels



Phonemic Transcription

ih.t.s k.ay.n.d
ah.v ey
y.uw.n.iy.k
p.ah.z.ih.sh.ah.n

How do models perform given more naturalistic data?

- Models help explore what can be learned from the input, given some algorithm

How do models perform given more naturalistic data?

- Models help explore what can be learned from the input, given some algorithm
- Computational models aren't people; some simplification to input must be made

How do models perform given more naturalistic data?

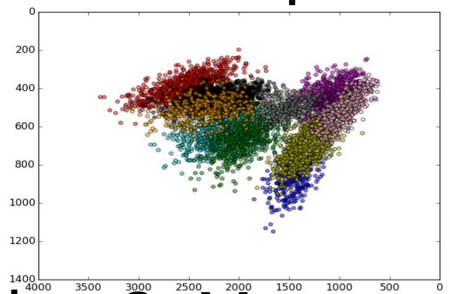
- Models help explore what can be learned from the input, given some algorithm
- Computational models aren't people; some simplification to input must be made
- What is the impact of input simplifications on model performance?

How do models perform given more naturalistic data?

- Models help explore what can be learned from the input, given some algorithm
- Computational models aren't people; some simplification to input must be made
- What is the impact of input simplifications on model performance?
- Are conclusions drawn from these models reliable?

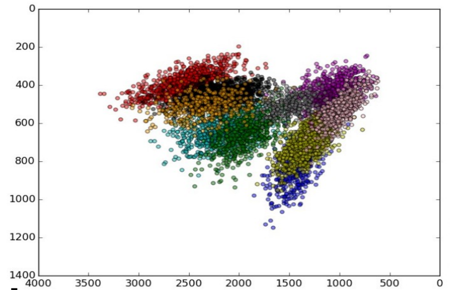
Overview

- Simulation 1: Replication of Simplified Input



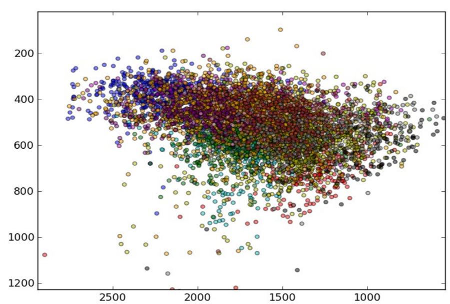
k.ay.n.d

- Simulation 2: More realistic lexical information



k.ah.nx

- Simulation 3: More realistic acoustic information



k.ah.nx

Corpora

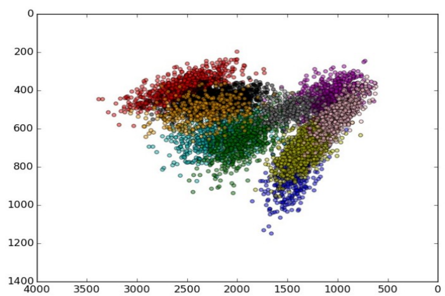
- Laboratory vowel productions:
  - English: Hillenbrand et al. 1995
  - Japanese: Mokhtari & Tanaka 2000
- Natural Speech:
  - English: Buckeye Speech corpus (Pitt et al. 2007)
  - Japanese: R-JMICC corpus (Mazuka et al. 2006)

Simulation 1:
Replication of Simplified Input

Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Resampled



Bayesian Model

Phonemic

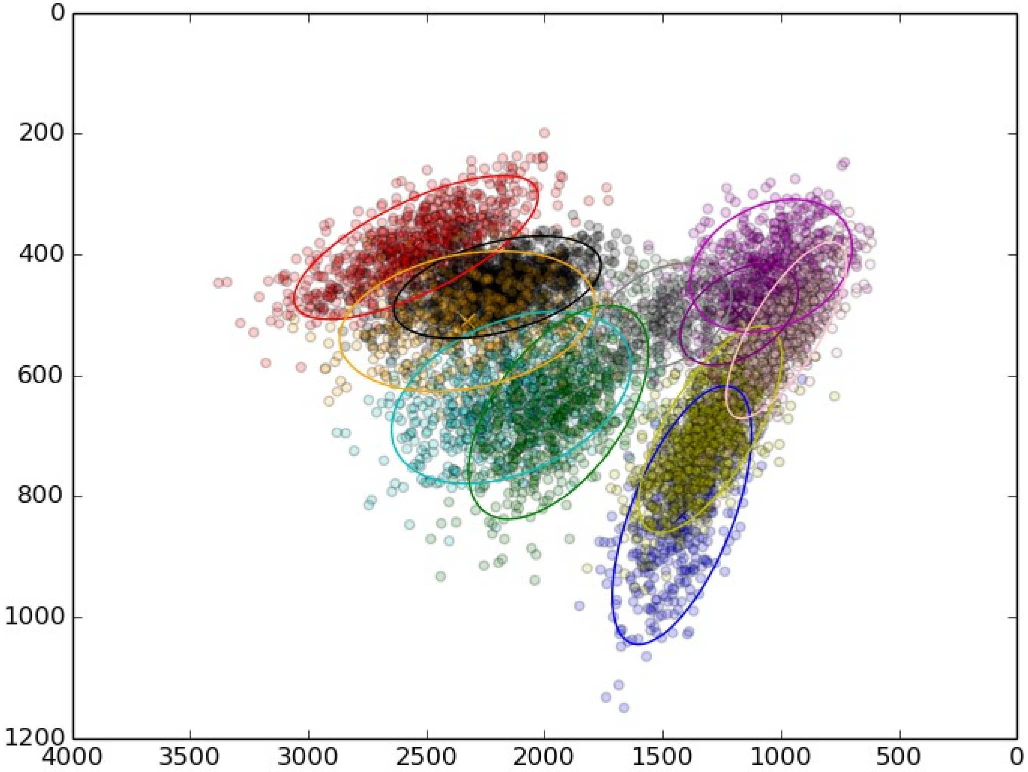
k.ay.n.d

Hillenbrand et al. 1995
Mokhtari & Tanaka 2000

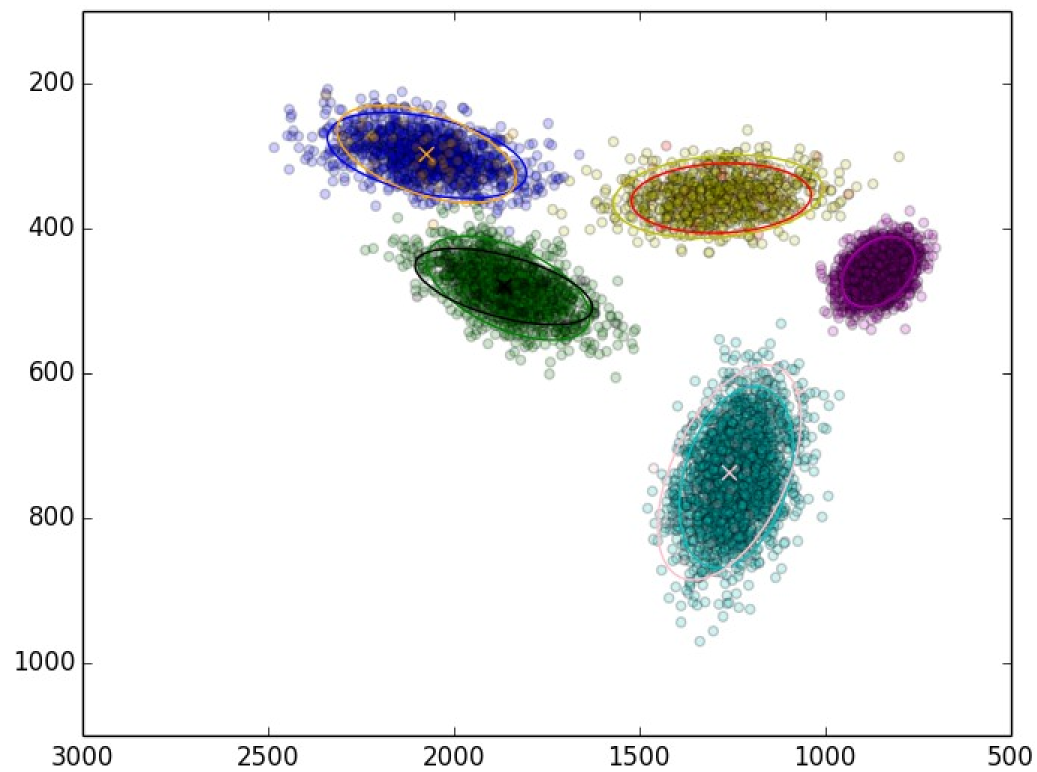
Buckeye Speech corpus (Pitt et al. 2007)
R-JMICC corpus (Mazuka et al. 2006)

