The Speaker-Affect Model

Measuring Emotion
in Political Speech
with Audio Data

17 March 2017

Dean Knox MIT
Christopher Lucas Harvard

Does emotion matter in politics?

Your browser does not support the video tag.

Probably.

How can we analyze it?

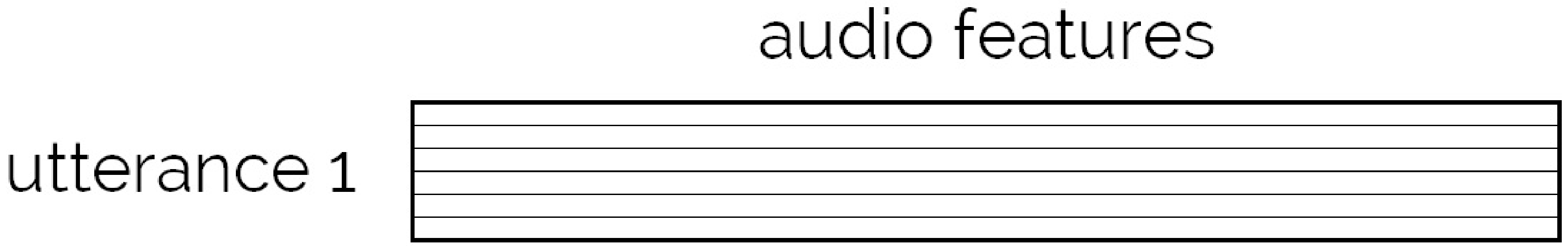
Roadmap

1. Intro: what is audio data?
2. Model: classifying audio with SAM
3. Dataset: the Supreme Court audio corpus
4. Results:
    - Benchmark against currently available audio methods
    - Compare to text-only approach for emotion detection

What is audio data?

What is audio data?

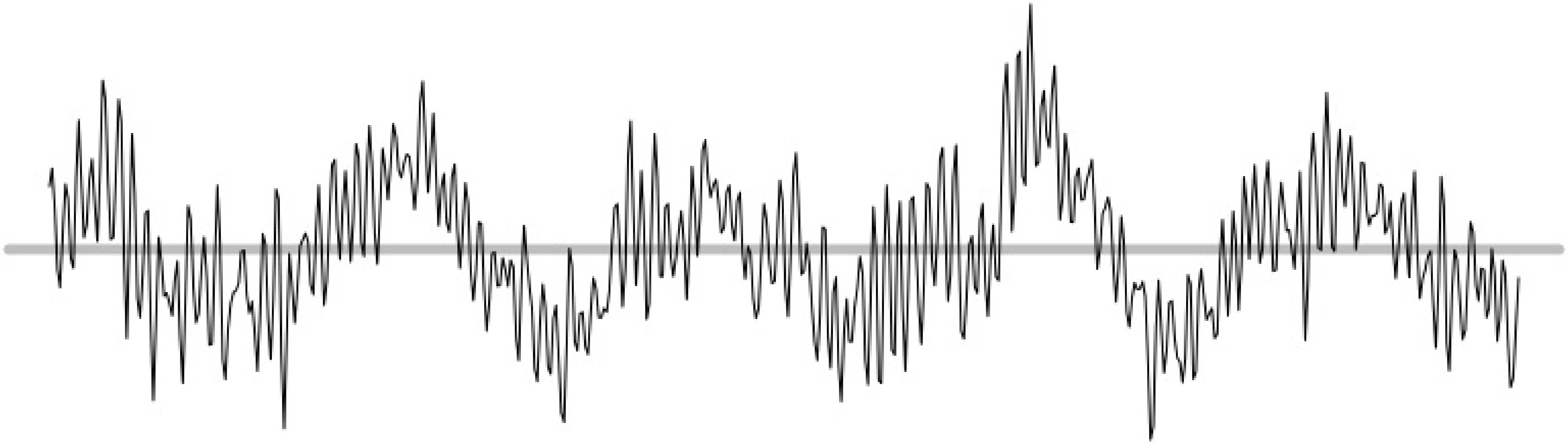
What is audio data?

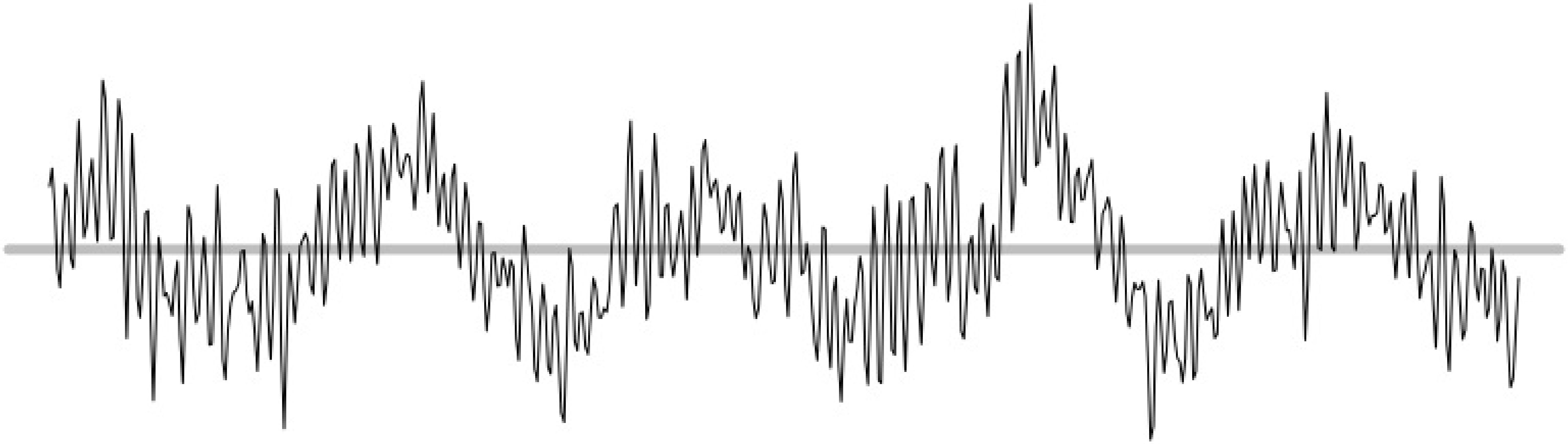


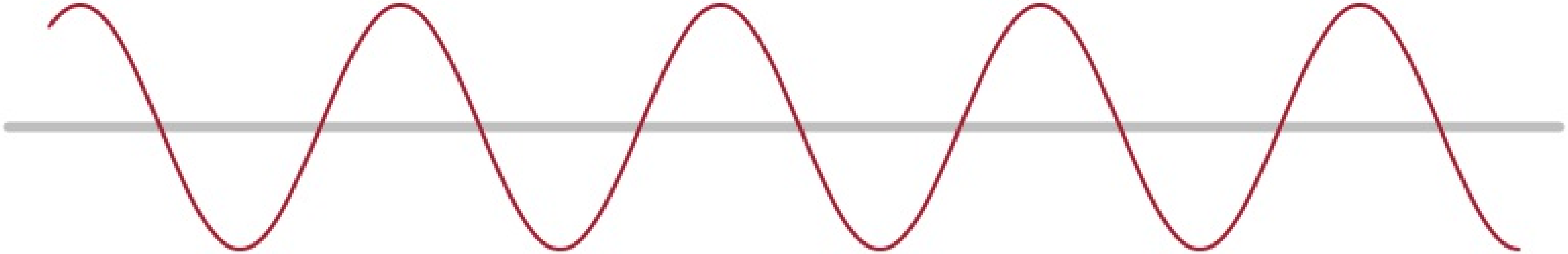
audio features

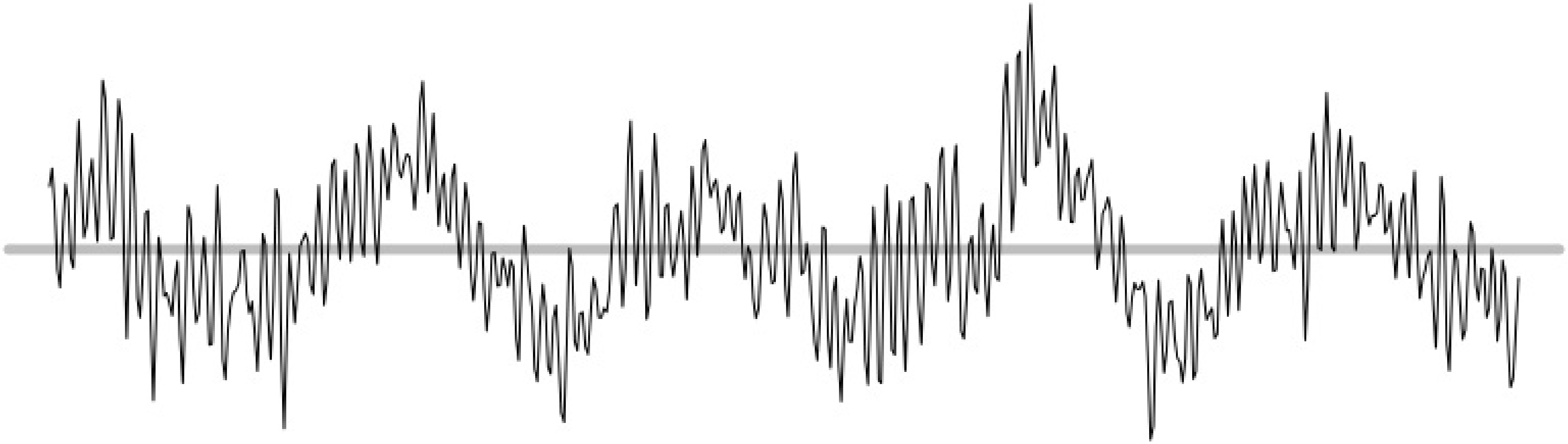
utterance 1

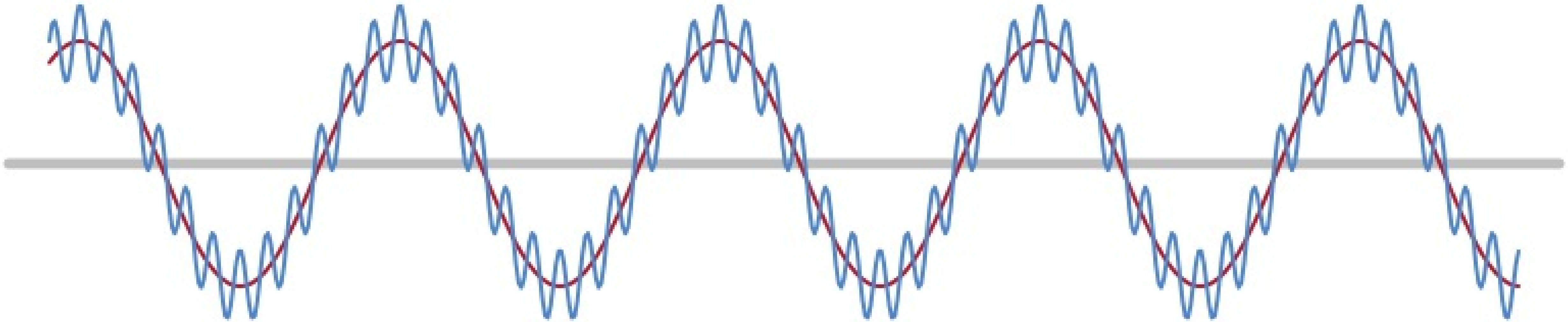
(not to scale)

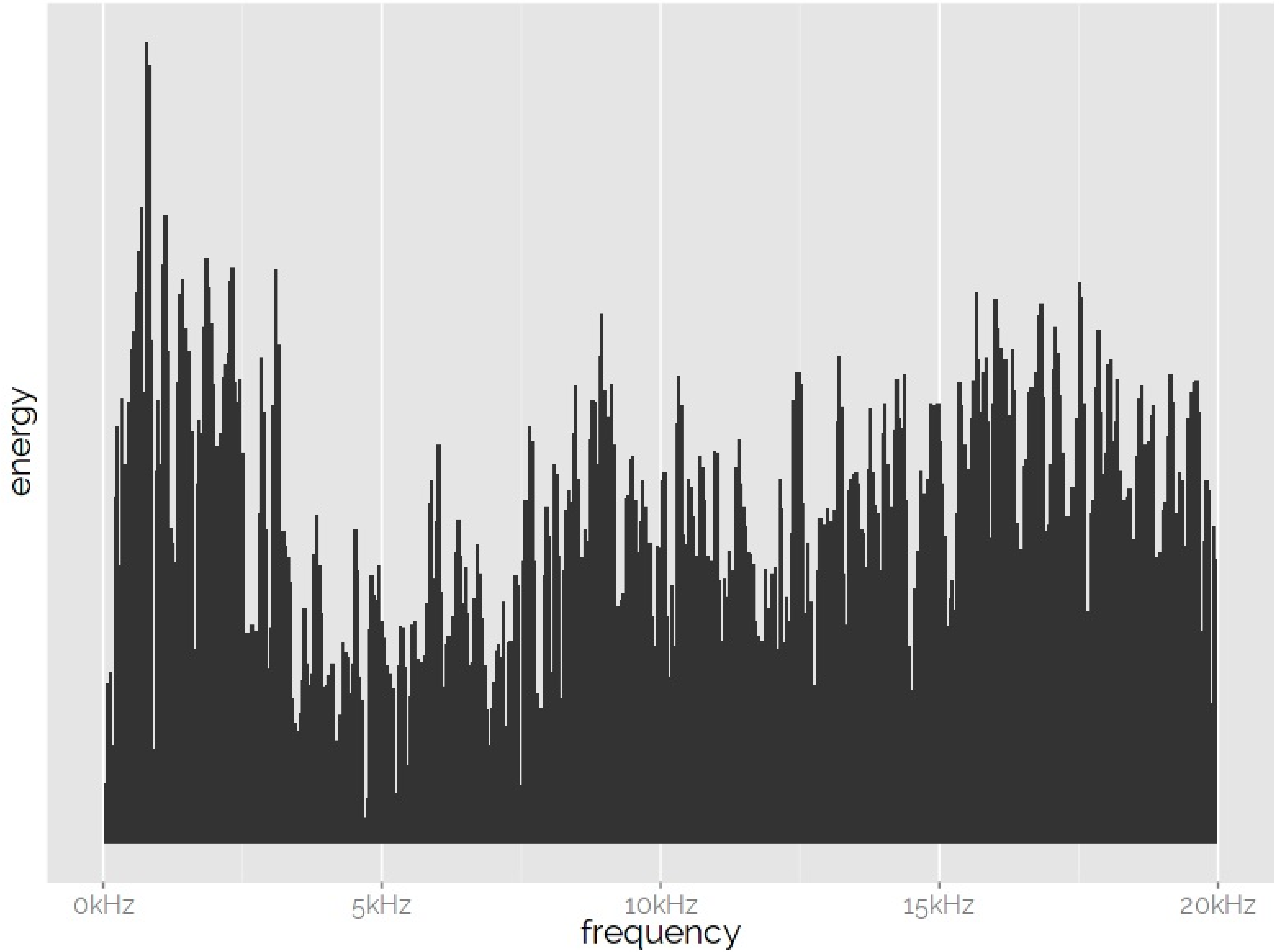




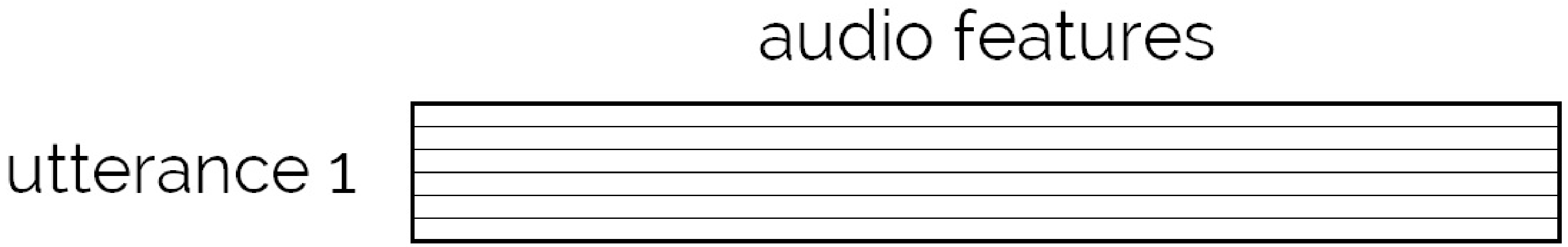







energy
0kHz
5kHz
10kHz
15kHz
20kHz
frequency

What is audio data?

