frontiers in Neuroscience

ORIGINAL RESEARCH
published: 02 September 2015
doi: 10.3389/fnins.2015.00309

Sound stream segregation: a neuromorphic approach to solve the "cocktail party problem" in real-time

Chetan Singh Thakur[1], Runchun M. Wang[1], Saeed Afshar[1], Tara J. Hamilton[1], Jonathan C. Tapson[1], Shihab A. Shamma[2] and André van Schaik[1]*

[1] *Biomedical Engineering and Neuroscience, The MARCS Institute, University of Western Sydney, Sydney, NSW, Australia,* [2] *Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, MD, USA*

OPEN ACCESS

Edited by:
*Themis Prodromakis,
University of Southampton, UK*

Reviewed by:
*Siddharth Joshi,
University of California, San Diego, USA
Leslie Samuel Smith,
University of Stirling, UK*

***Correspondence:**
*Chetan Singh Thakur,
Biomedical Engineering and Neuroscience, The MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, Sydney, NSW 2751, Australia
c.singhthakur@uws.edu.au*

Specialty section:
This article was submitted to Neuromorphic Engineering, a section of the journal Frontiers in Neuroscience

Received: *19 May 2015*
Accepted: *18 August 2015*
Published: *02 September 2015*

Citation:
Thakur CS, Wang RM, Afshar S, Hamilton TJ, Tapson JC, Shamma SA and van Schaik A (2015) Sound stream segregation: a neuromorphic approach to solve the "cocktail party problem" in real-time. Front. Neurosci. 9:309. doi: 10.3389/fnins.2015.00309

The human auditory system has the ability to segregate complex auditory scenes into a foreground component and a background, allowing us to listen to specific speech sounds from a mixture of sounds. Selective attention plays a crucial role in this process, colloquially known as the "cocktail party effect." It has not been possible to build a machine that can emulate this human ability in real-time. Here, we have developed a framework for the implementation of a neuromorphic sound segregation algorithm in a Field Programmable Gate Array (FPGA). This algorithm is based on the principles of temporal coherence and uses an attention signal to separate a target sound stream from background noise. Temporal coherence implies that auditory features belonging to the same sound source are coherently modulated and evoke highly correlated neural response patterns. The basis for this form of sound segregation is that responses from pairs of channels that are strongly positively correlated belong to the same stream, while channels that are uncorrelated or anti-correlated belong to different streams. In our framework, we have used a neuromorphic cochlea as a frontend sound analyser to extract spatial information of the sound input, which then passes through band pass filters that extract the sound envelope at various modulation rates. Further stages include feature extraction and mask generation, which is finally used to reconstruct the targeted sound. Using sample tonal and speech mixtures, we show that our FPGA architecture is able to segregate sound sources in real-time. The accuracy of segregation is indicated by the high signal-to-noise ratio (SNR) of the segregated stream (90, 77, and 55 dB for simple tone, complex tone, and speech, respectively) as compared to the SNR of the mixture waveform (0 dB). This system may be easily extended for the segregation of complex speech signals, and may thus find various applications in electronic devices such as for sound segregation and speech recognition.

Keywords: temporal coherence, cochlea, cocktail party problem, FPGA, machine-based speech recognition

Introduction

Humans can segregate sound sources and focus their attention on specific sounds, while filtering out a range of other background sounds with ease (Bregman, 1990). This attentional ability is known as the "cocktail party effect" (Cherry, 1953), for it enables one to focus on a single conversation in a noisy room. Intentions and attention play a key role in segregating complex auditory scenes into foregrounds and backgrounds, by directing sensory and cognitive processes to pertinent auditory features (Woldorff et al., 1993; Shinn-Cunningham, 2008; Elhilali et al., 2009b). Various acoustic characteristics of sound such as pitch, frequency, timbre, and spatial location may be the focal point of auditory attention in selective hearing (Lee et al., 2012).

The human auditory system is a highly efficient and sensitive sensory system. Sound waves collected in the outer ear travel through the middle ear to reach the cochlea, which serves as the front-end of the auditory system (Guinan et al., 2012). Different locations on the basilar membrane (BM) of the cochlea vibrate in response to specific sound frequencies, thus enabling the cochlea to function as a frequency spectrum analyser (Gold and Pumphrey, 1948; Plomp, 1964). The mechanical vibrations are transduced by the inner hair cells into neural impulses along the auditory nerve (LeMasurier and Gillespie, 2005). Subsequent processing in the brain includes pitch perception for complex tones (Hall and Plack, 2009), sound localisation (Grothe et al., 2010), sound segregation (Carlyon, 2004) and identification (Alain et al., 2001).

Machine-based speech recognition systems have so far not been able to match the functional efficiency of biological auditory systems (Lyon, 2010). It is especially desirable to develop a machine-based auditory system with the ability to segregate sound sources. Such systems would have a large number of applications such as speech recognition in a noisy background, source localisation, sound-based human computer interaction, the design of autonomous robots with the ability to hear and respond to sounds, mobile devices that can seamlessly use voice commands and the design of intelligent hearing aids, as sound segregation rapidly deteriorates in hearing impaired individuals. Several computational models have been proposed to solve the cocktail party problem of speech recognition in a noisy environment (Cooke and Ellis, 2001; Cooke et al., 2010; Shao et al., 2010; Shamma et al., 2011). However, all of these are software models and highly computationally intensive, which cannot process sound in real-time.

Here, we utilize a temporal coherence model of sound stream segregation (Krishnan et al., 2014) and adapt it for hardware implementation in a Field Programmable Gate Array (FPGA). The model works on the principle of temporal coherence (Elhilali et al., 2009a), meaning that the different types of features (e.g., pitch, location, loudness, etc.) belonging to a sound source fluctuate in strength at exactly the same times, while those belonging to different sound sources are rarely synchronized. The model also incorporates the feature that neural response patterns generated by the auditory features of a sound source are highly correlated (Shamma et al., 2011). Together, these principles allow separation of target speech from background noise using attentional mechanisms. A unique feature of the temporal coherence model is that it does not require any training or prior knowledge of target signal and background noise. Further, it is worth mentioning that since this model is highly computationally intensive, an FPGA implementation that runs in real-time is useful.

