# The Separation of Speech from Interfering Sounds: An Oscillatory Correlation Approach

Guy J. Brown
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield S1 4DP, UK
Email: g.brown@dcs.shef.ac.uk

DeLiang Wang
Department of Computer and Information Science
and Centre for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
Email: dwang@cis.ohio-state.edu

## Abstract

*A neural model is described which uses oscillatory correlation to segregate speech from interfering sound sources. The core of the model is a two-layer neural oscillator network. The first layer of the network identifies connected regions of energy in the time-frequency plane (segments). In the second layer, segments that have a common fundamental frequency are grouped into streams. A stream is represented by a synchronized population of relaxation oscillators, and different streams are represented by desynchronized oscillator populations. The model has been evaluated using a corpus of voiced speech mixed with interfering sounds, and produces an improvement in signal-to-noise ratio for every mixture.*

## 1. Introduction

Speech is seldom heard in isolation: usually, it is mixed with other environmental sounds. Hence, the auditory system must parse the acoustic mixture reaching the ears in order to retrieve a description of each sound source, a process termed *auditory scene analysis* (ASA) [2]. Conceptually, ASA may be regarded as a two-stage process. The first stage (which we term 'segmentation') decomposes the acoustic stimulus into a collection of sensory elements. In the second stage ('grouping'), elements that are likely to have arisen from the same environmental event are combined into a perceptual structure called a *stream*. Streams may be further interpreted by higher-level cognitive processes.

Recently, there has been a growing interest in the development of computational systems that mimic ASA (for example, see [4], [1], [5]). Most of these studies have been motivated by the need for robust front-end processors for automatic speech recognition. Such computational auditory scene analysis (CASA) systems are inspired by auditory function but do not model it closely; rather, they employ algorithms based on symbolic search or high-level inference engines. Although the performance of these systems is encouraging, they are no match for the abilities of a human listener; additionally, they tend to be complex and computationally intensive. In short, CASA currently remains an unsolved problem for real-time applications such as automatic speech recognition.

Given that human listeners can segregate concurrent sounds with apparent ease, computational systems that are more closely modelled on the neurobiological mechanisms of hearing may offer a performance advantage over existing CASA systems. This observation – together with a desire to understand the neurobiological basis of ASA – has led a number of investigators to propose neural network models of ASA. Most recently, Brown and Wang [3] have given an account of concurrent vowel separation based on *oscillatory correlation*. The oscillatory correlation framework, proposed by Wang [9], may be regarded as a special form of the temporal correlation theory elucidated by von der Malsburg [8]. In this framework, a set of auditory elements forms a perceptual stream if the corresponding oscillators are synchronized (phase locked with zero phase lag), and are desynchronized from oscillators that represent different streams. Evidence for the oscillatory correlation theory comes from neurobiological studies which report synchronised oscillations in the auditory cortex, visual system and olfactory system (see [11] for a review).

In this paper, we study ASA from a neurocomputational perspective and propose a neural network model that is able to segregate speech from interfering sounds. Our model uses oscillatory correlation as the underlying neural mechanism for ASA; streams are formed by synchronizing oscillators in a two-dimensional time-frequency network.

## 2. Model description

The input to the model consists of a mixture of speech and an interfering sound source, sampled at a rate of 16 kHz with 16 bit resolution. This input signal is processed in four stages, comprising peripheral auditory processing, mid-level auditory representations, neural oscillator network and resynthesis. The stages of the model are summarised in Figure 1 and described below (see [11] for a full account).

### 2.1. Peripheral auditory processing

Peripheral auditory frequency selectivity is modelled using a bank of bandpass filters with overlapping pass-bands. More specifically, we use a bank of 128 gammatone filters with center frequencies equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz [1]. The gains of the filters are chosen to simulate the frequency-dependent pressure gains of the outer and middle ears. Subsequently, the output of each filter is processed by a model of inner hair cell function. The output of the hair cell model is a probabilistic representation of auditory nerve firing activity, which exhibits saturation, two-component adaptation and frequency-limited phase locking effects.

### 2.2. Mid-level auditory representations

Mechanisms similar to those underlying pitch perception can contribute to the perceptual separation of sounds that have different fundamental frequencies (F0s). For example, listeners' ability to identify two concurrent vowel sounds is improved when the vowels have a different F0, relative to the case in which they have the same F0 [3].

