# Recurrent lateral inhibitory spiking networks for speech enhancement

Julie Wall
University of East London
London, UK
j.wall@uel.ac.uk

Cornelius Glackin
Intelligent Voice Ltd
London, UK
neil.glackin@intelligentvoice.com

Nigel Cannings
London, UK
Intelligent Voice Ltd
nigel.cannings@intelligentvoice.com

Gerard Chollet
Intelligent Voice Ltd
London,UK
gerard.chollet@intelligentvoice.com

Nazim Dugan
Intelligent Voice Ltd
London, UK
nazim.dugan@intelligentvoice.com

***Abstract*—Automatic speech recognition accuracy is affected adversely by the presence of noise. In this paper we present a novel noise removal and speech enhancement technique based on spiking neural network processing of speech data. The spiking network has a recurrent lateral topology that is biologically inspired, specifically by the inhibitory cells of the cochlear nucleus. The network can be configured for different acoustic environments and it will be demonstrated how the connectivity results in enhancement of temporal correlation between similar frequency bands and removal of uncorrelated noise sources. Demonstration of the speech enhancement capability will be provided with data taken from the TIMIT database with different levels of additive Gaussian white noise. Future directions for further development of this novel approach to noise removal and signal processing will also be discussed.**

***Keywords*—spiking neural networks; noise reduction; speech; lateral inhibition;***

## I. INTRODUCTION

Speech enhancement is a vital area of research in speech signal processing. Automatic speech recognition (ASR) accuracy is typically adversely affected by many noise sources. As the need for ASR applications to work in diverse environments increases, so too does the need to enhance speech by removing noise. In this context we will assume that noise is defined as anything other than speech. Traditionally, there are many existing techniques for removing noise from audio. The methodology advocated typically depends on the type of noise encountered, for example spectral subtraction [1] aims at removing wide-band noise whereas periodic speech enhancement is more suited to the removal of periodic noise such as that generated from the rotary motion of engines and other more structured types of sound.

In this paper we advocate an alternative ethos to speech enhancement in that we develop a speech enhancement approach which is responsive to the spectro-temporal nature of the speech, rather than the type of noise. We draw inspiration from the biological processing of sound by the cochlea and cochlear nucleus. The cochlea can be viewed as transforming sound into the frequency domain in a way broadly similar to a Fourier transform. Hence, the cochlea produces a tonotopically organised stimulus, but it also encodes the stimulus into spikes, which are reflective of the complex spatio-temporal features of the sound. This spiking input is then further processed by the cochlear nucleus using lateral inhibition [2] to extract features from it. Lateral inhibitory connectivity in conjunction with correctly configured transportation delays can be used to induce near-synchronous states that aid neural computation [3]. It has been previously demonstrated how the connectivity of lateral inhibitory networks can be successfully parameterised to promote edge enhancement and noise removal [4], and to minimise information loss between successive layers [5]. In this paper we build on this work and demonstrate how lateral inhibitory networks can be employed successfully as a practical technique for speech enhancement.

Section II discusses the transformation of the speech samples into spectrograms and the addition of noise. Section III describes the conversion of the spectrogram's continuous data into discrete spike timing. Section IV outlines the processing of the lateral inhibitory networks. Section V describes how the processed spiking representation is used to transform the spectral representation back to sound. Section VI presents the results of this technique while Section VII discusses conclusions and future directions of this research.

## II. PRE-PROCESSING OF SPEECH SAMPLES

The speech samples chosen to illustrate the lateral inhibitory networks are from the TIMIT database [6]. We have chosen four samples with different utterances of varying lengths, spoken by two male speakers and two female speakers, see TABLE I.

