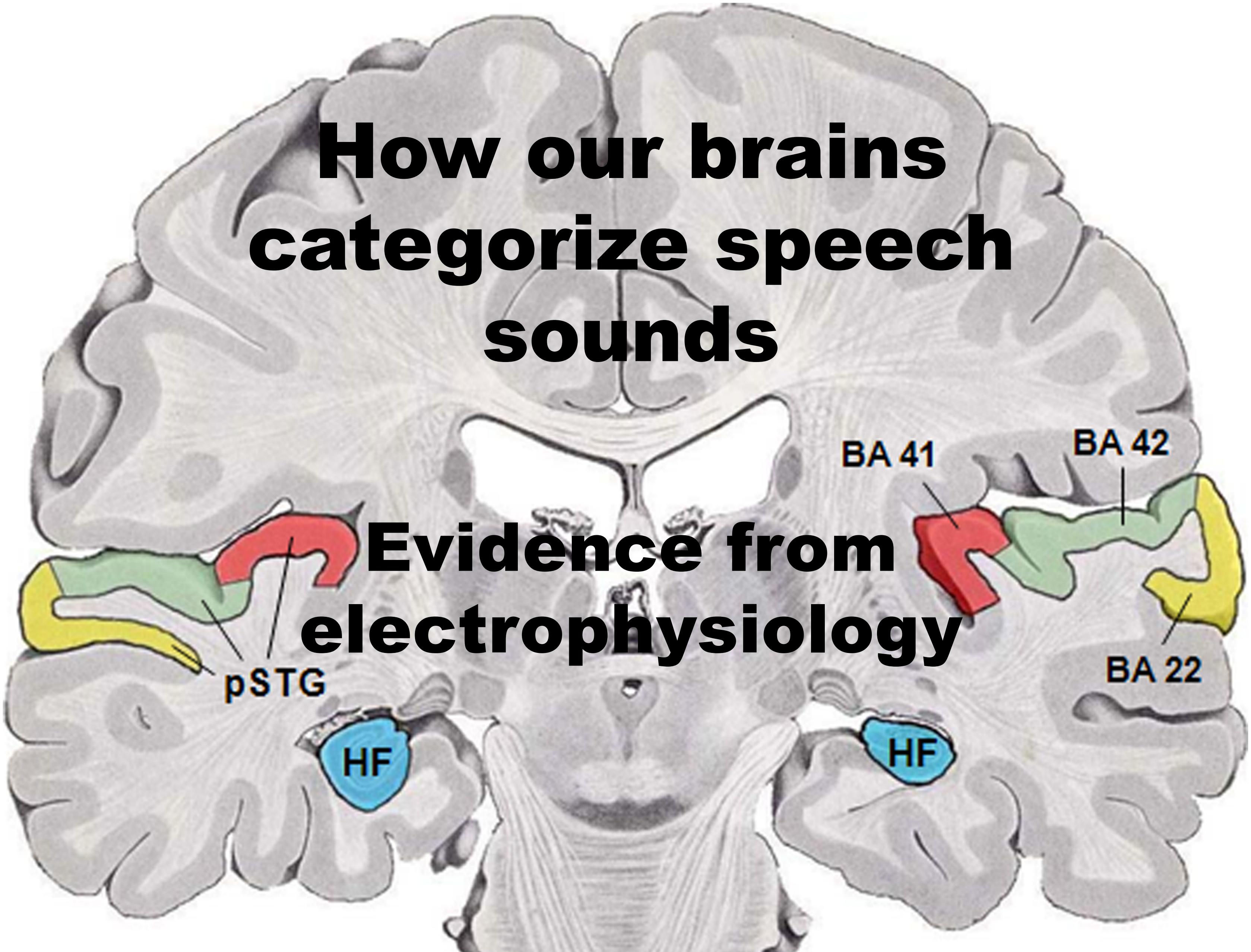
How our brains categorize speech sounds

Evidence from electrophysiology

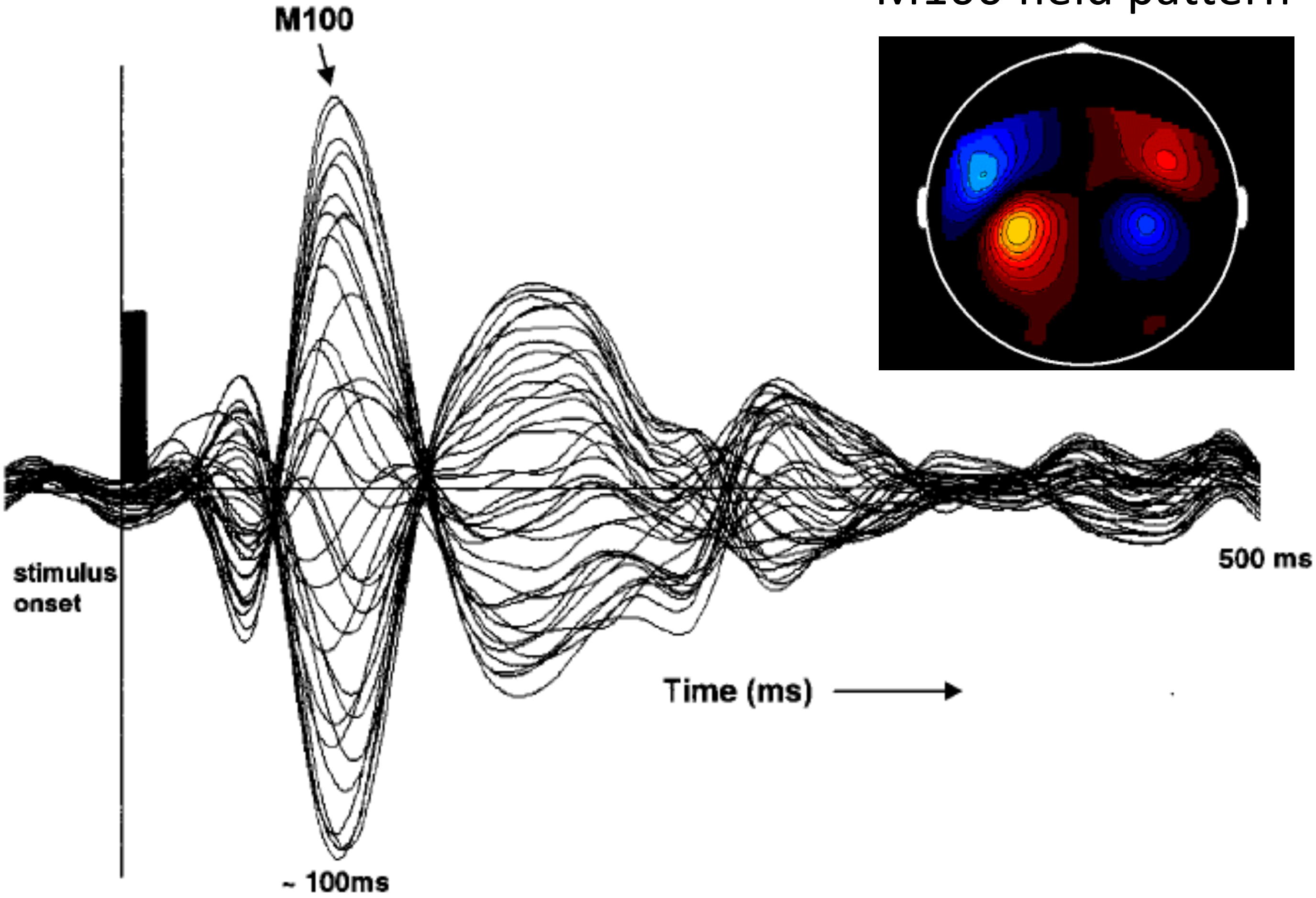


What do we know about how our brains categorize speech sounds into phonemes?