Accordingly, the second stage of the model extracts periodicity information from the simulated auditory nerve firing patterns. This is achieved by computing a running autocorrelation of the auditory nerve activity in each frequency channel, forming a representation known as a *correlogram* [1], [5]. At time step $j$, the autocorrelation $A(i,j,\tau)$ for channel $i$ with time lag $\tau$ is given by:

$$A(i, j, \tau) = \sum_{k=0}^{K-1} r(i, j-k)r(i, j-k-\tau)w(k) \quad (1)$$

Here, $r$ is the output of the hair cell model and $w$ is a rectangular window of width $K$ time steps. We use $K = 320$, corresponding to a window width of 20 ms. The autocorrelation lag $\tau$ is computed in $L$ steps of the sampling period between 0 and $L$-1; we use $L = 201$, corresponding to a maximum delay of 12.5 ms. Equation (1) is computed for $M$ time frames, taken at 10 ms intervals (i.e., at intervals of 160 steps of the time index $j$). Hence, the correlogram is a three-dimensional volume of size $N \times M \times L$.

For periodic sounds, a characteristic 'spine' appears in the correlogram which is centered on the lag corresponding to the stimulus period. This pitch-related structure can be emphasized by summing the channels of the correlogram across frequency, giving a 'pooled' correlogram $s(j,\tau)$:

$$s(j, \tau) = \sum_{i=1}^{N} A(i, j, \tau) \quad (2)$$

For periodic sounds, $s(j,\tau)$ exhibits a prominent peak whose position on the delay axis corresponds to perceived pitch.

It is also possible to extract harmonics and formants from the correlogram, since frequency channels that are excited by the same acoustic component share a similar pattern of periodicity. Bands of coherent periodicity can be identified by cross-correlating adjacent correlogram channels; regions of high correlation indicate a harmonic or formant [1]. We define the cross-correlation $C(i,j)$ between channels $i$ and $i$+1 at time frame $j$ as follows:

$$C(i, j) = \frac{1}{L}\sum_{\tau=0}^{L-1} \hat{A}(i, j, \tau)\hat{A}(i+1, j, \tau) \quad (1 \le i \le N-1) \quad (3)$$

Here, $\hat{A}(i, j, \tau)$ is the autocorrelation function of (1) which has been normalized to have zero mean and unity variance (this ensures that $C(i,j)$ is sensitive only to the pattern of periodicity in the correlogram, and not to the mean firing rate in each channel).

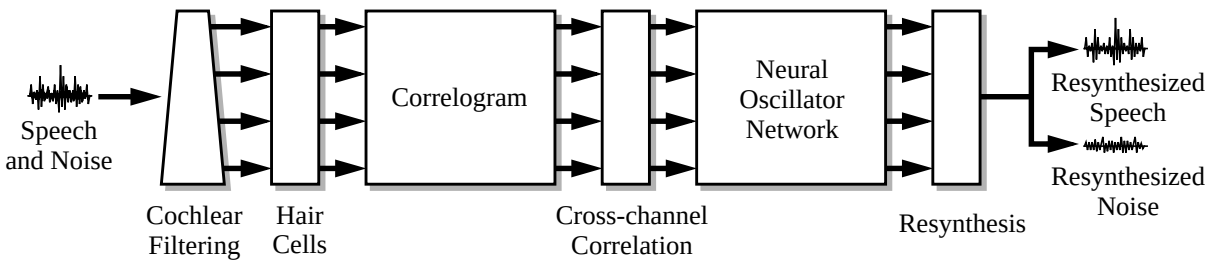


**Figure 1**: Schematic diagram of the auditory model. The model consists of four main stages; peripheral auditory processing (cochlear filtering and inner hair cell model), mid-level auditory representations (correlogram and cross-channel correlation), neural oscillator network and resynthesis pathway.

