Hearing through the noise: biologically inspired noise reduction

by

Tyler Paul Lee

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Neuroscience

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Frederic Theunissen, Co-chair
Professor Bruno Olshausen, Co-chair
Professor Jack Gallant
Professor Keith Johnson
Professor Friedrich Sommer

Summer 2016

Hearing through the noise: biologically inspired noise reduction

Abstract

Hearing through the noise: biologically inspired noise reduction

by

Tyler Paul Lee

Doctor of Philosophy in Neuroscience

University of California, Berkeley

Professor Frederic Theunissen, Co-chair

Professor Bruno Olshausen, Co-chair

Vocal communication in the natural world demands that a listener perform a remarkably complicated task in real-time. Vocalizations mix with all other sounds in the environment as they travel to the listener, arriving as a jumbled low-dimensional signal. A listener must then use this signal to extract the structure corresponding to individual sound sources. How this computation is implemented in the brain remains poorly understood, yet an accurate description of such mechanisms would impact a variety of medical and technological applications of sound processing. In this thesis, I describe initial work on how neurons in the secondary auditory cortex of the Zebra Finch extract song from naturalistic background noise. I then build on our understanding of the function of these neurons by creating an algorithm that extracts speech from natural background noise using spectrotemporal modulations. The algorithm, implemented as an artificial neural network, can be flexibly applied to any class of signal or noise and performs better than an optimal frequency-based noise reduction algorithm for a variety of background noises and signal-to-noise ratios. One potential drawback to using spectrotemporal modulations for noise reduction, though, is that analyzing the modulations present in an ongoing sound requires a latency set by the slowest temporal modulation computed. The algorithm avoids this problem by reducing noise predictively, taking advantage of the large amount of temporal structure present in natural sounds. This predictive denoising has ties to recent work suggesting that the auditory system uses attention to focus on predicted regions of spectrotemporal space when performing auditory scene analysis.

Contents

Acknowledgments

This thesis would not have been possible without the unwavering support from so many people. For this, I must first acknowledge the HWNI for setting up such a great community of wonderful people.

Of course, a great many thanks go to my two advisors, Frederic and Bruno. Frederic showed me what it means to love the intricacies inherent in any complicated analysis, even, or perhaps especially, when those intricacies don't love you back. Science is messy and hard, and I have never seen anyone who enjoyed those facts as much as Frederic. Bruno was essential in keeping me going whenever my research was not working out according to plan. Every time I came to his office with a problem, in research or life, I came out excited and clear-headed about what to do next. The mentorship I received from each of them has been invaluable to my growth as a scientist and as a person. Thank you to them and the rest of my thesis committee for their guidance throughout the years.

I must also thank the rest of the Theunissen lab. Never have a I met a nicer group of people. Wendy, Mike and Solveig were wonderful labmates and friends. Channing Moore helped me a lot during my first year and was a great collaborator on all of the work done in chapter 2. A very special thank you goes to Yuka Minton for all of her hard work caring for the Zebra Finches, not to mention being a great person and a close friend to everyone in lab. Thank you to Julie as well for being a great scientific role model. Her hard work and breadth of knowledge is truly impressive.

I am grateful to the Jack and the rest of the Gallant lab for always being ready to advise on confusing points in math and neural networks and for their hard work on our computer infrastructure. The work presented here literally might not have finished if not for their help. Additionally, I must acknowledge Fritz Sommer and the entire Redwood Center for so many great discussions and a tremendous amount of feedback. I have learned a lot from them all.

Finally, none of this would have been possible without the support from my entire family and my loving partner, Yasyn. I would not have made it nearly this far if it were not for all of the love and confidence they have given me. It's impossible to express how important they are to me. Thank you so much.

Chapter 1

Introduction

The natural world is full of sound. The fluctuations in air pressure evoked by an ever-changing array of sources interfere with each other both constructively and destructively. It is only this linear mixture that is available to any organism attempting to make sense of sound. It is thus a wonder that one of the primary modes of communication for many organisms, humans included, is vocalization. As challenging as extracting information regarding a single sound source from a sea of sounds may seem, this problem is thought to be solved by a surprisingly diverse set of organisms (Bee and Micheyl 2008; Hulse 2002; Fay 2008), perhaps indicating that the computational principles used to solve it are ubiquitous in nervous systems. While these computational principles remain poorly understood, their fundamentals have become increasingly clear. Auditory scene analysis relies heavily on the ability to robustly extract patterns from noisy inputs and the ability to predict temporal sequences of such patterns. Though these computations are not unique to sound processing (Lewicki et al. 2014), the auditory system will be the focus of the present work.

Reliably recognizing sounds in noisy environments is no simple feat. Auditory scenes are composed of sounds from a variety of sound classes, ranging from relatively unstructured wind noise to highly structured sounds like music and speech. For organisms to survive and communicate in the natural world, it is imperative that they are able to recognize the specific signals of an approaching predator or a nearby loved one from within the scene. Auditory scene analysis is the general problem of segregating a composite signal of many sound sources into a set of signals each corresponding to a single sound source. Oftentimes the more relevant challenge, however, is the problem of signal extraction or noise reduction. Here, the composite signal is separated into two components: a single sound source foreground signal and a (potentially) multi-sound source background noise. It is important to keep in mind that the classification of individual sound sources as belonging to signal or noise is entirely dependent on their behavioral relevance at a specific point in time. Even so, many classes of sounds are structured more like sound textures and are more readily classified as background noise. Among these textures are common noises of wind, rain, running water, and large-scale summation of individual vocalizations in a crowd, which are not able to be identified individually unless other strong cues about identity are provided (McDermott,

Wrobleski, and Oxenham 2011). The work presented here focuses primarily on mechanisms, both neuronal and algorithmic, for extracting vocalizations from these types of noises.

Recognizing sounds in noisy environments

Specific sounds must often be recognized in the presence of significant background noise (Lengagne et al. 1999; Hulse 2002), necessitating the ability to filter out sounds that are not likely to be behaviorally relevant (Bee and Micheyl 2008; Fay 2008). This build through each stage of the auditory system, as individual neurons extract the maximally informative features present in their inputs (for a review see Theunissen and Elie 2014). Neurons as early as the auditory nerve compute a sparse code of incoming sound waveforms, providing a compact representation of repeated structure in this low-dimensional signal (Smith and Lewicki 2006). Complex, behaviorally-relevant patterns have structure on many scales, however. The auditory system gradually extracts this more complicated, slower structure as information ascends the auditory pathway (Sarah M N Woolley, P. R. Gill, et al. 2009; Kim and A. Doupe 2011; Sharpee, Atencio, and Schreiner 2011). Neurons in the inferior colliculus appear to represent a sparse code of the spectrotemporal features of sound (Rodríguez et al. 2010; Carlson, Ming, and DeWeese 2012). Neurons in the auditory thalamus and primary auditory cortex have been shown to efficiently code spectrotemporal features (Sarah M N Woolley, Fremouw, et al. 2005). As tuning becomes increasingly abstracted from the low-dimensional sound waveform, neurons in higher auditory areas like primary and secondary auditory begin to show significant levels of robustness to background noise. This feature of higher auditory processing is the first major thread of the present work. The build-up of robust and efficient neural codes of natural sounds is reviewed more thoroughly in sections 2.4 and 3.2.

Aiding recognition with prediction

Extending the model that the auditory system encodes a compact representation of the informative features on incoming sounds, it seems inevitable that temporal prediction should play a significant role in complex sound processing. Sound is an inherently temporal stimulus with predictable structure at many time-scales (Voss and Clarke 1978). Since multiple sound sources are often largely independent of one another, the amount with which incoming sound features cohere with temporal predictions based on past sounds should be a reliable indicator on which to separate sources (Daniel P W Ellis 1999). This hypothesis is consistent with a large body of psychoacoustical evidence that demonstrates that temporal context has a direct consequential role in the perception of noisy or ambigous stimuli (Warren 1970; G. A. Miller, Heise, and Lichten 1951). This is most striking in the case of *phonemic restoration*. If a single phoneme in a sentence is replaced with a bout of silence, the change is perceptually salient and intelligibility of the affected word is decreased (Warren 1970). However, if the silent period is replaced by some form of transient broadband noise (e.g. a cough), the noise is perceived to come from a separate sound source and, most importantly, the sentence is

perceived to have continued uninterrupted. The brain is somehow capable of filling in regions where pattern recognition is impossible with a pattern inferred from temporal context. This principle suggests a more general role of predictive tuning in robust pattern recognition and forms the backbone of some computational auditory scene analysis methods (Daniel PW W Ellis 1996; Daniel P W Ellis 1999). More recently, neurophysiological evidence suggests that attentional selection in auditory scene analysis may be performed predictively (Bendixen 2014). Auditory cortex exhibits neuronal oscillations that encode incoming sound envelopes (Luo and Poeppel 2007). This encoding is preferential for the attended auditory stream in a multi-stream environment (Mesgarani and Chang 2012). These findings, combined with the developing hypothesis of attentional selection through coherent oscillations (Lakatos et al. 2008; Charles E. Schroeder and Lakatos 2009), suggest that the ability of the auditory system to predictively align high excitability regions of ongoing oscillations to incoming sounds from an attended sound source is critically important in natural listening conditions. This is the second main thread of the present work and is a burgeoning field for further research.

Noise reduction in the real world

Normal hearing individuals are remarkably good at extracting information from noisy sounds. Speech intelligibility in noise takes advantage of a great variety of cues and is generally extremely proficient (Bronkhorst 2000). However, not all individuals have normal hearing, nor are all systems that must process noisy sounds individuals. As much as 5.3% of people have some form of debilitating hearing impairment ("WHO global estimates on prevalence of hearing loss" 2012), which has a profound effect on ones' ability to hold a conversation in even moderate background noise (Palmer 2009). This has a dramatic impact on quality of life and is a principal complaint of individuals with assistive hearing devices (Edwards 2004). Another domain that struggles with sound processing in noise is automatic speech recognition (ASR) and related computational technologies. In recent years, ASR has dramatically improved, approaching even human-level abilities, yet robust recogntion in noisy environments remains a signficant challenge (J. Li et al. 2014). In the present work, we sought to use our knowledge of how the auditory system extracts natural sounds from background noise to create a frontend noise reduction algorithm with applications in both hearing aid and computer perception technologies. Existing, often less biologically-inspired, algorithms for frontend noise reduction are reviewed in sections 2.4 and 3.2.

Outline

In chapter 2 we describe one of the first neurophysiological studies into the processing of natural sounds in background noise. This work was published as Moore, Lee, and Theunissen 2013 and built upon a strong body of literature demonstrating that the principle tuning of higher level auditory neurons, especially beyond the inferior colliculus, is in the domain of spectrotemporal modulations. Different categories of sounds map clearly into distinct but overlapping regions of spectrotemporal modulations, providing a good basis in which

to discriminate communications sounds from other naturalistic sounds. We showed that a neuron's tuning in the space of spectrotemporal modulations correlates with its degree of invariance to the presence of background noise in the stimulus. Noise-invariant neurons were more likely to be found in the CaudoMedial Nidopalium (NCM), an avian secondary auditory region, and had receptive fields that focussed on slow temporal modulations and fast spectral modulations. To demonstrate that an ensemble of neurons with noise invariant responses provides a sufficient basis for separating signal from noise, we designed a real-time noise reduction algorithm that employed artificial neurons with a variety of spectrotemporal receptive field shapes. As with the noise-invariant neurons in area NCM, the artificial neurons in the algorithm that were found to be most important for the task of separating signal from noise were sensitive to fast spectral modulations and slow temporal modulations.

In chapter 3 we describe a large extension of the noise reduction algorithm introduced in chapter 2 and published in Lee and Theunissen 2015. The algorithm was used to separate human speech from a variety of background noises. To do so, we expanded the optimization to learn the spectrotemporal features best suited both for detecting speech and noise structure and reconstructing discriminative gains that map noisy speech to clean speech. The algorithm was the first, to our knowledge, to explicitly use spectrotemporal features for both detection and reconstruction, as well as explore the role of temporal prediction in real-time denoising. We showed that it outperformed a standard noise reduction algorithm in multiple noise conditions and speaker conditions, across a wide range of signal-to-noise ratios.

Chapter 2

Noise-invariant Neurons in the Avian Auditory Cortex: Hearing the Song in Noise

Moore, R. Channing, Lee, Tyler & Theunissen, Frederic

2.1 Abstract

Given the extraordinary ability of humans and animals to recognize communication signals over a background of noise, describing noise invariant neural responses is critical not only to pinpoint the brain regions that are mediating our robust perceptions but also to understand the neural computations that are performing these tasks and the underlying circuitry. Although invariant neural responses, such as rotation-invariant face cells are well described in the visual system, high-level auditory neurons that can represent the same behaviorally relevant signal in a range of listening conditions have yet to be discovered. Here we found neurons in a secondary area of the avian auditory cortex that exhibit noise-invariant responses in the sense that they responded with similar spike patterns to song stimuli presented in silence and over a background of naturalistic noise. By characterizing the neurons tuning in terms of their responses to modulations in the temporal and spectral envelope of the sound, we then show that noise invariance is partly achieved by selectively responding to long sounds with sharp spectral structure. Finally, to demonstrate that such computations could explain noise invariance, we

designed a biologically inspired noise-filtering algorithm that can be used to separate song or speech from noise. This novel noise-filtering method performs as well as other state-of-the-art de-noising algorithms and could be used in clinical or consumer oriented applications. Our biologically inspired model also shows how high-level noise-invariant responses could be created from neural responses typically found in primary auditory cortex.

2.2 Introduction

Invariant neural representations of behaviorally relevant objects are a hallmark of high-level sensory regions and are interpreted as the outcome of a series of computations that would allow us to recognize and categorize objects in *real life* situations. For example, view-invariant face neurons have been found in the inferior temporal cortex (Freiwald and Tsao 2010) and are thought to reflect our abilities to recognize the same face from different orientations and scales. The representation of auditory objects by the auditory system is less well understood although neurons in high-level auditory areas can be very selective for complex sounds and, in particular, communication signals (Rauschecker et al. 1995). It has also been shown that auditory neurons can be sound level invariant (Sadagopan and Xiaoqin Wang 2008; Billimoria et al. 2008) or pitch sensitive (Bendor and Xiaoqin Wang 2005). As is the case for all neurons labeled as invariant, pitch sensitive neurons respond similarly to many different stimuli as long as these sounds yield the same pitch percept. Both sound level invariant and pitch sensitive neurons could therefore be building blocks in the computations required to produce invariant responses to particular auditory signals subject to distortions due to propagations or corruption by other auditory signals. The existence of such distortion invariant auditory neurons, however, remains unknown. Similarly, the neuronal computations required to recognize communication signals embedded in noise are not well understood although it is known that humans (Bronkhorst 2000) and other animals (Bee and Micheyl 2008) excel at this task. In this study, we examined how neurons in the secondary avian auditory cortical area NCM (*CaudoMedial Nidopalium*) responded to song signals embedded in background noise to test whether this region presents noise-invariant char-

acteristics that could be involved in robust song recognition. We chose the avian model system because birds excel at recognizing individuals based on their communication calls (Vignal, Mathevon, and Mottin 2004), often in very difficult situations (Aubin and Jouventin 2002). Moreover, the avian auditory system is well characterized and it is known that neurons in higher-level auditory regions can respond selectively to particular conspecific songs (Knudsen and Gentner 2010). We focused our study on NCM because a series of neurophysiological (Stripling, Volman, and D. F. Clayton 1997) and immediate early gene studies (Mello, Nottebohm, and D. Clayton 1995; Bolhuis et al. 2000) have implicated this secondary auditory area in the recognition of familiar songs. In addition, although neuronal responses in the primary avian auditory cortex regions are systematically degraded by noise (Narayan et al. 2007), studies using immediate early gene activation suggested that responses to conspecific song in NCM were relatively constant for a range of behaviorally relevant noise levels (Vignal, Attia, et al. 2004).

2.3 Methods

Neurophysiology and Histology

All animal procedures were approved by our institutional Animal Care and Use Committee. Neurophysiological recordings were performed in four, urethane anesthetized adult zebra finches to obtain 50 single unit recordings in areas NCM and potentially field L (see below). We used similar neurophysiological and histological methods to characterize other regions of the avian auditory processing stream and detailed descriptions can be found there (Sarah M N Woolley, P. R. Gill, et al. 2009). The methods are summarized here and differences when they exist are noted.

To obtain recordings from NCM, we used more medial coordinates than our previous experiments. With the birds beak fixed at a 55° angle to the vertical, electrodes were inserted roughly 1.2mm rostral and 0.5mm lateral to the Y-sinus. We made extracellular recordings from tungsten-parylene electrodes having impedance between 1 and 3 MΩ (A-M Systems). Electrodes were advanced in 0.5μm steps with a microdrive (Newport), and extracellular

voltages were recorded with a system from Tucker-Davis Technologies (TDT).

In all cases, the extracellular voltages were thresholded to collect candidate spikes. Each time the voltage crossed the threshold, the timestamp was saved along with a high-resolution waveform of the voltage around that time (0.29ms before and 0.86ms after for a total of 1.15ms). After the experiment, these waveforms were sorted using SpikePak (TDT) to assess unit quality. We sorted spike waveforms using a combination of PCA and waveform features (maximum and minimum voltage, maximum slope, area). We assessed clustering qualitatively and verified afterwards that the resulting units had Inter-Spike-Interval distributions where no more than 0.5% of the intervals were less than 1.5ms.

In each bird, we advanced the electrode in 50 μm steps until we found auditory responses. At that point we recorded activity in 100 μm steps. When we no longer found auditory responses, we moved the electrode 300 μm further, made an electrolytic lesion (2μA x 10s), advanced another 300 μm, and made a second identical lesion. These lesions were used to find the electrode track post-mortem and to calibrate the depth measurements. At the end of the recording session, the bird was euthanized with an overdose of Equithesin and transcardially perfused with 0.9% saline, followed by 3.7% formalin in 0.025M phosphate buffer. The skullcap was removed and the brain was post-fixed in 30% sucrose and 3.7% formalin to prepare it for histological procedures. The brain was sliced parasagittally in 40 μm thick sections using a freezing microtome. Alternating brain sections were stained with both cresyl violet and silver stain, which were then used to visualize electrode tracks, electrolytic lesions and brain regions.

All of our electrode tracks sampled NCM from dorsal to ventral regions. Some of the more dorsal recordings (shallower depths) could have been in subregions L or L2b of the Field L complex as the boundary between either of these two regions and NCM proper is difficult to establish (Vates et al. 1996; Fortune and Margoliash 1992). It is possible therefore that the correlation between degree of invariance and depth also reflects lower invariance observed in the field L complex and higher invariance in NCM proper.

Sound Stimuli

Stimuli consisted of zebra-finch songs, roughly 1.6-2.6 seconds in length, recorded from 40 unfamiliar adult male zebra finches played either in isolation or in combination with a background of synthetic noise (song+ml-noise stimuli in main text).

The masking noise in the neurophysiological experiments was synthetic and obtained by low-pass filtering white noise in the modulation domain following the procedure described in (T. M. Elliott and Theunissen 2009). This modulation low-pass filter had cutoff frequencies of $\omega_f = 1.0$ cycles/kHz and $\omega_t = 50$ Hz and gain roll off of 10dB/(cycle/kHz) and 10dB/10Hz. The cutoff modulation frequencies were chosen in order to generate noisy sounds with similar range of modulation frequencies found in environmental noise (Singh and Theunissen 2003). In addition, most of the modulations found in zebra finch song are well masked by this synthetic noise although it should be noted that song also includes sounds features with high spectral modulation frequencies (above 2 cycles/kHz) and high temporal modulation frequencies (above 60Hz). The frequency spectrum of the ml-noise was flat from 250 Hz to 8 kHz completely overlapping the entire range of the band-passed filtered songs we used in the experiments. Thus, although, different results could be found with noise stimuli with different statistics, we carefully designed our masking noise stimulus to both capture the modulation found in natural environmental noise while at the same time completely overlapping the frequency spectrum of our signal. The frequency power spectrum of these signals can be found in A. Hsu, Sarah M N Woolley, et al. 2004.

We have also shown that such ml-noise is an effective stimuli for midbrain and cortical avian auditory neurons in a sense that it drives neuron with high response rates and high information rates (A. Hsu, Sarah M N Woolley, et al. 2004). ML-noise is also very similar to the dynamic noise ripples described in (Escabí et al. 2003) and used in many neurophysiological studies to characterize high-level mammalian auditory neurons. We also recorded responses to the ml-noise masker alone but these data were not analyzed for this study.

All song and ml-noise stimuli were processed to be band limited between 250Hz and 8 kHz and to have equal loudness using custom code in MATLAB.

The sounds were presented using software and electronics from TDT. Stimuli were played over a speaker at 72dB C-weighted average SPL in a double-walled anechoic chamber (Acoustic Systems). The bird was positioned 20cm in front of the speaker for free-field binaural stimulation.

Each of the combined stimuli consisted of a different ml-noise sound sample, randomly paired with one of the songs. The noise stimulus began five to seven seconds after the previous stimulus, and the song began after a random delay of 0.5 to 1.5 seconds after the onset of the noise. Thus for each trial the same song is paired with a different noise sample and at a different delay. In the combined presentations, the noise stimuli were attenuated by 3dB to obtain a signal to noise ratio (SNR) of 3 dB.

We played four trials at each recording location, each consisting of a randomized sequence of 40 songs, 40 masking noise stimuli, and 40 combined stimuli. Stimuli were separated by a period of silence with a length uniformly and randomly distributed between five and seven seconds.

Neural Data Analysis

We used custom code written in MATLAB, Python and R for all of our analyses.

We assessed responsiveness using an average z-score metric for each stimulus class. The z-score is calculated as follows:

$$z = \frac{\mu_S - \mu_{BG}}{\sqrt{\sigma_S^2 + \sigma_{BG}^2 - 2covar(S, BG)}}$$

, where μ_S is the mean response during the stimulus, μ_{BG} is the mean response during the background, σ_S^2 is the variance of the response during the stimulus, and σ_{BG}^2 the variance of the response during baseline. The background rates were calculated using the 500ms periods preceding and following each stimulus. Using a cutoff of $z \geq 1.5$ for either ml-noise or song stimuli, 32 of the 50 single units were determined to be responsive.

To measure invariance, we evaluated the similarity between the responses to song and song + ml-noise by computing two measures: 1) the correlation coefficient between the PSTH for each corresponding response and 2) the

ratio of the SNR in the neural response to song+noise and the SNR in the response to song alone.

If the PSTH for song is called $r_s(t)$ and the PSTH obtained in response to song+noise is called $r_{s+n}(t)$, then the correlation coefficient is given by:

$$I_{CC} = \frac{\langle (r_s(t) - \bar{r}_s)(r_{s+n}(t) - \bar{r}_{s+n}) \rangle_t}{\sqrt{\langle (r_s(t) - \bar{r}_s)^2 \rangle_t \langle (r_{s+n}(t) - \bar{r}_{s+n})^2 \rangle_t}}$$

, where the $\langle\rangle$ are averages across time samples. We called this correlation coefficient, the correlation invariance or the invariance for short. The correlation coefficient is bounded between -1 and 1 and measures the linear similarity in the response after mean subtracted and scaling. Thus a response to song+noise with a deviation from its mean rate that is similar in shape but much smaller than the time-varying response to song alone will have a very high CC invariance. A better measure of invariance might therefore take into account both the mean PSTH rate as a proxy for noise and the deviations from this rate as a measure of signal. Thus, for the response to song alone, we define the signal power as $S_s = \langle (r_s(t) - \bar{r}_s)^2 \rangle$ and the noise power as $N_s = \bar{r}_s^2$ for a signal to noise ratio of:

$$SNR_s = \frac{\langle (r_s(t) - \bar{r}_s)^2 \rangle}{\bar{r}_s^2}$$

. For the response to the song+noise, we wanted to determine the fraction of the time varying-response that was related to the song. For that purpose, we used $r_s(t)$ as a regressor to obtain an estimate of $r_{s+n}(t)$:

$$\begin{aligned} \hat{r}_{s+n}(t) &= \beta_0 + \beta_1 r_s(t) \\ &= (\beta_0 + \beta_1 \bar{r}_s) + (\beta_1 r_s(t) - \bar{r}_s) \end{aligned}$$

, where β_0 and β_1 are the coefficients obtained from the normal solution for linear regression.

The signal to noise ratio for the response to song+noise is then:

$$SNR_{s+n} = \frac{\beta_1^2 \langle (r_s(t) - \bar{r}_s)^2 \rangle}{(\beta_0 + \beta_1 \bar{r}_s)^2}$$

. And the SNR invariance is given by the ratio of the two SNRs:

$$I_{SNR} = \frac{SNR_{s+n}}{SNR_s} = \frac{\beta_1^2 \bar{r}_s^2}{(\beta_0 + \beta_1 \bar{r}_s)^2}$$

As shown on fig. 2.1A, the two metrics ended up being highly correlated: the correlation coefficient between I_{CC} and the log of I_{SNR} is $r = 0.94 (p < 10^{-6})$ and we decided to use I_{CC} in the main text. However, the calculation of I_{SNR} also provides useful information in terms of the absolute magnitude of the invariance. For example, it shows that the SNR in the response for the seven most invariant cells is decreased by 5 to 10 dB when the song in presented in noise. Thus, even for these noise-robust neurons the loss of signal quality is present. Similarly, one can examine the value of the linear regression coefficient, β_1 on fig. 2.1B. This coefficient is always less than one showing that the responses to the song signal in the song+noise stimulus is always reduced. β_1 is also highly correlated with I_{CC} but always smaller. Together this shows that although the shape of the time-varying response is often very well preserved in noise-invariant neurons, that the magnitude of this response is decreased resulting in significant losses in signal power (informative time-varying firing rate) relative to noise power (mean firing rate).

In the calculations above, the PSTH was obtained by smoothing spike arrival times using a 31 ms Hanning window. The bias introduced by the small number of trials used to compute each PSTH was correcting by jackknifing. The single-stimulus results indicate a small but consistent negative bias in the four-trial estimates. We then computed the invariance as the mean of the individual bias-corrected correlations obtained for each 40 stimulus.

For each responsive single unit, we estimated the neurons STRF from their responses to song alone. The STRF were obtained using the *strfLab* neural data analysis suite developed in our laboratory (strflab.berkeley.edu). The STRFs were estimated by regularized linear regression. The algorithm is implemented as a Ridge Regression in *strfLab* (*directfit* training option). Because of the $1/f^2$ statistics of song, the ridge regression hyper parameter acts as a smoothing factor on the STRF. In addition, we used a sparseness hyper-parameter that controls the number of non-zero coefficients in the STRF.

