ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Signal Processing in Acoustics
Session 1pSPc: Miscellaneous Topics in Signal Processing in Acoustics (Poster Session)

1pSPc14. A novel noise-reduction algorithm for real-time speech processing
Frédéric E. Theunissen* and Tyler Lee

*Corresponding author's address: UC Berkeley, UC Berkeley, Berkeley, CA 94720, theunissen@berkeley.edu

We developed a new noise-reduction algorithm based on a joint spectro-temporal representation of signals. The algorithm was inspired by the discovery in our laboratory of higher-level avian auditory cortical neurons that showed invariant responses to communication signals. The algorithm consists of an analysis step and a synthesis step. In the analysis step, the sound is first decomposed into narrow band signals by a frequency filter bank. These time-frequency waveforms are then further analyzed using a spectro-temporal modulation filter bank to obtain a representation that is akin to the one generated by cortical neurons. In our algorithm, this modulation filter bank was obtained from the principal component analysis of the speech signal in the time-frequency representation. We then learned which subset of the modulation filters provided the best information to extract the signal from the noise. In the synthesis step, we then used this subset of spectral-temporal modulation feature detectors to generate a set of time-varying frequency gains that could be applied directly to the original time frequency decomposition. In this manner, we were able to perform noise reduction in real time and with minimal delay. Our algorithm yielded similar noise reduction but better quality speech quality than current state-of-the-art algorithms.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Noise reduction is an essential component of many audio processing technologies. For example, noise reduction is used in automatic voice recognition algorithms used in consumer products[1] as well as in medical devices such as hearing aids or cochlear implants. In hearing aids, various forms of noise reduction have been shown to offer an incremental improvement in the listening experience[2, 3] though listening to speech in noisy environments remains the principal complaint of hearing aid users[4]. In addition, none of the current noise reduction algorithms have led to improvements in speech intelligibility[5, 6]. We have developed a new single microphone noise reduction (SMNR) algorithm inspired by our neurophysiological research in the avian auditory system. These new algorithm performs a filtering operation in the modulation domain: the Fourier domain of a time-frequency representation of the sound. By operating in this domain, we are able to separate signal and noise sources that overlap in their frequency spectrum but distinguish themselves by having different temporal or spectral modulation structure. The goal of this project was to optimize this operation for processing speech in noise in real-time and with minimal delay. This optimized algorithm was then validated in psychophysical experiments where our algorithm was compared to other state-of-the-art SMNRs.

CORTICAL PROCESSING OF SIGNALS IN NOISE AND MODULATION FILTERING.

Our noise reduction algorithm was inspired by our neurophysiological research in the avian auditory system. In recent years we have demonstrated two motivating features of avian auditory processing. First, the avian auditory system shows sensitivity to sound features that efficiently encode behaviorally relevant sounds and bird song, in particular[7]. Secondly, we found noise-invariant neurons in a high-level auditory processing region. These neurons exhibited very similar response patterns to a bird song played in quiet and the same song played on a background of noise. When we further examined the tuning properties of these noise-invariant neurons, we found that we could in part explain their ability to extract signals from noise by their spectral-temporal receptive fields (STRFs). 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 a in 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 sound that contain particular time-frequency patterns and while filtering out other sound that might have similar frequency content but lack this spectral-temporal structure.

To show that noise invariance and thus noise filtering can be obtained from such STRF filters, we engineered a noise filtering algorithm based on a decomposition of the sound by an ensemble of “artificial” neurons described by realistic STRFs, or, in other words, a modulation filter bank. Similar decompositions have also been proposed and used by others for the efficient processing of speech and other complex signals [8, 9]. 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 [10] that would prevent such 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. This data driven optimization also finds the best filtering possible for our re-synthesis procedure as opposed to deriving the filters that yield the best signal in the spectrographic domain.

**FIGURE 1.** Schematic illustrating all the majors steps in the STRF based noise reduction algorithm. In the top row, the signal + noise (S+N) is decomposed in narrow-band signals by a frequency filter bank. That same decomposition can be used to generate a spectrogram of the sound shown on the second row. That spectrogram is process by the STRF or modulation filter bank. The output of that STRF bank is used to generate time varying gains for each frequency channel. Those gains are then used to synthesize a new signal.

In the next section, we describe in more detail the current implementation of our algorithm and our initial assessments of its performance relative to other SMNR algorithms.
THE STRF NOISE REDUCTION ALGORITHM.

