Efficient coding in human auditory perception

Vivienne L. Ming, and Lori L. Holt

Citation: The Journal of the Acoustical Society of America **126**, 1312 (2009);
View online: https://doi.org/10.1121/1.3158939
View Table of Contents: http://asa.scitation.org/toc/jas/126/3
Published by the Acoustical Society of America

Efficient coding in human auditory perception

Vivienne L. Ming[a)]
Redwood Center for Theoretical Neuroscience, University of California at Berkeley, 156 Stanley Hall, MC 3220, Berkeley, California 94720

Lori L. Holt
Department of Psychology and the Center for the Neural Basis of Cognition, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213

(Received 8 February 2008; revised 14 March 2009; accepted 3 June 2009)

Natural sounds possess characteristic statistical regularities. Recent research suggests that mammalian auditory processing maximizes information about these regularities in its internal representation while minimizing encoding cost [Smith, E. C. and Lewicki, M. S. (2006). Nature (London) **439**, 978–982]. Evidence for this "efficient coding hypothesis" comes largely from neurophysiology and theoretical modeling [Olshausen, B. A., and Field, D. (2004). Curr. Opin. Neurobiol. **14**, 481–487; DeWeese, M., *et al.* (2003). J. Neurosci. **23**, 7940–7949; Klein, D. J., *et al.* (2003). EURASIP J. Appl. Signal Process. **7**, 659–667]. The present research provides behavioral evidence for efficient coding in human auditory perception using six-channel noise-vocoded speech, which drastically limits spectral information and degrades recognition accuracy. Two experiments compared recognition accuracy of vocoder speech created using theoretically-motivated, efficient coding filterbanks derived from the statistical regularities of speech against recognition using standard cochleotopic (logarithmic) or linear filterbanks. Recognition of the speech created using efficient encoding filterbanks was significantly more accurate than either of the other classes. These findings suggest potential applications to cochlear implant design. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3158939]

PACS number(s): 43.66.Ba, 43.71.An, 43.66.Ts, 43.72.Gy [RSN] Pages: 1312–1320

I. INTRODUCTION

Perceptual systems are limited capacity channels in that they may encode and transmit only a finite amount of information over any period of time. Mirroring the bandwidth issues that plagued early telecommunications and now electronic information exchange, perceptual systems face the seemingly intractable dilemma of coding efficiency; they must balance high-fidelity information transmission against the overall encoding cost to the system.

Although the problem of transmitting a high-fidelity, low-cost code may seem intractable, information theory states that optimally efficient codes, which carry the most information at the lowest cost, should match the statistics of the signals they represent (Shannon, 1948; MacKay, 2003). A large body of evidence from theoretical and empirical research in vision (Olshausen and Field, 1996; Sharpee *et al.*, 2006) suggests that efficiency may be central to perceptual encoding (Barlow, 1961; Atick, 1992; Simoncelli and Olshausen, 2001; Laughlin and Sejnowski, 2003).

Recent empirical and theoretical research (Rieke *et al.*, 1995; Attias and Schreiner, 1998; Lewicki, 2002; Klein *et al.*, 2003) has indicated that these principles extend to the auditory system. Smith and Lewicki (2006), for example, showed that auditory nerve response matches a theoretically-predicted efficient code for representing the diverse sounds of natural acoustic environments. In other words, the cochlear code reflects the statistics, both spectral power and higher-order (phase) statistics, of natural sounds. At a neural level, increased coding efficiency of natural signals has been repeatedly demonstrated (Rieke *et al.*, 1995; Attias and Schreiner, 1998; Vinje and Gallant, 2002; Sharpee *et al.*, 2006). Afferent fibers from the peripheral auditory system of the bullfrog better encode sounds with the spectrum of mating calls than broad-band noise (Rieke *et al.*, 1995). Similarly, neurons in the cat's inferior colliculus exhibit increased coding efficiency for narrow-band noise with "naturalistic" amplitude modulations versus "non-naturalistic" modulations (Attias and Schreiner, 1998; Escabi *et al.*, 2003). Although these results support the efficient coding hypothesis in neural auditory processing, they provide no direct insight into the extent to which observed neural coding differences have behavioral consequences in human perception.

In the present research, we examine this question directly by measuring human speech recognition under challenging perceptual circumstances. The underlying hypothesis guiding this work is that if coding efficiency has behavioral consequences, complex sounds created to match the statistics of natural sounds should have a perceptual advantage over sounds that diverge from environmental statistics. We use noise-excited vocoder speech (often used to mimic cochlear implant output in normal-hearing listeners; Shannon *et al.*, 1995), to create a challenging auditory perceptual task within which this advantage might be measured as gradations in speech intelligibility.

[a)]Author to whom correspondence should be addressed. Electronic mail: neuraltheory@gmail.com

II. VOCODING

In noise-vocoded speech, sounds are stripped of their fine spectral resolution and left with only their amplitude envelope via an algorithm similar to that used in cochlear implants (Zeng *et al.*, 2004b). A filterbank composed of limited number of filters (6 in the present work) separates speech sounds into a set of band-limited channels, with the choice of filterbank determining the frequency bands (e.g., linear versus logarithmic frequency tiling). The amplitude envelope of each channel, the slowly time-varying dynamics of the speech within that frequency band, is separated from its fine spectral detail via half-wave rectification followed by a 150 Hz low-pass filter. Each of these resulting envelopes is used to modulate the output of the Gaussian noise, giving the noise the low-frequency temporal dynamics of the original speech. Finally, each channel of modulated noise is again filtered so that its frequency range matches that of the original channel, and they are added back together, producing a single waveform. Through this process, noise-vocoded speech preserves the temporal dynamics of its limited number of frequency channels but has no spectral resolution within each channel, though some spectral information can be recovered by integrating information across channels (Nie and Zeng, 2004) allowing listeners hear some or all of the original speech steam. The change between original speech and its vocoder counterpart is illustrated in Fig. 1 where four spectrograms show how the acoustic frequency changes across time. Natural speech has complex spectral characteristics [Fig. 1(a)]. After vocoder transformation with six frequency channels, the sound loses nearly all fine spectral detail; however, the temporal envelopes of the six channels remain [Figs. 1(b)–1(d) showing three different choices of filterbanks]. Although the frequency information is severely degraded, the envelope retains many important cues for speech perception (Shannon *et al.*, 1995; Smith *et al.*, 2002). The resulting sounds can be quite difficult to understand but are clearly speech-like and, for a six-channel vocoder, reasonably intelligible with some practice.

