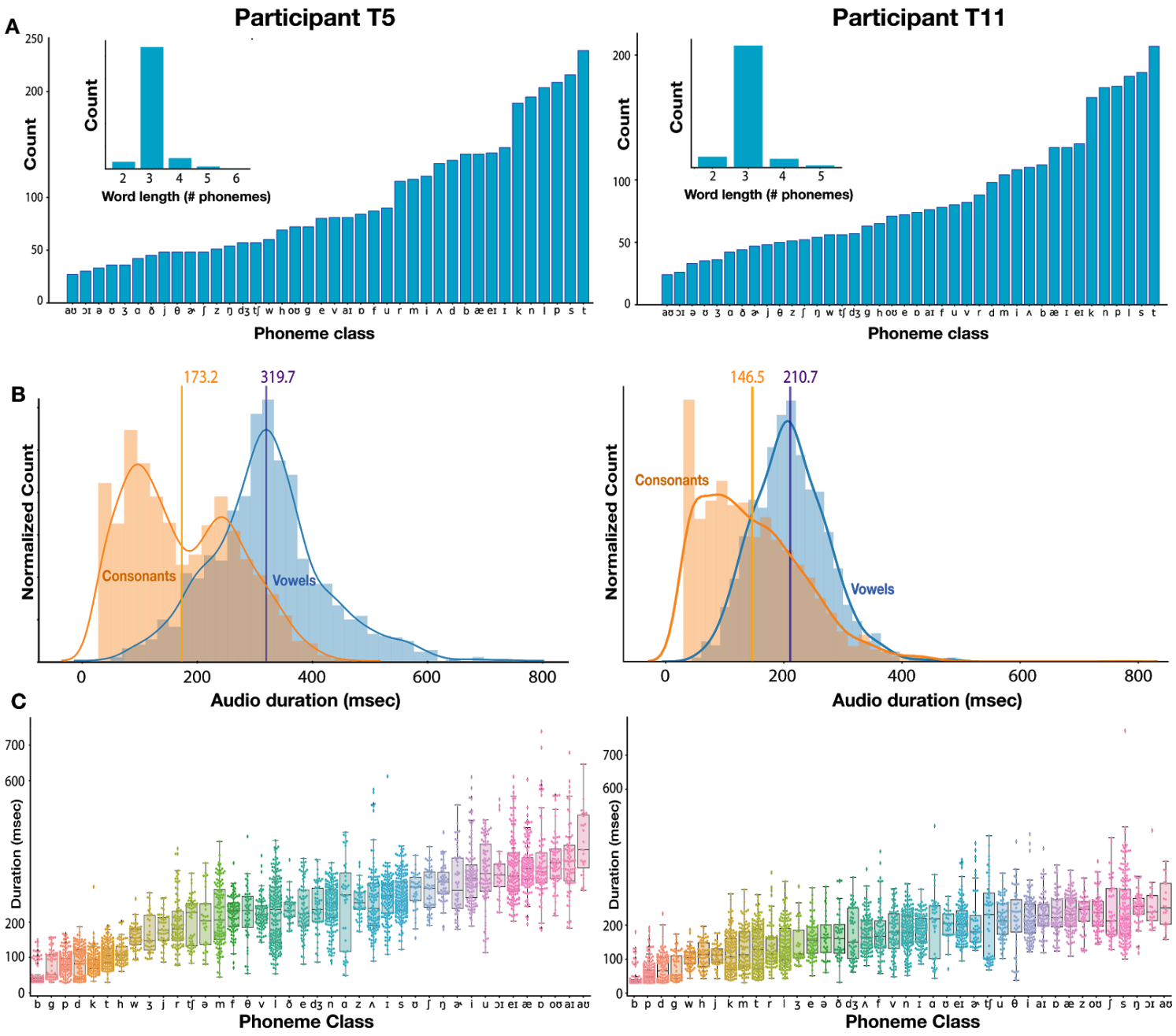