Replication of Simplified Input: English vs. Japanese Lab Vowels

English

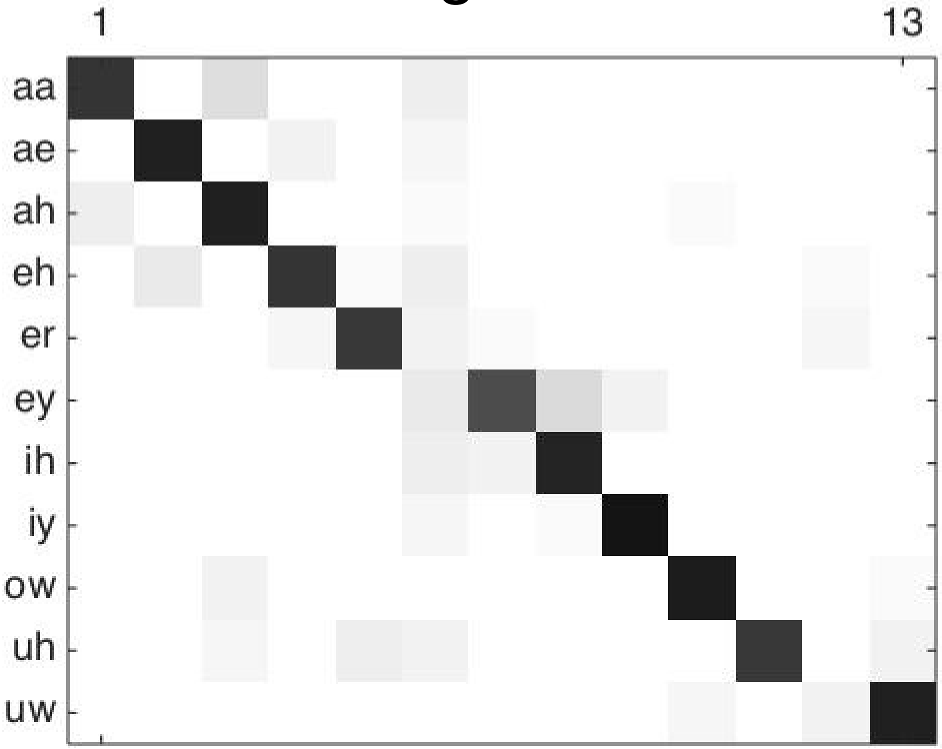


Japanese

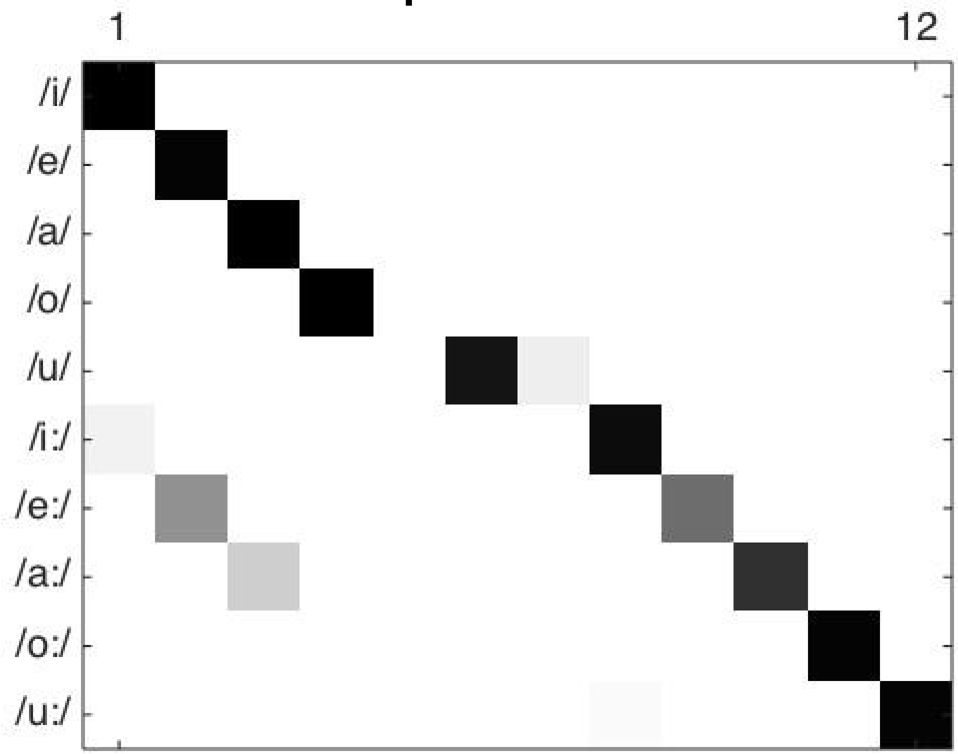


Simplified Input: Successful Category Recovery

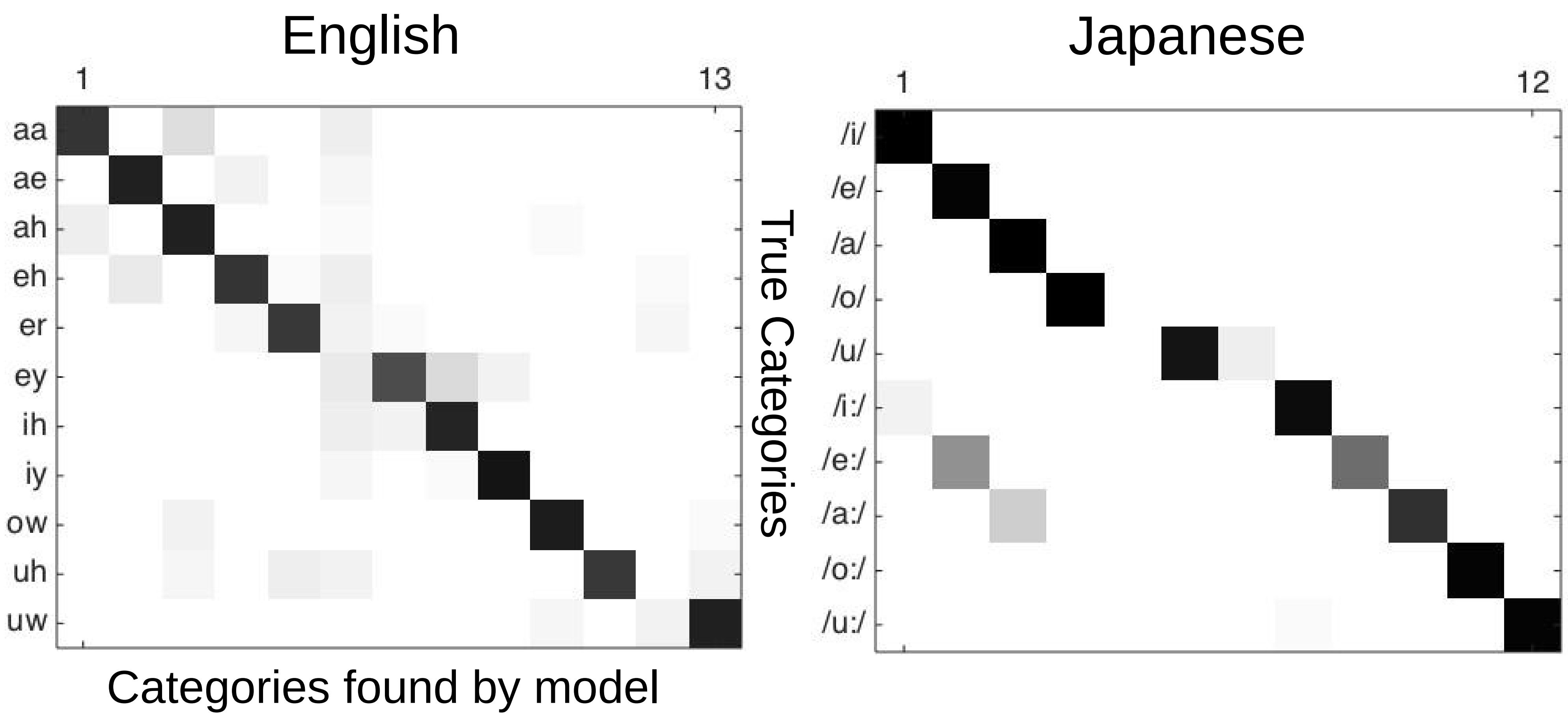
English



Japanese

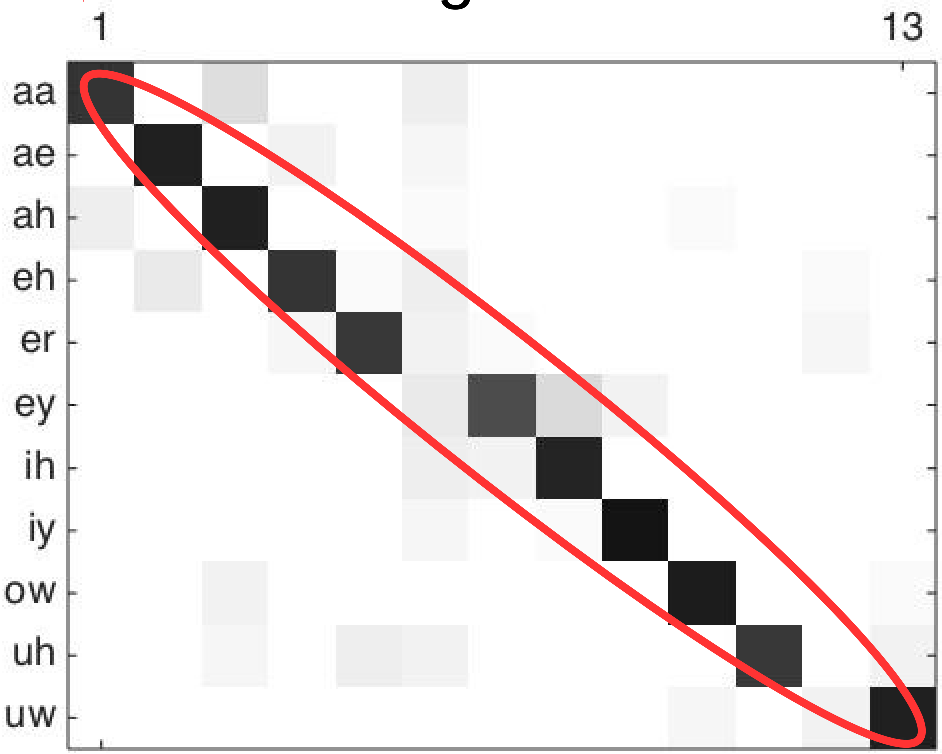


Simplified Input: Successful Category Recovery

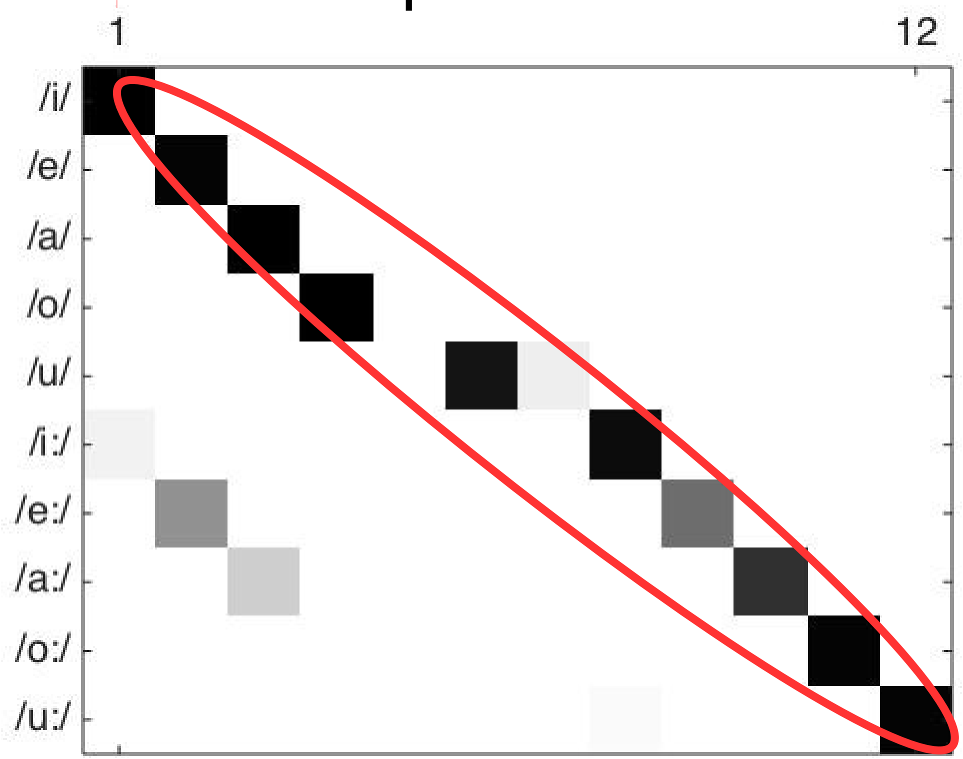


Simplified Input: Successful Category Recovery

English

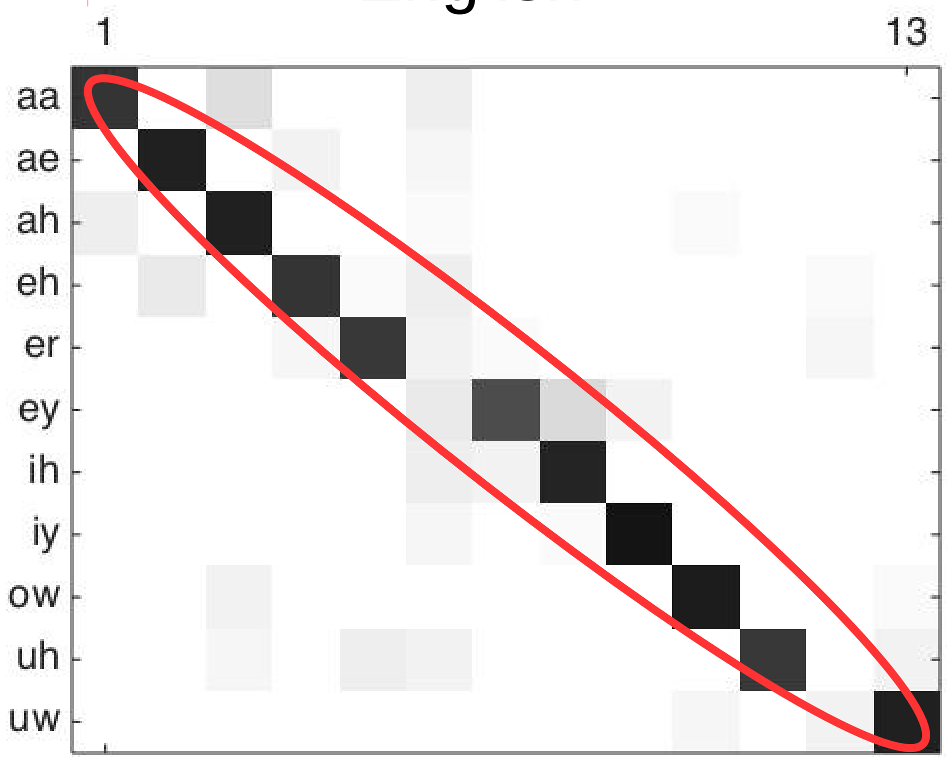


Japanese

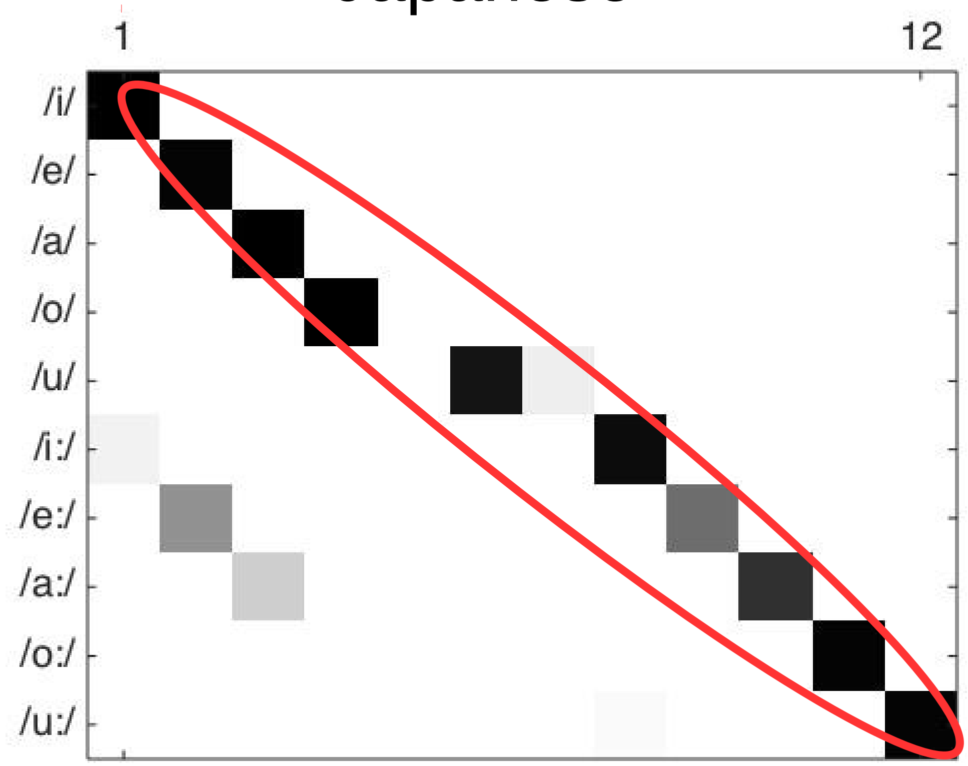