Key to the our investigation of the consequences of efficient coding on human auditory perception, the content of each channel (and the qualities of sound produced) depends on the characteristics of the filterbank. Their experiments manipulate the form of the six filters comprising the filterbank affecting, for example, how they tile the frequency dimension, as illustrated in Fig. 2. The set of filters shown in Fig. 2(a) ("linear") simply tiles temporally-symmetric, equal-bandwidth band-pass filters linearly across the frequency dimension. The second set of filters [Fig. 2(b), "cochleotopic"] is more natural in that it mimics the near-logarithmic frequency-coding characteristics of the cochlea whereby lower frequencies are sampled with finer resolution (smaller bandwidths) than are higher frequencies (Bekesy, 1960; Greenwood, 1961). A spectrogram of vocoder speech using this filterbank is shown in Fig. 1(b). Noise-excited vocoder speech processed with a cochleotopic filterbank is generally better understood than speech processed using a linearly-tiled filterbank (Shannon *et al.*, 2003), but it is unclear whether the advantage reflects a better match to the frequency representation of the auditory system or the spectral statistics of natural sounds.

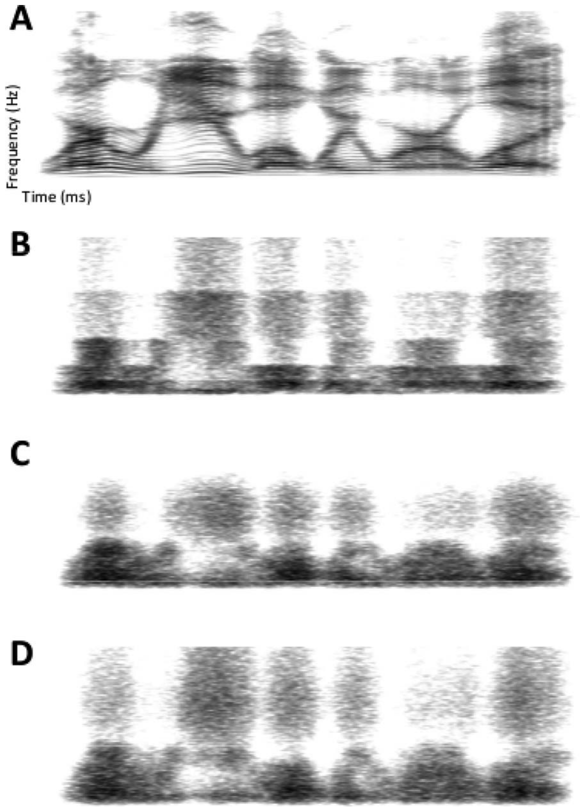

FIG. 1. Spectrograms of the utterance "Where were you while we were away?" Time unfolds along the *x*-axis, and frequency is presented along the *y*-axis and amplitude is illustrated with intensity (dark is higher amplitude). Unmodified speech (a) possesses fine spectral detail. Vocoder transformations of this utterance using three different, six-channel filterbanks [(b) cochleotopic; (c) efficient gammatone-smoothed; (d) efficient spline-smoothed, here with six channels] compress the spectral information within a channel so that only temporal modulation of six coarse frequency bands remains. The choice of filterbank determines how the spectral information is partitioned and influences the amplitude modulations extracted from each channel.

It is also possible, however, that the cochleotopic filterbank better reflects the statistical structure of speech acoustics, and that it is this quality which drives the improved performance. It is not possible to distinguish these two hypotheses comparing perception using cochleotopic versus linear filterbanks as the cochleotopic set matches both the biology and the sound statistics better than does the linear set. To address this confound, we will use a machine learning algorithm to analyze the statistics of speech acoustics and design a new filterbank to reflect the statistical structure.

III. EFFICIENT CODING HYPOTHESIS

According to the efficient coding hypothesis, perception should be optimally adapted to the statistics of natural signals such that they carry the most information at the least cost. Therefore, perceptual performance should be best when sensory codes match the statistics of environmental stimuli. To test this prediction, we use a computational model of efficient auditory coding (Smith and Lewicki, 2006) to opti-

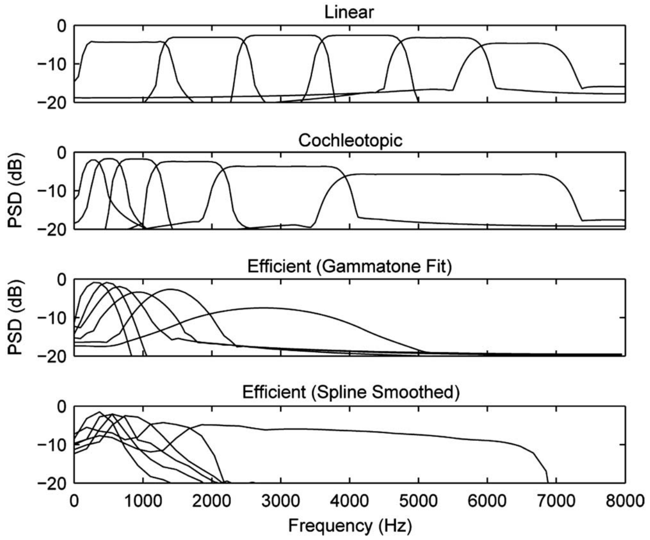

FIG. 2. The power spectra for all six filters from the four different filterbanks are shown. The top row shows the frequency tiling of the "linear" filterbank. The linear frequency tiling of the set can be clearly seen to the right. The cochleotopic frequency tiling is shown in row B. The next two rows show the learned "efficient" filters smoothed either by gammatone fitting or spline-smoothing.

mize a set of functions with respect to the information carried by the large TIMIT speech corpus training set (Garofolo *et al.*, 1990). This model allows an explicit prediction of the dimensions of perceptual sensitivity. In it, sound, $x(t)$, is generated by a linear superposition of a set of functions, $\varphi_1, \ldots, \varphi_M$, which can be positioned arbitrarily and independently in time. The mathematical form of the representation with additive noise is

$$x(t) = \sum_{m=1}^{M}\sum_{i=1}^{n_m} s_i^m \phi_m(t-\tau_i^m) + \varepsilon(t), \tag{1}$$

where τ_i^m and s_i^m are the temporal position and coefficient of the ith instance of kernel ϕ_m, respectively. The notation n_m indicates the number of instances of ϕ_m, which need not be the same across kernel functions. The kernel functions are not restricted in form or length, and both the kernel shapes and their lengths were adapted to optimize coding efficiency; in the results below, the kernels take on a variety of shapes and range in length from 10 to 100 ms. This provides a mathematical description of sound waveforms that has sufficient flexibility to encode arbitrary acoustic signals and encompass a broad range of potential auditory codes.