The temporal coherence model consists of two stages—feature extraction and clustering (Krishnan et al., 2014). The feature extraction stage employs an electronic cochlea along with rate filters. For the hardware implementation, we employ a neuromorphic model of the cochlea called CAR-FAC (Cascade of Asymmetric Resonators with Fast-Acting Compression) (Lyon, 2011). We have previously implemented the BM module of the CAR-FAC model in an FPGA (Thakur et al., 2014a). Here, we have further improved the FPGA implementation by incorporating a simplified inner hair cell module in addition to the BM module. This electronic cochlea, with the BM, and inner hair cell modules, extracts the auditory features of input sound stimuli. The rate filters then carry out a multi-resolution analysis of the cochlear output. The output of each rate filter is referred to as a channel. In the clustering stage, correlation among the channels is computed to identify coherent features, and the attention signal is utilized to select target features that serve as a mask for segregating and reconstructing source of interest.

We have tested our sound segregation model by using an alternating-tone sequence and an alternating-harmonic sequence in the hardware model, and a mixture of alternating speech in the software model. The FPGA system was able to segregate the target sound stream from the mixture of sounds in real-time. Our work demonstrates that the temporal coherence model of auditory filtering can be implemented on an FPGA for segregation of sound sources in real-time. The FPGA implementation of the temporal coherence model described here may find applications in various machine-hearing applications. This paper is organized as follows: the computational model and the system architecture are described in the Materials and Methods Section. Sound segregation from pure tone mixtures, complex tone mixtures and speech mixtures are presented in the Results Section, which is followed by the Discussion Section.

Materials and Methods

Temporal Coherence Model of Auditory Streaming

We have used a biological plausible temporal coherence model for our hardware implementation (Krishnan et al., 2014). This model exploits two characteristics of a sound source for auditory filtering—first, the acoustic features of a sound source are coherently modulated in a temporal manner, and second, the neural patterns generated in response to a sound source are highly correlated. The guiding principle, based on one of the Gestalt Principles, is that auditory channels highly correlated over a short time period represent a common fate (Bregman, 1990; Blake and Lee, 2005). This algorithm does not require any training or prior knowledge of the sound sources. This model also employs attention to specific attributes of a source to segregate it from the background. The model comprises of two stages:

Feature Extraction Stage

First, a cochlear model is used to compute an auditory spectrogram of the input sound. Next, features extracted from the cochlear stage are gone through a temporal analysis with multi- rate filters. The filters are selective to different temporal modulation rates ranging from slow to fast (2, 4, 8, 16 Hz), covering the cortical time-scale. In the current FPGA solution, we have implemented only one rate filter of 4 Hz, since it is sufficient for the tested sound mixtures. We can easily extend our system to multi-rate filters, and as the rate filters work in parallel, increasing their number will not affect the performance of the system.

Clustering Stage

In the actual model, a correlation matrix is calculated by computing the outer product of the multidimensional channels (output of the rate filters), and is updated for each time-step. The pair-wise correlation between channels is indicative of their degree of synchrony. The next process is to employ an attention signal to select correlation coefficients for the target stream, which then act as a mask to segregate and reconstruct the target stream from uncorrelated streams. In our system, we use an attention signal as an input, which avoids the calculation of the correlation matrix and reduces the computational burden (Section Attention Signal and Mask for Reconstruction). The attention signal is an exclusive feature present only in the target stream and it is used computationally to facilitate the identification of the target stream. The attention signal acts as an anchor to segregate the target stream and this anchor could be a pitch signal, an envelope of the target speech, or an envelope of lip movement etc.

Design Methodology

Figure 1 depicts the block diagram for the FPGA implementation of the temporal coherence model. The cochlea receives the auditory input, and transforms the sound signal into a frequency spectrum. The cochlear output is passed on to the rate filters that perform a temporal multi-rate analysis, integrating the history of cochlear channel responses. An attention signal, which is a part of the target stream influences stream formation by initiating binding. A correlation matrix that measures the similarity of auditory responses across channels needs to be computed, but this is computationally very costly for FPGA implementation. To ease the computational burden of this calculation, we use the attention signal to choose particular channels of interest. Pair-wise correlation of all channels is computed with the attention signal, which acts as a mask in stream reconstruction and is referred to as correlation column. Negative values of correlation coefficient indicate that other tones are highly uncorrelated with the attention signal. This allows us to compute only a single column of the correlation matrix, which represents correlation of the attention channel with all other channels. This single column is referred to as a mask. Finally, the mask is used to separate the target stream from the background interference and the filtering representations are converted back to the acoustic domain. Each of the modules of the architecture is described in detail below.

Cochlea

The cochlea functions as a front-end analyser for sound by transforming the sound input into a frequency spectrum. We utilize the CAR-FAC model of the cochlea (Lyon, 2011) in our system, as its speed and efficiency is superior to the more conventional parallel filter bank approach (Lyon, 1998). The asymmetric resonators in the cascade of asymmetric resonators (CAR) are quasi-linear transfer functions that model the motion of the BM. The outer hair cell module provides dynamic non-linearity or fast-acting compression (FAC). The inner hair cells are encoded using sigmoidal or half-wave rectification function, and introduce non-linearity in the outputs of the CAR. They function to connect the mechanical waves on the BM to neural signals on the auditory nerve.