Simplified Input: Successful Category Recovery

English



Japanese



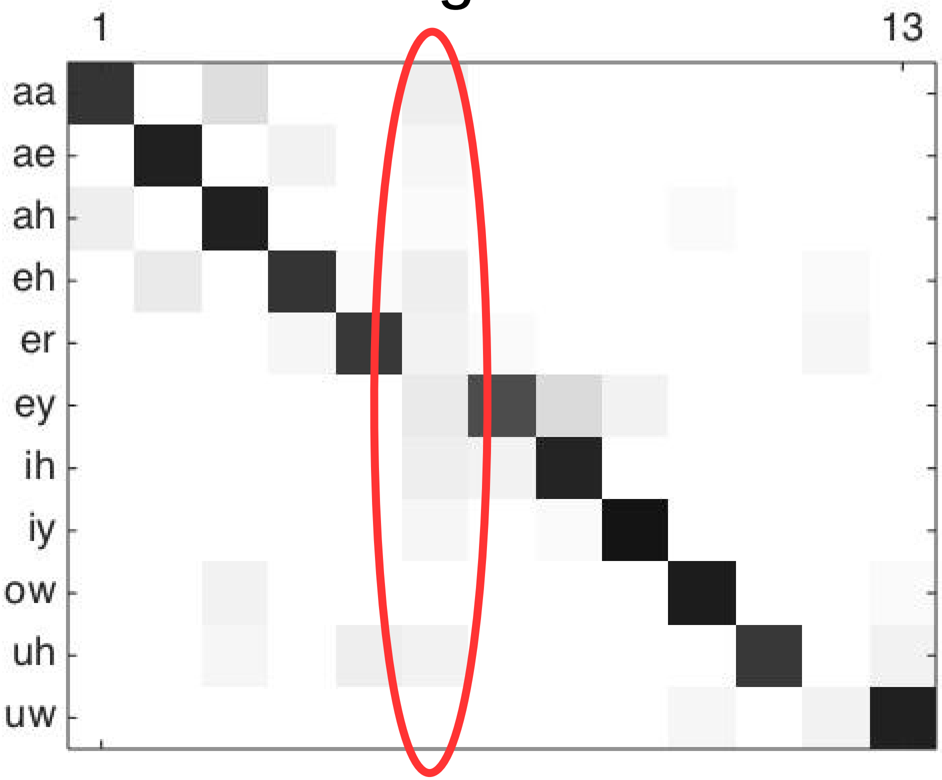
Phonetic F-Score: 0.78

Phonetic F-Score: 0.98

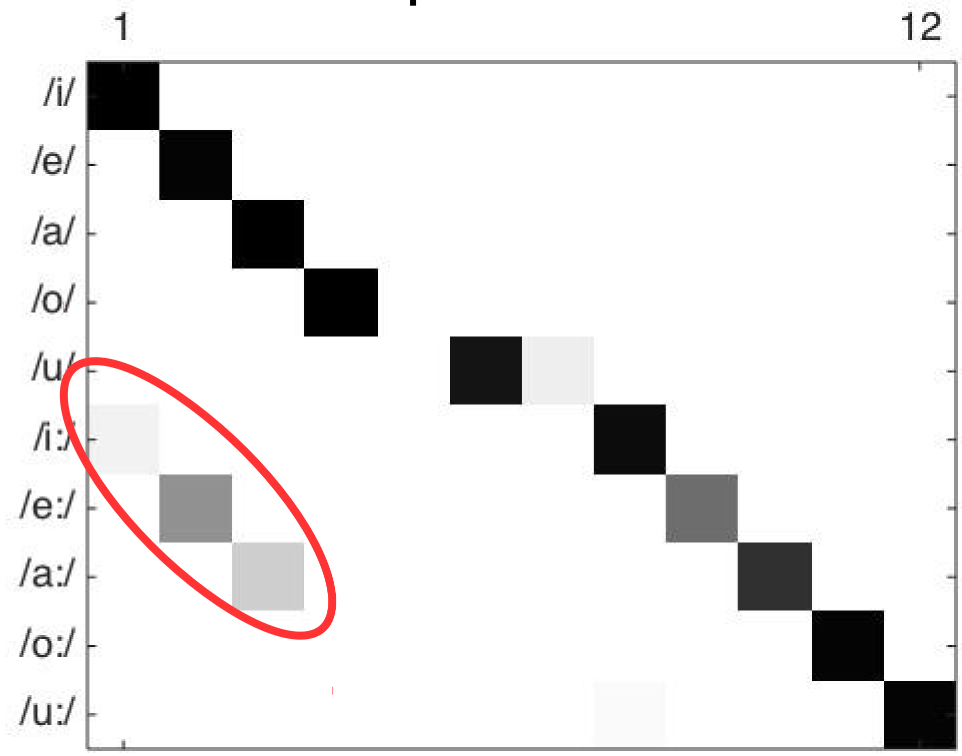
Original Feldman et al. 2013 results: Phonetic F-Score: 0.76

Simplified Input: Successful Category Recovery

English



Japanese

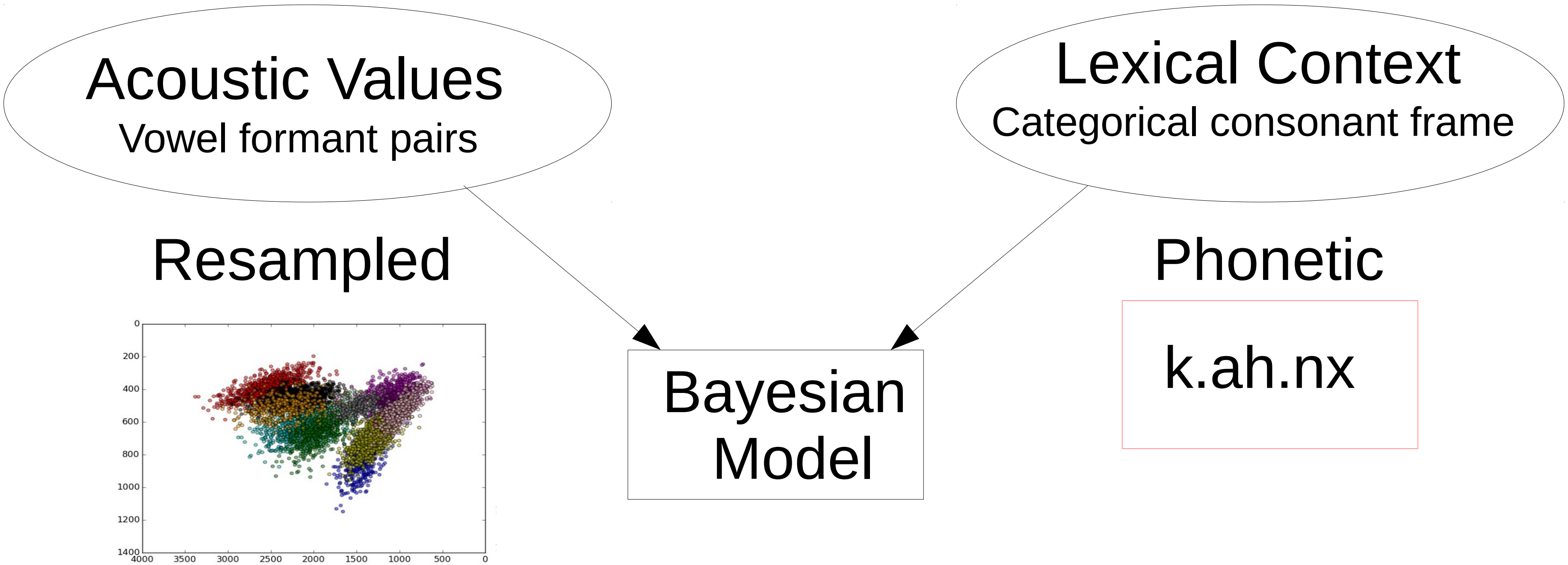


Phonetic F-Score: 0.78

Phonetic F-Score: 0.98

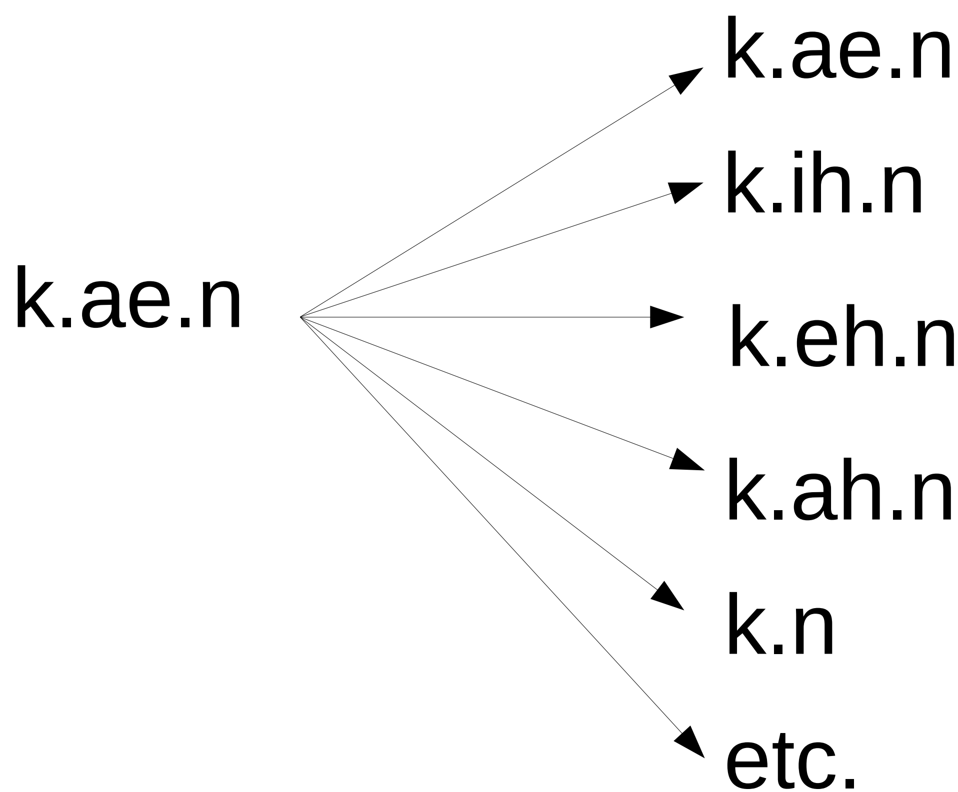
Original Feldman et al. 2013 results: Phonetic F-Score: 0.76