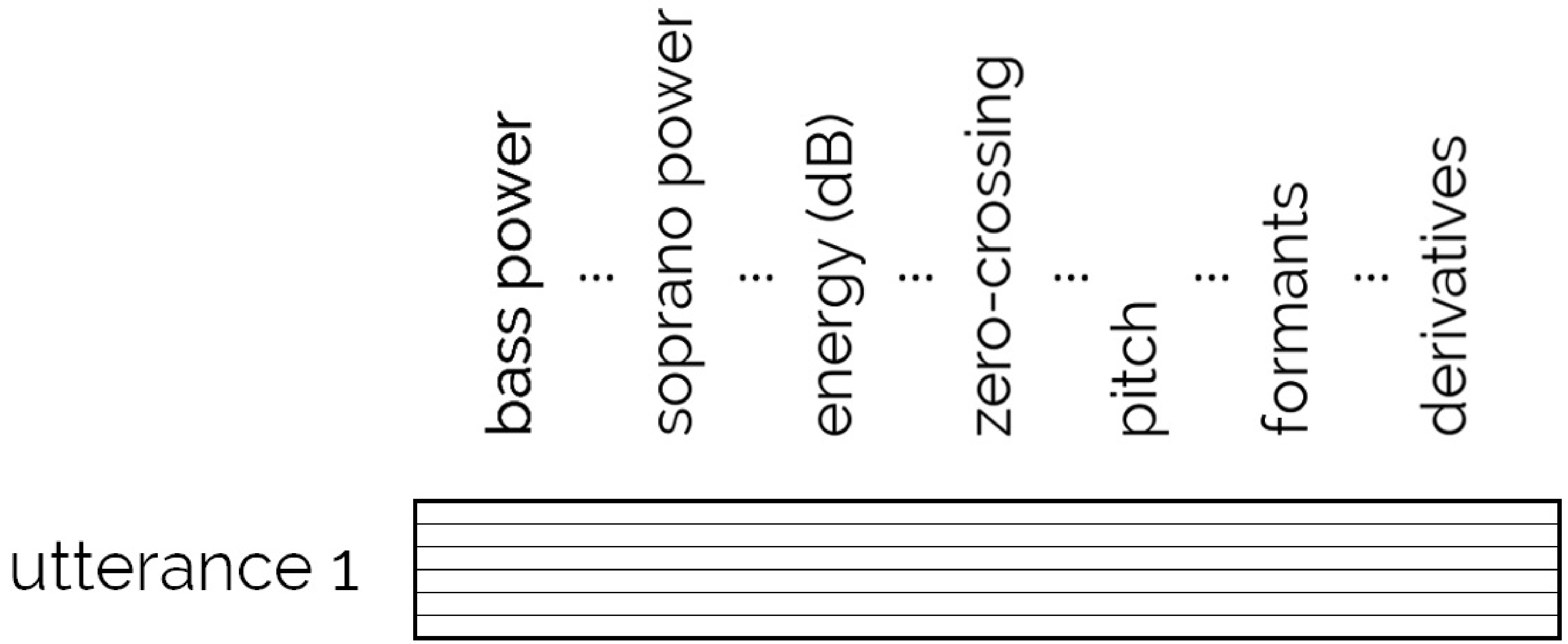


audio features

utterance 1

(not to scale)

What is audio data?

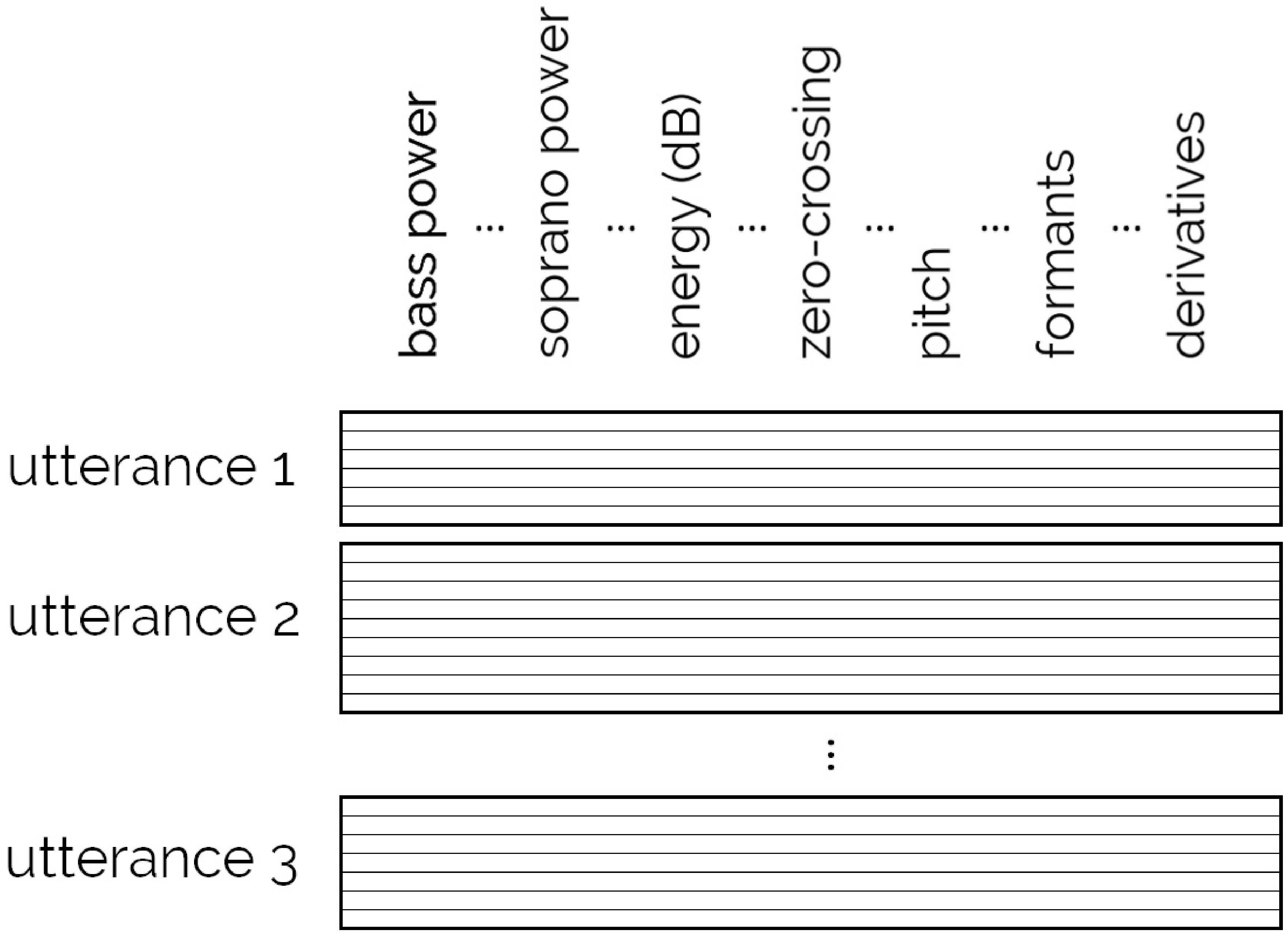


bass power … soprano power … energy (dB) … zero-crossing … pitch … formants … derivatives

utterance 1

(not to scale)

What is audio data?



bass power … soprano power … energy (dB) … zero-crossing … pitch … formants … derivatives

utterance 1

utterance 2

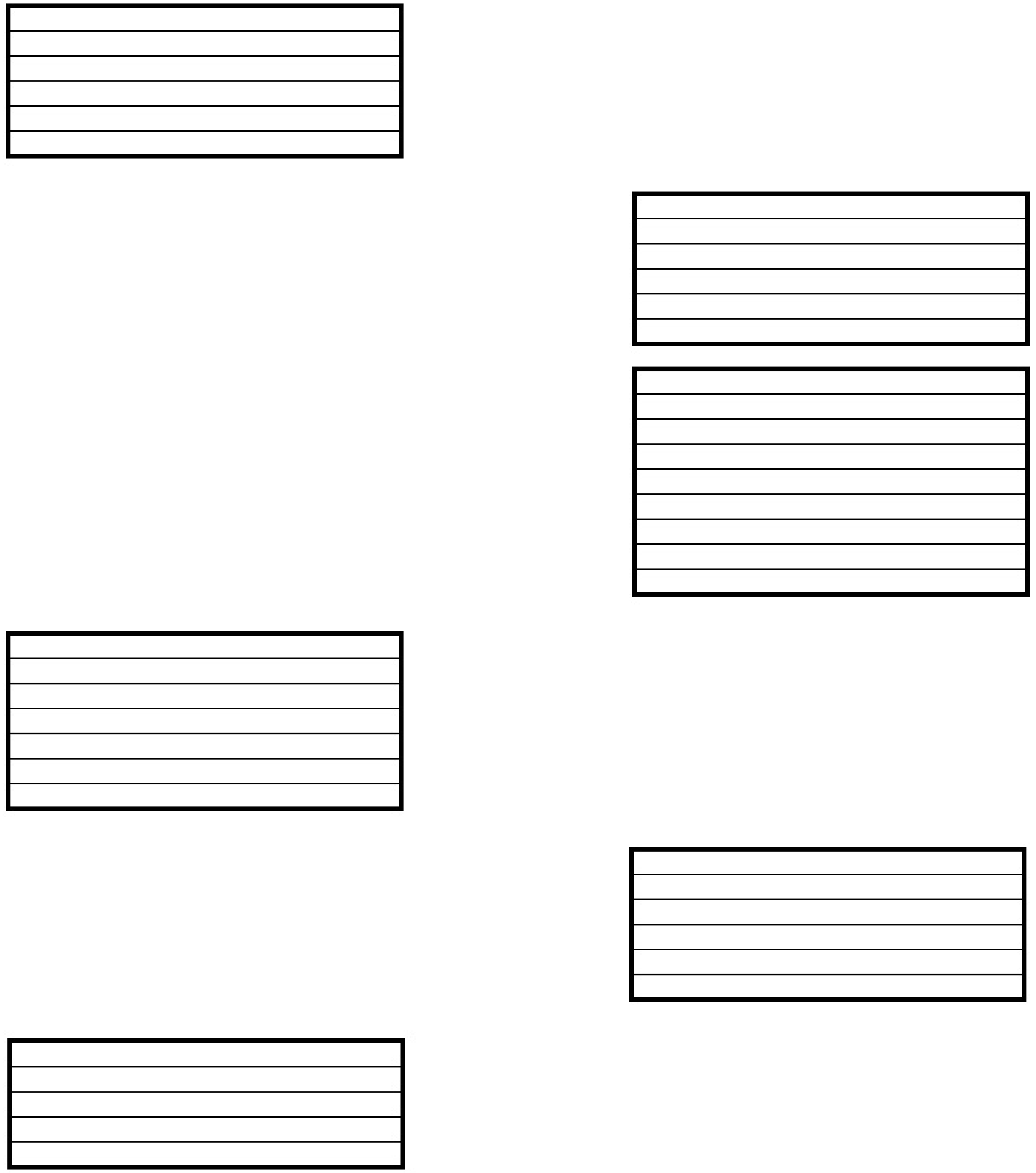
⋮

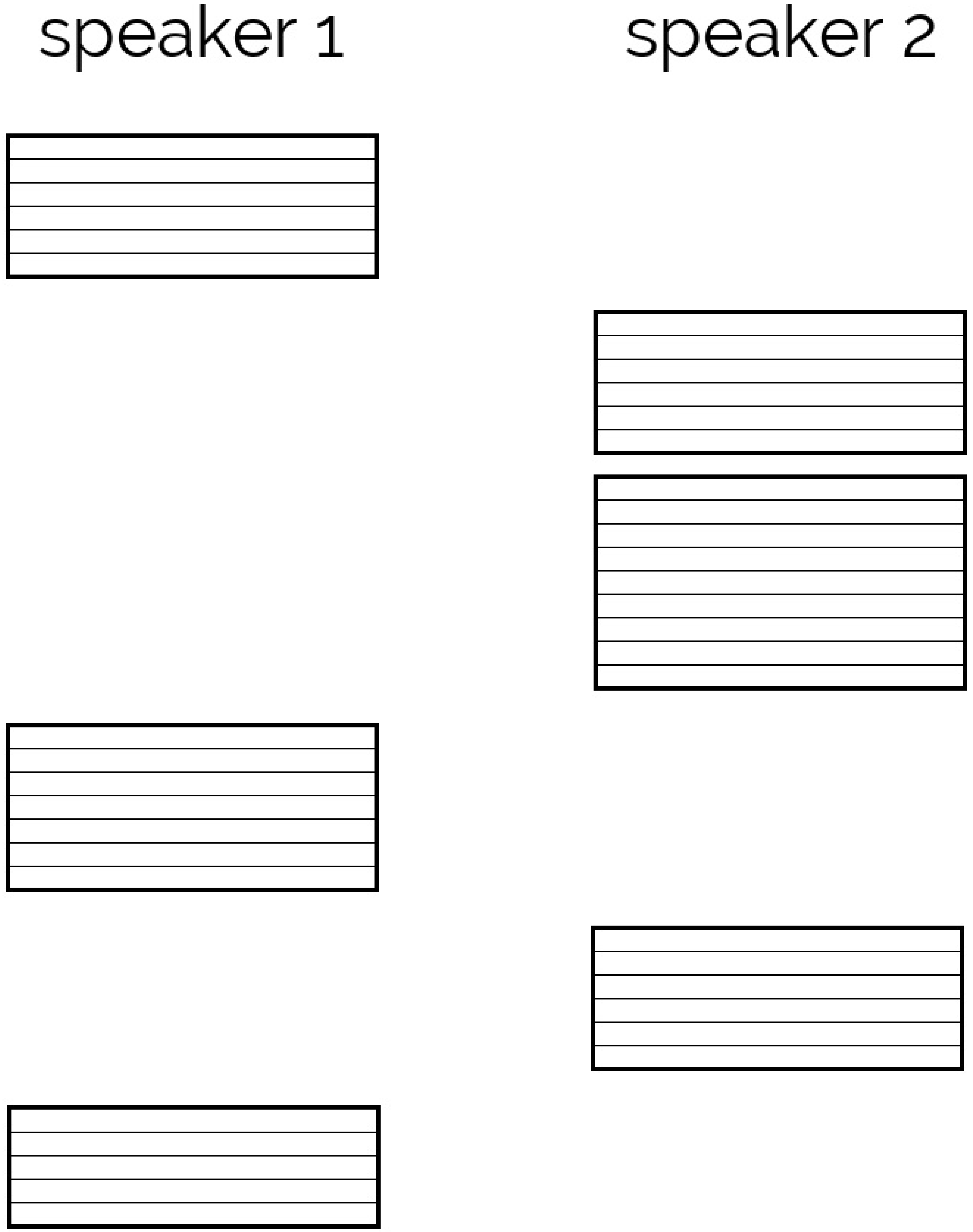
utterance 3

(not to scale)

