An Optimum Inexact Design for an Energy Efficient Hearing Aid

Sai Praveen Kadiyala, Aritra Sen, Shubham Mahajan, Quingyun Wang, Avinash Lingamaneni, James Sneed German, Hong Xu, Krishna V Palem, Arindam Basu

Journal of
Low Power Electronics
Vol. 15, 129–143, 2019

An Optimum Inexact Design for an Energy Efficient Hearing Aid

Sai Praveen Kadiyala[1,*], Aritra Sen[2], Shubham Mahajan[2], Quingyun Wang[2], Avinash Lingamaneni[3], James Sneed German[4], Hong Xu[5], Krishna V. Palem[6], and Arindam Basu[2]

[1] *School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore*
[2] *School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore*
[3] *Google, 94043, USA*
[4] *CNRS Aix-Marseille Université, 13007, France*
[5] *School of Humanities and Social Sciences, Nanyang Technological University, 639798, Singapore*
[6] *Department of Computer Science, Rice University, Houston, 77005, USA*

(Received: 1 February 2019; Accepted: 26 April 2019)

Inexact design has proved to be an efficient way of obtaining power savings with a marginal penalty on performance in applications which do not need high degree of output accuracy. Such applications are often found in domains which deal with human sensorial systems. The 'not so perfect' state of human senses like sight, hearing which are important for video and audio related appliances can compensate for the error introduced due to inexact design. An efficient analysis and modelling of these compensation capabilities of human senses can help the designers to build optimum inexact architectures meeting the redefined requirements with low power. In this work, we choose hearing aid as our test application, which is based on human sense of hearing. We estimate a metric Intelligibility, for a set of audio samples, which is obtained from multiple surveys on human subjects to model the sensorial processing. Our methodology uses a novel way of introducing inexactness in an optimum manner. This includes a fine grained analysis of the error that is being introduced in the FIR filters of the DSP present in hearing aid. The resulting inexact FIR filter bank is 1.92× or 2.56× more efficient in terms of power consumed while producing 10% or 20% less intelligible speech respectively when compared with a hearing-aid using exact filters.

Keywords: Inexact Design, Approximate Computing, Hearing Aid, Digital VLSI, PESQ.

1. INTRODUCTION

Digital electronics is playing an ever-increasing role in being part of systems that manipulate information that is sensorially consumed as sound and images. Typically, though the information being produced by these systems is heavily processed by our auditory and visual pathways—processing that compensates for error and other glitches—existing digital system design does not take advantage of this. In particular, what if glitchy images and sound clips are produced at lower energy cost for example, while our compensatory neurocognitive processing can either tolerate or overlook entirely? Dubbed inexact design or approximate computing, this counterintuitive approach has been shown to yield significant gains in the context of hardware for addition, multiplication and DSP primitives derived from these operations,[1–3] also in atmospheric modeling[4,5] MPEG coding,[6,7] recognition and classification tasks[8,9]—the basic principle involving trading the "quality" or accuracy for gains in the energy consumed, area as well as computing speed. The interesting fact about inexact design based architectures is that they can be used as an addition to most of the existing low power design approaches such as adaptive resolution, lossless compression,[10] self clocking,[11] hardware reuse,[12] circuit optimization,[13] truncated coding[14] etc.

In general, the concept of approximate computing (inexact design) is being used in many fields. Energy efficient implementation of approximate computing using resistive switching RAM (RRAM) is highlighted in Ref. [15]. The precision limitation caused by approximation in digital systems and the additional resolution achieved by RRAM is discussed in this work. Usage of approximate computations at block and system level targeting fixed point applications is presented in Ref. [16]. They aimed at developing inexact hardware for functional unit allocation,

*Author to whom correspondence should be addressed.
Email: saipraveen@ntu.edu.sg

resource scheduling and binding algorithms with a special attention towards precision. In multimedia applications compression using discrete cosine transform (DCT) is common. In Ref. [17] the error resilience of JPEG is taken advantage and an approximate method is proposed to carry out low power compression. The effect of low precision fixed point data on training of deep neural networks is studied in Ref. [18]. An energy efficient hardware accelerator is developed in this work which carries out the low precision arithmetic. An approximate method for reconfigurable memory based pattern matching in highly parallel architectures is discussed in Ref. [19]. Here the input patterns are matched with either selective bits or selective pre existing patterns so that matching energy can be minimized.