The key theoretical abstraction of the model is that the acoustic signal can be encoded most efficiently by decomposing it in terms of discrete acoustic elements, each of which has a precise amplitude and temporal position. This also yields a code that is time-relative and does not depend on artificial blocking of the signal (Smith and Lewicki, 2005a, 2005b). One interpretation of each analog τ_i^m, s_i^m pair is that it represents a local population of (binary) auditory nerve spikes firing probabilistically in proportion to the underlying analog value.

To code speech sounds efficiently, we need to determine both the optimal values of τ_i^m and s_i^m (*encoding*) and the optimal kernel functions ϕ_m (*learning*). From Eq. (1), coding efficiency can be defined approximately as the number of "spikes" (nonzero coefficient values) required to achieve a desired level of precision, which is defined by the variance of the additive noise $\varepsilon(t)$. This assumes that the goal of coding is to represent the entire acoustic signal and that coding efficiency is most closely related to the number of spikes in the code. Other definitions are possible within this framework, but this definition has the advantage of starting from a minimal set of assumptions.

Although the generative form of the model is linear, in other words the signal is a linear function of the representation, inferring the optimal representation for a signal is highly non-linear and computationally complex. Here we compute the values of τ_i^m and s_i^m for a given signal by using a matching pursuit algorithm (Mallat and Zhang, 1993), which iteratively approximates the input signal and has been shown to yield highly efficient representations for a broad range of sounds (Smith and Lewicki, 2005a, 2005b). In matching pursuit, the current residual signal (initialized as the original sound) is projected onto the dictionary of kernel functions. The projection with the largest inner product is subtracted out, and its coefficient and time recorded. For the results reported here, the encoding halts when s_i^m falls below a pre-set "spiking" threshold.

The goal of learning in the efficient coding model is to find a set of functions for which the coefficients are maximally efficient (i.e., carry the most information about the sound at the lowest cost) with respect to the given training data. We can rewrite Eq. (1) in probabilistic form in which we assume that the noise is Gaussian and the prior probability of a spike, $p(s)$, is sparse (i.e., comes from a probability distribution which produces very few nonzero values). The kernel functions are optimized by performing gradient ascent on the approximate log-data probability,

$$\begin{aligned}\frac{\partial}{\partial\phi_m}\log(p(x|\phi)) &= \frac{\partial}{\partial\phi_m}\log(p(x|\phi,\hat{s})) + \log(p(\hat{s}))\\ &= \frac{1}{2\sigma_\varepsilon}\frac{\partial}{\partial\phi_m}\left[x - \sum_{m=1}^{M}\sum_{i=1}^{n_m}\hat{s}_i^m[x-\hat{x}_{\tau_i^m}]\right]^2\\ &= \frac{1}{\sigma_\varepsilon}\sum_i \hat{s}_j^m[x-\hat{x}]_{\tau_i^m},\end{aligned} \tag{2}$$

where $[x-\hat{x}]_{\tau_i^m}$ indicates the residual error over the extent of kernel ϕ_m at position τ_i^m. The estimated kernel gradient is thus a weighted average of the residual error. For training here, we restrict the set to six functions, which were initialized as 100-sample Gaussian noise, and the spiking threshold (minimum value of s_i^m) was set at 0. Filters were derived from the resulting kernel functions using reverse correlation (Smith and Lewicki, 2006).

The filterbanks shown in Fig. 2(c) ["efficient (gammatone fit)"] and Fig. 2(d) ["efficient (spline smoothed)"] were learned using the efficient coding model (Smith and Lewicki, 2006) using two different smoothing methods to regularize the functions. These filters represent an optimal code for the statistical properties of the speech database when only six channels are available. The frequency tiling from the efficient coding model [Figs. 2(c) and 2(d)] is much more biased to the low frequencies than the cochleotopic model [Fig. 2(b)]. This can also be seen in the difference between the vocoder spectrograms in Fig. 1, which shows the difference between speech transformed using the cochleotopic vocoder [Fig. 1(b)], the gammantone-smoothed "efficient" vocoder [Fig. 1(c)] and the spline-smoothed efficient vocoder [Fig. 1(d)]. Moreover, the form of the efficient functions is not fixed; it combines both gammatone-like filters in the lower frequencies with broadly-tuned, symmetric filters at the higher frequencies.

If there is efficient coding in auditory processing, one would expect perceptual performance to align with the efficient filters. The distinction between the filters in Figs. 2(c) and 2(d) and the cochleotopic filters of Fig. 2(b) may seem counter-intuitive given that for larger filterbanks (30 +filters), the optimal filterbank very closely match both individual structure and population statistics of filters estimated from single-unit recordings of auditory nerve fibers (Smith and Lewicki, 2006). This correspondence is lost when many fewer are channels available, as with noise-vocoded speech or cochlear implants. Spectral resolution is limited and the resulting filter characteristics change from the cochleotopic filterbank typically thought to best characterize the frequency processing of the cochlea. For six-channel noise-vocoded speech, the optimally efficient code and the cochlear code diverge, providing a means to dissociate them experimentally.

If efficient coding carries perceptual benefits, speech recognition accuracy should be greatest for noise-vocoded speech created with efficient filters because these filters best characterize the statistics of speech within the limited capacity of six channels, preserving the available information. We explicitly tested this prediction by having adult human listeners transcribe noise-vocoded speech produced with the filterbanks shown in Fig. 2.

IV. METHODS

Following the approach of previous vocoder experiments (Shannon *et al.*, 1995), we measured speech intelligibility in two distinct tasks: identifying words in continuous speech (sentences, Experiment 1) and identifying phonemes from non-word utterances (non-words, Experiment 2).

