PROCEEDINGS A

rspa.royalsocietypublishing.org

Research

Cite this article: Lee T, Theunissen F. 2015 A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features. *Proc. R. Soc. A* **471**: 20150309.
http://dx.doi.org/10.1098/rspa.2015.0309

Received: 8 May 2015
Accepted: 13 November 2015

Subject Areas:
acoustics, computational biology

Keywords:
single microphone noise reduction, speech enhancement, spectro-temporal features, signal processing

Author for correspondence:
Frédéric Theunissen
e-mail: theunissen@berkeley.edu

A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features

Tyler Lee[1,2] and Frédéric Theunissen[1,3]

[1]Helen Wills Neuroscience Institute, [2]Redwood Center for Theoretical Neuroscience, and [3]Department of Psychology, University of California, Berkeley, CA 94720, USA

Animals throughout the animal kingdom excel at extracting individual sounds from competing background sounds, yet current state-of-the-art signal processing algorithms struggle to process speech in the presence of even modest background noise. Recent psychophysical experiments in humans and electrophysiological recordings in animal models suggest that the brain is adapted to process sounds within the restricted domain of spectro-temporal modulations found in natural sounds. Here, we describe a novel single microphone noise reduction algorithm called spectro-temporal detection–reconstruction (STDR) that relies on an artificial neural network trained to detect, extract and reconstruct the spectro-temporal features found in speech. STDR can significantly reduce the level of the background noise while preserving the foreground speech quality and improving estimates of speech intelligibility. In addition, by leveraging the strong temporal correlations present in speech, the STDR algorithm can also operate on predictions of upcoming speech features, retaining similar performance levels while minimizing inherent throughput delays. STDR performs better than a competing state-of-the-art algorithm for a wide range of signal-to-noise ratios and has the potential for real-time applications such as hearing aids and automatic speech recognition.

1. Introduction

Humans, as social beings, rely heavily on spoken language for communication. The fluctuations in air

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspa.2015.0309 or via http://rspa.royalsocietypublishing.org.

THE ROYAL SOCIETY
PUBLISHING

pressure through which speech is transmitted, however, are regularly corrupted by a variety of sounds from other sources, including the bustling noises of a crowded street, the ambient whoosh of wind in an open field or the speech babbles of other individuals at a social gathering. Human brains, and, indeed, the brains of many other species [1], are adept at extracting an individual sound source from these complex mixtures. How the brain performs this task remains poorly understood, yet a solution to this problem is critical to many important applications. Individuals with hearing aids struggle to understand speech in crowded spaces [2]; the optimal amplification and processing in quiet environments are often detrimental to the listening experience in noisy environments [3]. Similarly, artificial speech recognition systems, such as those used in smartphones, often fail in relatively low levels of background noise [4]. These difficulties have led to great interest in the field of noise reduction from auditory scientists and engineers. Although spatial cues can be used to facilitate the separation of speech in noise [5], we will be focusing on algorithms that record sound from a single location: single microphone noise reduction (SMNR) algorithms.

Recent work in auditory neurophysiology has shed light on how the brain parses sounds in noise. To parse the auditory scene, the brain must analyse incoming sounds in a feature space that reliably separates the particular sound of interest from the current background noise. One way that this is performed is by preferentially encoding behaviourally relevant sounds. This class of sounds, often broadly declared ‘natural sounds’, lies in a particular subspace of all possible sounds [6]. Indeed, there is a good deal of evidence showing that natural sounds activate neurons most strongly, especially in higher regions of the auditory system (reviewed in [7]). In an attempt to understand the relevant feature space for these higher-level neurons, many researchers have looked to reverse correlation and other methods to build encoding models capable of predicting a neuron’s response from an incoming sound [8–11]. Studies using these models have shown that the spectro-temporal modulations can account for large fractions of the sound-induced responses of neurons in many regions of the auditory system [9] (reviewed in [7]). This body of work has demonstrated that the set of spectro-temporal modulations that neurons detect is also not uniformly distributed throughout the entire space but instead lies in a subspace that lends an efficient encoding of behaviourally relevant sounds [12–14].

Extrapolating these results to the problem of analysing sound in noise leads to the postulate that when the brain is presented with a behaviourally relevant sound (e.g. a communication signal) in background noise, the preferential encoding of the behaviourally relevant sound leads to an underrepresentation of noise: a noise reduction. There is some evidence to believe this is the case. For example, a study by Moore *et al*. [15] showed that neurons sensitive to fast spectral modulations and slow temporal modulations responded to bird song presented in noise with greater levels of noise robustness. Other work builds on this preferential encoding hypothesis, but prescribes more important roles for nonlinear processing (e.g. neural adaptation) and attentional feedback [16–20].

Parallel work studying the relevant feature space to predict speech intelligibility has shown the importance of both temporal and spectro-temporal modulations. Degradation of the slow temporal modulations present in speech is known to correlate with a loss in speech intelligibility [21,22]. Other studies indicate that the signal-to-noise ratio (SNR) in the spectro-temporal modulation domain correlates strongly with the intelligibility of speech in a wide range of situations [23]. More specifically, the low-pass region of spectro-temporal modulations below 7.75 Hz (temporal) and 3.75 cycles per kHz (spectral) seems particularly important for speech intelligibility [24]. While some research has called into question the role of cross-frequency integration, or the ‘spectro’ of spectro-temporal modulations, it seems clear that the modulation space is a good candidate for the analysis of noisy and corrupted speech [25].

In addition, neural sensory systems are affected by top-down processes either in the form of attentive mechanisms or expectations. For example, neural processing of speech in auditory cortical areas has been shown to be selective for the attended speech stream [17]. Expecting linguistic information also changes the properties of neural responses to degraded speech in lower cortical areas [26,27]. Both attention and expectation rely on high-order statistical structure in the speech stream that can be used to make predictions about future sounds and in this manner facilitate the computations needed for detection and interpretation.

Here, we introduce an algorithm that performs SMNR, extracting speech from background noise by simultaneously learning a spectro-temporal feature space in which to project noisy speech, applying a static nonlinearity, and then decoding jointly time-frequency gains that modify the noisy speech to produce a clean speech estimate. This algorithm outperforms a standard optimal noise reduction scheme, the Ephraim–Malah (EM) algorithm [28] with a minimum statistics noise estimator [29,30], across multiple measures of sound quality and intelligibility. Furthermore, we explore the role that predicting upcoming spectro-temporal features can play in producing a system with strong noise reduction and minimal throughput delay.

2. Methods

(a) Stimuli

We trained our algorithm on clean speech recordings of the hearing-in-noise test (HINT) sentence corpus embedded in multiple noise types [31,32]. All stimuli were single channel, sampled at 16 kHz and band-limited between 25 Hz and 7.5 kHz, with durations averaging 1.9 s, ranging from 0.8 to 7.3 s. The algorithm was trained in multiple noise conditions. We first describe results on training sets with 100 stimuli from a single noise type: speech-shaped noise and babble noise. We then describe results on a training set with 280 stimuli from seven different noise types: speech-shaped noise, babble noise and all five noise types from the QUT database [33]. Testing was done either on held-out stimuli from the same noise types used in training, or on a separate dataset using 12 noise types: 10 gathered from www.freesound.org, along with white noise and pink noise. Training was done using sentences from either one speaker or 16 speakers at 0 dB SNR. A detailed description of each stimulus set is provided in the electronic supplementary material, Methods.

(b) Spectro-temporal detection–reconstruction noise reduction algorithm

The goal of any noise reduction scheme is to take a noisy signal, $x(t) = s(t) + n(t)$ (e.g. an individual speaker in a crowded room) and isolate, as well as possible, the sound components corresponding to the clean signal, $s(t)$ (e.g. the individual speaker) from the noise, $n(t)$. This is commonly done by first applying a collection of bandpass filters to the noisy signal to produce a set of narrowband channels, $y(f, t)$. Then, each narrowband signal is scaled by an estimated gain factor, $\hat{g}(f, t)$ that is proportional to the SNR of the channel. Finally, these scaled signals are summed to produce an estimate of the clean signal, $\hat{s}(t)$,

$$\hat{s}(t) \sum_{f} y(f, t)\hat{g}(f, t).$$

This scheme is often called an analysis–synthesis design and has been used successfully for decades in many SMNR algorithms [34–36]. Where these algorithms differ is in the method of estimating signal-to-noise and the functional form of the gains. Here, we use an artificial neural network that attempts to analyse the spectro-temporal modulations present in the noisy signal (detection stage) to estimate the optimal time-varying gains (reconstruction stage; figure 1). Both detection and reconstruction stages are inspired by auditory, and more generally sensory, computations performed by the brain. This novel network architecture provides explicit representation of the joint spectro-temporal structure present in both the noisy signal and the time-varying gains.