mode 1

mode 2





speaker 1

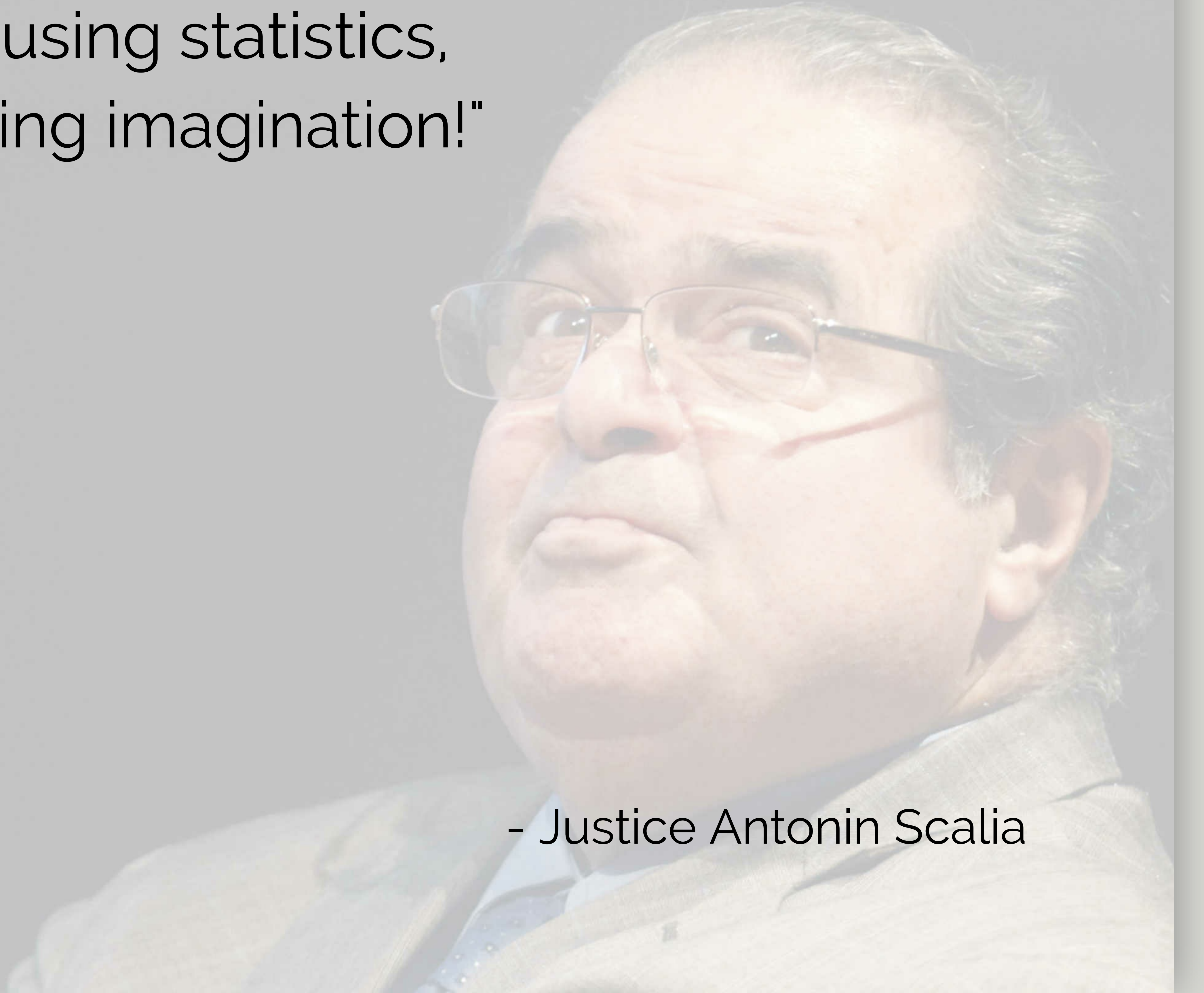
speaker 2

emotion 1

emotion 2

What is the Speaker Affect Model?

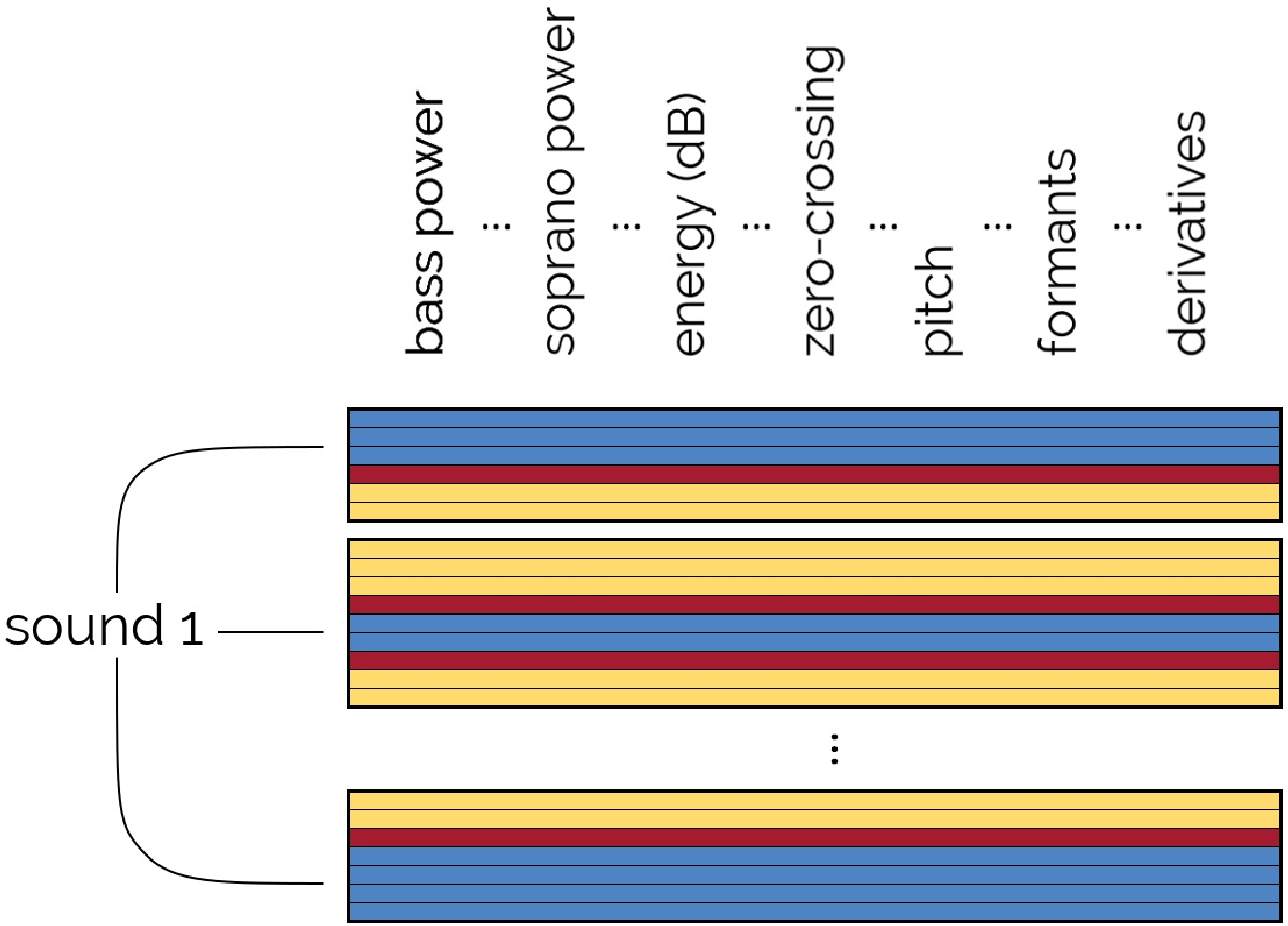
"It's not using statistics,
it's using imagination!"



- Justice Antonin Scalia

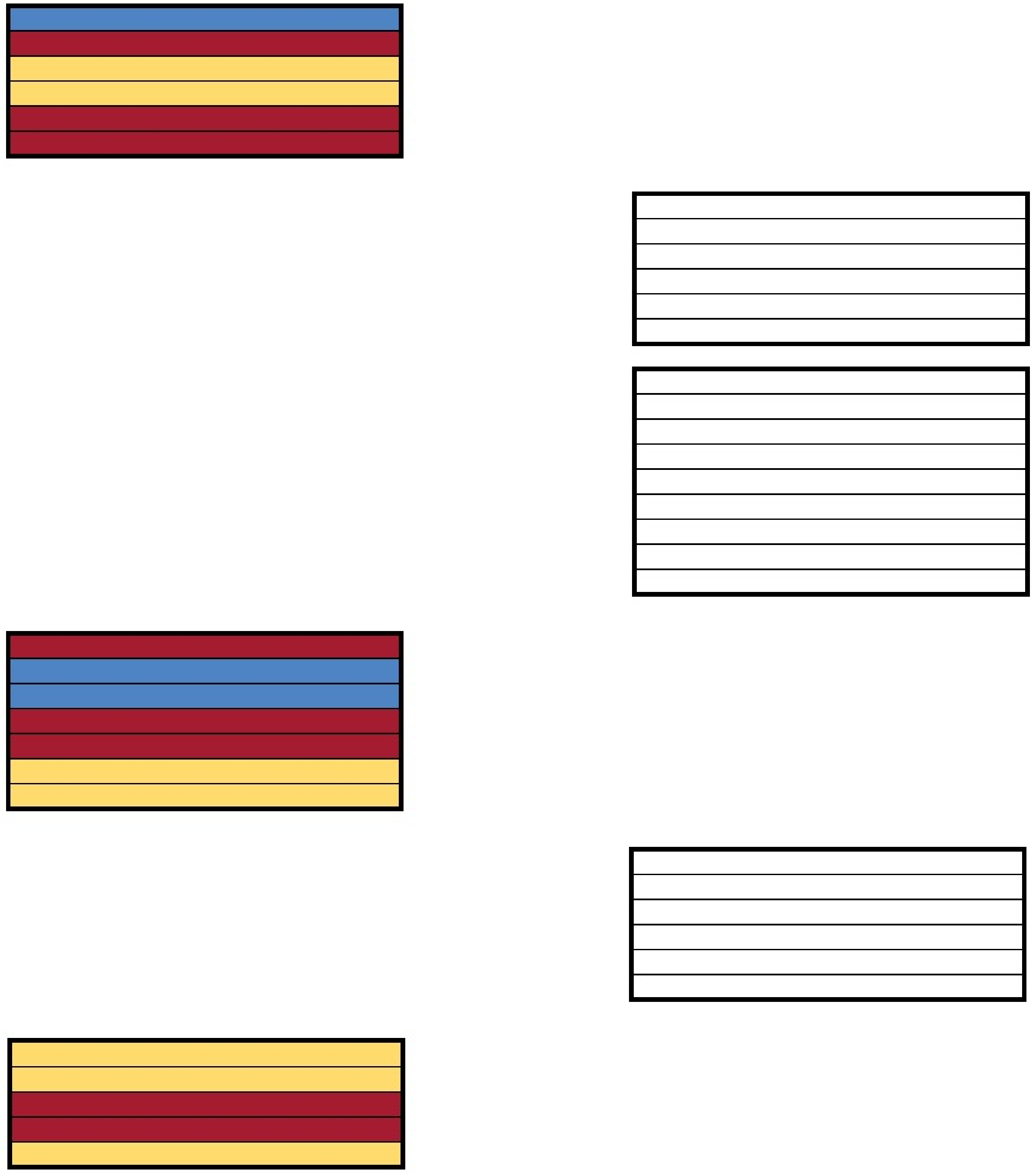
OK, it's using statistics.

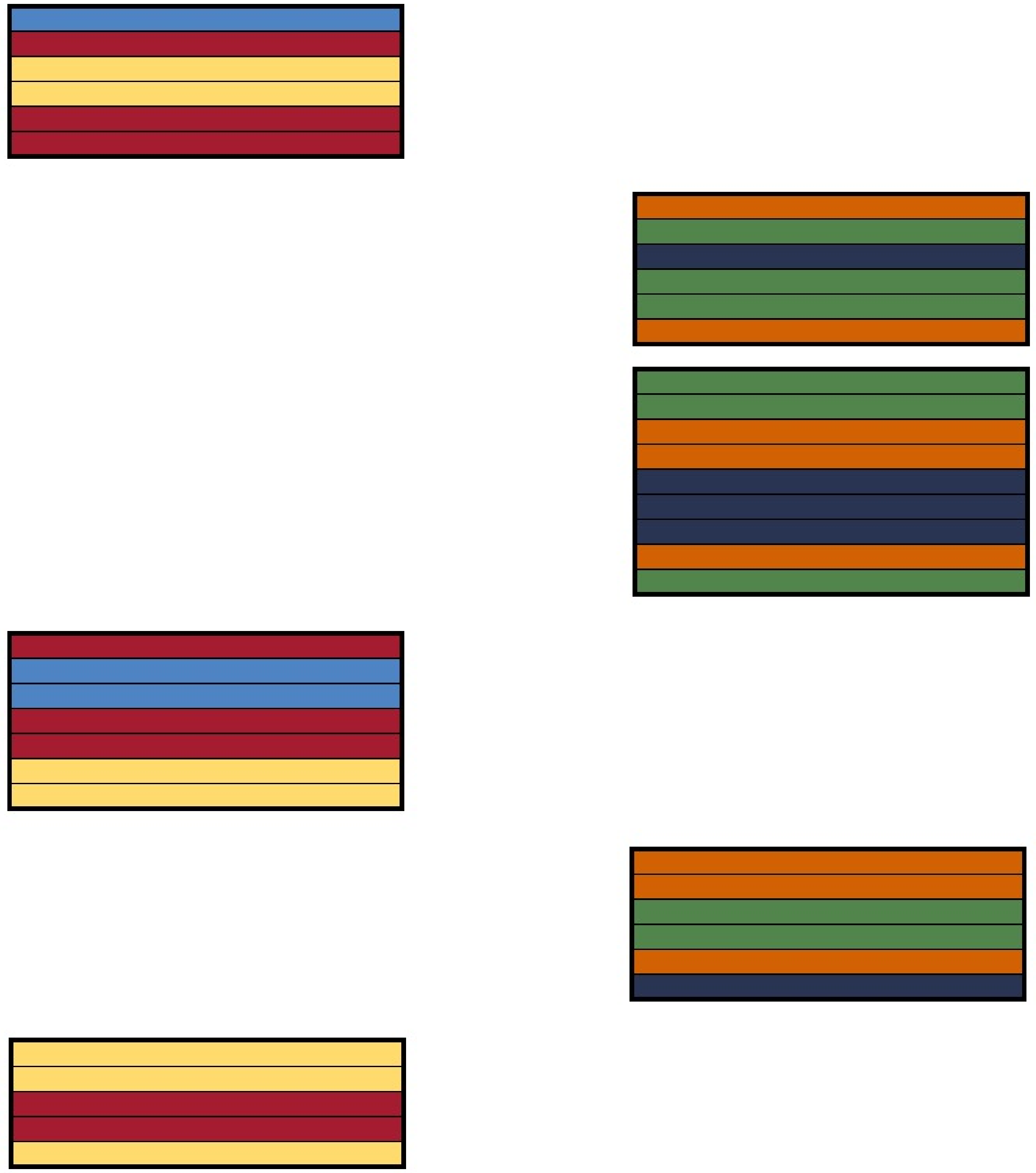
A model of speech



(not to scale)

emotion 1 emotion 2





emotion 1

emotion 2

Notation

Notation

- Speech is sequence of "utterances" $(u = 1, 2, \cdots)$

Notation

- Speech is sequence of "utterances" $(u = 1, 2, \cdots)$
    - each utterance has a "mode of speech" (S_u)

Notation

- Speech is sequence of "utterances" $(u = 1, 2, \cdots)$
  - each utterance has a "mode of speech" (S_u)

- Utterance is sequence of "frames" $(t = 1, 2, \cdots)$

Notation

- Speech is sequence of "utterances" $(u = 1, 2, \cdots)$
  - each utterance has a "mode of speech" (S_u)