### 2.3. Neural oscillator network: overview

In our model, segmentation and grouping take place within a two-layer oscillator network. The basic unit of the network is a single oscillator, which is defined as a reciprocally connected excitatory variable $x$ and inhibitory variable $y$ [7]. The oscillator may be interpreted as a model of action potential generation or oscillatory burst envelope, where $x$ represents membrane potential and $y$ represents the level of activation of a number of ion channels. Since each layer of the network takes the form of a time-frequency grid (see Figure 2), we index each oscillator according to its frequency channel ($i$) and time frame ($j$):

$$\dot{x}_{ij} = 3x_{ij} - x_{ij}^3 + 2 - y_{ij} + I_{ij} + S_{ij} + \rho \quad (4a)$$

$$\dot{y}_{ij} = \varepsilon(\gamma(1 + \tanh(x_{ij}/\beta)) - y_{ij}) \quad (4b)$$

Here, $I_{ij}$ represents external input to the oscillator, $S_{ij}$ denotes the coupling from other oscillators in the network, $\varepsilon$, $\gamma$ and $\beta$ are parameters, and $\rho$ is the amplitude of a Gaussian noise term. The inclusion of noise allows the robustness of the system to be tested, and assists desynchronization among different oscillator blocks.

If coupling and noise are ignored and $I_{ij}$ is held constant, (4) defines a relaxation oscillator with two time scales. The $x$-nullcline, i.e. $\dot{x}_{ij} = 0$, is a cubic function and the $y$-nullcline is a sigmoid function. If $I_{ij} > 0$, the two nullclines intersect only at a point along the middle branch of the cubic with $\beta$ chosen small. In this case, the oscillator gives rise to a stable limit cycle for all sufficiently small values of $\varepsilon$, and is referred to as *enabled*. The limit cycle alternates between *silent* and *active* phases of near steady-state behaviour.

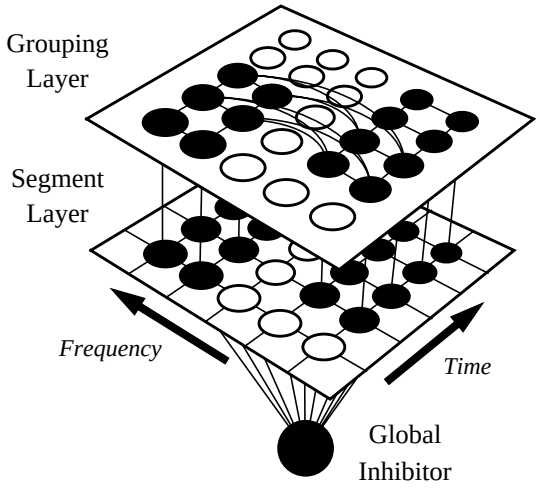


**Figure 2**: Structure of the oscillator network. The network has two layers, which mirror the two conceptual stages of auditory scene analysis: segmentation and grouping.

Compared to motion within each phase, the alternation between phases takes place rapidly, and is referred to as *jumping*. If $I_{ij} < 0$, the two nullclines intersect at a stable fixed point. In this case, no oscillation occurs. It is clear, therefore, that oscillations in (4) are stimulus-dependent.

### 2.4. Neural oscillator network: segment layer

In the first layer of the network, *segments* are formed – blocks of synchronised oscillators that trace the evolution of an acoustic component through time and frequency. The first layer is a two-dimensional time-frequency grid of oscillators with a global inhibitor (see Figure 2). Hence, $S_{ij}$ in (4a) is defined as

$$S_{ij} = \sum_{kl \in N(i,j)} W_{ij,kl} H(x_{kl} - \theta_x) - W_z H(z - \theta_z) \quad (5)$$

where $W_{ij,kl}$ is the connection weight from an oscillator ($i$,$j$) to an oscillator ($k$,$l$) and $N$($i$,$j$) is the four nearest neighbors of the grid location ($i$,$j$). The threshold $\theta_x$ is chosen so that an oscillator has no influence on its neighbors unless it is in the active phase. The weight of neighboring connections along the time axis is uniformly set to 1. The weight of vertical connections between an oscillator ($i$,$j$) and its neighbor ($i$+1,$j$) is set to 1 if the cross-correlation $C$($i$,$j$) exceeds a threshold $\theta_c$; otherwise it is set to 0. $W_z$ is the weight of inhibition from the global inhibitor $z$, defined as

$$\dot{z} = \sigma_\infty - z \quad (6)$$

where $\sigma_\infty = 1$ if $x_{ij} \geq \theta_z$ for at least one oscillator ($i$,$j$), and $\sigma_\infty = 0$ otherwise. Hence $\theta_z$ is a threshold. If $\sigma_\infty = 1$, $z \to 1$.