(i) Analysis and spectrogram computation

To compute the narrowband signals, $y(f, t)$, we created a filterbank with 223 bandpass Gaussian-shaped filters with centre frequencies linearly spaced between 25 Hz and 7.5 kHz and bandwidths

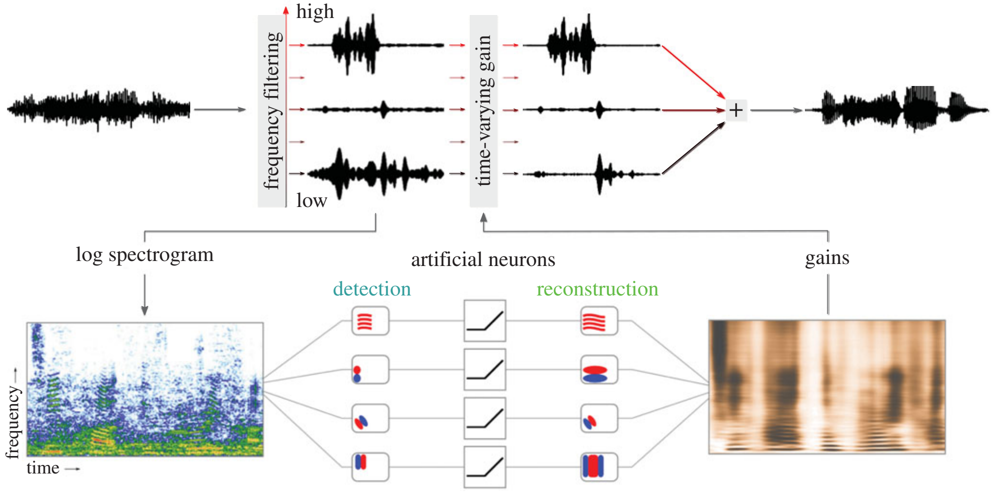

Figure 1. The spectro-temporal detection–reconstruction (STDR) algorithm is composed of two chains: the analysis–synthesis chain and the gain estimation chain (see Methods). Top row: a noisy signal waveform is first bandpass filtered into a set of narrowband channels. Each narrowband channel is then scaled by a time-varying gain, found in the gain estimation chain. The scaled channels are then summed to create an estimate of the original clean signal. Bottom row: the gains are produced using an artificial neural network. Each unit in the network is characterized by a spectro-temporal detection kernel (i.e. its receptive field) that determines its output given the spectrogram of a segment of noisy signal, and a gain reconstruction kernel that it uses to generate time-varying gains for estimating the denoised signal.

of 32 Hz each, corresponding to a time-domain Gaussian window with a 5 ms bandwidth. We computed the analytic signal from each narrowband signal and extracted the envelope. A Gaussian-shaped frequency filter with standard deviation of 32 Hz effectively limits the bandwidth of each channel's amplitude envelope below 192 Hz (6 × 32 Hz, because 6 standard deviations accounts for approx. 99.8% of the density of the window) [37]. Each channel's envelope was then extracted by computing the analytic signal and then downsampled to 1 kHz, producing a spectrogram $X_{\mathrm{lin}}(f,t)$. The spectrogram was then log transformed with a floor value set at −80 or −50 dB from the maximum power. Results were very similar for both floor values except for the babble noise where performance was slightly but consistently better at −50 dB. Finally, we subtracted the mean log spectrogram value for each frequency band before all later processing stages. This time-frequency analysis is qualitatively similar to the analysis performed by the cochlea, which is often modelled using a set of bandpass filters, followed by a half-wave rectification, low-pass filtering and adaptive gain control [38]. This complete preprocessing was applied to each individual sound before being fed into the network as $X(f,t)$.

(ii) Artificial neural network

The artificial neural network was structured as a three-layer autoencoder [39]. The input to the network was the processed spectrogram, $X(f,t)$. Each first layer unit operated on this time-frequency representation using a spectro-temporal filter

$$a_m(t) = \sum_{f=1}^{N} \sum_{\tau=0}^{L_{\mathrm{D}}-1} X(f,t-\tau)\phi_m(f,\tau),$$

where $a_m(t)$ is the response of input unit m, $\phi_m(f,\tau)$ is its spectro-temporal filter and L_{D} is its filter duration. The activation of each input unit was scaled to have unit standard deviation to help with optimization. This was done for each individual sentence, though the rescaling could instead be done on the next layer's input weight matrix, if desired. The number of units in the first layer was

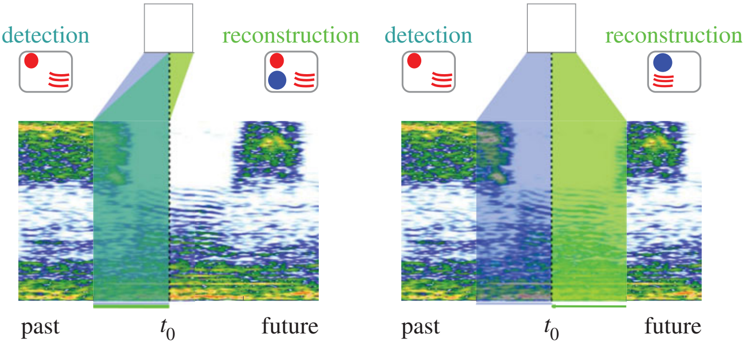

Figure 2. The reconstruction kernels can be used to apply gains completely in the past, overlapping the window used in detection (*causal* detection combined with *acausal* reconstruction), completely in a predictive mode where the detection window is in the past and the reconstruction kernel window is in the future (*causal* detection followed by *causal* reconstruction) or anywhere in between. This is done by shifting the delays of the reconstruction kernel window while maintaining a fixed window duration. For *acausal* reconstruction, the real-time algorithm would have a minimum delay given by the extent of the reconstruction window in the past.

chosen to be 100 and τ ranged from 0 to 99 ms, yielding a completely causal filter with 100 ms duration. The middle layer performed a weighted linear combination of the first layer responses followed by a pointwise threshold nonlinearity

$$r_i(t) = \max(\boldsymbol{w}_i \cdot \boldsymbol{a}(t) + \beta_i, 0).$$

Here, $r_i(t)$ is the response unit i, $\boldsymbol{w}_i$ is the ith row of the weight matrix W, $\boldsymbol{a}(t)$ is the vector of first layer unit responses, and β_i is the unit's threshold. The number of units in the middle layer was set to 80. The final layer performed a simple weighted linear combination of the middle layer's responses

$$o_n(t) = \boldsymbol{v}_n \cdot \boldsymbol{r}(t),$$

where, again, $o_n(t)$ is the unit response and $\boldsymbol{v}_n$ is the nth row of the weight matrix V. The time-varying gains were then reconstructed from the final layer activities by convolving with a spectro-temporal gain reconstruction kernel and summing across all units. Lastly, we applied a sigmoid function to the gain, bounding it between 0 and 1.

$$\hat{g}(f,t) = \sigma\left(\gamma_f + \sum_n \sum_{\tau=\tau_0}^{\tau_0+L_R-1} o_n(t-\tau)\psi_n(f,\tau)\right)$$

Here, σ is a sigmoid function (here the logistic function), γ_f is a bias term for frequency band f, $\psi_n(f,\tau)$ is the spectro-temporal gain reconstruction filter for unit n, and L_R is the duration of the reconstruction filters. Although it is not required, the parameters of the final layer were taken to be the same as those chosen for the first layer: the number of units was set to 100, and the duration of the filters was 100 ms. In contrast to the first layer, however, we explored several different ranges for τ, beginning with the completely *acausal* regime of −99 to 0 ms where the reconstructed gains are entirely in the past, and sliding the window to the completely *predictive* regime of 0–99 ms where the reconstructed gains are entirely in the future (figure 2). Also note that it is a common practice to set the minimum asymptotic value of the sigmoid to some small, non-zero value: $g' = g_{\min} + (1 - g_{\min}) \cdot \hat{g}$. While we found that this subjectively offers some benefits, we have kept this value at zero for sake of clarity. The number of units in each layer was chosen such that an increase provided no further qualitative benefit. The duration of the filters was varied symmetrically from 10 to 200 ms, and 100 ms was chosen as the value that provided the best overall performance for all noise types, though the differences were small.

To understand how the network processed the signals, we found it helpful to break down the computations into two functional phases: a detection phase, corresponding to the first and second layers, and a gain reconstruction phase, corresponding to the second and third layers (figure 1). In this view, each unit in the middle layer can be said to perform a spectro-temporal feature detection on the input using it 'spectro-temporal detection kernel', defined as

$$D_i(f,\tau) = \sum_m W_{i,m}\phi_m(f,\tau).$$

These filters are commonly called spectro-temporal receptive fields (STRFs) by auditory neurophysiologists and have been shown to effectively represent speech [40]. We will use this nomenclature here when appropriate. Similarly, each unit in the middle layer makes its own contribution to the estimated gains using its 'gain reconstruction kernel', defined as

$$R_i(f,\tau) = \sum_n V_{n,i}\psi_n(f,\tau).$$

For this reason, we call our algorithm the spectro-temporal detection–reconstruction (STDR) algorithm.

(iii) Optimization

The spectro-temporal filters of the first and third layers were learned using principal components analysis (PCA) on separate examples of clean speech and noise. PCA was performed on sections of spectrogram taken by sliding a 100 ms rectangular window with a stride of 50% of the window duration. We used a total of 100 principal components, 50 from clean speech and 50 from noise.

Optimizations were performed on a training set of 100 examples, for speech-shaped noise and babble noise, or 280 examples, for seven noise type training, of signal in background noise, each less than 5 s in duration and where ground truth signal and noise were known. All performance metrics, described in the next section, were computed on a held-out set of noisy stimuli not seen during training. The weight matrices, W and V, unit biases, β_i, and frequency band biases, γ_{f}, were all updated in order to minimize the mean squared error between the estimated gains, $\hat{g}$ and the optimal gains, $\tilde{g}$ computed as

$$E = \frac{1}{NT}\sum_{t=1}^{T}\sum_{f=1}^{N}(\tilde{g}(f,t) - \hat{g}(f,t))^2$$

and

$$\tilde{g}(f,t) = \frac{|S_{\mathrm{lin}}(f,t)|}{|X_{\mathrm{lin}}(f,t)|}.$$

Here, $\tilde{g}$ is the optimal time-frequency gain that maps the linear noisy spectrogram X_{lin} (i.e. pre-logarithm) to the linear clean spectrogram S_{lin}. T is the total number of time points. Parameters were updated using gradient descent and optimization ceased when the error had increased for five consecutive iterations on a held-out portion of 10% of the training data. All filter weights were initialized to small, uniform random values centred on zero. The range for the weights was chosen using the 'normalized initialization' heuristic from [41], which has been shown to alleviate discrepancies in learning between layers and to perform well in simulations with deep networks. The biases were all initialized to zero. Only one random initialization was done, as multiple randomizations produced qualitatively similar results.

(c) Performance metrics

We assessed the performance of our algorithm using objective measures of sound quality and speech intelligibility. Sound quality was quantified using three composite ratings as proposed by Hu & Loizou [42]. These three ratings predict the subjective evaluations of normal hearing listeners for the speech distortion, background noise intrusiveness and overall quality of a processed sound. These three ratings are obtained, in turn, from linear combinations of four other objective measures: the segmental SNR [43], the weighted spectral slope [44], the log likelihood ratio [45] and the perceptual estimate of sound quality [46]. The three ratings showed correlations of 0.73, 0.64 and 0.73 between objective and subjective quality judgements along each of the corresponding three dimensions (speech, background and overall). Code for the algorithms was downloaded from http://ecs.utdallas.edu/loizou/speech/software.htm.