TABLE I. TIMIT SPEECH SAMPLES

| *Name* | *Speaker* | *Description* | *Length (s)* |
|---|---|---|---|
| **SI969** | Male | *'If any of us miss, they can pick up the pieces.'* | 2.713 |
| **SA2** | Female | *'Don't ask me to carry an oily rag like that.'* | 2.841 |
| **SX52** | Female | *'Her classical performance gained critical acclaim.'* | 2.963 |
| **SI976** | Male | *'Now a distinguished old man called on nine divinities to come and join us.'* | 3.718 |

The processing of speech advocated in this paper is biologically inspired. In the brain, speech is processed by the cochlea which performs a spectral transformation to extract frequency information from sound waves. We will use a short-term Fourier transform (STFT) to perform the analogue of this operation. Strictly speaking a true biological cochlea performs this using frequency bands tonotopically arranged on the Mel scale. We simplify this and use an STFT, we make this simplification for the purposes of faster computation of the spectral transformation. We use the STFT version of the Fourier transform for its improved temporal rather than frequency resolution.

Once the STFT is calculated the magnitude of the complex Fourier transform is determined and then log scaled to obtain the spectrogram. Fig. 1 shows the spectrogram generated for the SI969 speech sample. This particular sample was chosen arbitrarily and will be used throughout this paper to illustrate the noise removal steps.

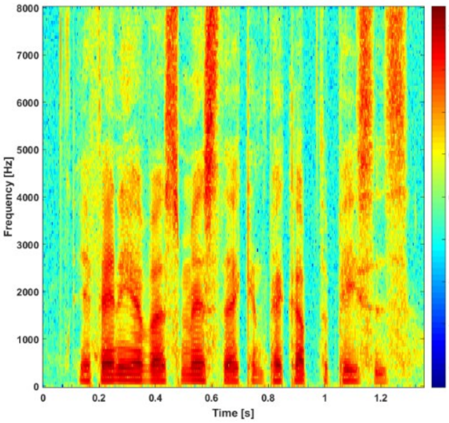

Fig. 1. Spectrogram of original speech sample (SI969 from TIMIT)

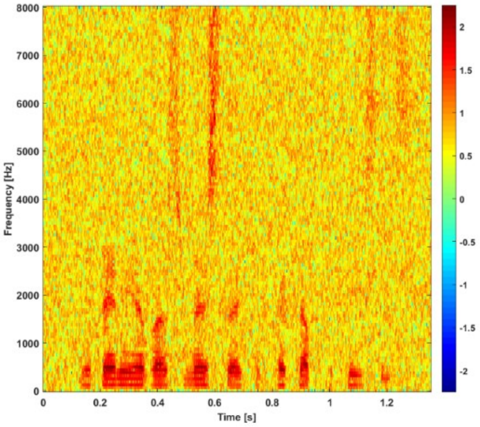

Fig. 2. Spectrogram of the original speech with noise added, the noise is Gaussian white and has the the same amount of power as the orginal speech.

The signal to noise ratio (SNR) defines how much of the original signal has been corrupted by noise. In the case of this research, the original signal corresponds to the TIMIT database speech samples and noise relates to the addition of white Gaussian noise. SNR can be defined as:

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{1}$$

where $P$ is the average power.

A range of SNRs were chosen for this task: (0.1; 1; 10), dB levels for the additive noise were based on the measurement of the power of the clean signal. The Matlab *awgn* function was used to incorporate the white Gaussian noise into the speech signals. The top left panel of Fig. 12 maps the original speech signal against the same input when noise is added resulting in a signal with an SNR of 1, i.e. the added noise has the same amount of power as the original speech. Consequently, the spectrogram of the speech signal has been greatly affected, as seen in Fig. 2.

## III. Spike Encoding

The cochlea in addition to the spectral transformation encodes the frequency information into trains of action potentials or 'spikes'. Arguably it is the timing of the individual spikes that is responsible for the encoding of the sound stimulus and not the spiking amplitude. Hence we employ a deconvolution algorithm to convert the continuous spectrogram data into digital spikes. We simplify the spiking representation and simply use the integer 1 to represent a spike and 0 otherwise in our data storage. Despite the additional memory overhead this creates in the data structures that store the spikes as compared to a sparse representation that simply stores the spike times, we have found that this provides the ability to pre-allocate the data structures and simplify the subsequent network programming. Similar to our previously published experiments, we have used Ben's Spiker Algorithm (BSA) [7] to convert this continuous data into discrete spike timing. This algorithm utilises a convolution/deconvolution filter which was optimised for encoding/decoding using a genetic algorithm optimised to minimise the error between the deconvolution encoding using the BSA algorithm, and the convolution decoding, which can be performed using a linear filter with the same BSA filter coefficients.