- Our brains get the job done within 180ms from sound onset.
  - **Evidence from mismatch studies**

- Superior temporal cortex represents distinctive features prior to 180ms.
  - **Evidence from natural listening during ECoG/iEEG**
    - Direct recordings from the cortex of pre-surgical patients

Auditory evoked response (MEG)



M100 field pattern

Roberts, Ferrari, Stufflebeam & Poeppel, 2000, *J Clin Neurophysiol*, Vol 17, No 2, 2000

M100 (N1 in EEG)

- Generated in auditory cortex bilaterally.
- Part of the response to **all sounds**.

Affected by:

- Intensity
  - □ **higher intensity → shorter latency, larger amplitude**
- Frequency
  - □ **higher frequency → shorter latency**
- Phonetic identity
  - □ **e.g., shorter latencies for *a* than for *u* (explainable in spectral terms)**
- **Role of the M100 in phoneme perception was unclear for a long time.**

Auditory Mismatch Field

- A change detector.
- Not part of the basic processing of sounds. (though various theories do exist)

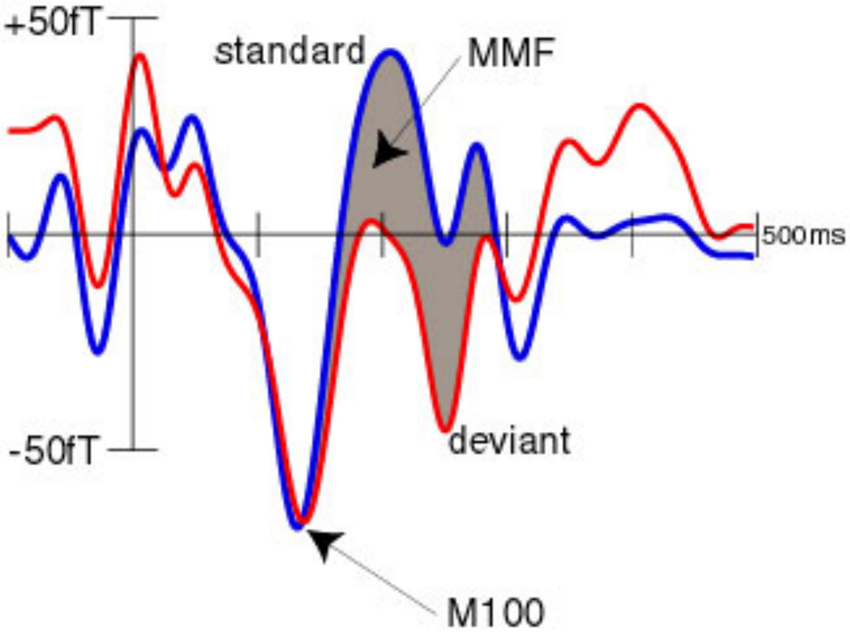
Auditory Mismatch Field

- Elicited in auditory cortex at ~180ms post stimulus onset by deviant stimuli in an oddball paradigm.

XXXYXXXXXYXXXXXYXXX

X = “standard”

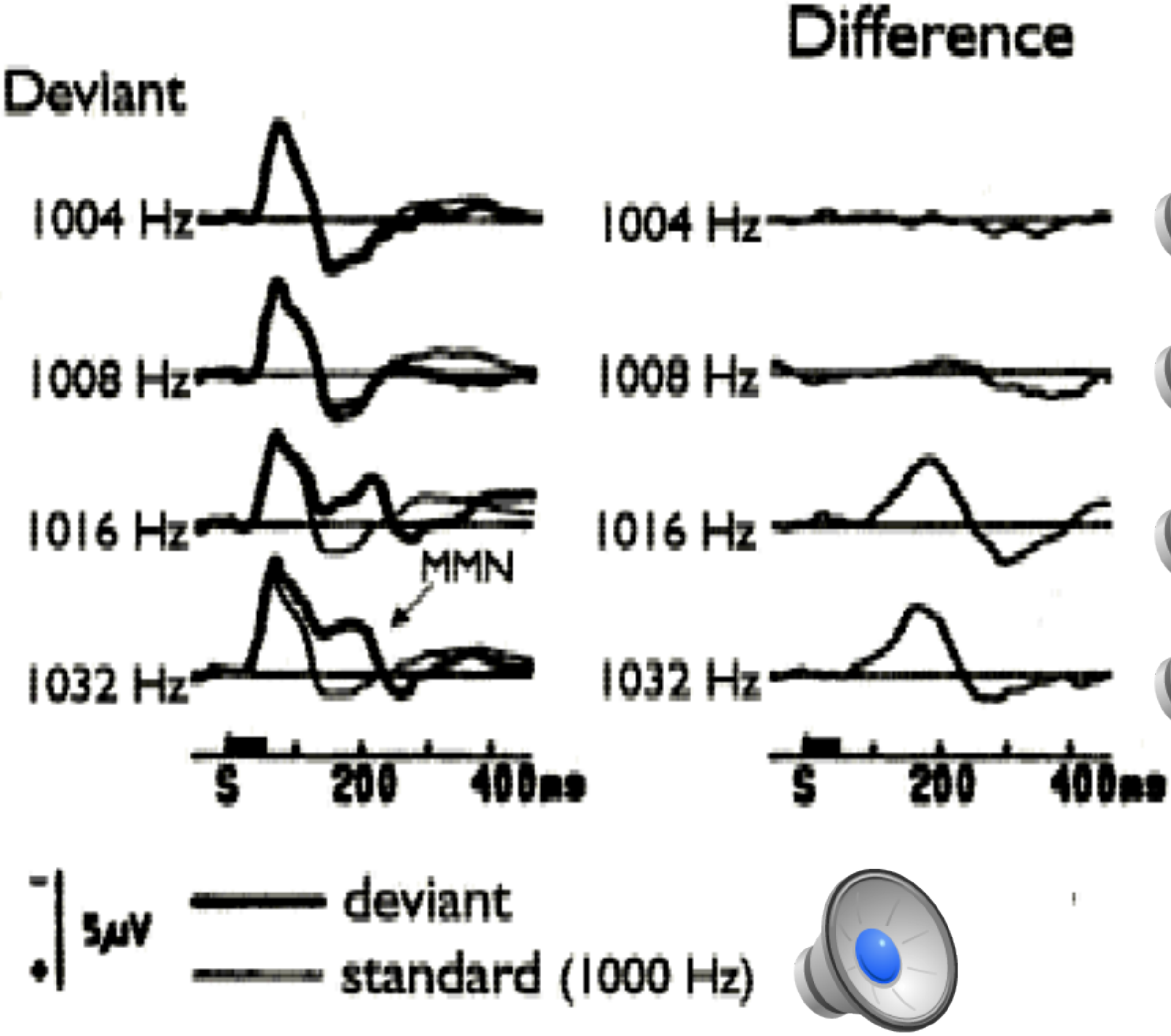
Y = “deviant”



Auditory Mismatch Field

- A tool for investigating what counts as a change for auditory cortex.
- Has been heavily used in the study of the neural bases of categorical perception.

