

Supplementary Materials for

## **Towards reconstructing intelligible speech from the human auditory cortex**

Hassan Akbari<sup>1,2</sup>, Bahar Khalighinejad<sup>1,2</sup>, Jose L. Herrero<sup>3,4</sup>, Ashesh D. Mehta<sup>3,4</sup>, Nima Mesgarani<sup>1,2</sup>

<sup>1</sup>*Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY*

<sup>2</sup>*Department of Electrical Engineering, Columbia University, New York, NY*

<sup>3</sup>*Hofstra Northwell School of Medicine, Manhasset, NY, United States*

<sup>4</sup>*The Feinstein Institute for Medical Research, Manhasset, NY, United States*

correspondence to: [nima@ee.columbia.edu](mailto:nima@ee.columbia.edu)

**This PDF file includes:** Tables. T1 to T6, Figures. S1 to S4, audio samples

<i>Input</i>		
<b>Feature Extraction</b>		
FCN	CNN	LCN
Flatten()	-	-
Dense (256) BN LeakyReLU(0.25) DropOut(0.3)	Conv2D (32) , [3×3] BN LeakyReLU(0.25) DropOut(0.3)	LCN(1), [5×5] BN LeakyReLU(0.25) DropOut(0.3)
Dense (256) BN LeakyReLU(0.25) DropOut(0.3)	Conv2D (64) , [3×3] BN LeakyReLU(0.25) DropOut(0.3)	LCN(1), [3×3] BN LeakyReLU(0.25) DropOut(0.3)
Dense (256) BN LeakyReLU(0.25) DropOut(0.3)	Conv2D (64) , [3×3] BN LeakyReLU(0.25) DropOut(0.3)	LCN(2), [1×1] BN LeakyReLU(0.25) DropOut(0.3)
-	Conv2D (32) , [1×1] BN LeakyReLU(0.25) DropOut(0.3)	-
-	Flatten()	Flatten()
<b>Feature Summation</b>		
Dense (256) BN ELU(alpha=1)		
<i>Output</i>		
ReLU OR tanh		

Table I. Structure of the proposed FCN, CNN, and LCN neural networks in feature summation and feature extraction networks.

Measure	LR (spectrogram)	DNN (spectrogram)	DNN (Vocoder)	DNN+AEC (Vocoder)
ESTOI	0.31	0.40	0.30	<b>0.47</b>

Table 2. Objective measures for intelligibility of the reconstructions using linear regression (LR) and DNN with spectrogram and vocoder parameters, with and without the autoencoder (AEC) network. The network with the autoencoder achieves higher reconstruction accuracy.

Bottleneck size	32	64	128	256
ESTOI	<b>0.47</b>	0.46	<b>0.47</b>	<b>0.47</b>

Table 3. Objective intelligibility and quality measures of the synthesized speech based on coded-decoded vocoder parameters using AEC with different number of nodes in the bottleneck layer. The number of bottleneck nodes was chosen to maximize the objective and subjective intelligibility, which was highest for both for 256 nodes.

Network	FCN	CNN	LCN	FCN+CNN	FCN+LCN
ESTOI	0.45	0.41	0.43	0.42	<b>0.47</b>

Table 4. Objective intelligibility measures of the synthesized speech based on coded-decoded vocoder parameters using AEC with different types of network. The feature summation network in all cases is the same (FCN), but different networks architectures were used for feature extraction. Reconstructing the vocoder parameters is best when a FCN+LCN network is used as feature extraction.

Network	FCN	CNN	LCN	FCN+CNN	FCN+LCN
ESTOI	<b>0.40</b>	0.41	0.37	0.41	0.40

Table 5. Objective intelligibility measures of the synthesized speech based on spectrogram with different types of network. Feature summation network in all cases is the same (FCN). Because no significant difference was observed, we chose the network with lowest complexity which was the FCN model.

Input(513+1+1+1)			
Dense(512)			
LeakyReLU			
Dense(400)			
LeakyReLU			
Dense(300)			
LeakyReLU			
Dense(256)			
Tanh (Bottleneck Features)			
AdditiveGaussianNoise			
Dense(300)			
LeakyReLU			
Dense(400)			
LeakyReLU			
Dense(512)			
LeakyReLU			
Dense(512)	Dense(32)	Dense(32)	Dense(32)
LeakyReLU	LeakyReLU	LeakyReLU	LeakyReLU
Dense(513)	Dense(8)	Dense(8)	Dense(8)
ReLU	LeakyReLU	LeakyReLU	LeakyReLU
	Dense(1)	Dense(1)	Dense(1)
	ReLU	ReLU	ReLU