To gauge speech intelligibility, we used the short-time objective intelligibility (STOI) rating, which measures the similarity between time-frequency representations of the clean speech and the processed noisy speech [47]. This measure was shown to significantly correlate with subjective reports of speech intelligibility, with a correlation coefficient of 0.92 for speech processed using SMNR techniques. Code for STOI was downloaded from http://www.ceestaal.nl/matlab.html.

To determine if the performance of our STDR algorithm was significantly better than either the unfiltered noisy signal or a comparison algorithm (the EM algorithm), we used a linear mixed-effects model. Both the comparison algorithm and the mixed-effects model are described in more detail in the electronic supplementary material, Methods.

(d) Normalized performance

Normalized performance values shown in figures 5 and 7 were computed as

$$NP_{\mathrm{alg}} = 100 \times \frac{P_{\mathrm{alg}} - P_{\mathrm{unfilt}}}{P_{\mathrm{opt}} - P_{\mathrm{unfilt}}}.$$

Here, P_{alg} is the performance of a particular algorithm, P_{unfilt} is the metric computed on the noisy speech signal and P_{opt} is the optimal performance using time-frequency gains, $\tilde{g}(f,t)$ Thus, the normalized performance of the unfiltered noisy speech is set to 0, the optimal performance is set to 100 and a specific algorithm's performance is the percentage of improvement over the unfiltered noisy speech on any particular metric that the algorithm could hope to achieve.

3. Results

As described in the Methods, we developed a novel algorithm for SMNR called STDR. STDR relies on the detection of spectro-temporal features that are useful for separating signal from noise and uses those detections to adjust time-varying gains on each frequency band in a predictive manner. In the Results section, we further describe how the algorithm works by examining the role of its components in specific speech-in-noise situations and compare its performance to the EM algorithm.

(a) Role of individual detection and gain reconstruction filters

As we will further describe below, our algorithm showed improvements on most of the metrics we tested across a wide range of input SNR when compared with both the unfiltered sound and the sound processed by the EM algorithm. STDR achieves this feat by detecting characteristic structure in both the signal and the noise and attempting to maintain high gains in signal-heavy regions of the time-frequency plane and to decrease the gains in noise-heavy regions. This 'push–pull' action manifests in learned detection and reconstruction gain filters that can clearly be interpreted as signal-selective and noise-suppressive units. Figure 3 shows four example units trained on speech from a single speaker embedded in babble noise. The first unit (figure 3*c*) functions primarily to suppress noise. The detection filter is strongly inhibited by the broadband onsets and more sustained energy in high frequencies that are characteristic of isolated speech. When the unit is not inhibited, it yields a broadband negative gain suppressing sound. The other three units select for specific speech features, with sparse activations that are non-zero only when their particular feature is present in the stimulus. The filter of the second unit (figure 3*d*) detects short bursts of high-frequency power often associated with fricatives in speech. The reconstruction gain is almost a perfect match to the detection filter boosting those specific sections

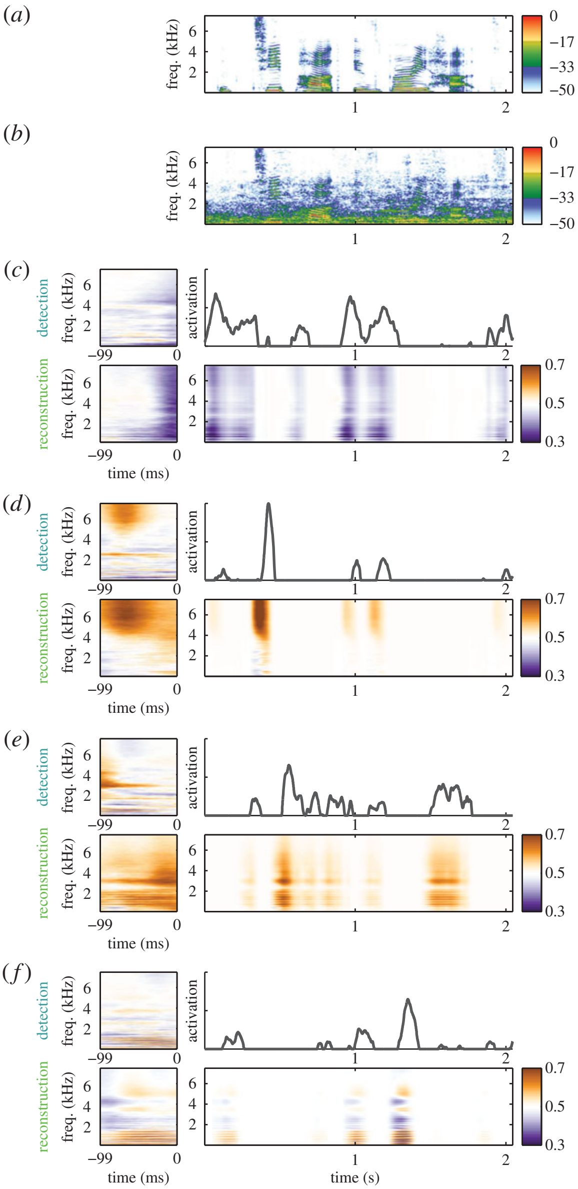

Figure 3. Individual units in the network detect and reconstruct different types of spectro-temporal features. (*a*) A spectrogram of the sentence 'The teapot was very hot'. (*b*) A spectrogram of the sentence from (*a*) added to babble noise at 0 dB SNR. (*c–f*) Example responses from four individual units showing, for each, its detection filter (top left), its thresholded activation in response to the spectrogram in (*b*) (top right), its gain reconstruction filter (bottom left), and the resulting reconstructed gains (bottom right). For the filters and reconstructed gains, blue represents a decrease in gain value, whereas orange represents an increase in gain value. (*c*) This unit predominately lowers the gain on noisy periods and is strongly inhibited by the broadband onsets of speech. (*d*) This unit detects power and reconstructs gains in the mid-range frequencies with additional selectivity for specific harmonic structure. (*e*) This unit detects sharp onsets in the high frequencies, a feature present only in the consonants of the foreground speech. (*f*) This unit shows strong selectivity for the specific harmonics of the trained speaker.

of the speech signal. The filter of the third unit (figure 3*e*) detects coarse power in the mid-frequencies with some harmonic structure. The filter of the fourth unit (figure 3*f*) is highly specific, responding selectively to the harmonic structure of the trained speaker's voice. The detection filters and reconstruction gain filters for all units are shown in the electronic supplementary material, figure S1. In general, there was a consistent dichotomy between noise-suppressive and signal-selective units with a continuum of filter types within each category.

(b) Performance on speech in speech-shaped noise

By using an array of individual units with different feature selectivity learned from a representative dataset, the algorithm is able to produce accurate reconstructions to novel noisy stimuli. We tested the performance of our algorithm using sentences from the HINT embedded in two types of background noise: speech-shaped noise and babble noise. The STDR algorithm was trained on a set of 100 sentences from either one or 16 individuals, chosen randomly from a large database of speakers, at 0 dB SNR (see electronic supplementary material, Methods). Figure 4 shows the performance of the STDR algorithm on a novel sentence from speaker 1 when trained on speech from one speaker (figure 4*c*,*e*) and 16 speakers (figure 4*d*,*f*). A few features stand out when looking at the time-frequency gains. First, it captures precisely the low-frequency harmonics corresponding to the speaker's pitch. This effect is much stronger when the algorithm was trained only on the speaker shown, but is still also present when trained on 16 speakers (figure 4*c*,*d*, inset). Second, the STDR algorithm reconstructs the slowly changing spectro-temporal contours of stimulus power in the formants. This is evident in the dark regions in the low-to-mid frequencies. Third, it precisely amplifies the high-frequency power found in many consonants. Because this level of high-frequency power is transient and only present in the speech itself, it represents a very specific cue for clean speech and is robustly detected. Lastly, the temporal structure, in general, of the voice is very reliably detected, demonstrated by the strong onsets and offsets in gains. Sound files for both the noisy speech and the denoised estimates can be found in the electronic supplementary material.

To quantify the performance, we processed 15 novel sentences from each trained speaker at seven different SNRs and then computed several objective measures of performance that have been used in the field (see Methods). Note that the algorithm was only trained at 0 dB SNR and that our performance quantification not only uses novel sentences, but also a range of SNR around 0. To assess the intelligibility of the processed speech, we computed the STOI measure [47]. Our STDR algorithm showed slight but significant improvements on this measure over the unfiltered noisy speech (0.03 ($p < 10^{-4}$), d.f. = 207, linear mixed-effects model, see electronic supplementary material, Method; figure 5*a*,*b*, left column). It also significantly outperformed a standard noise reduction algorithm that uses minimum statistics noise estimation and log minimum mean squared error optimal frequency filtering, the EM algorithm (0.05 ($p < 10^{-4}$), see electronic supplementary material, Methods) [28]. These benefits were seen on a large majority of individual sentences (figure 5*b*, left column).

To assess the resulting quality of the processed stimulus, we computed a set of three composite measures [42]. These measures combine multiple pre-existing objective measures to best estimate the subjective sound quality ratings of human listeners along three axes, namely speech quality, background noise intrusiveness and overall quality (see also Methods). The STDR algorithm performed well, significantly improving each rating over the unfiltered stimulus (estimated improvement of 0.54 ($p < 10^{-4}$), 0.30 ($p < 10^{-4}$) and 0.44 ($p < 10^{-4}$) for signal, background and overall, respectively). It also provided significant improvements over the EM algorithm on all three measures (0.32 ($p = 6 \times 10^{-4}$), 0.07 ($p < 10^{-4}$), 0.15 ($p < 10^{-4}$); figure 5*a*,*b*, centre and right columns show background noise intrusiveness and overall quality, respectively). The performance increases were not just in aggregate, but were found for the vast majority of the sentences (figure 5*b*, centre and right columns). The mean performance for each processing on the complete set of eight metrics computed (STOI, three composite measures and the four component measures they comprise) is shown in electronic supplementary material, table S1.