Mismatch field as a function of frequency change

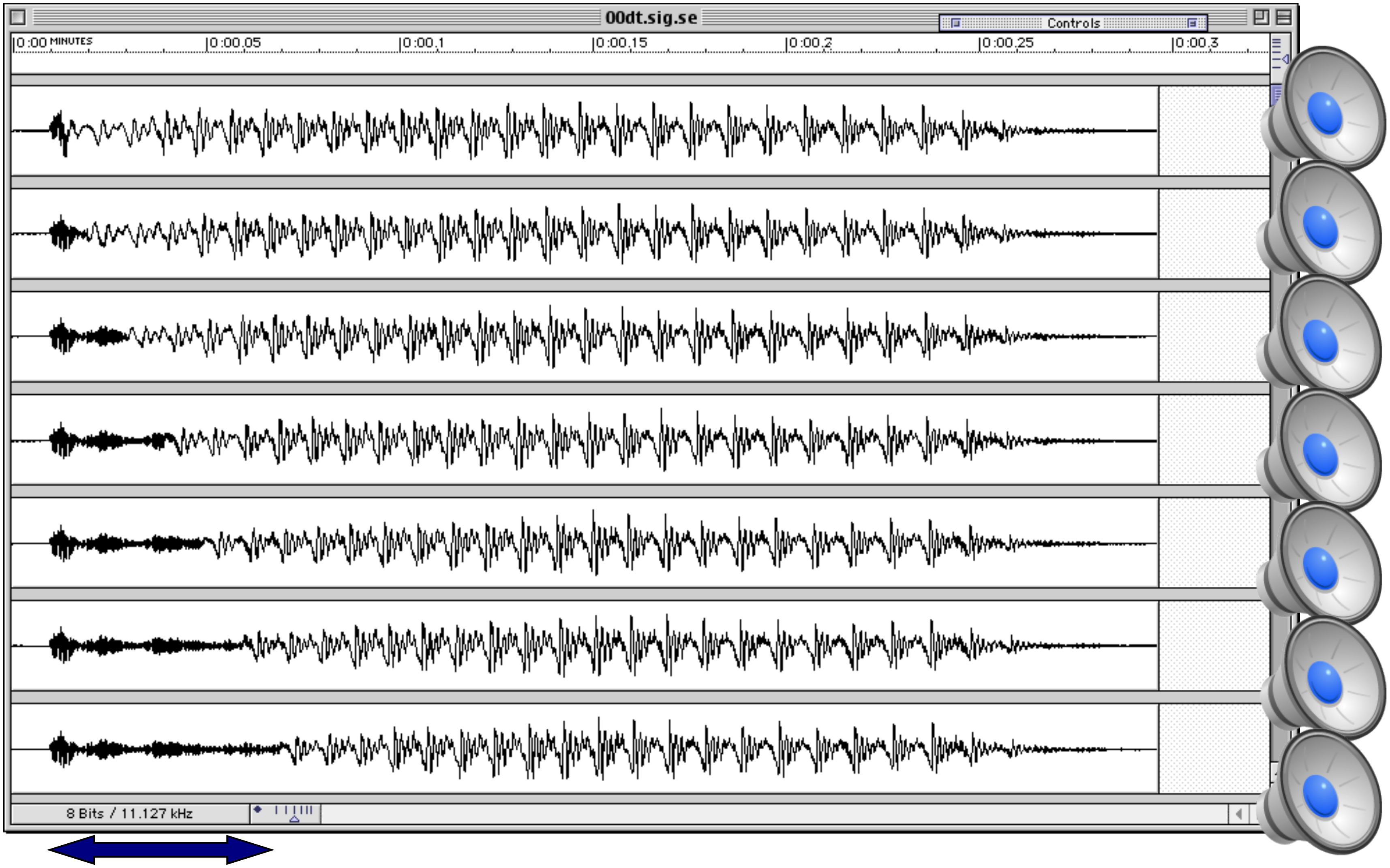


Sams et al. 1985

The Mismatch Field and Phonemic Categorization

- Would crossing a phoneme boundary elicit a mismatch field?
- If yes, this would tell us
  - □ **that auditory cortex has access to phonemic categories**
  - □ **that category information is accessed by 180ms.**
- **But how can we cross a phoneme boundary without also creating a physical distance that on its own, would also elicit a mismatch effect?**