- Utterance is sequence of "frames" $(t = 1, 2, \cdots)$
  - in each frame, a sound is being pronounced $(R_{u,t})$

Notation

- Speech is sequence of "utterances" $(u = 1, 2, \cdots)$
  - each utterance has a "mode of speech" (S_u)

- Utterance is sequence of "frames" $(t = 1, 2, \cdots)$
  - in each frame, a sound is being pronounced $(R_{u,t})$

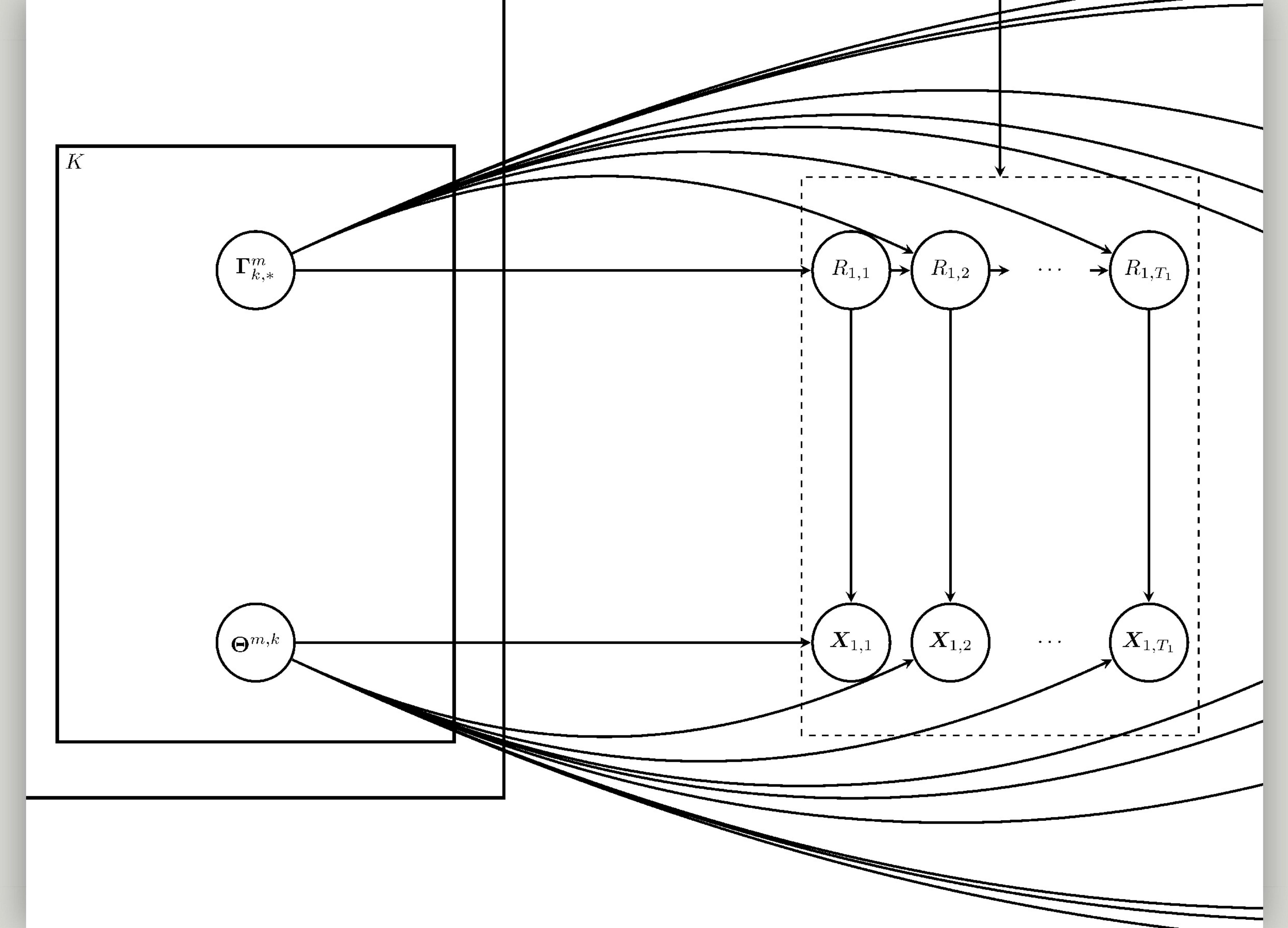
- Sound generates audio features $(\mathbf{X}_{u,t})$

Speaker Affect Model

$$\begin{aligned}
\text{mode of speech:} \quad & S_u \sim \text{Cat}\left(\boldsymbol{\Delta}_{S_{u-1},*}\right) \\
\text{sounds:} \quad & (R_{u,t} \mid S_u) \sim \text{Cat}\left(\boldsymbol{\Gamma}^{S_u}_{R_{u,t-1},*}\right) \\
\text{audio features:} \quad & (\mathbf{X}_{u,t} \mid S_u, R_{u,t}) \sim \ N\left(\mu_{S_u,R_{u,t}}, \boldsymbol{\Sigma}_{S_u,R_{u,t}}\right)
\end{aligned}$$

Δ : mode-of-speech transition matrix

Γ^m : sound transition matrix for mode-of-speech m


K
$\mathbf{\Gamma}^{m}_{k,*}$
$\mathbf{\Theta}^{m,k}$
$R_{1,1}$
$R_{1,2}$
$\cdots$
R_{1,T_1}
$\boldsymbol{X}_{1,1}$
$\boldsymbol{X}_{1,2}$
$\cdots$
$\boldsymbol{X}_{1,T_1}$

Your browser does not support the video tag.

Estimation: Single Mode

- Estimate by EM with forward-backward algorithm

Estimation: Single Mode

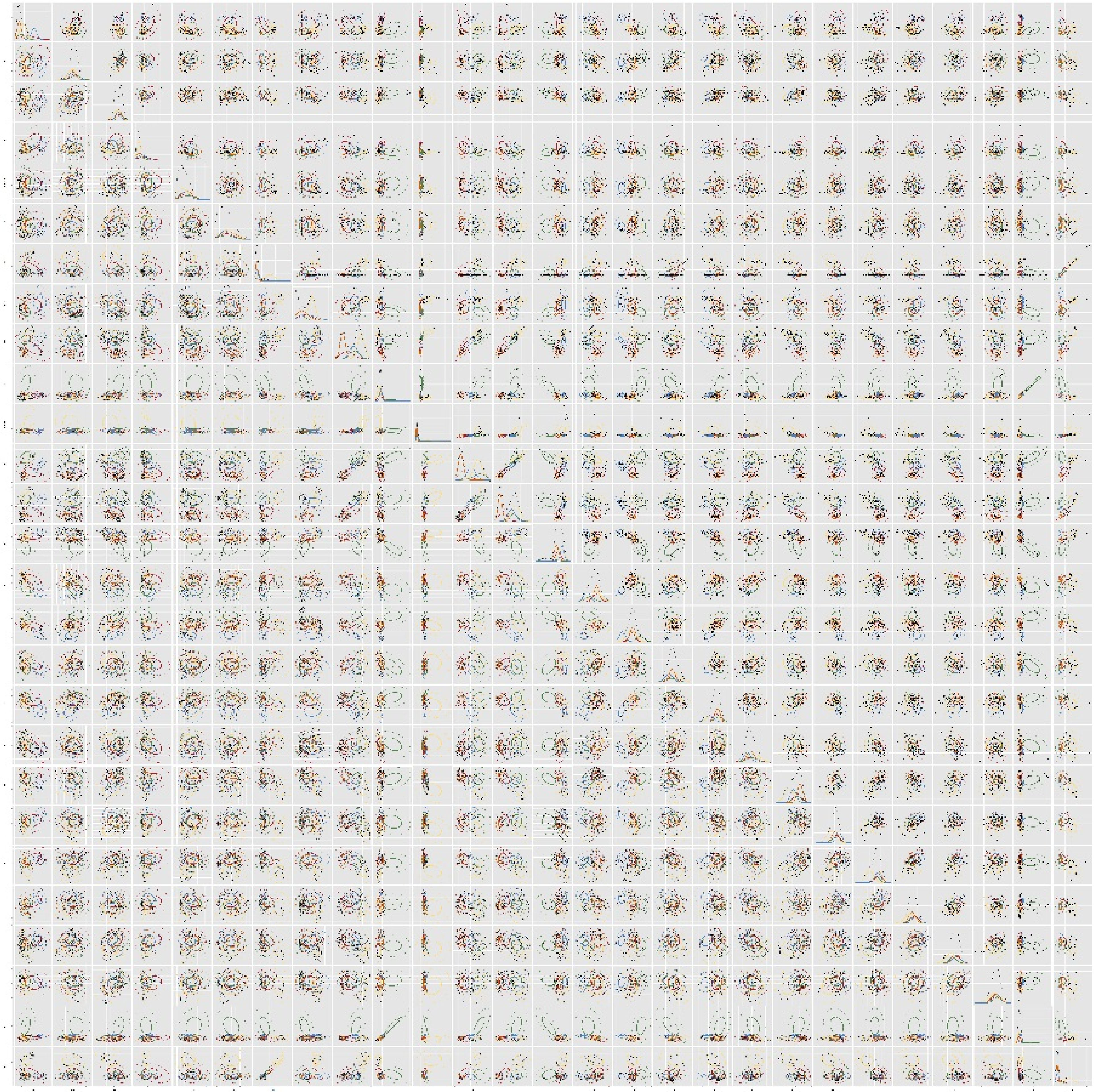
- Estimate by EM with forward-backward algorithm
- E-step
    - Expected emotion labels
    - Expected emotional transitions

Estimation: Single Mode

- Estimate by EM with forward-backward algorithm
- E-step
  - Expected emotion labels
  - Expected emotional transitions
- M-step
  - Sound distributions
  - Transition probabilities

Estimation: Single Mode

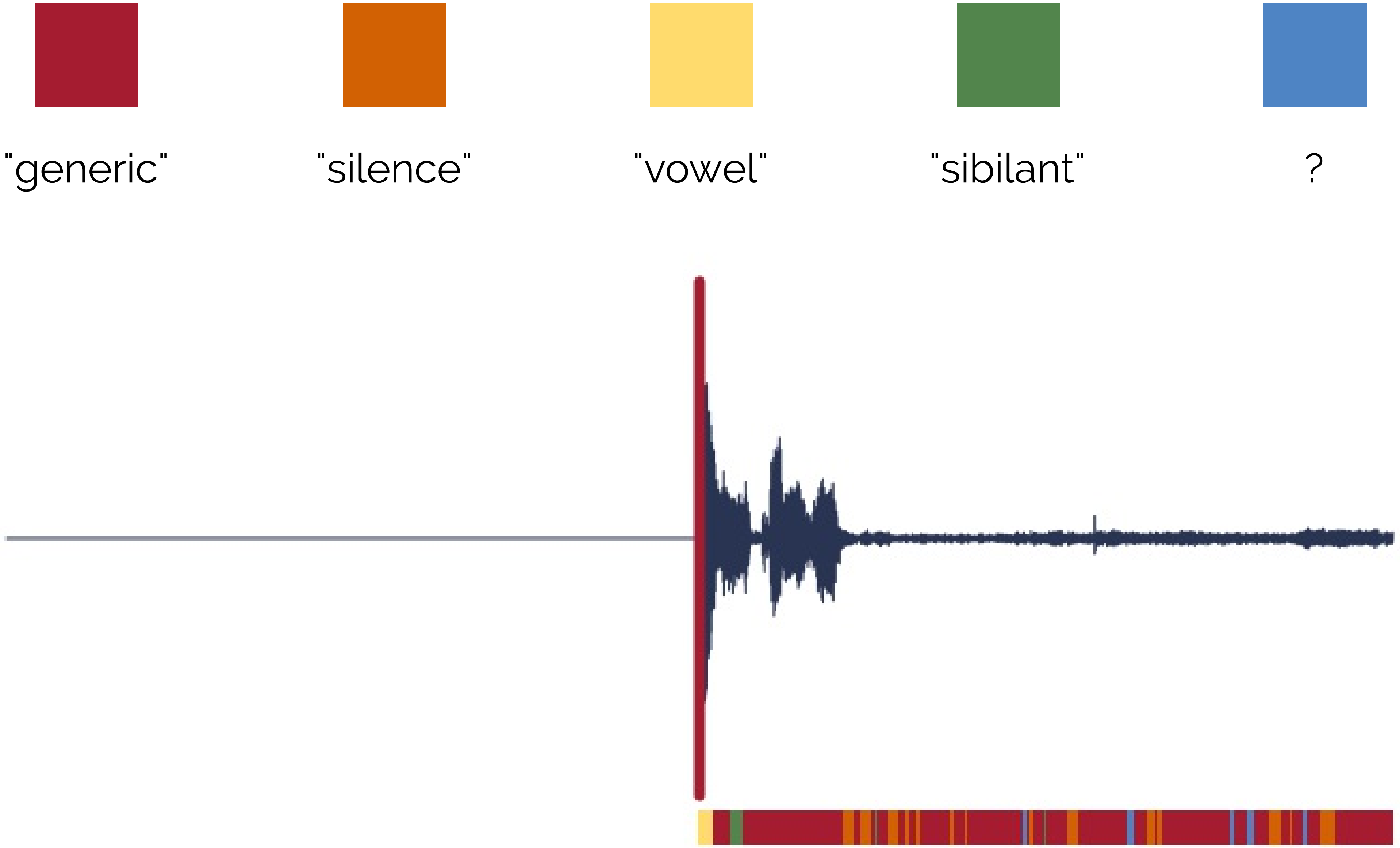
- Estimate by EM with forward-backward algorithm
- E-step
  - Expected emotion labels
  - Expected emotional transitions
- M-step
  - Sound distributions
  - Transition probabilities
- Rcpp implementation in our package, SAM (alpha)



Your browser does not support the video tag.

Your browser does not support the video tag.

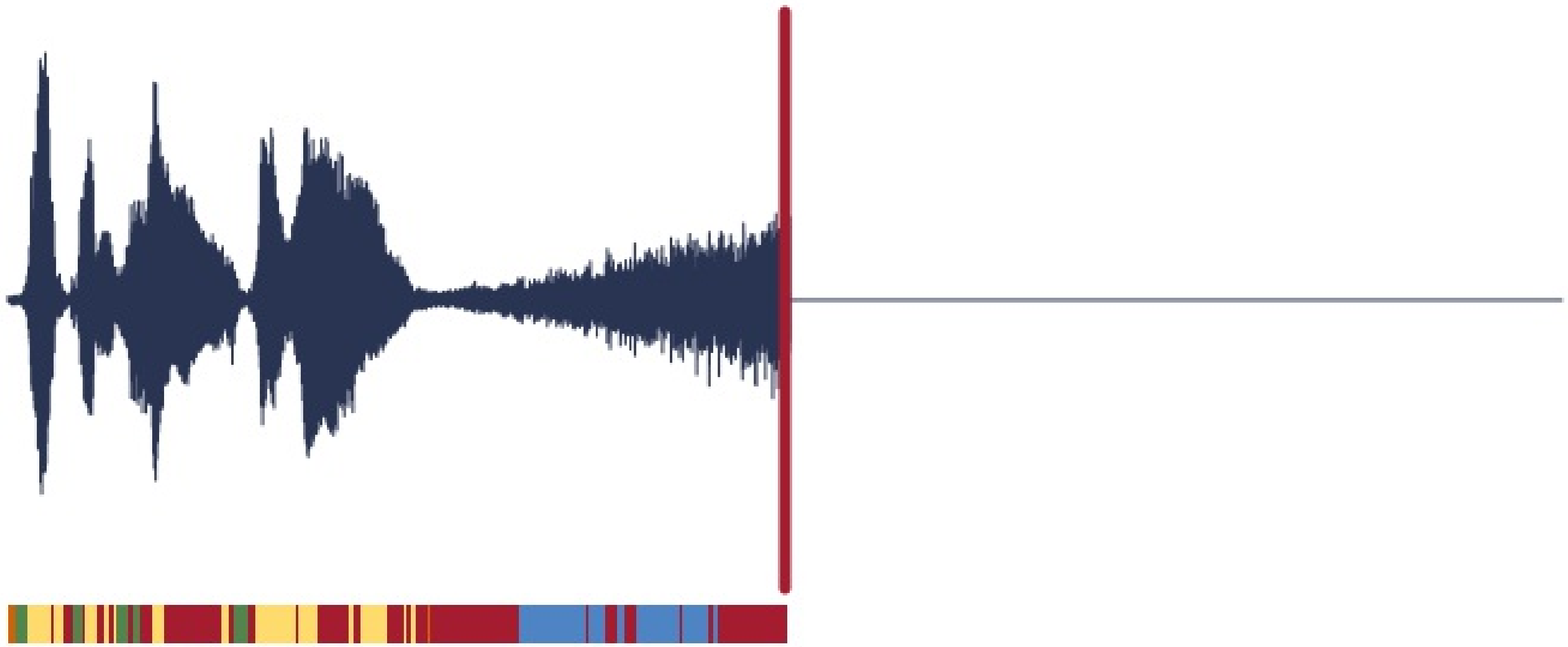
🟥	high variance	"generic"
🟧	low intensity	"silence"
🟨	loud, mid-range 1st formant	"vowel"
🟩	high zero-crossing rate	"sibilant"
🟦	high resonance	?