Small segments may form which do not correspond to perceptually significant acoustic components. In order to remove these noisy fragments, we introduce a lateral potential $p_{ij}$ for oscillator ($i$,$j$), defined as [12]:

$$\dot{p}_{ij} = (1 - p_{ij}) H\left[\sum_{kl \in N_p(i,j)} H(x_{kl} - \theta_x) - \theta_p\right] - \varepsilon p_{ij} \quad (7)$$

Here, $N_p$($i$,$j$) is called the potential neighborhood of ($i$,$j$), which is chosen to be the left neighbor ($i$,$j$-1) and the right neighbor ($i$,$j$+1). $\theta_p$ is a threshold. If both neighbors of ($i$,$j$) are active, $p_{ij}$ approaches 1 on a fast time scale; otherwise, $p_{ij}$ relaxes to 0 on a slow time scale determined by $\varepsilon$.

The lateral potential plays its role by gating the input to an oscillator. More specifically, we replace (4a) with

$$\dot{x}_{ij} = 3x_{ij} - x_{ij}^3 + 2 - y_{ij} + I_{ij} H(p_{ij} - \theta) + S_{ij} + \rho \quad (4a')$$

With $p_{ij}$ initialized to 1, it follows that $p_{ij}$ will drop below the threshold $\theta$ unless the oscillator ($i$,$j$) receives excitation from its entire potential neighborhood. Given our choice of neighborhood in (5), this implies that a segment must extend

for at least three consecutive time frames. Oscillators that are stimulated but cannot maintain a high potential are relegated to a discontiguous 'background' of noisy activity.

An oscillator ($i$,$j$) is stimulated if its corresponding input $I_{ij} > 0$. Oscillators are stimulated only if the energy in their corresponding correlogram channel exceeds a threshold $\theta_a$. It is evident from (1) that the energy in a correlogram channel $i$ at time $j$ corresponds to $A(i,j,0)$; thus we set $I_{ij}$ = 0.2 if $A(i,j,0) > \theta_a$, and $I_{ij}$ = -5 otherwise.

Figure 3 shows the results of segmentation by the first layer of the network for a mixture of speech and trill telephone. The network was simulated using the LEGION algorithm [9] which follows the major steps in the dynamic evolution of the differential equations, but offers considerable savings in computation time. The system produces 94 segments plus the background, which consists of small fragments lasting just one or two time frames. Each segment is represented by a distinct gray level. It should be noted that although all segments are shown together in Figure 3 for convenience, each arises during a unique time interval in accordance with the principle of oscillatory correlation.

### 2.5. Neural oscillator network: grouping layer

The second layer is a two-dimensional network of laterally coupled oscillators without global inhibition, which embodies the grouping stage of ASA. Oscillators in this layer are stimulated if the corresponding oscillator in the first layer is stimulated and does not form part of the background. Initially, all oscillators have the same phase,

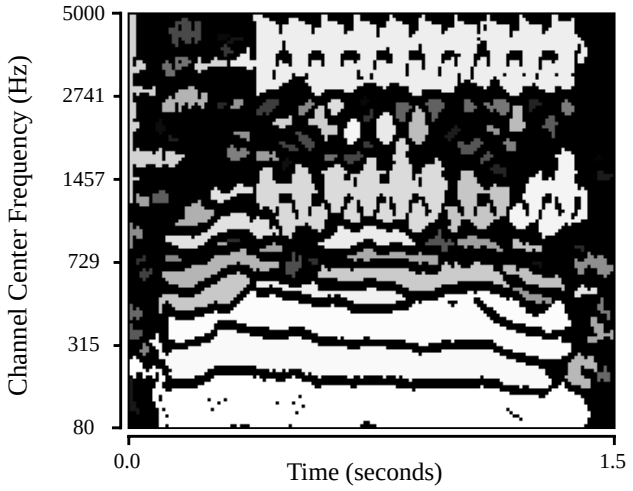


**Figure 3**: Segments formed by the first layer of the network for a mixture of speech and trill telephone. The utterance is 'why were you all weary?' spoken by a male speaker. Each segment is indicated by a unique gray level. Unstimulated oscillators and the background are shown in black.

