Supplementary Materials for

Reconstructing intelligible speech from human auditory cortex

Hassan Akbari[1,2], Bahar Khalighinejad[1,2], Jose L. Herrero[3,4], Ashesh D. Mehta[3,4], Nima Mesgarani[1,2]

[1]*Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY*

[2]*Department of Electrical Engineering, Columbia University, New York, NY*

[3]*Hofstra Northwell School of Medicine, Manhasset, NY, United States*

[4]*The Feinstein Institute for Medical Research, Manhasset, NY, United States*

correspondence to: nima@ee.columbia.edu

This PDF file includes: Tables. T1 to T6, audio samples

Input		
Feature Extraction		
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()
Feature Summation		
Dense (256) BN ELU(alpha=1)		
Output 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	**0.47**

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	**0.47**	0.46	**0.47**	**0.47**

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 (256 nodes).

Network	FCN	CNN	LCN	FCN+CNN	FCN+LCN
ESTOI	0.45	0.41	0.43	0.42	**0.47**

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	**0.40**	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.

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, TI-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_ln_aud:	female speaker, linear regression to spectrogram
F8_ln_vocoder:	female speaker, linear regression to vocoder
F8_dnn_aud:	female speaker, DNN to spectrogram
F8_dnn_vocoder:	female speaker, DNN to vocoder

M5_ln_aud:	male speaker, linear regression to spectrogram
M5_ln_vocoder:	male speaker, linear regression to vocoder
M5_dnn_aud:	male speaker, DNN to spectrogram
M5_dnn_vocoder:	male speaker, DNN to vocoder