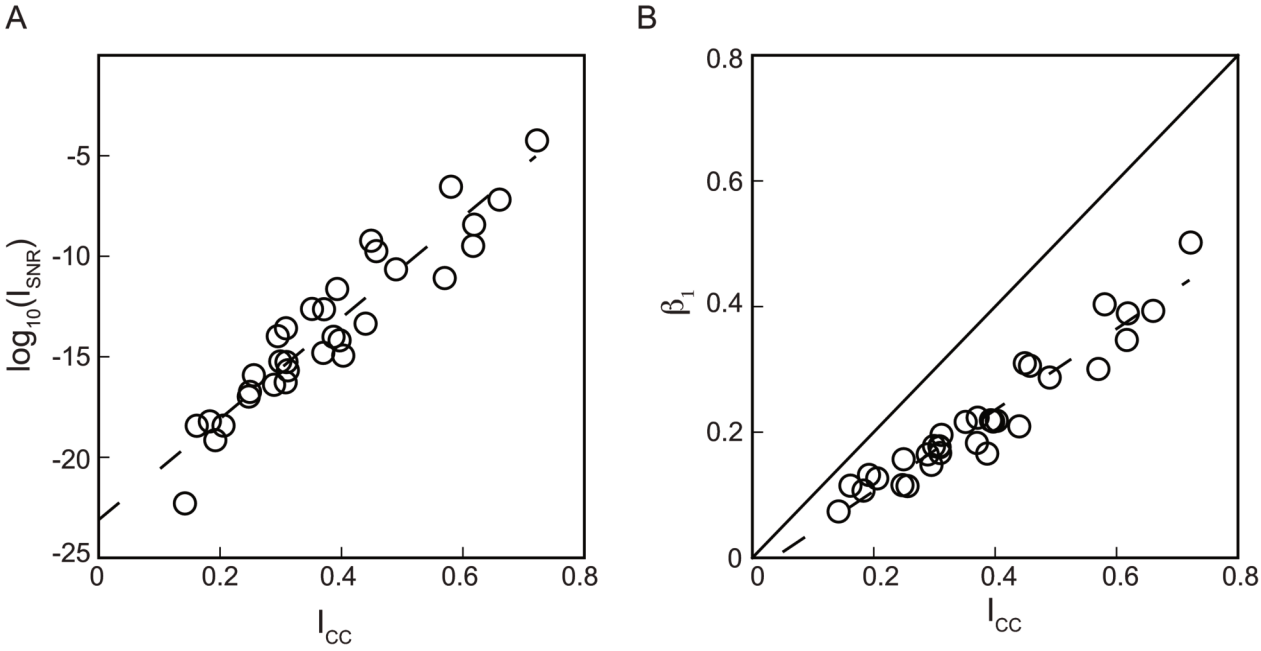

Figure 2.1: **Comparison of the Correlation Invariance (I_{CC}) and the SNR Invariance (I_{SNR}).** A. Scatter plot showing the strong correlation between the I_{SNR} (in dB units) and I_{CC} : $r = 0.96, p < 10^{-6}$. B. Scatter plot between the non-normalized linear regression coefficient, β_1, and the normalized measure of invariance, I_{CC}. These two measures are also highly correlated: $r = 0.96, p < 10^{-6}$.

Optimal values of the two hyperparameters were found by Jackknife cross-validation (see Theunissen, Stephen V David, et al. 2001; S. M. N. Woolley 2006, for more details). The stimulus representation used for the STRF was the log of the amplitude of the spectrogram of the sound obtained with a Gaussian shaped filter bank of 125 Hz wide frequency bands. Time delays of up to 100 ms were used to assess the cross-correlation between the stimulus and the response. Performance of the estimated final best STRF was then quantified with a separate validation data set.

We assessed the performance of each STRF using coherence and the normal mutual information as described in Borst and Theunissen 1999; A. Hsu, Borst, and Theunissen 2004. First, we compute the expected coherence between two single response trials; we then compute the coherence between the STRF prediction and the average response. The coherence is a function of frequency between zero and 1 that measures the correlation of two signals

at each frequency. To obtain a single measure of correlation, one can compute the normal mutual information (MI). We then computed the normal MI for the two coherences, calling the first the response information and the second the predicted information. The ratio of the predicted information to the response information is the performance ratio, and provides a measure of model performance that is independent of the variability of the neuron (A. Hsu, Sarah M N Woolley, et al. 2004). In all of our receptive field analyses, we used only STRFs that predict sufficiently well, defined here as having predicted information of at least 1.2 bits/second and a performance ratio of at least 20%. The STRF performance was not correlated with either the responsiveness of the neuron, as measured by their z-score, or the degree of invariance (data not shown).

To further examine the gain of the neuronal response as a function of temporal and spectral modulations, we also represented each STRF in terms of its Modulation Transfer Function (MTF). The MTF is obtained by taking the amplitude of 2 dimensional Fourier Transform of the STRF (Sarah M N Woolley, Fremouw, et al. 2005). For each neuron, we also computed the center of mass of its MTF to estimate its best spectral and temporal modulation frequencies.

To calculate the invariance metrics for the STRF model, we first obtained the predicted response to the song+ml-noise stimulus for each trial. Using these in place of the actual responses, we then computed an invariance metrics for the STRF model by comparing the predicted responses to the actual response obtained for song alone. In this manner, we were able to directly compare the STRF model invariance with the invariance calculated for the actual neuron. We used a two-tailed t-test to compare the distribution of similarity values for the 40, four-trial linear predictions to the 40 actual four-trial responses.

Figure 2.4 illustrates the methodology and shows the STRF, MTF, neural responses and predictions to both song and song+ml-noise for two additional example neurons: one with relatively low noise-invariance and one with relatively high noise-invariance.

Noise Filtering Algorithm Using the Modulation Filter Bank Model

Following directly from the premise that neurons in area NCM selectively respond to spectral-temporal modulations present in zebra finch songs, even in the presence of corrupting background noise, we developed a noise reduction scheme that would exploit this property. Our algorithm falls in the general class of single microphone noise reduction (SMNR) algorithms using spectral subtraction. The core idea in spectral subtraction is to estimate the frequency components of the signal from the short time Fourier components of the corrupted signal. The estimated signal frequency components are obtained by multiplying the Fourier components of signal+noise by a gain function. This is the synthesis part of the algorithm. The gain function can vary both in frequency and time. The form and estimation of the *optimal* gain function is the analysis step of the algorithm and its design is the principal focus of the novel development of the state-of-the art SMNR algorithms.

Both the analysis and synthesis step in our algorithm used a complete (amplitude and phase) time-frequency decomposition of the sound stimuli (fig. 2.7). This time-frequency decomposition was obtained from a frequency filter bank of N linearly-spaced band-pass filter Gaussian shaped channels located between 250 Hz and 8 kHz (BW=125Hz). N was set at 60 for all simulations. The amplitude of these N narrow-band signals could then be obtained using the Hilbert transform to generate a spectrogram of the sound. The analysis step in the algorithm involved generating an additional representation of the sounds based on an ensemble of M model neurons fully characterized by their STRF. The model STRFs were parameterized as the product of two Gabor functions describing the temporal and spectral response of the neuron:

$$STRF(t, f) = H(t) \cdot G(f)$$

, where

$$H(t) = A_t \exp^{-\frac{(t-t_0)^2}{2\sigma_t^2}} \cdot \cos(2\pi\Omega_t(t - t_0) + P_t)$$

and

$$G(f) = A_f \exp^{-\frac{(f-f_0)^2}{2\sigma_f^2}} \cdot \cos(s\pi\Omega_f(f - f_0) + P_f)$$

.

The parameters of these Gabor functions (e.g. for time: t_0, the temporal latency; σ_t, the temporal bandwidth; Ω_t, the best temporal modulation frequency; and P_t, the temporal phase) were randomly chosen using a uniform distribution over the range of those found in area NCM (present study) and Field L (Sarah M N Woolley, P. R. Gill, et al. 2009). The number of model neurons, M, was not found to be critical as long as the population of STRFs sufficiently tiled the relevant modulation space. M was set to be 140 for the results shown. To obtain the representation of sounds in this neural space, the log spectrogram of the stimuli was convolved by each STRF to obtain the model neural response: $\overrightarrow{\boldsymbol{a}}\boldsymbol{(t)}$ of dimension M. As explained in the main text, we then used these activation functions to obtain a set of optimal time varying frequency gains, $\overrightarrow{\boldsymbol{g}}\boldsymbol{(t)}$ of dimension N. These frequency gains are then be applied to the corresponding frequency slices in the time-frequency decomposition of the sound to synthesize the processed signal using:

$$\hat{s}(t) = \sum_{j=1}^{N} g_j(t) \cdot y_j(t)$$

, where $y_j(t)$ is the narrow-band signal from the frequency filter j obtained in the time-frequency decomposition of the song + noise stimulus, $x(t)$.

The optimal set of weights, d_i, needed to obtain the optimal gains, $\overrightarrow{\boldsymbol{g}}\boldsymbol{(t)}$ (see section 2.4) was learned by minimizing the squared error $e^2(t) = (s(t) - \hat{s}(t))^2$ through gradient descent. For this purpose, training stimuli were generated by summing together a 1.5 s song clip and a randomly selected chunk of either ml-noise or zebra finch colony noise of the same duration. To match the experimental results, both the song, $s(t)$, and the noise, $n(t)$, were first high-pass filtered above 250 Hz and low-pass filtered below 8 kHz, and then resampled to a sampling rate of 16 kHz. The song and noise were weighted to obtain a SNR of 3 dB, although similar results were found with lower SNR's. Training was performed on all instances of the signal + noise samples. Weights were determined by averaging across values obtained through

jack-knifing across this data set ten times with 10% of the data held out as an early stopping set. Noise reduction was then validated and quantified on a novel song in novel noise. Examples of noise corrupted signals and filtered signals that correspond to the spectrograms shown in fig. 2.8 can be found in the supplemental online material.

To assess the performance of our model, we computed the cross-correlation between the estimate and the clean signal in the log spectrogram domain. We then took the ratio of this cross-correlation and the value obtained prior to attempting to de-noise the stimulus to obtain a performance ratio. As summarized in the text, we then compared our algorithm to other noise reduction schemes. For this purpose, we also estimated the performance ratio for three other spectral subtraction noise algorithms: the optimal Wiener filter (OWF), a variable gain algorithm patented by Sonic Innovations (SINR) and the ideal binary mask (IBM). The optimal Wiener filter is a frequency filter whose static gain depends solely of the ratio of the power spectrum of the signal and signal + noise. In our implementation, the Wiener filter was constructed using the frequency power spectrum of signal and noise from the training set and then applied to a stimulus from the testing set (of the same class). The spectral subtraction algorithm for Sonic Innovations used a time variable gain just as in our implementation. Also, as in our implementation, the analysis step for estimating this gain was based on the log of the amplitude of the Fourier components. However, the gain function itself was estimated not from a modulation filter bank but estimating the statistical properties of the envelope of the signal and noise in each frequency band (US Patent 6,757,395 B1). We used a MATLAB implementation of the SINR algorithm provided to us by Dr. William Woods of Starkey Hearing Research Center, Berkeley, CA. Optimal parameters for the level of noise reduction and the estimation of the noise envelope for that algorithm were also obtained on the training signal and noise stimuli and the performance was cross-validated with the test stimuli. The IBM procedure used a zero-one mask applied to the sounds in the spectrogram domain. The mask is adapted to specific signals by setting an amplitude threshold. Binary masks require prior knowledge of the desired signal and thus should be seen as an approximate upper bound on the potential performance of general noise reduction algorithms. Although

these simulations are far from comprehensive, they allowed us to compare our algorithm to optimal classical approaches for Gaussian distributed signals (OWF), to a very recent state-of-the-art algorithm (SINR) and to an upper bound (IBM). For commercial applications, our noise-reduction algorithm is available for licensing via UC Berkeleys Office of Technology Licensing (Technology: Modulation-Domain Speech Filtering For Noise Reduction; Tech ID: 22197; Lead Case: 2012-034-0).

2.4 Results/Discussion

We recorded neural responses from single neurons in NCM of anesthetized adult male Zebra Finches. We obtained responses to 40 different unfamiliar conspecific songs and to the same songs embedded in naturalistic synthetic noise also called modulation-limited noise (ml-noise from here on). Ml-noise is broadband white-noise that has been filtered in the modulation domain to mimic the structure that is found in environmental sounds by restricting the power of modulations in the envelope to low spectral-temporal frequencies (Singh and Theunissen 2003). Ml-noise has also been shown to be an efficient stimulus for driving high-level auditory neurons (see section 2.3 for additional details). The signal to noise ratio (SNR) was set at 3dB.

Noise Invariant Neurons in NCM

As illustrated on the left panels in fig. 2.2, responses of some neurons to song signal were almost completely masked by the addition of noise. In these situations, the post-stimulus time histogram (PSTH) obtained for song only (third row) is very different than the one obtained for song + ml-noise (fifth row). However, some neurons also showed strong robustness to noise degradation as illustrated on the right panels of fig. 2.2. Those neurons had similar PSTHs for both conditions.

To quantify the degree of noise robustness, we calculated two measures of noise-invariance: a de-biased correlation coefficient between the PSTHs obtained for the song alone and song + ml-noise stimuli (called I_{CC}) and the ratio of the SNR estimated for the song + noise response and the song +

ml-noise response (I_{SNR} invariance). The I_{CC} metric is a normalized measure that ranges in values between -1 and 1. It is 1 when the response pattern observed to song+ml-noise is identical to the one observed to song, irrespective of the relative magnitude of the two responses. For I_{SNR}, we defined the response SNR as follows. For the response to song alone, the signal power was defined as the variance in the PSTH across time and the noise was defined as the mean firing rate. For the response to song plus noise, the signal was taken to be the time-varying response that could be predicted linearly from the response to song alone and the noise was the mean of this predicted response (see section 2.3). This second value of invariance is bounded between 0 and 1 and captures not only the similarities in response patterns but also magnitudes of time-varying responses that carry information about the song. As shown in the supplemental material, the two measures were highly correlated and subsequent analyses resulted in very similar results and identical conclusions. For brevity, we show the analysis using the I_{CC} metric in the main paper. Some of the results with the I_{SNR} metric are included in the supplemental material.

Noise Invariance and Frequency Tuning

We found neurons with different degrees of noise invariance throughout NCM but the neurons in the ventral region tended to have highest I_{CC} (fig. 2.3B). NCM also exhibits some degree of frequency tonotopy along this dimension with higher frequency tuning found in more ventral regions (Ribeiro et al. 1998; Terleph, Mello, and Vicario 2006). Indeed, in our data set, we also found a strong correlation between dorsal/ventral position and the best frequency (BF) of the neuron (fig. 2.3C). We estimated a neurons best frequency from the peak of the frequency marginal of its spectral-temporal receptive field (STRF). We found a range of BF from 1300 Hz to 3300 Hz with a dorsal-ventral gradient (adjusted $R^2 = 0.34, p < 10^{-3}$). Although the frequency range of our song stimulus and ml-noise stimulus was identical, the frequency power spectrum of song has a peak around 4 kHz (A. Hsu, Sarah M N Woolley, et al. 2004) that could have lead to stronger and thus potentially more noise invariant responses to song for neurons with higher best fre-

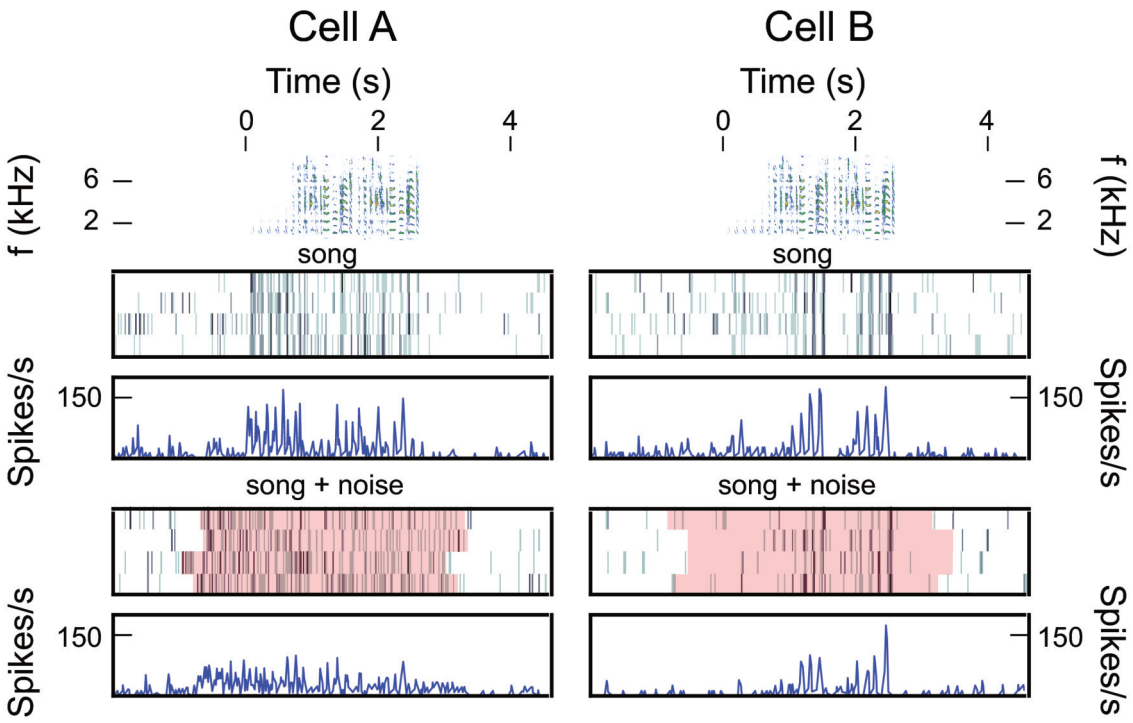

Figure 2.2: **Noise-invariant responses in the avian NCM.** Responses of two neurons (Cell A and Cell B) to song presented alone and over noise. The top row shows the spectrogram of the same zebra finch song used in the two recordings. Song starts at 0s. Below the spectrogram are raster plots and corresponding smoothed PSTHs. The first raster and PSTH correspond to the response of each neuron to the song alone presented at 70 dB SPL. Clear temporal synchrony across the four trials can be seen illustrative of an equally robust response to song stimuli. The second raster and PSTH correspond to the responses to song+ modulation limited noise (ml-noise) presented at 3dB signal to noise ratio. Ml-noise is synthesized by low-pass filtering white noise in the space of temporal and spectral modulations (see section 2.3). The pink highlights show the duration of the stimulus (song + noise). The onset and offset of the stimulus is different in each trial because the trials are aligned to the onset of the song and the noise masker began and ended with a different delay in each trial. The noise was also different in each trial. This addition of naturalistic noise destroys the cross- trial synchrony in the response for the neuron shown in the left column but not for the neuron shown in the right column.

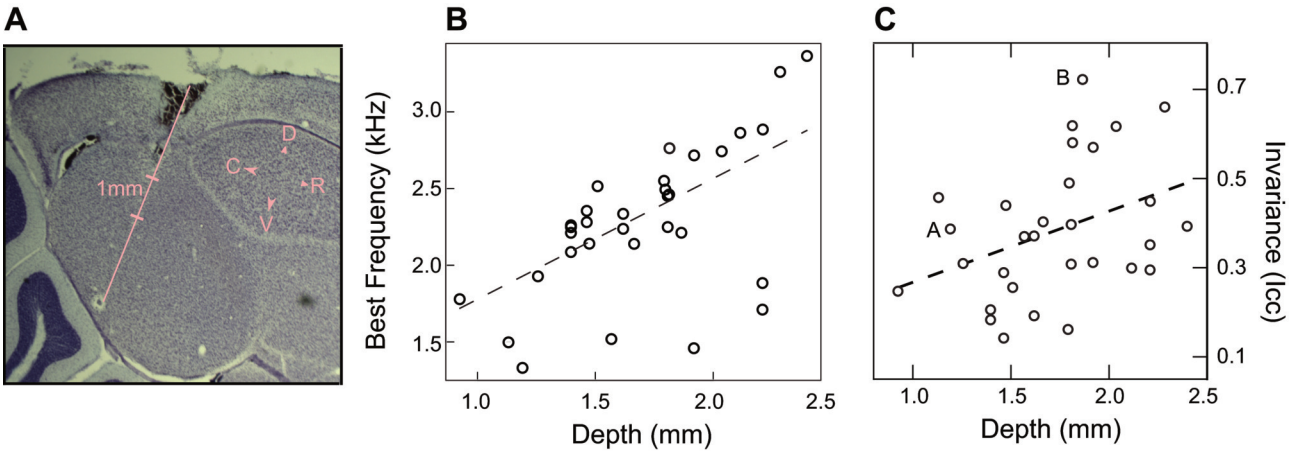

Figure 2.3: **Location of noise invariant neurons in NCM.** A. Photomicrograph of Nissl-stained brain slice in one bird showing the typical trajectory of the electrode penetration. By carefully orienting our electrode angle, we were able to sample NCM along its entire dorsal to ventral extent. B. Scatter plot of noise invariance against stereotactic depth of neural recordings. Noise invariance and recording depth were significantly correlated (slope $= 0.15$/mm, adjusted $R^2 = 0.13, p = 0.02$). The example neurons are labeled A and B on the scatter plot. C. Scatter plot showing the relationship between the best frequency (Y-axis) and the depth of the recording along the dorsal to ventral axis of NCM (X-axis). The solid line is the linear regression between these two variables (adjusted $R^2 = 0.34, p < 10^{-3}$).

quency. A linear regression analysis between invariance and the neurons best frequency could not confirm that hypothesis (adjusted $R^2 = 0.06, p = 0.1$). Thus, if this relationship exists, it can only have a very small effect size.

Noise Invariance and Spectral-temporal Tuning.

To further attempt to understand how noise invariance was achieved in this system, we examined how the neurons responses for particular joint spectral-temporal patterns that are unique to song could have contributed to robust coding of song in noisy conditions. To do so we estimated the STRF of each neuron and examined the predicted response to song and to song plus noise. The STRF describes how acoustical patterns in time and frequency correlate with the neurons response (Theunissen, Stephen V David, et al. 2001; Sarah M N Woolley, P. R. Gill, et al. 2009). The STRF can also be used as a model of the neuron to estimate predicted responses for arbitrary sound stimuli. The STRF model is often described as linear but can include both input

and output non-linearities. In this study, the stimulus was represented as a log spectrogram and the output of the linear filter was half-wave rectified (see section 2.3). Although we have shown that better response predictions could be obtained using additional non-linear elements such as gain control (P. Gill, Zhang, et al. 2006), in this study we have used the simpler STRF model to more explicitly describe the spectral-temporal tuning of each neuron (examples of STRF predictions are found in fig. 2.4).

To determine whether a neurons tuning for particular spectral-temporal features characteristic of song and less common in noise could explain the observed invariance, we use the STRF to obtain estimated responses to song and noise. We then regressed the I_{CC} values that we measured directly from the neurons response against the I_{CC} values obtained from the predictions of STRF model (fig. 2.5A). Two results come out of this analysis. First, the measured invariance and the model invariance are positively but weakly correlated showing that the neurons STRFs can in part explain the observed noise-invariance (Adjusted $R^2 = 0.12, p = 0.034$). Second, we found that, for most neurons, the degree of invariance predicted by the STRF model was greater than the one found in actual neurons. In other words, non-linearities not captured in the STRF model made these neurons less invariant. Although this result might seem surprising for an auditory region believed to be important for song recognition, it has a simple explanation. Many high-level neurons show adapting responses to sound intensity levels (I. Dean, Harper, and McAlpine 2005) and this common non-linear response property is not captured in this STRF model. Intensity adapting neurons would exhibit a decrease in response to the song in noise relative to the song alone due to the adaptive changes in gain. This decrease in response gain without a corresponding decrease in background rate would result in a decrease of the responses SNR.

Therefore, for the task of extracting the song from noise, the most effective non-linearities appear to be the simple thresholding non-linearity (i.e. for neurons with STRFs closest to the x=y line in fig. 2.5A) or a yet to be described additional non-linearity boosts invariance (n=3/32). Although the specific non-linearities that could be beneficial for preserving signal in noise still need to be described, previous research have characterized higher-order

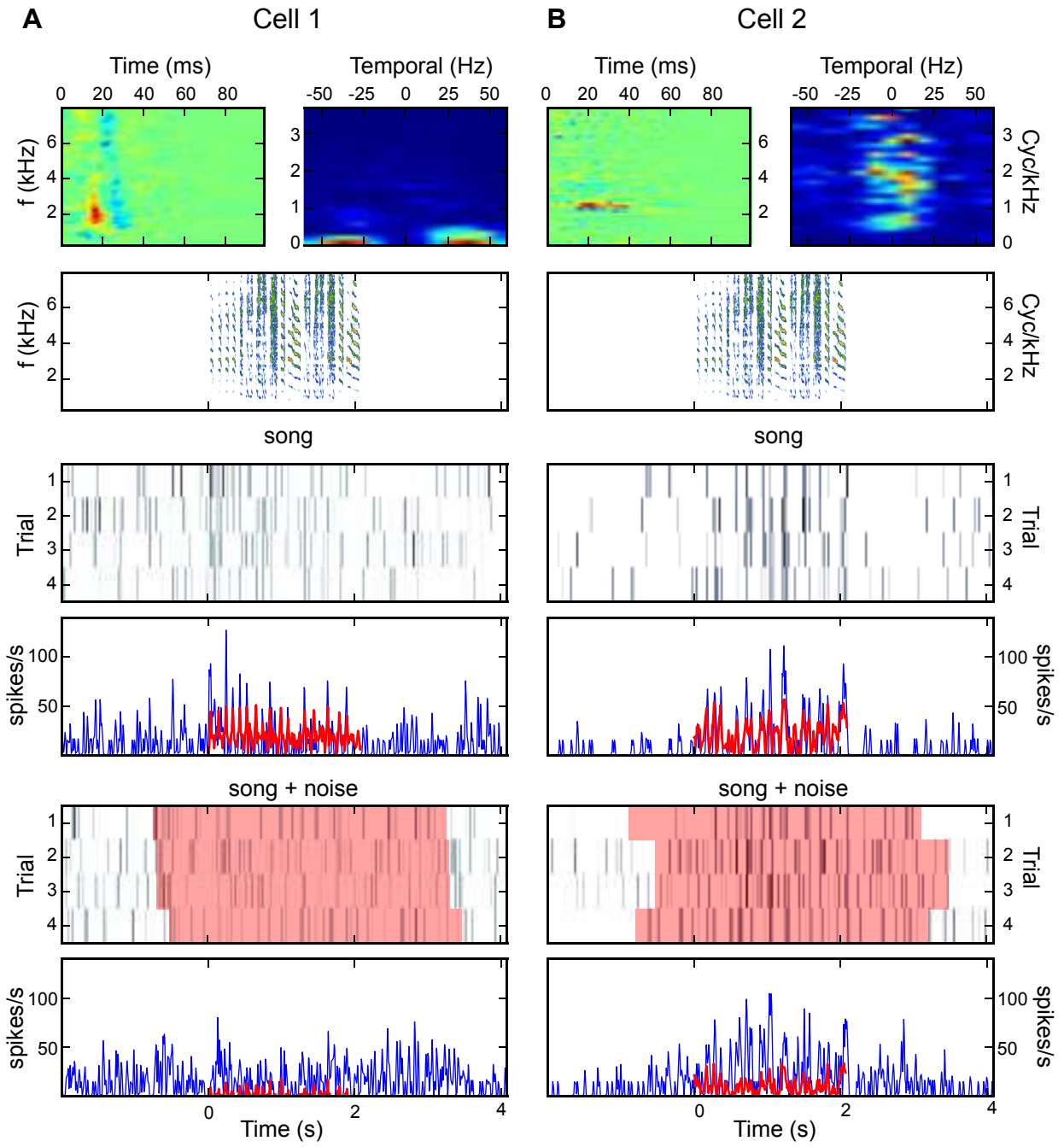

Figure 2.4: **Example of STRF and STRF Predictions for Two Cells. Column A:** A low noise-invariant cell (invariance = 0.25), Cell 1 and **Column B:** a high noise-invariant cell (invariance = 0.65), Cell 2. Top row shows the STRF and the corresponding MTF. Second row shows the spectrogram of one song stimulus. Third and fourth row show the neural responses as a spike raster (top) and a PSTH (below) to the song presented alone. In the PSTH plot, the actual neural response is in blue and the prediction obtained from the STRF is in red. Spike raster for response of low noise-invariant cell to masked song. The fifth and sixth row show the responses to the song presented over a masker of ml-noise.