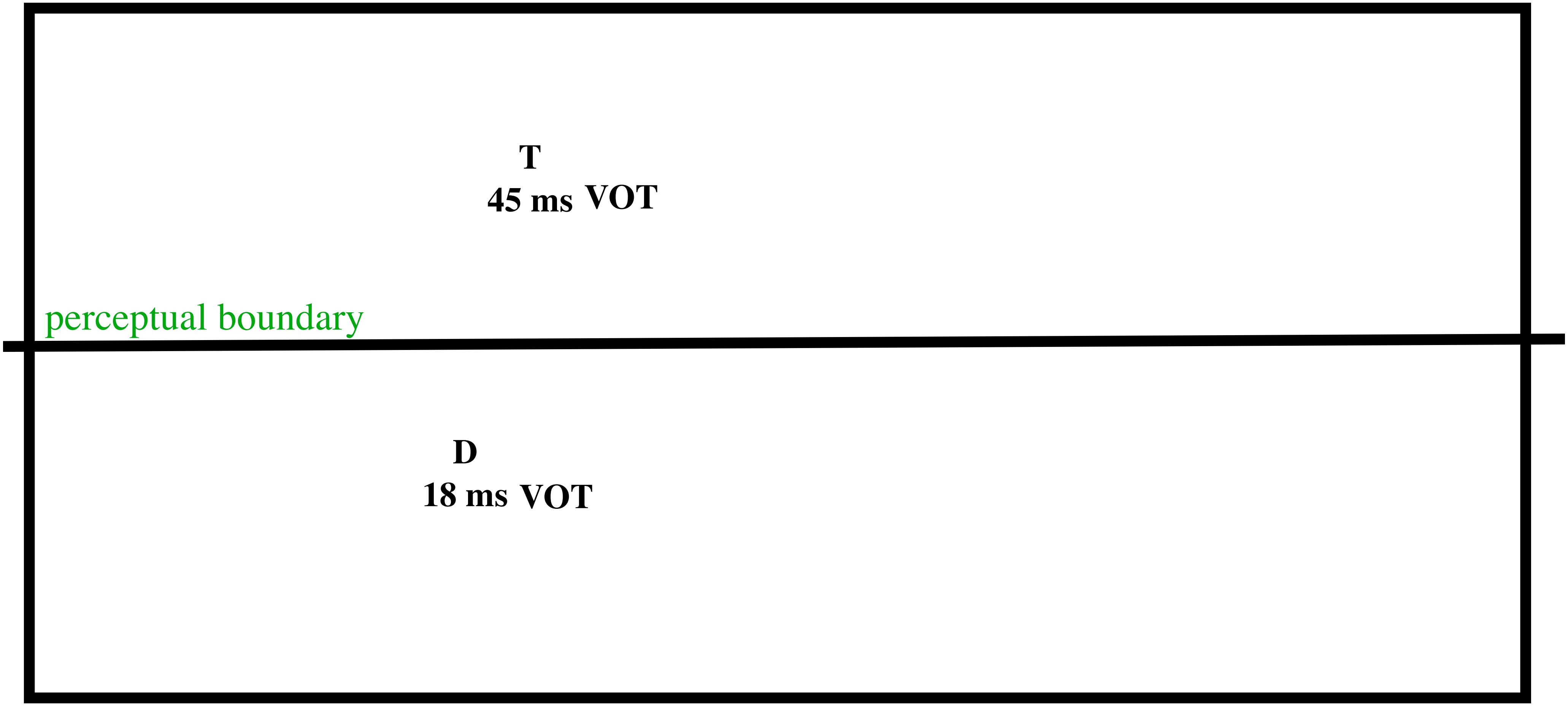
VOT continuum for ta-da



60 msec

English boundary

deviants



T
45 ms VOT

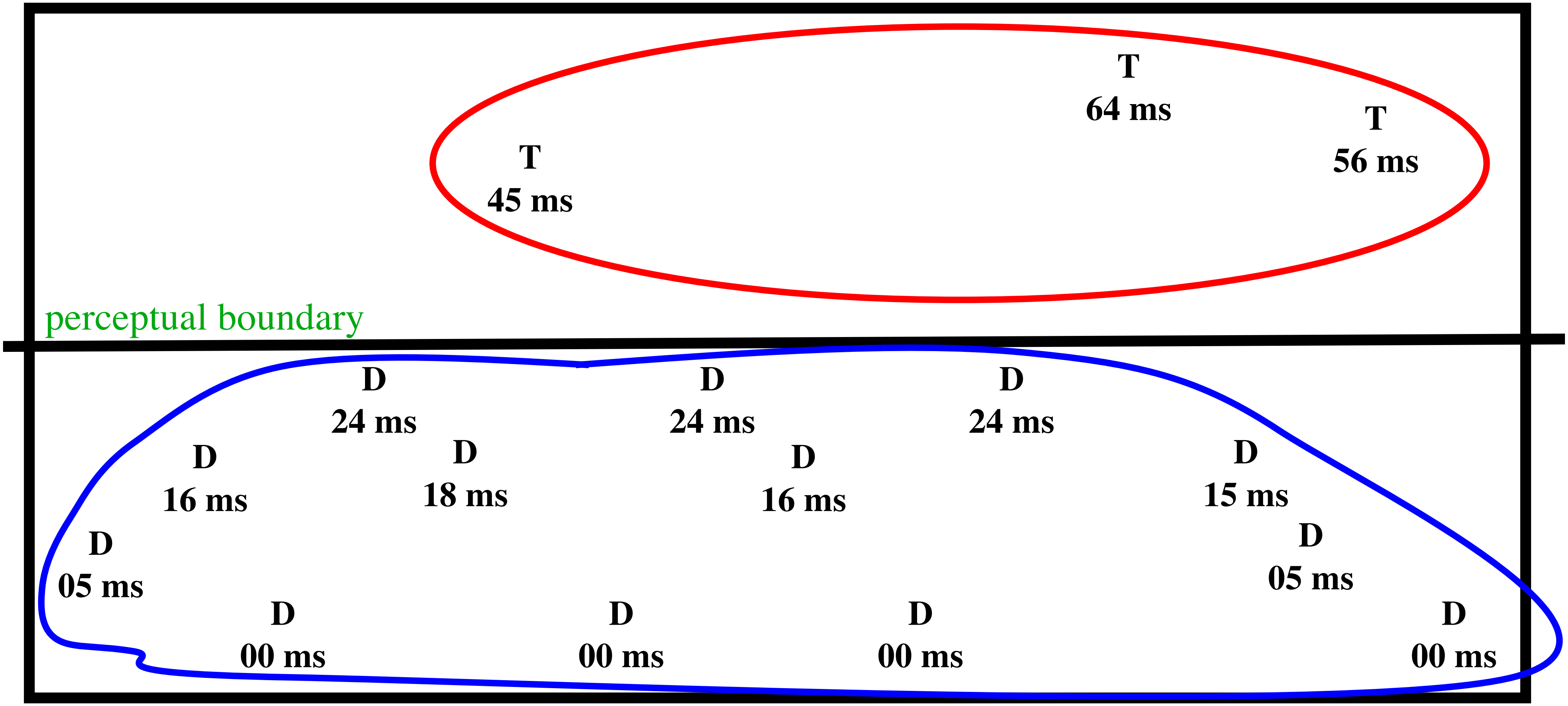
perceptual boundary

D
18 ms VOT

standards

Phillips et al. (2000, *Journal of Cognitive Neuroscience*)