In an effort to ensure our results are more in line with the speech recognition domain of research, we have generated spike trains for 128 frequency channels ranging between 0 *Hz* and 8 *kHz*. Fig. 3 illustrates the different spike trains that begin firing, end firing, and fire maximally at specific times. The literature describes these features of sound signals as onsets, offsets and peak firing rates respectively [8]. The aim of this research is to remove noise, such as background sound, from the speech signal using spiking neural network connectivity regimes; thus facilitating the extraction of these features for enhanced speech processing.

All neurons in the lateral inhibitory layer are implemented using a leaky integrate and fire (LIF) neuron [9], it is one of the simplest, computationally efficient and most popular models of a spiking neuron:

$$\tau_{mem}\frac{dv}{dt} = -v + R_{in}I_{syn}(t) \tag{2}$$

where $\tau_{mem}$ refers to the membrane time constant of the neuron, $v$ is the membrane potential, and $R_{in}$ is the membrane resistance, driven by the synaptic current $I_{syn}(t)$. Fig. 3 illustrates the spike encoded output of the original signal and Fig. 4 illustrates the spike encoded output of the original speech signal when it has been embedded with white Gaussian noise. Reflecting the differences between the two spectrograms from Fig. 1 and Fig. 2, the task for the lateral inhibitory layer of spiking neurons is to remove the spikes resulting from the noise and enhance the original speech signal.

For comparison purposes the spiking representation of the original speech and the noisy speech are shown in Fig. 3 and

Fig 4 respectively. Hence it will be the aim of the spiking neural network processing to process the noisy spiking input to realise the clean spiking representation. Whilst this is a somewhat artificial example (since we have artificially added noise to clean speech) we do this to demonstrate how the spiking processing enhances speech and removes noise.