Biquadratic filters represent asymmetric resonators in the CAR model of the BM. The number of filter sections and their coefficients are optimized to match a linearised model of the

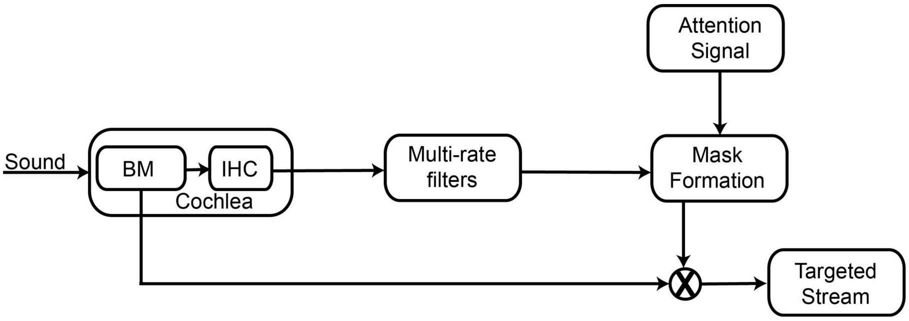

FIGURE 1 | Schematic of the FPGA implementation of the temporal coherence model of sound segregation. The cochlea receives the auditory input, and transforms the sound signal into a frequency spectrum. The cochlear output of each frequency channel is passed on to a rate filter that performs a temporal analysis of the rate of variation of the amplitude of the cochlear output at that frequency channel. An externally determined attention signal, which is temporally correlated with the target stream, is correlated with the output of the rate filters to create a mask for stream segregation. This mask is used with the output of the cochlear frequency channels to separate the target stream from the background interference.

cochlea. The filter poles are equally spaced along the length of the cochlea. For a normalized position x along the cochlea, the pole frequency, f, is obtained using the Greenwood function (Greenwood, 1990):

$$f = 165.4(10^{2.1x} - 1)$$

where, x varies from 0 at the apex of the BM, to 1 at its base. **Figure 2** shows a biquadratic filter section. Parameters a_0 and c_0 are functions of position x, and represent the analog pole position in the zero-damping case. An explicit parameter, r, can be modulated to vary the pole and zero radius in the z plane, thus modulating the damping factor. The relationship between these parameters is given using the following equations:

$$a_0 = \cos(\theta_R) = a/r$$
$$c_0 = \sin(\theta_R) = c/r$$

where, θ_R is the normalized pole ringing frequency or pole angle in the z plane. The transfer function is given as:

$$\frac{Y}{X} = g\frac{z^2 + (-2a_0 + hc_0)\,rz + r^2}{z^2 - 2a_0 rz + r^2}$$

The h coefficient controls the difference between zeros and the pole frequency, and the g coefficient is used to adjust the overall gain. The zeros will be at the same radius r as the poles, if h is small enough that the zeros remain complex. For high-frequency channels, $\cos\theta_R < 0$. In that case:

$$h_0 < \frac{2 + 2a_0}{c_0}$$

To get unity gain at DC, we can solve for g:

$$g = \frac{1 - 2a_0 r + r^2}{1 - (2a_0 - hc_0)\,r + r^2}$$

The combination of cascaded stages creates a family of filters at the output taps between the stages. The resulting filters may have high peak gains, depending on the stage damping parameters.

We have previously implemented the CAR module of the CAR-FAC model in FPGA, which represents the cochlear basilar membrane (BM) (Thakur et al., 2014a). Here we have added a simplified inner hair cell (IHC) model using a half-wave rectifier at the output of the CAR filters followed by two first order low-pass filters with a 8 kHz cut-off frequency, to generate an approximate neural activity pattern. Although this is a very simplified model of the IHC function, it suffices for our application.

Rate Filter

While the cochlea is selective only for the frequency content of sound, cortical auditory neurons are additionally selective for temporal modulations found in natural sounds (Kowalski et al., 1996; Theunissen et al., 2000; Lu et al., 2001; Escabí et al., 2003; Woolley et al., 2005). Hence, a temporal analysis of the auditory spectrogram generated by the cochlea is carried out with multi-range dynamics covering a frequency range of 2–16 Hz. For this, the output of each cochlear channel is connected to a rate filter. In our FPGA design, we have implemented only a 4 Hz rate filter using two sets of low pass filters (LPF) and high pass filters (HPF), each with the same cut-off frequency of 4 Hz and connected in series to obtain steeper slopes. The LPF and HPF are represented using the following equations:

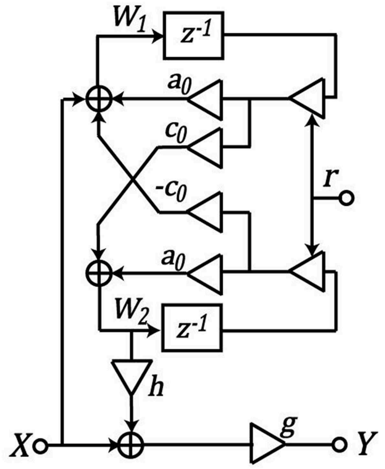

FIGURE 2 | Two-pole–two-zero filter in the CAR model. a_0 and c_0 are functions of position x along the cochlea. h controls the difference between zeros and the pole frequency, g is used to adjust the overall gain. X represents the input signal, Y represents the output signal, and W_1 and W_2 are internal state variables. Figure adapted from Thakur et al. (2014a).

$$bpl_t = ((1 - c_l) * bpl_{t-1} + c_l * \vartheta_t$$
$$bph_t = ((1 - c_h) * bph_{t-1} + c_h * (bpl_t - bph_{t-1})$$

where, ϑ_t is input to the rate filters coming from the cochlea at time t. bpl and bph represent the LPF and HPF function, respectively. c_l and c_h denote the coefficients for LPF and HPF corresponding to a cut-off frequency of 4 Hz, respectively, and are given by:

$$c_l = c_h = 2\pi(\frac{4}{f_s})$$

where, f_s is the sampling frequency.