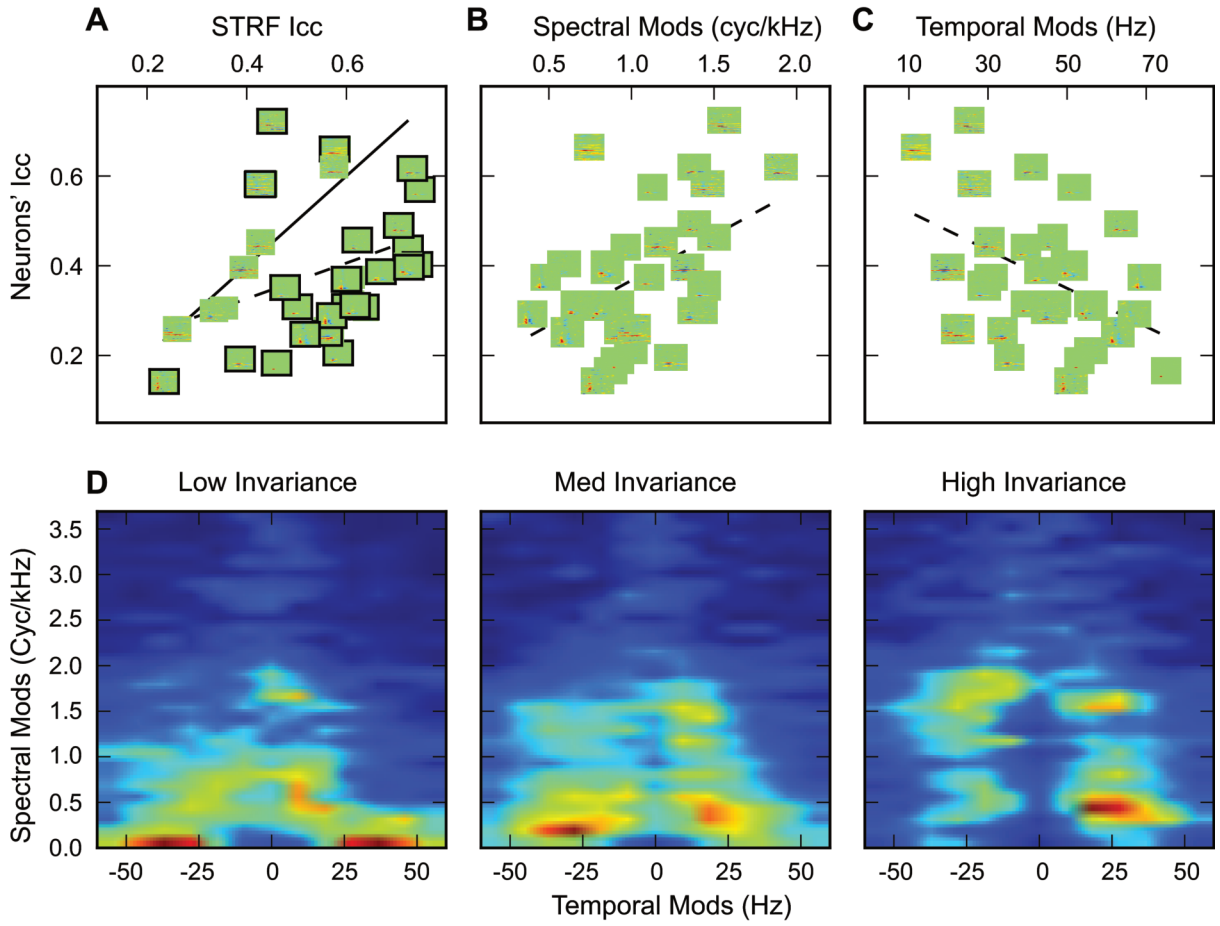

Figure 2.5: **Spectral-temporal tuning and invariance.** Vertical axis in AC shows the noise invariance in the neural response. Each neuron (each point on the scatterplots) is represented by its STRF (0.25 - 8 kHz on the vertical axis, 0 - 60 ms on the horizontal). A. Invariance vs STRF Model Invariance. The solid line has slope 1.0, showing equal performance between the STRF model and the neural response. Neurons with significantly different performance ($p < 0.05$, two-tailed t-test) have their receptive fields outlined. Dashed line shows regression fit (slope = 0.40, Adjusted $R^2 = 0.12, p = 0.034$), indicating the positive correlation between the invariance predicted by the STRF and actual invariance. B. Invariance vs Spectral Modulation Tuning. Neurons sensitive to higher spectral modulations are more invariant (Adjusted $R^2 = 0.192, p = 0.007$). C. Invariance vs Temporal Modulation Tuning. Neurons sensitive to lower temporal modulations are more invariant (Adjusted $R^2 = 0.15, p = 0.015$). D. Ensemble modulation transfer functions for neurons grouped by invariance. Low invariance neurons (left panel, *invariance* < 0.3, n = 11) respond to high temporal and low spectral frequency modulations. Neurons with moderate invariance (middle panel, $0.3 <$ *invariance* < 0.4, n = 11) transmit faster, sharper modulations. Neurons with high invariance (right panel, *invariance* > 0.4, n = 10) respond mostly to slower and spectrally sharp sounds.

non-linearities response that could play an important role: neurons in NCM exhibit stimulus specific adaptation (Stripling, Volman, and D. F. Clayton 1997) and neurons in another avian secondary auditory area, CM (Caudal Mesopallium), respond preferentially to surprising stimuli (P. Gill, Sarah M N Woolley, et al. 2008). These non-linearities could facilitate noise invariant responses since they tend to de-emphasize the current or expected stimulus (in this case noise like sounds) without decreasing the gain of the neuron to sound at the same frequency.

Since the STRF could partially explain the observed noise-invariance, we asked what feature of the neurons spectral-temporal tuning was important for this computation. By estimating the modulation gain from the neurons STRFs, we found that tuning for high spectral modulations and low temporal modulations correlate with neural invariance (fig. 2.5B-C). Neurons sensitive to higher spectral modulations are more invariant (Adjusted $R^2 = 0.192, p = 0.007$) and neurons sensitive to lower temporal modulations are more invariant (Adjusted $R^2 = 0.15, p = 0.015$). To assess the effect size of these two relationships taken together, we used multiple linear-regression with spectral and temporal modulation tuning as regressors used to explain the neurons invariance and found an adjusted R^2 of 0.23 (p=0.009). Thus the contributions of spectral and temporal tuning to invariance are not completely additive. The ensemble modulation transfer functions further illustrate how the spectral and temporal modulation tuning co-vary along the noise-invariance dimension (fig. 2.5D). Noise invariant neurons exhibit the combination of longer integration times and sharp spectral tuning. In addition, the sharp excitatory spectral tuning was often combined with sharp inhibitory spectral tuning as well. These properties make noise-invariant neurons particularly sensitive to the longer harmonic stacks present in song (and other communication signals) even when these are embedded in noise as illustrated in the example neuron in fig. 2.2 (right panel).

The generation of the observed modulation tuning properties of the more noise invariant neurons described in this study is not a trivial task: most neurons in lower auditory areas have much shorter integration times and lack the sharp excitation and inhibition along the spectral dimension that we observed here. From comprehensive surveys of tuning properties in the

avian primary auditory cortex (Field L) (Sarah M N Woolley, P. R. Gill, et al. 2009; Nagel and A. J. Doupe 2008), we know that a small number of neurons with similar characteristics exist in these pre-synaptic areas (Sarah M N Woolley, P. R. Gill, et al. 2009). Similarly, in the mammalian system, neurons in A1 have been shown to have a range of spectral-temporal tuning similar to that seen in birds but few with the sharp spectral tuning seen here (Depireux et al. 2001; L. M. Miller et al. 2002). Thus it is reasonable to postulate that noise-invariance in NCM (and putatively in mammalian secondary auditory cortical regions) is the result of a series of computations that are occurring along the auditory processing stream. However, it is also known that NCM possesses a complex network of inhibitory neurons and that these play an important role in shaping spectral and temporal response properties (Pinaud et al. 2008). We also found a higher concentration of noise invariant neurons in the more ventral regions of NCM but failed to find a correlation between invariance and best frequency. On the other, we found that both temporal modulation tuning (adjusted $R^2 = 0.12, p = 0.02$) and spectral modulation tuning (adjusted $R^2 = 0.15, p = 0.01$) were also correlated with depth: lower temporal and higher spectral modulation tuning is found in ventral regions of NCM. This organization of tuning properties is reminiscent of the organization of the primary auditory areas, field L, where the output layers have a higher concentration of neurons with longer integration times (Kim and A. Doupe 2011). Thus both upstream and local circuitry are almost certainly involved in the creation of noise-invariant neural representations.

Invariance and Song Selectivity

Since the tuning of noise invariant neurons described by their STRF and the threshold non-linearity only describes a fraction of the invariance, we were interested in assessing whether noise invariant neurons were selective for longer sound segments such as those that might be useful to distinguish one song from another. To begin to investigate this idea, we examined the invariance of all the neurons for each song and calculated the standard deviation and the coefficient of variation (CV) of the invariance metric for each

neuron. These results are shown as a two-dimensional heat map on fig. 2.6. Although, the degree of invariance varied somewhat across songs (and the most invariant neurons could have invariances above 0.9 for certain sounds), the variability was remarkably low: highly invariant neurons tended to show noise- invariance to most song stimuli. The CVs for the 10 most invariant neurons were similar and all below 0.5. We therefore conclude that neurons that show a high degree of invariance could be useful to extract signal from noise not only for a specific song but also for an entire stimulus class. For example, noise invariant neurons could detect short acoustical features that are characteristic of many zebra finch songs. The STRF analysis shows that sensitivity to features up to 100 ms in duration is more than sufficient to generate in model neurons noise invariance of similar magnitude to that observed in the actual data. However, since the STRF only explains a fraction of both the observed invariance and the response, selective response properties that involve longer integration times could also be involved in the generation of noise invariant responses.

Biologically Inspired Noise Reduction Algorithm

Inspired by our discovery of noise invariant neurons in NCM, we engineered a noise filtering algorithm based on a decomposition of the sound by an ensemble of artificial neurons described by realistic STRFs. We developed this algorithm both for biological and engineering purposes. Our biological goal was to demonstrate that an ensemble of noise-invariant responses such as the one observed here could indeed be used to recover a signal from noise. We also wanted to show whether an optimization process designed to extract signals from noise would rely on responses of particular artificial neurons with properties that are similar to those found in the biology. Finally, we also wanted to explore to what extent is the invariance of signal in noise dependent on the exact statistics of the signal and noise stimuli. Our engineering goal was to develop a real-time algorithm inspired by the biology that could potentially be used in clinical applications such as hearing aids and cochlear implants or in commercial applications involving automatic speech recognition (Hermus 2007). In hearing aids, various forms of noise reduction have

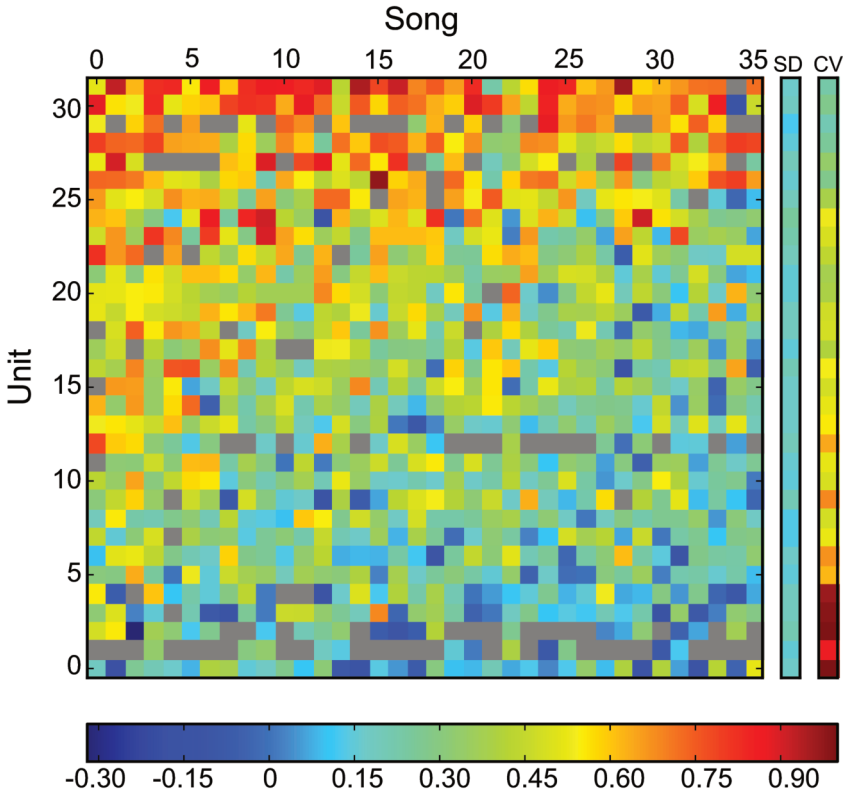

Figure 2.6: **Range of invariance observed across neurons and song stimuli.** Two dimensional heat plot that shows the value of the variance metric obtained for each neuron (n = 32) and each song stimuli (n = 36). The neurons are sorted from low mean invariance (bottom row) to high mean invariance (top row). The columns on the left show the standard deviation of the variance and the coefficient of variation for each neuron. The color bar is placed at the bottom of the graph and is the same for the variance, the standard deviation and the coefficient of variation. The grey cells in the matrix correspond to (neuron, stimulus) where we were not able to calculate the invariance either because of missing data or very low response rates.

been shown to offer an incremental improvement in the listening experience (Luts et al. 2010; DiGiovanni, Davlin, and Nagaraj 2011) though listening to speech in noisy environments remains the principal complaint of hearing aid users (Palmer 2009). In addition, none of the current noise reduction algorithms have led to improvements in speech intelligibility (Alcántara et al. 2003; Bentler et al. 2008).

Our ensemble of artificial neurons can be thought of as a modulation filter bank because the response of each neuron quantifies the presence and absence of particular spectral-temporal patterns as observed in a spectrogram and, contrary to a frequency filter bank, not solely the presence or absence of energy at a particular frequency band. In other words, the STRFs can be thought of as higher-level sound filters: if lower-level sound filters operate in the frequency domain (for example removing low frequency noise such as the hum of airplane engines), these high-level filters operate in the spectral-temporal modulation domain. In this joint modulation domain, sounds that have structure in time (such as beats) or structure in frequency (such as in a musical note composed of a fundamental tone and its harmonically related overtones) are characterized by specific temporal and spectral modulations. A spectral-temporal modulation filter could then be used to detect sounds that contain particular time-frequency patterns while filtering out other sounds that might have similar frequency content but lack this spectral-temporal structure. Similar decompositions have also been proposed and used by others for the efficient processing of speech and other complex signals (Mesgarani, Slaney, and Shihab A. Shamma 2006; David J. Klein, König, and Kording 2003; Chi, Ru, and Shihab A Shamma 2005).

Noise filtering with such a modulation filter bank can be described as series of signal processing steps: i) decompose the signal into frequency channels using a frequency filter bank; ii) represent the sound as the envelope in each of the frequency channels, as it is done in a spectrogram; iii) filter this time-frequency amplitude representation by a modulation filter bank to effectively obtain a filtered spectrogram; iv) invert this filtered spectrogram to recover the desired signal. Although each of these steps involves relatively simple signal processing, two significant issues remain. First, one has to choose the appropriate gain on the modulation filters in order to detect behaviorally

relevant signals over noise. Second, the spectrogram inversion step requires a computationally intensive iterative procedure (Griffin and Lim 1984) that would prevent such a modulation filtering procedure to operate in real time or with minimal delays. Our algorithm solves these two issues. We have eliminated the spectrographic inversion step and instead use the output of the modulation filter bank to generate a time-varying gain vector that can directly operate on the output of the initial frequency filter bank. Second, we propose to find optimal fixed gains on the modulation filter bank by minimizing the error between a desired signal and the output of the filtering process in the time domain. Then once the modulation filter weights are fixed, the algorithm can operate in real-time with a delay that is only dependent on the width of the STRF in the modulation filter bank.

The various steps in our algorithm are illustrated on fig. 2.7. Both the analysis and synthesis steps of the algorithm use a complete (amplitude and phase) time-frequency decomposition of the sound stimuli. This time-frequency decomposition is obtained from a frequency filter bank of N linearly-spaced band-pass filter Gaussian shaped channels located between 250 Hz and 8 kHz. The amplitude of these N narrow-band signals is obtained using the Hilbert transform (or rectification and low-pass filtering) to generate a spectrogram of the sound. This spectrographic transformation is identical to the one that we use for the estimation of the STRFs (see section 2.3).

The analysis step in the algorithm involves generating an additional representation of the sounds based on an ensemble of model neurons fully characterized by their STRF. These STRFs are designed to efficiently encode the structure of the signal and the noise, allowing them to be useful indicators of the time-course of signal in a noisy sound. For this study, we used a bank of STRFs that were designed to model the STRFs found throughout the auditory pallium, including STRFs not only from neurons in NCM but also the field L complex (Sarah M N Woolley, P. R. Gill, et al. 2009). The log spectrogram of the stimulus is convolved with each STRF to obtain model neural responses: $\overrightarrow{\boldsymbol{a}}(\boldsymbol{t})$ of dimension M. The crux of our algorithm is to transform these neural responses back into a set of time varying frequency gains, $\overrightarrow{\boldsymbol{g}}(\boldsymbol{t})$ of dimension N. These frequency gains will then be applied to the corresponding frequency slices in the time-frequency decomposition of the

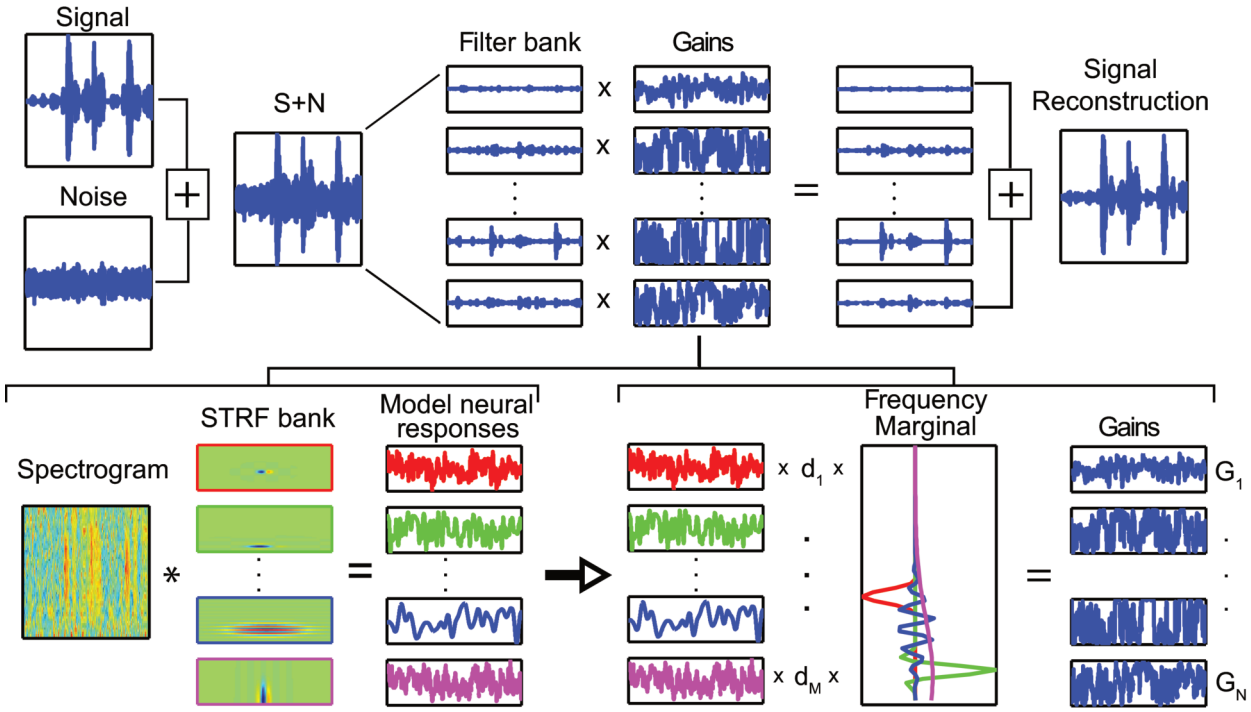

Figure 2.7: **Noise Reduction Algorithm.** We implemented a biologically inspired noise-filtering algorithm using an analysis/synthesis paradigm (top row) where the synthesis step is based on a STRF filter bank decomposition. The bottom row shows the model neural responses obtained from a sound (spectrogram of noise-corrupted song) using the filter bank of biologically realistic STRFs. These responses are then weighed optimally with weights $d_1, \ldots, d_M$ to select the combination of responses that are most noise-invariant. The weighted responses are then transformed into frequency space by multiplying the weighted responses by the frequency marginal of the corresponding STRF (color-matched on the figure) to obtain gains as a function of frequency. The top row illustrates how these time-varying frequency gains can then be applied to a decomposition of the sound into frequency channels allowing for the synthesis step and an estimate of the clean signal. This technology is available for licensing via UC Berkeleys Office of Technology Licensing (Technology: Modulation-Domain Speech Filtering For Noise Reduction; Tech ID: 22197; Lead Case: 2012-034-0)

sound to synthesize the processed signal. $\overrightarrow{\boldsymbol{g}}(\boldsymbol{t})$ is a function of the sum of all model neural responses each scaled by an importance weighting, d_i, and then multiplied by the frequency marginal of the corresponding neuron's STRF:

$$g_j(t) = f\left(\sum_{i=1}^{M} d_i \cdot a_i(t) \cdot K_{i,j}\right)$$

with $j \in \{1, N\}$. The function f was chosen to be the logistic function in order to restrict the gains to lie between a lower bound, representing maximal attenuation, and 0 dB, representing no attenuation. $K_{i,j}$ is the frequency marginal value of neuron i for the frequency band centered at j, and it was obtained from the frequency marginal of each STRF. Using these gains, we then synthesized a processed signal:

$$\hat{s}(t) = \sum_{j=1}^{N} g_j(t) \cdot y_j(t)$$

, where $y_j(t)$ is the narrow-band signal from from the frequency filter j obtained in the time-frequency decomposition of the song + noise stimulus, $x(t)$. The optimal set of weights, d_i, was learned by minimizing the squared error, $e^2(t) = (s(t) - \hat{s}(t))^2$ through gradient descent.

To assess the quality of our algorithm, we compared it to 3 other noise reduction schemes: the optimal classical frequency Wiener filter for stationary Gaussian signals (OWF), a state-of-the-art spectral subtraction algorithm (SINR) used by a hearing aid company, and the upper bound obtained by an ideal binary mask (IBM). The optimal Wiener filter is a frequency filter whose static gain depends solely on the ratio of the power spectrum of the signal and signal + noise. The state-of-the-art spectral subtraction algorithm uses a time variable gain just as in our algorithm but based on a running estimate of noise and signal spectrum. This algorithm was patented by Sonic Innovations (US Patent 6,757,395 B1) and is currently used in hearing aids. The IBM procedure used a zero-one mask applied to the sounds in the spectrogram domain. The mask is adapted to specific signals by setting an amplitude threshold. Ideal binary masks require prior knowledge of the desired signal and thus can be considered as an approximate upper bound on the potential

performance of general noise reduction algorithms (Y. Li and D. L. Wang 2009).

As shown on fig. 2.8A, with relatively little customization and exploration (for example in the choice of the set of artificial STRFs) our algorithm performed strikingly well: our algorithm performed significantly better than both the classical frequency Wiener filter and the SINR algorithm for a song embedded in ml-noise and similarly to the SINR algorithm for a song embedded in colony noise. The quality of the noise filtering can also be assessed by examining the time-varying gains shown on bottom row in fig. 2.8B and C: without any *a priori* knowledge of the location of the signal in time (and contrary to the IBM), the time-varying gains can pick out when the signal occurs in the noise. Moreover, the gains are not constant for all frequencies but instead are also able to pick out harmonic structure in the sound. The quality of the reconstruction can also be visually assessed by examining the spectrograms shown in that figure or listening to the demos provided as supplemental material.