For Experiment 1, there were 168 distinct English sentences (42 sentences/condition), each spoken by a different native-English speaker (TIMIT corpus: Garofolo *et al.*, 1990). The assignment of sentences to filtering conditions was counter-balanced such that, across participants, each sentence was presented in each of the four conditions but no sentence or speaker was repeated for an individual participant. Sentences ranged in length from 8–16 words (approximately 1–6 s) for a total of 1564 words. The sentences used in the experiment were drawn from the TIMIT testing set and were distinct from those used in the training of the computational model that produced the efficient filterbanks.

Four stimulus conditions were created by synthesizing vocoder versions of each item using one of three filterbanks plus using the original, unmodified speech as a control. The filterbanks were composed of six finite impulse response filters, with the number of filters chosen to produce sufficiently challenging stimuli so as to avoid ceiling effects. The linear and cochleotopic filterbanks were composed of six Hanning-window band-pass filters. The filters were tiled across 0–7 kHz with either linear or cochleotopic (logarithmic) placement (see Fig. 2).

For the efficient filterbanks, the "raw" (unsmoothed) filters comprising them were identical in both experiments. Kernel functions were trained on the 4956 sentences from the TIMIT training set. Training involved encoding a batch of 100 full sentences on each iteration and then updating the kernel functions based on the gradient estimated from the batch. Training continued until the set reach convergence, about 10,000 iterations. Filters were then derived from the functions. These filters were then smoothed to regularize the filters, removing residual noise from the learning algorithm. In experiment 1, the efficient filters were fitted with gammatone functions, a parametrized approximation of the learned functions composed of sine wave modulated by a gamma function [Fig. 2(c)].

Sixteen listeners participated in Experiment 1. Participants were college-age native-English speakers from Carnegie Mellon University with no reported or obvious speaking or hearing disorders. Participants received undergraduate Psychology course credit for participation. Seated in individual sound-attenuated booths, participants listened to each stimulus and typed what they heard. In Experiment 1, participants were told that some of the sentences may be difficult to understand, but a response must be made on each trial. They were allowed to hear each sentence only once. In neither experiment was there a pre-exposure or training period for the participants with the vocoder speech. Each participant listened to 62 stimuli from each condition (186 total). In both experiments, order of stimulus presentation was randomly permuted for each participant.

The ALVIN experiment-control software (Hillenbrand and Gayvert, 2005) was used for stimulus presentation and data collection. Acoustic presentation was under the control of Tucker Davis Technologies (Alachua, FL) System II hardware; stimuli were converted from digital to analog, amplified, and presented dichotically over linear headphones (Beyer DT-150, Berlin, Germany) at approximately 70 dB SPL(A).

The stimuli for Experiment 2 consisted of non-word syllables spoken in isolation (Shannon *et al.*, 1999). Each stimulus was composed of two distinct utterances of the same syllable separated by 500 ms of silence. The stimuli consisted of both vowel-consonant-vowel (VCV) syllables such as "aba" and consonant-vowel (CV) syllables such as "bi" representing the full range of combination described in Shannon *et al.*, 1999. Stimuli were drawn from a corpus of ten speakers (five male, five female) to include 46 unique syllables from each speaker. Both the sentence and non-word stimuli were sampled at 16 kHz with 16-bit resolution.

In Experiment 2, we used spline-smoothing to regularize the raw filters [Fig. 2(d)], which offers a less biased estimate of the underlying function than gammatone fitting. As can be seen in Fig. 2(d), the gammatone fitting rounded the power spectra of the filters, even cutting off the high-end of the

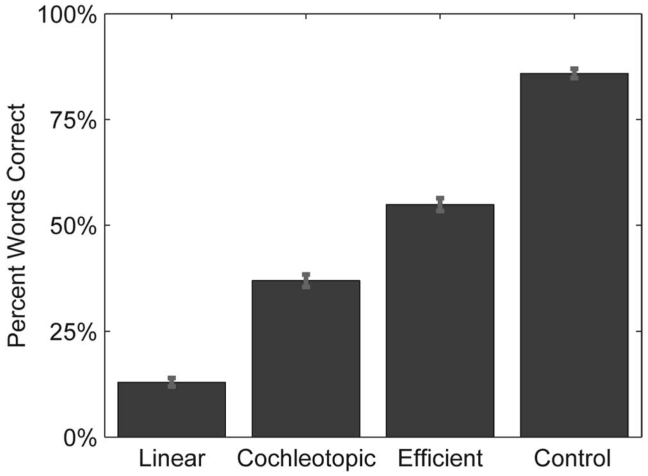

FIG. 3. Average speech intelligibility across participants as a function of condition. The control condition is unaltered speech. Error bars show 95% confidence interval of the mean.

highest-frequency filter. The spline smoothed set better-preserved these features. In both cases, however, the exact same set of raw filters, derived from learning the statistics of the TIMIT training set, were used. Only the choice of smoothing techniques changed.

Fourteen CMU undergraduate students, none of whom had participated in Experiment 1, participated in Experiment 2. They were instructed that they would hear a speech sound repeated twice and they were to type what they heard. For simplicity, the linear condition was removed to focus on the comparison of interest, cochleotopic versus efficient. All other methodological details were identical to Experiment 1.

V. RESULTS

Experiment 1. Each participant's performance was evaluated by comparing every word transcribed against the set of words in the original sentence. The typed responses were hand coded as "correct" if a match could be found. Minor alterations, such as adding -s or -ed, were not scored as correct but homophones were (e.g., sea versus see). Compound words (e.g., houseboat, bittersweet, sleepwalk, etc.) were treated as multiple words. Data from three participants were coded independently by two coders; the two sets of scores were highly correlated ($r>0.99$).

Each word in the original sentence was treated as independent (see below for further discussion of this issue) and the overall probability of a correct response was computed for each filterbank condition. As shown in Fig. 3, although intelligibility was greatly degraded for vocoded speech relative to original speech, intelligibility of vocoded speech was highly influenced by filterbank choice. The mean percent correct across participants for each condition were 13±4%, 37±6%, 56±6%, and 86±4% (mean±95% CI) for the linear, cochleotopic, efficient, and control conditions, respectively (planned Bonferroni-corrected pairwise comparisons for all results were highly significant, $p<0.001$). 15 of the 16 participants were significantly more accurate at transcribing speech synthesized with efficient versus cochleotopic filters ($p<0.01$). Across participants, performance differed by an average of 19% (efficient and cochleotopic filters, 56% versus 37%, respectively; $p<0.0001$).