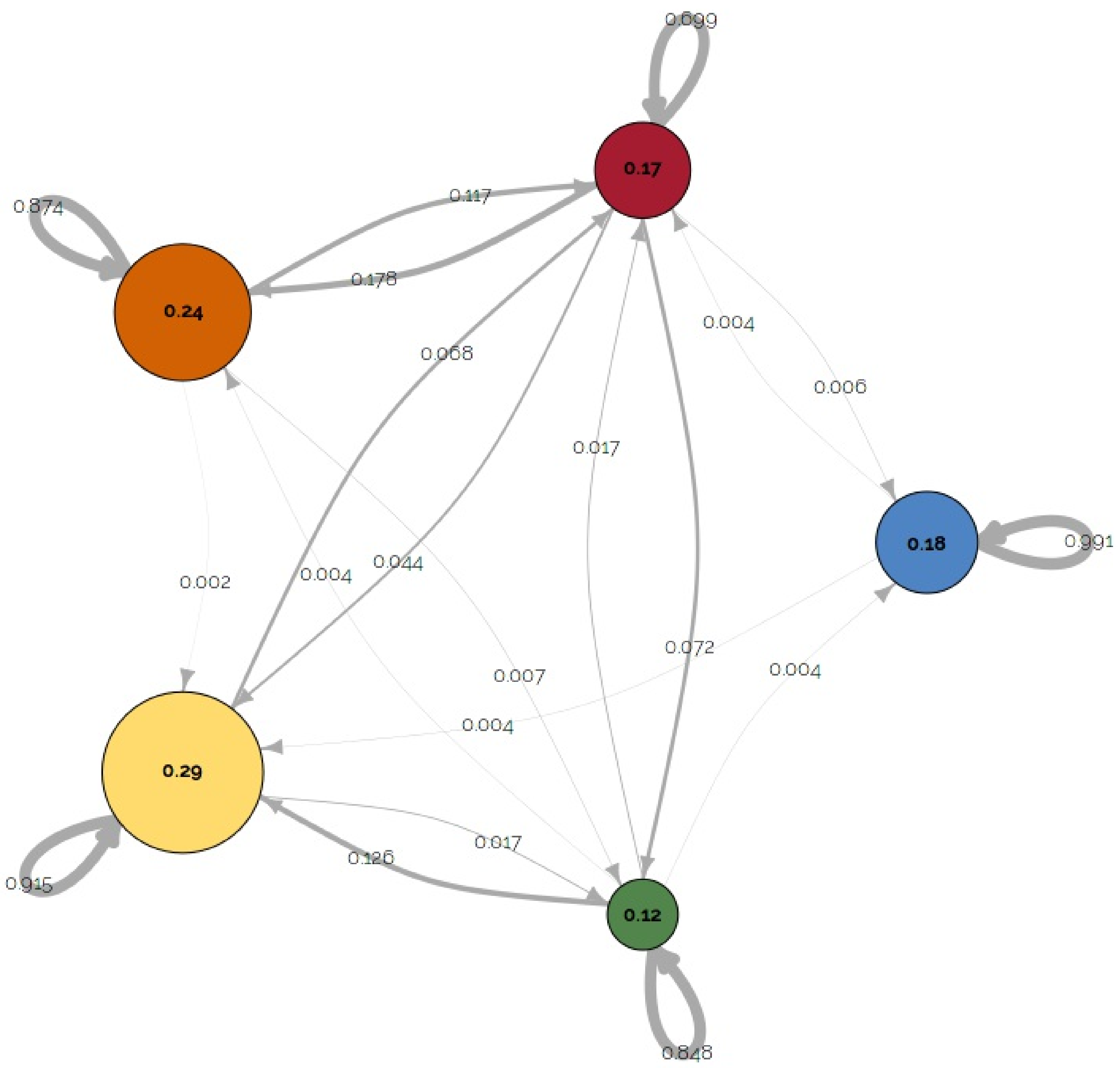

🟥 "generic"
🟧 "silence"
🟨 "vowel"
🟩 "sibilant"
🟦 ?

🟥	🟧	🟨	🟩	🟦
"generic"	"silence"	"vowel"	"sibilant"	?

Your browser does not support the video tag.

🟥 "generic" 🟧 "silence" 🟨 "vowel" 🟩 "sibilant" 🟦 "audience"




0.699
0.17
0.874
0.117
0.178
0.24
0.004
0.068
0.006
0.017
0.18
0.991
0.002
0.004
0.044
0.072
0.007
0.004
0.004
0.29
0.017
0.126
0.915
0.12
0.848

Estimation: Multiple Modes

1. Experts determine speaking modes & rubric

Estimation: Multiple Modes

1. Experts determine speaking modes & rubric

2. Humans code "speaking mode" for training set

Estimation: Multiple Modes

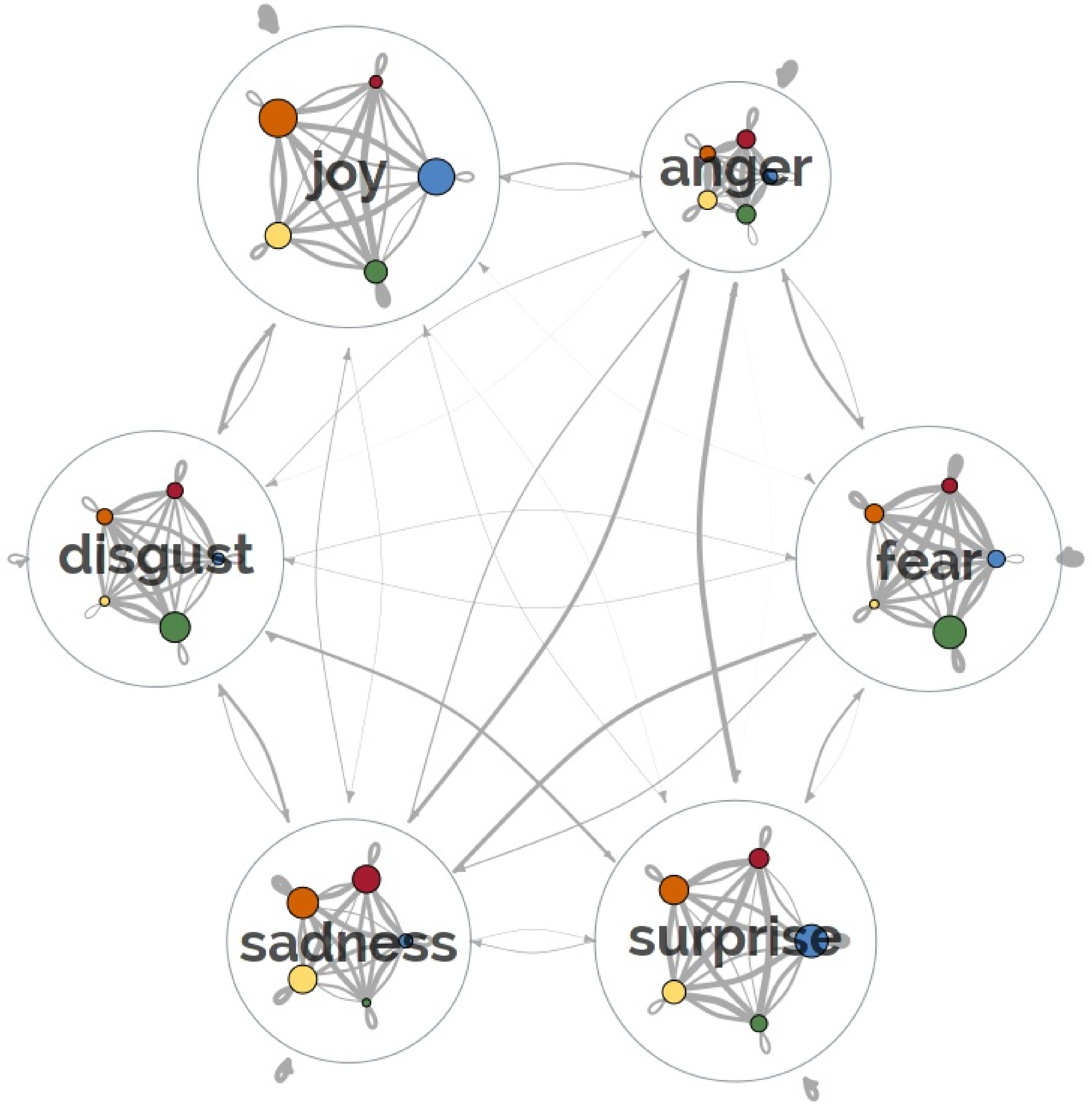
1. Experts determine speaking modes & rubric

2. Humans code "speaking mode" for training set

3. Unsupervised HMM for each speaking mode
    - Automatically classify sounds, estimate content/usage

Estimation: Multiple Modes

1. Experts determine speaking modes & rubric
2. Humans code "speaking mode" for training set
3. Unsupervised HMM for each speaking mode
   - Automatically classify sounds, estimate content/usage
4. Supervised HMM for changes in mode of speech (estimate flow of speech)
   - Usage of different speaking modes
   - How speaking modes change over course of speech
   - Interplay in speaking modes between people


joy
anger
disgust
fear
sadness
surprise

Supreme Court Audio Corpus

Oral Arguments

- Supreme Court data from Oyez Project
  - 782 recordings from Roberts court, ~800 hours total
  - Timestamped transcripts with speaker labels
- Segment into 454k utterances
  - Pool lawyers together, analyze each justice separately
- Extract 81 features for each 25-millisec. window

Validating the Model with Supreme Court Data

An Easy Task: Speaker ID

- Distinguish between 11 coarse modes of speech:
  - Speech by Alito, speech by Breyer, …

- Practical application: deliberation experiments
  1. Record audio of deliberation in lab or field
  2. Have participants self-introduce at beginning
  3. Automatically generate transcript with transcribeR
  4. Learn a model of each participant's speech
  5. Use participant models to label the transcript

An Easy Task: Speaker ID

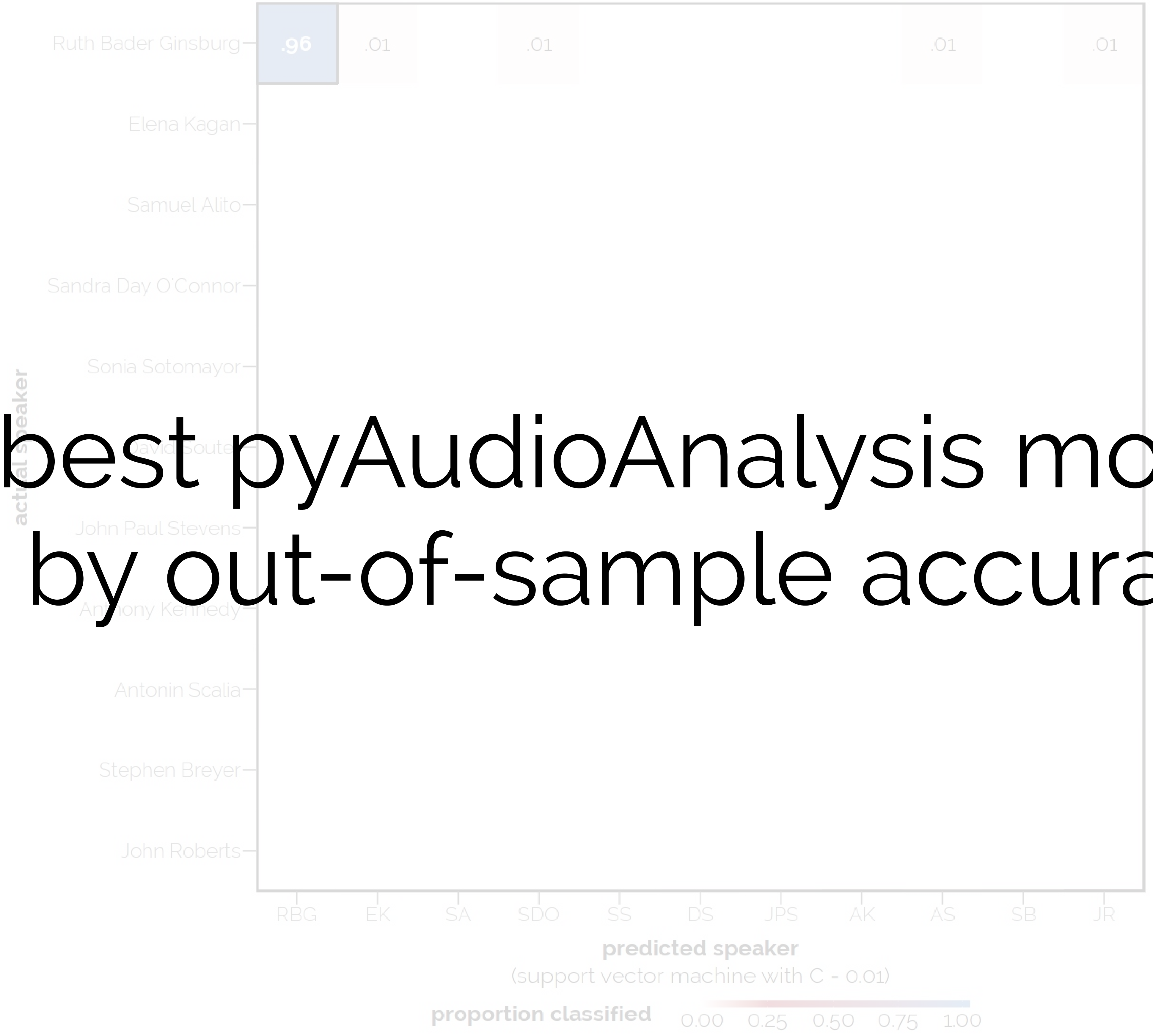
- Draw 100 utterances per justice (1100 total)