implying that all the segments from the first layer are assumed to be in the same stream. This initialization is consistent with psychophysical evidence suggesting that perceptual fusion is the default state of auditory organisation [2]. In the second layer, a single oscillator has the same form as in (4), except that $x_{ij}$ is changed to:

$$\dot{x}_{ij} = 3x_{ij} - x_{ij}^3 + 2 - y_{ij} + I_{ij}[1 + \mu H(p_{ij} - \theta)] + S_{ij} + \rho \qquad (4a'')$$

Here, $\mu$ is a small positive parameter (0.01); this implies that an oscillator with a high lateral potential gets a slightly higher external input. We choose $N_p(i,j)$ and $\theta_p$ (see equation (7)) so that oscillators which correspond to the longest segment from the first layer are the first to jump to the active phase. The longest segment can be identified by using the selection mechanism described in [10].

The coupling term in (4a') consists of two types of coupling:

$$S_{ij} = S_{ij}^e + S_{ij}^v \qquad (8)$$

Here, $S_{ij}^e$ represents mutual excitation between oscillators within each segment. Specifically, we set $S_{ij}^e = 4$ if the active oscillators from the same segment occupy more than half of the length of the segment; otherwise $S_{ij}^e = 0.1$ if there is at least one active oscillator from the same segment.

The coupling term $S_{ij}^v$ denotes vertical connections between oscillators corresponding to different frequency channels and different segments, but within the same time frame. At each time frame, an F0 is estimated from the pooled correlogram (2) and this is used to classify frequency channels into two categories: a set of channels, $P$, that are consistent with the F0, and a set of channels that are not. Specifically, given the delay $\tau_m$ at which the largest peak occurs in the pooled correlogram, for each channel $i$ at time frame $j$, $i \in P$ if

$$A(i, j, \tau_m) / A(i, j, 0) > \theta_d \qquad (9)$$

Since $A(i,j,0)$ corresponds to the energy in correlogram channel $i$ at time $j$, (9) amounts to classification on the basis of an energy threshold. We use $\theta_d$ = 0.95. The delay $\tau_m$ can be found by using a winner-take-all network, although for simplicity we apply a maximum selector in the current implementation.

The classification process described above operates on channels, rather than segments. As a result, channels within the same segment at a particular time frame may be allocated to different pitch categories. Since we do not allow segments to be decomposed, we enforce a rule that all channels of the same frame within each segment must belong to the same pitch category as that of the majority of

channels. After this conformational step, vertical connections are formed such that, at each time frame, two oscillators of different segments have mutual excitatory links if the two corresponding channels belong to the same pitch category; otherwise they have mutual inhibitory links. $S_{ij}^{v}$ is set to -0.5 if (*i*,*j*) receives an input from its inhibitory links; similarly, $S_{ij}^{v}$ is set to 0.5 if (*i*,*j*) receives an input from its vertical excitatory links.

At present, our model has no mechanism for grouping segments that do not overlap in time. Accordingly, we limit operation of the second layer to the time span of the longest segment. After forming lateral connections and trimming by the longest segment, the network is numerically solved using the singular limit method [6].

Figure 4 shows the response of the second layer to the mixture of speech and trill telephone. The figure shows two snapshots of the second layer, where a white pixel indicates an active oscillator and a black pixel indicates a silent oscillator. The network quickly (in the first cycle) forms two synchronous blocks, which desynchronize from each other. Figure 4A shows a snapshot taken when the oscillator block (stream) corresponding to the segregated speech is in the active phase; Figure 4B shows a subsequent snapshot when the oscillator block (stream) corresponding to the trill telephone is in the active phase. This successive 'pop-out' of streams continues in a periodic fashion. Hence, the activity in this layer of the network embodies the result of the ASA process; the individual sources in an acoustic mixture have been separated using F0 information and represented by oscillatory correlation.

### 2.6. Resynthesis

The last stage of the model is a resynthesis path. For each block of oscillators (stream), resynthesis proceeds by reconstructing a waveform from only those time-frequency regions in which the corresponding oscillators are in the active phase. Phase-corrected output from the gammatone filterbank is divided into 20 ms sections, overlapping by 10 ms and windowed with a raised cosine. A binary weighting is then applied to each section, which is unity if the corresponding oscillator is in its active phase, and zero otherwise. These weighted filter outputs are summed across all frequency channels to yield a resynthesized waveform.