We are now able to answer our questions. First, as quantified above, using an ensemble of physiologically realistic noise-invariant responses, we show that one is able to recover the distorted signal with remarkable accuracy. Second, we were also able to compare the properties of the STRFs in the model that had the biggest importance gains (d_i) with those found in noise-invariant neurons in NCM. As shown on fig. 2.9A & B, these STRFs are composed both of narrow band neurons with long integration times as observed in our data set and also broad band neurons with very short integration time. The eMTF shown in fig. 2.9C & D further quantify these results. Thus, the noise invariant neurons found in NCM are well represented in by the model STRFs tuned for high spectral modulation and low temporal modulations. NCM also has neurons tuned to faster temporal modulations but the majority of these neurons had narrow band frequency tuning (or high spectral modulations) and these neurons are therefore not particularly effective at rejecting noise stimuli. Fast broad-band neurons are however found in the avian primary auditory forebrain (Sarah M N Woolley, P. R. Gill, et al. 2009; Nagel and A. J. Doupe 2008) and could thus play a role, as part of an ensemble, in the signal and noise separation. Our third question regarded the sensitivity of

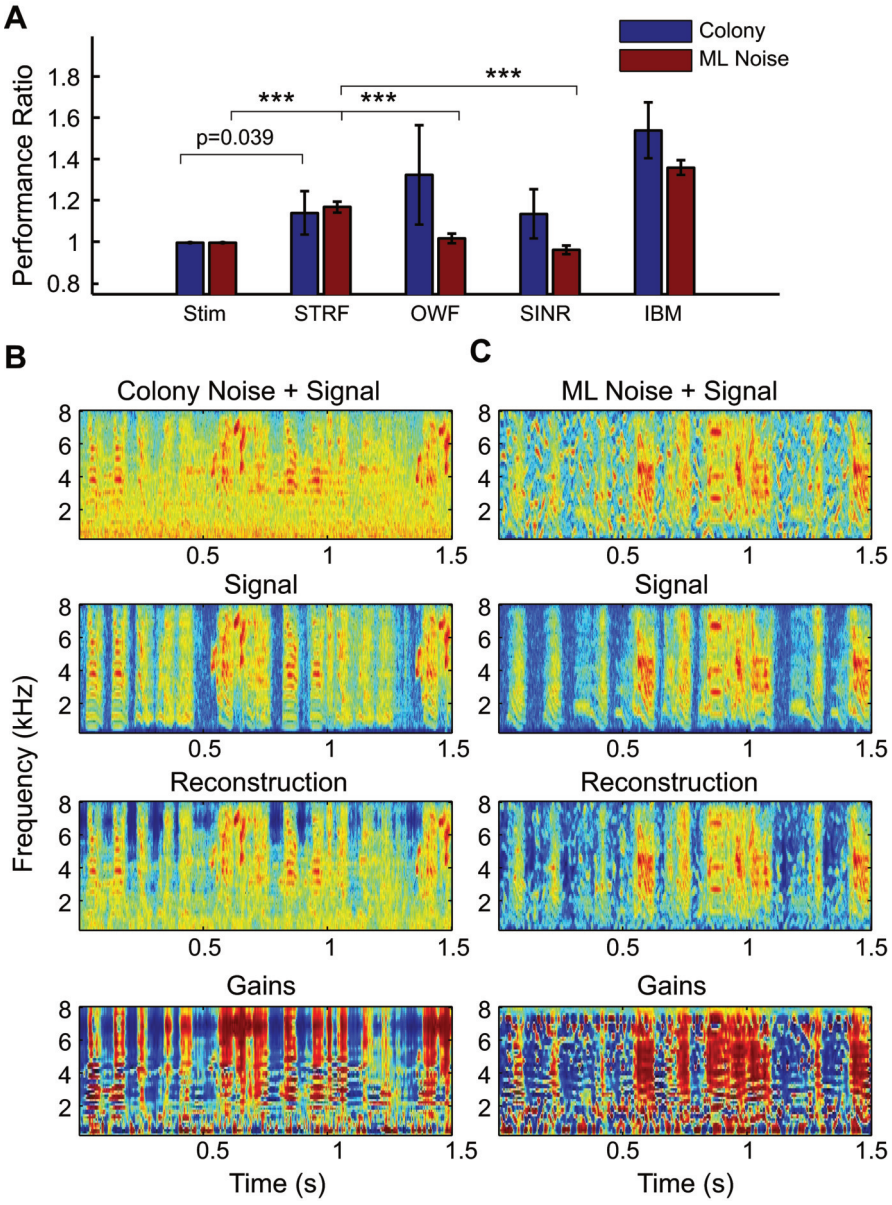

Figure 2.8: **Performance of STRF Based Noise-Reduction.** A. Performance of three noise reduction algorithms (STRF, OWF, SINR) and lower and upper bounds (Stim, IBM) on song embedded in colony noise or modulation-limited (ML) noise. The performance ratio (y-axis) depicts the improvement in noise levels over the noise-corrupted signal, as measured by the cross-correlation in the log spectrogram domain, with the error bars representing one standard deviation across five noisy stimuli. On the x-axis are the models we have tested, where Stim is the noise-corrupted signal, STRF is the model presented here, OWF is the optimal Wiener filter, SINR is a spectral subtraction algorithm similar to STRF but based on engineering constructs, and IBM is an ideal binary mask. B. Spectrograms of the signal masked with noise from the zebra finch colony, the clean zebra finch song, and our signal reconstruction, followed by the time-frequency gains. C. Same as B but for modulation-limited noise. Sounds can be found in the supplemental materials.

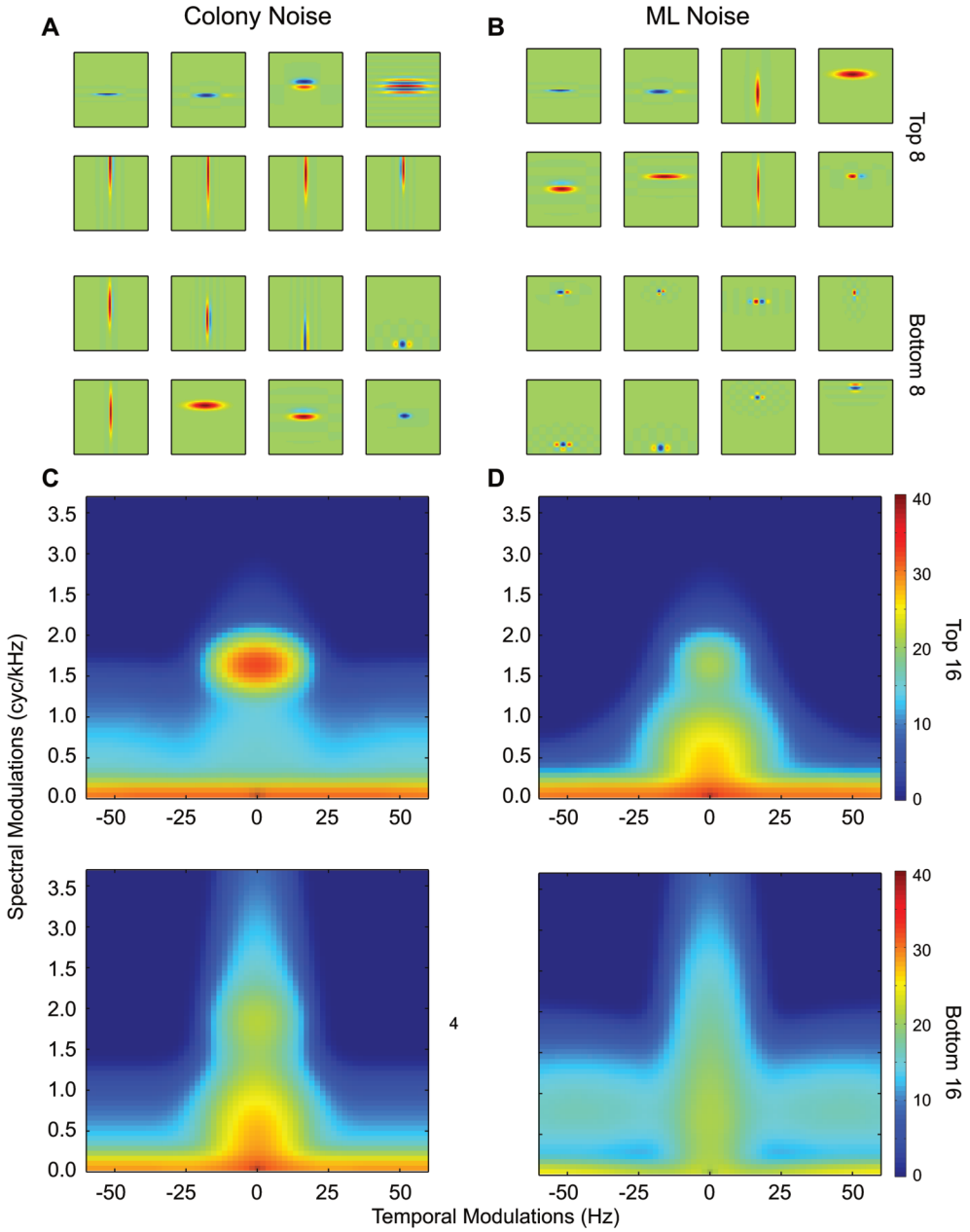

Figure 2.9: **Model STRFs for Noise Reduction.** A. The eight most positively (top) and most negatively (bottom) weighted STRFs from the noise reduction algorithm trained with a background of colony noise. B, Same as in A, but for the model trained with a background of modulation-limited noise. C. The ensemble modulation transfer functions for the top 16 and bottom 16 STRFs for the model trained in colony noise, sorted as in A. D Same as in C, but for the model trained in modulation-limited noise.

noise-invariant neurons to the particular choice of signal and noise. The modeling shows that the importance weights obtained for filtering out ml-noise were slightly different that the weights obtained for filtering colony noise. This relatively small effect can be visually assessed by comparing the highest weighted STRFs for each noise class shown in fig. 2.9A versus fig. 2.9B. These results suggest that slightly different sets of invariant-neurons depending on the statistical nature of the signal and noise but that these effects might be rather small. In addition, we found no correlation between the magnitude of importance weights of the artificial neurons and their BF. Thus, we also predict that the modulation tuning properties of noise-invariant neurons that we described here would apply to a relatively large relevant set of natural signals and noise. This is in part possible because many forms of environmental noise, including noise resulting from the summation of multiple sound signals, have similar modulation structure characterized by a concentration of energy at very low spectral modulations and low to intermediate temporal modulations. In converse, communication signals can have significant energy in regions combining either high spectral modulations with low temporal modulations or high temporal modulations with low spectral modulations (Singh and Theunissen 2003; Sarah M N Woolley, Fremouw, et al. 2005).

Both in the model and in the biological system, given a complete modulation filter bank, the importance weights for a given signal and noise could be learned quickly through supervised learning. Moreover, after learning, the algorithm can easily be implemented in real-time with minimal delay. Thus, the algorithm is particularly useful with adaptive weights or if the statistics of the noise and signal are known, both of which are true in the biological system. Finally given its performance and the advantages described above, we also believe that this noise filtering approach could be useful in clinical applications, such as hearing aids or cochlear implants, or in consumer applications such as noise canceling preprocessing for automatic speech recognition.

In summary, we have shown the presence of noise-invariant neurons in a secondary auditory cortical area. We show that a fraction of the noise-rejecting property can be explained by the spectral-temporal tuning of the neurons. However, tuning properties that are not well captured by the STRF can also both increase or decrease noise-invariance and these properties will

have to be examined in future work. We have also described a novel noise reduction algorithm that uses a modulation filter-bank akin to the STRFs found in the avian auditory system. The performance of this algorithm in noise reduction was excellent and similar or better than the current state-of-the-art algorithms used in hearing aids. The model also illustrates some fundamental principles and allowed us to make stronger statements on the scope of our biological findings. The fundamental principles are, first, that signal and noises can have a distinct signature in the modulation space while overlapping in the frequency space and that therefore filtering in this domain can be advantageous. Second, that although modulation filtering is a linear operation in the spectrogram domain, that both the generation of a spectrogram and the re-synthesis of a clean signal require non-linear computations. We argue that the spectral-temporal properties that are found in higher auditory areas and that are particularly efficient at distinguishing noise modulations from signal modulations are the result of a series of non-linear computations that occurred in the ascending auditory processing stream. The model also shows that a real-time re-synthesis of a cleaned signal could be obtained with additional non-linear operations or, in other words, that a real-time spectrographic inversion is possible. Finally, our modeling efforts show that the noise-invariant findings described here for a song as a chosen prototypical signal and a modulation-limited noise as the chosen prototypical noise would also apply to other signals and noise. However, the involvement of neurons with slightly different tuning or adaptive properties would be needed to obtain optimal signal detection. Given the behavioral experiments that have shown that birds excel at auditory scene analysis tasks both in the wild (Aubin and Jouventin 2002) and in the lab (MacDougall-Shackleton et al. 1998; Benney and Braaten 2000) and given our increasing understating of the underlying neural mechanisms (Bee, Micheyl, et al. 2010), the birdsong model shows great promise to tackle one of the most difficult and fascinating problems in auditory sciences: the analysis of a sound scape into distinct sound objects.

Chapter 3

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

Lee, Tyler & Theunissen, Frederic

3.1 Abstract

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 fea-

tures, retaining similar performance levels while minimizing inherent throughput delays. STDR performs better than a competing state-of-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.

3.2 Introduction

Humans, as social beings, rely heavily on spoken language for communication. The fluctuations in air 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 (Fay 2008), 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 (Palmer 2009); the optimal amplification and processing in quiet environments are often detrimental to the listening experience in noisy environments (Edwards 2004). Similarly, artificial speech recognition (ASR) systems, such as those used in smartphones, often fail in relatively low levels of background noise (Stern and Morgan 2012). 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 (Litovsky 2005) 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 analyze 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 behaviorally relevant sounds. This class of sounds, often broadly declared natural sounds, lies in a particular subspace of all possible sounds (Singh and Theunissen 2003). 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 Theunissen and Elie 2014). 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 neurons response from an incoming sound (Eggermont et al. 1981; David J Klein et al. 2006; Theunissen, Sen, and a. J. Doupe 2000; Sharpee, Atencio, and Schreiner 2011). 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 (reviewed in Theunissen and Elie 2014; David J Klein et al. 2006). 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 behaviorally relevant sounds (Sarah M N Woolley, Fremouw, et al. 2005; Rodríguez et al. 2010; Escabí et al. 2003).

Extrapolating these results to the problem of analyzing sound in noise leads to the postulate that when the brain is presented with a behaviorally relevant sound (e.g. a communication signal) in background noise, the preferential encoding of the behaviorally 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, Lee, and Theunissen 2013 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 (Moore, Lee, and Theunissen 2013). Other work builds on this preferential encoding hypothesis but prescribes more important roles for nonlinear processing (e.g. neural adaptation) and attentional feedback (Mesgarani, Stephen V. David, et al. 2014; Mesgarani and Chang 2012; Zion Golumbic, Poeppel, and Charles E. Schroeder 2012; Rabinowitz et al. 2013; Schneider and Sarah M N Woolley 2013).

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 (Dubbelboer and Houtgast 2008). Other studies indicate that the signal-to-noise ratio in the

spectro-temporal modulation domain correlates strongly with the intelligibility of speech in a wide range of situations (Elhilali, Chi, and S. a. Shamma 2003). More specifically, the lowpass region of spectro-temporal modulations below 7.75 Hz (temporal) and 3.75 cycles / kHz (spectral) seems particularly important for speech intelligibility (T. M. Elliott and Theunissen 2009). 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 (Chabot-Leclerc, Jørgensen, and Dau 2014).

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 (Mesgarani and Chang 2012). Expecting linguistic information also changes the properties of neural responses to degraded speech in lower cortical areas (Hannemann, Obleser, and Eulitz 2007; Holdgraf et al. n.d.). 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 single microphone noise reduction, 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 algorithm (Ephraim and Malah 1985) with a minimum statistics noise estimator (Martin 1994; Martin 2001), across multiple measures of sound quality and intelligibility. Further, we explore the role that predicting upcoming spectro-temporal features can play in producing a system with strong noise reduction and minimal throughput delay.

3.3 Methods

Stimuli

We trained our algorithm on clean speech recordings of the HINT sentence corpus embedded in multiple noise types (Nilsson, Soli, and Sullivan 1994; Bradlow et al. 2011). All stimuli were single channel, sampled at 16 kHz, and band-limited between 25 Hz and 7.5 kHz, with durations averaging 1.9 seconds, ranging from 0.8 to 7.3 seconds. 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 7 different noise types: speech-shaped noise, babble noise and all 5 noise types from the QUT database (D. Dean et al. 2010). 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 freesound.org, along with white noise and pink noise. Training was done using sentences from either 1 speaker or 16 speakers at 0 dB SNR. A detailed description of each stimulus set is provided below.

HINT sentences

The speech used was from the ALLSTAR corpus, a large set of recordings from 128 multi-lingual speakers (Bradlow et al. 2011). We specifically used all recordings of the HINT sentences, a stimulus set that includes 120 short, phonetically balanced sentences (Nilsson, Soli, and Sullivan 1994). Our algorithm was trained on a subset of either 1 female speaker or 16 speakers (8 male and 8 female, including the 1) chosen randomly from the entire database. Though the speakers were multi-lingual, all sentences were spoken in American English. Testing was then done either on held-out sentences from the same speakers, or on 112 novel speakers from the database.

Speech-shaped noise

We first tested the algorithm on speech embedded in speech-shaped noise. Speech-shaped noise is Gaussian amplitude noise filtered to have the same

long-term average sound spectrum as speech. The long-term average spectrum was taken over the entire ALLSTAR corpus. Each noise segment was one second longer than the corresponding speech segment. The speech was then centered and added to the noise at 0 dB SNR, unless otherwise stated.

Babble noise

Babble noise was taken from a 235 second segment from the Noisex-92 database (Varga and Steeneken 1993). For each speech stimulus a random chunk of babble noise 1 second longer than the speech was extracted. Just as the speech-shaped noise stimuli, the speech was then centered and added to the babble noise at 0 dB SNR, unless otherwise stated.

QUT noise

Additional training noise was taken from the QUT noise database (D. Dean et al. 2010). The database contains more than 30 minutes of recording from each of 10 different locations in 5 different noise types: caf, home, street, car, reverb. Training data was taken only from group A locations (food court, kitchen, city, windows down, indoor pool), while group B locations were left for testing. For each speech stimulus a random chunk 1 second longer than the speech was extracted. The speech was then centered and added to the noise at 0 dB SNR, unless otherwise stated.

Freesound.org

Additional noise types used in testing the algorithm were downloaded from the website `freesound.org`. The sounds were made available by users of the site under a variety of Creative Commons licenses. We selected multiple examples from each of 10 different ambient noise types: airport, bird, machine, ocean, rain, rain forest, theatre, train, train station, and wind (table 3.1). For each noise type we chose 5 clips of 5 seconds duration each. This was done using Audacity audio editing program, and the selections were made to provide a representative sampling of each noise type.

		URL (http://www.freesound.org/people/)	Start time (min:sec)
Airport	1		
	2	polymorpheva/sounds/104545	04:42.43
	3	polymorpheva/sounds/122196	27.306
	4	joedeshon/sounds/273468	11.451
	5	joedeshon/sounds/273468	01:46.74
Birds	1	saphe/sounds/194889	01:18.96
	2	jus/sounds/73617	2.383
	3	flio191/sounds/269333	22.62
	4	flio191/sounds/269333	44.072
	5	frankie01234/sounds/214869	43.178
Machine	1	Personal library	02:50.15
	2	rutgermuller/sounds/104079	1.041
	3	viertelnachvier/sounds/249636	6.676
	4	atmowav/sounds/126289	1.041
	5	felipelnv/sounds/153299	2.228
Ocean	1	viertelnachvier/sounds/249637	8.675
	2	slanesh/sounds/31762	19.677
	3	slanesh/sounds/31762	42.129
	4	xserra/sounds/161700	4.588
	5	abcopen/sounds/166214	58.624
Rain	1	abcopen/sounds/166214	01:56.42
	2	inchadney/sounds/22132	4.911
	3	inchadney/sounds/22132	34.04
	4	giddykipper/sounds/53489	15.018
	5	giddykipper/sounds/53489	43.052
Rainforest	1	inchadney/sounds/221059	03:07.83
	2	laurent/sounds/15467	38.074
	3	laurent/sounds/15468	58
	4	bulj/sounds/93710	5.649
	5	bulj/sounds/93710	01:44.77
Theater	1	laurent/sounds/163355	46.633
	2	corsica-s/sounds/28422	22.374
	3	corsica-s/sounds/28422	55.507
	4	edhutschek/sounds/242604	7.323
	5	edhutschek/sounds/242604	42.64
Train	1	temawas/sounds/179875	03:59.49
	2	sound-of-silenced/sounds/127956	31.317
	3	sound-of-silenced/sounds/127956	02:48.23
	4	rollingmill/sounds/262262	14.15
	5	rollingmill/sounds/262262	02:13.73
Train station	1	rollingmill/sounds/262262	58.284
	2	erh/sounds/58179	5.875
	3	erh/sounds/61496	33.311
	4	volivieri/sounds/50678	01:48.33
	5	kyster/sounds/169722	05:08.59
Wind	1	erh/sounds/58179	27.349
	2	incarnadine/sounds/13234	33.877
	3	incarnadine/sounds/13234	21.772
	4	incarnadine/sounds/16109	20.845
	5	raremess/sounds/238038	1.82

Table 3.1: Composition of 10 categories of untrained noise types downloaded from `freesound.org`.

Spectro-Temporal Detection-Reconstruction (STDR) 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 signal-to-noise ratio of the channel. Finally, these scaled signals are summed to produce an estimate of the clean signal, $\hat{s}(t)$: $\hat{s}(t) = \sum_f y(f,t) \cdot \hat{g}(f,t)$.

This scheme is often called an analysis-synthesis design and has been used successfully for decades in many single microphone noise reduction algorithms (Boll 1979; McAulay and Malpass 1980; Ephraim and Malah 1984). Where these algorithms differ is in the method of estimating signal-to-noise and the functional form of the gains. Here we utilize an artificial neural network that attempts to analyze the spectro-temporal modulations present in the noisy signal (Detection stage) to estimate the optimal time-varying gains (Reconstruction stage) (fig. 3.1). 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 required in both the noisy signal and the time-varying gains.

Analysis and spectrogram computation

To compute the narrowband signals, $y(f,t)$, we created a filterbank with 223 bandpass Gaussian-shaped filters with center frequencies linearly spaced between 25 Hz to 7.5 kHz and bandwidths of 32 Hz each, corresponding to a time-domain window 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 channels amplitude envelope below 192 Hz (6 x 32 Hz, since 6 standard deviations accounts for approximately 99.8%

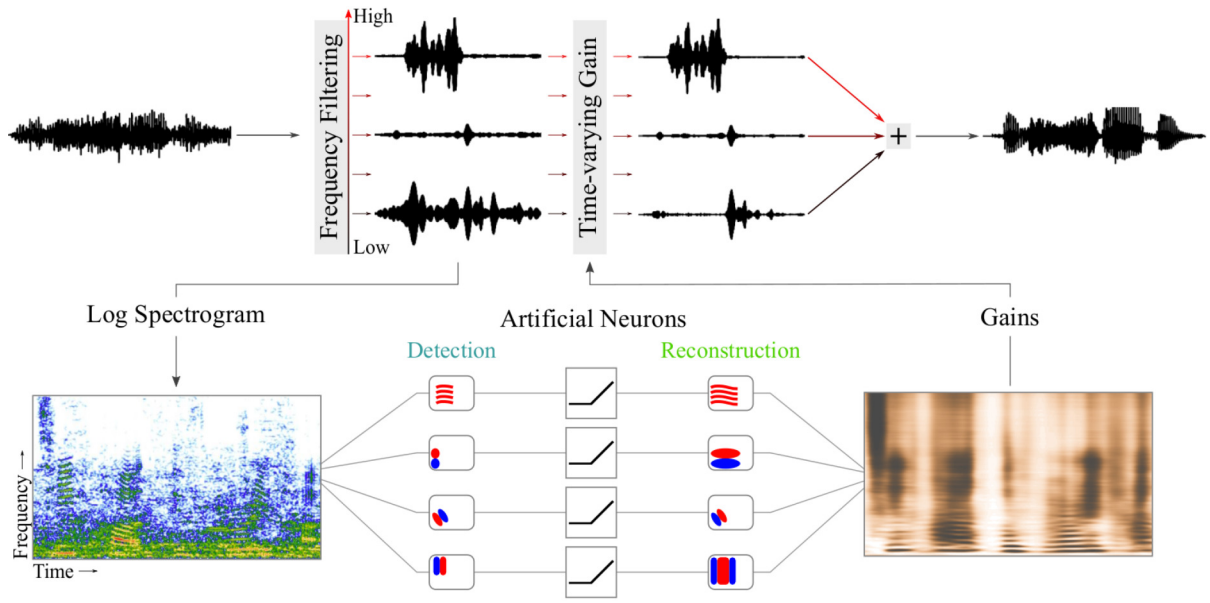

Figure 3.1: The spectro-temporal detection-reconstruction (STDR) algorithm is composed of two chains: the analysis-synthesis chain and the gain estimation chain (see section 3.3). Top row: a noisy signal waveform is first bandpass filtered into a set of narrowband channels. Each narrowband channel is 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 the density of the window) (Flanagan 1980). Each channel's envelope was then extracted by computing the analytic signal and then downsampled to 1 kHz, producing a spectrogram $X_{lin}(f,t)$. The spectrogram was then log transformed with a floor value set at -80 dB 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 modeled using a set of bandpass filters, followed by a half-wave rectification, low-pass filtering, and adaptive gain control (Lyon 1982). This complete preprocess-

ing was applied to each individual sound before being fed into the network as $X(f,t)$.

Artificial neural network

The artificial neural network was structured as a three-layer autoencoder (Hinton and Salakhutdinov 2006). The input to the network was processed spectrogram, $X(f,t)$. Each first layer unit operated on this time-frequency representation using a spectro-temporal filter:

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

, 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 layers input weight matrix, if desired. The number of units in the first layer was chosen to be 100 and τ ranged from 0 ms 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 layers 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.

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

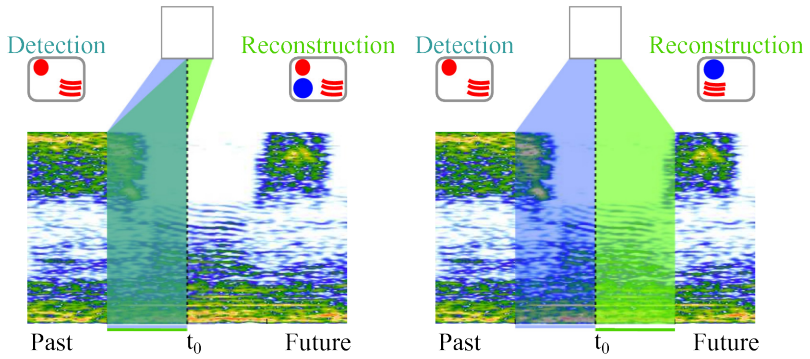

Figure 3.2: The reconstruction kernels can be used to apply gains completely in the past, overlapping the window used in detection (*causal* detection combined with *acausal* reconstruction), completely in a predictive mode where the detection window is in the past and the reconstruction kernel window 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.

. 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. Though 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 ms to 0 ms where the reconstructed gains are entirely in the past, and sliding the window to the completely *predictive* regime of 0 ms to 99 ms where the reconstructed gains are entirely in the future (fig. 3.2). Also note that is a common practice to set the minimum asymptotic value of the sigmoid to some small, nonzero value: $g = g_{min} + (1 - g_{min})\hat{g}$. 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 were varied symmetrically from 10 ms 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 (fig. 2.2). 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 (STRF) by auditory neurophysiologists and have been shown to effectively represent speech (Mesgarani 2008). 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, or STDR, algorithm.

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 7 noise type training, of signal in background noise, each less than 5 seconds 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

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

, with

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

. 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 5 consecutive iterations on a held-out portion of 10% of the training data. All filter weights were initialized to small, uniform random values centered on zero. The range for the weights was chosen using the normalized initialization heuristic from (Glorot and Bengio 2010), 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.

Performance metrics

We assessed the performance of our algorithm using objective measures of sound quality, speech intelligibility. Sound quality was quantified using three composite ratings as proposed by Hu and Loizou 2008. 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 signal-to-noise ratio (Hansen and Pellom 1998), the weighted spectral slope (Klatt 1982), the log likelihood ratio (Quackenbush, Barnwell, and Clements 1988) and the perceptual estimate of sound quality (PESQ) (Rix et al. 2001). The three ratings showed correlations of 0.73, 0.64, and 0.73 between objective and subjective quality judgments 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 (Taal et al. 2011). This measure was shown to significantly correlate with subjective reports of speech intelligibility, with a correlation coefficient of 0.92 for speech processed using single microphone noise reduction 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 Ephraim-Malah algorithm) we used a linear mixed-effects model. Both the comparison algorithm and the mixed-effects model are described in more detail in sections 3.3 and 3.3.