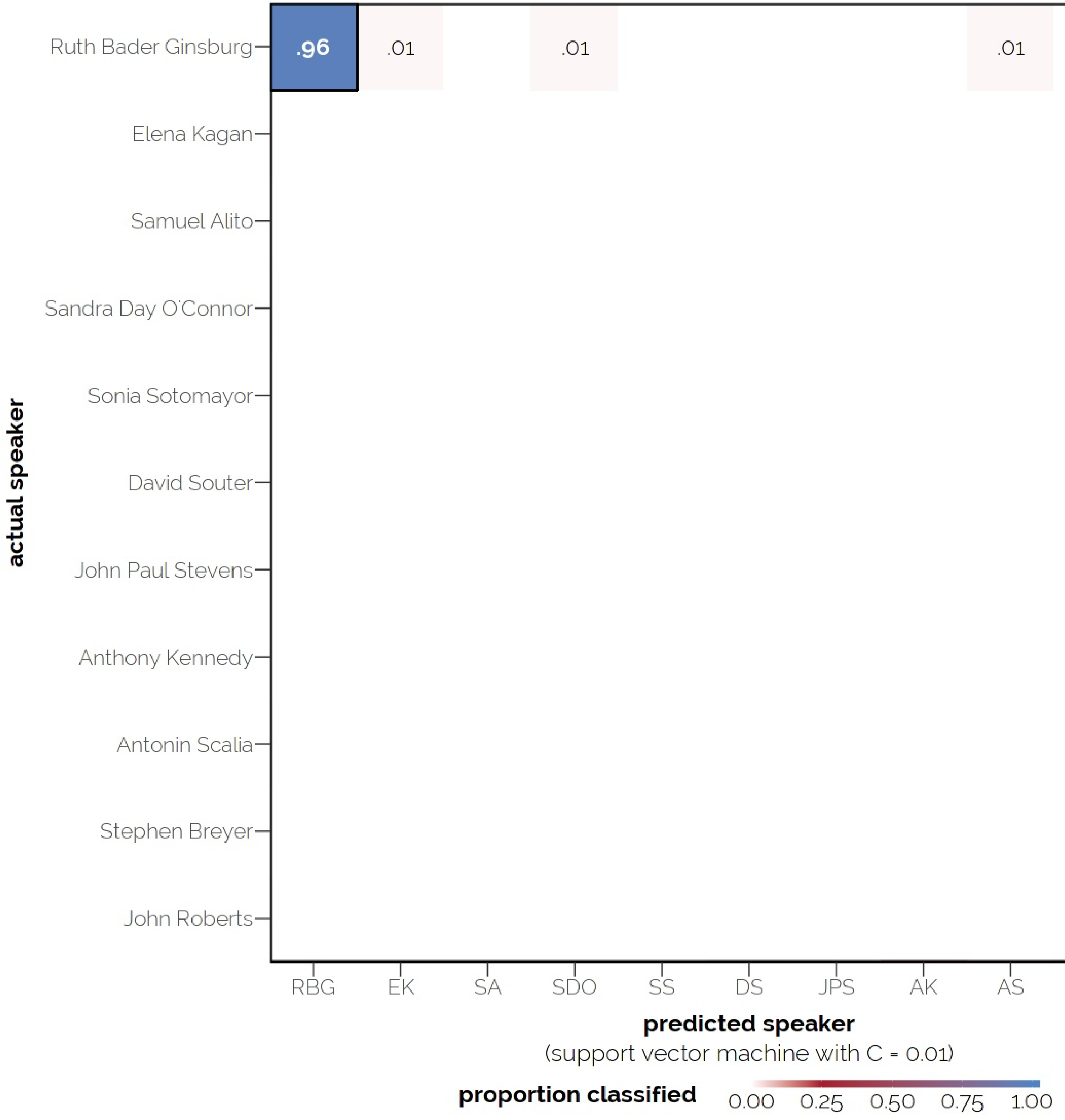
- Evaluate our model's out-of-sample predictive accuracy by K-fold cross-validation
  - Split the data into K balanced folds. For each fold:
  - Hold out the 1/K utterances from this fold for testing
  - Divide the remaining (K-1)/K utterances by speaker
  - For each speaker, train a speaker-specific HMM
  - Calculate log-lik. of held-out utterances under each model → predict speaker based on the most likely model

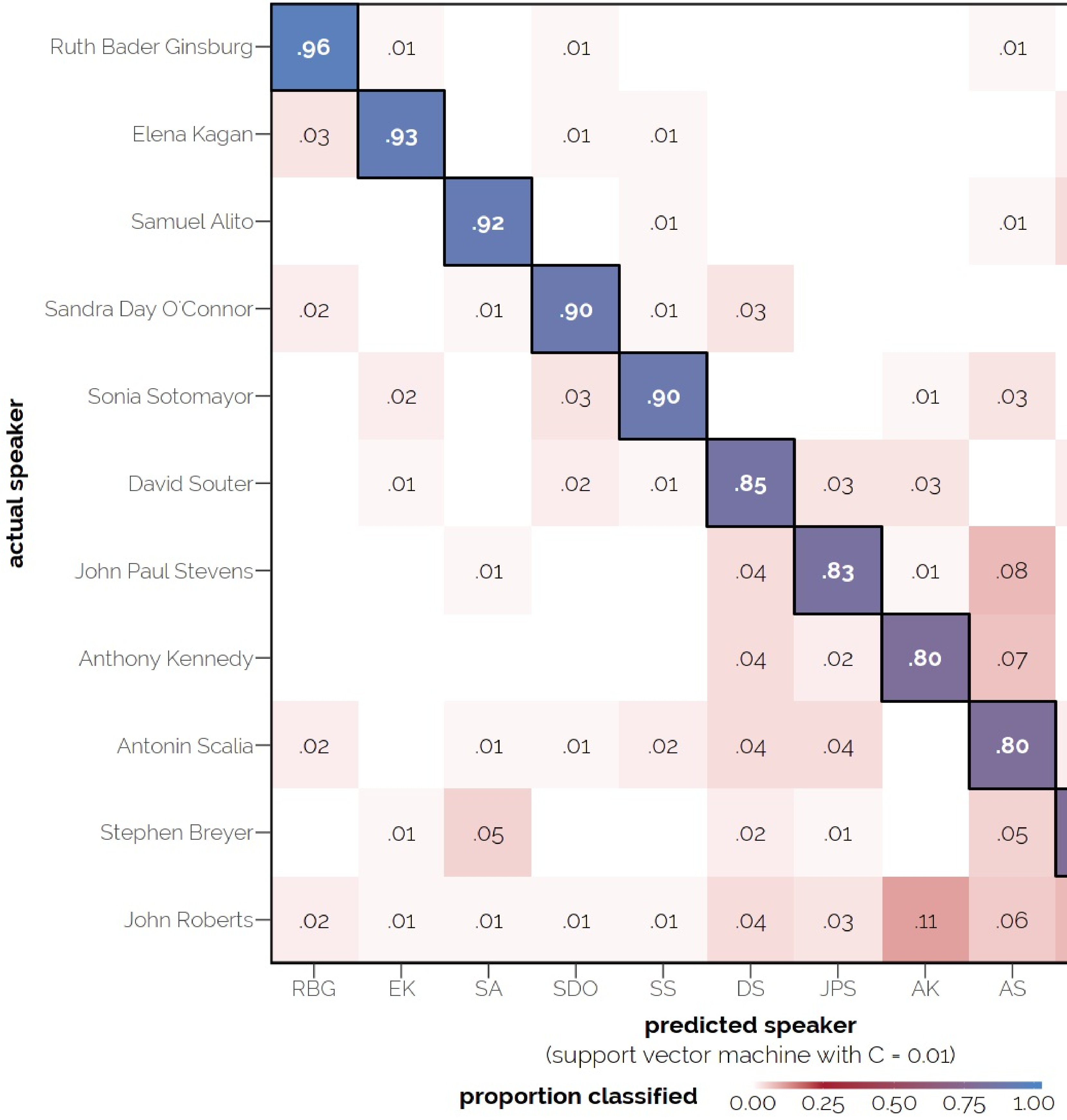
Audio Model Horse Race!

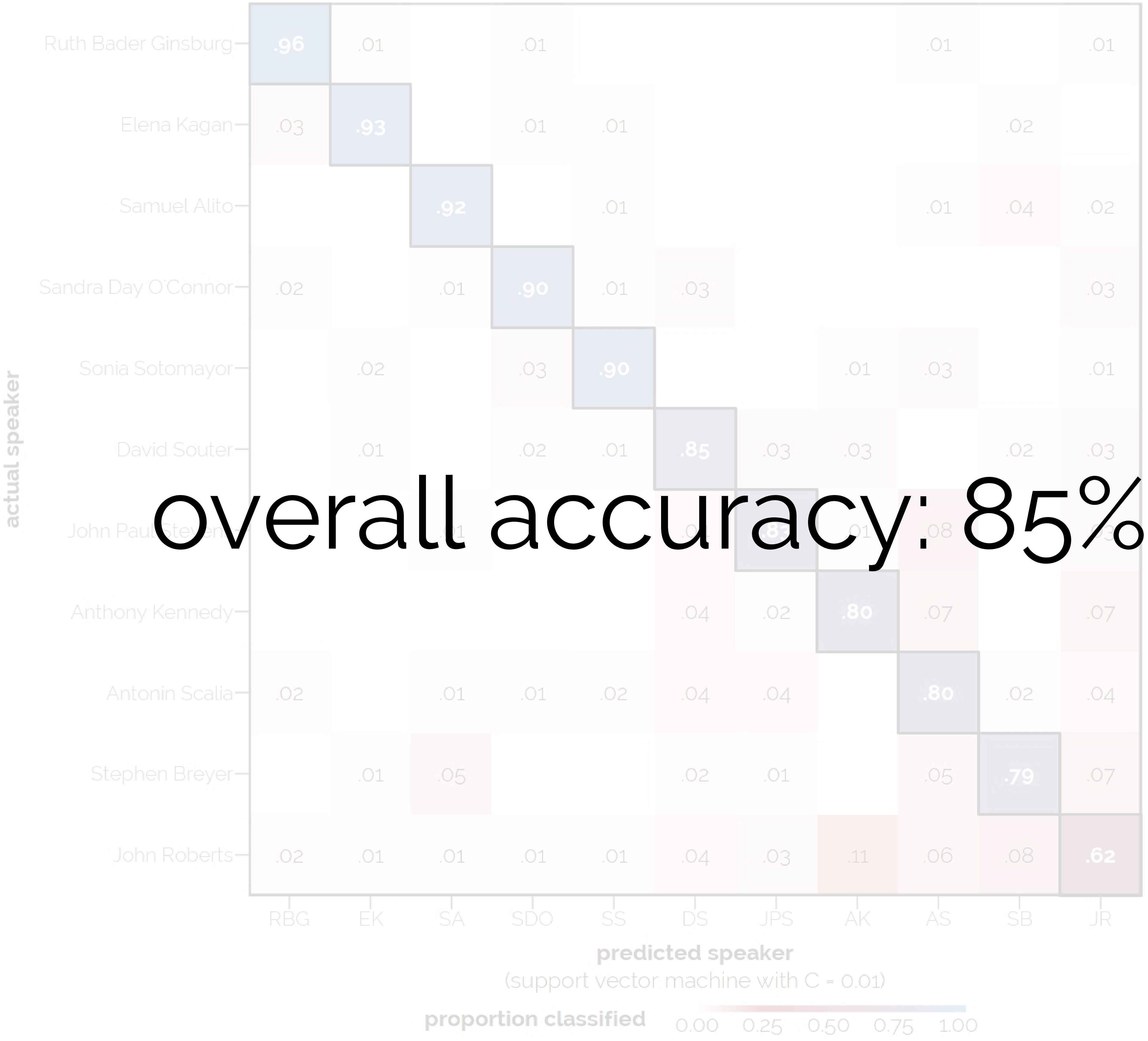
- Comparison with pyAudioAnalysis:
  - Widely used Python library for audio classification
  - Only alternative package in R or Python
- Benchmark performance vs. all available models:
  - Support vector machines
  - Gradient boosting
  - Random forest
  - Extremely randomized trees
- These methods do not model speech dynamics

best pyAudioAnalysis model by out-of-sample accuracy

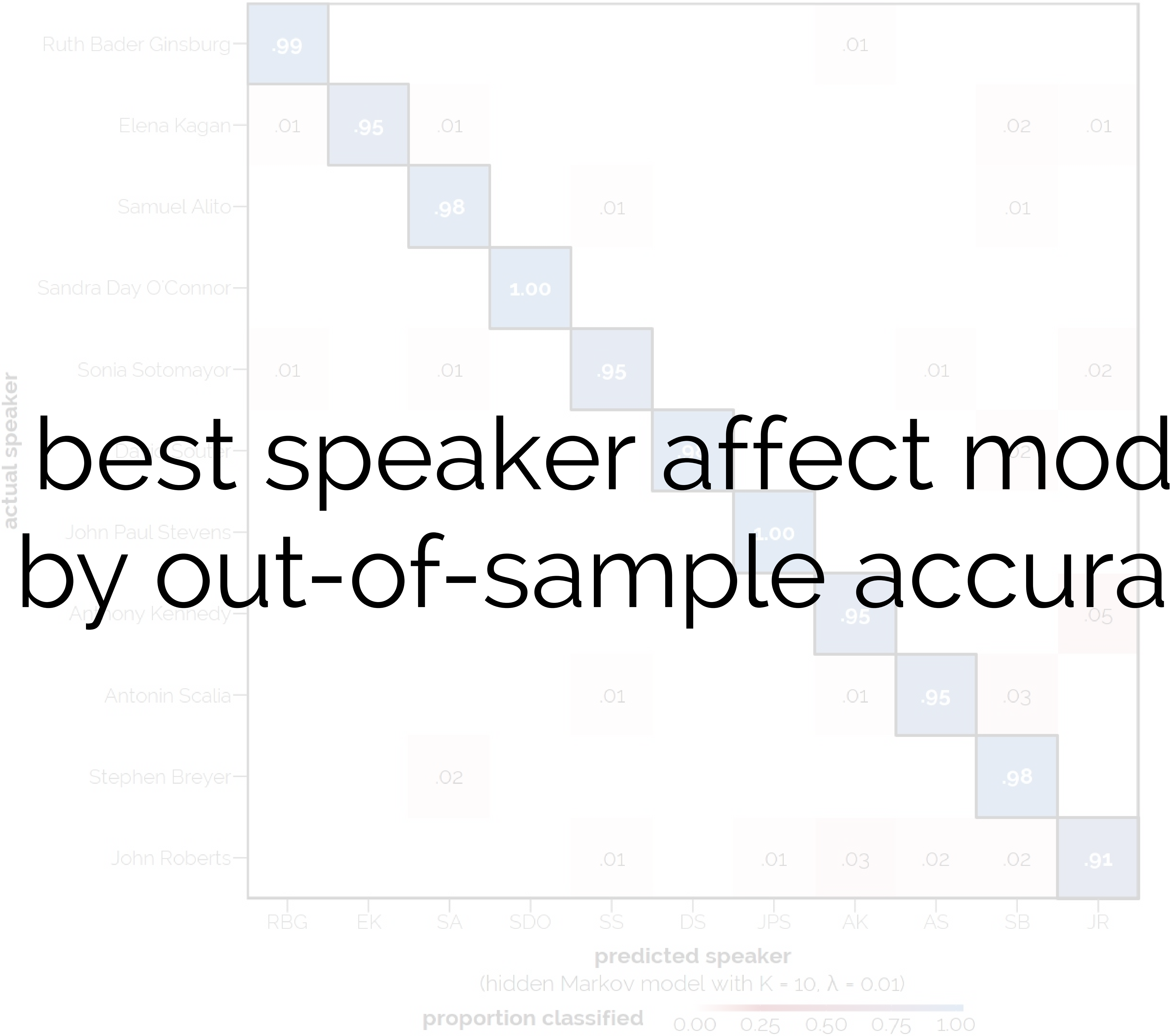



Ruth Bader Ginsburg
Elena Kagan
Samuel Alito
Sandra Day O'Connor
Sonia Sotomayor
David Souter
John Paul Stevens
Anthony Kennedy
Antonin Scalia
Stephen Breyer
John Roberts
actual speaker
.96
.01
.01
.01
.01
RBG
EK
SA
SDO
SS
DS
JPS
AK
AS
SB
JR
predicted speaker
(support vector machine with C = 0.01)
proportion classified
0.00
0.25
0.50
0.75
1.00