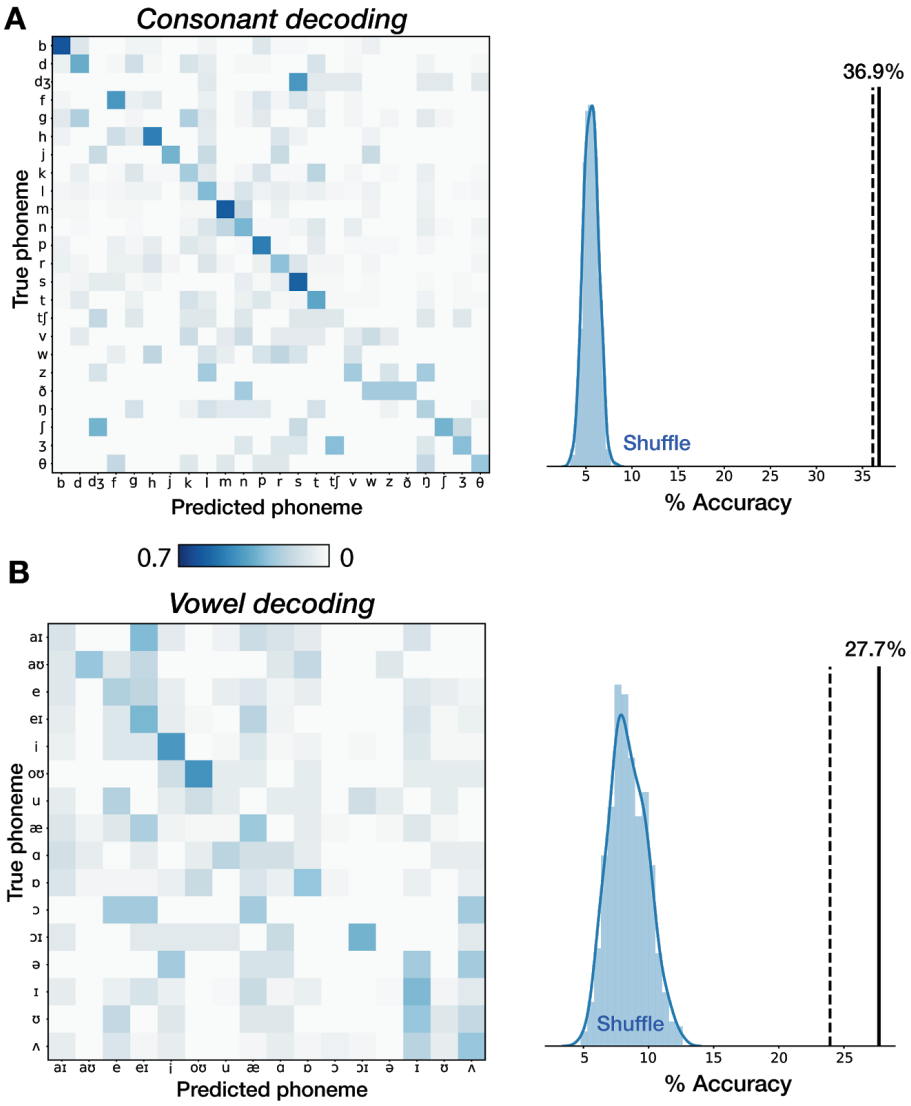
Supplementary Figures