Simulation 2: Phonetic Transcription



Corpus effects of phonetic transcription

- English: vowels of frequent words reduced to schwa in natural speech → increased number of phonetic variants

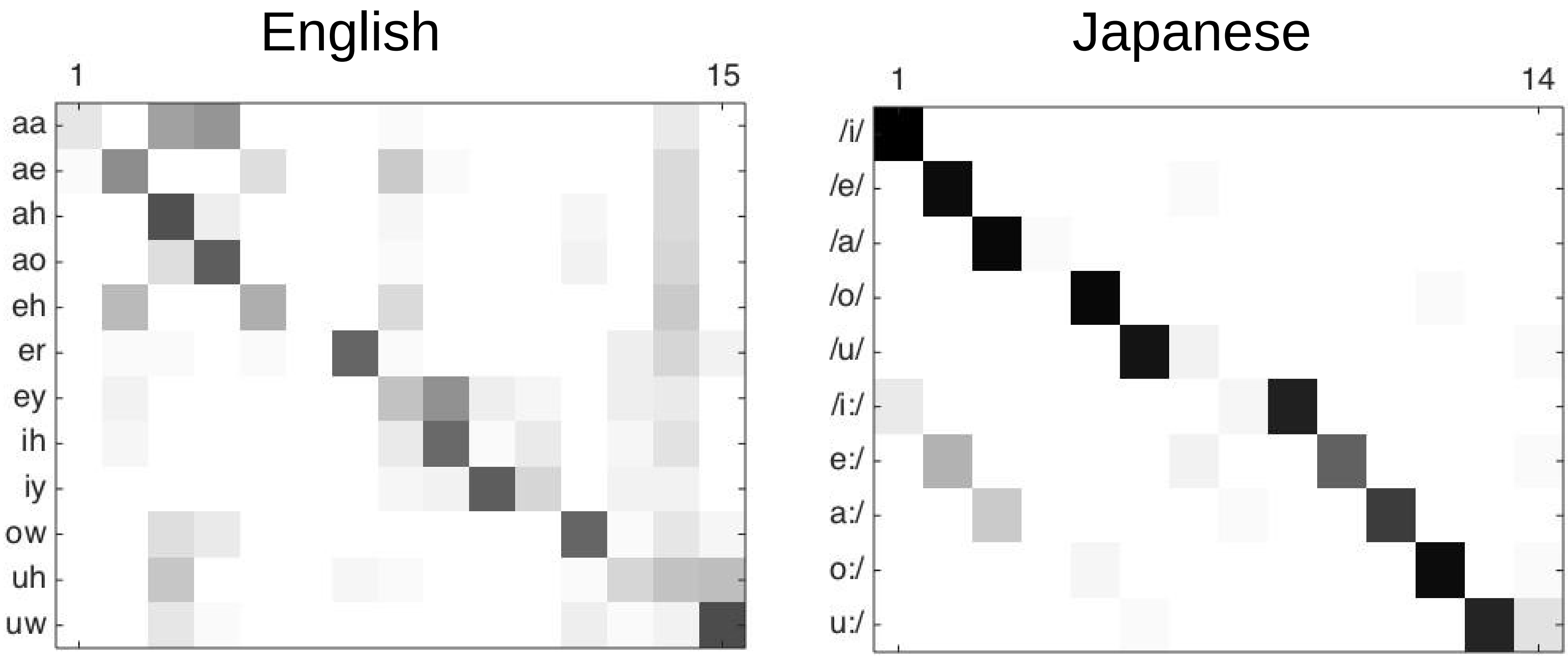


Corpus effects of phonetic transcription

- English: vowels of frequent words reduced to schwa in natural speech → increased number of phonetic variants
- Japanese: less phonetic reduction

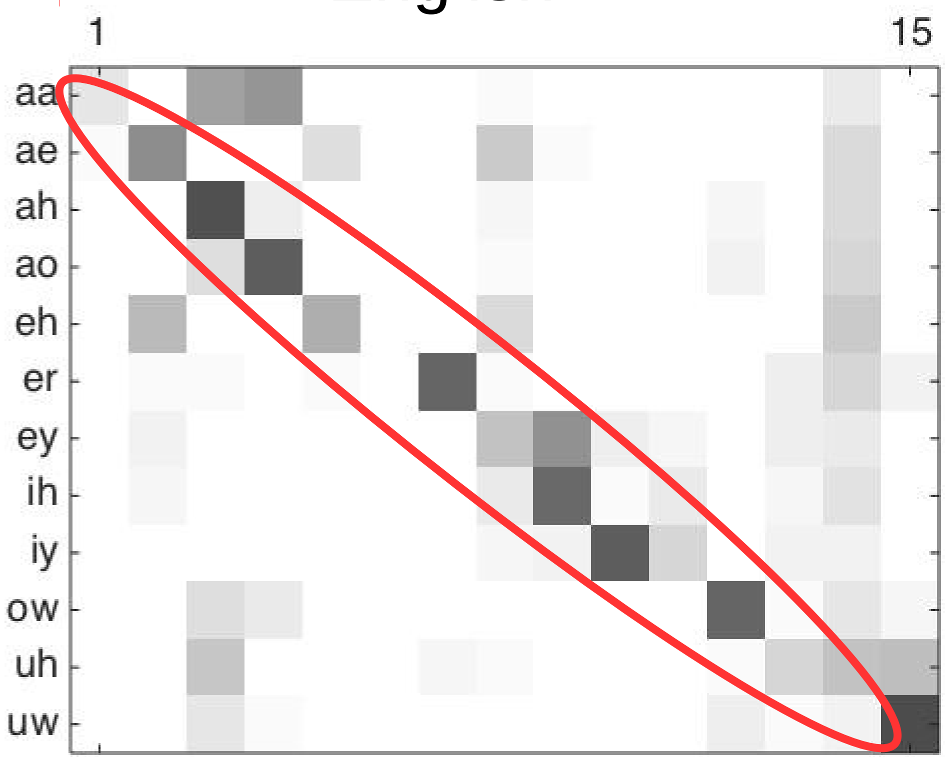
English Word Types		Japanese Word Types	
Phonemic Transcription:	1099	Phonemic Transcription:	751
Phonetic Transcription:	1813	Phonetic Transcription:	791

Phonetic Transcription: Decline in English Performance



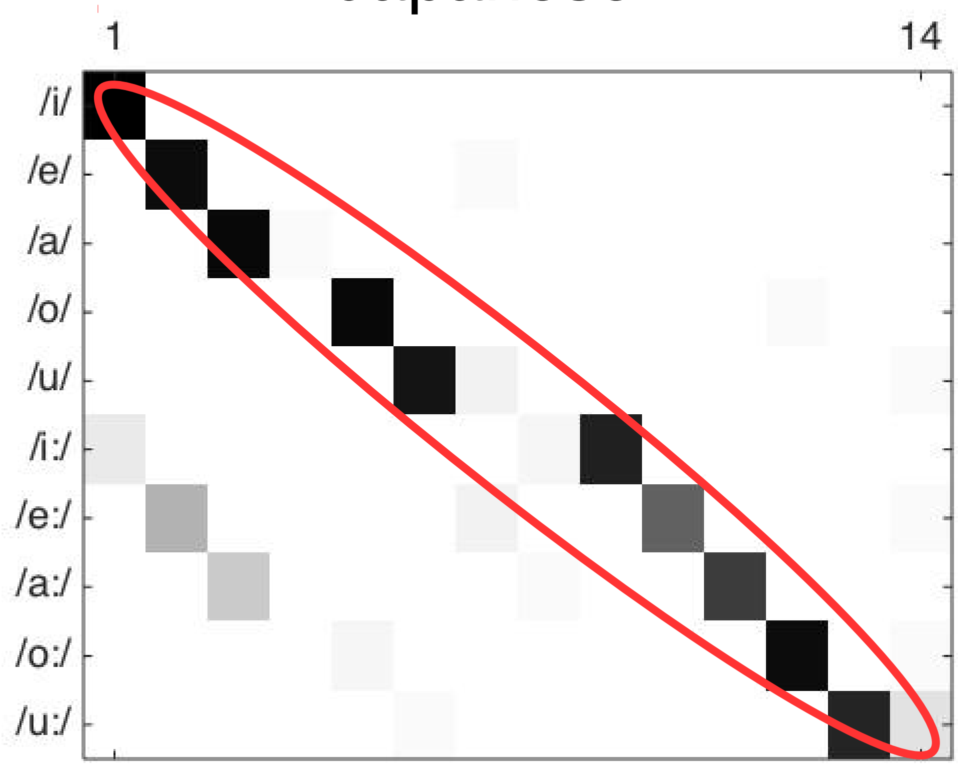
Phonetic Transcription: Decline in English Performance