While approximate computing is especially appealing in the context of systems for digital signal processing, past work falls short of being truly applicable. Specifically, glitches introduced through inexactness have been quantified as arithmetic magnitude—the amount by which the sum or product of two numbers deviates from the correct answer since the operations used are addition and multiplication—as opposed to using metrics that capture their impact on our senses in a natural way. However, there is a chasm between any metric quantifying arithmetic error and its relationship to the degree to which it affects our sense of hearing or sight. Naturally, this adds to the complexity of design significantly since the human designer has to bridge the gap between the magnitudes of arithmetic error on the one hand, and its impact on our senses on the other. The concept is explained in Figure 1 where it is obvious that the traditional method of field testing prototypes is impractical in terms of time and effort due to the large number of possible combinations of inexact designs to be tested. In this paper, we present a neurocognitive model of human hearing which is used to quantify glitches or loss in accuracy, through a novel metric of intelligibility of spoken sound. The model and the intelligibility metric are used to guide the design of inexact versions of standard digital building blocks of a hearing aid.

Once we adopt the notion that designing an inexact (or inaccurate) DSP primitive in return for significant savings or efficiency is acceptable, the amount of inexactness or inaccuracy that is admissible becomes an immediate question. The more inexact the system is, the more efficient it is in terms of area, size and energy consumption. However, beyond a certain point, the level of inexactness will render the design so glitchy that it become unusable. Where is the threshold above which the design is as inexact as possible, while remaining useful? We will refer to such a design as a "good enough" design. Once again, the concept of intelligibility will be used to characterize the threshold above which a design is "good-enough" and the digital building blocks will serve as vehicles for demonstrating the utility of our approach. The concept of a good-enough product—a hearing aid in our case—as a basis for achieving efficiency through inexact design, supported by a quantitative metric such as intelligibility for guiding this process is a second significant contribution of this paper.

The cognitive model we use in this work is a combination of a well-known model for evaluating quality of speech—Perceptual Evaluation of Speech Quality or PESQ[20]—and a custom calibration technique based on a set of behavioural experiments to determine intelligibility. In our setup, sound samples are processed by an inexact hearing aid and their intelligibility is determined by the cognitive model. Based on this result, an optimization loop determines how much more inexactness can be