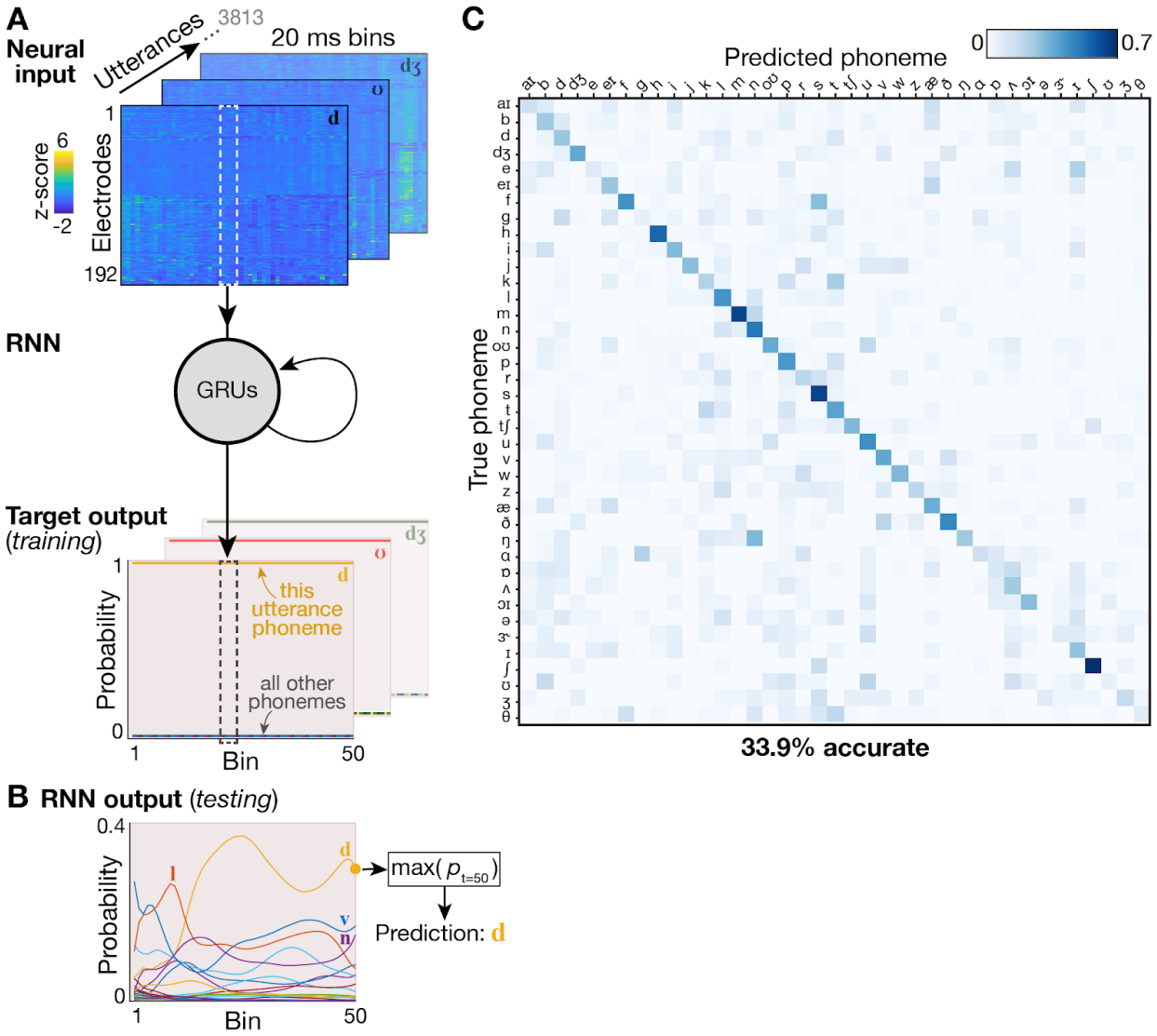
Supplementary Figure 1. Spoken word and phoneme statistics.