Performance remained high when the model was trained on 16 speakers and tested on 15 held-out sentences from each of those same 16 speakers (figure 5*c*,*d*). For the STOI ratings, the STDR algorithm showed slight, but significant improvements over both the unfiltered speech (0.01 ($p < 10^{-4}$), d.f. = 3357) and the EM algorithm (0.04 ($p < 10^{-4}$)). Similarly, STDR improved

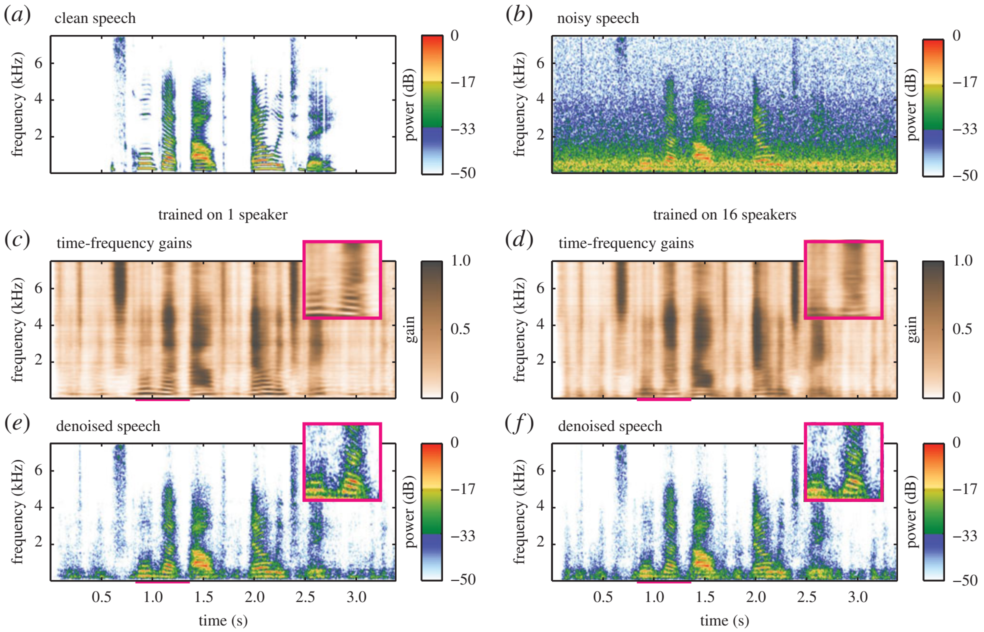

Figure 4. An example of filtering by the STDR algorithm when trained on one or 16 speakers in stationary speech-shaped noise. (*a*) A spectrogram of the sentence 'School got out early today'. (*b*) A spectrogram of the same sentence after the addition of speech-shaped noise at 0 dB SNR. (*c*) The predicted time-frequency gains generated by a model trained only on the speaker of the sentence. (*d*) The predicted time-frequency gains generated by a model trained on 16 different speakers, including the speaker of the sentence. The resulting gains in (*c*,*d*) are similar with similar coarse spectral and temporal structure. Differences between the two are found in their finer spectral structure: the model trained with only a single speaker shows more finely resolved harmonic structure (inset), indicating that the model is more finely tuned to the speaker's characteristic pitches and pitch transitions. (*e*,*f*) The resulting estimated clean speech spectrogram obtained by applying the gains from (*c*) and (*d*), respectively, to (*b*).

the composite ratings of quality over unfiltered speech (0.38 ($p < 10^{-4}$), 0.11 ($p < 10^{-4}$), 0.26 ($p < 10^{-4}$), for signal, noise and overall quality, respectively) and the EM algorithm for all three metrics (0.27 ($p < 10^{-4}$), 0.02 ($p = 5 \times 10^{-4}$), 0.12 ($p < 10^{-4}$), for signal, noise and overall quality, respectively).

(c) Performance on speech in babble noise

A more challenging stimulus set is shown in figure 6. Here, the sentences from the same database were added to babble noise from the NOISEX corpus. Babble noise, being roughly equivalent to the summation of many individual speakers, has the same long-term spectrum as clean speech, but with spectral and temporal modulations somewhere in between individual speakers and speech-shaped noise. Here again, the STDR algorithm extracted complex joint spectro-temporal structure, with better resolution of individual harmonics when trained on a single speaker than trained on 16 speakers (figure 6*c*,*d*, inset). The model trained on a single speaker also showed greater overall levels of contrast, indicating more specificity its ability to detect the target speech. Sound files for both the noisy speech and the denoised estimate can be found in the electronic supplementary material.

Looking again at the entire set of 15 sentences per speaker, composite quality ratings were significantly increased over unfiltered speech (0.61 ($p < 10^{-4}$), 0.66 ($p < 10^{-4}$), 0.60 ($p < 10^{-4}$), speech distortion, noise intrusiveness and overall quality, respectively; figure 7*a*,*b*, composite rating of noise intrusiveness in centre column) and over the EM algorithm (0.16 ($p < 10^{-4}$),

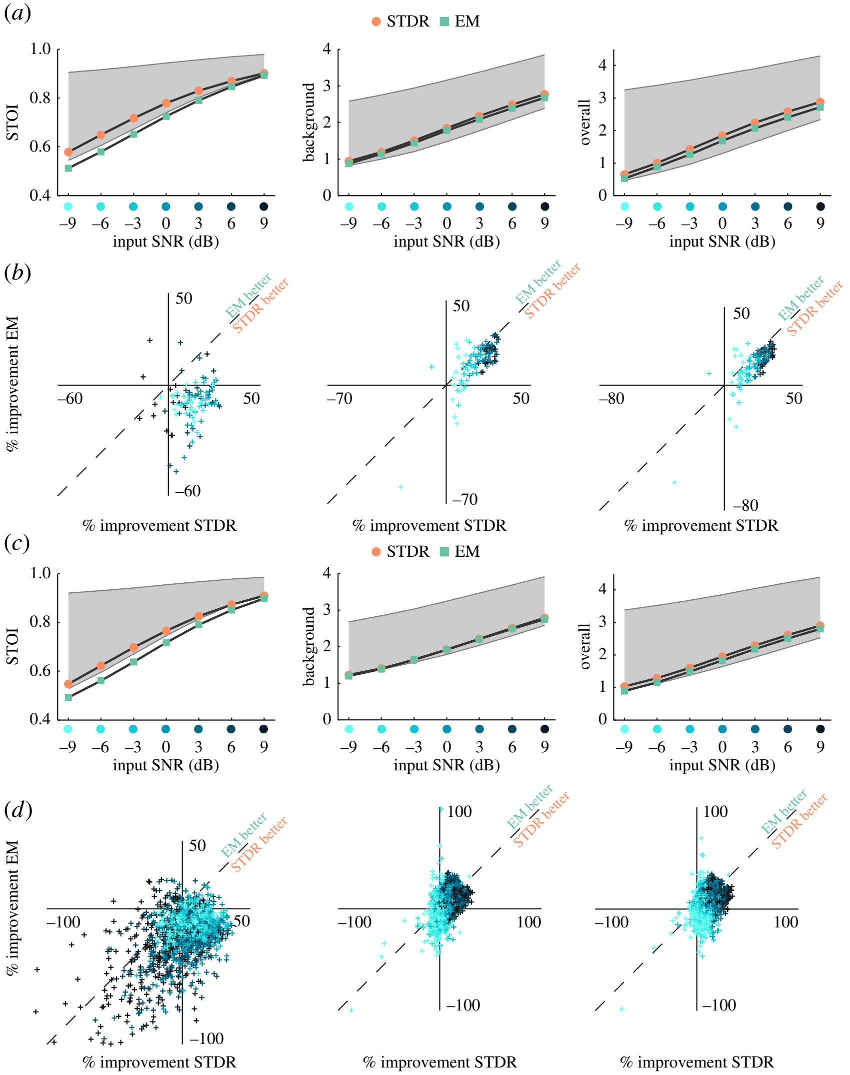

Figure 5. The performance of the algorithm on speech in speech-shaped noise was measured using four different objective measures (three shown). Stimuli consisted of a holdout set of 15 sentences from each speaker processed at seven different SNRs, ranging from −9 to +9 dB SNR. Results shown in rows (*a*) and (*b*) were obtained from the algorithm trained on a single speaker, whereas results in rows (*c*,*d*) were obtained from the algorithm trained on 16 speakers. The measures shown here are the STOI rating, the composite rating of background intrusiveness, and the composite rating of overall quality (see Methods). (*a*) Summary plots of the results obtained on each rating were obtained by averaging over all 15 sentences per SNR. In these plots, the lower bound of the shaded region shows the rating of the unfiltered, noisy speech and the upper bound depicts performance using the optimal time-frequency mask (the ideal gains used as the objective during training). The two lines represent the performance of our algorithm (marked by circles) and the EM algorithm (squares). (*b*) Scatter plots of the normalized ratings (improvement in performance) obtained for each individual sentence (see Methods): the *x*-value corresponds to the sentence processed by STDR, the *y*-value corresponds to the sentence processed by EM, and the shade corresponds to the input SNR for that sentence. Values to the right of the *y*-axis indicate that processing with our algorithm improves the rating over unfiltered. Values above the *x*-axis indicate that processing with EM improves the rating over unfiltered. Values to the right of $y = x$ represent sentences where our algorithm is superior to the EM algorithm. For each metric, the STDR algorithm performed significantly better than both the unfiltered stimulus and the EM algorithm. (*c*, *d*) Same plots as in (*a*, *b*) but for 240 sentences from 16 speakers. The STDR algorithm improved both composite metrics over unfiltered, outperforming the EM algorithm on background intrusiveness. The mean performance for each processing on the complete set of eight metrics computed (STOI, three composite measures and the four component measures they comprise) is shown in electronic supplementary material, tables S1 and S2. (Online version in colour.)

