Supplemental Methods

Dataset and pre-processing

The Toronto Emotional Speech Set (TESS), containing 200 spoken utterances for each of 7 different emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral) and 2 speakers (14), was used for training and testing our networks. Individual audio files spanned durations of 1-3 seconds and were zero-padded to be of identical length. The dataset was divided into 10 equal-sized folds, yielding 10 different data splits with train/test set ratios of 90/10% each. The networks were trained and tested on all 10 folds, to investigate the stability of the findings. Figures 2A, 2B, 2C, and 3A depict results of a single fold (the first one), so as to allow the visualization of individual filters' data points, instead of providing summary statistics, while all other figures reflect results obtained by pooling (Figures 3B, 3C, and 4C) or averaging

(Figures 4A&B) across all 10 folds. The dataset has a native sampling frequency of 24414 Hz, which was used for training in 'full-frequency' conditions. The training on 'low-frequency'-conditions followed the application of a low-pass filter in the frequency domain, using a cut-off frequency of 500 Hz, as inspired by previous auditory recordings (2). For the testing of classification performances reported in Figure 4A, a wide range of low-pass filters were applied, with cut-off frequencies of 500, 550, 600, 650, 700, 750, 1000, 1250, 1500, 1750, and 2000 Hz.

Network architecture, parameters, and training procedure

We utilized the "M5" model by (13), equipped with 4 convolutional layers (each involving convolution, batch normalization, application of the ReLU activation function, and max pooling), followed by global average pooling, and connected to the output nodes of the network, representing the 7 different emotion classes, through a single dense layer with softmax activation function. The number of units in the different layers were taken from (13), with the exception of the kernel size in the first convolutional layer, which, due to a different sampling frequency in the dataset, was adjusted to 244, to span a 10ms window of the auditory input – a time window frequently used in MFCC (Mel Frequency Cepstral Coefficients)-based computational audition models as suggested by the authors of the computational model used (13). The network was implemented in Keras and trained on a single GPU, using stochastic gradient descent with a batch size of 32 and a standard learning rate of 0.01. All four regimens were trained for a total of 100 epochs ('low-to-full' for 50 epochs on low-frequency and 50 epochs on full-frequency inputs, 'full-to-low' for 50 epochs on full-frequency and 50 epochs on low-frequency inputs, 'exclusively-full' for 100 epochs on full-frequency inputs, and 'exclusively-low' for 100 epochs on low-frequency inputs). The number of epochs was chosen to ensure that the regimens had converged and showed little variation between runs, to ensure comparability and stability of results.

Analyses of representations and activations

For the receptive field analysis reported in Figures 2 and 3, we examined the filters in the first convolutional layer of the networks by subjecting them to the Fourier transform. Figures 2A-B depict the power values extracted for frequencies between 0.1 and 6 kHz, in steps of 0.1 kHz, with the sum up to 6 kHz normalized to 1. We thereby chose a maximum of 6 kHz (representing approximately half of the Nyquist frequency) to increase the reliability of the analysis and peak frequency detection, with only a small fraction of the signal, but an increased potential for artifacts, left between half of the Nyquist frequency and the Nyquist frequency. Note that these frequency values, and the ones depicted on the axes in Figure 2, are rounded: With a frequency increment of 24414/244, the rounded frequency values (0.1, 0.5, 1, 2, 3, 4, and 6 kHz) correspond to 100.0574, 500.2869, 1000.5738, 2001.1475, 3001.7213, 4002.2951, and 6003.4426 Hz. For the correlational analysis reported in Figure 4C, for each trained network and each of the 10 folds, correlations were computed between the activations of units, concatenated for all 280 test set items, following full-frequency vs. low-frequency inputs. Figure 4C depicts the distribution of these individual units' correlations.