(A) Distribution of different phoneme class frequencies (T5 - min: 27, max: 239; T11 - min: 24, max: 207). The exact utterance distributions differ between participants due to occasional missed trials or misspeaking. Insets show the distribution of word lengths. A majority (87% for T5, 85% for T11) of words are 3 phonemes long. **(B)** Distribution of phoneme audio durations in milliseconds, split by vowels and consonants. Vertical lines with number labels show each class' mean. Vowels are longer on average. **(C)** Distribution of phoneme durations, broken down by specific class. Box plots display the median (middle horizontal lines), interquartile range (upper and lower lines), and outliers (Lilliputian diamonds).



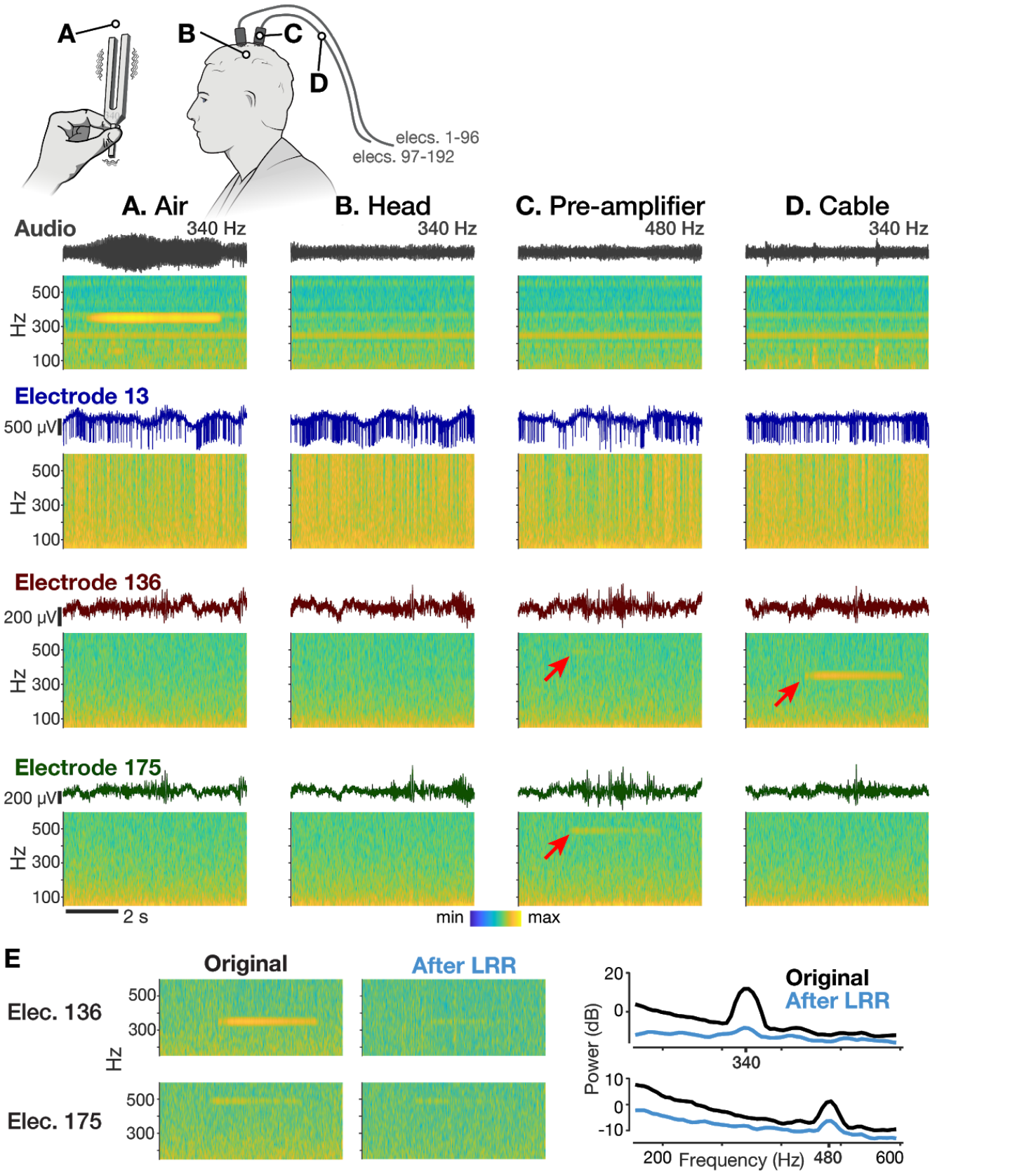
Supplementary Figure 2. Comparing decoding accuracy to a previous ECoG study with similar spoken words

Phonemes were classified using a 600 ms neural window with 50 ms, non-overlapping bins. Cross-validated decoding accuracy across 20 folds is reported. (**A**) Confusion matrix when decoding 24 consonants using a phoneme set closely matched to (Mugler *et al.* 2014). Measured intracortical decoding performance (36.8%, solid line) is comparable with that of Mugler *et al.* 2014 (36.1%, dashed line) and well above chance ($p < 0.002$; permutation test, 500 permutations). (**B**) same as (A) but with 17 vowels. Measured performance (27.7%) is comparable with that of Mugler *et al.* 2014 (23.9%) and well above chance ($p < 0.002$; permutation test, 500 permutations).