## 3. Evaluation

The model has been evaluated using a corpus of 100 mixtures of speech and noise described by Cooke [4]. The mixtures are obtained by adding the waveforms of each of ten voiced utterances to each of ten intrusive sounds. Since separate speech and noise waveforms are available, a signal-to-noise ratio (SNR) can be computed for each mixture. Also, the SNR can be estimated *after* processing by the model, since separated speech and noise waveforms can be obtained by resynthesis (for details, see [11]).

The SNR before and after segregation by the model is shown in Figure 5, averaged across the ten utterances for each noise condition. Dramatic improvements in SNR are obtained when the interfering noise is narrowband (1 kHz tone and siren); such intrusions tend to be represented as a single segment because of their compact spectral structure,

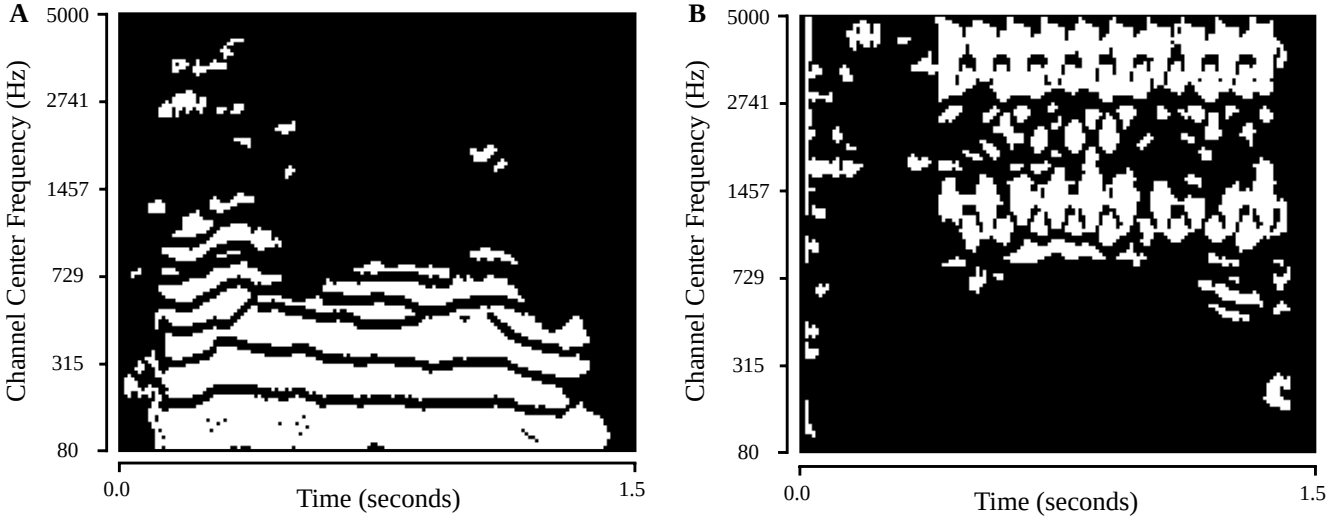


**Figure 4**: The result of separation for the mixture of speech and telephone. **A**. Snapshot showing the activity of the second layer shortly after the start of simulation. White pixels indicate active oscillators; at this time, the active oscillators correspond to the speech stream. **B**. Another snapshot, taken shortly after A. Active oscillators correspond to the telephone stream.

and hence they can be segregated very effectively from the speech source. Informal listening tests suggest that the intelligibility of the resynthesized speech is good, particularly for narrowband intrusions. Also, we have quantified the percentage of speech energy that is recovered by the segregation process; typically, between 55% and 80% of the speech energy is recovered, depending on the type of intrusion [11].

## 4. Discussion

A significant feature of the model proposed here is that each stage has a neurobiological foundation. The peripheral auditory model is based upon the gammatone filter, which is derived from physiological measurement of auditory nerve impulse responses. Similarly, our mid-level auditory representations are consistent with the physiology of the higher auditory system [1]. Overall, the model is based on a framework – oscillatory correlation – which is supported by recent neurophysiological findings.