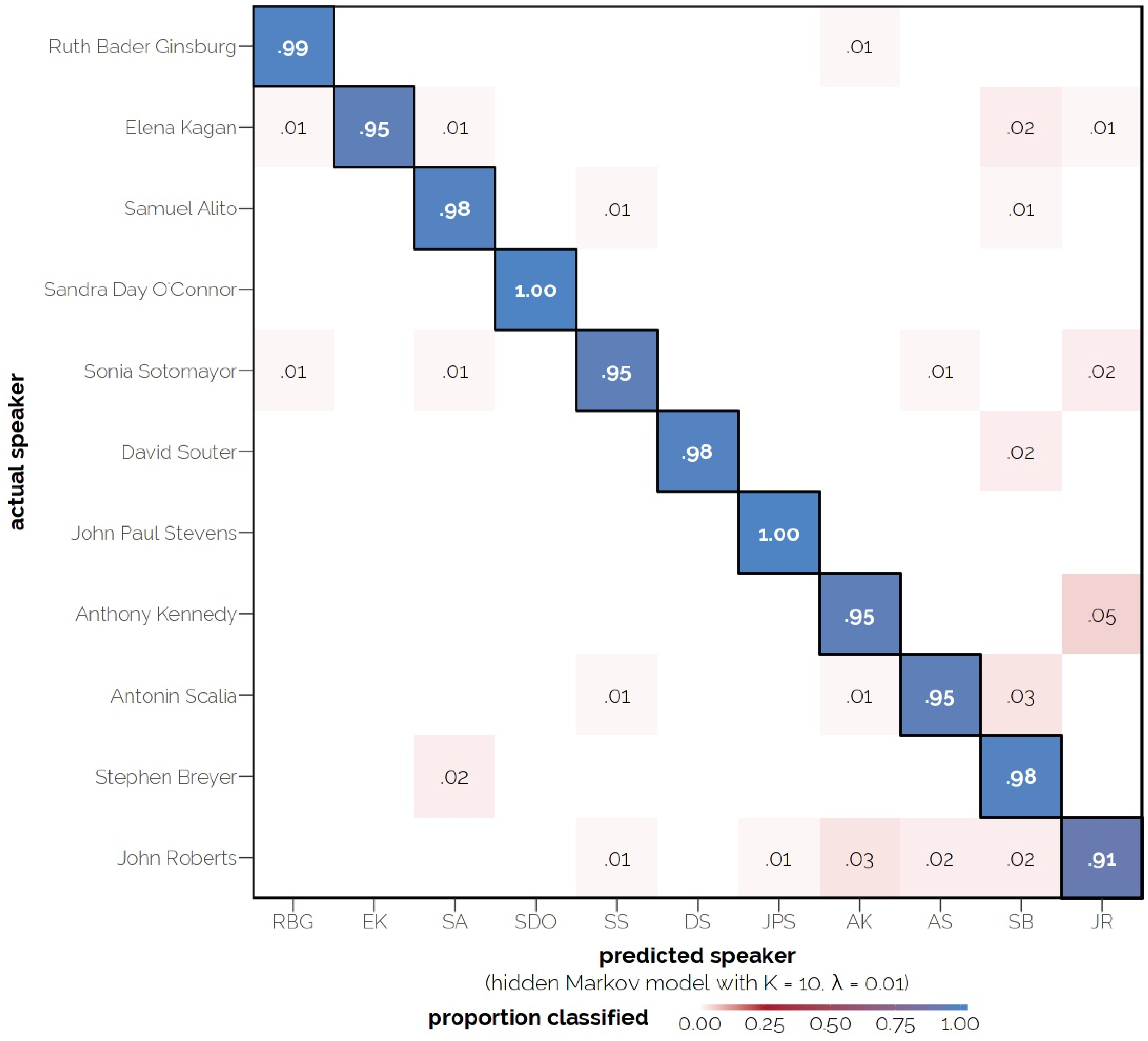

actual speaker
Ruth Bader Ginsburg
.96
.01
.01
.01
.01
Elena Kagan
.03
.93
.01
.01
.02
Samuel Alito
.92
.01
.01
.04
.02
Sandra Day O'Connor
.02
.01
.90
.01
.03
.03
Sonia Sotomayor
.02
.03
.90
.01
.03
.01
David Souter
.01
.02
.01
.85
.03
.03
.02
.03
John Paul Stevens
.01
.04
.83
.01
.08
.03
Anthony Kennedy
.04
.02
.80
.07
.07
Antonin Scalia
.02
.01
.01
.02
.04
.04
.80
.02
.04
Stephen Breyer
.01
.05
.02
.01
.05
.79
.07
John Roberts
.02
.01
.01
.01
.01
.04
.03
.11
.06
.08
.62
RBG
EK
SA
SDO
SS
DS
JPS
AK
AS
SB
JR
predicted speaker
(support vector machine with C = 0.01)
proportion classified
0.00
0.25
0.50
0.75
1.00

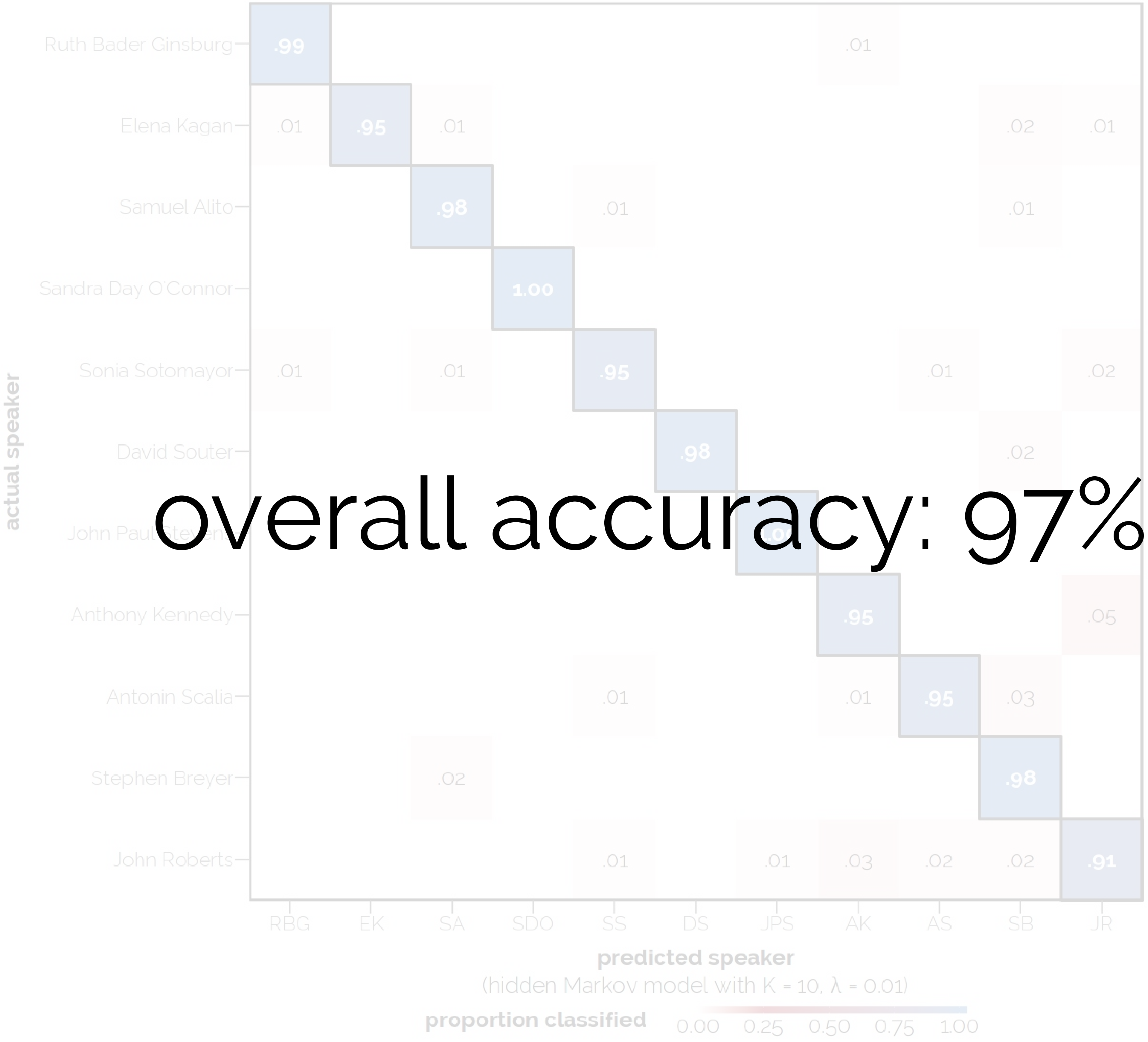


overall accuracy: 85%

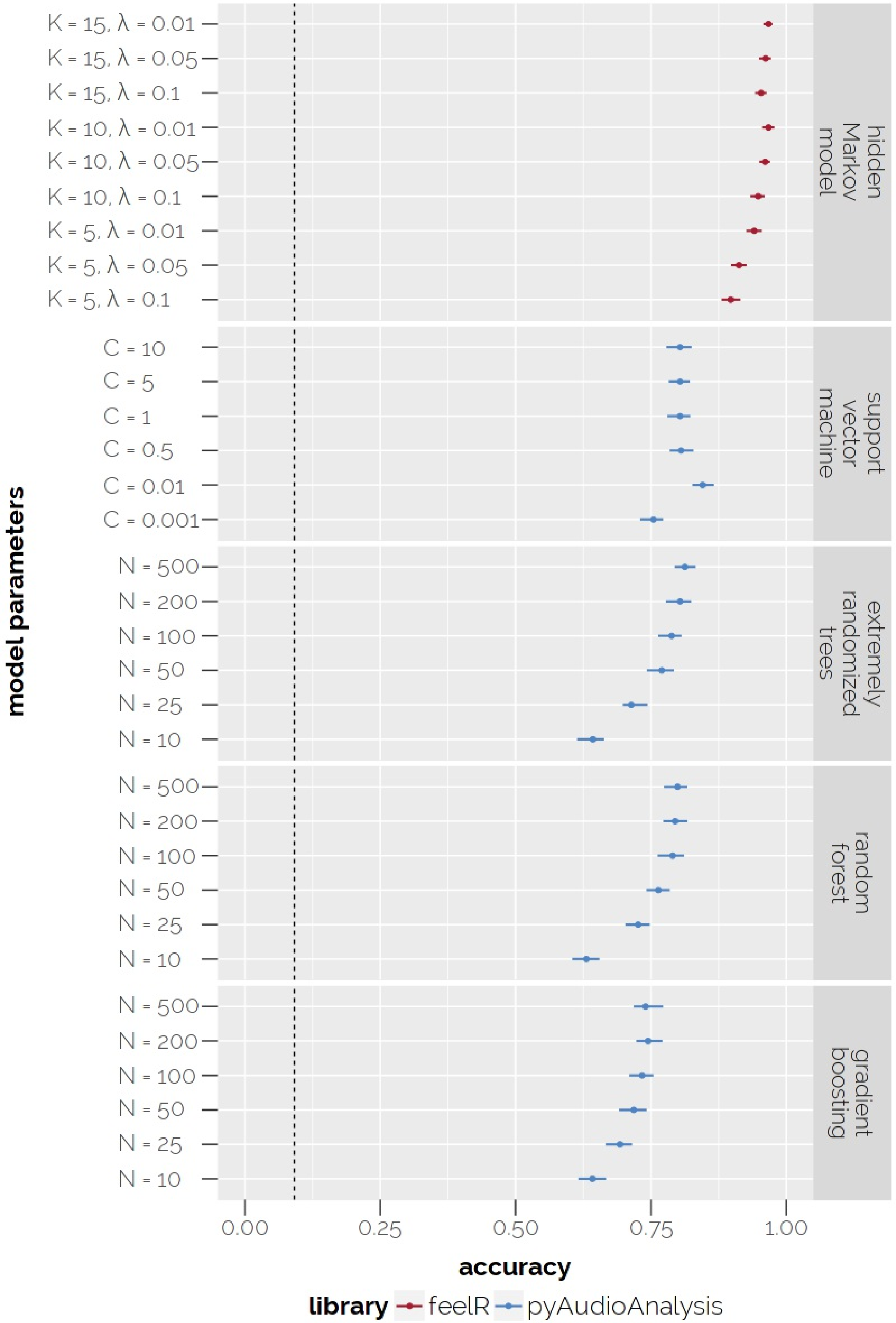


best speaker affect model by out-of-sample accuracy

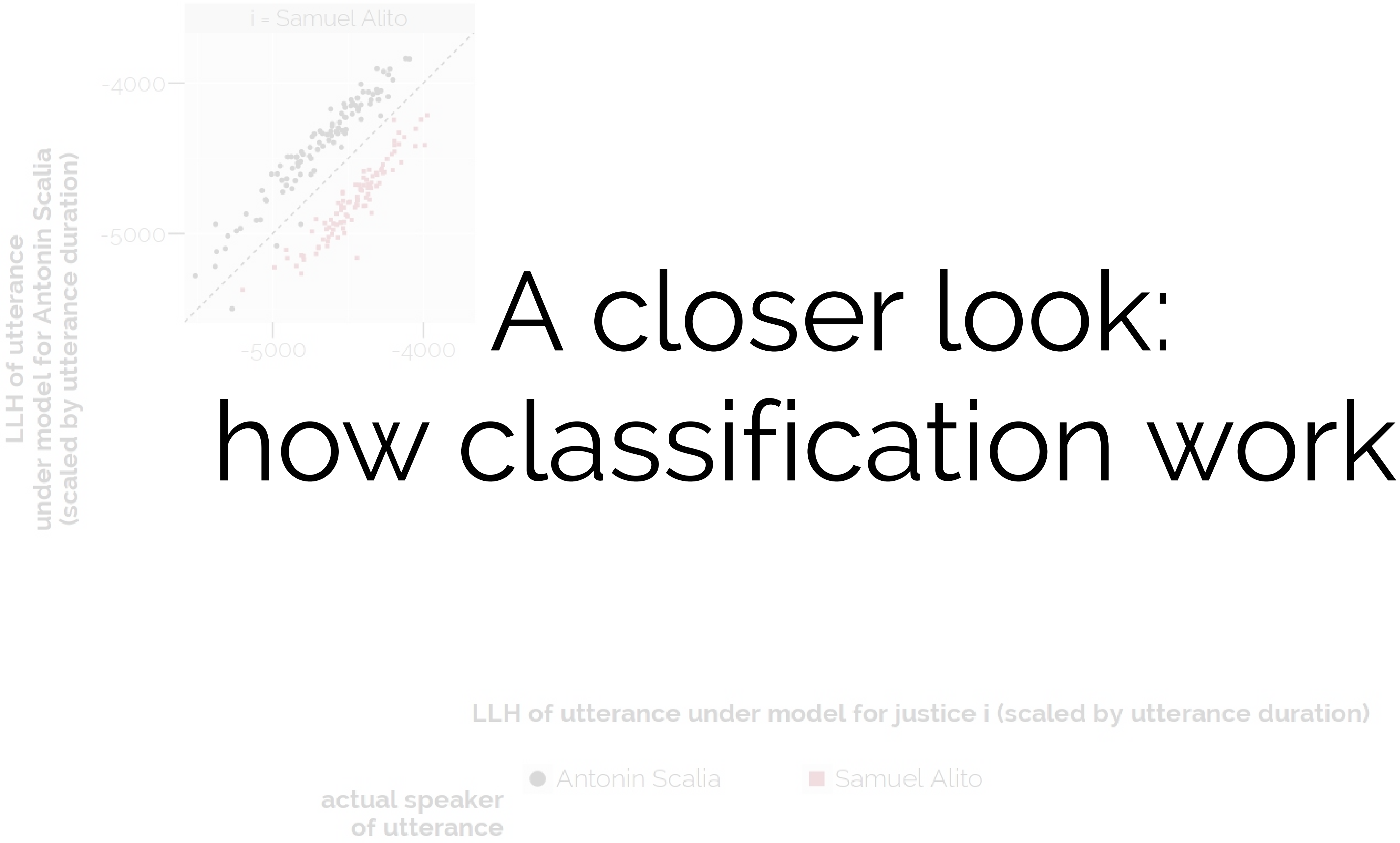

actual speaker
Ruth Bader Ginsburg
Elena Kagan
Samuel Alito
Sandra Day O'Connor
Sonia Sotomayor
David Souter
John Paul Stevens
Anthony Kennedy
Antonin Scalia
Stephen Breyer
John Roberts
.99
.01
.01
.95
.01
.02
.01
.98
.01
.01
1.00
.01
.01
.95
.01
.02
.98
.02
1.00
.95
.05
.01
.01
.95
.03
.02
.98
.01
.01
.03
.02
.02
.91
RBG
EK
SA
SDO
SS
DS
JPS
AK
AS
SB
JR
predicted speaker
(hidden Markov model with K = 10, λ = 0.01)
proportion classified
0.00
0.25
0.50
0.75
1.00