Supplementary Figure 3. Phoneme decoding using a recurrent neural network (RNN).

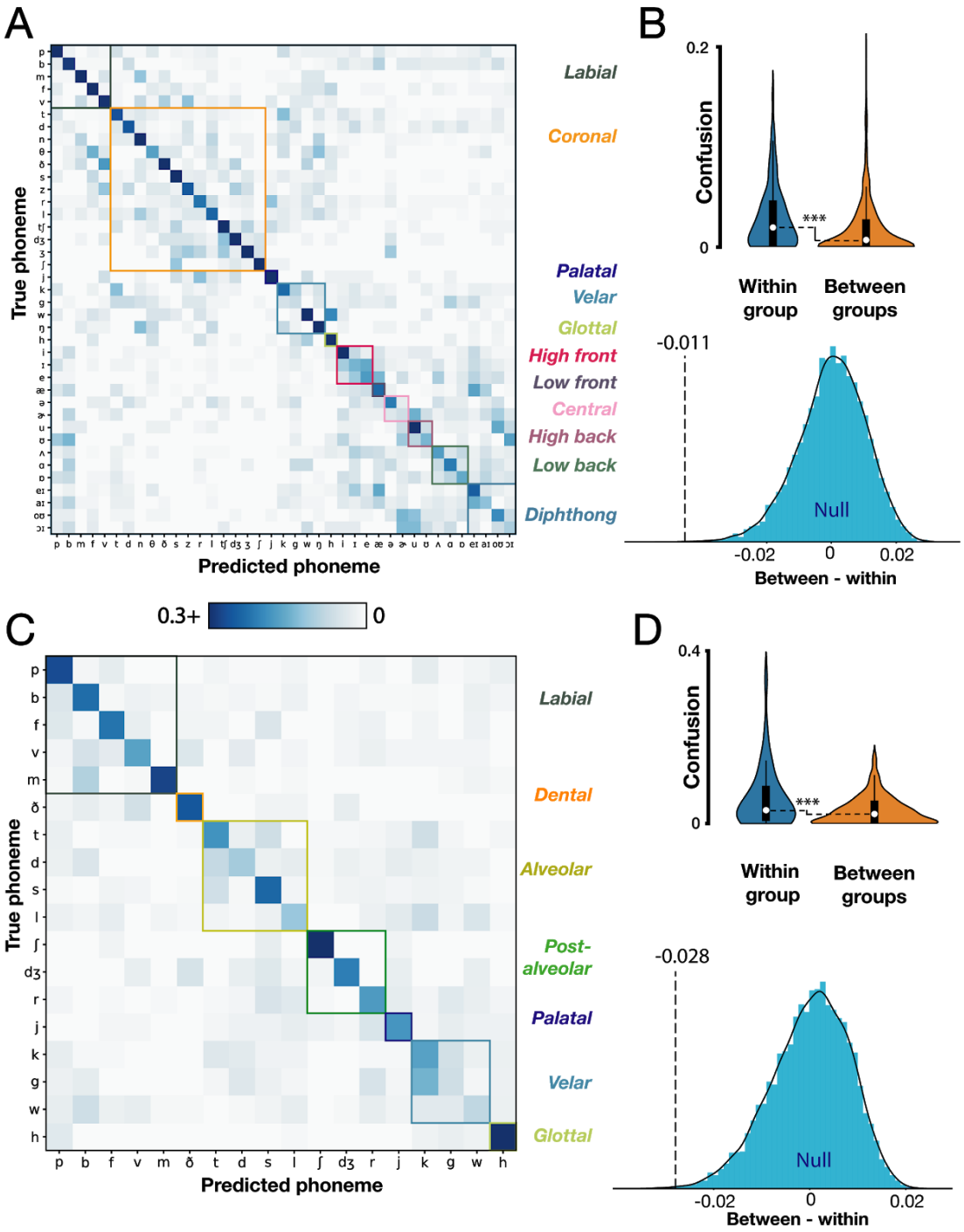
(A) Schematic overview of the RNN decoding approach. A single-layer RNN consisting of 512 gated recurrent units (GRUs) was trained to map neural inputs (top) at each time step to a ones-hot output (bottom) representing which phoneme utterance this neural data snippet came from. Each utterance provided one input snippet consisting of 1000 ms of HLFP activity per electrode (divided into fifty 20 ms bins) centered on the phoneme onset. **(B)** Example network output when the RNN was provided held-out neural data as input. Although estimated probabilities for each phoneme are read out during every time bin, for the utterance's final discrete output we used the most probable phoneme at the last time bin (here, that would be the phoneme /d/). **(C)** Cross-validated confusion matrix using the RNN to classify all of T5's phonemes. The 33.9% overall accuracy is slightly improved compared to the 29.6% accuracy when using a linear decoder (logistic regression, as in Fig. 3) with identical cross-validation folds, 1000 ms of HLFP activity, and using all electrodes.