deviants



T
64 ms

T
56 ms

T
45 ms

perceptual boundary

D
24 ms

D
24 ms

D
24 ms

D
16 ms

D
18 ms

D
16 ms

D
15 ms

D
05 ms

D
05 ms

D
00 ms

D
00 ms

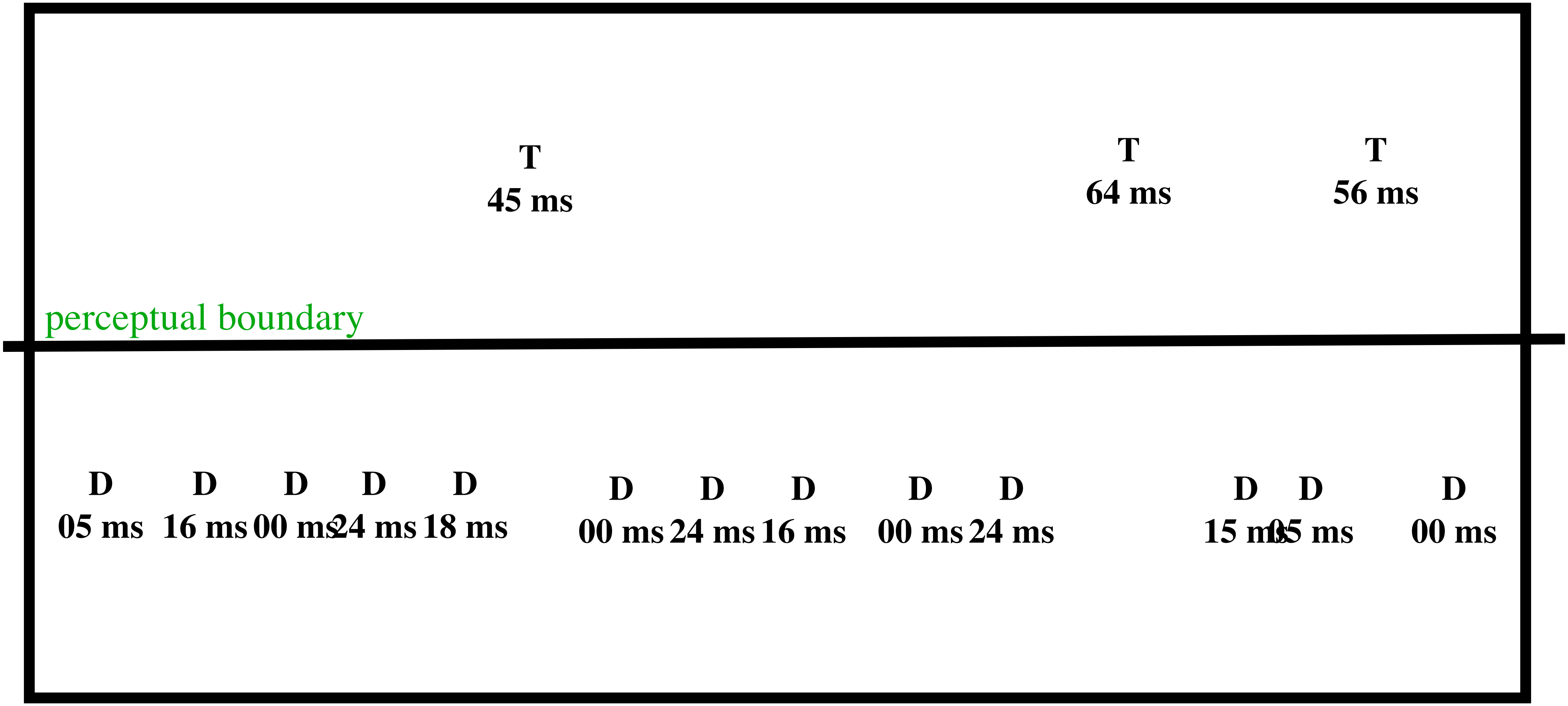
D
00 ms

D
00 ms

standards

Acoustic representation

Phillips et al. (2000, *Journal of Cognitive Neuroscience*)



T 45 ms

T 64 ms

T 56 ms

perceptual boundary

D 05 ms

D 16 ms

D 00 ms

D 24 ms

D 18 ms

D 00 ms

D 24 ms

D 16 ms

D 00 ms

D 24 ms

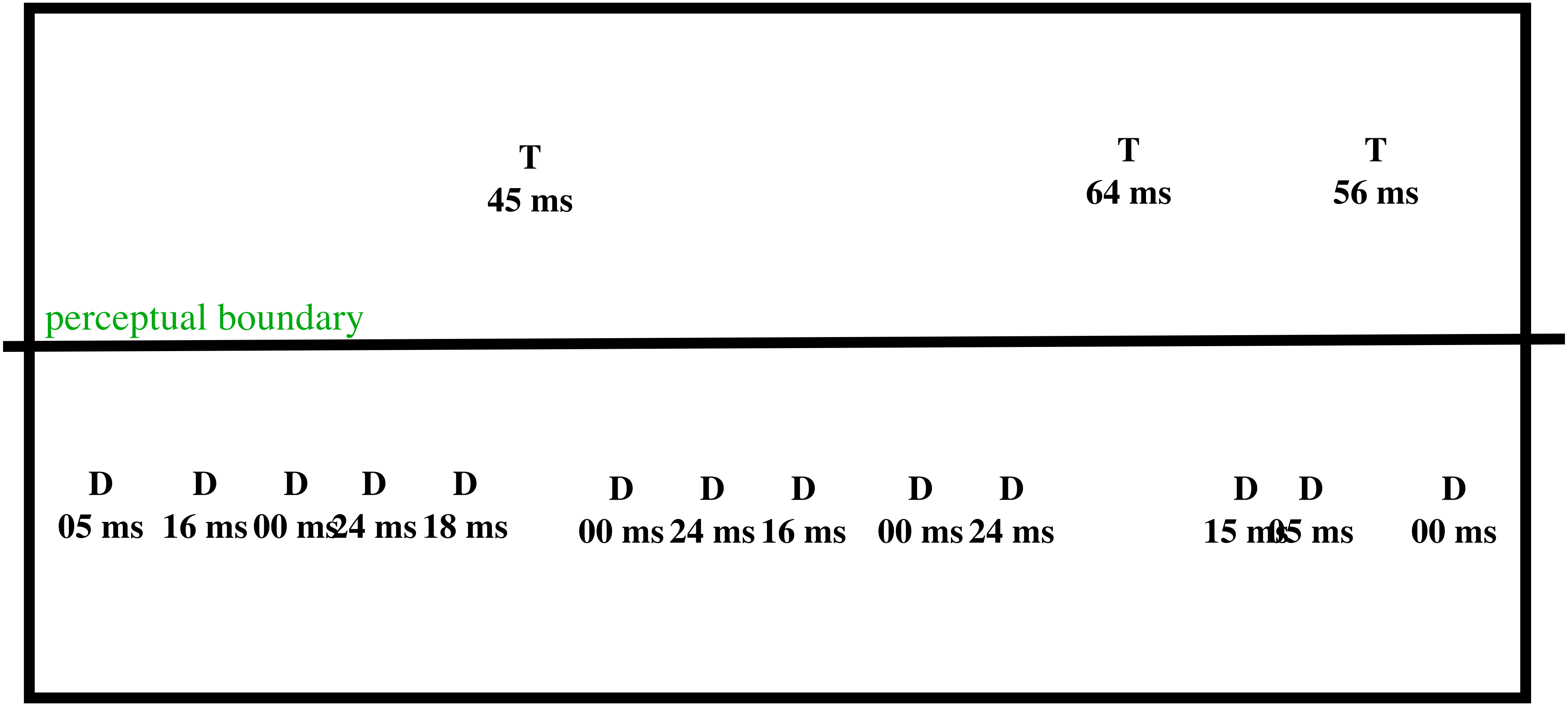
D 15 ms

D 05 ms

D 00 ms

Phonological representation

Phillips et al. (2000, *Journal of Cognitive Neuroscience*)