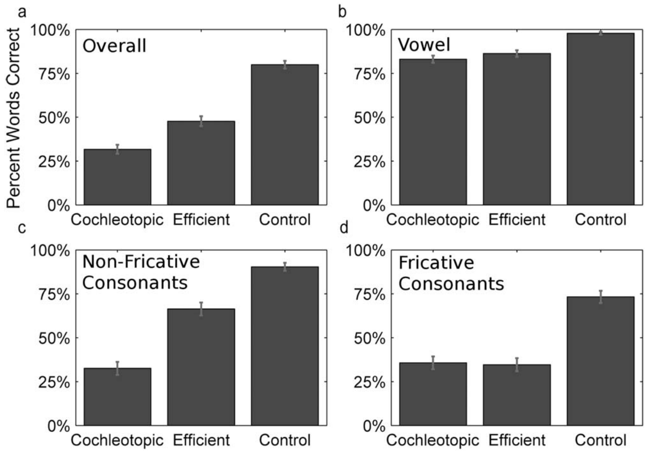

FIG. 4. Speech intelligibility for non-word stimuli. Each subfigure shows the accuracy of non-word speech identification across three conditions: cochleotopic filtering, efficient filtering, or control speech. The subplots show mean performance for (a) whole items, (b) vowels, (c) non-fricative consonants, and (d) fricatives. Error bars show the 95% confidence interval of the mean.

Participants' performance was little influenced by various lexical variables. A small correlation was found between word frequency (averaged from Kucera and Francis, 1967; Brown, 1984) and participant accuracy ($r=0.23$). There was no significant relationship of word type (noun/verb/other) and intelligibility ($p=0.33$). There was a small, but significant, increase in accuracy for words near the end of a sentence versus those occurring near the beginning or middle (5.4%, $p<0.0001$). There was no significant effect of either speaker or participant gender ($p=0.13$ and 0.33, respectively). Comparing performance between the first and second half of the experiment shows a significant increase in percent correct for both the cochleotopic (+4.2%; p-value$=0.007$) and efficient conditions (+4.9% increase; p-value<0.0001). There was, however, no significant interaction between early-late training and filter condition (p-value$=0.3$).

Experiment 2. For the non-word task, performance was measured by hand coding the response to each vowel and consonant in an item as correct or "incorrect." The entire item was coded correct if all of its phonetic elements were correct. As shown in Fig. 4(a), overall accuracy was slightly lower than Experiment 1: 31±8%, 49±9%, and 80±7%, for the cochleotopic, efficient, and control conditions, respectively (mean±95% CI). Performance with vocoded speech remained best with efficient filters (18% greater than cochleotopic filters; $p<0.0001$).

Performance gains differed based on the acoustic properties of the speech sounds. Participants were very accurate at identifying vowels [Fig. 4(b)], nearing ceiling in the control condition (98% correct) and achieving 83% and 86% correct in the efficient and cochleotopic conditions, respectively. The small difference between vocoded-speech conditions was not reliable ($p=0.10$). Relative to vowels, accuracy was much lower for consonants across all conditions (34% for cochleotopic, 51% for efficient, and 82% for control).

Performance in the efficient condition improved significantly between the first and second half of the experiment (+15%; $p<0.0001$) and a more modest improvement in performance across experiment halves was observed for the cochleotopic condition (+5.5% increase; $p=0.03$). There was no reliable difference in this learning effect across the different filter types ($p=0.1$).

Results of the theoretical modeling of Smith and Lewicki (2006) suggest that noise-like, ambient natural sounds represent a dimension in natural sound statistics distinct from acoustic transients. Based on this distinction, we separated the consonant stimuli into two classes, fricative and non-fricative consonants. Reanalyzing the data based on these classes produced very different results. As shown in Fig. 4(c), performance on non-fricatives (e.g., stop consonants, nasals and glides, such as /b/, /n/, and /l/) differed markedly, 33% and 66% ($p<0.0001$) in the cochleotopic and efficient conditions, respectively. In contrast, performance with fricatives [Fig. 4(d)] was quite low in all conditions (35%, 36%, and 73% for cochleotopic, efficient, and control) and it did not differ significantly between the vocoded-speech conditions ($p=0.4$).

There was a small, but unreliable, trend for better overall performance with the VCV-stimuli compared to CV (53% versus 50%; $p=0.10$). As with the sentences, there was no significant effect of either speaker ($p=0.08$) or participant gender ($p=0.19$).

VI. DISCUSSION

The results are consistent with a marked perceptual benefit of efficient coding. Speech recognition accuracy was greatest for noise-vocoded speech created with efficient filters. Given that these filters were created such that they best characterized the regularities of speech within the limits of six channels, it appears that the acoustic dimensions conveyed by the efficient filters provided listeners with more information with which to identify words and phonemes. The effect was dramatic. In the linear condition of the continuous speech task, participants typically understood only one word per sentence, consistent with previous findings that linear frequency mapping degrades perceptual performance (Fu and Shannon, 1999). On average, participants understood nearly twice as many words synthesized with the cochleotopic filters. In accordance with the efficient coding hypothesis, though, the efficient representation further increased performance, with participants identifying more than half the words in each sentence, four times more than the linear condition.

Even with a completely different stimulus set and nonsense syllables in Experiment 2, the efficient filters produced greater accuracy than cochleotopic filters for the non-word stimuli. The non-word task also revealed that the benefit of the efficient filters stemmed largely from benefits in non-fricative consonant intelligibility. Nearly the entire increase in performance between the filter conditions came from those items. However, it should be noted that it is possible that there is a ceiling effect confounding any effect on vowel recognition.

The pronounced increase in speech intelligibility in both experiments strongly suggests that participants are sensitive to the dimensions of speech acoustics predicted by the efficient coding hypothesis, but it is not yet clear what drives these improvements. One simple possibility is that the frequency tiling learned by the efficient filters maximizes its channel capacity (i.e., each channel carries an equal amount of information). To test this, we computed the variance of the envelope output from each channel for each filterbank across all stimuli used in the continuous speech task. Ideally, assuming independent, equal capacity channels, the variance across all six channels should be equal, implying that each channel is carrying equal amounts of independent information. If the variance in any channel is low relative to the others, then the total capacity of the system is underutilized.