Supplementary Figure 4. Neural, tuning fork recordings reveal microphonic pickup.

In this positive control, a 340 Hz or 480 Hz vibrating tuning fork was held in the air (**A**), gently pressed against participant T5's head (**B**), pressed against the pre-amplifier (**C**), or pressed against the cable (**D**). The top row shows a snippet of audio channel recording and corresponding acoustic power spectrogram for each condition. The remaining rows show simultaneous raw voltages and power spectrograms from three example electrodes. Electrode 13 was chosen as an example with no apparent microphonic artifact (it also has prominent action potentials). Electrode 136 showed a

microphonic artifact at the tuning fork's frequency (marked with a red arrow) in the pre-amplifier (C) and cable (D) vibration conditions. Electrode 175 showed an artifact in the pre-amplifier (C) condition, but not the cable condition. We only observed this artifact on the medial array (electrode numbers ≥ 97). The artifact could be generated with either the 340 Hz or 480 Hz tuning forks applied to either of the two pre-amplifiers/cables, but was stronger when applied to the medial pre-amplifier and cable, as in the examples shown. (**E**) Applying LRR "decontamination" to the electrode array recordings as in **Fig. 5** substantially attenuates this microphonic artifact. Spectrograms (left) and power spectra (right) are shown for the example electrode 136 from the cable condition (top row) and electrode 175 from the pre-amplifier condition (bottom row).



Supplementary Figure 5. T5 classifier errors reflect articulatory groupings of phonemes.

(A) Decoder confusion matrix sorted by articulatory groups. Note that the color axis is rescaled to [0 : 0.3] to highlight confusion patterns. **(B)** top: empirical within- and between-group normalized errors (Gaussian kernel densities, box plots with medians and interquartile ranges overlaid). Comparison of within and between group means reveals significantly higher confusions within groups ($p < 0.0005$); bottom: corresponding null distribution when articulatory groupings are shuffled (blue) compared to the true empirical difference (black dotted line). **(C)** Confusion matrix for predicting neurally-realigned first phonemes of each word. **(D)** Associated significance testing results, presented as in panel B. A permutation test revealed higher confusion within groups ($p < 0.002$) even after correcting for the biases in audio-derived voice onsets.

Supplementary Audio 1. Good examples of Brain-to-Speech synthesized audio.
Forty seven example word utterances were chosen to showcase reconstruction quality approaching intelligibility. For each example word, the true recorded audio spoken by T5 is played first, followed by the audio synthesized from intracortical HLFP neural data. Duration: 3m29s.

Supplementary Audio 2. Random examples of Brain-to-Speech synthesized audio.
Similar to Supplementary Audio 1, except these forty seven example word utterances were randomly chosen from all trials *except* those included in Supplementary Audio 1. Duration: 3m28s.