Normalized performance

Normalized performance values shown in fig. 3.6 and fig. 3.8 were computed as:

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

. 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 algorithms performance is the percentage of improvement over the unfiltered noisy speech on any particular metric that the algorithm could hope to achieve.

Comparison algorithm

We compared our algorithm with a current standard method for SMNR. The algorithm computes the optimal gain to map the noisy speech log spectral amplitude to the clean speech log spectral amplitude, as put forth by Ephraim and Malah (EM) (Ephraim and Malah 1985). A minimum statistics noise es-

timator provides the EM algorithm with the required estimate of the noise spectrum. The minimum statistics algorithm estimates the noise spectrum as proportional to the minimum power within each frequency band across a short time window (Martin 2001). The algorithm is founded on the expectation that the target speech envelope is highly modulated with brief bouts of silence between words and syllables. If the time window over which you estimate the minimum power is long enough to reliably include these silent periods, the noise spectrum can be estimated as the minimum values. Because the noise estimator works on a relatively short time window (1.5 seconds), the algorithm is capable of handling nonstationary noises and continuous speech without any large silent periods. The algorithm used was the ssubmmse.m MATLAB routine implemented in the freely available package VOICEBOX (Brookes 2002). Code for the EM algorithm was downloaded from: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`. All user dependent parameters of the EM algorithm were left at their default values.

Significance testing

To determine if the performance of our STDR algorithm was significantly better than either the unfiltered noisy signal or the EM algorithm, we used a linear mixed-effects model with algorithm and stimulus SNR as fixed effects and sentence ID as a random effect. This model was implemented using the MATLAB function fitlme.m. The model computes the best estimate for the difference in performance on any given metric, that processing with STDR adds over either the unfiltered stimulus or the EM filtered stimulus, as well as confidence intervals and p-values for this difference being nonzero. All statistics are reported as: Delta in performance (p-value). The number of degrees of freedom for each LME model were: 207 for all single speaker tests and 3357 for all 16 speaker tests.

3.4 Results

As described in the methods, we developed a novel algorithm for single microphone noise reduction called spectro-temporal detection-reconstruction

(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.

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 signal-to-noise ratios (SNR) as compared to 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. Section 3.4 shows four example units trained on speech from a single speaker embedded in babble noise. The first unit (section 3.4c) 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 nonzero only when their particular feature is present in the stimulus. The filter of the second unit (section 3.4d) 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 of the speech signal. The filter of the third unit (section 3.4e) detects coarse power in the mid-frequencies with some harmonic structure. The filter of the fourth unit (section 3.4f) is highly specific, responding selectively to the harmonic structure of the trained speakers voice. The detection filters and reconstruction gain filters for all units are shown in fig. 3.4. In general, there was a consistent dichotomy between noise-suppressive and signal-selective

units with a continuum of filter types within each category.

Performance on speech in speech-shaped noise

By utilizing an array of individual units with different feature selectivity learned from a representative data set, the algorithm is able to produce accurate reconstructions to novel noisy stimuli. We tested the performance of our algorithm using sentences from the hearing-in-noise test (HINT) embedded in two types of background noise: speech-shaped noise and babble noise. The speech-shaped noise tested was computed to match the spectrum of each individual sentence. The STDR algorithm was trained on a set of 100 sentences from either 1 or 16 individuals, chosen randomly from a large database of speakers, at 0 dB SNR (See section 3.3). fig. 3.5 shows the performance of the STDR algorithm on a novel sentence from speaker 1 when trained on speech from 1 speaker (fig. 3.5c and fig. 3.5e) and 16 speakers (fig. 3.5d and fig. 3.5f). A few features stand out when looking at the time-frequency gains. Firstly, it captures precisely the low frequency harmonics corresponding to the speakers pitch. This effect is much stronger when the algorithm was trained only on the speaker shown but is still present when trained on 16 speakers (fig. 3.5c and fig. 3.5d, inset). Secondly, 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. Thirdly, 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 supplemental materials.

To quantify the performance, we processed 15 novel sentences from each trained speaker at 7 different SNRs and then computed several objective measures of performance that have been used in the field (see section 3.3). 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

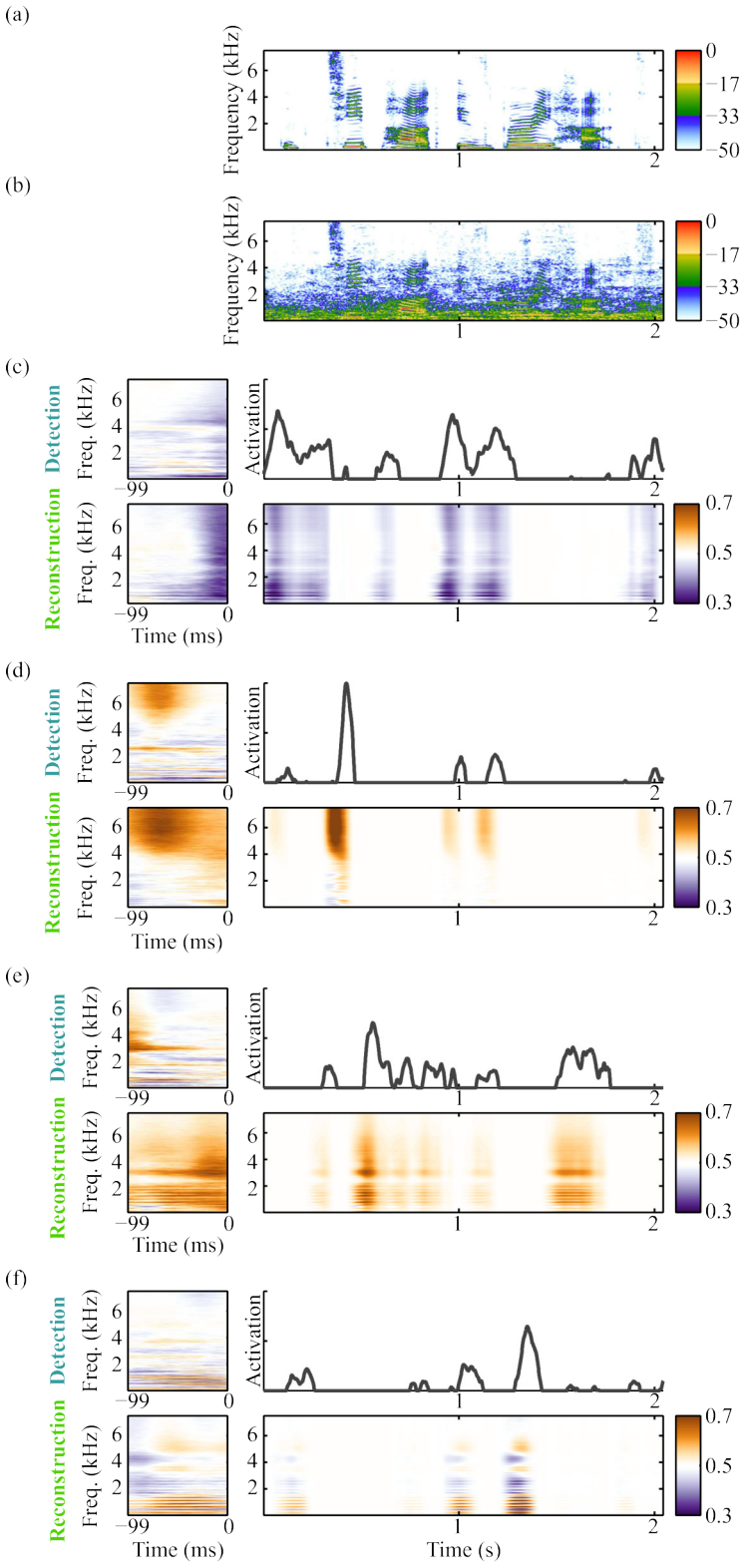
(a)
(b)
(c)
(d)
(e)
(f)
Frequency (kHz)
Detection
Reconstruction
Freq. (kHz)
Activation
Time (ms)
Time (s)

Figure 3.3: 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.

around 0. To assess the intelligibility of the processed speech, we computed the short-time objective intelligibility (STOI) measure (Taal et al. 2011). Our STDR algorithm showed slight but significant improvements on this measure over the unfiltered noisy speech (.03 ($p < 10^{-4}$), df=207, linear mixed-effects model, see section 3.3)(fig. 3.6a,b, left column). It also significantly outperformed a standard noise reduction algorithm that utilizes minimum statistics noise estimation and log MMSE optimal frequency filtering, the Ephraim Malah (EM) algorithm (.05 ($p < 10^{-4}$), see section 3.3)(Ephraim and Malah 1985). These benefits were seen on a large majority of individual sentences (fig. 3.6b, left column).

To assess the resulting quality of the processed stimulus, we computed a set of three composite measures (Hu and Loizou 2008). 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 section 3.3). The STDR algorithm performed well, significantly improving each rating over the unfiltered stimulus (estimated improvement of .54 ($p < 10^{-4}$), .30 ($p < 10^{-4}$) and .44 ($p < 10^{-4}$) for signal, background and overall, respectively). It also provided significant improvements over the EM algorithm on all three measures (.32 ($p = 6 \times 10^{-4}$), .07 ($p < 10^{-4}$), .15 ($p < 10^{-4}$))(fig. 3.6a,b, center and right columns show background noise intrusiveness and overall

Figure 3.4: All detection and gain reconstruction filters for the model trained on a single speaker in babble noise. Filters from the four units highlighted in section 3.4 are bounded with a thick box.

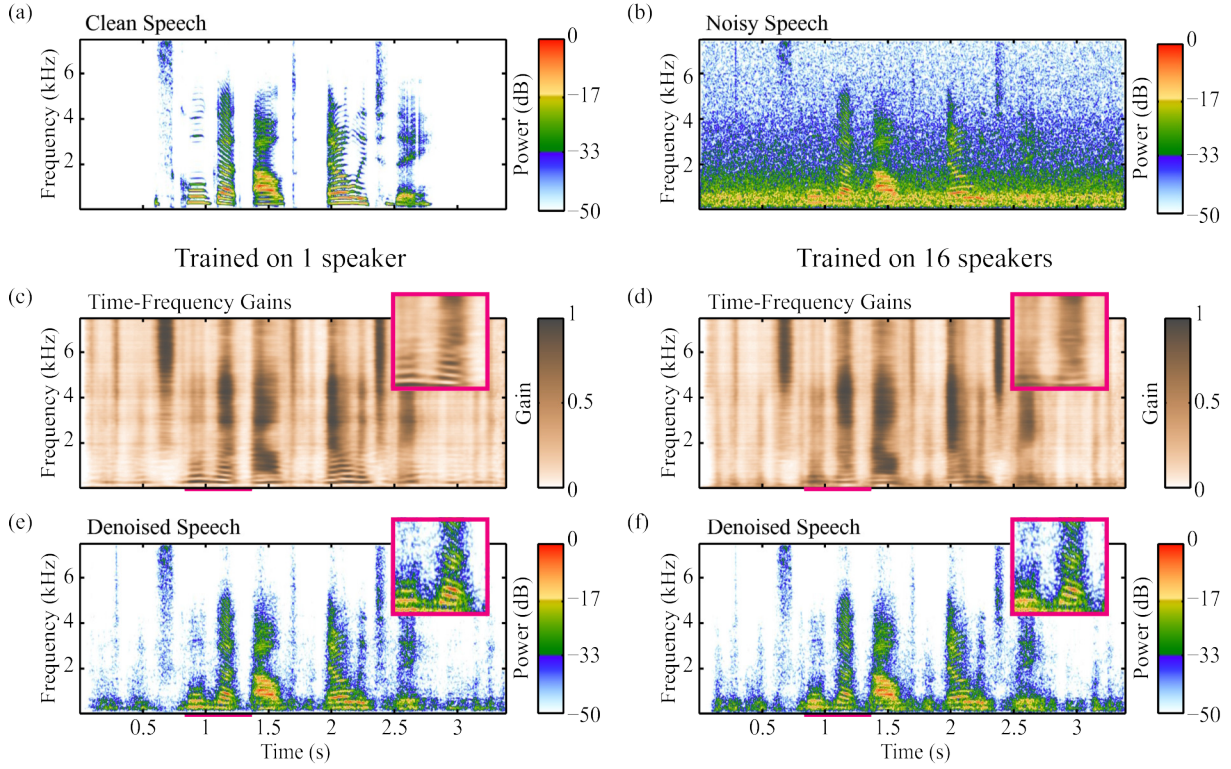

Figure 3.5: An example of filtering by the STDR algorithm when trained on 1 or 16 speakers in stationary speech-shaped noise. (a) A spectrogram of the sentence "School got out early today". (b) A spectrogram of the same sentence after the addition of speech-shaped noise at 0 dB SNR. (c) The predicted 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) and (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 speakers characteristic pitches and pitch transitions. (e) and (f) The resulting estimated clean speech spectrogram obtained by applying the gains from (c) and (d), respectively, to (b).

quality, respectively). The performance increases were not just in aggregate but were found for the vast majority of the sentences (fig. 3.6b, center and right columns). The mean performance for each processing on the complete set of 8 metrics computed (STOI, 3 composite measures and the 4 component measures they comprise) is shown in table 3.2.

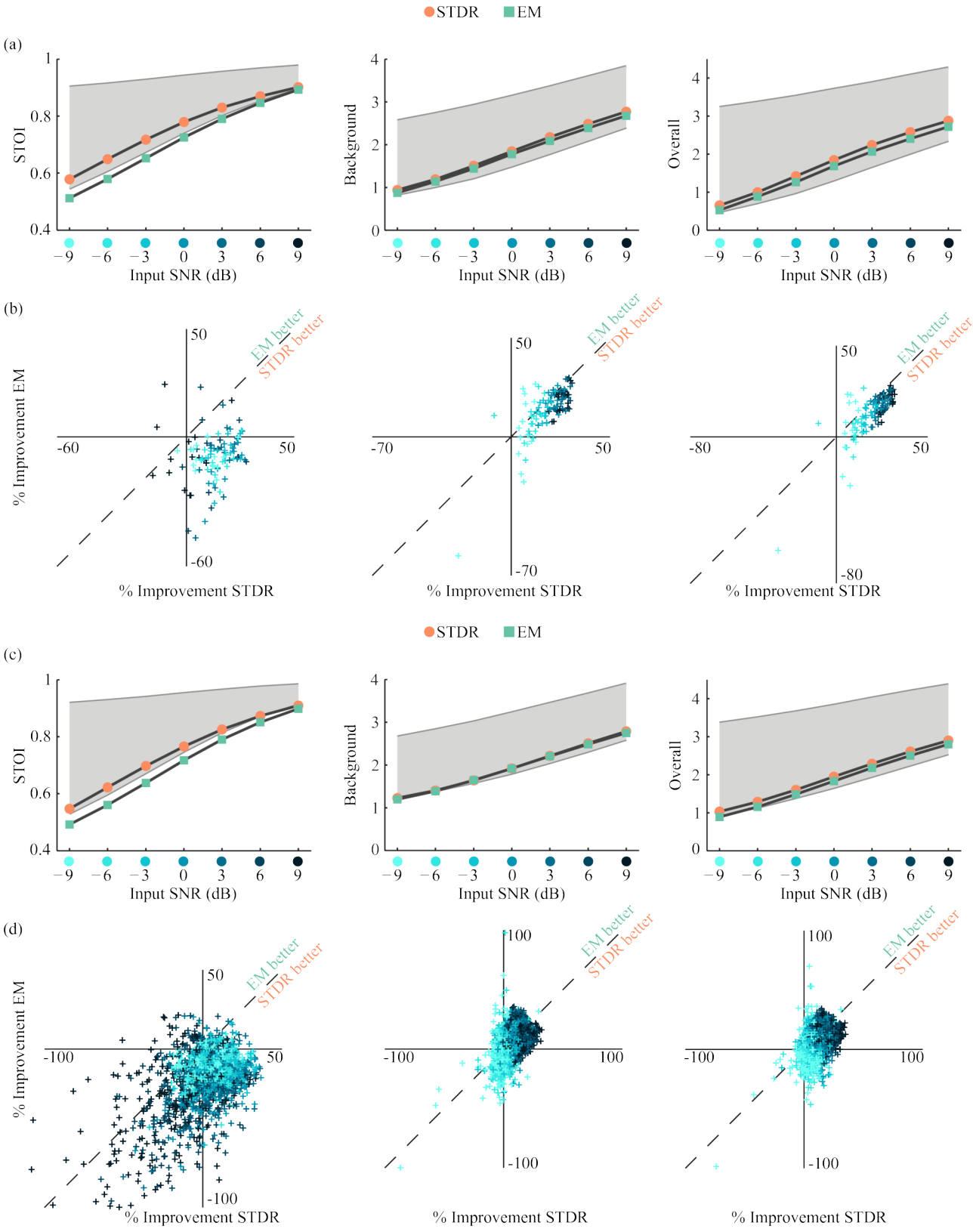
STDR
EM
(a)
STOI
Background
Overall
Input SNR (dB)
(b)
% Improvement EM
% Improvement STDR
EM better
STDR better
(c)
(d)

Figure 3.6: The performance of the algorithm on speech in speech-shaped noise was measured using 4 different objective measures (3 shown). Stimuli consisted of a hold-out set of 15 sentences from each speaker processed at 7 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) and (d) were obtained from the algorithm trained on 16 speakers. The measures shown here are: the short-time objective intelligibility (STOI) rating, the composite rating of background intrusiveness, and the composite rating of overall quality (see section 3.3). (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 section 3.3): 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) and (d) Same plots as in (a) and (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 8 metrics computed (STOI, 3 composite measures and the 4 component measures they comprise) is shown in table 3.2 and table 3.3.

Algorithm	Unfiltered	EM	STDR
Metric			
Composite Signal	1.76	2.0	**2.31**
Composite Background	1.55	1.77	**1.84**
Composite Overall	1.36	1.65	**1.8**
STOI	0.73	0.71	**0.76**
LLR	1.29	1.23	**1.0**
Segmental SNR	-1.6	-0.19	**-0.06**
PESQ	1.27	**1.64**	1.55
WSS	84.64	91.5	**75.69**

Table 3.2: Performance of EM and STDR algorithms on 15 sentences from a single speaker embedded in speech-shaped noise. Values are averaged over all 15 sentences and 7 different SNRs. Bold values indicate best performance. For all but log-likelihood ratio (LLR) and weighted spectral slope (WSS) larger values are better.

Algorithm	Unfiltered	EM	STDR
Metric			
Composite Signal	2.08	2.18	**2.45**
Composite Background	1.85	1.94	**1.96**
Composite Overall	1.7	1.84	**1.95**
STOI	0.74	0.71	**0.75**
LLR	1.35	1.28	**1.05**
Segmental SNR	-1.34	**-0.41**	-0.53
PESQ	1.5	**1.73**	1.65
WSS	**59.34**	71.04	62.16

Table 3.3: Performance of EM and STDR algorithms on 15 sentences each from 16 speakers embedded in speech-shaped noise. Values are averaged over all 240 sentences and 7 different SNRs. Bold values indicate best performance.

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 (fig. 3.6c,d). For the STOI ratings, the STDR algorithm showed slight but significant improvements over both the unfiltered speech (.01 ($p < 10^{-4}$), df=3357) and the EM algorithm (.04 ($p < 10^{-4}$)). Similarly, STDR improved

the composite ratings of quality over unfiltered speech (.38 ($p < 10^{-4}$), .11 ($p < 10^{-4}$), .26 ($p < 10^{-4}$), for signal, noise and overall quality, respectively) and the EM algorithm for all three metrics (.27 ($p < 10^{-4}$), .02 ($p = 5 \times 10^{-4}$), .12 ($p < 10^{-4}$), for signal, noise and overall quality, respectively).

Performance on speech in babble noise

A more challenging stimulus set is shown in fig. 3.7. 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 (fig. 3.7 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 supplementary materials.

Looking again at the entire set of 15 sentences per speaker, composite quality ratings were significantly increased over unfiltered speech (.61 ($p < 10^{-4}$), .66 ($p < 10^{-4}$), .60 ($p < 10^{-4}$), speech distortion, noise intrusiveness, and overall quality, respectively)(fig. 3.8a,b, composite rating of noise intrusiveness in center column) and over the EM algorithm (.16 ($p < 10^{-4}$), .45 ($p < 10^{-4}$), .23 ($p < 10^{-4}$)). For the composite ratings, the STDR algorithm showed larger benefits over the EM algorithm when processing babble noise instead of speechshaped noise. This is due primarily to the EM algorithm showing smaller, though still significant, benefits from processing, likely because of the temporal nonstationarity of the noise. Performance gains on the STOI measure were lessened, though still significant, for the STDR algorithm (.02 ($p < 10^{-4}$)) over both unfiltered and EM processed speech.

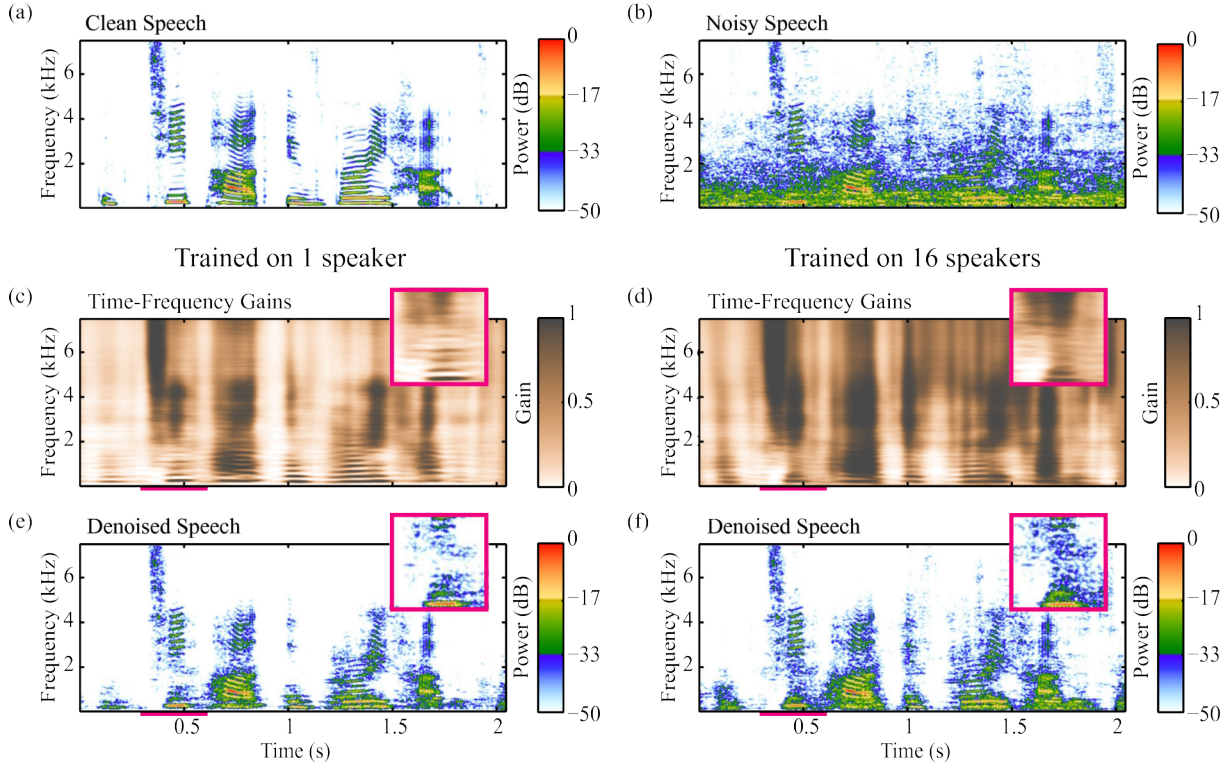

Figure 3.7: An example of filtering by the STDR algorithm when trained on 1 or 16 speakers in nonstationary babble noise. The figure layout is identical to fig. 3.5. (a) The clean speech spectrogram for the sentence "The teapot was very hot." (b) Spectrogram for the sentence from (a) added to babble noise at 0 dB SNR. (c) and (d) Again, the resulting gains in both the 1 and 8 speaker case are very similar but the 1 speaker model is able to capture more precise harmonic structure (inset). (e) and (f) The resulting estimated clean speech spectrogram obtained by applying the gains from (c) and (d), respectively, to (b).

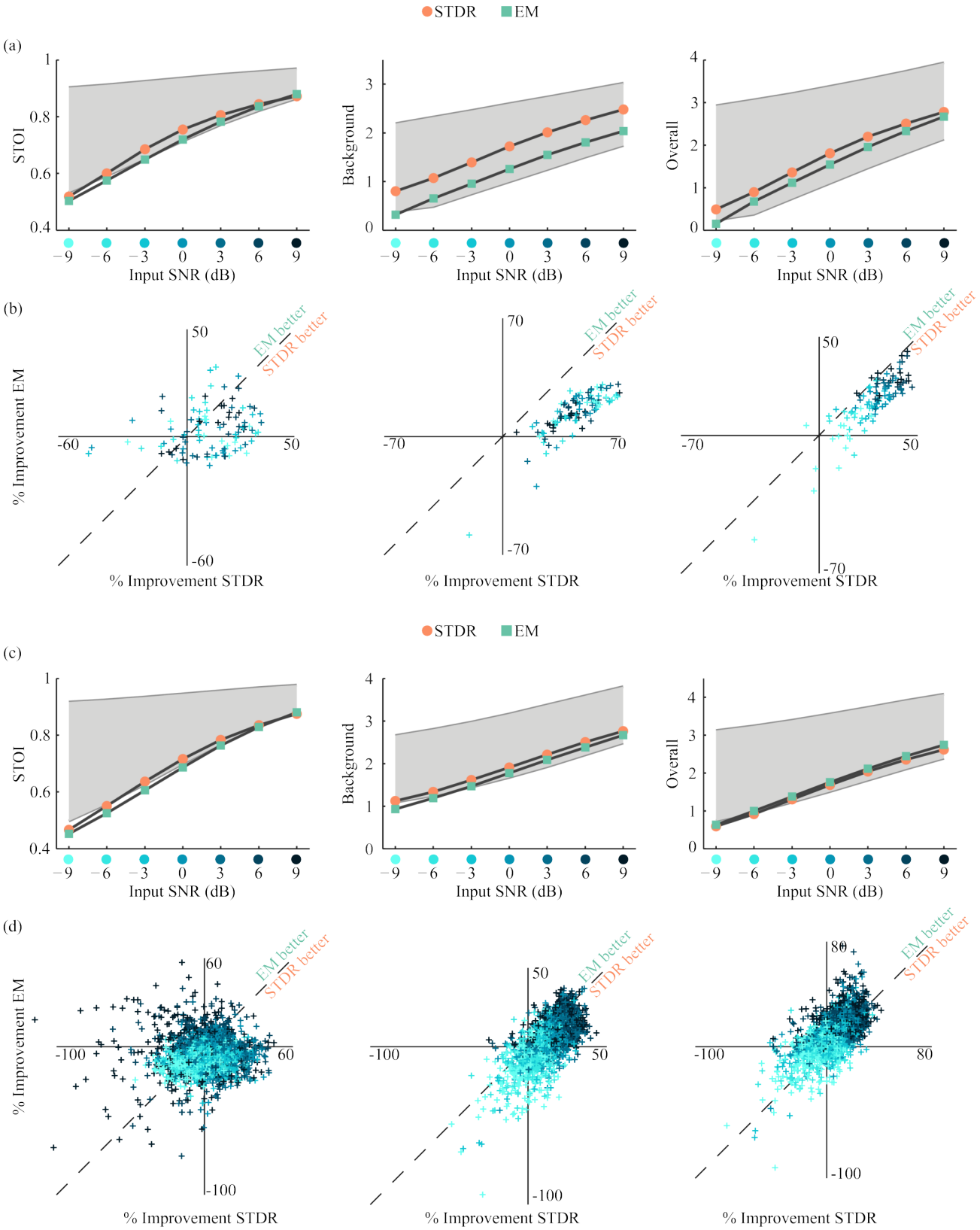
STDR
EM
(a)
STOI
Background
Overall
Input SNR (dB)
(b)
% Improvement EM
% Improvement STDR
EM better
STDR better
(c)
(d)

Figure 3.8: The performance of the algorithm on speech in nonstationary babble noise was measured using 4 different objective measures (3 shown). The figure layout is identical to fig. 3.6. The mean performance for each processing on the complete set of 8 metrics are shown in table 3.4 and table 3.5.