English



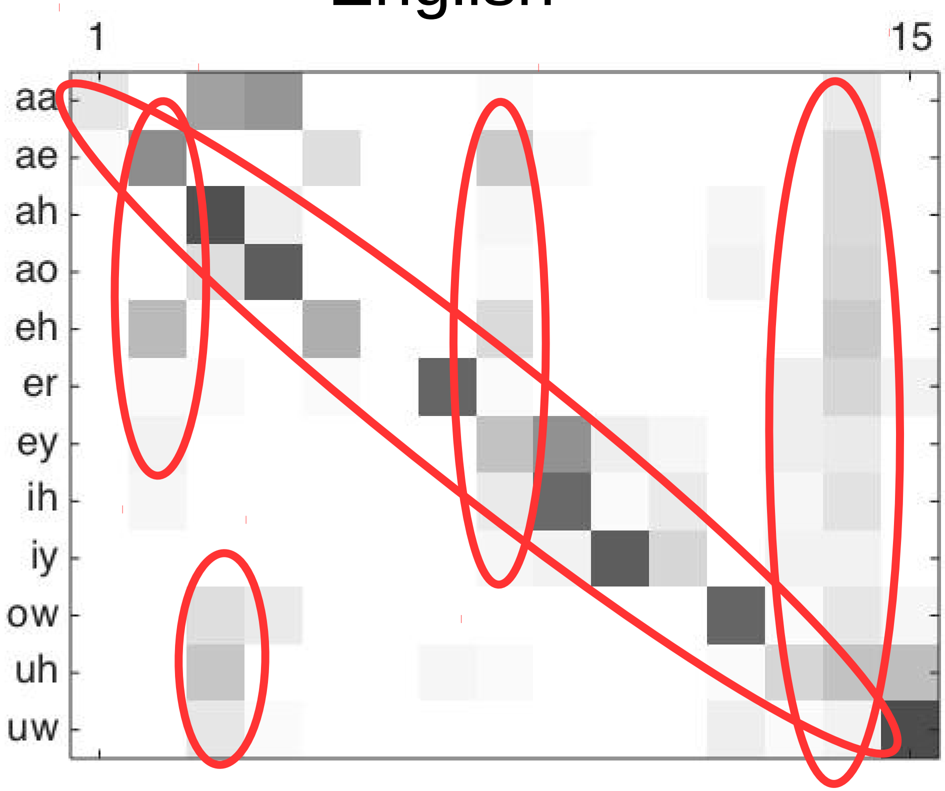
Phonetic F-Score: 0.46

Japanese

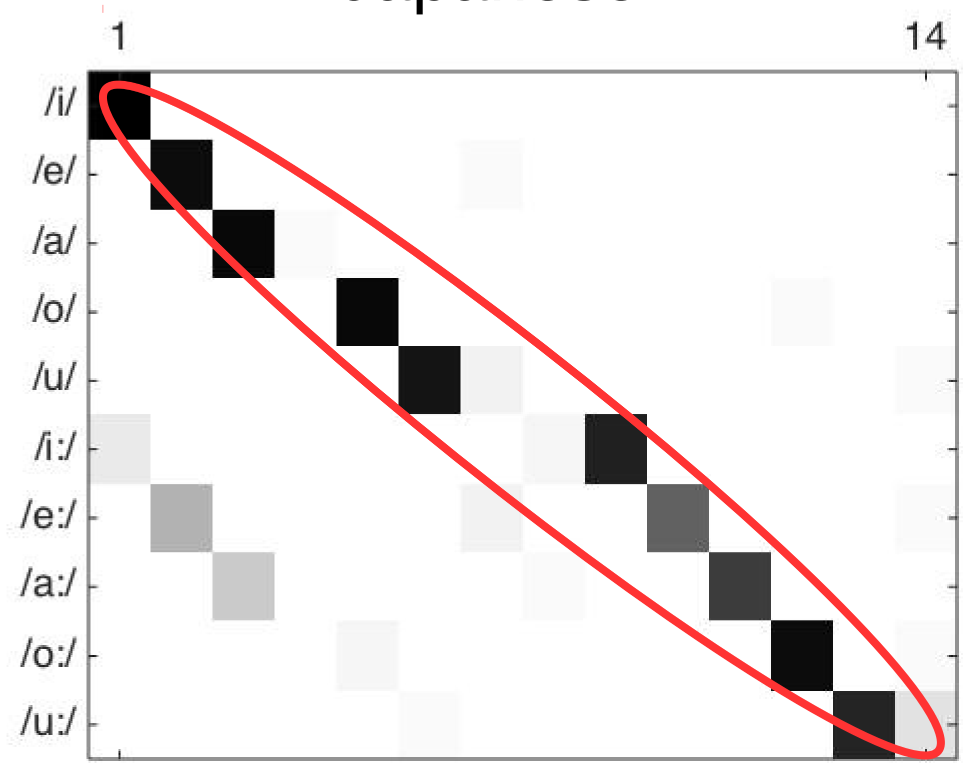


Phonetic F-Score: 0.95

Phonetic Transcription: Decline in English Performance



Phonetic F-Score: 0.46



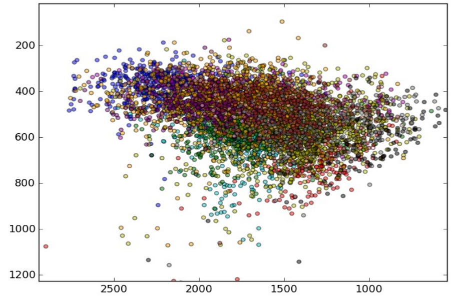
Phonetic F-Score: 0.95

Simulation 3: Realistic Vowels From Corpus

Acoustic Values
Vowel formant pairs

Lexical Context
Categorical consonant frame

Corpus Vowels



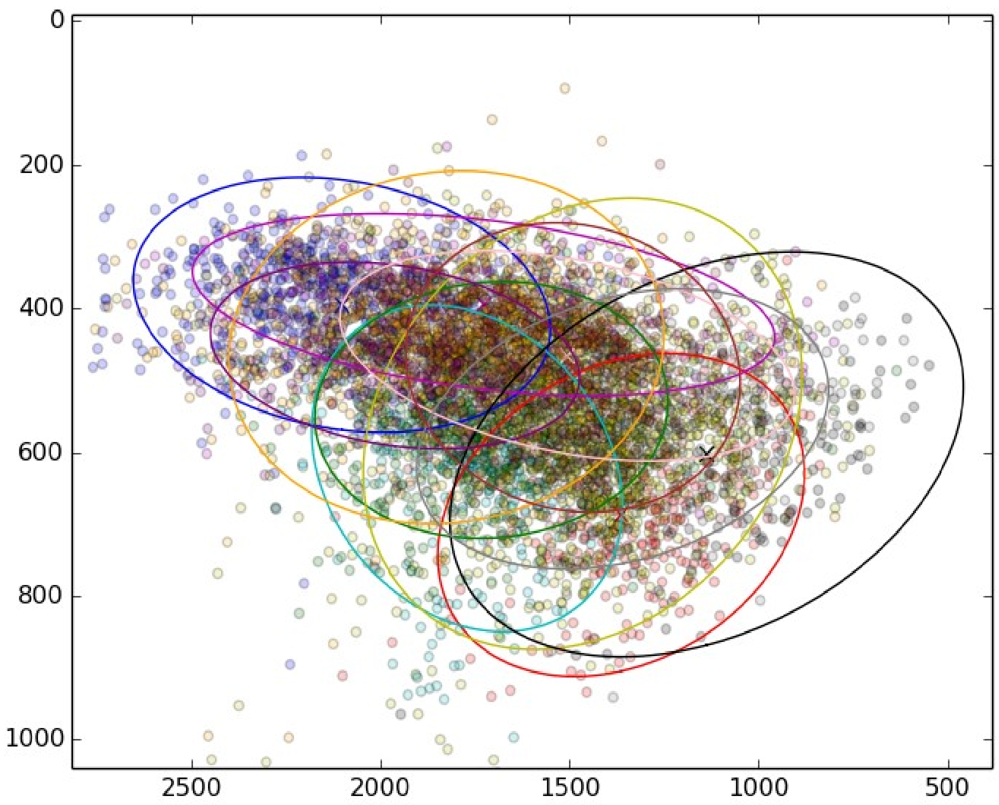
Bayesian Model

Phonetic

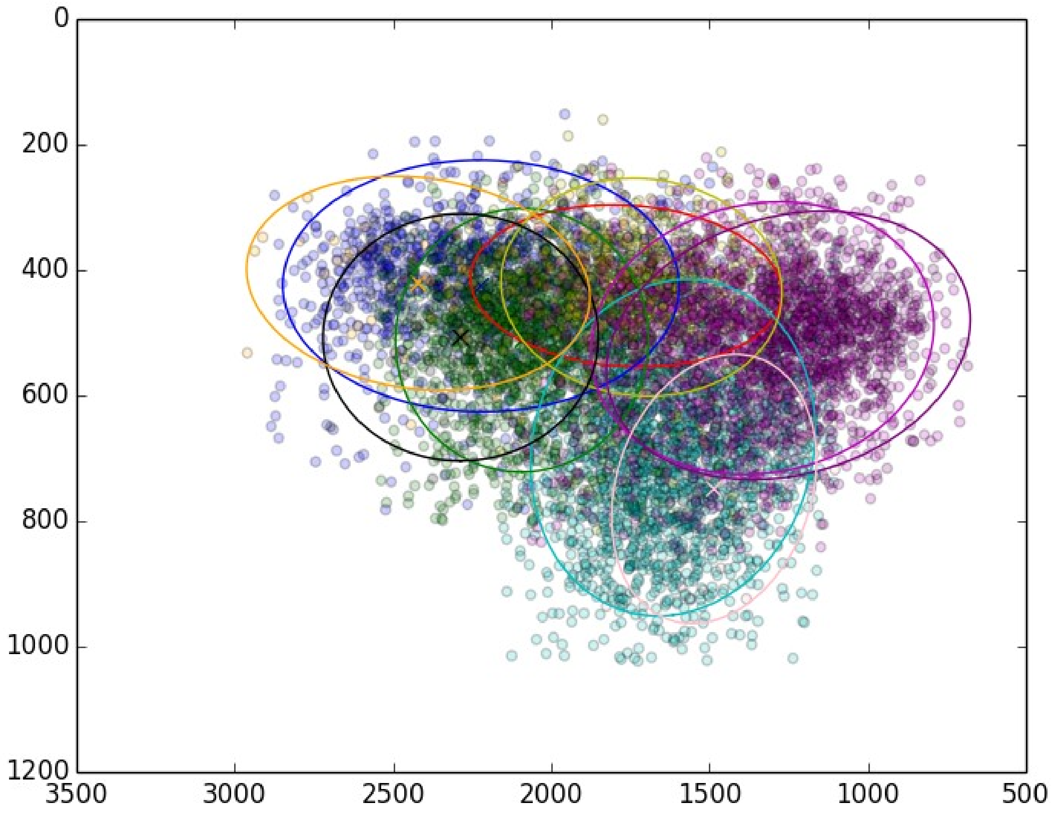
k.ah.nx

Simulation 3:
Realistic Vowels From Corpus

English



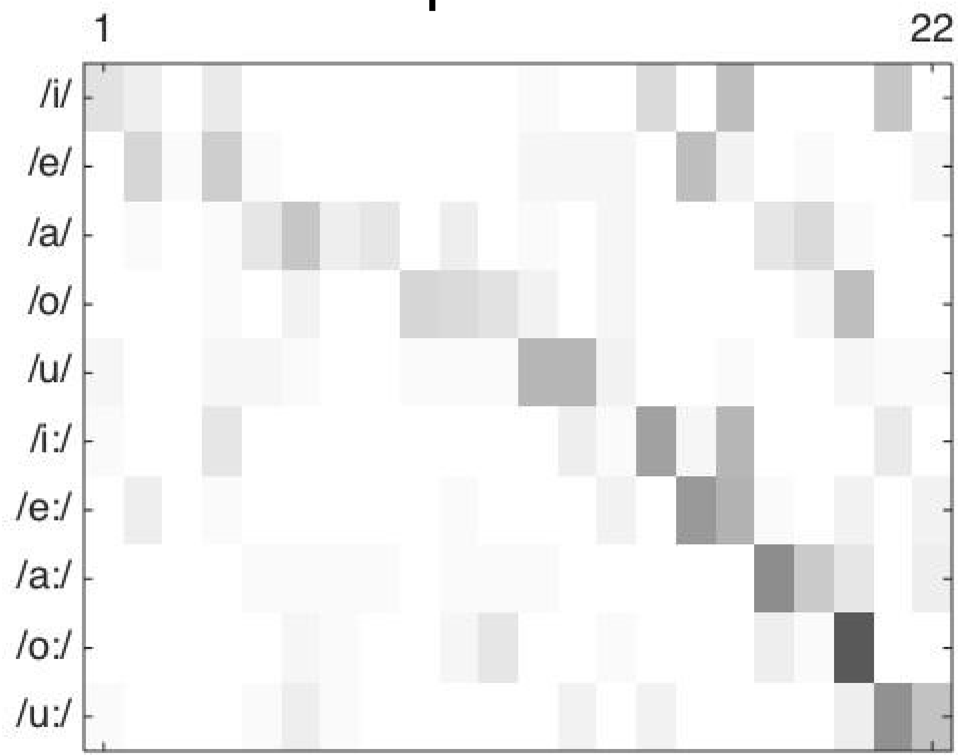
Japanese



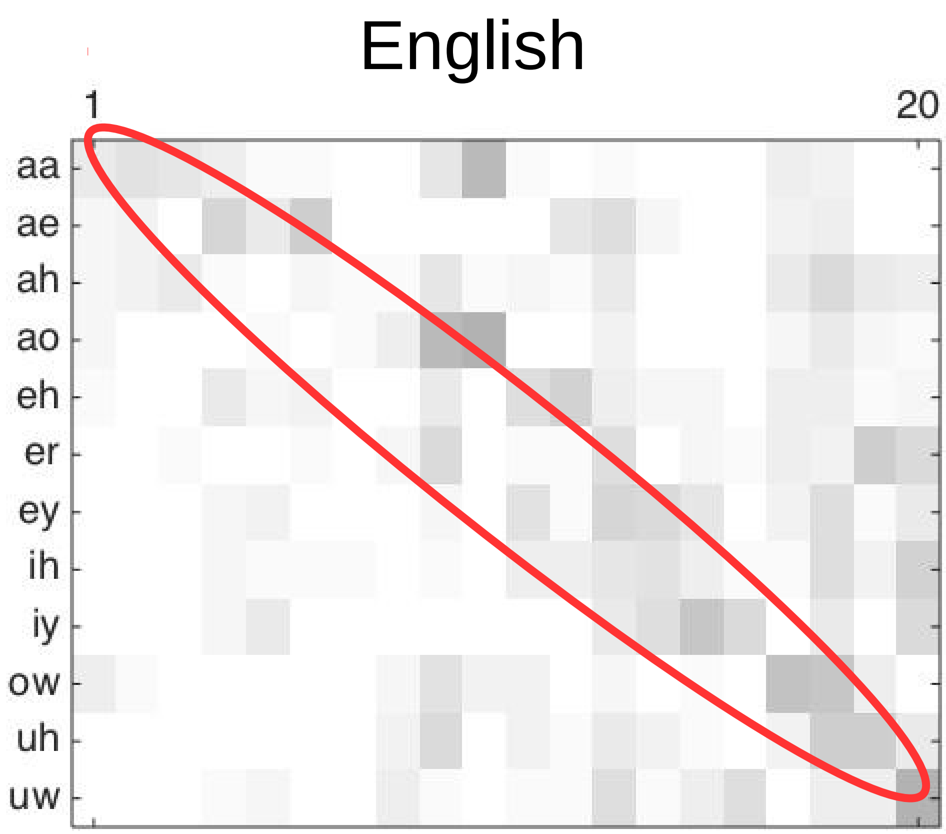
Realistic Vowels: Poor Category Recovery