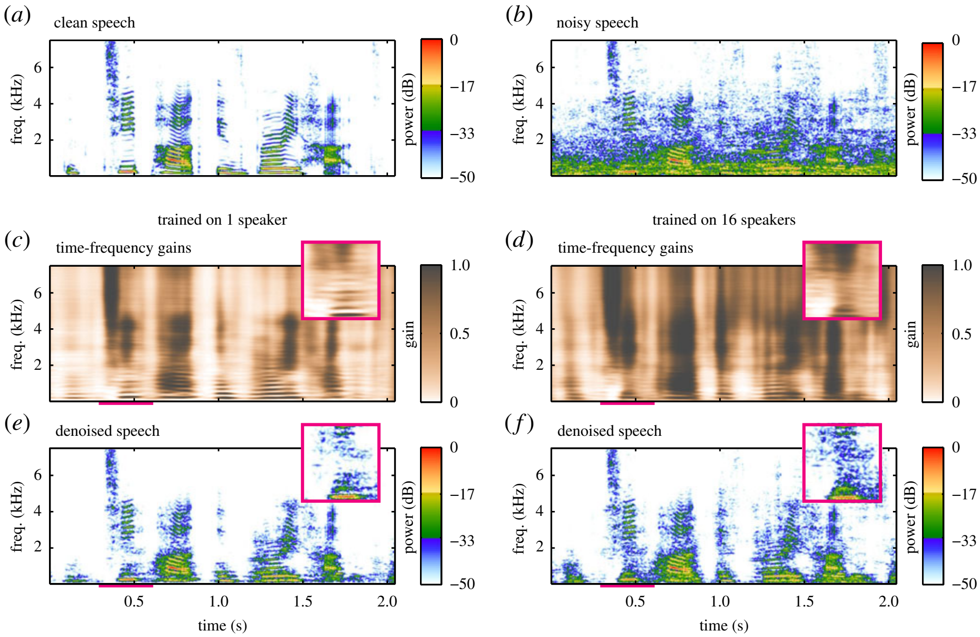

Figure 6. An example of filtering by the STDR algorithm when trained on one or 16 speakers in non-stationary babble noise. The figure layout is identical to figure 4. (*a*) The clean speech spectrogram for the sentence 'The teapot was very hot'. (*b*) Spectrogram for the sentence from (*a*) added to babble noise at 0 dB SNR. (*c*,*d*) Again, the resulting gains in both the one and eight speaker case are very similar but the one speaker model is able to capture more precise harmonic structure (inset). (*e*,*f*) The resulting estimated clean speech spectrogram obtained by applying the gains from (*c*) and (*d*), respectively, to (*b*).

0.45 ($p < 10^{-4}$), 0.23 ($p < 10^{-4}$)). For the composite ratings, the STDR algorithm showed larger benefits over the EM algorithm when processing babble noise instead of speech-shaped noise. This is due primarily to the EM algorithm showing smaller, though still significant, benefits from processing, likely to be because of the temporal non-stationarity of the noise. Performance gains on the STOI measure were lessened, though still significant, for the STDR algorithm (0.02 ($p < 10^{-4}$)) over both unfiltered and EM processed speech.

Performance of STDR trained on 16 speakers was generally similar (figure 7*c*,*d*), so we will focus on the differences. In total, composite ratings of quality were elevated for all processing types, with EM processing improving most for signal and overall quality and STDR improving most for noise intrusiveness (−0.17 ($p < 10^{-4}$), 0.14 ($p < 10^{-4}$), −0.08 ($p < 10^{-4}$), for signal, noise and overall quality, respectively). Both processing types improved quality above unfiltered speech, however. STOI ratings for STDR were insignificantly different than unfiltered speech ($p = 0.98$), whereas the EM algorithm was significantly worse (−0.01 ($p < 10^{-4}$)).

(d) Testing performance generalization

To gauge the flexibility and generalization ability of the STDR algorithm, we trained the algorithm on a total of 280 sentences from 16 speakers embedded in seven different noise types (five QUT noises (see electronic supplementary material, Methods), NOISEX babble noise and speech-shaped noise). The algorithm was then tested on novel sentences from each speaker and noise type, as well as from 114 untrained speakers embedded in 12 different untrained noise types gathered from www.freesound.org (see electronic supplementary material, Methods). The STDR algorithm, as presented here, has too few parameters to effectively handle such diverse and large datasets. In these situations, filtering shows little improvement over unfiltered, though rarely acts as a detriment (electronic supplementary material, figures S2 and S3). In general, STOI was

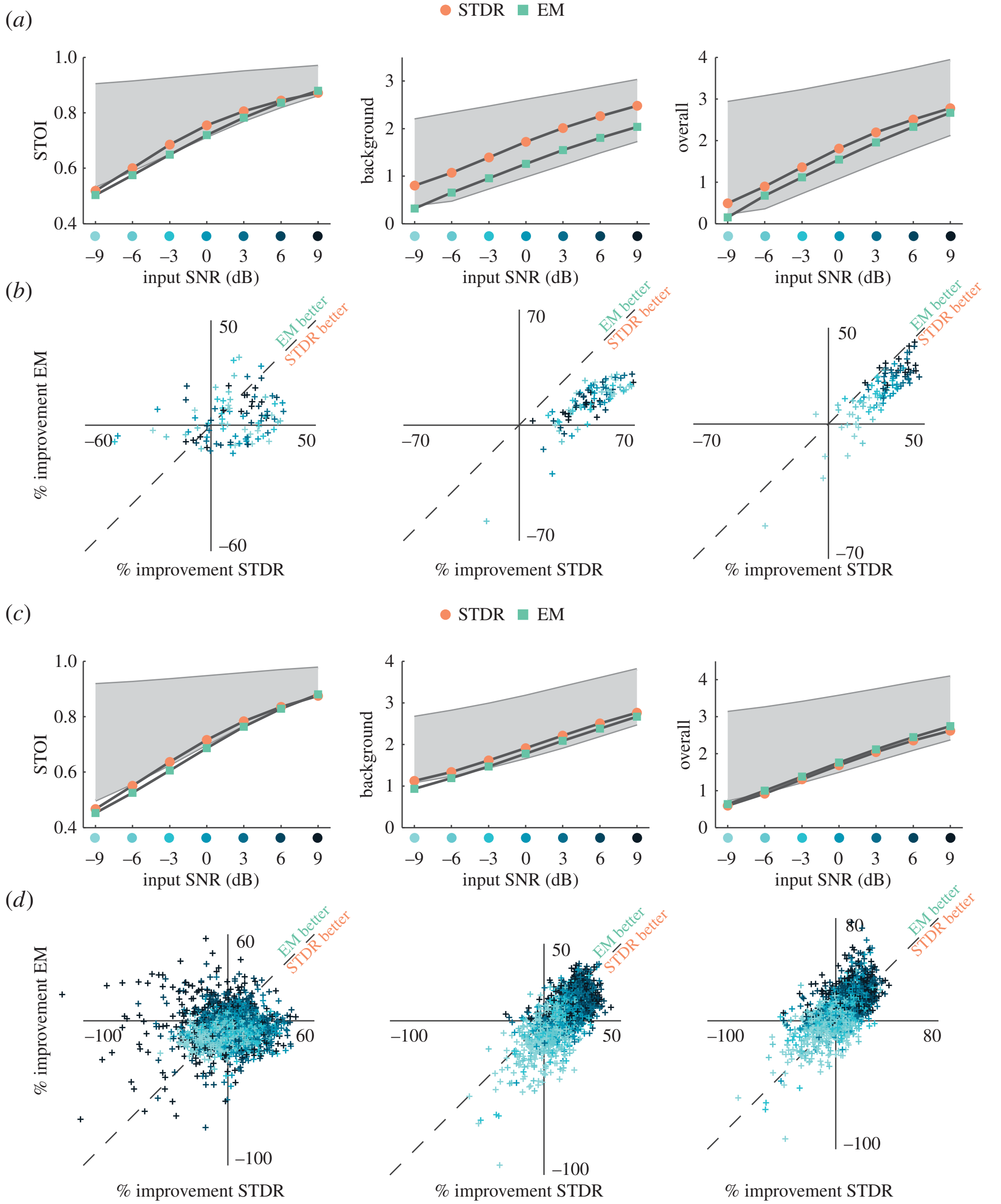

Figure 7. (*a*–*d*) The performance of the algorithm on speech in non-stationary babble noise was measured using four different objective measures (three shown). The figure layout is identical to figure 5. The mean performance for each processing on the complete set of eight metrics are shown in electronic supplementary material, tables 3 and 4. (Online version in colour.)

unaffected or slightly decreased, whereas composite measures were unaffected or significantly improved. Specifically, six of seven noise types (all but babble noise) showed improvement on multiple composite measures. Generalization to novel noise types and speakers was best for stationary noises (white noise and pink noise), as well as backgrounds of birds and rainforest sounds. Again, improvements were primarily seen for composite measures, with STOI showing either no difference or small detriments. These findings are consistent with the differences seen above when training on a single speaker versus 16 speakers. For single speaker instances, the detection and reconstruction kernels can be tuned to very precise structure. Increasing the number of speakers loses some of this precise structure but maintains much of the coarse spectro-temporal structure characteristic of speech. By increasing the diversity of the dataset under investigation, the set of features that can reliably distinguish speech from noise decreases. As discussed below, one could increase the sensitivity to these diminishing discriminating features by increasing the number of units (PCs) or adding intermediate layers in our network.

(e) Performance for different reconstruction delays

Speech contains strong correlations at the timescale of tens to hundreds of milliseconds. These correlations imply that one could build an effective noise reduction algorithm with minimal throughput delay by using mostly *predictive* gains. As the time-frequency gains produced by our algorithm result from the convolution of gain reconstruction kernels with artificial unit activations, we need only adjust the time delays used for the reconstruction window. All previous results displayed an algorithm that was entirely *acausal* in its reconstruction—that is the model-detected features in the past and then attempted to produce gains for those same past time points. The application of such an algorithm would result in a minimum time delay that would correspond to the duration of the gain reconstruction kernel (here 100 ms). We also explored the ability of our algorithm to function using prediction by varying the delay window. For reconstruction kernels of 100 ms duration, entirely *acausal* delays correspond to a central delay of −50 ms, whereas entirely *predictive* delays correspond to a central delay of +50 ms. We tested three additional delays in the middle of these two extremes. Figure 8 shows the results of these experiments using the same performance metrics as before. Here, we have plotted the average performance across 15 novel sentences from a single speaker in both speech-shaped noise and babble noise. All ratings are plotted for sentences at 0 dB SNR. The schematic labels graphically depict the purview of the detection filters and reconstruction gain filters for each condition. Generally, performance was best for entirely *acausal* delays, with gradually decreasing, though still significantly positive, performance with more predictive delays. For both background noises, the STOI was the measure most affected by shifting to predictive delays. For both, the two most predictive algorithms no longer showed a benefit, with the most predictive algorithm decreasing the rating. Conversely, STDR showed significant improvements over unprocessed stimuli for all of the composite ratings at each set of delays used ($p < 10^{-4}$ for all ratings).