Algorithm	Unfiltered	EM	STDR
Metric			
Composite Signal	1.4	1.84	**2.0**
Composite Background	1.01	1.22	**1.68**
Composite Overall	1.12	1.49	**1.72**
STOI	0.71	0.71	**0.73**
LLR	1.43	**1.15**	1.34
Segmental SNR	-7.26	-5.81	**-3.12**
PESQ	1.29	1.6	**1.74**
WSS	111.6	115.83	**84.84**

Table 3.4: Performance of EM and STDR algorithms on 15 sentences from a single speaker embedded in babble noise. Values are averaged over all 15 sentences and 7 different SNRs. Bold values indicate best performance.

Performance of STDR trained on 16 speakers was generally similar (fig. 3.8c,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 (-.17 ($p < 10^{-4}$), .14 ($p < 10^{-4}$), -.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=.98), while the EM algorithm was significantly worse (-.01 ($p < 10^{-4}$)).

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 7 different noise types (5 QUT noises (see section 3.3), Noisex babble noise and speech-shaped noise). The algorithm was then tested on

Algorithm	Unfiltered	EM	STDR
Metric			
Composite Signal	1.79	**2.06**	1.89
Composite Background	1.72	1.78	**1.92**
Composite Overall	1.53	**1.72**	1.64
STOI	0.69	0.68	**0.69**
LLR	1.51	**1.23**	1.5
Segmental SNR	-1.93	-0.74	**0.11**
PESQ	1.53	**1.72**	1.64
WSS	74.61	89.2	**72.23**

Table 3.5: Performance of EM and STDR algorithms on 15 sentences each from 16 speakers embedded in babble noise. Values are averaged over all 240 sentences and 7 different SNRs. Bold values indicate best performance.

novel sentences from each speaker and noise type, as well as from 114 untrained speakers embedded in 12 different untrained noise types gathered from freesound.org (see section 3.3). 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 (fig. 3.9 and fig. 3.10). In general, STOI was unaffected or slightly decreased, while composite measures were unaffected or significantly improved. Specifically, 6 of 7 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

		EM	STDR	Unfiltered
Babble				
	Composite Signal	2.08	2.24	**2.41**
	Composite Background	1.73	1.78	**1.87**
	Composite Overall	1.66	1.76	**1.88**
	STOI	0.68	0.68	**0.72**
	LLR	**1.15**	1.1	1.0
	Segmental SNR	**-0.77**	-1.39	-0.84
	PESQ	1.55	1.53	**1.57**
Cafe		**84.55**	71.7	65.99
	Composite Signal	1.23	**1.84**	1.37
	Composite Background	1.79	**1.92**	1.84
	Composite Overall	1.21	**1.56**	1.3
	STOI	0.7	0.73	**0.74**
	LLR	**2.0**	1.56	1.99
	Segmental SNR	-0.14	**0.02**	-0.96
	PESQ	1.46	**1.46**	1.41
Car		**76.72**	59.36	57.82
	Composite Signal	**2.51**	2.5	2.45
	Composite Background	**2.03**	1.95	1.96
	Composite Overall	**2.04**	1.98	1.94
	STOI	0.74	0.74	**0.77**
	LLR	1.02	1.01	**1.05**
	Segmental SNR	**0.31**	-0.61	-0.54
	PESQ	**1.82**	1.67	1.62
Home		**69.83**	63.1	59.86
	Composite Signal	2.05	**2.35**	2.19
	Composite Background	1.91	2.01	**2.02**
	Composite Overall	1.69	**1.89**	1.8
	STOI	0.72	0.73	**0.76**
	LLR	1.3	1.14	**1.3**
	Segmental SNR	**0.57**	0.49	0.51
	PESQ	1.61	**1.63**	1.6

		EM	STDR	Unfiltered
Reverberant		**75.64**	61.67	58.86
	Composite Signal	1.4	**1.7**	1.14
	Composite Background	**1.95**	1.95	1.85
	Composite Overall	1.4	**1.51**	1.2
	STOI	0.72	0.74	**0.76**
	LLR	2.01	1.73	**2.25**
	Segmental SNR	**0.29**	-0.04	-1.19
	PESQ	**1.63**	1.49	1.42
SSN		**68.4**	56.98	55.25
	Composite Signal	2.18	**2.3**	2.07
	Composite Background	**1.93**	1.86	1.84
	Composite Overall	**1.83**	1.82	1.69
	STOI	0.71	0.72	**0.73**
	LLR	1.28	1.12	**1.35**
	Segmental SNR	**-0.43**	-1.2	-1.4
	PESQ	**1.73**	1.55	1.5
Street		**71.56**	63.51	59.59
	Composite Signal	1.59	**1.92**	1.6
	Composite Background	1.8	**1.86**	1.81
	Composite Overall	1.43	**1.6**	1.41
	STOI	0.7	0.72	**0.74**
	LLR	1.69	1.45	**1.74**
	Segmental SNR	**-0.35**	-0.55	-0.99
	PESQ	**1.55**	1.48	1.41

Table 3.6: Performance of EM and STDR algorithms on 5 sentences each from 16 speakers embedded in 7 different training noise types. Values are averaged over all 80 sentences and 7 different SNRs per noise type. Bold values indicate best performance.

could increase the sensitivity to these diminishing discriminating features by increasing the number of units (PCs) or adding intermediate layers in our network.

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 utilizing 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 applica-

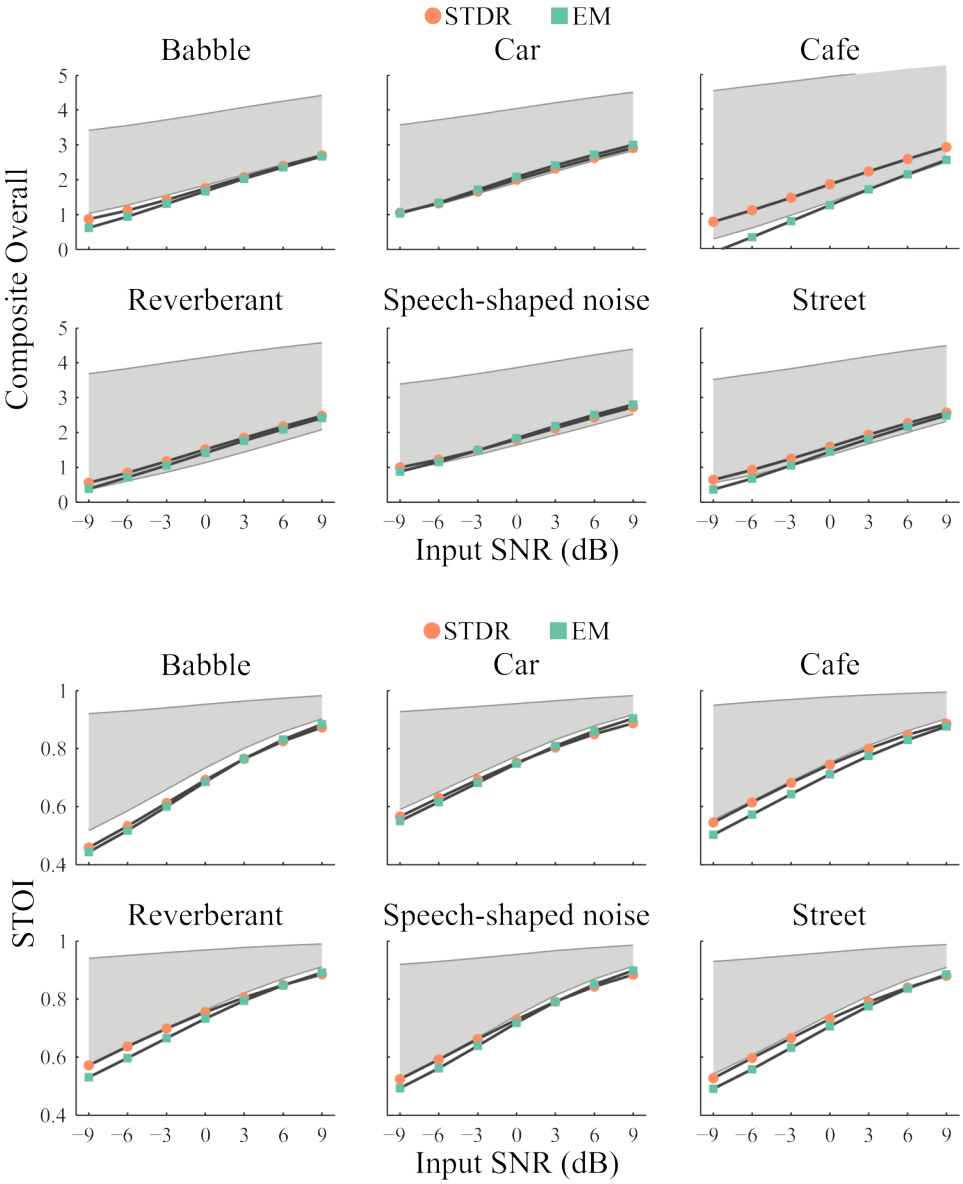

Figure 3.9: Mean performance on 6 of 7 training noise types (Home noise not shown) for both the STDR and EM algorithms. Testing was done on 15 novel sentences from each of the 16 training speakers, each embedded in a novel example from the specified noise type. Car, Cafe, Home, Reverberant and Street noises are taken from the QUT database, babble noise was taken from the Noisex-92 database, and speech-shaped noise was created to match the long-term average speech spectrum of the entire ALLSTAR database of 128 speakers. Plots are formatted as in fig. 3.6 and fig. 3.8. In the interest of space, only the composite overall quality and STOI measures are shown. Average values across all SNRs are shown in table 3.6.

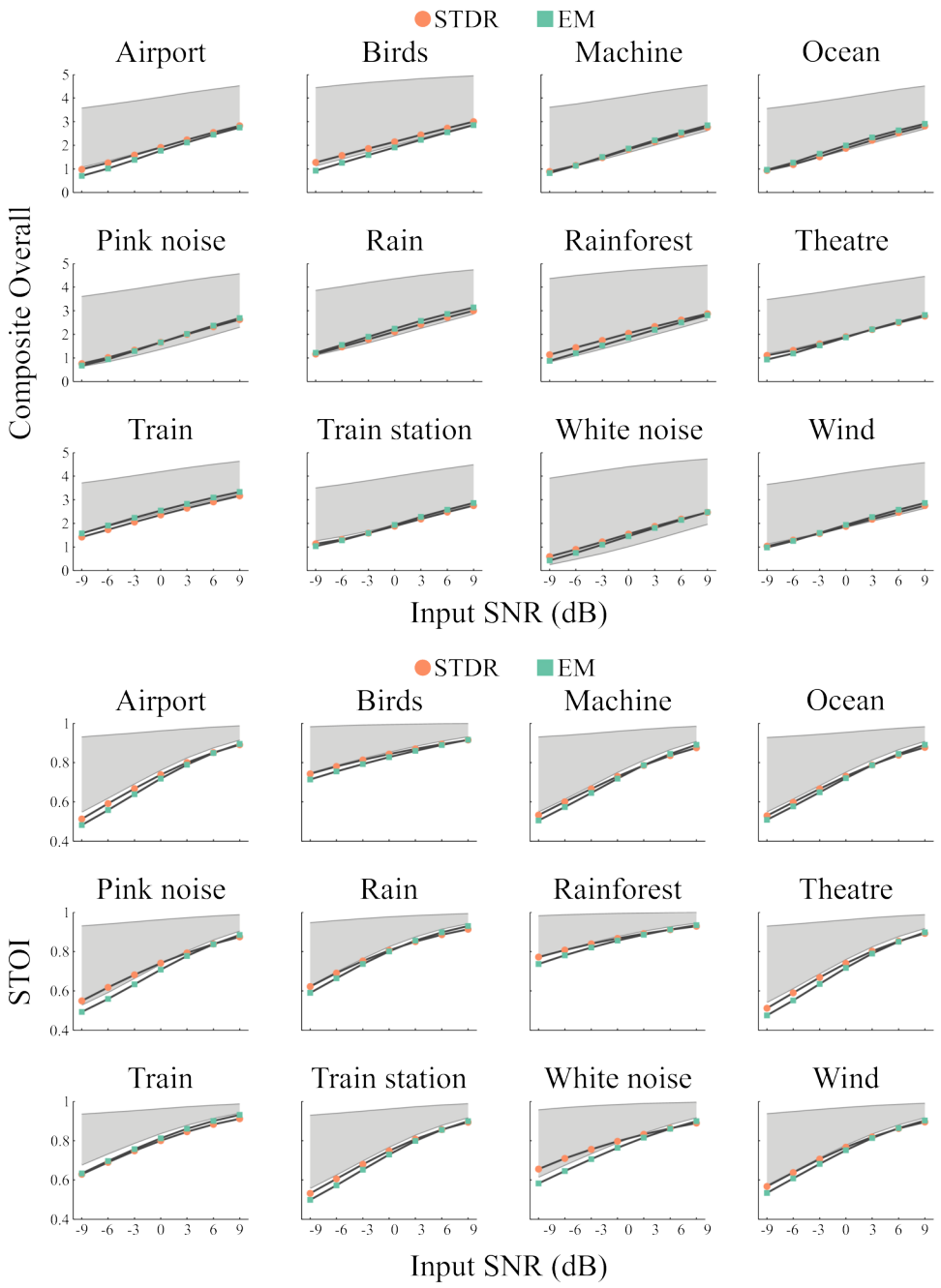

Figure 3.10: Mean performance on all 12 novel noise types for both the STDR and EM algorithms. Testing was done on 5 sentences from each of 112 novel speakers, each embedded in an example of the specified noise type. Ten of twelve noise types were downloaded from freesound.org and are listed in table 3.1. White noise and pink noise were generated. In the interest of space, only the composite overall quality and STOI measures are shown. Average values across all SNRs are shown in table 3.7.

		EM	STDR	Unfiltered
Airport				
	Composite Signal	2.16	2.4	**2.49**
	Composite Background	1.79	1.88	**1.92**
	Composite Overall	1.74	1.9	**1.95**
	STOI	0.7	0.72	**0.75**
	LLR	**1.13**	1.06	1.0
	Segmental SNR	**-0.67**	-1.15	-0.91
	PESQ	**1.62**	1.62	1.61
Birds		**82.3**	64.98	60.77
	Composite Signal	1.83	**2.21**	2.04
	Composite Background	2.26	**2.46**	2.34
	Composite Overall	1.9	**2.14**	2.02
	STOI	0.82	0.84	**0.85**
	LLR	**1.99**	1.76	1.91
	Segmental SNR	-0.03	**1.32**	-0.36
	PESQ	2.12	**2.17**	2.09
Machine		**54.55**	41.56	38.83
	Composite Signal	2.17	**2.25**	2.11
	Composite Background	**2.0**	1.91	1.92
	Composite Overall	**1.84**	1.82	1.74
	STOI	0.71	0.72	**0.74**
	LLR	1.33	1.22	**1.37**
	Segmental SNR	**0.13**	-0.92	-1.02
	PESQ	**1.75**	1.57	1.54
Ocean		**67.61**	59.31	55.53
	Composite Signal	**2.4**	2.38	2.33
	Composite Background	**1.98**	1.87	1.89
	Composite Overall	**1.96**	1.86	1.83
	STOI	0.71	0.72	**0.74**
	LLR	1.1	1.07	**1.14**
	Segmental SNR	**-0.16**	-1.23	-1.1
	PESQ	**1.74**	1.54	1.51
Pink noise		**67.64**	60.58	56.54
	Composite Signal	1.83	**2.03**	1.6
	Composite Background	**1.98**	1.89	1.85
	Composite Overall	1.66	**1.67**	1.43
	STOI	0.7	0.73	**0.73**
	LLR	1.64	1.4	**1.8**
	Segmental SNR	**-0.03**	-0.78	-1.3
	PESQ	**1.71**	1.5	1.41
Rain		**66.45**	58.58	54.56
	Composite Signal	**2.62**	2.55	2.39
	Composite Background	**2.23**	2.03	2.03
	Composite Overall	**2.21**	2.1	1.98
	STOI	0.78	0.79	**0.81**
	LLR	1.1	1.08	**1.22**
	Segmental SNR	**1.12**	-1.01	-0.79
	PESQ	**2.0**	1.83	1.73

		EM	STDR	Unfiltered
Rainforest		**61.55**	59.94	54.22
	Composite Signal	1.71	**1.97**	1.45
	Composite Background	2.37	**2.51**	2.3
	Composite Overall	1.85	**2.02**	1.71
	STOI	0.85	0.86	**0.87**
	LLR	2.16	2.02	**2.47**
	Segmental SNR	0.9	**1.8**	-0.85
	PESQ	2.12	**2.14**	2.03
Theatre		**48.01**	37.34	35.61
	Composite Signal	2.33	2.41	**2.48**
	Composite Background	1.86	1.91	**1.91**
	Composite Overall	1.87	1.91	**1.95**
	STOI	0.7	0.72	**0.75**
	LLR	1.05	**1.08**	1.02
	Segmental SNR	**-0.55**	-1.06	-1.2
	PESQ	**1.68**	1.61	1.59
Train		**76.75**	61.17	58.01
	Composite Signal	**3.13**	2.93	3.13
	Composite Background	**2.21**	2.05	2.12
	Composite Overall	**2.5**	2.33	2.45
	STOI	0.8	0.79	**0.83**
	LLR	0.63	**0.77**	0.63
	Segmental SNR	**0.3**	-1.44	-0.87
	PESQ	**2.07**	1.9	1.92
Train station		**62.33**	58.3	53.2
	Composite Signal	2.44	2.4	**2.54**
	Composite Background	1.89	1.88	**1.91**
	Composite Overall	1.94	1.9	**1.98**
	STOI	0.72	0.73	**0.76**
	LLR	0.97	**1.06**	0.96
	Segmental SNR	**-0.4**	-1.31	-1.24
	PESQ	**1.72**	1.61	1.61
White noise		**77.05**	63.39	59.49
	Composite Signal	1.3	**1.59**	0.78
	Composite Background	**2.16**	2.14	1.95
	Composite Overall	1.45	**1.54**	1.07
	STOI	0.75	**0.79**	0.78
	LLR	2.3	2.01	**2.73**
	Segmental SNR	**1.06**	0.84	-1.38
	PESQ	**1.75**	1.6	1.46
Wind		**53.95**	45.29	42.85
	Composite Signal	**2.41**	2.41	2.39
	Composite Background	**1.97**	1.92	1.92
	Composite Overall	**1.93**	1.88	1.88
	STOI	0.74	0.75	**0.76**
	LLR	1.05	1.06	**1.09**
	Segmental SNR	**0.1**	-0.65	-0.93
	PESQ	**1.66**	1.53	1.52

Table 3.7: Performance of EM and STDR algorithms on 5 sentences each from 112 speakers embedded in 12 different untrained noise types. Values are averaged over all 660 sentences and 7 different SNRs per noise type. Bold values indicate best performance.

tion 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 milliseconds duration, entirely acausal delays correspond to a central delay of -50 milliseconds, whereas entirely predictive delays correspond to a central delay of +50 milliseconds. We tested three additional delays in the middle of these two extremes. fig. 3.11 shows the results of these experiments using the same performance metrics as before. Here we have plotted the average performance across fifteen 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 below 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).

3.5 Discussion

We developed a novel algorithm for single-microphone noise reduction that performs well on several objective measures, across two noise types, and several signal-to-noise ratios and speaker counts. The spectro-temporal detection-reconstruction (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-temoral 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.

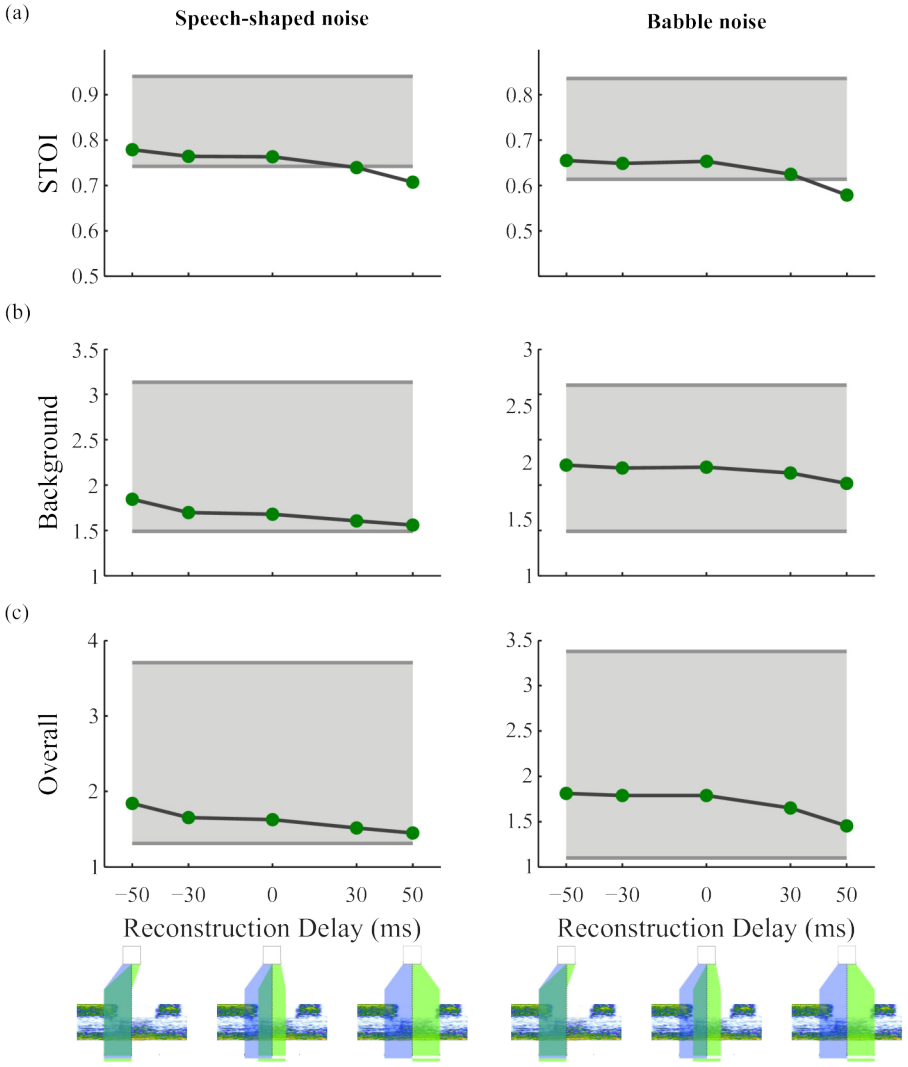

Figure 3.11: The algorithm performed well using time-frequency gains produced from reconstruction kernels with windows ranging from entirely a causal to entirely predictive. (a) Performance using the objective measure for speech intelligibility (STOI). Values on the x-axis correspond to the center time 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 figs. 3.6 and 3.8, 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 8 metrics are shown in table 3.8 and table 3.9.

Delay	-50	-30	0	30	50	Unfiltered
Metric						
Composite Signal	**2.37**	2.17	2.14	2.0	1.92	1.71
Composite Background	**1.84**	1.69	1.68	1.6	1.56	1.49
Composite Overall	**1.84**	1.65	1.62	1.52	1.45	1.31
STOI	**0.78**	0.76	0.76	0.74	0.71	0.74
LLR	**0.97**	1.03	1.05	1.12	1.17	1.32
Segmental SNR	**-0.37**	-0.99	-1.09	-1.51	-1.79	-2.22
PESQ	**1.58**	1.42	1.4	1.33	1.28	1.23
WSS	**74.72**	79.31	79.84	81.22	81.82	84.77

Table 3.8: Performance of the STDR algorithm on sentences from a single speaker embedded in speech-shaped noise at 0 dB SNR for 5 different gain reconstruction kernel center delays, ranging from -50 ms for an entirely acausal reconstruction to +50 ms for an entirely predictive reconstruction.

Delay	-50	-30	0	30	50	Unfiltered
Metric						
Composite Signal	**2.12**	2.1	2.12	1.97	1.76	1.37
Composite Background	**1.72**	1.69	1.7	1.64	1.52	0.99
Composite Overall	**1.81**	1.79	1.79	1.65	1.45	1.1
STOI	**0.75**	0.75	0.75	0.72	0.68	0.71
LLR	1.27	1.27	**1.25**	1.29	1.37	1.43
Segmental SNR	-3.05	-3.3	-3.07	**-2.78**	-3.04	-7.33
PESQ	**1.8**	1.79	1.77	1.66	1.49	1.27
WSS	**83.1**	84.66	84.53	87.93	91.16	112.94

Table 3.9: Performance of the STDR algorithm on sentences from a single speaker embedded in babble noise at 0 dB SNR for 5 different gain reconstruction kernel center delays, ranging from -50 ms for an entirely acausal reconstruction to +50 ms for an entirely predictive reconstruction.