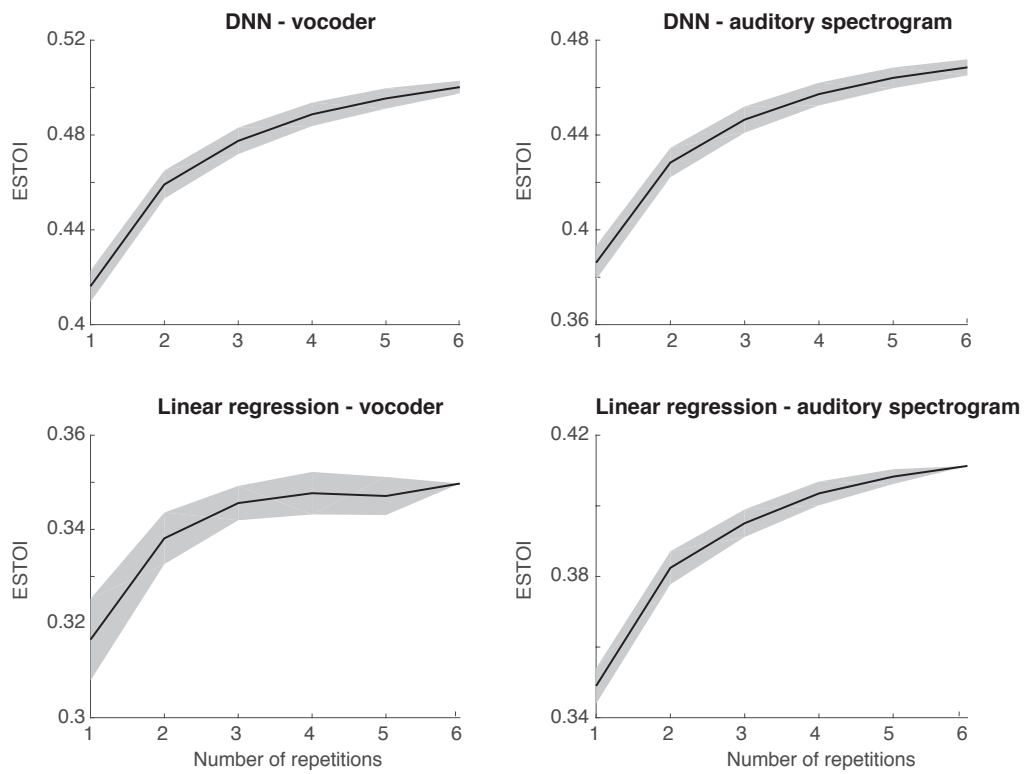
Spectrogram

Aperiodicity

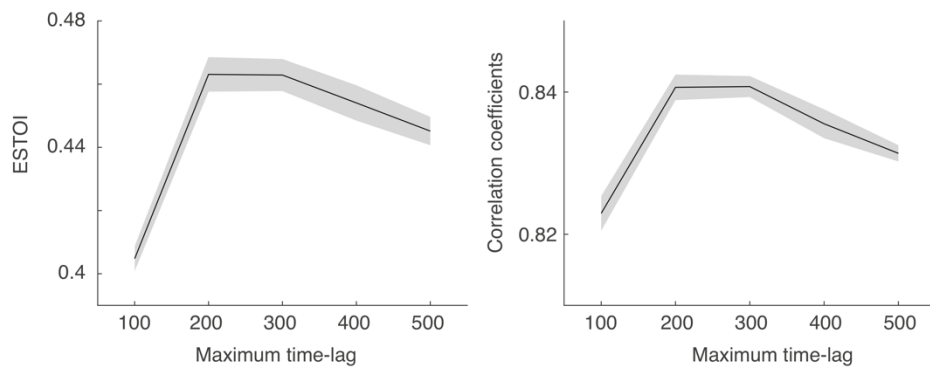
F0

VUV

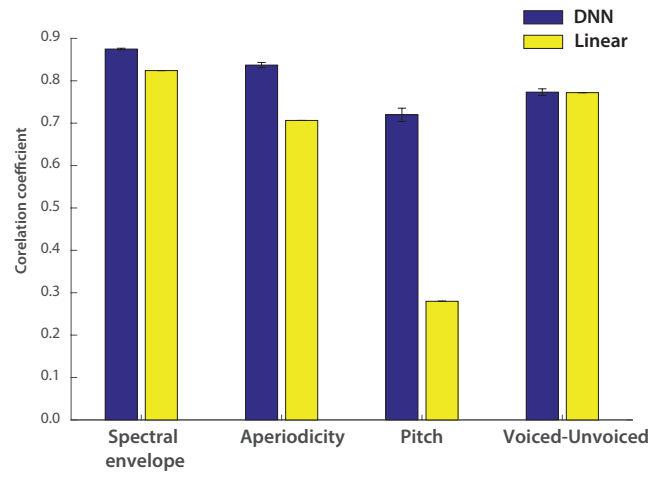
Table 6. The architecture and the parameters of the autoencoder (AEC) network.