Attention Signal and Mask for Reconstruction

Attention is a cognitive process that allows one to focus on a group of features of an auditory stimulus. This enhances their relative amplitudes as compared to unattended stimuli, thus playing an important role in auditory stream perception and segregation (Snyder et al., 2006; Bidet-Caulet et al., 2007). It has been shown that there are attention-dependent changes in the spectro-temporal receptive fields of the auditory cortex, such as frequency selective enhancement (Fritz et al., 2003, 2007). Additionally, attention can influence streaming by modulating the temporal coherence of neural populations (Niebur et al., 2002). Attention is an exclusive feature present only in the target stream and it is used computationally to facilitate identification of the target stream.

The correlation vectors are computed as the product of the rate filter channels with the attention channels, and these vectors act as a mask. Only the instantaneous correlation across all pairs of channels is considered. Currently, we have implemented only the 4 Hz rate filter. In the case of multi-rate filters, we sum all the correlation coefficients of each cochlear channel and zero the negative coefficients of the resultant vector, which represents the mask. Usage of an attention signal eases the computational burden of this calculation, otherwise we would have to calculate the complete correlation matrix followed by decomposition into principle components using a non-linear auto-encoder (Krishnan et al., 2014). The computed mask is used for the reconstruction or segregation of the speech of interest. Since the time-scale of the rate filters is very slow (<20 Hz), and FPGA processing is very fast (∼hundreds of MHz), we are able to use a single rate filter of 4 Hz across all the channels using a time-multiplexing technique.

Stream Reconstruction

In stream reconstruction, the target stream present in the input auditory stimulus is resynthesised computationally using the output of early auditory and cortical stages. A detailed mathematical explanation of stream reconstruction is published by Chi et al. (2005). It should be noted that reconstruction of sound is not a biological process, but for sound segregation applications we need to reconstruct the target sound. In this final stage, point to point multiplication of the formed mask with the output of the BM channels of the cochlea is carried out to segregate the target stream from the background, and to reconstruct the stream. This process would require many multipliers, but our FPGA implementation requires only one multiplier, as we are using a time-multiplexing technique. Rate filtering introduces some latency and each BM filter also introduces a different phase delay. Currently, we are using a 100 MHz system clock. The BM block introduces a delay of 49 clock cycles (490 ns) and the total delay introduced by the IHC and the rate filter is 45 clock cycles (450 ns). These could be compensated for by delaying each BM output channel by an appropriate amount before reconstruction, but in the current implementation, we have not done this as the quality of the reconstructed signal without delay compensation is good enough for our purpose.

FPGA Implementation

First, we simulated a software floating-point implementation of the model in Python. Next, we adapted the Python code for fixed-point implementation, and determined the word length of the input, output, and internal variables required for FPGA implementation without loss of accuracy. The system architecture for the hardware implementation is shown in **Figure 3**. Here, we have used time-multiplexing to share the hardware resources on the FPGA. We have implemented a single hardware block as shown in **Figure 3**, and reused it for all the 70 filter sections, given an audio sampling clock of 8 KHz and a system clock of 100 MHz (Thakur et al., 2014b). The cochlear filter section that is processed at a particular time is determined by a global state machine. The latter also controls the coefficients and data for the filter section. For each filter section, the coefficients *a*, *c*, *g*, and *h* are calculated externally, and uploaded into FPGA memory from a file at the start of execution. Each input sound sample passes via the global state machine to the BM for processing. Two parallel state machines are contained in the BM block. These control and calculate internal variables W1 and W2, which are further used to calculate the transfer function, *Y/X*. A delay element (z^{-1} block in **Figure 2**) requires the variables, W1 and W2, to be stored for each filter section. The output of the BM is passed on to the inner hair cell and rate filter block for each filter section, as explained in Section Rate Filter. The output of each cycle of operation of one filter section is passed to the global state machine. This serves as the input for the next filter section stage, resulting in a cascading of the filter sections. The completion of processing of an input sample by one filter, inner hair cell and rate filter section is denoted as "Done," while "Done_Sample" by the global state machine denotes the completion of processing by all the filter sections including inner hair cell and rate filter. We have successfully implemented the proposed system on an Altera Cyclone V FPGA (on a Terasic Cyclone GX starter kit) with the utilization area as shown in **Table 1**.

Results

Here, we present test results for the performance of the system. We first tested the model on typical auditory stimuli (single and complex tones) widely used to study the perceptual formation of auditory streams. Further, we tested the model on complex speech sound for speaker separation. The results presented in Sections Segregation of Streams from Alternating-tone Sequence and Segregation of Streams from Alternating Complex Tones were obtained from hardware implementation, and those in Section Segregation of Speech from Mixtures were obtained from simulation of the model in software, which is a replica of our hardware model. The performance of the model is quantified by comparing the original separate streams to the segregated stream. The signal-to-noise ratio (SNR) is computed as:

$$\text{SNR\_segregated\_stream} = 10log\left(\frac{|S1{*}O1|^2}{|S1{*}O2|^2}\right) \quad (1)$$

$$\text{SNR\_mixture} = 10log\left(\frac{|M{*}O1|^2}{|M{*}O2|^2}\right) \quad (2)$$

where, *S1* is the segregated (output) stream; *O1* and *O2* are the original separate streams; and *M* is the mixture stream provided as the input. The "$*$" operator represents the dot product of the two sound vectors.