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 (Theunissen and Elie 2014). All natural sounds reside in a restricted subspace of possible spectro-temporal modulations (Singh and Theunissen 2003). 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 (David J. Klein, König, and Kording 2003; Carlson, Ming, and DeWeese 2012). 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 neurons invariance to noise (Moore, Lee, and Theunissen 2013). The slower spectro-temporal modulations present in formants are important features for vowel discrimination (Liberman et al. 1967). They are modified during clear speech to increase speech intelligibility (Amano-Kusumoto and Hosom 2011) and are an interesting target for modern speech enhancement algorithms (Rao and Carney 2014). 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 (Byrne et al. 1994), a part of the spectrum dominated by voiceless consonants (Heinz and Kenneth N Stevens 1961). These slow spectral modulations and fast temporal modulations can be used to discriminate speech or other animal communication signals from environmental sounds (Singh and Theunissen 2003; Sarah M N Woolley, Fremouw, et al. 2005). Spectro-temporal receptive fields 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 (L. M. Miller et al. 2002; Nagel and A. J. Doupe 2008; Sarah M N Woolley, P. R. Gill, et al. 2009) 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 (Smith and Lewicki 2006). 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 Ephraim-Malah algorithm (EM) (Ephraim and Malah 1985), 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 to 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 (Mesgarani 2005; C.-C. Hsu et al. 2015). 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 (Wan and Nelson 1998; Healy et al. 2013; Narayanan and D. L. Wang 2013; Xu et al. 2015). 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 net-

works 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 remains unclear in the literature. In some studies, reconstructing the clean speech spectrogram performed better than attempting to reconstruct the ideal ratio mask (IRM), a closely related metric to the optimal gains used in our study (Xu et al. 2015). However, other studies have found the opposite (Weninger, Hershey, and Roux 2014). 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 (Huang et al. 2014). Independent of the type of output reconstructed (i.e the nature of the objective function), multiple network architectures have been proposed, with autoencoders (Xia and Bao 2014), stacked autoencoders (Lu et al. 2013), deep neural networks (Healy et al. 2013; Narayanan and D. L. Wang 2013; Xu et al. 2015), and deep recurrent neural networks (Weninger, Eyben, and Schuller 2014) as the most common. Our STDR algorithm can best be described as a shallow neural network that operates on a 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 figs. 3.5 and 3.7 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 time frames. In contrast, existing algorithms commonly reconstruct a single frame or time-frequency point using sound from either past frames or several frames centered 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, 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 (Hinton, Osindero, and Teh 2006). 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 (Xu et al. 2015). 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 auto-encoders, 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 behaviorally 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 fig. 3.11, 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 ($\leq 8Hz$) oscillations and power of high gamma oscillations (Mesgarani and Chang 2012; Ding and Simon 2013; Zion Golumbic, Ding, et al. 2013; Ding and Simon 2012) 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 (Zion Golumbic, Ding, et al. 2013). These oscillations may represent the alignment of high-excitability periods with predictions of upcoming

auditory events (Charles E. Schroeder and Lakatos 2009), 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 (Shihab A Shamma, Elhilali, and Micheyl 2011).

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 (Pichora-Fuller 2008; Kalikow, K N Stevens, and L. L. Elliott 1977; G. A. Miller, Heise, and Lichten 1951; Bronkhorst, Bosman, and Smoorenburg 1993). For example, increasing the amount of context in a sentence can increase the intelligibility of the final word in the sentence by nearly 50% (Kalikow, K N Stevens, and L. L. Elliott 1977). More generally predictive coding has been shown to play an essential role for perceptual computations in many sensory modalities (Summerfield and Lange 2014; Clark 2013).

Since 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 feed-forward fashion; they provide no feedback and do not modify the activations of the detection filters in any way. The brain, however, appears to utilize 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. Due 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. Since 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 out performing 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.

Chapter 4

Conclusion

Here we have demonstrated the existence of neurons in a secondary auditory region of the zebra finch that are invariant to the addition of background noise. Their invariance can be partially explained by their tuning to particular spectrotemporal modulations, principally fast spectral modulations and slow temporal modulations. By developing an algorithm that used artifical neurons with similar receptive fields to extract bird song from background noise, we showed they are sufficient to represent a de-noised version of the stimulus. An expanded version of this noise reduction algorithm, still functioning in the domain of spectrotemporal modulations, outperformed optimal frequency filtering when applied to speech from many speakers in a variety of noise conditions. In addition, we found that noisy speech represented in the spectrotemporal feature space can be predictively denoised, enabling numerous real-time applications. These findings, in total, further cement the domain of spectrotemporal modulations as an intermediate level representation crucial for a noise-robust representation of communication sounds and open promising avenues of future research that utilizes prediction for real-time sound processing.

To follow up on a few questions raised in section 3.5, we performed a set of preliminary experiments to test the effects of dataset size, neural network depth, and single timestep reconstruction filters on noise reduction using the full noise dataset from fig. 3.9. We expected that dataset size would have a significant effect, since exposure to a wider variety of training data can only help when generalizing to new noise types and speakers. We also expected

that increasing network depth would provide a large boost. This is because a large network has many more degrees of freedom with which to fit complicated input-output functions. Lastly, we expected that a single timestep reconstruction filter would perform much worse, as the neural network must find a way to encode the joint spectrotemporal structure of the output gains implicitly, rather than allowing the reconstruction filter to explicitly code these dependencies. As these tests were very preliminary, I will only make a few qualitative statements here regarding our findings. Firstly, as expected, using a much larger training dataset (13241 sentences or 14.7 hours) gave much better generalization. However, increasing the network depth by including additional hidden layers had little effect on overall performance. To better understand this lack of improvement, one would have to fully map out the relationship between training set size and network size. It is possible that the dataset used was still not large enough to properly take advantage of the larger network, though I find this unlikely. Further clarification of this is left for future work. Lastly, for the same network size given in chapter 3, a single timestep reconstruction filter showed much degraded performance. To fully understand the necessity of an explicitly *spectrotemporal* reconstruction filter, one would have to check larger network sizes capable of implicitly storing the necessary dependencies. This is also left as future work.

Neural processing for auditory scene analysis

Several studies have explored noise invariance and auditory scene analysis in the brain in the years since the completion of chapter 2. Though much of this work is described in section 3.2, I will provide a brief summary here. The research has progressed along two main lines: processing by single neurons at different stages in the auditory pathway and attentionally-modulated entrainment of oscillations in the human brain to a particular speaker in a multi-speaker environment. In mammals, single units in primary auditory cortex show increased invariance to background noise over units in the inferior colliculus or auditory nerve (Rabinowitz et al. 2013; Mesgarani, Stephen V. David, et al. 2014). These changes are due primarily to increasing levels of short-term firing rate adaptation and secondarily to increased levels

of response gain normalization. Interestingly, it appears that the increase in noise-invariance does not appear until secondary auditory regions in the avian brain, as neurons in Field L show similar levels of corruption to noise as neurons in MLD, the IC analogue (Schneider and Sarah M N Woolley 2013; Narayan et al. 2007). The work by Schneider and Sarah M N Woolley 2013 replicated our finding of noise invariant neurons in NCM, extending it with a simple circuit model where the sparse noise invariant neurons receive excitatory drive from Field L and slower inhibitory input from interneurons in NCM. Whether such a circuit could underly the slow temporal modulation tuning we found to be important remains an open question.

Work in humans has also developed considerably in recent years. Using MEG, EEG, and ECog, it has been shown repeatedly that low frequency phase and high frequency amplitude modulations in cortical oscillations entrain to the envelope of the attended speaker in a multi-speaker stimulus (Mesgarani and Chang 2012; Zion Golumbic, Ding, et al. 2013; Ding and Simon 2012; Ding and Simon 2013; O'Sullivan et al. 2014). By decoding from these frequency bands, researchers have discovered many details about the information stored in different regions and the action of attention on the representation of a single sound source. The entrainment can be used to decode the amplitude envelope (eg. Zion Golumbic, Ding, et al. 2013) or full spectrogram of the attended speaker (eg. Mesgarani and Chang 2012). Early in auditory cortex, attention acts as a gain that increases the representation of the attended source and decreases the representation of the unattended source. In later auditory areas, the representation becomes entirely selective for the target sound source (Zion Golumbic, Ding, et al. 2013). The representation of the target sound source is independent of the target-to-masker ratio, further indicating that multiple sound sources are represented separately (Ding and Simon 2012). Though the effect is often small, it is reliable at the level of a single trial (O'Sullivan et al. 2014), which opens the tantalizing possibility for real-time tracking of attention allocation for medical applications (Van Eyndhoven, Francart, and Bertrand 2016).

Taken together, future research will hopefully bridge the gap between these two domains. It has already been postulated that such entrainment of oscillations could play a functional role in sensory selection during attention

(Lakatos et al. 2008; Charles E. Schroeder and Lakatos 2009; Charles E Schroeder et al. 2010), but how this impacts single neurons and how the representation of a single sound source is built up such that selection can occur in the first place remains an exciting avenue of research. From the work in chapter 2 and the work reviewed earlier, we know that bottom-up processing and the efficient encoding of natural sounds provides a jumping off point for active attentional selection. Further, we know that many of the principles presumed to underlie auditory scene analysis, at least for simple tone stimuli, arise as early as the cochlear nucleus (Pressnitzer et al. 2008). If I may speculate: this suggests that low-level features could be used early on in auditory processing to provide an initial segregation of the auditory scene into proposed sources. Then, given a small amount of early segregation, top-down processing effected through entrainment to a particular sound source could reinforce the representation of a single target sound. Studying this process requires a better understanding of how cortical entrainment builds up following the onset of a particular sound stream. Thus far, the temporal evolution of this process has not been studied, but I would hypothesize that entrainment of cortical oscillations, neural response invariance, and phase-locking of spikes would all increase during this initial time period.

Continued applications for frontend noise reduction

Noise reduction continues to be a valuable technology for frontend signal processing. It already shows significant promise in aiding individuals with hearing aids by easing listening effort (Sarampalis et al. 2009), though there is much room for improvement (Kochkin 2010). Recent efforts using deep neural networks provided the first significant increases in speech intelligibility in both normal hearing and hearing-impaired individuals (Healy et al. 2013). It can also provide significant improvements when used prior to automatic speech recognition. Indeed, there is a lot of recent work that has tried to fuse speech enhancement with ASR (Du et al. 2016; Xiaofei Wang et al. 2015). The best new algorithms take advantage of the continued growth of artificial neural networks to solve complicated regression and classification tasks. Some of this work was reviewed in section 3.5. More recently, al-

gorithms have been developed that combine multiple deep neural networks to tackle different aspects of the speech enhancement problem by, for instance, using one network for voice activity detection and another for clean speech reconstruction (Gao et al. 2015). Other efforts have begun to develop more specialized deep neural networks that function best on particular sets of stimuli and combine their estimates to provide good generalization to novel stimuli (Du et al. 2016). While such efforts work well for offline speech enhancement, they face significant challenges for implementation in real-time, low-power situations like those faced in hearing aids. The continued growth of neural networks in low-power environments such as cell phones provides some hope in this respect. Yet, in approaching the problem of real-time processing, the brain's reliance on temporal context and the immense amount of structure in natural sounds strongly suggests that temporal prediction should be explored much further in future noise reduction algorithms.

To this end, one avenue we hope to explore in continuing the research laid out here is to use the predictive gains as a method of pre-filtering the incoming noisy spectrogram. This would allow more robust pattern recognition in low SNR environments. The desired clean speech spectrogram could then also be used as a training signal, asking the algorithm to reconstruct clean speech from filtered noisy speech. These two signals would provide both discriminative and generative teacher signals, as was recently found to be beneficial (Weninger, Hershey, and Roux 2014).

Auditory scene analysis will remain an interesting problem for years to come because it is at the intersection of so many interesting and challenging questions. It merges the basic building blocks of auditory perception with impressive top-down cognitive processing, integrating grouping effects that begin at the auditory nerve with high-level concepts of speech comprehension and music appreciation. A better understanding at every level could improve the quality of life for individuals with hearing deficits and irrevocably change the way we interact with technology around us.

Bibliography

Alcántara, José L et al. (2003). *Evaluation of the noise reduction system in a commercial digital hearing aid.* DOI: 12564514. URL: http://www.ncbi.nlm.nih.gov/pubmed/12564514.

Amano-Kusumoto, Akiko and John-paul Hosom (2011). "A review of research on speech intelligibility and correlations with acoustic features". In: *Science*, pp. 1–16.

Aubin, Thierry and Pierre Jouventin (2002). "How to vocally identify kin in a crowd: the penguin model". In: *Advances in the Study of Behavior* 31. Ed. by Jay S. Rosenblatt Charles T. Snowdon Peter J. B. Slater and Timothy J. Roper, pp. 243–277. ISSN: 00653454. DOI: 10.1016/S0065-3454(02)80010-9. URL: http://www.sciencedirect.com/science/article/pii/S0065345402800109%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorization/login?request%7B%5C_%7Did=CEA47994-6FBF-11E4-B203-E307DB31B14D.

Bee, Mark A. and Christophe Micheyl (2008). "The "Cocktail Party Problem": What Is It? How Can It Be Solved? And Why Should Animal Behaviorists Study It?" In: *Journal of Comparative Psychology* 122.3, pp. 235–251. ISSN: 0735-7036. DOI: 10.1037/0735-7036.122.3.235. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2692487/%20http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2692487/pdf/nihms-112262.pdf.

Bee, Mark A., Christophe Micheyl, et al. (2010). "Neural adaptation to tone sequences in the songbird forebrain: Patterns, determinants, and relation to the build-up of auditory streaming". In: *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 196.8, pp. 543–557. ISSN: 03407594. DOI: 10.1007/s00359-010-0542-4.

Bendixen, Alexandra (2014). "Predictability effects in auditory scene analysis: A review". In: *Frontiers in Neuroscience* 8.8 MAR, pp. 1–16. ISSN: 1662453X. DOI: 10.3389/fnins.2014.00060.

Bendor, Daniel and Xiaoqin Wang (2005). "The neuronal representation of pitch in primate auditory cortex". In: *Nature* 436.7054, pp. 1161–1165. ISSN: 1476-4687. DOI: 10.1038/nature03867. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1038/nature03867.

Benney, Kristen Stoll and Richard F. Braaten (2000). "Auditory scene analysis in Estrildid finches (Taeniopygia guttata and striata domestica): A species advantage for detection

of conspecific song." In: *Journal of Comparative Psychology* 114.2, p. 174. URL: http://psycnet.apa.org/journals/com/114/2/174/.

Bentler, Ruth et al. (2008). "Digital noise reduction: outcomes from laboratory and field studies." In: *International journal of audiology* 47.8, pp. 447–60. ISSN: 1708-8186. DOI: 10.1080/14992020802033091.

Billimoria, Cyrus P. et al. (2008). "Invariance and Sensitivity to Intensity in Neural Discrimination of Natural Sounds". en. In: *Journal of Neuroscience* 28.25, pp. 6304–6308. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.0961-08.2008. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0961-08.2008%20http://www.jneurosci.org/content/28/25/6304%20http://www.ncbi.nlm.nih.gov/pubmed/18562600.

Bolhuis, J J et al. (2000). "Localized neuronal activation in the zebra finch brain is related to the strength of song learning." In: *Proceedings of the National Academy of Sciences* 97.5, pp. 2282–2285. ISSN: 0027-8424. DOI: 10.1073/pnas.030539097.

Boll, S. (1979). "Suppression of acoustic noise in speech using spectral subtraction". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27.2, pp. 113–120. ISSN: 0096-3518. DOI: 10.1109/TASSP.1979.1163209. URL: http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=1163209%20http://ieeexplore.ieee.org/Xplore/cookiedetectresponse.jsp?reload=true%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorization/login?request%7B%5C_%7Did=76B6BF96-9772-11E4-BB4F-C21CDB31B14D.

Borst, Alexander and Frédéric E. Theunissen (1999). "Information theory and neural coding". In: *Nature neuroscience* 2.11, pp. 947–957. ISSN: 1097-6256. DOI: 10.1038/14731. URL: http://www.ncbi.nlm.nih.gov/pubmed/10526332%20http://www.nature.com/neuro/journal/v2/n11/abs/nn1199%7B%5C_%7D947.html.

Bradlow, Ann R et al. (2011). "Language- and Talker-Dependent Variation in Global Features of Native and Non-Native Speech." ENG. In: *Proceedings of the International Congress of Phonetic Sciences*. Vol. 17. Hong Kong: International Congress of Phonetic Sciences, pp. 356–359. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3594809%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Bronkhorst, Adelbert W. (2000). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions". In: *Acta Acustica united with Acustica* 86.1, pp. 117–128. URL: http://www.ingentaconnect.com/content/dav/aaua/2000/00000086/00000001/art00016.

Bronkhorst, Adelbert W., Arjan J. Bosman, and Guido F. Smoorenburg (1993). "A model for context effects in speech recognition." In: *Journal of the Acoustical Society of America* 93.1, pp. 499–509. ISSN: 0001-4966.

Brookes, Mike (2002). *Voicebox: Speech Processing Toolbox for Matlab.*

Byrne, Denis et al. (1994). "An international comparison of long-term average speech spectra". In: *Journal of the Acoustical Society of America* 96.4, pp. 2108–2120. ISSN: 00014966. DOI: 10.1121/1.410152. URL: http://scitation.aip.org/content/asa/journal/

jasa/96/4/10.1121/1.410152%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorizat login?request%7B%5C_%7Did=BDFB47DE-95C7-11E4-93DE-08EEDA31B14D.

Carlson, Nicole L., Vivienne L. Ming, and Michael Robert DeWeese (2012). "Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus". In: *PLoS Computational Biology* 8.7, pp. 1–15. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002594. arXiv: 1209.5029.

Chabot-Leclerc, Alexandre, Søren Jørgensen, and Torsten Dau (2014). "The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction." In: *Journal of the Acoustical Society of America* 135.6, pp. 3502–12. ISSN: 1520-8524. DOI: 10.1121/1.4873517. URL: http://www.ncbi.nlm.nih.gov/pubmed/24907813.

Chi, Taishih, Powen Ru, and Shihab A Shamma (2005). "Multiresolution spectrotemporal analysis of complex sounds". In: *The Journal of the Acoustical Society of* 118, p. 887. URL: http://link.aip.org/link/?JASMAN/118/887/1.

Clark, Andy (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science." In: *The Behavioral and brain sciences* 36.3, pp. 181–204. ISSN: 1469-1825. DOI: 10.1017/S0140525X12000477. URL: http://www.ncbi.nlm.nih.gov/pubmed/23663408.

Dean, David et al. (2010). "The QUT-NOISE-TIMIT Corpus for the Evaluation of Voice Activity Detection Algorithms". In: *Interspeech*. 26-30 September. Makuhari, pp. 6–8.

Dean, Isabel, Nicol S Harper, and David McAlpine (2005). "Neural population coding of sound level adapts to stimulus statistics". In: *Nature neuroscience* 8.12, pp. 1684–1689. ISSN: 1097-6256. DOI: 10.1038/nn1541. URL: http://www.nature.com/neuro/journal/v8/n12/abs/nn1541.html%20http://www.ncbi.nlm.nih.gov/pubmed/16286934.

Depireux, D.A. et al. (2001). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex". In: *Journal of Neurophysiology* 85.3, p. 1220. URL: http://jn.physiology.org/content/85/3/1220.short.

DiGiovanni, Jeffrey J., Erin A. Davlin, and Naveen K. Nagaraj (2011). "Effects of Transient Noise Reduction Algorithms on Speech Intelligibility and Ratings of Hearing Aid Users". In: *American journal of audiology* 20.2, pp. 140–150. URL: http://perspectives.pubs.asha.org/article.aspx?articleid=1781670%20http://aja.pubs.asha.org/article.aspx?articleid=1781670.

Ding, Nai and Jonathan Z Simon (2012). "Emergence of neural encoding of auditory objects while listening to competing speakers". In: *Proceedings of the National Academy of Sciences* 109.29, pp. 11854–11859. ISSN: 0027-8424. DOI: 10.1073/pnas.1205381109.

— (2013). "Adaptive temporal encoding leads to a background-insensitive cortical representation of speech." In: *Journal of Neuroscience* 33.13, pp. 5728–35. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.5297-12.2013. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3643795%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%20http://www.ncbi.nlm.nih.gov/pubmed/23536086.

Du, Jun et al. (2016). "A Regression Approach to Single-Channel Speech Neural Networks". In: *IEEE Trans on Audio, Speech, and Language Processing* 24.8, pp. 1424–1437.

Dubbelboer, Finn and Tammo Houtgast (2008). "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility." In: *Journal of the Acoustical Society of America* 124.6, pp. 3937–3946. ISSN: 00014966. DOI: 10.1121/1.3001713. URL: http://www.ncbi.nlm.nih.gov/pubmed/19206818.

Edwards, Brent (2004). "Hearing Aids and Hearing Impairment". In: *Speech processing in the auditory system*. Ed. by Steven Greenberg et al. Springer. Chap. 7, pp. 339–421.

Eggermont, J. J. et al. (1981). "Spectro-temporal characterization auditory neurons: Redundant or necessary?" In: *Hearing Research* 5.1, pp. 109–121. ISSN: 03785955. DOI: 10.1016/0378-5955(81)90030-7.

Elhilali, Mounya, Taishih Chi, and Shihab a. Shamma (2003). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility". In: *Speech Communication* 41.2-3, pp. 331–348. ISSN: 01676393. DOI: 10.1016/S0167-6393(02)00134-6. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167639302001346.

Elliott, Taffeta M. and Frédéric E. Theunissen (2009). "The modulation transfer function for speech intelligibility". In: *PLoS Computational Biology* 5.3, e1000302. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000302. URL: http://dx.plos.org/10.1371/journal.pcbi.1000302%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2639724%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Ellis, Daniel P W (1999). "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures". In: *Speech Communication* 27.3, pp. 281–298. ISSN: 01676393. DOI: 10.1016/S0167-6393(98)00083-1. URL: http://www.sciencedirect.com/science/article/pii/S0167639398000831.

Ellis, Daniel PW W (1996). "Prediction-driven computational auditory scene analysis". PhD thesis. Massachusetts Institute of Technology. URL: http://sound.media.mit.edu/Papers/dpwe-phd-pdcasa%7B%5C_%7Dexternal.pdf.

Ephraim, Y. and D. Malah (1984). "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.6. ISSN: 0096-3518. DOI: 10.1109/TASSP.1984.1164453. URL: http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=1164453.

— (1985). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33.2, pp. 443–445. ISSN: 0096-3518. DOI: 10.1109/TASSP.1985.1164550. URL: http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=1164550%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorization/login?request%7B%5C_%7Did=3B78B8A6-95DF-11E4-841E-C922DB31B14D.

Escabí, Monty A. et al. (2003). "Naturalistic Auditory Contrast Improves Spectrotemporal Coding in the Cat Inferior Colliculus". en. In: *Journal of Neuroscience* 23.37, pp. 11489–11504. ISSN: 0270-6474, 1529-2401. DOI: 23/37/11489[pii]. URL: http://www.jneurosci.org/content/23/37/11489%20http://www.ncbi.nlm.nih.gov/pubmed/14684853.

Fay, Richard R. (2008). "Auditory Scene Analysis". In: *Bioacoustics* 17.1-3, pp. 106–109. ISSN: 0952-4622. DOI: 10.1080/09524622.2008.9753783. URL: http://www.tandfonline.com/doi/abs/10.1080/09524622.2008.9753783.

Flanagan, J L (1980). "Parametric coding of speech spectra". In: *Journal of the Acoustical Society of America* 68.2, pp. 412–419.

Fortune, E. S. and D. Margoliash (1992). "Cytoarchitectonic organization and morphology of cells of the field L complex in male zebra finches (Taenopygia guttata)". In: *Journal of Comparative Neurology* 325.3, pp. 388–404. ISSN: 00219967. DOI: 10.1002/cne.903250306.

Freiwald, Winrich A. and Doris Y. Tsao (2010). "Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System". In: *Science* 20.2, pp. 172–176. DOI: 10.1016/j.conb.2010.02.010.Dynamics.

Gao, Tian et al. (2015). "Improving Deep Neural Network Based Speech Enhancement in Low SNR Environments". In: *Latent Variable Analysis and Signal Separation*, pp. 75–82.

Gill, Patrick, Sarah M N Woolley, et al. (2008). "What's that sound? Auditory area CLM encodes stimulus surprise, not intensity or intensity changes." In: *Journal of neurophysiology* 99.6, pp. 2809–2820. ISSN: 0022-3077. DOI: 10.1152/jn.90270.2008. URL: http://www.ncbi.nlm.nih.gov/pubmed/18287545.

Gill, Patrick, Junli Zhang, et al. (2006). "Sound representation methods for spectro-temporal receptive field estimation." In: *Journal of computational neuroscience* 21.1, pp. 5–20. ISSN: 0929-5313. DOI: 10.1007/s10827-006-7059-4. URL: http://link.springer.com/10.1007/s10827-006-7059-4%20http://www.ncbi.nlm.nih.gov/pubmed/16633939.

Glorot, Xavier and Yoshua Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 9. Sardinia, pp. 249–256. DOI: 10.1.1.207.2059. URL: http://machinelearning.wustl.edu/mlpapers/paper%7B%5C_%7Dfiles/AISTATS2010%7B%5C_%7DGlorotB10.pdf.

Griffin, D. and Jae Lim (1984). "Signal estimation from modified short-time Fourier transform". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2, pp. 236–243. ISSN: 0096-3518. DOI: 10.1109/TASSP.1984.1164317. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1164317.

Hannemann, R., J. Obleser, and C. Eulitz (2007). "Top-down knowledge supports the retrieval of lexical information from degraded speech". In: *Brain Research* 1153.1, pp. 134–143. ISSN: 00068993. DOI: 10.1016/j.brainres.2007.03.069.

Hansen, John H L and Bryan L Pellom (1998). "An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms". In: *Proc. Int. Conf. on Spoken Language Processing (ICSLP)1*. Sydney, pp. 2819–2822.

Healy, Eric W et al. (2013). "An algorithm to improve speech recognition in noise for hearing-impaired listeners." In: *Journal of the Acoustical Society of America* 134.4, pp. 3029–38. ISSN: 1520-8524. DOI: 10.1121/1.4820893. URL: http://www.ncbi.nlm.nih.gov/pubmed/24116438.

Heinz, John M and Kenneth N Stevens (1961). *On the Properties of Voiceless Fricative Consonants*. DOI: 10.1121/1.1908734.

Hermus, Kris (2007). "A review of signal subspace speech enhancement and its application to noise robust speech recognition". In: *Eurasip Journal on Advances in Signal Processing* 2007.Mv. ISSN: 11108657. DOI: 10.1155/2007/45821.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh (2006). "A fast learning algorithm for deep belief nets". In: *Neural computation* 18.7, pp. 1527–1554.

Hinton, Geoffrey E. and R. R. Salakhutdinov (2006). "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786, pp. 504–507. ISSN: 0036-8075. DOI: 10.1126/science.1127647.

Holdgraf, Chris et al. "Rapid tuning adaptation in human auditory cortex enhances speech intelligibility".

Hsu, Anne, Alexander Borst, and Frédéric E. Theunissen (2004). "Quantifying variability in neural responses and its application for the validation of model predictions". In: *Network: Computation in Neural Systems* 10.2, pp. 123–32. DOI: 10.1088/0954-898X.

Hsu, Anne, Sarah M N Woolley, et al. (2004). "Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons." In: *Journal of Neuroscience* 24.41, pp. 9201–11. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.2449-04.2004. URL: http://www.ncbi.nlm.nih.gov/pubmed/15483139.