4. Discussion

We developed a novel algorithm for single-microphone noise reduction that performs well on several objective measures, across two noise types, and several SNRs and speaker counts. The STDR algorithm functions by detecting joint spectro-temporal features present in either the speech or the noise and using that information to selectively enhance the spectro-temporal features of speech and reduce the spectro-temporal features of noise. The STDR algorithm can be used acausally, providing its best noise reduction at the cost of an inherent time delay. It can also be used predictively, preserving significant noise reduction and with minimal inherent time delay.

This work builds on a large body of research in auditory science that has demonstrated the importance of spectro-temporal modulations in the processing of speech and other natural sounds [7]. All natural sounds reside in a restricted subspace of possible spectro-temporal modulations [6]. The STDR algorithm operates within this subspace, finding the features that allow it to best discriminate between the trained speech and the trained noise. These features, not surprisingly, fall into a few well-known categories. Harmonic stacks are robust indicators of the presence of speech and have been found by many studies to be key sparse features of speech [48,49]. They also provide a basis for noise robust coding in higher auditory brain regions, where selectivity for fast spectral modulations and slow temporal modulations correlates with a neuron's invariance to noise [15]. The slower spectro-temporal modulations present in formants are important features for vowel discrimination [50]. They are modified during clear speech to increase speech intelligibility [51] and are an interesting target for modern speech enhancement

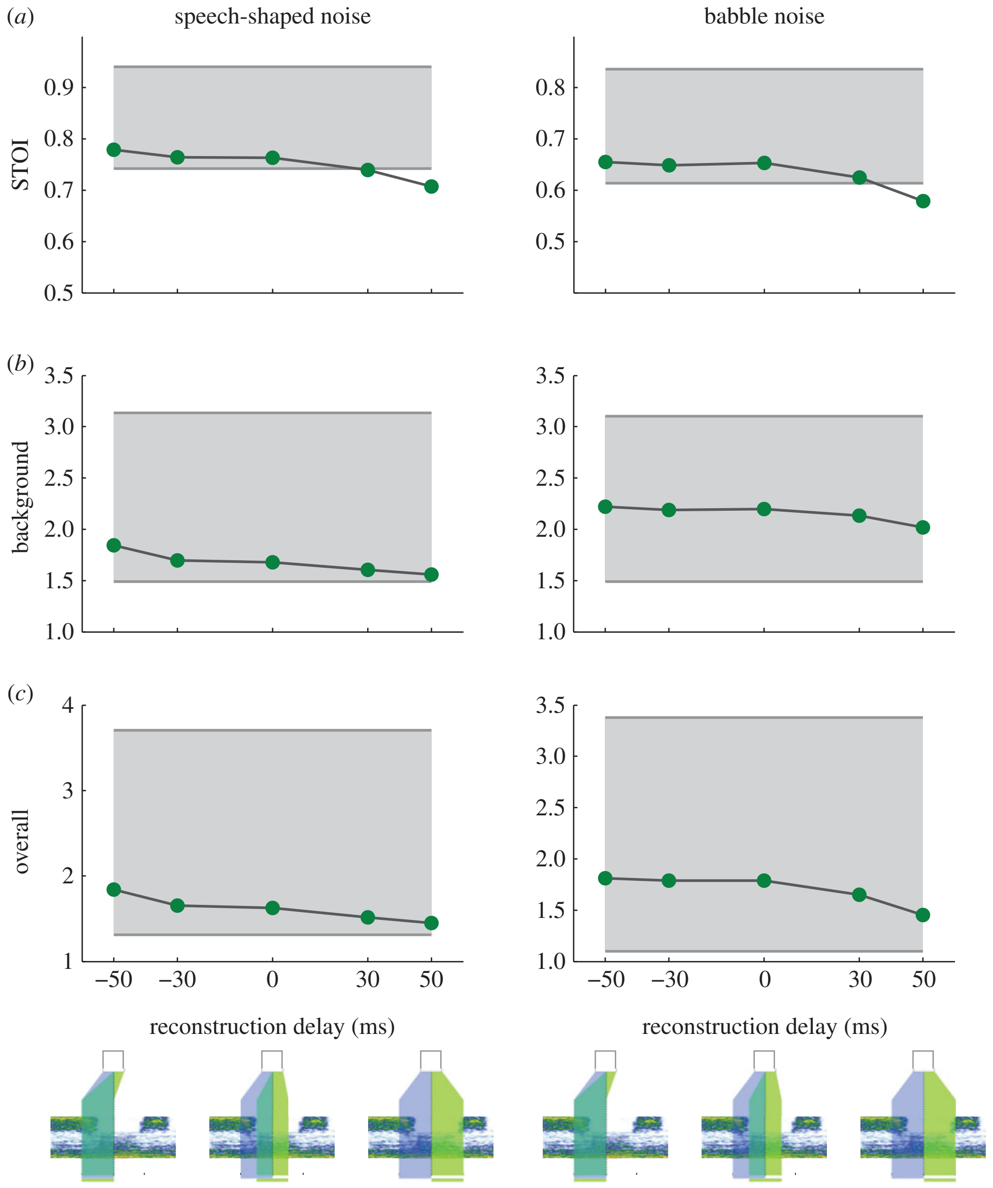

Figure 8. The algorithm performed well using time-frequency gains produced from reconstruction kernels with windows ranging from entirely acausal to entirely predictive. (*a*) Performance using the objective measure for speech intelligibility (STOI). Values on the *x*-axis correspond to the centre time delay of the reconstruction kernel window, characterizing kernels that are entirely *acausal* (left), equally *acausal* and predictive (middle), and entirely predictive (right). Performance values used were from stimuli processed at 0 dB SNR from the speech in speech-shaped noise dataset (left column) and the speech in babble noise dataset (right column). As in figures 5 and 7, the baseline of the shaded region represents the rating of the unfiltered noisy stimulus and the top edge represents the optimal performance obtained using ideal gains. (*b*) Same as in (*a*) but for the composite noise metric. (*c*) Same as in (*a*) but for the composite overall metric. Our algorithm produced significant improvements for all but four cases: STOI using the two most predictive sets of delays in either column. The mean performance for each delay on the complete set of eight metrics is shown in electronic supplementary material, tables S8 and S9.

algorithms [52]. Lastly, the sharp, broadband onsets and offsets of voiceless consonants is a robust feature. Speech-shaped noise and speech averaged across many speakers has a general dearth of high-frequency power [53], a part of the spectrum dominated by voiceless consonants [54]. These slow spectral modulations and fast temporal modulations can be used to discriminate speech or other animal communication signals from environmental sounds [6,12]. STRFs found in both the avian and mammalian auditory cortex have also been shown to cluster, specializing in the detection of slower but more harmonic sound features and faster but spectrally coarse features [55–57] providing a filter bank tuned for extracting the characteristic slow and fast speech features also found in our STDR algorithm. It has also been suggested that the frequency filters in the mammalian auditory periphery are already optimized in this dual task of representing the slower harmonic and more broadband transient sounds of speech [58]. Thus, the detection filters in our STDR algorithm whose structure were originally inspired by research in auditory neuroscience also exhibit, after learning, a distribution of modulation tuning that is akin to what is found in the auditory system.

Our biologically inspired STDR algorithm performs better than a standard model for speech enhancement, the EM algorithm [28], across a wide range of SNRs for each metric tested. The EM algorithm is one of many methods for unsupervised speech enhancement. We have chosen it here because it is a useful and standard benchmark given its simplicity and generality. The EM algorithm is based on reasonable assumptions about the properties of speech and noise. More precisely, it assumes that noise is relatively stationary compared with speech. These assumptions can be modified using carefully designed heuristics, such as automatic voice activity detection or running noise spectrum estimates. While these methods can be quite effective given their simplicity, they are rooted in objectives that treat each time frame as an independent sample, omitting any explicit reference to the joint spectro-temporal structures of sound, which are known to be important both physiologically and psychophysically, as described above. To address this shortcoming, some unsupervised algorithms have worked in the domain of spectro-temporal modulations, with moderate success [59,60]. However, all these approaches remain limited, because, being unsupervised, they necessarily rely on stationary properties of relatively low-level features of the signal and noise: a single estimate of the speech and noise in a particular feature space (e.g. the power spectrum) is assumed to hold across time.

Many studies, including this one, have instead opted to perform supervised learning-based speech enhancement using artificial neural networks [61–64]. Artificial neural networks are a general class of function approximators that make very few assumptions on the nature of the relevant statistics characterizing speech or noise. Moreover, artificial neural networks with proper regularization to prevent over-fitting can work in a large variety of feature spaces. Recently, many algorithms have been proposed that use neural networks to map a time-frequency representation of noisy speech to either a representation of clean speech or a set of time-frequency gains, as performed here. The relative merits of predicting clean speech versus time-frequency gains remain unclear in the literature. In some studies, reconstructing the clean speech spectrogram performed better than attempting to reconstruct the ideal ratio mask, a closely related metric to the optimal gains used in our study [64]. However, other studies have found the opposite [65]. The argument for reconstructing a mask comes from the fact that noise reduction is an inherently discriminative process and thus including a term representing the reconstruction of both the noise and the speech (as is the case when computing a gain) should improve performance [66]. Independent of the type of output reconstructed (i.e. the nature of the objective function), multiple network architectures have been proposed, with autoencoders [67], stacked autoencoders [68], deep neural networks [62–64] and deep recurrent neural networks [69] as the most common. Our STDR algorithm can best be described as a shallow neural network that operates on high-level and time-dependent input and output features: our algorithm is the first to explore the role of spectro-temporal reconstruction in producing optimal gains. Moving to the spectro-temporal domain allows our algorithm to naturally and explicitly capture spectral changes over time, as can be seen in figures 4 and 6 where the time-frequency gains follow the complex spectro-temporal structure of the formants. As far as we know, this is also the first algorithm where the output units operate explicitly on many timeframes. In contrast, existing algorithms commonly reconstruct a single frame or time-frequency point using sound from either past frames or several frames centred on the output. These approaches, unfortunately, leave any coding of the joint spectro-temporal structure of the output embedded implicitly within the network; not only is such implicit coding difficult to visualize or understand, but it will also necessarily lead to more difficult training. It is true that with the advent of recent and more capable training algorithms, the learning the parameters of a many-layered neural network has become possible [70]. These deep networks show significant promise because of their impressive flexibility. Given a large enough training dataset, they can be trained to generalize effectively to a large number of untrained speakers and noise classes [64]. The STDR algorithm, as currently implemented, showed limited generalizability and performed much better on specific tasks. However, given the similarity between STDR and more traditional autoencoders, our algorithm can easily be expanded to include more layers and, in doing so, could further its generalizable performance. Deeper networks greatly expand the feature space where a model can distinguish speech from noise by producing increasingly abstract, combination-sensitive units. In this manner, one could combine the power of deep networks with the biologically inspired architecture of our STDR algorithm that relies on mid-level acoustical features known to be behaviourally relevant and used by the brain.