Segregation of Streams from Alternating-tone Sequence

An alternating-tone sequence is composed of two continuously repeated pure tones of different frequencies, A and B. Such

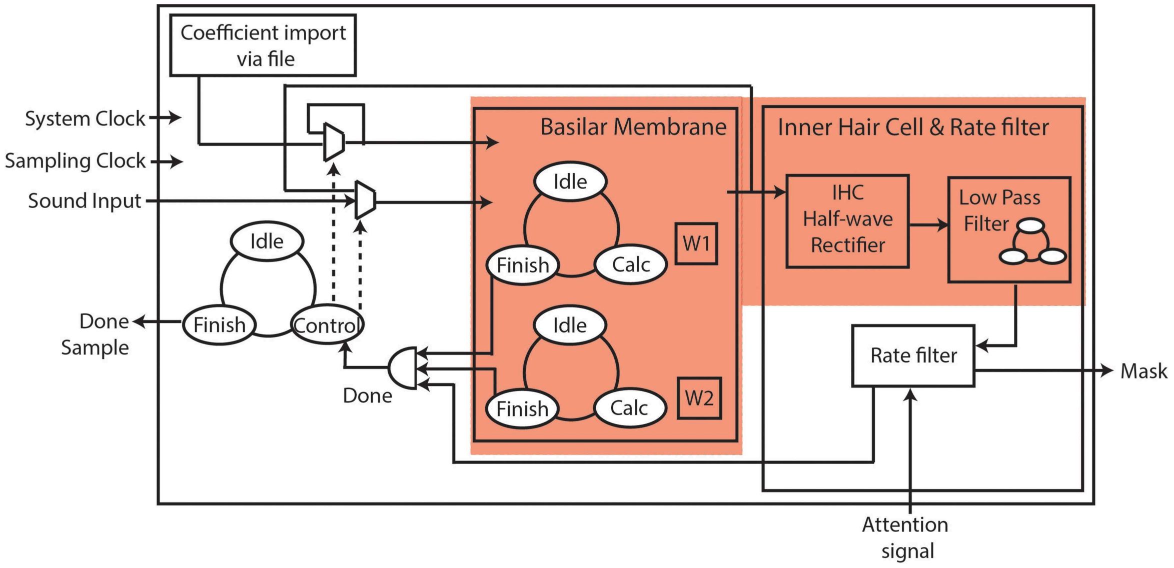

FIGURE 3 | System Architecture for FPGA implementation. The system includes a basilar membrane block, and an inner hair cell and rate filter block. Each block is only instantiated once, but is time multiplexed to create 70 instances in our system. The cochlea is shown using a red background.

TABLE 1 | Device utilization Altera Cyclone-V 5CGXFC7C7F23C8.

Adaptive Logic Modules (ALMs)	Total registers	DSPs
1793/56480 (3%)	3899	10/156 (6%)

a sequence is commonly used in studies of auditory stream segregation. The frequency separation between the two tones, Δf, and the inter-tone interval, ΔT, determine the percept evoked by such sequences. If Δf is small and ΔT is long, the sequence is perceived as a single stream of tones alternating in frequency (ABAB). This phenomenon is known as temporal coherence (Van Noorden and Schouten, 1975). On the contrary, if Δf is large and ΔT is short, the sequence is perceived as two separate streams of tones of constant frequencies (A's and B's). This phenomenon is known as stream segregation. We have used a sequence composed of frequencies 440 and 1000 Hz, with a presentation rate (ΔT) of 4 Hz. **Figure 4** shows the result for this alternating-tone sequence. The input mixture is transformed into an auditory spectrogram using cochlea as described in Section Cochlea, and a particular channel that is a feature of the targeted stream is used as attention signal. Pair-wise correlation of all channels is computed with the attention signal which we refer as correlation column. The negative value of correlation coefficient indicates that the tone of frequency 1000 Hz is highly uncorrelated with the attention signal. The targeted stream of 440 Hz tone is segregated, and reconstructed using mask which is created based on the attention signal. The effectiveness of segregation is measured using Equations (1) and (2). The SNR for the segregated stream is calculated as 91 dB compared to the input mixture of two simple tones, mixed at 0 dB.

Segregation of Streams from Alternating Complex Tones

Here, we have used a sequence of two complex tones alternating with a presentation rate of 4 Hz. This presentation rate lies within the range (2–20 Hz) of the presentation rate of auditory signals over which auditory stream formation takes place in the brain (Fishman et al., 2004; Chakalov et al., 2013). The first complex tone is a mixture of pure tones of frequencies 300 and 900 Hz, and the second complex tone consists of tones of frequencies 600 and 1500 Hz. The results are shown in **Figure 5**. The tone with frequency of 600 Hz is used as the attention signal. As this tone is temporally coherent with the tone of frequency 1500 Hz, all two tones become segregated as one stream. It can be seen in **Figure 5** that all frequencies comprising the first mixture—600 and 1500 Hz, show positive correlation coefficient because they are temporally coherent with the attention signal. In contrast, the other complex tone of frequencies 300 and 900 Hz show negative correlation coefficients, suggesting that they belong to a different acoustic source and are incoherent with the attention signal. The targeted stream is segregated and reconstructed using the mask generated based on the attention signal. The effectiveness of segregation is measured using Equations (1) and (2). The SNR for the segregated stream is calculated as 77 dB compared to the input mixture of two complex tones, mixed at 0 dB.

Segregation of Speech from Mixtures

Here, we have used a mixture of two female utterances, one saying "good morning" and the other saying "game over." The target speech is the female speech corresponding to "good morning." Our system needs an attention signal, which should be present exclusively in the target speech. Normally, this might be channels tuned roughly near the pitch range, or location responses