Hsu, Chung-Chien et al. (2015). "Modulation Wiener filter for improving speech intelligibility". In: *ICASSP '15. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, pp. 370–374. ISBN: 9781467369978.

Hu, Yi and Philipos C. Loizou (2008). "Evaluation of objective quality measures for speech enhancement". In: *IEEE Transactions on Audio, Speech and Language Processing* 16.1, pp. 229–238. ISSN: 15587916. DOI: 10.1109/TASL.2007.911054. URL: http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=4389058%20http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=4389058%7B%5C&%7Dtag=1.

Huang, Po-Sen et al. (2014). "Deep learning for monaural speech separation". In: *ICASSP '14. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1562–1566.

Hulse, S H (2002). "Auditory scene analysis in animal communication". In: *Advances in the Study of Behavior* 31, pp. 163–200. DOI: 10.1016/s0065-3454(02)80008-0. URL: http://www.sciencedirect.com/science/article/pii/S0065345402800080.

Kalikow, D. N., K N Stevens, and L L Elliott (1977). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability". In: *Journal of the Acoustical Society of America* 61.5, pp. 1337–1351. ISSN: 00014966. DOI: 10.1121/1.381436. URL: http://scitation.aip.org/content/asa/journal/jasa/61/5/10.1121/1.381436.

Kim, Gunsoo and Allison Doupe (2011). "Organized representation of spectrotemporal features in songbird auditory forebrain." In: *Journal of Neuroscience* 31.47, pp. 16977–90. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.2003-11.2011. URL: http://www.ncbi.nlm.nih.gov/pubmed/22114268.

Klatt, Dennis (1982). "Prediction of perceived phonetic distance from critical-band spectra: a first step". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.* IEEE, pp. 1278–1281.

Klein, David J., Peter König, and Konrad P Kording (2003). "Sparse spectrotemporal coding of sounds". In: *EURASIP Journal on Applied Signal Processing* 7, pp. 659–667. URL: http://portal.acm.org/citation.cfm?id=1283319.

Klein, David J et al. (2006). *Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex.* Tech. rep. 2. DTIC Document, pp. 111–136. DOI: 10.1007/s10827-005-3589-4. arXiv: 0508039 [q-bio]. URL: http://oai.dtic.mil/oai/oai?verb=getRecord%7B%5C&%7DmetadataPrefix=html%7B%5C&%7Didentifier=ADA438561%20http://link.springer.com/article/10.1007/s10827-005-3589-4.

Knudsen, Daniel P. and Timothy Q. Gentner (2010). "Mechanisms of song perception in oscine birds." In: *Brain and language* 115.1, pp. 59–68. ISSN: 1090-2155. DOI: 10.1016/j.bandl.2009.09.008. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2932808%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%20http://dx.doi.org/10.1016/j.bandl.2009.09.008.

Kochkin, S (2010). "MarkeTrak VIII: Consumer satisfaction with hearing aids is slowly increasing". In: *The Hearing Journal* 63.1, pp. 19–27. ISSN: 07457472. DOI: 10.1097/01.HJ.0000366912.40173.76. URL: http://journals.lww.com/thehearingjournal/Fulltext/2010/01000/MarkeTrak%7B%5C_%7DVIII%7B%5C_%7D%7B%5C_%7DConsumer%7B%5C_%7Dsatisfaction%7B%5C_%7Dwith%7B%5C_%7Dhearing.4.asp.

Lakatos, Peter et al. (2008). "Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection". In: *Science* 320.5872, pp. 110–113. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1154735. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1154735.

Lee, Tyler P. and Frédéric E. Theunissen (2015). "A single microphone noise reduction algorithm based on the detection and reconstruction of spectro-temporal features". In: *Proceedings of the Royal Society A* 471.2184, p. 20150309. ISSN: 1364-5021. DOI: 10.1098/rspa.2015.0309. URL: http://dx.doi.org/10.1098/rspa.2015.0309.

Lengagne, T. et al. (1999). "How do king penguins (Aptenodytes patagonicus apply the mathematical theory of information to communicate in windy conditions?" en. In: *Proceedings of the Royal Society of London B: Biological Sciences* 266.1429, pp. 1623–1628. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.1999.0824. URL: http://rspb.royalsocietypublishing.org/content/266/1429/1623%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorization/login?request%7B%5C_%7Did=0CFA6970-6FBF-11E4-BEF4-B228DB31B14D.

Lewicki, Michael S. et al. (2014). "Scene analysis in the natural environment". In: *Frontiers in Psychology* 5.APR, pp. 1–21. ISSN: 16641078. DOI: 10.3389/fpsyg.2014.00199.

Li, Jinyu et al. (2014). "An overview of noise-robust automatic speech recognition". In: *IEEE Transactions on Audio, Speech and Language Processing* 22.4, pp. 745–777. ISSN: 15587916. DOI: 10.1109/TASLP.2014.2304637.

Li, Yipeng and DeLiang L. Wang (2009). "On the optimality of ideal binary time-frequency masks". In: *Speech Communication* 51.3, pp. 230–239. ISSN: 01676393. DOI: 10.1016/j.specom.2008.09.001. URL: http://linkinghub.elsevier.com/retrieve/pii/S0167639308001325%20http://dx.doi.org/10.1016/j.specom.2008.09.001.

Liberman, a M et al. (1967). *Perception of the speech code.* DOI: 10.1037/h0020279.

Litovsky, Ruth Y (2005). "Speech intelligibility and spatial release from masking in young children." In: *Journal of the Acoustical Society of America* 117.5, pp. 3091–3099. ISSN: 00014966. DOI: 10.1121/1.1873913. URL: http://scitation.aip.org/content/asa/journal/jasa/117/5/10.1121/1.1873913%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorization/login?request%7B%5C_%7Did=E47B323A-976F-11E4-8860-C022DB31B14D.

Lu, Xugang et al. (2013). "Speech Enhancement Based on Deep Denoising Autoencoder". In: *Interspeech*. August 25-29. Lyon, pp. 436–440.

Luo, Huan and David Poeppel (2007). "Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex". In: *Neuron* 54.6, pp. 1001–1010. ISSN: 08966273. DOI: 10.1016/j.neuron.2007.06.004. URL: http://linkinghub.elsevier.com/retrieve/pii/S0896627307004138.

Luts, H et al. (2010). "Multicenter evaluation of signal enhancement algorithms for hearing aids". In: *Journal of the Acoustical Society of America* 127.3, pp. 1491–1505. ISSN: 00014966 (ISSN). DOI: 10.1121/1.3299168. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve%7B%5C&%7Ddb=PubMed%7B%5C&%7Ddopt=Citation%7B%5C&%7Dlist%7B%5C_%7Duids=20329849.

Lyon, R. (1982). "A computational model of filtering, detection, and compression in the cochlea". In: *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Paris, pp. 1282–1285. DOI: 10.1109/ICASSP.1982.1171644.

MacDougall-Shackleton, S a et al. (1998). "Auditory scene analysis by European starlings (Sturnus vulgaris): perceptual segregation of tone sequences." In: *Journal of the Acoustical Society of America* 103.6, pp. 3581–7. ISSN: 0001-4966. URL: http://www.ncbi.nlm.nih.gov/pubmed/9637040.

Martin, Rainer (1994). "Spectral Subtraction Based on Minimum Statistics". In: *Proc. EUSIPCO 94*. 1. Edinburgh, pp. 1182–1185.

— (2001). "Noise power spectral density estimation based on optimal smoothing and minimum statistics". In: *IEEE Transactions on Speech and Audio Processing* 9.5, pp. 504–512. ISSN: 10636676. DOI: 10.1109/89.928915. URL: http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=928915.

McAulay, R. and M. Malpass (1980). "Speech enhancement using a soft-decision noise suppression filter". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.2, pp. 137–145. ISSN: 0096-3518. DOI: 10.1109/TASSP.1980.1163394.

McDermott, Josh H, David Wrobleski, and Andrew J Oxenham (2011). "Recovering sound sources from embedded repetition". In: *Proceedings of the National Academy of Sciences* 108.3, pp. 1188–1193. ISSN: 1091-6490. DOI: 10.1073/pnas.1004765108. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3024660%7B%5C&

%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%20http://www.pnas.org/content/108/3/1188.short.

Mello, Claudio V., F Nottebohm, and D Clayton (1995). "Repeated exposure to one song leads to a rapid and persistent decline in an immediate early gene's response to that song in zebra finch telencephalon." In: *Journal of Neuroscience* 15.October, pp. 6919–6925. ISSN: 0270-6474.

Mesgarani, Nima (2005). "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations". Master of Science. University of Maryland, College Park.

— (2008). *Representation of speech in the primary auditory cortex and its implications for robust speech processing*. ProQuest. URL: http://books.google.com/books?hl=en%7B%5C&%7Dlr=%7B%5C&%7Did=8YBlejp%7B%5C_%7DIgIC%7B%5C&%7Doi=fnd%7B%5C&%7Dpg=PR2%7B%5C&%7Ddq=%7B%5C%%7D22and+then+a+lowpass+%7B%5C%%7DEF%7B%5C%%7DAC%7B%5C%%7D81lter+(hair+cell+membrane+leakage).+(3)+Finally,+a%7B%5C%%7D22+%7B%5C%%7D22which+corresponds+to+a+neuron+that+responds+well+to+a+ripple+of+4Hz+rate%7B%5C%%7D22+%7B%5C&%7Dots=b7ovWTFlCM%7B%5C&%7Dsig.

Mesgarani, Nima and Edward F. Chang (2012). "Selective cortical representation of attended speaker in multi-talker speech perception". In: *Nature* 485.7397, pp. 233–236. ISSN: 0028-0836. DOI: 10.1038/nature11020. URL: http://www.nature.com/doifinder/10.1038/nature11020.

Mesgarani, Nima, Stephen V. David, et al. (2014). "Mechanisms of noise robust representation of speech in primary auditory cortex". In: *Proceedings of the National Academy of Sciences* 111.18, pp. 6792–6797. ISSN: 00014966. DOI: 10.1121/1.3385147. URL: http://www.pnas.org/content/111/18/6792.short.

Mesgarani, Nima, Malcolm Slaney, and Shihab A. Shamma (2006). "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations". In: *IEEE Transactions on Audio, Speech and Language Processing* 14.3, pp. 920–930. ISSN: 15587916. DOI: 10.1109/TSA.2005.858055.

Miller, George A., George A. Heise, and William Lichten (1951). "The intelligibility of speech as a function of the context of the test materials." In: *Journal of experimental psychology* 41.5, pp. 329–335. ISSN: 0022-1015. DOI: 10.1037/h0062491.

Miller, Lee M. et al. (2002). "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex." en. In: *Journal of Neurophysiology* 87.1, pp. 516–527. ISSN: 0022-3077. DOI: 10.1152/jn.00395.2001. URL: http://jn.physiology.org/content/87/1/516%20http://www.ncbi.nlm.nih.gov/pubmed/11784767%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorization/login?request%7B%5C_%7Did=36834F30-9771-11E4-9A6C-C6F0DA31B14D.

Moore, R. Channing, Tyler P. Lee, and Frédéric E. Theunissen (2013). "Noise-invariant Neurons in the Avian Auditory Cortex: Hearing the Song in Noise". In: *PLoS Computational Biology* 9.3. Ed. by Konrad P. Kording, e1002942. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002942. URL: http://dx.doi.org/10.1371/journal.pcbi.1002942%20http://dx.plos.org/10.1371/journal.pcbi.1002942.g007%20http://dx.plos.org/10.1371/journal.pcbi.1002942.

Nagel, Katherine I and Allison J Doupe (2008). "Organizing principles of spectro-temporal encoding in the avian primary auditory area field L." In: *Neuron* 58.6, pp. 938–55. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2008.04.028. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2547416%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Narayan, Rajiv et al. (2007). "Cortical interference effects in the cocktail party problem". In: *Nature Neuroscience* 10.12, pp. 1601–1607. ISSN: 1097-6256. DOI: 10.1038/nn2009. URL: http://www.nature.com/doifinder/10.1038/nn2009%20http://www.ncbi.nlm.nih.gov/pubmed/17994016.

Narayanan, Arun and DeLiang L. Wang (2013). "Ideal Ratio Mask Estimation using Deep Neural Networks for Robust Speech Recognition". In: *ICASSP '13. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7092–7096.

Nilsson, M, S D Soli, and J a Sullivan (1994). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise." In: *Journal of the Acoustical Society of America* 95.2, pp. 1085–1099. ISSN: 00014966. DOI: 10.1121/1.408469. URL: http://www.ncbi.nlm.nih.gov/pubmed/8132902.

O'Sullivan, James A. et al. (2014). "Attentional selection in a cocktail party environment can be decoded from single-trial EEG". In: *Cerebral Cortex*, bht355. ISSN: 1460-2199. DOI: 10.1093/cercor/bht355. URL: http://cercor.oxfordjournals.org/content/early/2014/01/14/cercor.bht355.short%20http://www.ncbi.nlm.nih.gov/pubmed/24429136.

Palmer, Catherine V. (2009). "A Contemporary Review of Hearing Aids". en. In: *The Laryngoscope* 119.11, pp. 2195–2204. ISSN: 1531-4995. DOI: 10.1002/lary.20690. URL: http://onlinelibrary.wiley.com/doi/10.1002/lary.20690/abstract%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorization/login?request%7B%5C_%7Did=DC8E832A-9778-11E4-9D3E-E9C1DA31B14D.

Pichora-Fuller, M Kathleen (2008). "Use of supportive context by younger and older adult listeners: balancing bottom-up and top-down information processing." In: *International journal of audiology* 47 Suppl 2, S72–S82. ISSN: 1708-8186. DOI: 10.1080/14992020802307404.

Pinaud, R. et al. (2008). "Inhibitory network interactions shape the auditory processing of natural communication signals in the songbird auditory forebrain." In: *Journal of neurophysiology* 100.1, pp. 441–55. ISSN: 0022-3077. DOI: 10.1152/jn.01239.2007. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2493480%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Pressnitzer, Daniel et al. (2008). "Perceptual organization of sound begins in the auditory periphery." In: *Current biology : CB* 18.15, pp. 1124–8. ISSN: 0960-9822. DOI: 10.1016/j.cub.2008.06.053. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2559912%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Quackenbush, Schuyler R., Thomas Pinkney Barnwell, and Mark A. Clements (1988). *Objective measures of speech quality*. Englewood Cliffs, NJ: Prentice Hall PTR, p. 377.

Rabinowitz, Neil C. et al. (2013). "Constructing Noise-Invariant Representations of Sound in the Auditory Pathway". In: *PLoS Biology* 11.11, e1001710. ISSN: 15449173. DOI: 10.1371/journal.pbio.1001710. URL: http://dx.doi.org/10.1371/journal.pbio.1001710.

Rao, Akshay and Laurel H. Carney (2014). "Speech enhancement for listeners with hearing loss based on a model for vowel coding in the auditory midbrain". In: *IEEE Transactions on Biomedical Engineering* 61.7, pp. 2081–2091. ISSN: 15582531. DOI: 10.1109/TBME.2014.2313618.

Rauschecker, Josef P. et al. (1995). "Processing of complex sounds in the macaque nonprimary auditory cortex". In: *Science* 268.5207, pp. 111–114. URL: http://www.sciencemag.org/content/268/5207/111.short.

Ribeiro, S et al. (1998). "Toward a song code: evidence for a syllabic representation in the canary brain." In: *Neuron* 21.2, pp. 359–71. ISSN: 0896-6273. DOI: DOI:10.1016/S0896-6273(00)80545-0. URL: http://www.ncbi.nlm.nih.gov/pubmed/9728917.

Rix, A.W. et al. (2001). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *ICASSP '01. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. Salt Lake City, UT, pp. 2–5. ISBN: 0-7803-7041-4. DOI: 10.1109/ICASSP.2001.941023.

Rodríguez, Francisco a et al. (2010). "Neural Modulation Tuning Characteristics Scale to Efficiently Encode Natural Sound Statistics". In: *Journal of Neuroscience* 30.47, pp. 15969–15980. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.0966-10.2010. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0966-10.2010%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3351116%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Sadagopan, Srivatsun and Xiaoqin Wang (2008). "Level Invariant Representation of Sounds by Populations of Neurons in Primary Auditory Cortex". en. In: *Journal of Neuroscience* 28.13, pp. 3415–3426. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.2743-07.2008. URL: http://www.jneurosci.org/content/28/13/3415%20http://www.ncbi.nlm.nih.gov/pubmed/18367608.

Sarampalis, Anastasios et al. (2009). "Objective measures of listening effort: effects of background noise and noise reduction." In: *Journal of speech, language, and hearing research : JSLHR* 52.5, pp. 1230–1240. ISSN: 1092-4388. DOI: 10.1044/1092-4388(2009/08-0111).

Schneider, David M and Sarah M N Woolley (2013). "Sparse and Background-Invariant Coding of Vocalizations in Auditory Scenes". In: *Neuron* 79.1, pp. 141–152. ISSN: 08966273. DOI: 10.1016/j.neuron.2013.04.038. URL: http://www.sciencedirect.com/science/article/pii/S0896627313003693%20http://dx.doi.org/10.1016/j.neuron.2013.04.038%20http://www.sciencedirect.com/science/article/pii/S0896627313003693%7B%5C#%7D.

Schroeder, Charles E. and Peter Lakatos (2009). "Low-frequency neuronal oscillations as instruments of sensory selection". In: *Trends in Neurosciences* 32.1, pp. 9–18. ISSN: 01662236. DOI: 10.1016/j.tins.2008.09.012. URL: http://linkinghub.elsevier.com/retrieve/pii/S0166223608002506.

Schroeder, Charles E et al. (2010). "Dynamics of Active Sensing and perceptual selection". In: *Current Opinion in Neurobiology*. Cognitive neuroscience 20.2, pp. 172–176. ISSN: 0959-4388. DOI: 10.1016/j.conb.2010.02.010. URL: http://www.sciencedirect.com/science/article/pii/S0959438810000322%20http://linkinghub.elsevier.com/retrieve/pii/S0959438810000322.

Shamma, Shihab A, Mounya Elhilali, and Christophe Micheyl (2011). "Temporal coherence and attention in auditory scene analysis." In: *Trends in neurosciences* 34.3, pp. 114–23. ISSN: 1878-108X. DOI: 10.1016/j.tins.2010.11.002. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3073558%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%20http://linkinghub.elsevier.com/retrieve/pii/S0166223610001670.

Sharpee, Tatyana O., Craig a. Atencio, and Christoph E. Schreiner (2011). "Hierarchical representations in the auditory cortex". In: *Current Opinion in Neurobiology*. Networks, circuits and computation 21.5, pp. 761–767. ISSN: 09594388. DOI: 10.1016/j.conb.2011.05.027. URL: http://www.ncbi.nlm.nih.gov/pubmed/21704508%20http://www.sciencedirect.com/science/article/pii/S095943881100095X.

Singh, Nandini C and Frédéric E. Theunissen (2003). "Modulation spectra of natural sounds and ethological theories of auditory processing." In: *Journal of the Acoustical Society of America* 114.6 Pt 1, pp. 3394–3411. ISSN: 00014966. DOI: 10.1121/1.1624067. URL: http://link.aip.org/link/JASMAN/v114/i6/p3394/s1%7B%5C&%7DAgg=doi.

Smith, Evan C and Michael S. Lewicki (2006). "Efficient auditory coding." In: *Nature* 439.7079, pp. 978–82. ISSN: 1476-4687. DOI: 10.1038/nature04485. URL: http://www.nature.com/doifinder/10.1038/nature04485%20http://www.ncbi.nlm.nih.gov/pubmed/16495999.

Stern, Richard M. and Nelson Morgan (2012). "Hearing is believing - biologically-inspired feature extraction for robust automatic speech recognition". In: *Signal Processing Magazine, IEEE* 29.6, pp. 34–43.

Stripling, R, S F Volman, and D F Clayton (1997). "Response modulation in the zebra finch neostriatum: relationship to nuclear gene regulation." In: *Journal of Neuroscience* 17.10, pp. 3883–3893. ISSN: 0270-6474.

Summerfield, Christopher and Floris P. de Lange (2014). "Expectation in perceptual decision making: neural and computational mechanisms". In: *Nature Reviews Neuroscience* 15.October, pp. 745–756. ISSN: 1471-003X. DOI: 10.1038/nrn3838. URL: http://dx.doi.org/10.1038/nrn3838.

Taal, Cees H. et al. (2011). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech". In: *IEEE Transactions on Audio, Speech and Language Processing* 19.7, pp. 2125–2136. ISSN: 15587916. DOI: 10.1109/TASL.2011.2114881.

Terleph, Thomas A., Claudio V. Mello, and David S. Vicario (2006). "Auditory topography and temporal response dynamics of canary caudal telencephalon". In: *Journal of Neurobiology* 66.3, pp. 281–292. ISSN: 00223034. DOI: 10.1002/neu.20219.

Theunissen, Frédéric E., Stephen V David, et al. (2001). "Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli".

In: *Network: Computation in Neural Systems* 12.3, pp. 289–316. ISSN: 0954-898X. URL: http://www.ncbi.nlm.nih.gov/pubmed/11563531%20http://informahealthcare.com/doi/abs/10.1080/net.12.3.289.316.

Theunissen, Frédéric E. and Julie E Elie (2014). "Neural processing of natural sounds." In: *Nature Reviews Neuroscience* 15.6, pp. 355–66. ISSN: 1471-0048. DOI: 10.1038/nrn3731. URL: http://www.ncbi.nlm.nih.gov/pubmed/24840800.

Theunissen, Frédéric E., K Sen, and a J Doupe (2000). "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds." In: *Journal of Neuroscience* 20.6, pp. 2315–2331. ISSN: 1529-2401. URL: http://www.ncbi.nlm.nih.gov/pubmed/10704507.

Van Eyndhoven, Simon, Tom Francart, and Alexander Bertrand (2016). "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses". In: pp. 1–12. arXiv: 1602.05702. URL: http://arxiv.org/abs/1602.05702.

Varga, Andrew and Herman J.M. Steeneken (1993). "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems". In: *Speech Communication* 12.3, pp. 247–251. ISSN: 01676393. DOI: 10.1016/0167-6393(93)90095-3. URL: http://www.sciencedirect.com/science/article/pii/0167639393900953%20https://proxy.lib.berkeley.edu/ucblibraryproxyauthorization/login?request%7B%5C_%7Did=75A84F92-A5B0-11E4-9CEB-0722DB31B14D.

Vates, G. Edward et al. (1996). "Auditory pathways of caudal telencephalon and their relation to the song system of adult male zebra finches (Taenopygia guttata)". In: *Journal of Comparative Neurology* 366.4, pp. 613–642. ISSN: 00219967. DOI: 10.1002/(SICI)1096-9861(19960318)366:4<613::AID-CNE5>3.0.CO;2-7.

Vignal, Clémentine, Joël Attia, et al. (2004). "Background noise does not modify song-induced genic activation in the bird brain". In: *Behavioural Brain Research* 153.1, pp. 241–248. ISSN: 01664328. DOI: 10.1016/j.bbr.2003.12.006. URL: http://www.ncbi.nlm.nih.gov/pubmed/15219725%20http://linkinghub.elsevier.com/retrieve/pii/S0166432803004753.

Vignal, Clémentine, Nicolas Mathevon, and Stéphane Mottin (2004). "Audience drives male songbird response to partner's voice." In: *Nature* 430.July, pp. 448–451. ISSN: 0028-0836. DOI: 10.1038/nature02645. arXiv: 0911.3184.

Voss, Richard F and John Clarke (1978). ""1 / f noise" in music: Music from 1 / f noise". In: *Journal of the Acoustical Society of America* 63.1, pp. 258–263.

Wan, E.A. and A.T. Nelson (1998). *Networks for speech enhancement*. Ed. by S Katagiri. 1st ed. Artech House, pp. 1–27. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.9812%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf.

Wang, Xiaofei et al. (2015). "Noise Robust IOA/CAS Speech Separation and Recognition System For The Third 'CHIME' Challenge". In: arXiv: 1509.06103. URL: http://arxiv.org/abs/1509.06103.

Warren, R M (1970). *Perceptual restoration of missing speech sounds.* DOI: 10.1126/science.167.3917.392.

Weninger, Felix, Florian Eyben, and Bjorn Schuller (2014). "Single-channel speech separation with memory-enhanced recurrent neural networks". In: *ICASSP '14. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3709–3713. ISSN: 15206149. DOI: 10.1109/ICASSP.2014.6854294.

Weninger, Felix, John R Hershey, and Jonathan Le Roux (2014). "Discriminatively Trained Recurrent Neural Networks for Single-Channel Speech Separation". In: *GlobalSIP 2014: Machine Learning Applications in Speech Processing*, pp. 577–581.

"WHO global estimates on prevalence of hearing loss" (2012). In: *Mortality and Burden of Diseases and Prevention of Blindness and Deafness*.

Woolley, S. M. N. (2006). "Stimulus-Dependent Auditory Tuning Results in Synchronous Population Coding of Vocalizations in the Songbird Midbrain". In: *Journal of Neuroscience* 26.9, pp. 2499–2512. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3731-05.2006. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3731-05.2006.

Woolley, Sarah M N, Thane E Fremouw, et al. (2005). "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds." In: *Nature neuroscience* 8.10, pp. 1371–1379. ISSN: 1097-6256. DOI: 10.1038/nn1536. URL: http://www.ncbi.nlm.nih.gov/pubmed/16136039.

Woolley, Sarah M N, Patrick R Gill, et al. (2009). "Functional groups in the avian auditory system." In: *Journal of Neuroscience* 29.9, pp. 2780–93. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.2042-08.2009. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2677621%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Xia, Bingyin and Changchun Bao (2014). "Wiener filtering based speech enhancement with Weighted Denoising Auto-encoder and noise classification". In: *Speech Communication* 60, pp. 13–29. ISSN: 01676393. DOI: 10.1016/j.specom.2014.02.001. URL: http://dx.doi.org/10.1016/j.specom.2014.02.001.

Xu, Yong et al. (2015). "A Regression Approach to Speech Enhancement Based on Deep Neural Networks". In: *IEEE Transactions on Audio, Speech and Language Processing* 23.1, pp. 7–19.

Zion Golumbic, Elana M., Nai Ding, et al. (2013). "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"". In: *Neuron* 77.5, pp. 980–991. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.12.037. URL: http://dx.doi.org/10.1016/j.neuron.2012.12.037%20http://linkinghub.elsevier.com/retrieve/pii/S0896627313000457%20http://www.ncbi.nlm.nih.gov/pubmed/23473326.

Zion Golumbic, Elana M., David Poeppel, and Charles E. Schroeder (2012). "Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective". In: *Brain and Language* 122.3, pp. 151–161. ISSN: 0093934X. DOI: 10.1016/j.bandl.2011.12.010. URL: http://www.ncbi.nlm.nih.gov/pubmed/22285024%20http://dx.doi.org/10.1016/j.bandl.2011.12.010.