T 45 ms

T 64 ms

T 56 ms

perceptual boundary

D 05 ms

D 16 ms

D 00 ms

D 24 ms

D 18 ms

D 00 ms

D 24 ms

D 16 ms

D 00 ms

D 24 ms

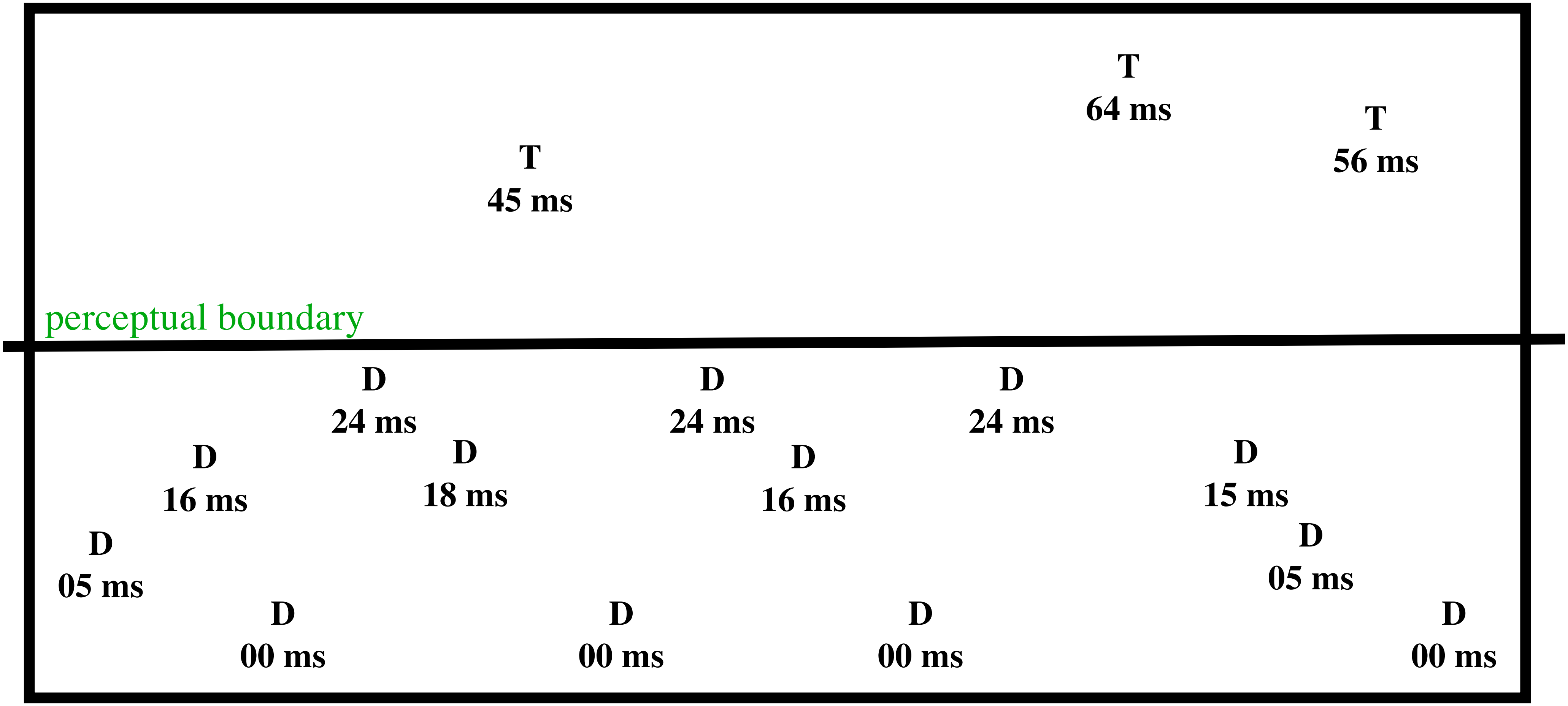
D 15 ms

D 05 ms

D 00 ms

If the auditory cortex is sensitive to phonological categories, deviant Ts should elicit a Mismatch Field.

What should happen if we keep everything else the same, but lift all the VOTs by 20ms? What does the phonological hypothesis predict?