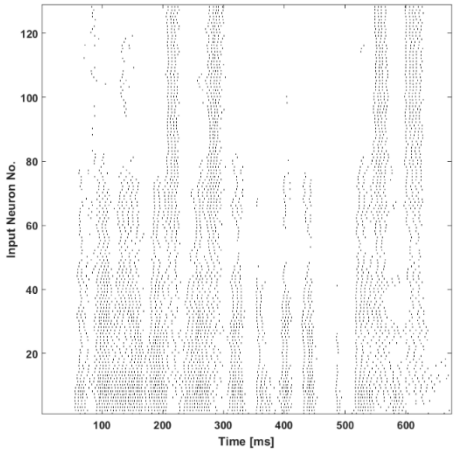

Fig. 3. Spike raster representation of the original (noiseless) speech sample

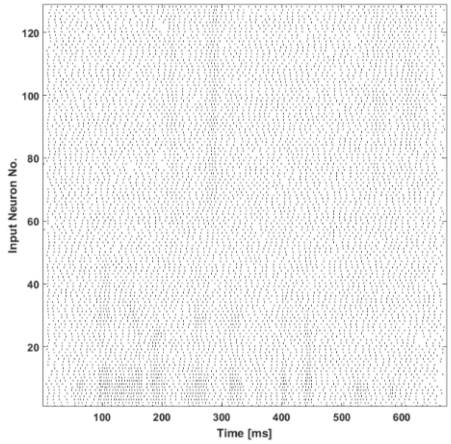

Fig. 4. Spike raster corresponding to the noisy speech sample

## IV. LATERAL INHIBITORY NETWORKS

The spiking topology that is used in this work to process the audio is inspired by the seminal paper by Abbot [10]. In this paper, Abbot clarifies the role of laterally inhibitory neurons that was successfully proposed in response to an international challenge by Hopfield and Brody [11] to propose a mechanism by which neurons in disparate parts of the brain synchronise their activity. This of course is attempting to address the long-held variable binding problem. The winning solution was proposed by Wills and Mackay [8] and then elegantly demonstrated by the experiment illustrated in Fig. 5 [10]. The figure shows the interaction between two LIF neurons with lateral inhibitory connections, connecting to an excitatory output neuron which sums the output. Each neuron in the input layer receives as input the inhibitory output of the other. Each of these neurons has its own excitatory input, one receives a fixed firing rate of 25 *Hz*, the other a firing rate linearly changing from 28.5 *Hz* to 21.5 *Hz*. When the two input neurons are firing at different frequencies, the input neurons take turns at suppressing the output of one another, depending on spike timing. The output neuron, neuron 3, fires maximally (coincidently) when the two input neurons are firing at the same frequency. Similar to two people trying to pass one another in a narrow space, the laterally connected neurons take turns inhibiting one another, preventing and delaying each other from firing until they are 'in-sync'.

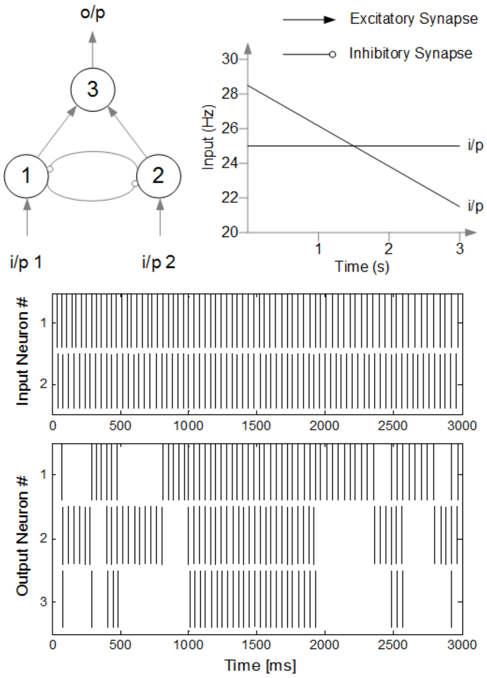

Fig. 5. A simple three-neuron SNN (top left) used to test the ability of laterally connected neurons to produce coincidental firing and synchrony [10]. The SNN receives two input spike trains, one with a firing rate reducing from 28.5 to 21.5 Hz, the other with a constant firing rate of 25 Hz (top right). The bottom two subplots show the actual input and output spike rasters of inputs 1 and 2 (i/p 1 and i/p 2) and resultant spike output for neurons 1, 2, and 3 respectively.

Expanding on this simple example from Abbot, Glackin et al. employed layers of neurons with similar lateral inhibitory connections to produce synchronised spiking activity from tonotopically arranged sound information [4-5]. The lateral connectivity of the input layer was designed in terms of a *connection length* parameter and a *neighbourhood radius*.

As the output of the STFT and subsequent BSA encoding of spike trains is tonotopically arranged, it does not necessarily make sense to associate every input neuron and hence sound frequency with every other, as this disregards the tonotopic arrangement. It seems more likely that the lateral connectivity of the input layer can be described in terms of a *connection length* parameter. A particular connection length of *c* would mean that each input layer neuron is laterally connected to *c* neurons either side of it. Fig. 6 illustrates this idea, the black lines of various styles represent connection lengths between 1 and 3 for an example layer of laterally connected neurons. In this way, a layer of $N$ neurons can have a maximum connection length of $c_{max} = N - 1$.

The concept of a neighbourhood in the connectivity of neurons was introduced by Kohonen with the 'winner-take-all' competitive learning algorithm [12]. Essentially, 'winning' neurons had their weights increased along with neurons topologically close to them, i.e. in the same neighbourhood. In terms of the spiking neural networks in this paper, the *neighbourhood radius* describes how neurons that are tonotopically close to one another are not connected laterally. Fig. 7 illustrates this idea.