Finally and importantly, our explicit representation with time extending causally (i.e. in the future) enables us to directly explore the role that spectro-temporal *predictions* might play in real-time speech enhancement. One of the challenges in constructing a real-time algorithm for filtering based on spectro-temporal modulations is that detecting slower temporal features takes time. To adequately detect, a 100 ms vowel from an individual speaker should conceivably require the algorithm to buffer at least 100 ms of sound before applying gains. Yet, because we are using spectro-temporal reconstruction kernels, we can detect predictable features and extrapolate gains into the future. As shown in figure 8, this can be done with little degradation in performance. This strategy is also related to many phenomena observed throughout the auditory system. Most directly, recent work on how humans process speech from multiple simultaneous speakers suggests that cortical oscillations entrain auditory neurons to the attended speaker. This entrainment occurs primarily in the phase of low-frequency (less than or equal to 8 Hz) oscillations and power of high gamma oscillations [17,71–73] and can result in the selective representation of the attended speaker at higher levels and decreased gain on the representation of the unattended speaker at lower levels [72]. These oscillations may represent the alignment of high-excitability periods with predictions of upcoming auditory events [74], synchronizing the neural response to the event. Synchronicity of neural responses is thought to be a critical mechanism by which components of a sound are grouped into coherent auditory objects [75].

At a higher level, prediction is known to play a strong role in the intelligibility of noisy and degraded speech. Reported levels of intelligibility for speech vary wildly depending on the size of the potential response set (e.g. individual phonemes, digits or open-ended words) as well as the amount of context in which a target word is embedded [76–79]. For example, increasing the amount of context in a sentence can increase the intelligibility of the final word in the sentence by nearly 50% [77]. More generally, predictive coding has been shown to play an essential role for perceptual computations in many sensory modalities [80,81].

Because prediction could be a key player in real-time processing of auditory scenes, one could also imagine further improvements to our STDR algorithm. Currently, the predictions are used strictly to generate gains in a feedforward fashion; they provide no feedback and do not modify the activations of the detection filters in any way. The brain, however, appears to use these temporal predictions to modulate the activity to ongoing stimuli. This could be implemented by applying the predicted gains immediately to the incoming stimulus and detecting features on the modified spectrogram. Also, our algorithm relies on prediction only at the level of spectro-temporal modulations. Owing to the modular design of the algorithm, including additional layers of detection and prediction on more abstract features such as phoneme transitions or even words is an intriguing possibility. Additionally, further layers would enable interactions among the detection filters. One current drawback to the algorithm is that, when trained on a sufficiently diverse set of voices, it will readily detect voice features in the background babble noise, despite the intermittent nature of the background voices. A higher layer that aggregates information across units will generally find more evidence for the foreground speaker in the synchronized activity of the detection filters and could weed out the sporadic activation of isolated voice features.

An additional advantage to using an algorithm optimized to the task at hand, such as the STDR, is that it makes no assumptions on the properties of the foreground and background. Because many noise reduction algorithms assume that the background noise is both more stationary and less modulated than the foreground speech, they cannot be flexibly applied to other standard sound source separation problems. The STDR algorithm retains the potential to be applied to situations where the intuitions about foreground and background no longer apply, such as the separation of two competing speakers or of voice from music.

In summary, we have shown that a biologically inspired noise reduction algorithm based on two properties found in the auditory system, the use of spectro-temporal modulation filter banks and adaptive and predictive gains, is capable of outperforming a benchmark noise reduction algorithm. Moreover, it can operate with minimal delay, making it an attractive solution for clinical or engineering applications requiring real-time processing, such as hearing aids and automatic speech recognition. Finally, its modular structure allows for flexibility in its use for signals and noise of different natures and its hierarchical structure will facilitate the implementation of more abstract rules for detection and prediction.

Data accessibility. The datasets supporting this article have been uploaded to the dryad repository with http://dx.doi.org/10.5061/dryad.bn410.

Authors' contributions. T.L. wrote the code and performed all analyses. T.L. and F.E.T. wrote the manuscript.

Competing interests. We have no competing interests.

Funding. This work was supported by funds from the UC Discovery grant no. 13-PC-268969 and the NIDCD grant no. DC007293.

Acknowledgements. We thank Sridhar Kalluri and Bill Woods of the Starkey Hearing Research Center for their collaboration and many thoughtful exchanges in developing this algorithm. We also thank R. Channing Moore, William Vinje and Akshay Rao for their work on this algorithm at various stages. Lastly, we thank the rest of the members of our laboratory and Bruno Olshausen for many discussions over the years.

References

1. Fay RR. 2008 Auditory scene analysis. *Bioacoustics* **17**, 106–109. (doi:10.1080/09524622.2008.9753783)
2. Palmer CV. 2009 A contemporary review of hearing aids. *Laryngoscope* **119**, 2195–2204. (doi:10.1002/lary.20690)
3. Edwards B. 2004 Hearing aids and hearing impairment. In *Speech processing in the auditory system* (eds S Greenberg, WA Ainsworth, AN Popper, RR Fay), pp. 339–421. Berlin, Germany: Springer.
4. Stern RM, Morgan N. 2012 Hearing is believing: biologically-inspired feature extraction for robust automatic speech recognition. *Signal Process. Mag. IEEE* **29**, 34–43. (doi:10.1109/MSP.2012.2207989)
5. Litovsky RY. 2005 Speech intelligibility and spatial release from masking in young children. *J. Acoust. Soc. Am.* **117**, 3091–3099. (doi:10.1121/1.1873913)
6. Singh NC, Theunissen FE. 2003 Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* **114**, 3394–3411. (doi:10.1121/1.1624067)
7. Theunissen FE, Elie JE. 2014 Neural processing of natural sounds. *Nat. Rev. Neurosci.* **15**, 355–366. (doi:10.1038/nrn3731)
8. Eggermont JJ, Aertsen AMHJ, Hermes DJ, Johannesma PIM. 1981 Spectro-temporal characterization auditory neurons: redundant or necessary? *Hear. Res.* **5**, 109–121. (doi:10.1016/0378-5955(81)90030-7)
9. Klein DJ, Simon JZ, Depireux DA, Shamma SA. 2006 Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex. *J. Comput. Neurosci.* **20**, 111–136. (doi:10.1007/s10827-005-3589-4)
10. Theunissen FE, Sen K, Doupe AJ. 2000 Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J. Neurosci.* **20**, 2315–2331.
11. Sharpee TO, Atencio CA, Schreiner CE. 2011 Hierarchical representations in the auditory cortex. *Curr. Opin. Neurobiol.* **21**, 761–767. (doi:10.1016/j.conb.2011.05.027)

12. Woolley SMN, Fremouw TE, Hsu A, Theunissen FE. 2005 Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat. Neurosci.* **8**, 1371–1379. (doi:10.1038/nn1536)
13. Rodríguez FA, Chen C, Read HL, Escabí MA, Rodriguez FA, Chen C, Read HL, Escabi MA. 2010 Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. *J. Neurosci.* **30**, 15 969–15 980. (doi:10.1523/JNEUROSCI.0966-10.2010)
14. Escabí MA, Miller LM, Read HL, Schreiner CE. 2003 Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *J. Neurosci.* **23**, 11 489–11 504.
15. Moore RC, Lee T, Theunissen FE. 2013 Noise-invariant neurons in the avian auditory cortex: hearing the song in noise. *PLoS Comput. Biol.* **9**, e1002942. (doi:10.1371/journal.pcbi.1002942)
16. Mesgarani N, David SV, Fritz JB, Shamma SA. 2014 Mechanisms of noise robust representation of speech in primary auditory cortex. *Proc. Natl Acad. Sci. USA* **111**, 6792–6797. (doi:10.1073/pnas.1318017111)
17. Mesgarani N, Chang EF. 2012 Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236. (doi:10.1038/nature11020)
18. Zion Golumbic EM, Poeppel D, Schroeder CE. 2012 Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain Lang.* **122**, 151–161. (doi:10.1016/j.bandl.2011.12.010)
19. Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH. 2013 Constructing noise-invariant representations of sound in the auditory pathway. *PLoS Biol.* **11**, e1001710. (doi:10.1371/journal.pbio.1001710)
20. Schneider DM, Woolley SMN. 2013 Sparse and background-invariant coding of vocalizations in auditory scenes. *Neuron* **79**, 141–152. (doi:10.1016/j.neuron.2013.04.038)
21. Jørgensen S, Dau T. 2011 Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J. Acoust. Soc. Am.* **130**, 1475. (doi:10.1121/1.3621502)
22. Dubbelboer F, Houtgast T. 2008 The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *J. Acoust. Soc. Am.* **124**, 3937–3946. (doi:10.1121/1.3001713)
23. Elhilali M, Chi T, Shamma SA. 2003 A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Commun.* **41**, 331–348. (doi:10.1016/S0167-6393(02)00134-6)
24. Elliott TM, Theunissen FE. 2009 The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* **5**, e1000302. (doi:10.1371/journal.pcbi.1000302)
25. Chabot-Leclerc A, Jørgensen S, Dau T. 2014 The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction. *J. Acoust. Soc. Am.* **135**, 3502–3512. (doi:10.1121/1.4873517)
26. Hannemann R, Obleser J, Eulitz C. 2007 Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Res.* **1153**, 134–143. (doi:10.1016/j.brainres.2007.03.069)
27. Holdgraf C, de Heer W, Knight RT, Theunissen FE. Submitted. Rapid tuning adaptation in human auditory cortex enhances speech intelligibility. *Nat. Neurosci.*
28. Ephraim Y, Malah D. 1985 Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust.* **33**, 443–445. (doi:10.1109/TASSP.1985.1164550)
29. Martin R. 1994 Spectral subtraction based on minimum statistics. In *Proc. EUSIPCO 94, Edinburgh, UK, 13–15 September*, pp. 1182–1185. Kessariani, Greece: EURASIP.
30. Martin R. 2001 Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* **9**, 504–512. (doi:10.1109/89.928915)
31. Nilsson M, Soli SD, Sullivan JA. 1994 Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.* **95**, 1085–1099. (doi:10.1121/1.408469)
32. Bradlow AR, Ackerman L, Burchfield LA, Hesterberg L, Luque J, Mok K. 2011 Language- and talker-dependent variation in global features of native and non-native speech. In *Proc. Int. Cong. Phonetic Sciences, Hong Kong, China, 17–21 August*, pp. 356–359.
33. Dean D, Sridharan S, Vogt R, Mason M. 2010 The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. In *Proc. Interspeech, Makuhari, Japan, 26–30 September*, pp. 6–8.
34. Boll S. 1979 Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust.* **27**, 113–120. (doi:10.1109/TASSP.1979.1163209)