T
64 ms

T
56 ms

T
45 ms

perceptual boundary

D
24 ms

D
24 ms

D
24 ms

D
16 ms

D
18 ms

D
16 ms

D
15 ms

D
05 ms

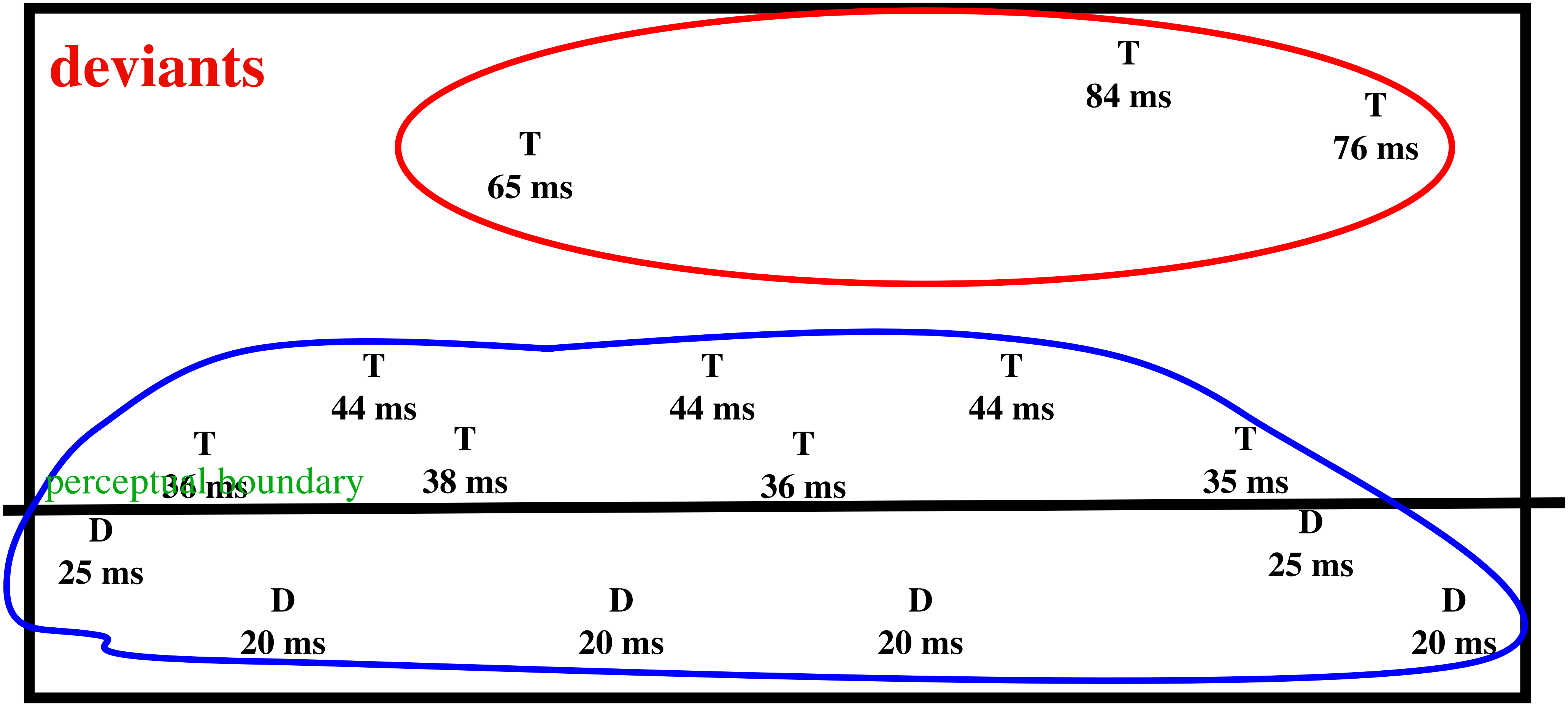
D
05 ms

D
00 ms

D
00 ms

D
00 ms

D
00 ms



deviants

T
84 ms

T
76 ms

T
65 ms

T
44 ms

T
44 ms

T
44 ms

T
36 ms

T
38 ms

T
36 ms

T
35 ms

perceptual boundary

D
25 ms

D
25 ms

D
20 ms

D
20 ms

D
20 ms

D
20 ms

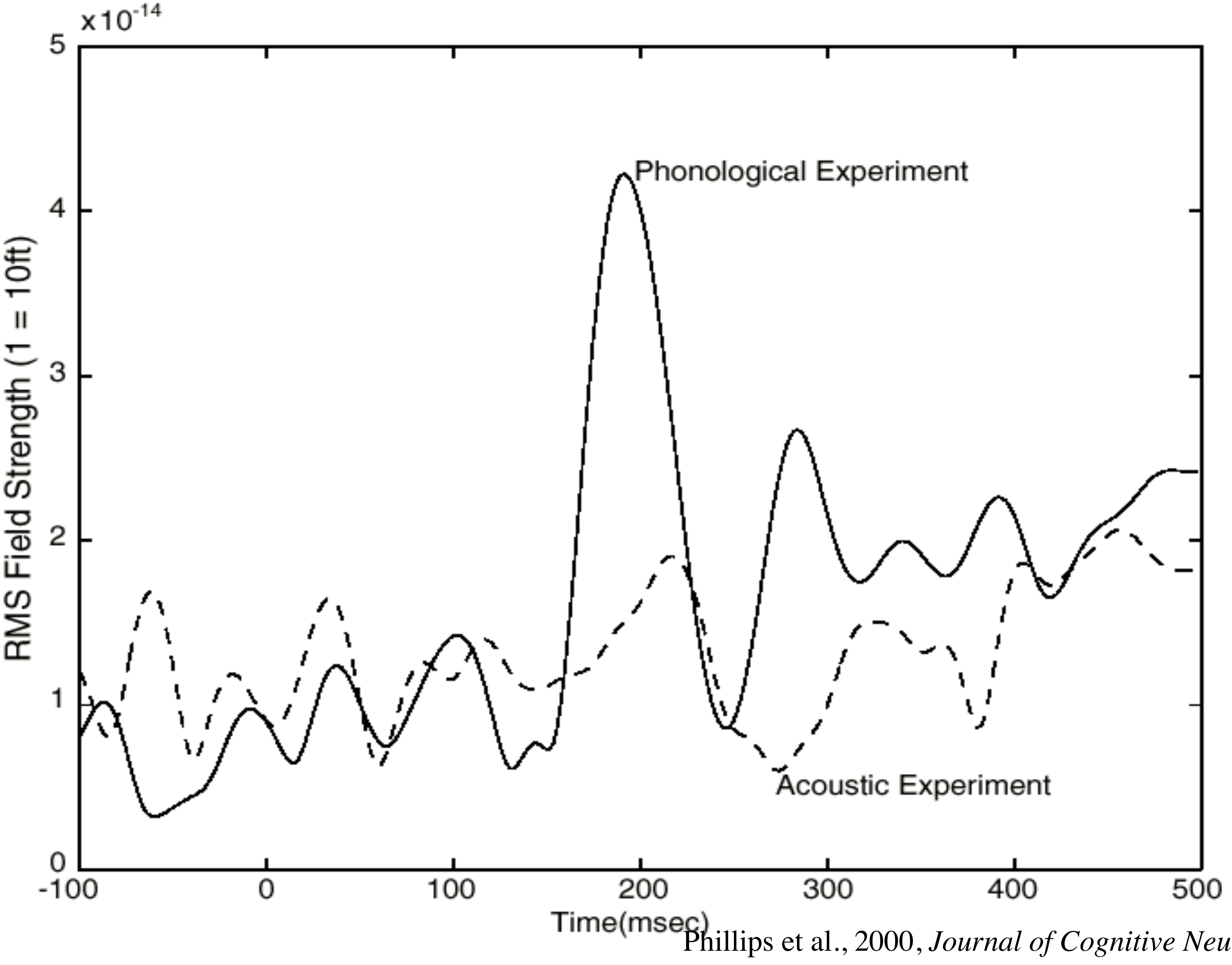
standards

The deviants should not elicit a phoneme driven Mismatch Field -- they are not perceived as different from the standards.

Phillips et al. (2000, *Journal of Cognitive Neuroscience*)

- Experiment 1: Standards fall on one side of a perceptual boundary, the deviants on the other.
  - □ **Standards and deviants differ BOTH in physical distance and phonemic category**
- Experiment 2: The physical distance between standards and deviants is as in Exp. 1, but now the standards and deviants are no longer differentiated by a perceptual boundary.
  - □ **Standards and deviants differ ONLY in physical distance**
- If a Mismatch Field elicited in Exp 1 is driven by the category difference between standards and deviants, it should disappear in Exp 2. **YES!**

Difference waves between deviants and standards in Exp 1 (phon) & Exp 2 (acoust)



Phillips et al., 2000, *Journal of Cognitive Neuroscience*

Mismatch Summary

- By 180ms, auditory cortex has extracted from the input the relevant features needed for categorizing sounds into phonemes.
- Since the Mismatch field is a surprise response, the necessary computational steps needed for categorization must occur prior to 180ms.
- How this happens has not been explained by research on the M100 peak.
- But recent recordings directly from the cortical surface of the superior temporal gyrus have shed light on this.

Mesgarani et al. 2014 *Science*:

Phonetic Feature Encoding in Human Superior Temporal Gyrus

- Electrode grid on cortex
- Subject listens to 500 sentences
- Sentences are segmented for phonemes
- The sensitivity of each electrode to the different speech sounds is studied