The efficient filterbanks (using either gammatone- or spline-smoothing) make fuller use of their channel capacity than do the standard filterbanks; the higher-frequency filters in the linear and cochleotopic filterbanks are relatively unused, forcing all of the information about the sound to be carried by only two to four channels. The pressure to fully utilize channel capacity explains why the frequency tiling of the efficient filterbanks (as shown in Fig. 2) appears biased to the low frequencies; these filters more equitably carry information about speech.

Whereas this analysis makes use of signal statistics to differentiate the information carried across filters in the filterbank, it is also possible to consider what drives listeners' sensitivity to the dimensions of speech acoustics predicted by the efficient coding hypothesis from a psychoacoustic perspective. The articulation index (AI) has long been used to evaluate the importance of different frequency bands for speech recognition using perceptual measures (Fletcher and Steinberg, 1929; French and Steinberg, 1947; Studebaker *et al.*, 1987; ANSI, 1997). It is possible that perceptual weighting across frequency of importance for speech is similar to the information-theoretic optimum. Figure 5(a) illustrates the band importance values for normal speech calculated using the AI (ANSI, 1997; 1/3 octave). For each filter in each filterbank, we computed the filter response at each frequency band. The importance weighting for each filter can be estimated as the dot product of the filters' frequency response (power spectrum) and the band importance values shown in Fig. 5(a). This provides a score for each filter that indicates its importance to intelligibility. As is clear from Fig. 5(b), the efficient filterbank more evenly distributes band importance across its constituent filters (see Table I), with a significantly higher mean and lower variance compared to the cochleotopic and linear filterbanks. Thus, like the analysis based on signal statistics, analysis based on speech psychophysics (via the AI) also indicates that the efficient coding filters make fuller use of the channel capacity.

The results of the non-word task suggest that intelligibility of non-fricative consonants, in particular, drives the increase in intelligibility. In the case of stop consonants, which make up 47% of the non-fricative consonants, the distinguishing characteristics are not purely functions of frequency resolution but reflect higher-order, temporal structure of sound (e.g., voice onset time, Lisker and Abramson, 1964;

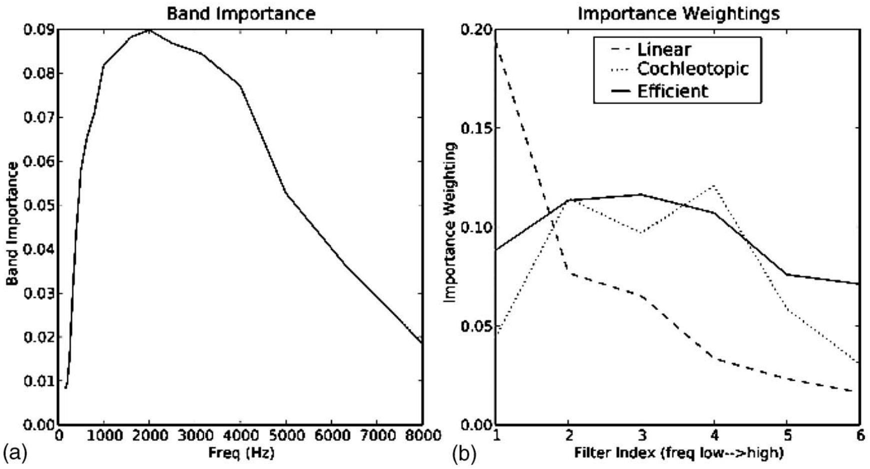

FIG. 5. (a) Band importance values for normal speech calculated using the AI (ANSI, 1997; 1/3 octave) indicates the contribution of different frequency bands to speech perception. The lowest and highest-frequency bands contribute much less to speech perception than the range from 1 to 3 kHz. (b) Importance weightings for each filter are computed as the dot product of the band importance values in (a) with the normalized filter's response at each frequency band and indicates how the filter pools information from each band.

Steinschneider *et al.*, 1999). The temporal asymmetry of the efficient filters may play a significant role here. This would agree with the finding that significant linguistic information is available in the temporal as well as spectral contents of speech (Shannon *et al.*, 1995). Increased sensitivity to temporal features like onsets suggests an influence of higher-order sound structure on the dimensions of perceptual sensitivity; second-order characteristics (i.e., the power spectrum) are not particularly sensitive to transient, edge-like signal structure (Field, 1987).

Our decision to treat each word in the continuous speech task as an independent measure greatly simplified analysis, but it is not realistic. It is known that syntactic and semantic contexts provided cues for sentence-level processing (Boothroyd and Nittrouer, 1988; Bronkhorst *et al.*, 1993; Gibson, 1998). Although the sentences in the TIMIT corpus have little predictability, participants clearly exhibited evidence of sentence-level processing. Correct responses were more likely to occur in pairs than would be expected at random and they were more likely to occur near the end of a sentence. In the non-word task, though, participants were required only to produce a single syllable. It is unlikely that cognitive load or context effects influenced the results, although there may be contrast effects (Lotto and Aravamudhan, 2004). Nonetheless, listeners' performance in the linear and cochleotopic conditions was significantly lower than that observed by Shannon *et al.* (1995); this is possibly a result of our choice of test materials (Zeng *et al.*, 2005). Syllables processed with an efficient filterbank were significantly better recognized than those processed with a cochleotopic filterbank.

In Experiment 1, we chose to smooth the learned efficient filters by fitting them with gammatones, allowing us to preserve the basic form of the learned filters using a model of auditory filters common to the auditory modeling literature (Patterson *et al.*, 1988; Slaney, 1993; Lyon, 1996). In Experiment 2, we aimed to address some limitations of the gammatone fitting by switching to spline-smoothing. For example, with the highest-frequency efficient filter, the best-fit gammatone truncated the highest frequencies whereas the best-fit spline did not. The smoothing in each experiment was performed on the same set of the raw filters produced by the computational model. Thus, the filters in both conditions reflect the statistics of the training set. The only difference between them was the smoothing technique. Spline smoothing, having many more free parameters, preserved more of the true spectral shape of the optimized filters. It should be noted that smoothing therefore introduced some differences between the experiments; specifically, whether (Experiment 1) or not (Experiment 2) the highest-frequency (5–7 kHz) information was incorporated into the sixth channel of the vocoder. The consistent patterning of results across the two experiments suggests that this difference did not have a significant impact on the results or their interpretation.

TABLE I. The mean and variance of the importance weightings across the filterbanks computed using the articulation index (AI). Higher means and smaller variances indicate greater and stronger correspondence between a given filterbank and the AI.