Our algorithm falls in the general class of 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 various steps in our algorithm are illustrated on figure 1 and described here in moderate detail. Both the analysis and synthesis step 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 50 Hz and 10 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.

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. Initially, we developed the algorithm using a bank of STRFs that were designed to model the STRFs found in the auditory cortex, as these are known to model zebra finch song well [11]. To apply this method to speech, we obtained STRFs by performing Principal Components Analysis (PCA) on the log spectrogram of the clean speech and noise; the principal components (PCs) obtained are STRF-like structures that account for the characteristic structure of the audio sample.

The log spectrogram of the stimulus is convolved with each STRF to obtain model neural responses: \( \tilde{a}(t) \) of dimension M. The crux of our algorithm is to transform these neural responses back into a set of time varying frequency gains, \( \hat{g}(t) \) of dimension N. These frequency gains will then be applied to the corresponding frequency slices in the time-frequency decomposition of the sound to synthesize the processed signal. To obtain \( \hat{g}(t) \), each model neural response was first scaled by an importance weighting, \( d_i \), and then multiplied by the frequency marginal of the neuron's STRF:

\[
g_j(t) = f\left( \sum_{i=1}^{M} d_i \cdot a_i(t) \cdot K_{i,j} \right) \quad \text{with} \quad j \in \{1, N\}.
\]

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. The lower bound was set between 10 and 20 dB of attenuation, as this range yielded the most natural sounding signals. \( K_{i,j} \) 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.
ALGORITHM PERFORMANCE

We evaluated the algorithm both in terms of objective noise reduction by calculating a goodness of fit and in terms of qualitative subjective assessments by human subjects. We assessed noise reduction performance by computing a measure of signal to noise ratio (SNR) for speech in noise for both the unprocessed signal and the filtered signals. Using this metric, we compared our algorithm to three other spectral subtraction noise algorithms: the optimal Wiener filter (OWF), a variable gain algorithm patented by Sonic Innovations (STA for state-of-the-art) and the 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 spectral subtraction algorithm for Sonic Innovations used a time variable gain just as in our implementation (US Patent 6,757,395 B1). This algorithm is currently used in hearing aids. We used a Matlab implementation of the STA algorithm provided to us by Starkey Hearing Aid Research (Berkeley, CA). 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 can be considered as an approximate upper bound on the potential performance of general noise reduction algorithms[12]. The improvement in SNR was significantly higher for our STRF-algorithm than for the Wiener filter or the STA algorithm (figure 2, top right panel). The improvement reached to almost half the ceiling value as specified by the performance of the IBM.

![Figure 2](image)

**FIGURE 2.** Left: Gains and spectrogram of filtered speech when the STRF noise algorithm is applied to speech in babble noise. Right top: Quantitative assessment based on the increase in SNR of the STRF algorithm relative to the STA, OWF and IBM (see text). Right bottom: Subjective ratings (number of Better/Worse responses) comparing the STRF noise reduction to original unprocessed signal (Unp) or STA algorithm.

We then performed psychophysical experiments with human subjects for speech in crowd noise. In this test, subjects were asked to judge the amount of the noise reduction, the overall quality of the signal and the overall quality of the listening experience. As shown in figure 2 (bottom right pane), subjects appreciated the noise reduction of the STRF noise filtering and
although the noise reduction was assessed to not be as significant as for the Sonic algorithm, the quality of the signal as well as the overall quality of the listening experience were determined to be far superior

CONCLUSIONS

We have developed a new algorithm for filtering speech in noise that can be implemented in real-time with minimal delay. Initial psychophysical experiments show that this algorithm performs as well as other state-of-the-art real-time noise reduction algorithms. The algorithm therefore shows promises for use in hearing aids or other clinical applications. Addition psychophysical experiments involving speech intelligibility tests in hearing impaired subjects are needed to confirm this potential. The algorithm could also be optimized for other applications such as voice recognition in consumer electronics. In such applications, and if longer processing delays can be tolerated, the performance of the algorithm would be even greater.

ACKNOWLEDGMENTS

This work was funded by a NIDCD R01-007293 grant to FET. We thank William S. Woods of Starkey Hearing Research Center, Berkeley, CA., for expert advice on current state-of-the-art noise reduction algorithms.

REFERENCES