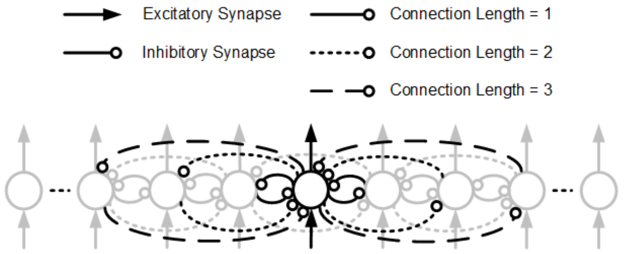

Fig. 6. Neurons connected laterally as determined by a connection length parameter

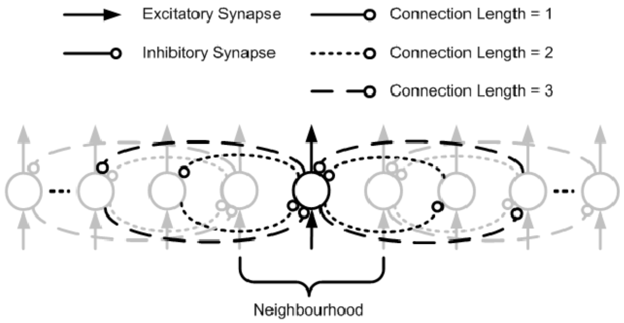

Fig. 7. Incorporating the neighbourhood radius parameter to layers of lateral connectivity specified by the connection length parameter

When the layer of spiking neurons is presented with the noisy spiking input from Fig. 4, the lateral inhibitory processing removes spikes. The spikes that remain are densely packed by necessity in order to survive this process. Contrastingly, areas of spikes which are not densely packed are removed. This removal of noisy spikes is an important feature for the pre-processing of speech signals which will then be used as input to an ASR system.

From previous experiments [5], we found that a less than maximum connection length has interesting effects on small neighbourhoods. The small neighbourhood radius and slightly larger connection length do not result in a synchronous state. Synchronous states typically show entire populations of neurons firing in phase. But if all neurons are firing at the same time, how could synchrony produce selectivity to particular stimuli? Using multiple spiking layers of lateral inhibitory processing with a sparser connectivity, noisy spikes can be removed and a clearer representation of the sound signal can be extracted by sharpening the main contours of the spike distribution. Fig. 8 and Fig. 9 show the output after two iterative layers of synchrony on the noisy spiking input.

The network used has fixed weights and connectivity parameters. The connection length and neighbourhood was set to 22 and 5 for layer 1 and 25 and 5 for layer 2. The connectivity parameters were tuned heuristically to gradually suppress noise whilst taking care not to suppress the speech parts of the signal. The weights were tuned according to the amount of lateral inhibitory connectivity dictated by the connection length and neighbourhood radius. The inhibitory and excitatory weights were set to 1.4 and -0.5 for layer 1 and 1.5 and -0.2 for layer 2. The slight difference in the parameters chosen for layer 2 ensured gentle inhibition with sharper more defined contours. The tuning process was not difficult but in future work we will investigate how to do this in a more principled automated way.

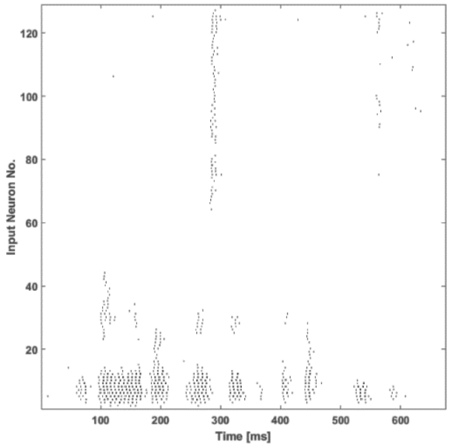

Fig. 8. Spike encoded output after passing through first spiking layer of lateral inhibitory processing. Areas of spikes which were not densely packed in the input, are removed.

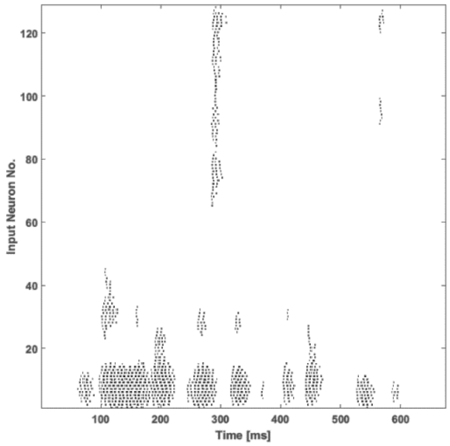

Fig. 9. Spike encoded output after passing through second spiking layer of lateral inhibitory processing. Spikes that remain after passing through the first layer, have become denser and more tightly packed.