Filter type	Mean (CI)	Variance (CI)
Linear	0.0682 (0.0138–0.1225)	0.0072 (0.0019–0.0190)
Cochleotopic	0.0776 (0.0462–0.1089)	0.0024 (0.0007–0.0063)
Efficient	0.0953 (0.0792–0.1114)	0.0006 (0.0001–0.0016)

A possible criticism of this research is the use of classic linguistic categories (vowel, fricatives, etc.) that presuppose a particular structure to speech. Phonemic categories were used for stimulus categories as a rough approximation of the natural structure of speech acoustics. Their use here should not be taken as an assumption that phonemes represent a fundamental of acoustic or cognitive representation. Rather, their role here is only as a loose stand-in for dimensions of perceptual variability. Given the efficient coding hypothesis, ultimately it may be preferable to identify these dimensions using efficient coding algorithms similar to those used to train the filters here.

VII. CONCLUSION

As a unique compliment to the growing body of empirical and theoretical literature on efficient coding in neural systems (Barlow, 1961, Olshausen and Field, 2004), these results provide direct behavioral evidence for the role of coding efficiency as a general principle in human auditory perception. Yet to be addressed is the relevance of coding efficiency to higher-level representation. Methodologies that further meld theoretical-experimental designs to test listeners' sensitivity to the statistics of complex everyday sounds will be important for future exploration of efficiency in auditory processing. For example, by adapting our generative model to the acoustics of different spoken languages, we can generate acoustic stimuli directly from the model that reflect the differing low-level statistics of sounds from different languages absent any high-level, linguistic content.

Experiments with normal-hearing participants and vocoder speech previously have been useful in modeling cochlear implant hearing (Shannon *et al.*, 2003). It is possible that consideration of the computational principles of efficient coding may provide insight in cochlear implant applications.

Cochlear implants are by far the most successful neuroprosthetic devices and the only one in standard clinical use. They employ direct, electrical stimulation of auditory nerve fibers along the tonotopic axis of the cochlea to restore some degree of hearing in individuals with peripheral hearing loss, even in cases of profound deafness (Wilson *et al.*, 1991; Zeng *et al.*, 2004a). Unfortunately, despite 20 years of research and wide clinical application, speech perception in cochlear implant users remains highly variable and often quite degraded (Shannon *et al.*, 2003).

In general, the present results emphasize the significance of perceptual theory in neuroprosthetic design. Mimicking the surface features of a perceptual system, as in the cochleotopic filtering scheme, may not provide as much leverage as understanding a perceptual system's computational principles. The efficient coding hypothesis claims a specific computational principle: optimally efficient codes which carry the most information at the lowest cost should match the statistics of the signals they represent. Here, we found that the set of filters derived from a computational model trained to optimally extract the statistics of a corpus of speech passed more information normal-hearing participants could use to identify speech in sentence and non-word contexts than did more standard filtering schemes (linear, cochleotopic).

Of course, there remain many open questions for this line of research, and the specific algorithm used here may not necessarily produce the same dramatic improvements in speech intelligibility outside that laboratory. For example, the algorithm used to learn the efficient filters has not taken issues of electrode placement into account, which are essential in optimizing cochlear implant performance. Perceptual performance is known to degrade sharply as the mismatch between the frequency of the input channel and tonotopy of the cochlea increases (Shannon *et al.*, 1998; Fu and Shannon, 2002; Baskent and Shannon, 2005). Although it is beyond the scope of the current work, exploring issues regarding adaptation by cochlear implant users to changes in place-frequency mapping (Rosen *et al.*, 1999; Fu *et al.*, 2002b) would be an important extension of this research. Alternatively, expanding the efficient coding algorithm to incorporate constraints relevant to cochlear implants, such as frequency-place mappings, might be even more valuable.

We have shown in this study that recognition performance for perceptually degraded, vocoder speech improves when the vocoder filterbank matches the statistical structure of speech acoustics. A machine learning algorithm based on the efficient coding hypothesis was used to adapt the filterbank to speech structure. In two experiments, using stimuli from two unrelated speech corpora, recognition accuracy was superior for speech generated by the adapted filterbanks than recognition using cochleotopic filterbanks. The adapted filterbanks show greater spectral resolution in the frequency range of speech formants, which plays a large role in the higher recognition accuracy.

ACKNOWLEDGMENTS

V.L.M. and L.L.H. designed the experiments, planned the analyses and wrote the paper. V.L.M. implemented the filters, produced the stimuli, and ran the experiments and analyses. He was supported by NSF IGERT training Grant No. DGE-9987588. This research was supported by Grant No. 2R01DC004674-04A2 from the National Institutes of Health and Grant No. 0345773 from the National Science Foundation. We would like to thank Christi Gomez for help collecting and coding data.

ANSI (**1997**). "Methods for calculation of the speech intelligibility index," American National Standards Institute, New York.

Atick, J. J. (**1992**). "Could information-theory provide an ecological theory of sensory processing?" Networks **3**, 213–251.

Attias, H., and Schreiner, C. E. (**1998**). "Coding of naturalistic stimuli by auditory midbrain neurons," in *Advances in Neural Information Processing Systems*, edited by M. I. Jordan, M. J. Kearns, and S. A. Solla (MIT, Cambridge, MA), Vol. **10**.

Barlow, H. B. (**1961**). "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, edited by W. A. Rosenbluth (MIT, Cambridge, MA), pp. 217–234.

Baskent, D. E., and Shannon, R. V. (**2005**). "Interactions between cochlear implant electrode insertion depth and frequency-place mapping," J. Acoust. Soc. Am. **117**, 1405–1416.

Bekesy, G. (**1960**). *Experiments in Hearing* (McGraw-Hill, New York), pp. 503–509.

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.

Bronkhorst, A. W., Bosman, A. J., and Smoorenburg, G. F. (**1993**). "A model for context effects in speech recognition," J. Acoust. Soc. Am. **93**, 499–509.

Brown, G. D. A. (**1984**). "A frequency count of 190,000 words in the London Lund corpus of English conversation," Behav. Res. Methods Instrum. **16**, 502–532.

Escabi, M. A., Miller, L. M., Read, H. L., and Schreiner, C. (**2003**). "Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus," J. Neurosci. **23**, 11489–11504.

Field, D. (**1987**). "Relations between the statistics of natural images and the response profiles of cortical cells," J. Opt. Soc. Am. A **4**, 2379–2394.