35. McAulay R, Malpass M. 1980 Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust*. **28**, 137–145. (doi:10.1109/TASSP.1980.1163394)
36. Ephraim Y, Malah D. 1984 Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust*. **32**. (doi:10.1109/TASSP.1984.1164311)
37. Flanagan JL. 1980 Parametric coding of speech spectra. *J. Acoust. Soc. Am*. **68**, 412–419. (doi:10.1121/1.384752)
38. Lyon R. 1982 A computational model of filtering, detection, and compression in the cochlea. In *ICASSP '82 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Paris, France, 3–5 May*, pp. 1282–1285. Piscataway, NJ: IEEE.
39. Hinton GE, Salakhutdinov RR. 2006 Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507. (doi:10.1126/science.1127647)
40. Mesgarani N, David SV, Fritz JB, Shamma SA. 2008 Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am*. **123**, 899. (doi:10.1121/1.2816572)
41. Glorot X, Bengio Y. 2010 Understanding the difficulty of training deep feedforward neural networks. In *Proc. 13th Int. Conf. Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy, 13–15 May*, pp. 249–256.
42. Hu Y, Loizou PC. 2008 Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio, Speech Lang. Process*. **16**, 229–238. (doi:10.1109/TASL.2007.911054)
43. Hansen JHL, Pellom BL. 1998 An effective quality evaluation protocol for speech enhancement algorithms. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP), Sydney, Australia, 30 November–4 December*, pp. 2819–2822.
44. Klatt D. 1982 Prediction of perceived phonetic distance from critical-band spectra: a first step. In *IEEE Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP '82.*, pp. 1278–1281. Piscataway, NJ: IEEE.
45. Quackenbush SR, Barnwell TP, Clements MA. 1988 *Objective measures of speech quality*. Englewood Cliffs, NJ: Prentice Hall PTR.
46. Rix AW, Beerends JG, Hollier MP, Hekstra AP. 2001 Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP '01 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, 7–11 May*, pp. 2–5.
47. Taal CH, Hendriks RC, Heusdens R, Jensen J. 2011 An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio, Speech Lang. Process*. **19**, 2125–2136. (doi:10.1109/TASL.2011.2114881)
48. Klein DJ, König P, Kording KP. 2003 Sparse spectrotemporal coding of sounds. *EURASIP J. Appl. Signal Process*. **2003**, 659–667. (doi:10.1155/S1110865703303051)
49. Carlson NL, Ming VL, DeWeese MR. 2012 Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput. Biol*. **8**, 1–15. (doi:10.1371/journal.pcbi.1002594)
50. Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. 1967 Perception of the speech code. *Psychol. Rev*. **74**, 431–461. (doi:10.1037/h0020279)
51. Amano-Kusumoto A, Hosom J-P. 2011 *A review of research on speech intelligibility and correlations with acoustic features*. Technical Rep. CSLU-011-001. Beaverton, OR: Oregon Health & Science University.
52. Rao A, Carney LH. 2014 Speech enhancement for listeners with hearing loss based on a model for vowel coding in the auditory midbrain. *IEEE Trans. Biomed. Eng*. **61**, 2081–2091. (doi:10.1109/TBME.2014.2313618)
53. Byrne D *et al*. 1994 An international comparison of long-term average speech spectra. *J. Acoust. Soc. Am*. **96**, 2108–2120. (doi:10.1121/1.410152)
54. Heinz JM, Stevens KN. 1961 On the properties of voiceless fricative consonants. *J. Acoust. Soc. Am*. **33**, 589. (doi:10.1121/1.1908734)
55. Miller LM, Escabí MA, Read HL, Schreiner CE. 2002 Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J. Neurophysiol*. **87**, 516–527.
56. Nagel KI, Doupe AJ. 2008 Organizing principles of spectro-temporal encoding in the avian primary auditory area field L. *Neuron* **58**, 938–955. (doi:10.1016/j.neuron.2008.04.028)
57. Woolley SMN, Gill PR, Fremouw T, Theunissen FE. 2009 Functional groups in the avian auditory system. *J. Neurosci*. **29**, 2780–2793. (doi:10.1523/JNEUROSCI.2042-08.2009)
58. Smith EC, Lewicki MS. 2006 Efficient auditory coding. *Nature* **439**, 978–982. (doi:10.1038/nature04485)

59. Mesgarani N, Shamma S. 2005 Speech enhancement based on filtering the spectrotemporal modulations. In *ICASSP '05. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Philadelphia, PA, 18–23 March*, pp. 1105–1108. Piscataway, NJ: IEEE.
60. Hsu C-C, Cheong K-M, Chien J-T, Chi T-S. 2015 Modulation Wiener filter for improving speech intelligibility. In *IEEE Int Conf. on Acoustics, Speech and Signal Processing* (*ICASSP*), 2015, pp. 370–374.
61. Wan EA, Nelson AT. 1998 Networks for speech enhancement. In *Handbook of neural networks for speech processing* (ed. S Katagiri), pp. 1–27, 1st edn. Norwood, MA: Artech House.
62. Healy EW, Yoho SE, Wang Y, Wang D. 2013 An algorithm to improve speech recognition in noise for hearing-impaired listeners. *J. Acoust. Soc. Am.* **134**, 3029–3038. (doi:10.1121/1.4820893)
63. Narayanan A, Wang D. 2013 Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *ICASSP '13. IEEE Int. Conf. Acoust. Speech, Signal Process.* pp. 7092–7096. Piscataway, NJ: IEEE.
64. Xu Y, Du J, Dai L, Lee C. 2015 A regression approach to speech enhancement based on deep neural networks. *IEEE Trans. Audio, Speech Lang. Process.* **23**, 7–19. (doi:10.1109/TASLP.2014.2364452)
65. Weninger F, Hershey JR, Le Roux J. 2014 Discriminatively trained recurrent neural networks for single-channel speech separation. In *Glob. 2014 Mach. Learn. Appl. Speech Process*, pp. 577–581.
66. Huang P-S, Kim M, Hasegawa-Johnson M, Smaragdis P. 2014 Deep learning for monaural speech separation. In *ICASSP '14. IEEE Int. Conf. Acoust. Speech, Signal Process*, pp. 1562–1566. Piscataway, NJ: IEEE.
67. Xia B, Bao C. 2014 Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Commun.* **60**, 13–29. (doi:10.1016/j.specom.2014.02.001)
68. Lu X, Tsao Y, Matsuda S, Hori C. 2013 Speech enhancement based on deep denoising autoencoder. In *Interspeech, Lyon, France, 25–29 August*, pp. 436–440.
69. Weninger F, Eyben F, Schuller B. 2014 Single-channel speech separation with memory-enhanced recurrent neural networks. In *ICASSP '14. IEEE Int. Conf. Acoust. Speech, Signal Process*, pp. 3709–3713. Piscataway, NJ: IEEE.
70. Hinton GE, Osindero S, Teh Y-W. 2006 A fast learning algorithm for deep belief nets. *Neural Comput.* **8**, 1527–1554.
71. Ding N, Simon JZ. 2013 Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J. Neurosci.* **33**, 5728–5735. (doi:10.1523/JNEUROSCI.5297-12.2013)
72. Zion Golumbic EM *et al.* 2013 Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party'. *Neuron* **77**, 980–991. (doi:10.1016/j.neuron.2012.12.037)
73. Ding N, Simon JZ. 2012 Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl Acad. Sci. USA* **109**, 11 854–11 859. (doi:10.1073/pnas.1205381109)
74. Schroeder CE, Lakatos P. 2009 Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci.* **32**, 9–18. (doi:10.1016/j.tins.2008.09.012)
75. Shamma SA, Elhilali M, Micheyl C. 2011 Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* **34**, 114–123. (doi:10.1016/j.tins.2010.11.002)
76. Pichora-Fuller MK. 2008 Use of supportive context by younger and older adult listeners: balancing bottom-up and top-down information processing. *Int. J. Audiol.* **47**(Suppl. 2), S72–S82. (doi:10.1080/14992020802307404)
77. Kalikow DN, Stevens KN, Elliott LL. 1977 Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* **61**, 1337–1351. (doi:10.1121/1.381436)
78. Miller GA, Heise GA, Lichten W. 1951 The intelligibility of speech as a function of the context of the test materials. *J. Exp. Psychol.* **41**, 329–335. (doi:10.1037/h0062491)
79. Bronkhorst AW, Bosman AJ, Smoorenburg GF. 1993 A model for context effects in speech recognition. *J. Acoust. Soc. Am.* **93**, 499–509. (doi:10.1121/1.406844)
80. Summerfield C, de Lange FP. 2014 Expectation in perceptual decision making: neural and computational mechanisms. *Nat. Rev. Neurosci.* **15**, 745–756. (doi:10.1038/nrn3838)
81. Clark A. 2013 Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204. (doi:10.1017/S0140525X12000477)