## V. DECODING

As mentioned previously, convolution with the linear filter containing the original BSA filter coefficients can accurately convert the processed spectral stimulus from the discrete spiking domain back to the continuous domain. However, the resulting processed spectrogram is not easily invertible. This of course has nothing to do with the spike decoding but instead is due to the lossy nature of spectrograms. Specifically the conversion of the complex Fourier spectrum to the power values results in a loss of phase information. Consequently, without modelling the phase in some way, performing the inverse STFT on the spectrogram can only be done by assuming the magnitude and phase are the same, which results in phase distortion in the inverse STFT.

Many ASR systems are now employing end-to-end deep neural network topologies which process spectrograms. Therefore the phase distortion issue for such ASR systems is irrelevant if the aim of this speech enhancement technique is to improve ASR accuracy in such systems. However for illustrative purposes we now use the processed spectrogram as a mask for the original STFT representation. The mask is simply applied by element-by-element multiplying the processed spectrogram to both the real and imaginary parts of the original noisy spectral representation. For illustrative purposes we decode and perform the masking operation at successive layers of the network. The reconstructed spectrograms from the spike encoded outputs of both iterative layers of synchrony can be seen in Fig. 10 and Fig. 11 for layers 1 and 2. The figures demonstrate how the spectrograms from layers 1 and 2 successively remove the noise and enhance the speech pattern.

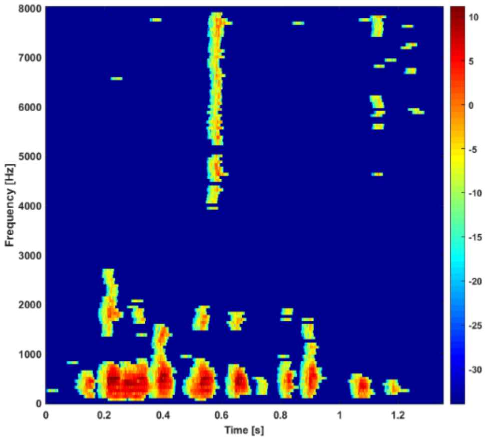

Fig. 10. Reconstructed spectrogram after the first spiking layer's output is used to mask the original STFT.

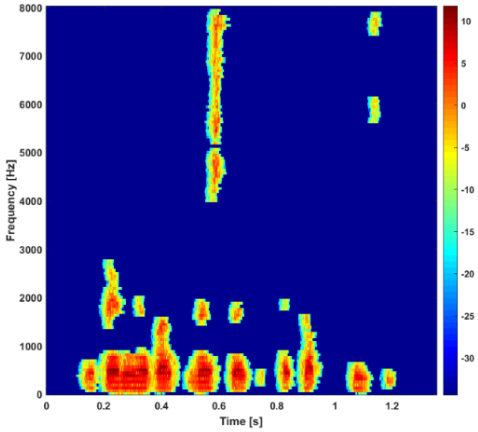

Fig. 11. Final reconstructed spectrogram after the noisy spiking has passed through two iterative layers of synchrony

By using the layer 2 output as a mask for the original noisy STFT we can then perform the inverse STFT operation and transform the spectral representation back to sound. Fig. 12 shows the noisy signal (top right), the reconstructed signal after processing by the SNN (bottom right), and for comparison purposes we also show the original noise free signal (top left), and the processed noise free signal output (bottom left). As can be seen the reconstructed signal is qualitatively similar to the noise-free signal.