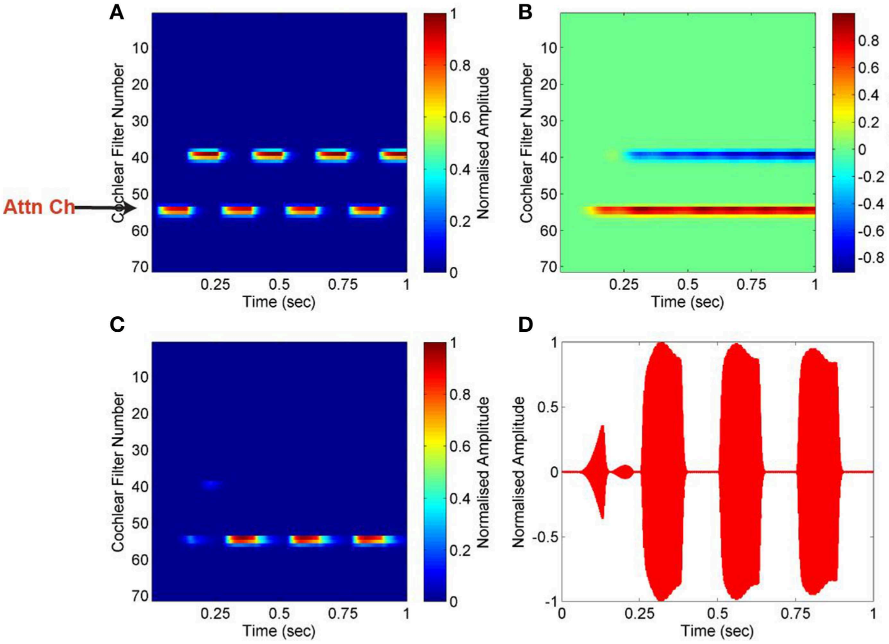

FIGURE 4 | Segregation of alternating-tone sequence. (A) Auditory spectrogram of the input mixture of two alternating pure tones (440 and 1000 Hz). A low cochlear filter number represents high frequency, and *vice versa*. One particular channel, of frequency 440 Hz, is used as the attention signal (Attn Ch). (**B)** Pair-wise correlation of all channels are computed using the attention signal. The negative value of correlation coefficient (*blue*) indicates that the second tone is highly uncorrelated with the attention signal. (**C,D)** The targeted stream is segregated, and reconstructed using mask which is created based on the attention signal, shown as a spectrogram in **(C)** and plot in (**D)**.

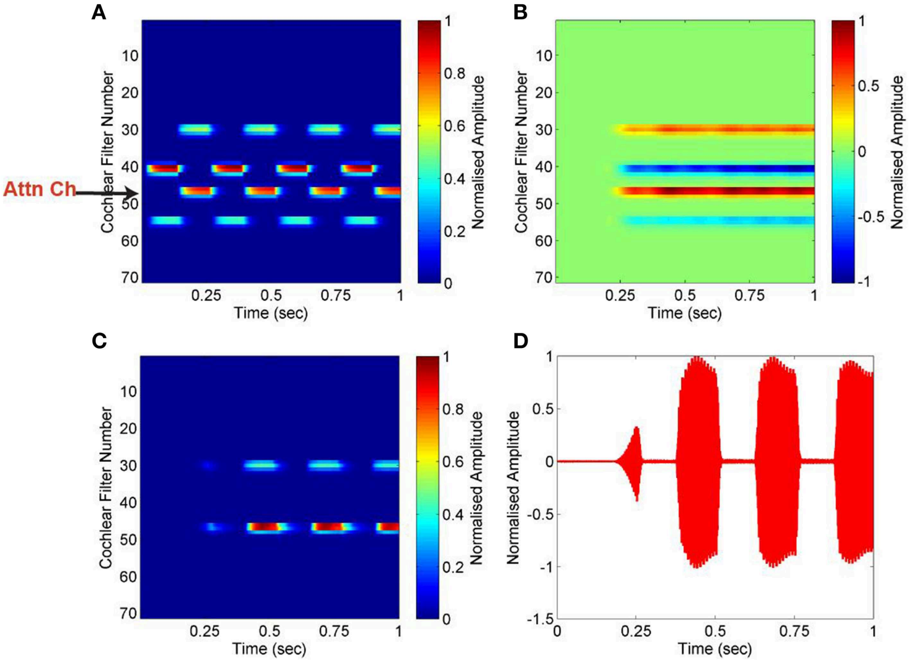

FIGURE 5 | Segregation of alternating complex tones. (A) Auditory spectrogram for an input mixture of complex tones [(300, 900 Hz) and (600, 1500 Hz)]. Low cochlear filter number represents high frequency, and *vice versa*. The channel corresponding to frequency of 600 Hz is used as attention signal (Attn Ch) and is marked with an arrow. (**B)** Pair-wise correlation of all channels is computed using the attention signal. Negative values of correlation coefficient (*blue*) indicate that other tones are highly uncorrelated with the attention signal. (**C,D)** The targeted stream (600, 1500 Hz) is segregated and reconstructed using mask generated based on the attention signal, shown as a spectrogram in (**C)** and plot in (**D)**.