Fletcher, H., and Steinberg, J. C. (**1929**). "Articulation testing methods," Bell Syst. Tech. J. **8**, 806–854.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Fu, Q.-J., and Shannon, R. V. (**1999**). "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," J. Acoust. Soc. Am. **105**, 1889–1900.

Fu, Q.-J., and Shannon, R. V. (**2002**). "Frequency mapping in cochlear implants," Ear Hear. **23**, 339–348.

Fu, Q.-J., Shannon, R. V., and Galvin, J. (**2002**). "Perceptual learning following changes in the frequency-to-electrode assignment with the Nucleus-22 cochlear implant," J. Acoust. Soc. Am. **112**, 1664–1674.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (**1990**). TIMIT Acoustic-Phonetic Continuous Speech Corpus.

Gibson, E. (**1998**). "Linguistic complexity: Locality of syntactic dependencies," Cognition **68**, 1–76.

Greenwood, D. (**1961**). "Critical bandwidth and the frequency coordinates of the basilar membrane," J. Acoust. Soc. Am. **33**, 1344–1356.

Hillenbrand, M. J., and Gayvert, T. R. (**2005**). "Open source software for experiment design and control," J. Speech Lang. Hear. Res. **48**, 45–60.

Klein, D. J., Konig, P., and Kording, K. P. (**2003**). "Sparse spectrotemporal coding of sounds," EURASIP J. Appl. Signal Process. **2003**(7), 659–667.

Kucera, H., and Francis, W. N. (**1967**). *Computational Analysis of Present-Day American English* (Brown University Press, Providence, RI).

Laughlin, S. B., and Sejnowski, T. J. (**2003**). "Communication in neuronal networks," Science **301**, 1870–1874.

Lewicki, M. S. (**2002**). "Efficient coding of natural sounds," Nat. Neurosci. **5**, 356–363.

Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422.

Lotto, A. J., and Aravamudhan, R. (**2004**). "Phonetic context effects in cochlear implant listeners," in Meeting of the Acoustical Society of America, San Diego, CA.

Lyon, R. F. (**1996**). "The all-pole gammatone filter and auditory models," Forum Acusticum, Antwerp, Belgium.

MacKay, D. J. C. (**2003**). *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge).

Mallat, S. G., and Zhang, Z. (**1993**). "Matching pursuits with time-frequency dictionaries," IEEE Trans. Signal Process. **41**, 3397–3415.

Nie, K. B., and Zeng, F. G. (**2004**). "Speech perception with temporal envelope cues: Study with an artificial neural network and principal component analysis," The 26th IEEE EMBS Conference, San Francisco.

Olshausen, B. A., and Field, D. (**1996**). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature (London) **381**, 607–609.

Olshausen, B. A., and Field, D. (**2004**). "Sparse coding of sensory inputs," Curr. Opin. Neurobiol. **14**, 481–487.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). "Implementing a gammatone filter bank. SVOS final report: The auditory filter bank," Report No. 2341, Medical Research Council Applied Psychology Unit, University of Cambridge Medical School.

Rieke, F., Bodnar, D. A., and Bialek, W. (**1995**). "Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory neurons," Proc. R. Soc. London, Ser. B **262**, 259–265.

Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **106**, 3629–3636.

Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423, 623–656.

Shannon, R. V., Fu, Q.-J., Galvin, J., and Friessen, L. (**2003**). "Speech perception with cochlear implants," in *Cochlear Implants: Auditory Prostheses and Electric Hearing*, Springer Handbook of Auditory Research, edited by F.-G. Zeng, A. N. Popper, and R. R. Fay (Springer, New York), pp. 334–376.

Shannon, R. V., Jensvold, A., Padilla, M., Robert, M., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Shannon, R. V., Zeng, F.-G., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**, 2467–2476.

Sharpee, T., Sugihara, H., Kurgansky, A., Rebrik, S., Stryker, M. P., and Miller, K. D. (**2006**). "Adaptive filtering enhances information transmission in visual cortex," Nature (London) **439**, 936–942.

Simoncelli, E., and Olshausen, B. (**2001**). "Natural image statistics and neural representation," Annu. Rev. Neurosci. **24**, 1193–1216.

Slaney, M. (**1993**). "An efficient implementation of the Patterson–Holdsworth auditory filter bank," Technical Report No. 35, Apple Computer, Cupertino, CA.

Smith, E. C., and Lewicki, M. S. (**2005a**). "Efficient coding of time-relative structure using spikes," Neural Comput. **17**, 19–45.

Smith, E. C., and Lewicki, M. S. (**2005b**). "Learning efficient auditory codes using spikes predicts cochlear filters," in *Advances in Neural Information Processing Systems*, edited by L. K. Saul, Y. Weiss, and L. Bottou (MIT, Cambridge, MA), Vol. **17**, pp. 1289–1296.

Smith, E. C., and Lewicki, M. S. (**2006**). "Efficient auditory coding," Nature (London) **439**, 978–982.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature (London) **416**, 87–90.

Steinschneider, M., Volkov, I. O., Noh, M. D., Garell, P. C., and Howard, M. A. (**1999**). "Temporal Encoding of the Voice Onset Time Phonetic Parameter by Field Potentials Recorded Directly From Human Auditory Cortex," J. Neurophysiol. **82**, 2346–2357.

Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (**1987**). "A frequency importance function for continuous discourse," J. Acoust. Soc. Am. **81**, 1130–1138.

Vinje, W. E., and Gallant, J. L. (**2002**). "Natural stimulation of the nonclassical receptive field increases information transmission efficiency in V1," J. Neurosci. **22**, 2904–2915.

Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (**1991**). "Better speech recognition with cochlear implants," Nature (London) **352**, 236–238.

Zeng, F.-G., Nie, K., Liu, S., Stickney, G., DelRio, E., Kong, Y.-Y., and Chen, H. (**2004a**). "On the dichotomy in auditory perception between temporal envelope and fine structure cues," J. Acoust. Soc. Am. **116**, 1351–1354.

Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations," Proc. Natl. Acad. Sci. U.S.A. **102**, 2293–2298.

Zeng, F.-G., Popper, A. N., and Fay, R. R. (**2004b**). *Cochlear Implants: Auditory Prostheses and Electric Hearing, Springer Handbook of Auditory Research* (Springer, New York).