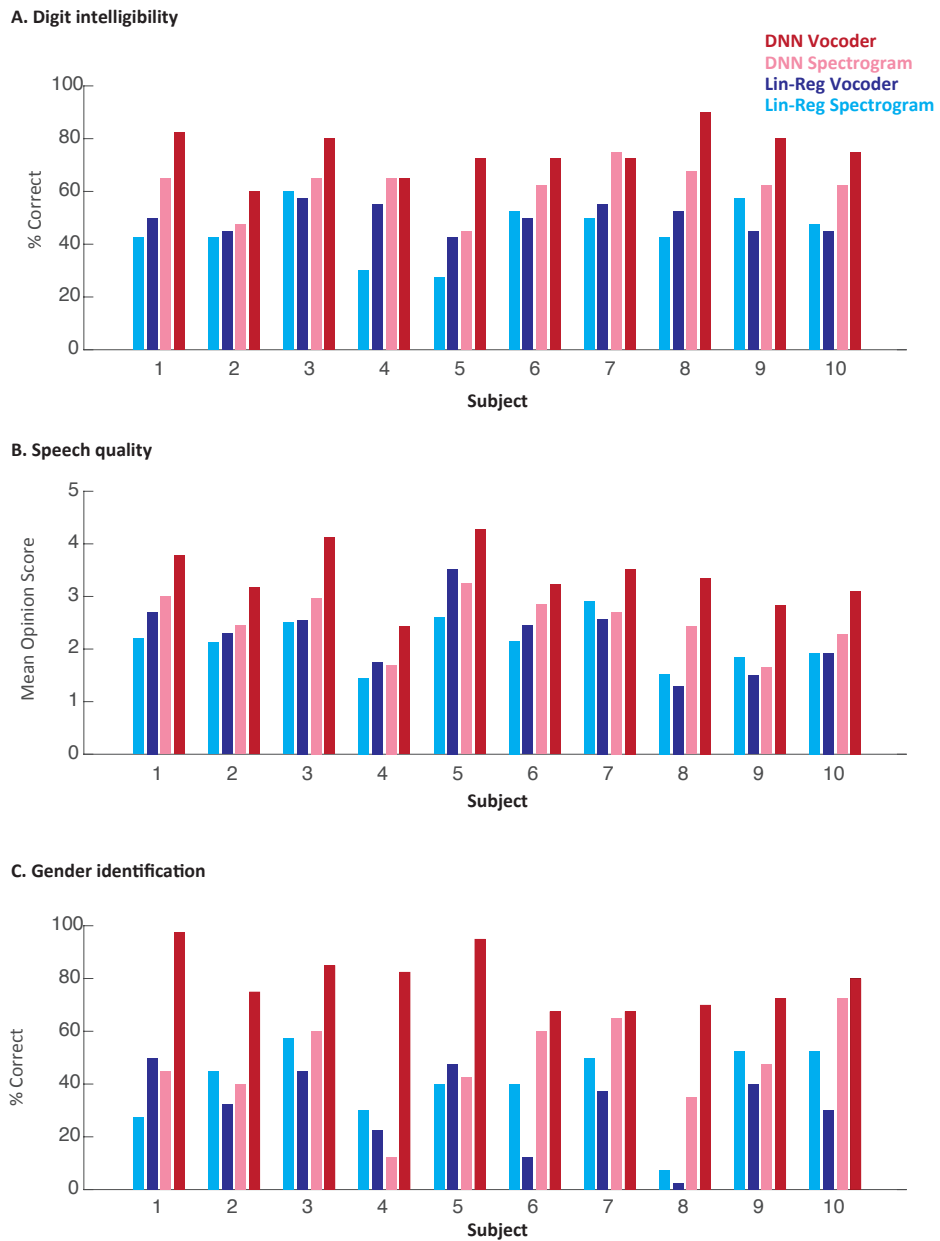
**Supplementary Figure 1:** ESTOI scores of the four models for speech utterances reconstructed from the averaged neural responses over an increasing number of stimulus repetitions. Shaded lines show the standard deviation.



**Supplementary Figure 2:** Reconstruction accuracy (Pearson's correlation) and ESTOI scores for reconstructed speech sounds when varying the length of the sliding window.



**Supplementary Figure 3:** Reconstruction accuracy (Pearson's correlation) for individual parameters of the WORLD vocoder model: spectral envelope, aperiodicity, pitch, and voiced-unvoiced.



**Supplementary Figure 4:** Subjective evaluation of the reconstruction models reported separately for the ten listeners. Scores are reported for A) intelligibility, B) quality, and C) gender identification.



### Supplemental Movie 1:

The audio files are reconstructed digits from the neural responses in the auditory cortex of listeners. These sounds were used in the subjective intelligibility and quality test (Fig. 3). The digits sounds were taken from a publicly available corpus, T1-46 (See methods). Samples below are from a female (F8) and a male (M5) speaker. For each speaker, the sounds are reconstructed using 4 models: Linear spectrogram, linear vocoder, DNN spectrogram, and DNN vocoder.

F8\_In\_aud: female speaker, linear regression to spectrogram  
F8\_In\_vocoder: female speaker, linear regression to vocoder  
F8\_dnn\_aud: female speaker, DNN to spectrogram  
F8\_dnn\_vocoder: female speaker, DNN to vocoder

M5\_In\_aud: male speaker, linear regression to spectrogram  
M5\_In\_vocoder: male speaker, linear regression to vocoder  
M5\_dnn\_aud: male speaker, DNN to spectrogram  
M5\_dnn\_vocoder: male speaker, DNN to vocoder