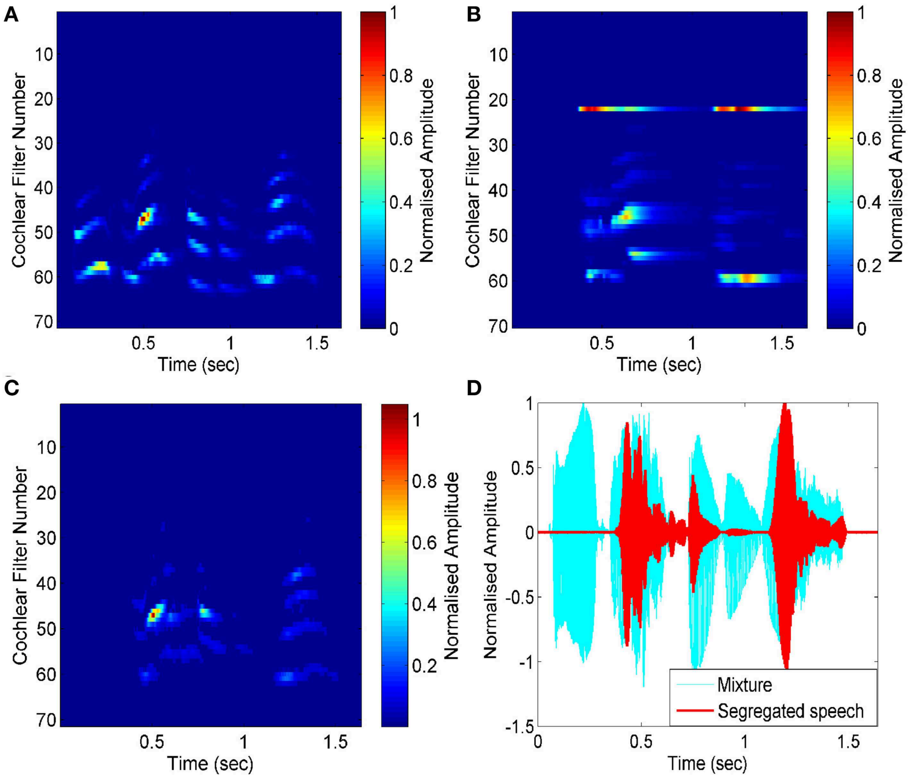

FIGURE 6 | Software simulation for segregation of speech from mixture. (A) Auditory spectrogram for an input mixture of two female utterances ("good morning" and "game over") is transformed into an auditory spectrogram. The envelope of the target speech ("good morning") acts as the attention signal. **(B)** Pair-wise correlation of all channels are computed with the attention signal (channel number 23). **(C)** The segregated target streams are shown in the spectrogram. **(D)** Input mixture (cyan graph) and segregated speech (red graph) are shown.

sensitive to the approximate direction of the target speaker. By using these channels, we can simply use their coherently-modulated power as the cue to the presence of the target speaker. Here, to simulate these extra computations, we have used the envelope (power) of the target speech to act as the attentional signal. The results are shown in **Figure 6**. Pair-wise correlation of all channels is computed with this simulated attention signal, resulting in the necessary mask in stream reconstruction. As shown in **Figure 6**, we are able to segregate the target speech signal from the speech mixture efficiently. The effectiveness of segregation is measured using Equations (1) and (2). The SNR for the segregated stream is calculated as 55 dB compared to the input mixture of the two female utterances, mixed at 0 dB.

Discussion

In this work, we have adapted a temporal coherence-based computational model of auditory scene analysis for neuromorphic hardware implementation in FPGA. We have validated our system by testing various sound stimuli such as alternating-tone sequence, alternating-harmonic complexes and mixtures of speech. We show that our FPGA architecture can successfully segregate sound streams in all these cases (see Supplementary Material). Our system implements a neuromorphic model of the segregation of sound sources in real-time. The system is easily scalable to incorporate higher number of cochlear channels and rate filters, and thus may serve as a feasible solution to the cocktail party problem.

The temporal coherence algorithm differs from other computational systems of auditory stream segregation in its close correspondence to the neurobiological cortical mechanisms of hearing (Shamma et al., 2011). The algorithm also requires no prior information or training on the sources, and can gracefully incorporate and benefit from attention as a criterion for stream segregation. Research has shown that attention plays an important role in segregation by enhancing the perception of a particular stream over others in the auditory scene (Hillyard et al., 1973; Tiitinen et al., 1993; Bidet-Caulet et al., 2007; Elhilali et al., 2009b). Here, we utilize an attention signal as a means to separate target sound from background noise.

In our previous work, we have implemented a cochlear model and demonstrated its ability to process sound in real-time (Thakur et al., 2014b). We have now integrated this cochlear implementation with the temporal coherence model, and this system is a novel prototype for multi-talker speech separation and recognition. Future work will aim to extend the existing model for the segregation of complex speech signals to incorporate various cues such as pitch, frequency, timbre, and spatial location etc. Additionally, the limitations of the current system will be addressed in the future. For example, attention signal is only one of the means to identify a feature of interest in the target stream for segregation. We will incorporate additional features that will make the system more robust to segregate sound. The correlation vector in our system is implemented using a multiplier. This is a simplified model for the auditory cortex, which will be improved by using spike-based computation to group coherent features. Finally, we could also expand the model to explore the effects of switching attention channel between two streams and look at how this switching affects the representation of streams.

Overall, our FPGA implementation of the temporal coherence algorithm establishes that it is feasible to develop a hardware system that can segregate sound sources in real-time. Our FPGA implementation is area efficient, since it reuses a single hardware block for all the filter sections. Our system will have several applications, such as robust front-end processors for automatic speech recognition.

Acknowledgments

The authors acknowledge the organizers of the 2014 NSF-funded Neuromorphic Cognition Engineering Workshop in Telluride, CO, USA where much of this work was conceived, and the Office of Naval Research grant on Figure-ground audio/visual segregation.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnins.2015.00309

References

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., and Grady, C. L. (2001). "What" and "where" in the human auditory system. *Proc. Natl. Acad. Sci. U.S.A.* 98, 12301–12306. doi: 10.1073/pnas.211209098

Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P.-E., Giard, M.-H., and Bertrand, O. (2007). Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *J. Neurosci.* 27, 9252–9261. doi: 10.1523/JNEUROSCI.1402-07.2007

Blake, R., and Lee, S.-H. (2005). The role of temporal structure in human vision. *Behav. Cogn. Neurosci. Rev.* 4, 21–42. doi: 10.1177/1534582305276839

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends Cogn. Sci.* 8, 465–471. doi: 10.1016/j.tics.2004.08.008

Chakalov, I., Draganova, R., Wollbrink, A., Preissl, H., and Pantev, C. (2013). Perceptual organization of auditory streaming-task relies on neural entrainment of the stimulus-presentation rate: MEG evidence. *BMC Neurosci.* 14:120. doi: 10.1186/1471-2202-14-120

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with 2 ears. *J. Acoust. Soc. Am.* 25, 975–979. doi: 10.1121/1.1907229

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887–906. doi: 10.1121/1.194580

Cooke, M., and Ellis, D. P. W. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Commun.* 35, 141–177. doi: 10.1016/S0167-6393(00)00078-9

Cooke, M., Hershey, J. R., and Rennie, S. J. (2010). Monaural speech separation and recognition challenge. *Comput. Speech Lang.* 24, 1–15. doi: 10.1016/j.csl.2009.02.006

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (2009a). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron* 61, 317–329. doi: 10.1016/j.neuron.2008.12.005

Elhilali, M., Xiang, J., Shamma, S. A., and Simon, J. Z. (2009b). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biol.* 7:e1000129. doi: 10.1371/journal.pbio.1000129

Escabí, M. A., Miller, L. M., Read, H. L., and Schreiner, C. E. (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *J. Neurosci.* 23, 11489–11504.