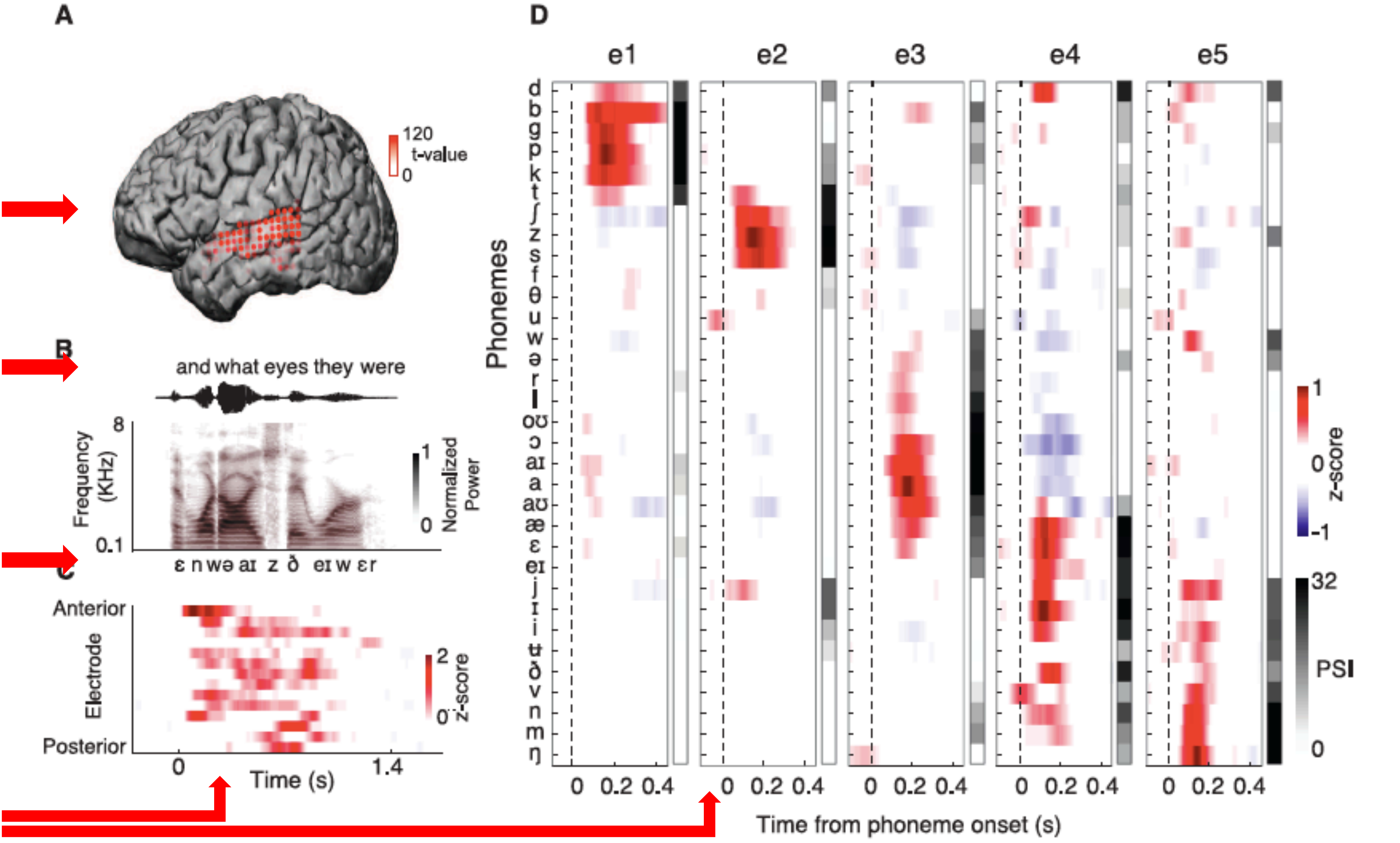
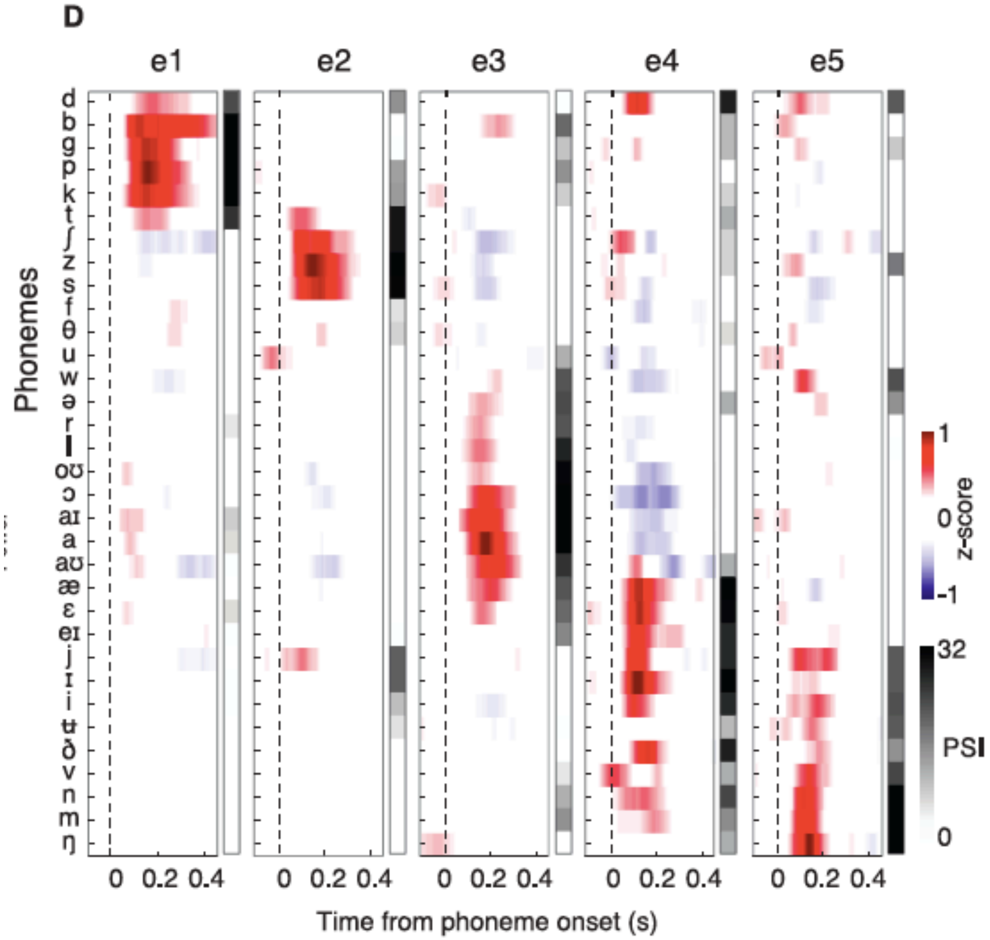


Fig. 1. Human STG cortical selectivity to speech sounds. (A) Magnetic resonance image surface reconstruction of one participant's cerebrum. Electrodes (red) are plotted with opacity signifying the *t* test value when comparing responses to silence and speech ($P < 0.01$, *t* test). **(B)** Example sentence and its acoustic waveform, spectrogram, and phonetic transcription. **(C)** Neural responses evoked by the sentence at selected electrodes. z score indicates normalized response. **(D)** Average responses at five example electrodes to all English phonemes and their PSI vectors.

Mesgarani et al. 2014 *Science*: Phonetic Feature Encoding in Human Superior Temporal Gyrus

- Result: most STG electrodes were sensitive not to individual phonemes but to groups of phonemes, sharing a feature (or features).
- Evidence that the feature space relevant for phonological categorization is represented in STG.



In sum:

- When the spatial distribution of responses to phonemes is studied in STG, we find evidence that (probably) the entire feature space of speech sounds is represented around auditory cortex at 100 – 200ms (and probably even earlier, but I didn't show you evidence for this).
- Given the mismatch literature indicating completion of categorization at 180ms, something like this had to be the case.
- While we still don't know *how* our brains perform phonological categorization, we know that the relevant information is represented in STG at around 100ms and that by 180ms, categorization has occurred.