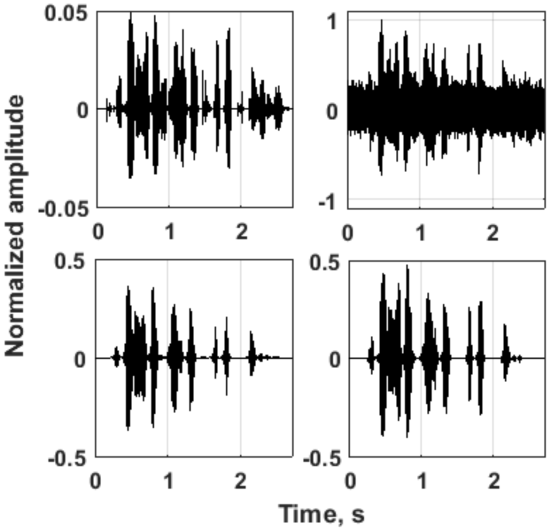

Fig. 12. Original signal (top left panel), noisy signal (top right panel), reconstructed signal after one pass through lateral inhibitory layer (bottom left panel), reconstructed signal after second pass through lateral inhibitory layer (bottom right panel)

## VI. Results

During the setup of these experiments, informal subjective listening tests were carried out at each stage of the algorithm's development to determine how successfully the lateral inhibitory networks performed when applied to the TIMIT speech samples outlined in TABLE I. However, once the algorithm was complete, from encoding the original noisy speech signal to reconstructing the processed speech output, objective tests using metrics could be performed to fully evaluate the output, taking advantage of the clean speech reference signal. Hence we can directly measure the improvement in SNR using Equations (3-5):

$$SNR_A = 10log_{10} * \left(\frac{var(x_k)}{var(n_k)}\right) \quad (3)$$

$$SNR_B = 10log_{10} * \left(\frac{var(y_k)}{var(n_k)}\right) \quad (4)$$

$$SNR_{IMP} = SNR_B - SNR_A \quad (5)$$

where $x_k$ is the clean speech signal; $n_k$ is the added noise; $y_k$ is the processed speech signal; and $SNR_{IMP}$ measures the improvement between the original clean speech signal $SNR_A$ and the output processed and reconstructed speech signal $SNR_B$.

TABLE II. presents the SNR values for the clean speech samples and the resulting SNR after each pass through the layer of spiking neurons for the three different levels of added white Gaussian noise. While TABLE III. presents the individual improvement between the original clean speech samples and the processed and reconstructed outputs and the average and standard deviation across the four speech samples tested. From both the figures included throughout this paper and the objective test results presented, it can be seen that layer 1 processing removes a considerable amount of noise from the speech signal while layer 2 processing continues to remove noise but also ensures that the portions of the signals that remain are bolstered and become sharper. In terms of SNR, a clear improvement across all speech samples tested has been demonstrated; with an average improvement of 19.003 dB with a standard deviation measure of 1.8682 *dB*.

TABLE II. EXPERIMENTAL RESULTS

| ***Sample*** | ***SNR 'measured'*** | ***Orig. SNR*** | ***Layer 1 output SNR*** | ***Layer 2 output SNR*** |
|---|---|---|---|---|
| **SI969** | 0.1 | 0.0985 | 18.337 | 20.363 |
| | 1 | 0.9985 | 19.734 | 21.697 |
| | 10 | 9.9985 | 30.779 | 32.491 |
| **SA2** | 0.1 | 0.1175 | 15.867 | 18.7 |
| | 1 | 1.0175 | 18.365 | 20.707 |
| | 10 | 10.0175 | 29.161 | 31.178 |
| **SX52** | 0.1 | 0.1235 | 14.985 | 17.503 |
| | 1 | 1.0235 | 16.4 | 18.296 |
| | 10 | 10.0235 | 23.545 | 26.652 |
| **SI976** | 0.1 | 0.1042 | 15.967 | 17.621 |
| | 1 | 1.0042 | 16.585 | 18.198 |
| | 10 | 10.0042 | 28.217 | 29.056 |