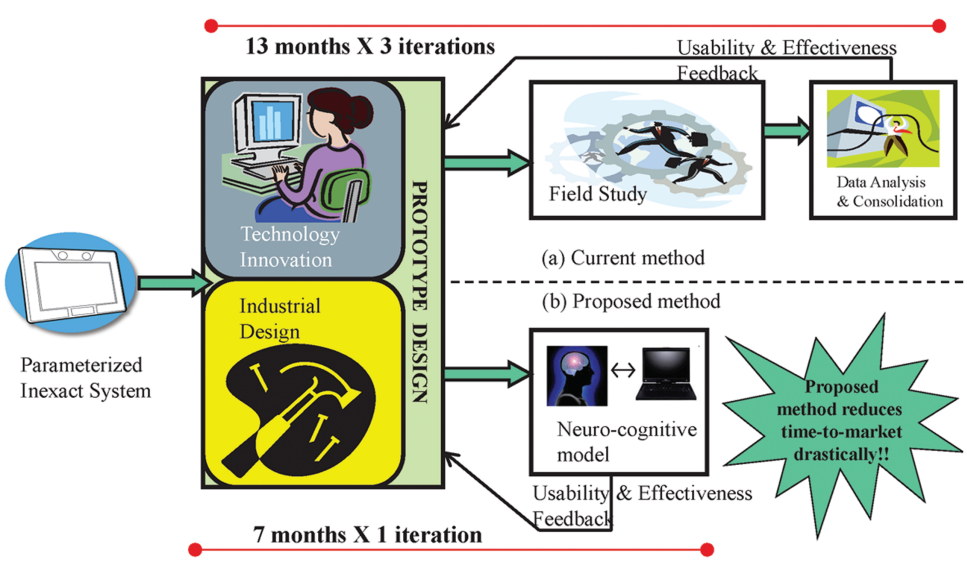

Fig. 1. Reduced time-to-market is achievable using the proposed method of designing inexact electronics that includes a cognitive model in the design loop.

introduced and this process is repeated till a certain threshold of performance is reached. In this work, we have introduced inexactness in the digital filter bank responsible for frequency decomposition; however, the principle can be applied to any other circuit block in the processing chain. To further reduce optimization time, a behavioural model of the inexact circuit is developed to quickly estimate power savings and errors incurred thus eliminating the need for time consuming digital synthesis in every step of the optimization. We followed an approach which combines both greedy and genetic algorithm in carrying out the optimization step. Using the techniques described earlier, we can reduce the power area product for the filter bank by 2:56× for an intelligibility loss of only 20% over conventional exact designs. We have earlier presented an initial version of this work in Ref. [21]. Compared to Ref. [21], we now present the following novelties in this work.

- Detailed explanation of the entire system and developing a level heuristic for component pruning
- Use of a more accurate, fine-grained error estimation technique
- Improvement of the previous optimization technique by involving genetic algorithm.

The organization of the paper is as follows: In Section 2, we shall describe our optimization framework for pruning circuits with a cognitive model in the loop. Details of the hearing aid architecture which will serve as a candidate for demonstrating the gains are given in Section 3. Next, Section 4, describes the results obtained by our design procedure and reports the performance gains over conventional digital design. Finally, we discuss some aspects of future work and conclude in the last Section 5.

2. PERCEPTUALLY GUIDED PRUNING FOR EFFICIENT INEXACT CIRCUITS

Probabilistic Pruning[22–24] is an inexact design technique that exploits the knowledge of the significance of a circuit component and its switching probability during circuit operation to derive a systematic approach to prune the "least useful" components in a circuit. When applied on data path elements, this technique has been shown to achieve significant savings ranging between 30%–50% in all of energy, delay and area without any implementation overheads in hardware for acceptable losses in the accuracy of the outputs. In this paper, we will use this technique as the basis for introducing "inexactness" and enabling the energy-accuracy trade-offs as opposed to other voltage scaling based methods like BiVOS[25,26] due to the practical problems (detailed in Ref. [22]) of having multiple power supplies and level converters for signals crossing these power domains. In this paper, we discuss three novel contributions in the domain of inexact circuit design: first, we develop a level heuristic for pruning of circuit components, replacing the exact components with inexact ones. Second, we formulate a fine grained error modelling technique associated with each of the inexact components. Finally, we present a combination of Greedy and Genetic algorithm based optimization strategy for pruning that is scalable to large digital systems. We shall demonstrate the gains achievable by our methods using a hearing aid (more specifically the FIR filters in it) as the driving example.