English

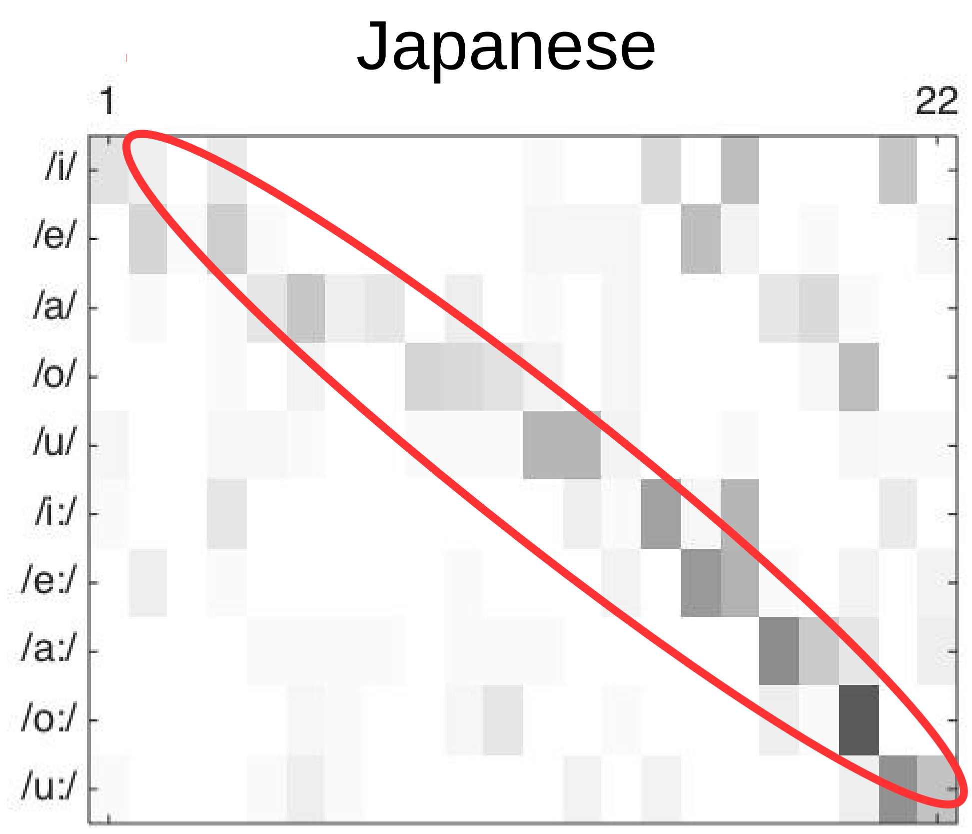
Japanese



Realistic Vowels: Poor Category Recovery



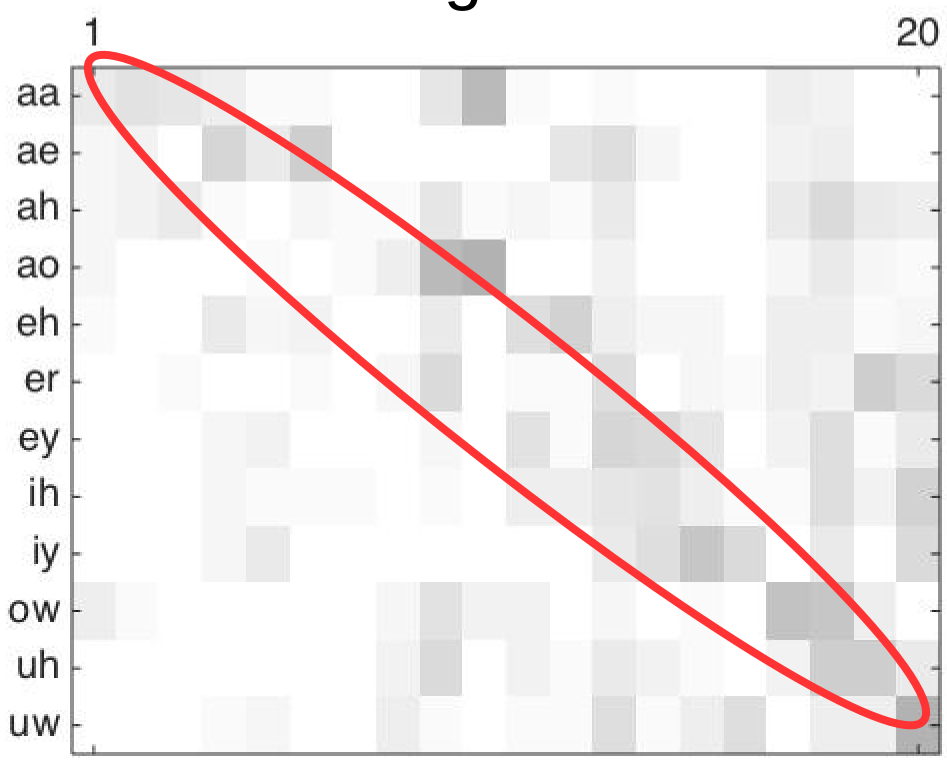
Phonetic F-Score: 0.13



Phonetic F-Score: 0.22

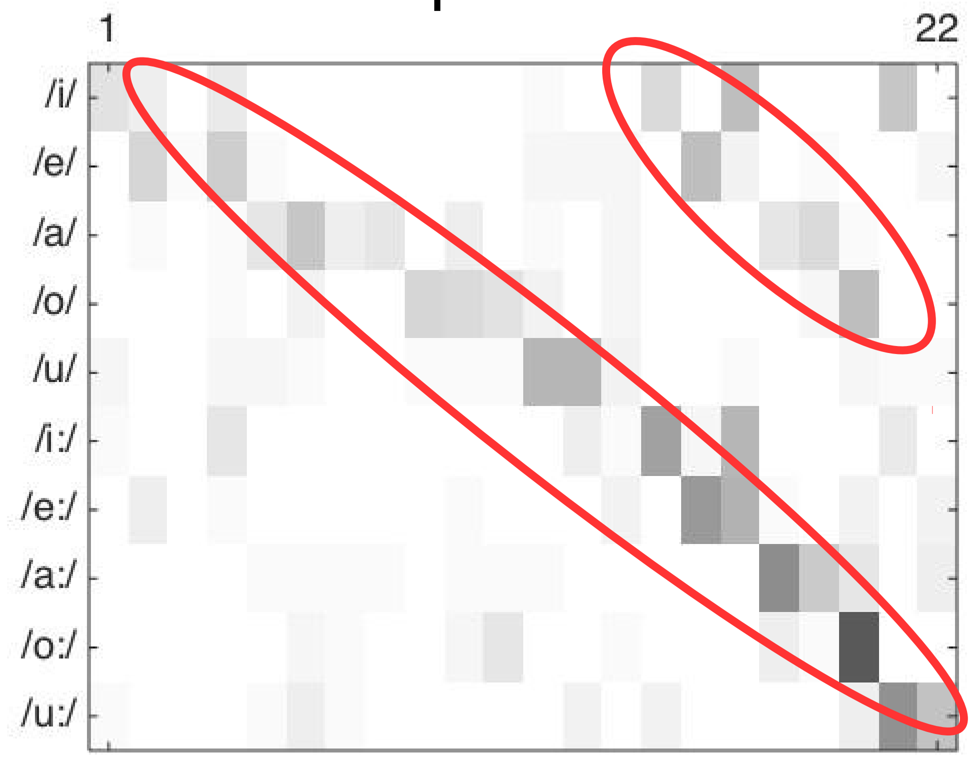
Realistic Vowels: Poor Category Recovery

English



Phonetic F-Score: 0.13

Japanese



Phonetic F-Score: 0.22

Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama

Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama

mama:

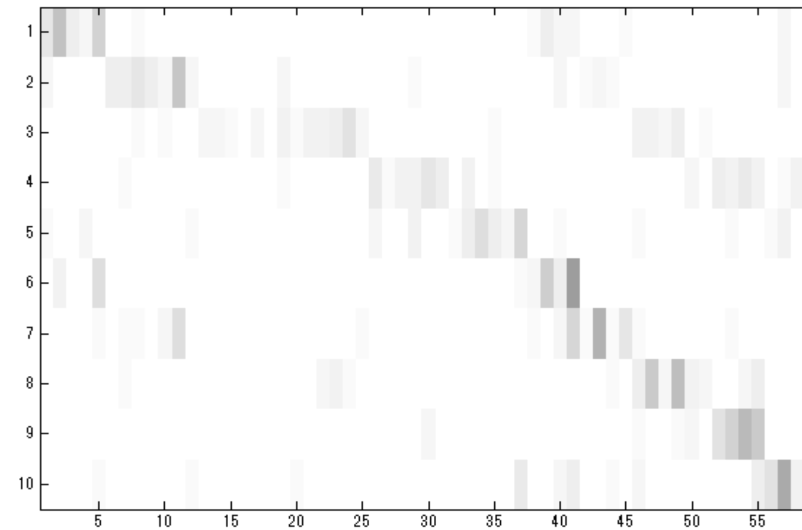
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

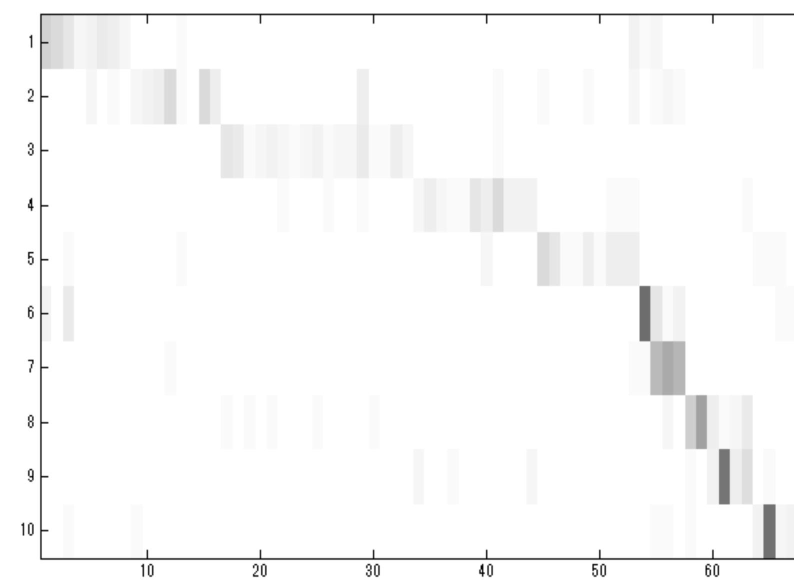
mama

mama:

Without Normalization



With Normalization



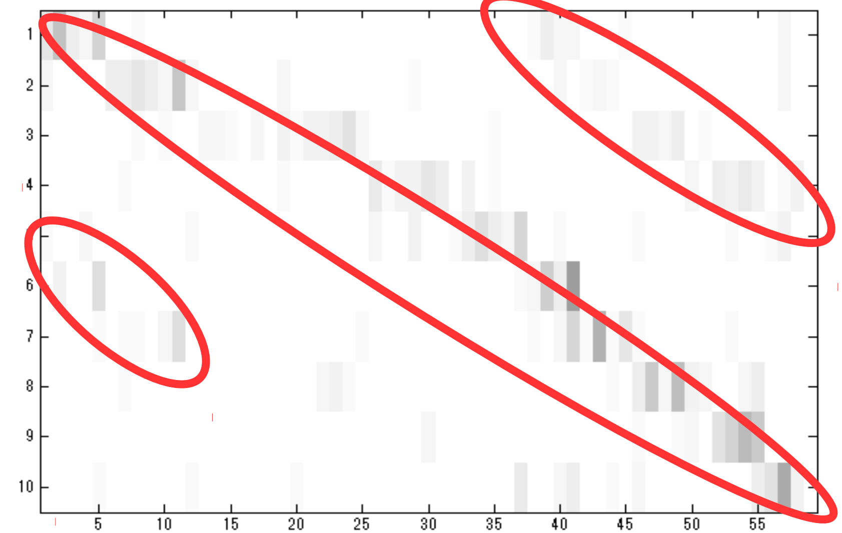
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

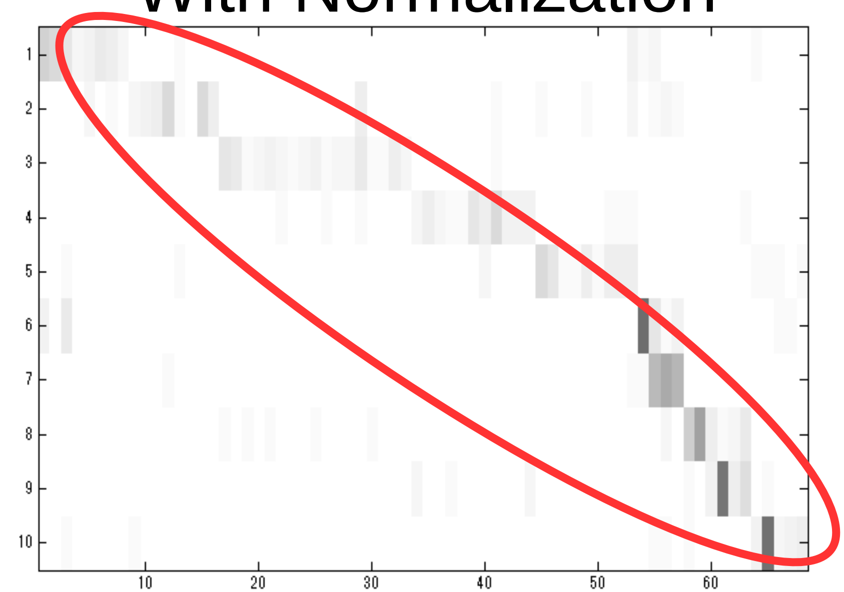
mama

mama:

Without Normalization



With Normalization



Summary of Results

	Phonetic F-Score	
	English	Japanese
Simulation 1: Simplified Input	0.78	0.98
Simulation 2: Phonetic Transcription	0.46	0.95
Simulation 3: Realistic Vowels	0.13	0.22

Discussion

- There is little variability in simplified input, but a lot in the input received by children
  - Lexical variability
  - Acoustic variability
- Adding this variability back to the input can drastically impact model performance, and may have different effects on different languages.
- To explore the learning problem we must have ecologically valid datasets

Thank You!

NSF IIS-1422987

NSF IIS-1421695

NSF DGE-1343012

OSU Lacqueys reading group

Japanese Long vs Short

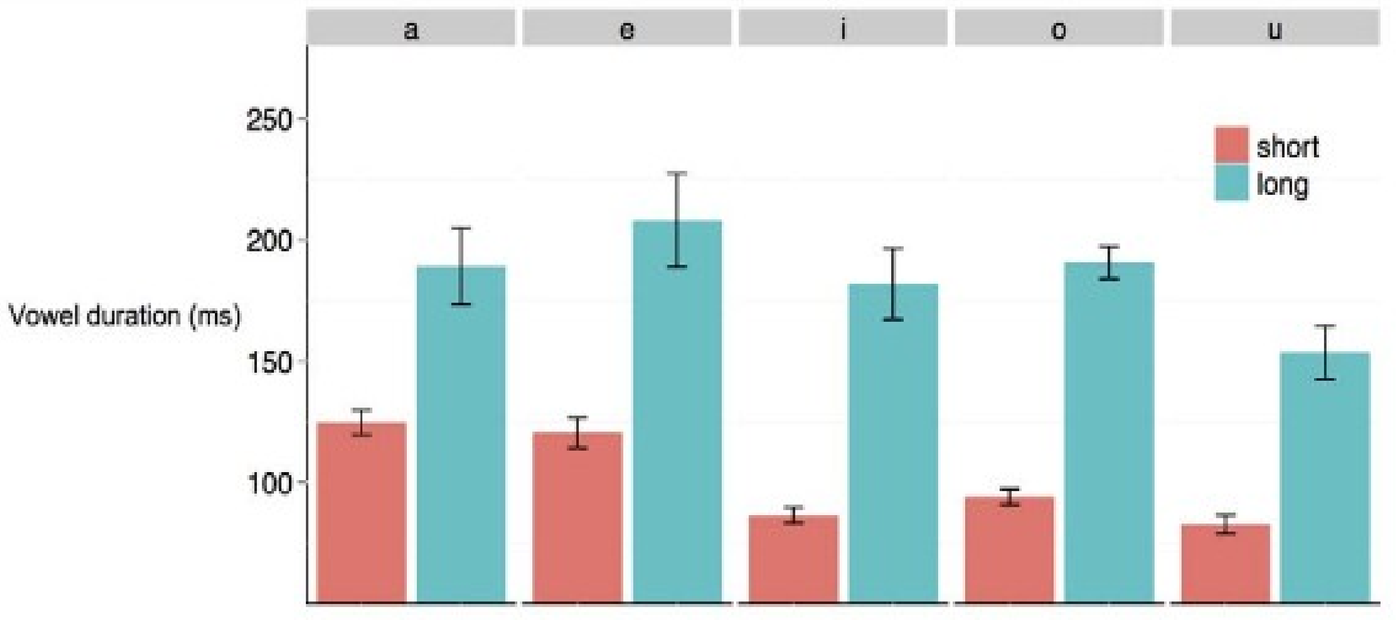
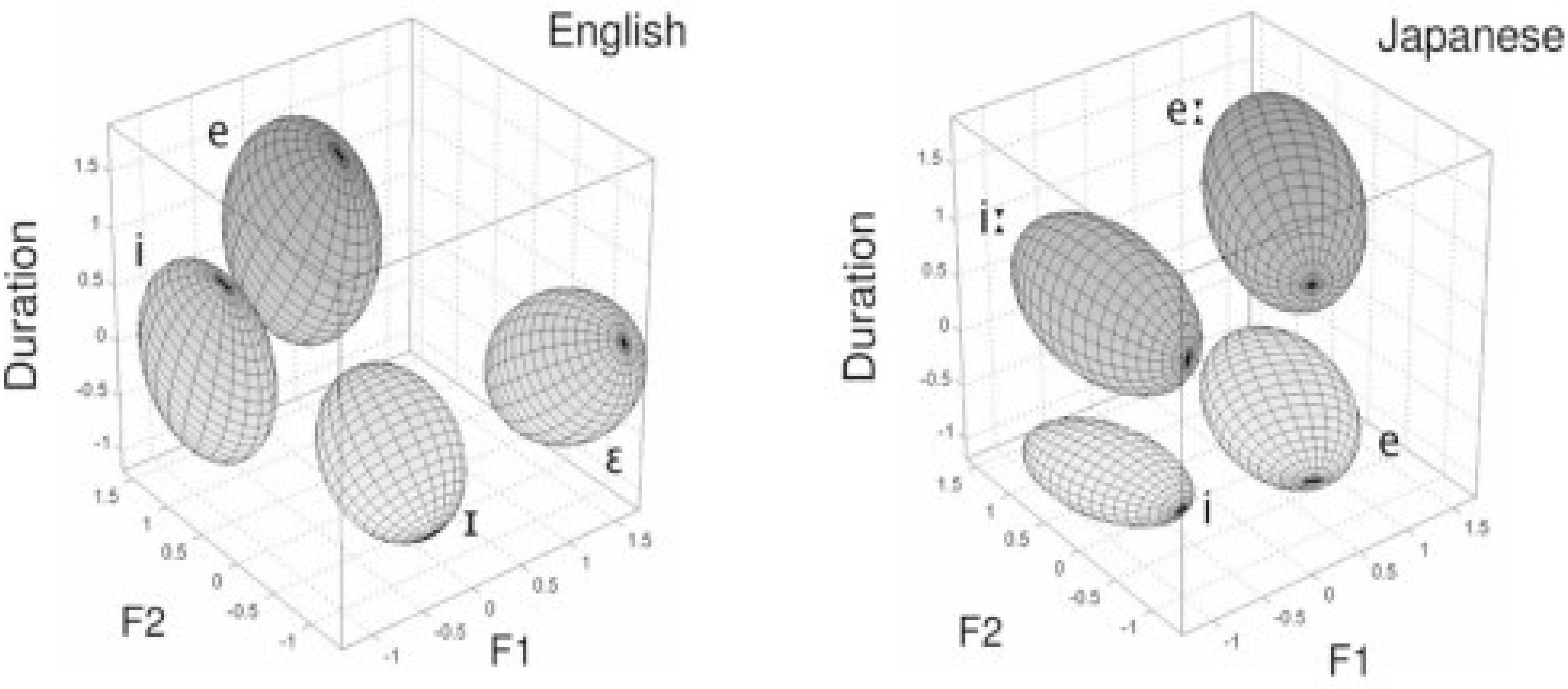


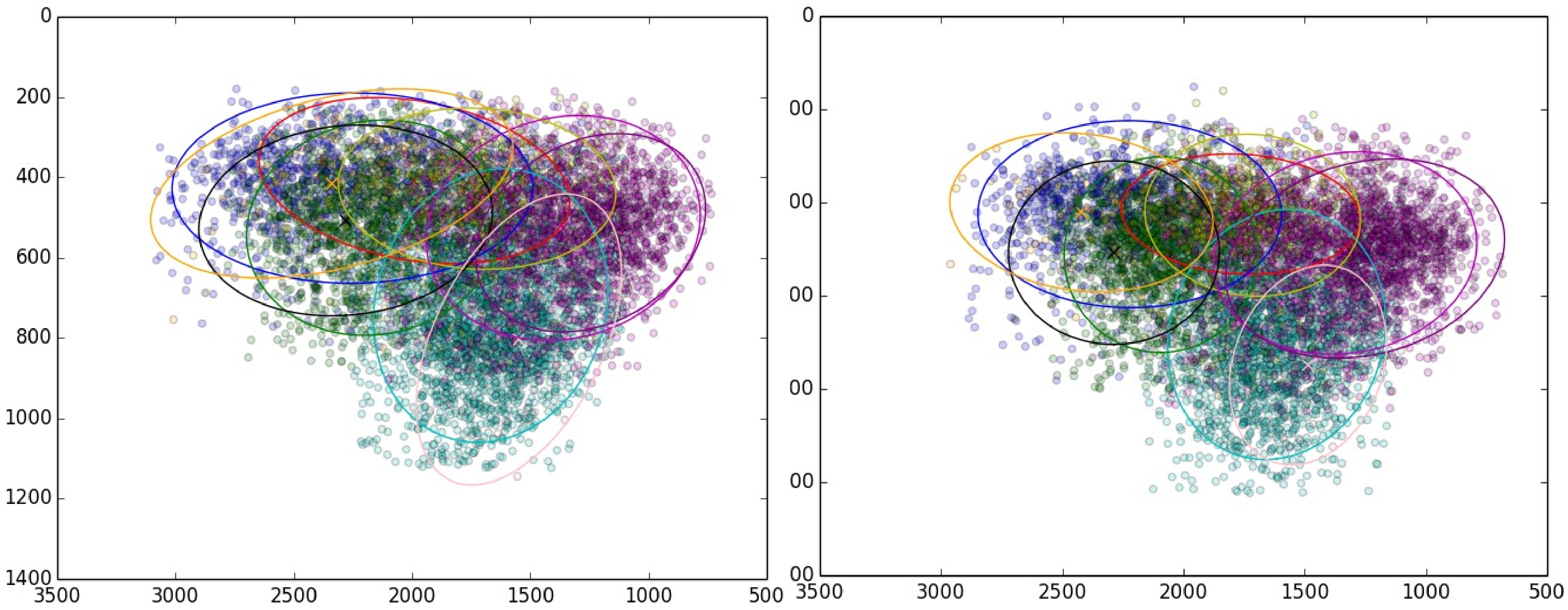
Figure 1. Mean duration of short and long vowels in the present Japanese IDS corpus. The difference in duration between short and long vowels is reliable and the effect size is large. The error bars represent the standard error of the mean for each vowel across participants.
doi:10.1371/journal.pone.0051594.g001