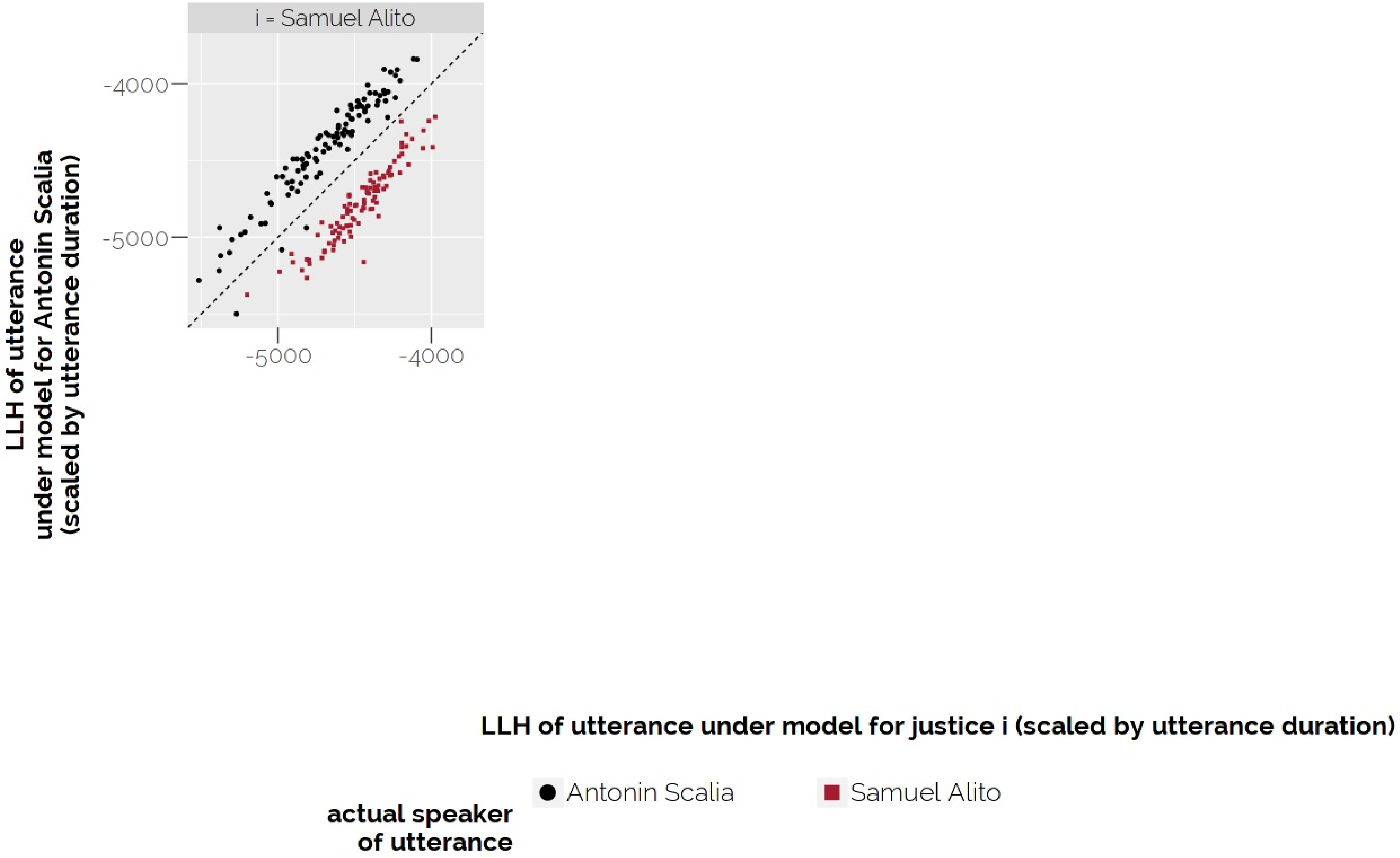


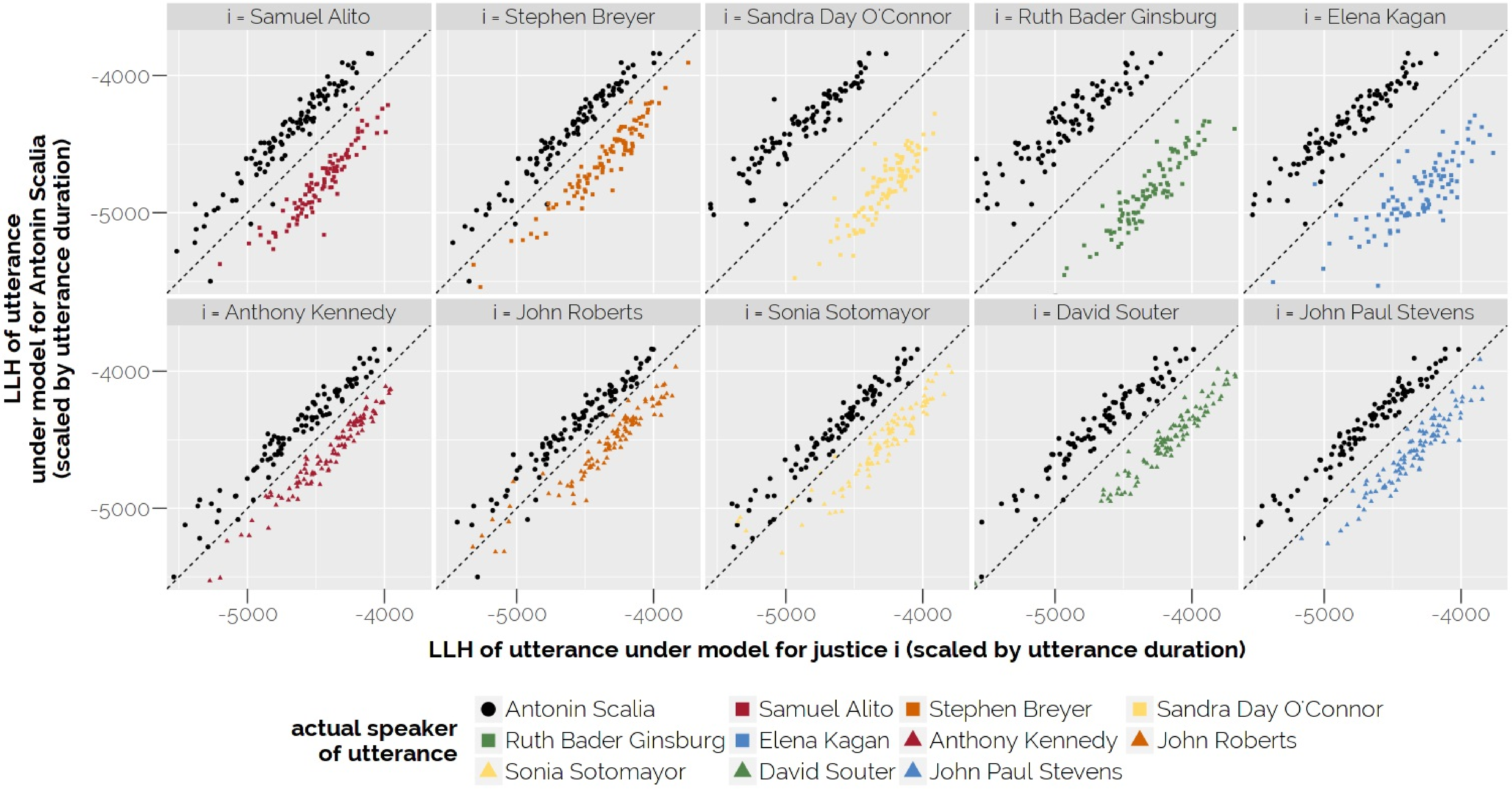
overall accuracy: 97%


K = 15, λ = 0.01
K = 15, λ = 0.05
K = 15, λ = 0.1
K = 10, λ = 0.01
K = 10, λ = 0.05
K = 10, λ = 0.1
K = 5, λ = 0.01
K = 5, λ = 0.05
K = 5, λ = 0.1
hidden Markov model
C = 10
C = 5
C = 1
C = 0.5
C = 0.01
C = 0.001
support vector machine
N = 500
N = 200
N = 100
N = 50
N = 25
N = 10
extremely randomized trees
N = 500
N = 200
N = 100
N = 50
N = 25
N = 10
random forest
N = 500
N = 200
N = 100
N = 50
N = 25
N = 10
gradient boosting
model parameters
0.00
0.25
0.50
0.75
1.00
accuracy
library
feelR
pyAudioAnalysis

A closer look:
how classification works




i = Samuel Alito
-4000
-5000
-5000
-4000
LLH of utterance under model for Antonin Scalia (scaled by utterance duration)
LLH of utterance under model for justice i (scaled by utterance duration)
actual speaker of utterance
● Antonin Scalia
■ Samuel Alito


i = Samuel Alito
i = Stephen Breyer
i = Sandra Day O'Connor
i = Ruth Bader Ginsburg
i = Elena Kagan
i = Anthony Kennedy
i = John Roberts
i = Sonia Sotomayor
i = David Souter
i = John Paul Stevens
-4000
-5000
LLH of utterance under model for Antonin Scalia (scaled by utterance duration)
LLH of utterance under model for justice i (scaled by utterance duration)
actual speaker of utterance
Antonin Scalia
Samuel Alito
Stephen Breyer
Sandra Day O'Connor
Ruth Bader Ginsburg
Elena Kagan
Anthony Kennedy
John Roberts
Sonia Sotomayor
David Souter
John Paul Stevens

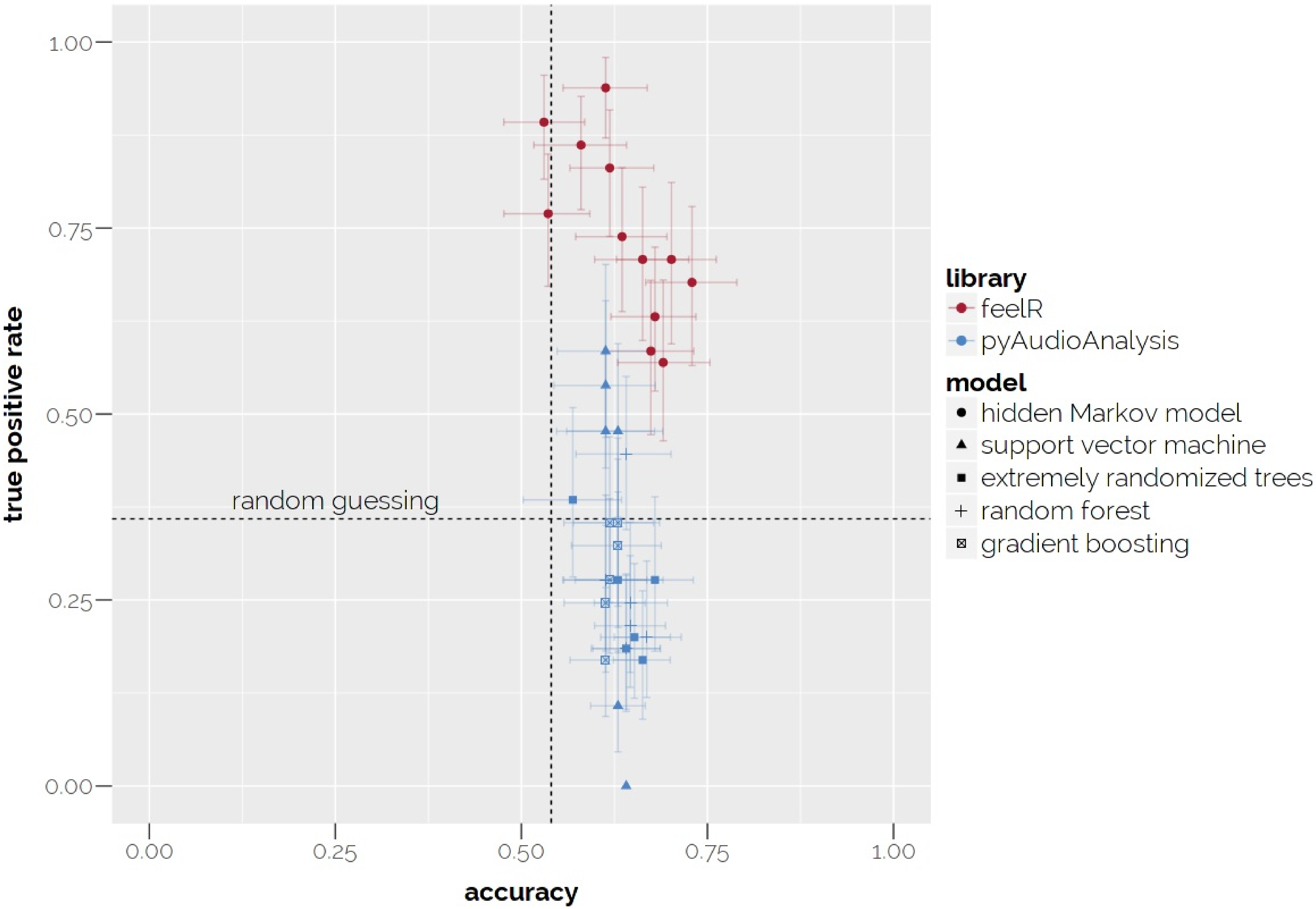
Supreme Court Emotion Classification

Preliminary Results

- Coded 200 utterances by Chief Justice Roberts
  - Modes of speech: "neutral" (64%) and "skeptical" (36%)
  - Perceived "skepticism" depends on both text & tone
- Existing Supreme Court sentiment analyses use text of utterances only
- Speaker affect model uses tone of utterances only

Preliminary Results

- HMM selected by K-fold CV: 15 states, $\lambda=0.01$
  - Out-of-sample accuracy: 70% accuracy
  - True positive rate (skepticism): 71%
  - True negative rate (neutral): 70%
- Best pyAudioAnalysis model: SVM with C=10
  - Overall accuracy 61%, TPR 58%, TNR 63%
- Stanford Core NLP deep learning model with text:
  - Vast majority (78%) classified as "negative" (≈ skepticism?)
  - Overall accuracy 45%, TPR 89%, TNR 20%



true positive rate
accuracy
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
random guessing
library
feelR
pyAudioAnalysis
model
● hidden Markov model
▲ support vector machine
■ extremely randomized trees
+ random forest
⊠ gradient boosting

Conclusion

- Recap
    - New sources of data for social scientists
    - New questions about political speech
    - Advances over state-of-the-art CS models
- Ongoing work
    - Incorporating text into audio analysis (Knox, Lucas)
    - Rhetoric of Parliamentary Debate (Goplerud, Knox, Lucas)
    - Analyzing visual features with text (Lucas)