Our model can potentially be implemented as a real-time system. The oscillator network performs ASA in a parallel and distributed fashion, where each oscillator behaves autonomously and in parallel with all the other oscillators in the network. Although there are issues regarding real-time implementation of the model that need to be resolved (see [11]), there is a real possibility that the oscillator network, with its continuous-time dynamics, can be implemented on an analog VLSI chip. This feature is particularly attractive because the high speed and compact size of analog VLSI are both required for real-time implementation.

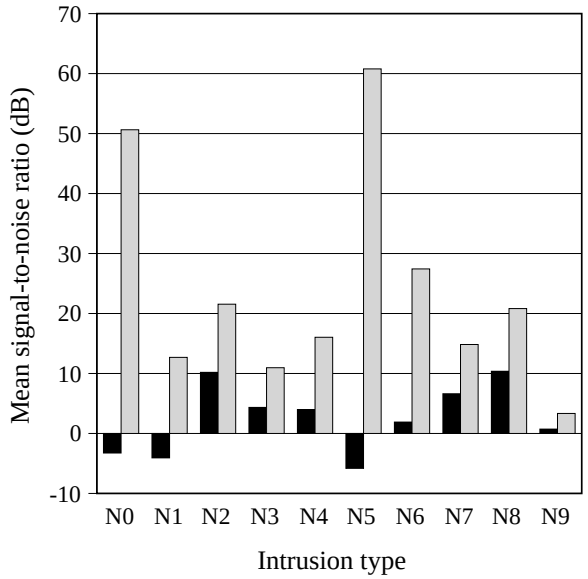


**Figure 5**: SNR before (black bar) and after (grey bar) separation by the model. Results are shown for voiced speech mixed with ten intrusions (N0 = 1 kHz tone; N1 = random noise; N2 = noise bursts; N3 = 'cocktail party' noise; N4 = rock music; N5 = siren; N6 = trill telephone; N7 = female speech; N8 = male speech; N9 = female speech).

In conclusion, we have studied ASA from a neuro-computational perspective and have proposed a multi-stage model for segregating speech from interfering sounds. The model employs a two-layer oscillator network to perform segmentation and grouping of the acoustic input. Lateral connections within the network encode proximity in time, proximity in frequency and harmonicity. Segments arise in the first layer of the network, which correspond to connected time-frequency regions that are atomic and perceptually relevant. Streams then emerge from the second layer of the network through the grouping of segments that have a common F0. The model is founded on neurobiology, and has been systematically evaluated using a corpus of voiced speech mixed with a variety of interfering sounds.

## References


[1] G. J. Brown & M. Cooke, 'Computational auditory scene analysis', *Computer Speech and Language*, **8**, pp. 297-336, 1994.

[2] A. S. Bregman, *Auditory scene analysis*. Cambridge MA: MIT Press, 1990.

[3] G. J. Brown & D. L. Wang, 'Modelling the perceptual segregation of double vowels with a network of neural oscillators', *Neural Networks*, **10**, pp. 1547-1558, 1997.

[4] M. Cooke, *Modelling auditory processing and organization*. Cambridge U.K.: Cambridge University Press, 1993.

[5] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*. Ph.D. Dissertation, MIT Department of Electrical Engineering and Computer Science, 1996.

[6] P. S. Linsay & D. L. Wang, 'Fast numerical integration of relaxation oscillator networks based on singular limit solutions', *IEEE Trans. Neural Net.*, **9**, pp. 523-532, 1998.

[7] D. Terman & D. L. Wang, 'Global competition and local cooperation in a network of neural oscillators', *Physica D*, **81**, pp. 148-176, 1995.

[8] C. von der Malsburg, *The correlation theory of brain function*, Internal Report 81-2, Max-Planck-Institute for Biophysical Chemistry, 1981.

[9] D. L. Wang, 'Primitive auditory segregation based on oscillatory correlation', *Cognit. Sci.*, **20**, pp. 409-456, 1996.

[10] D. L. Wang, 'Object selection based on oscillatory correlation', *Neural Networks*, in press, 1999.

[11] D. L. Wang & G. J. Brown, 'Separation of speech from interfering sounds based on oscillatory correlation', *IEEE Trans. Neural Net.*, in press, 1999.

[12] D. L. Wang & D. Terman, 'Image segmentation based on oscillatory correlation', *Neural Comp.*, **9**, pp. 805-836 (for errata see *Neural Comp.*, **9**, pp. 1623-1626, 1997), 1997.