Fishman, Y. I., Arezzo, J. C., and Steinschneider, M. (2004). Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *J. Acoust. Soc. Am.* 116, 1656. doi: 10.1121/1.1778903

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (2007). Auditory attention–focusing the searchlight on sound. *Curr. Opin. Neurobiol.* 17, 437–455. doi: 10.1016/j.conb.2007.07.011

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* 6, 1216–1223. doi: 10.1038/nn1141

Gold, T., and Pumphrey, R. J. (1948). Hearing. I. The cochlea as a frequency analyzer. *Proc. R. Soc. B* 135, 462–491. doi: 10.1098/rspb.1948.0024

Greenwood, D. D. (1990). A cochlear frequency-position function for several species–29 years later. *J. Acoust. Soc. Am.* 87, 2592–2605. doi: 10.1121/1.399052

Grothe, B., Pecka, M., and McAlpine, D. (2010). Mechanisms of sound localization in mammals. *Physiol. Rev.* 90, 983–1012. doi: 10.1152/physrev.00026.2009

Guinan, J. J., Salt, A., and Cheatham, M. A. (2012). Progress in cochlear physiology after Békésy. *Hear. Res.* 293, 12–20. doi: 10.1016/j.heares.2012.05.005

Hall, D. A., and Plack, C. J. (2009). Pitch processing sites in the human auditory brain. *Cereb. Cortex* 19, 576–585. doi: 10.1093/cercor/bhn108

Hillyard, S. A., Hink, R. F., Schwent, V. L., and Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science* 182, 177–180. doi: 10.1126/science.182.4108.177

Kowalski, N., Depireux, D. A., and Shamma, S. A. (1996). Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *J. Neurophysiol.* 76, 3503–3523.

Krishnan, L., Elhilali, M., and Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS Comput. Biol.* 10:e1003985. doi: 10.1371/journal.pcbi.1003985

Lee, A. K. C., Rajaram, S., Xia, J., Bharadwaj, H., Larson, E., Hämäläinen, M. S., et al. (2012). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Front. Neurosci.* 6:190. doi: 10.3389/fnins.2012.00190

LeMasurier, M., and Gillespie, P. G. (2005). Hair-cell mechanotransduction and cochlear amplification. *Neuron* 48, 403–415. doi: 10.1016/j.neuron.2005.10.017

Lu, T., Liang, L., and Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat. Neurosci.* 4, 1131–1138. doi: 10.1038/nn737

Lyon, R. (2010). Machine hearing: an emerging field. *IEEE Signal Process. Mag.* 27, 131–139. doi: 10.1109/MSP.2010.937498

Lyon, R. F. (1998). "Neuromorphic systems engineering," in *Neuromorphic Systems Engineering: Neural Networks in Silicon*, Vol. 447, ed T. S. Lande (Boston, MA: Springer), 3–18.

Lyon, R. F. (2011). "Using a cascade of asymmetric resonators with fast-acting compression as a cochlear model for machine-hearing applications," in *Autumn Meeting of the Acoustical Society of Japan* (Matsue), 509–512.

Niebur, E., Hsiao, S. S., and Johnson, K. O. (2002). Synchrony: a neuronal mechanism for attentional selection? *Curr. Opin. Neurobiol.* 12, 190–194. doi: 10.1016/S0959-4388(02)00310-0

Plomp, R. (1964). The ear as a frequency analyzer. *J. Acoust. Soc. Am.* 36, 1628. doi: 10.1121/1.1919256

Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* 34, 114–123. doi: 10.1016/j.tins.2010.11.002

Shao, Y., Srinivasan, S., Jin, Z., and Wang, D. (2010). A computational auditory scene analysis system for speech segregation and robust speech recognition. *Comput. Speech Lang.* 24, 77–93. doi: 10.1016/j.csl.2008.03.004

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn. Sci.* 12, 182–186. doi: 10.1016/j.tics.2008.02.003

Snyder, J. S., Alain, C., and Picton, T. W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. *J. Cogn. Neurosci.* 18, 1–13. doi: 10.1162/089892906775250021

Thakur, C. S., Hamilton, T. J., Tapson, J., van Schaik, A., and Lyon, R. F. (2014a). "FPGA implementation of the CAR model of the cochlea," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)* (Melbourne, VIC: IEEE), 1853–1856.

Thakur, C. S., Wright, J., Hamilton, T. J., Tapson, J., van Schaik, A. (2014b). "Live demonstration: FPGA implementation of the CAR model of the cochlea," in *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on* (Melbourne, VIC: IEEE). doi: 10.1109/ISCAS.2014.6865170

Theunissen, F., Sen, K., and Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons. *J. Neurosci.* 20, 2315–2331.

Tiitinen, H., Sinkkonen, J., Reinikainen, K., Alho, K., Lavikainen, J., and Näätänen, R. (1993). Selective attention enhances the auditory 40-Hz transient response in humans. *Nature* 364, 59–60. doi: 10.1038/364059a0

Van Noorden, L., and Schouten, J. F. (1975). *Temporal Coherence in the Perception of Tone Sequences*. Institute for Perception Research. Doctoral dissertation, Eindhoven University of Technology, Eindhoven.

Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., Pantev, C., Sobel, D., et al. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proc. Natl. Acad. Sci. U.S.A.* 90, 8722–8726. doi: 10.1073/pnas.90.18.8722

Woolley, S. M. N., Fremouw, T. E., Hsu, A., and Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat. Neurosci.* 8, 1371–1379. doi: 10.1038/nn1536

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.