English versus Japanese Duration

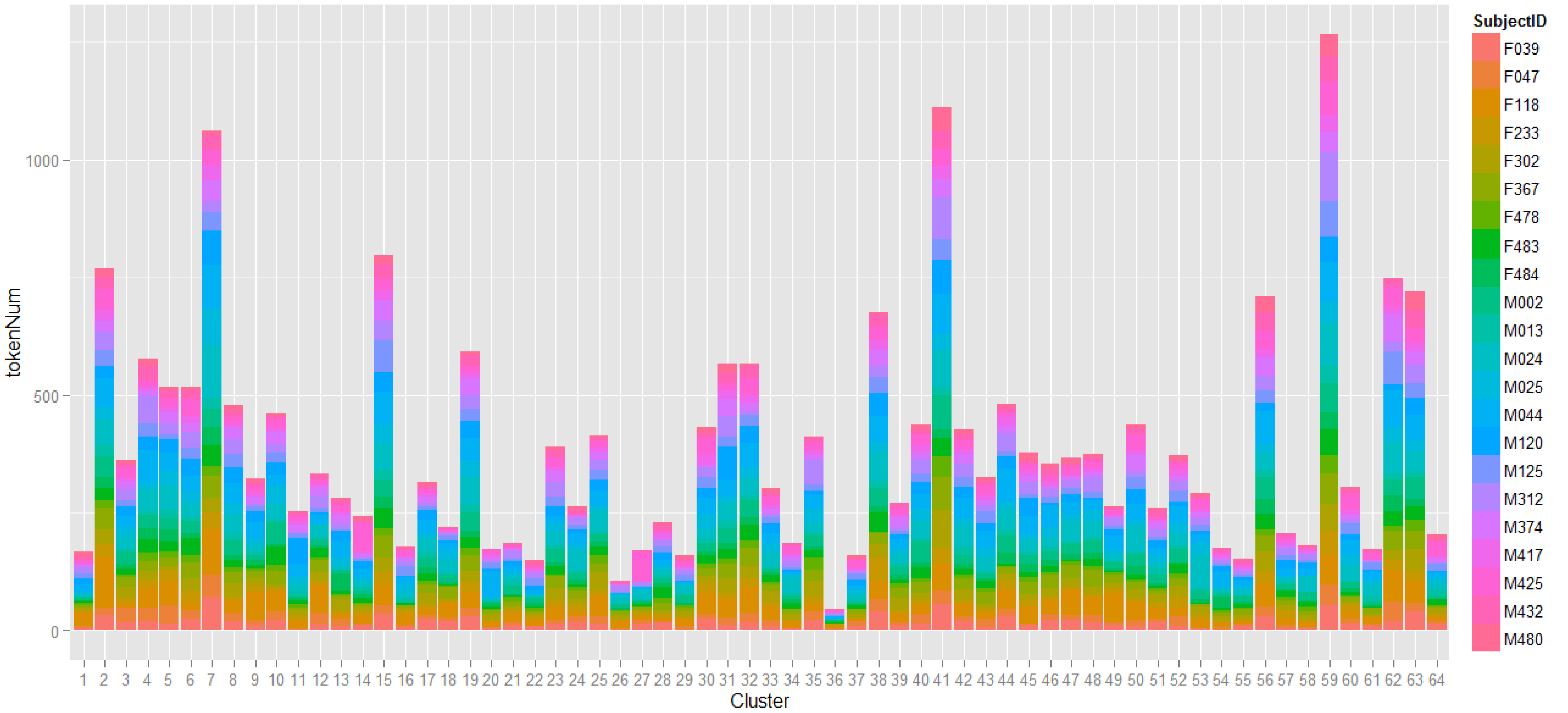
Gaussian mixture models (e.g., Vallabha, McClelland, Pons, Werker, & Amano, 2007; McMurray, Aslin, & Toscano, 2009)



CDS versus ADS



Speaker Variability



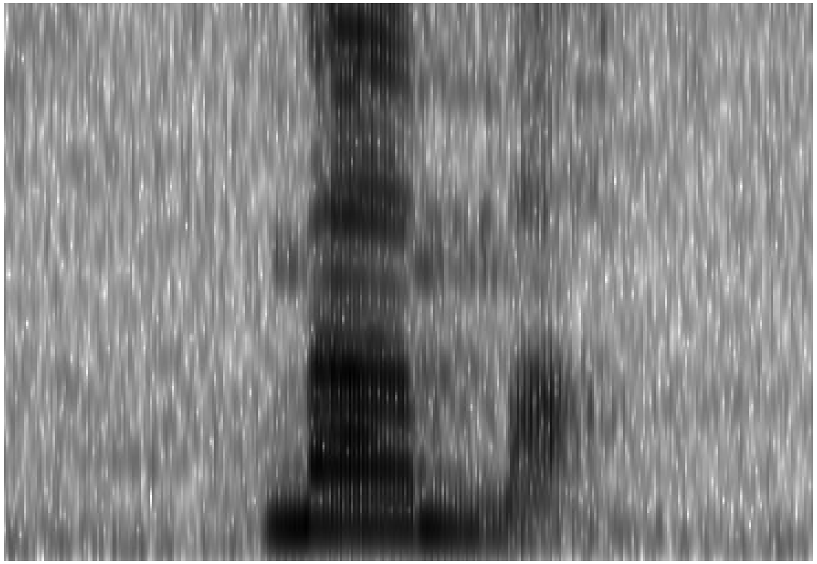
Summary of results

	Phonetic F-score				Lexical F-score			
	English		Japanese		English		Japanese	
	AD	CD	AD	CD	AD	CD	AD	CD
Corpus 1	0.78	0.80	0.96	0.98	0.96	0.94	0.98	0.98
Corpus 2	0.46	-	0.95	0.95	0.63	-	0.97	0.99
Corpus 3	0.13	-	0.24	0.22	0.41	-	0.59	0.61

Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

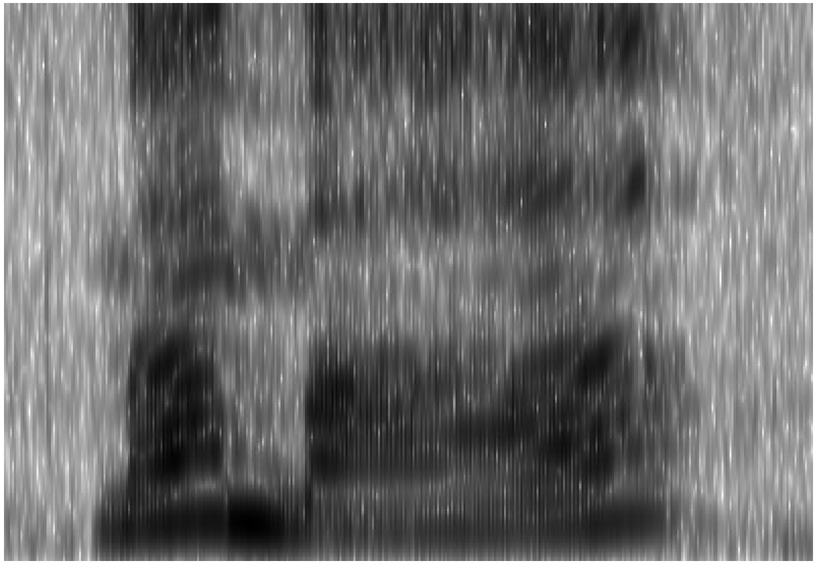
mama



Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

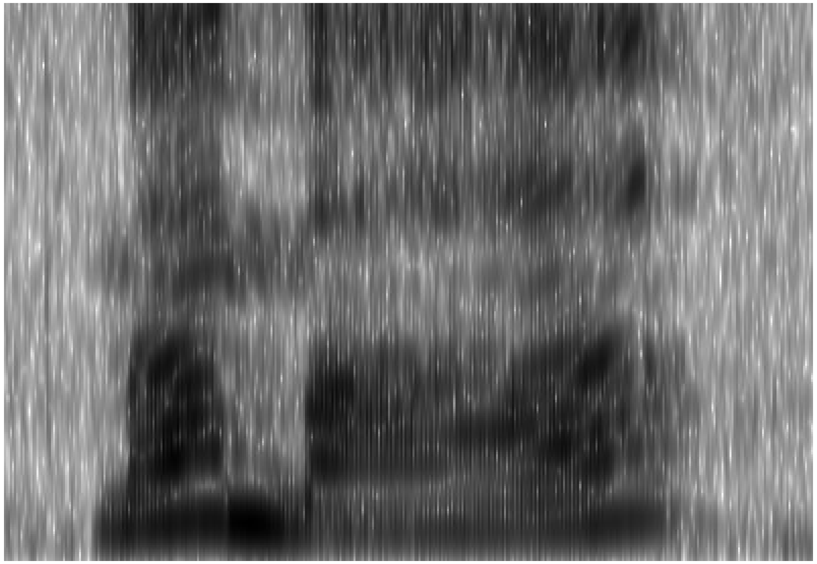
mama:



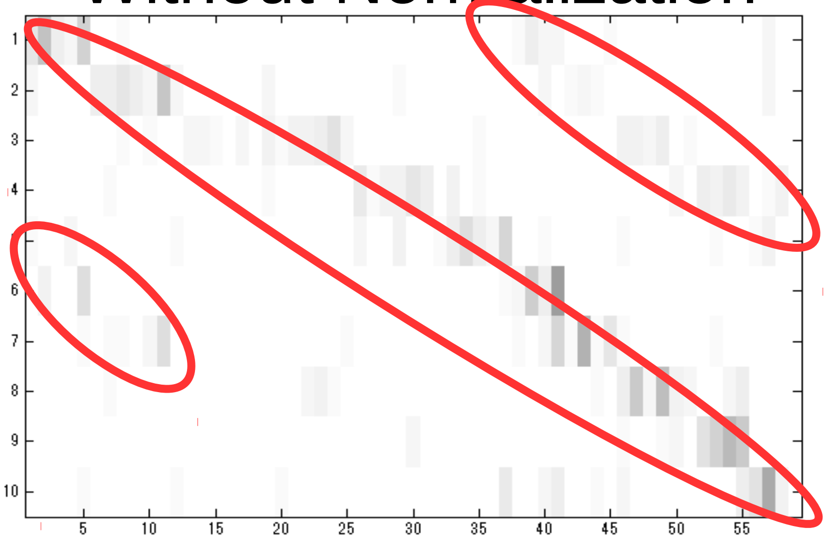
Japanese phrase-final lengthening

- Japanese drop in performance potentially partly due to phrase-final lengthening affecting vowel durations

mama:



Without Normalization



With Normalization