TABLE III. ANALYSIS

| ***Sample*** | ***SNR 'measured'*** | ***SNR Layer 1 Improvement*** | ***SNR Layer 2 Improvement*** |
|---|---|---|---|
| **SI969** | 0.1 | 18.2385 | 20.2649 |
| | 1 | 18.7359 | 20.6982 |
| | 10 | 20.78 | 22.492 |
| **SA2** | 0.1 | 15.7491 | 18.5829 |
| | 1 | 17.347 | 19.6898 |
| | 10 | 19.1437 | 21.1607 |
| **SX52** | 0.1 | 14.862 | 17.3797 |
| | 1 | 15.3767 | 17.2728 |
| | 10 | 13.5217 | 16.6289 |
| **SI976** | 0.1 | 15.8626 | 17.6211 |
| | 1 | 15.5808 | 17.1937 |
| | 10 | 18.2123 | 19.0516 |
| **Average** | | 16.9508 | 19.003 |
| **STD** | | 2.1142 | 1.8682 |

## VII. CONCLUSIONS

Most of the techniques in the speech enhancement literature advocate an ethos that models the noise so that it can be extracted from audio. Contrastingly with the spiking approach in this paper, the spiking network strengthens the spectro-temporal correlations in the audio and naturally suppresses any uncorrelated noise, and hence makes this approach a novel one in the literature. The spiking topology is one suggestion as to possible inhibitory connectivity that could exist in the biology. However in the biology a much larger population of neurons is devoted to auditory processing in the cochlear nucleus. Hence it is likely that many configurations of spiking networks could be used to extract and be attentive to different aspects of the different sounds they perceive.

In future work we would like to replace the described masking process with some form of phase reconstruction. In addition, in the future we will expose this technique to more varieties of noise.

The spiking networks presented do not employ any form of learning yet in the configuration of the network. As previously described, the weights were fixed and in future work we will investigate how to modify the weights to perform some form of unsupervised learning. Despite the minor limitations in some aspects of the methodology, there is little doubt as to the effectiveness of the noise removal and speech enhancement. Even for the extreme case where the additive Gaussian white noise has the same power as the underlying signal, the effect is dramatic.

## REFERENCES

[1] Y. Zhang, and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Communication,* vol. 55, no. 4, pp. 509-522, 2013.

[2] M. R. Martin and J. W. Dickson, "Lateral inhibition in the anteroventral cochlear nucleus of the cat: A microiontophoretic study," *Hearing Research*, vol. 9, no. 1, pp. 35-41.

[3] E. V. Lubenov and A. G. Siapas, "Decoupling through synchrony in neuronal circuits with propagation delays," *Neuron*, vol. 58, no. 1, pp. 118-131, 2008.

[4] C. Glackin, L. Maguire, L. McDaid and J. Wade, "Lateral inhibitory networks: synchrony, edge enhancement and noise reduction," *Proc. IEEE International Joint Conference on Neural Networks*, pp. 1003-1009, 2011.

[5] C. Glackin, L. Maguire, L. McDaid and J. Wade, "Synchrony: A spiking-based mechanism for processing sensory stimuli," *Neural Networks*, vol. 32, pp. 26-34, 2012.

[6] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, V. Zue, TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[7] B. Schrauwen and J. Van Campenhout, "BSA, a fast and accurate spike train encoding scheme," *Proc. International Joint Conference on Neural Networks*, vol. 4, pp. 2825-2830, 2003.

[8] S. A. Wills. "Recognising speech with biologically-plausible processors," *Hamilton Prize*, 2001, [Online], http://www.inference.phy.cam.ac.uk/saw27/hamilton.pdf.

[9] R. Stein, "The frequency of nerve action potentials generated by applied currents," *Proc. of the Royal Society of London. Series B, Biological Sciences (1934-1990)*, vol. 167, no. 1006, pp. 64–86, 1967

[10] L. F. Abbot, "The timing game," *Nature Neuroscience*, vol. 4, no. 2, pp. 115-116, 2001.

[11] "A rule of thumb that unscrambles the brain," *The New York Times*, 2000, [Online], http://www.nytimes.com/2000/10/03/science/03CHALL.html

[12] K. Samuel and T. Kohonen, "Winner-take-all networks for physiological models of competitive learning," *Neural Networks*, vol. 7, no. 6, pp. 973-984, 1994.