2.1. Optimization Framework

Earlier work on pruning[22,23] has considered the removal of gates in arithmetic circuits like adders and multipliers by using explicit cost functions involving switching probabilities and significance. However, such approaches are intractable for large systems producing outputs for perceptual consumption because of the non-availability of such explicit cost functions and the severe overheads involved in determining it computationally at the granularity of individual gates. Hence, we propose to have a library E of N_E elemental circuits, E_i, each of which can admit only one of several pre-characterized pruned topologies. These topologies can be indexed by an integer l that denotes the level or degree of pruning; larger values of l indicate more savings at the cost of increased error magnitude.[22] Let L_i denote an integer that indicates the maximum level of pruning of $E_i \in E$ while $l = 1$ corresponds to the unpruned structure. Now, we can define a set C of all components c_j allowed in our design by denoting $c_j = \{l_i, E_i\}$ where $E_i \in E$ and $l_i\ (\leq Li) \in N$. For example, if we have two elemental circuits, say adder and multiplier each with inexact levels 4 and 6 respectively. Then we have, $N_E = 2$ and set $C = (1, 1), (2, 1), (3, 1), (4, 1), (1, 2), (2, 2), (3, 2), (4, 2), (5, 2), (6, 2)$. The circuit we want to optimize can be represented as a directed acyclic graph G whose nodes are N_{G} components selected from C, inputs, or outputs and whose edges are wires. We can now formulate an optimization problem where the performance of G can be modified by varying $L = [l_1 l_2 \ldots l_{\text{NG}}]$ in exchange of cost savings. To make this dependence explicit, we shall henceforth refer to the graph as $G(L)$.

The quality of outputs $\{O\}$ of $G(L)$ for the same set of inputs $\{I\}$ depends on the value of L with the best quality of outputs being obtained for $L = [11\ldots1] = L_u$ corresponding to the unpruned circuit. We formally define the performance as a function of L by Γ:

$$\Gamma = \sum_{k=1}^{v} p_k Q_p(O_k(\vec{L}), O_k(\vec{L}_u)), \quad \Gamma \in R^+ \tag{1}$$

where $O_k(L)$ represents the output of $G(L)$ for input I_k that occurs with a probability p_k for $1 \leq k \leq \nu$.

Q_p denotes a function that measures the perceptual quality of the output of the pruned circuit with respect to the output of the unpruned one based on models of human sensory processing with larger values denoting better quality.

The problem can now be defined as finding the optimal value of L, L_{opt} such that:

$$\vec{L}_{\text{opt}} = \arg\min \text{ subject to } \Gamma \geq \Gamma_{\text{TH}} \tag{2}$$

where M is some cost metric of the circuit (like area or power) that needs to be minimized and Γ_{TH} denotes a threshold for minimum acceptable perceptual quality. Note that exact evaluation of Γ requires ν to grow exponentially with the dimension of inputs which is generally intractable for finding solutions in limited time. Hence, we evaluate the results on a smaller subset of inputs which the circuit is likely to encounter in real applications with the hope that the obtained optimal solution is good for practical use. Rigorously speaking, the optima we find is likely near the global optimum and we describe the optimization techniques next.

2.2. Greedy Optimization Algorithm

The optimization problem posed in (2) can be solved by a gradient descent approach by modifying the objective function to include the constraint using a Lagrange multiplier approach. We chose the Lagrange multiplier approach since it allows us to specify a weightage between circuit metric M and performance Γ. In that case, L is updated at every iteration by an increment that is proportional to the negative of the gradient or sensitivity of this modified cost function. To ensure convergence to a global optima, a stochastic version of the gradient descent algorithm can be used. However, this process entails a huge computational complexity that quickly becomes intractable for large values of N_{G} and k. Also, the number of iterations needed to converge to the global optima depends largely on the nature of the cost function and can often be prohibitively large. Hence, we propose a modified algorithm to reduce computational complexity with the following features:

- We evaluate only $s(<N_{\text{G}})$ randomly selected components of the gradient and this random candidate set is modified every iteration. This can also be viewed as taking the projection of the true gradient vector on a randomly selected lower dimensional sub-space of R^{NG}.
- To reduce the number of iterations, we use a 'greedy' approach to give preference to those components which can reduce M without compromising the perceptual quality too much (even when $\Gamma > \Gamma_{\text{TH}}$). Hence, we modify the cost function to be:

$$C = \frac{M}{\varepsilon + \Gamma} \lambda u(\Gamma - \Gamma_{\text{TH}}) \tag{3}$$

where ε is a small positive number for regularization, λ is a large positive number to prevent choices which reduce the performance below threshold and u is the Heaviside function. We modified Eq. (3) slightly when compared with the equation mentioned in Ref. [21]. The factor ϵ is added in this work in order to stabilize the overall cost function when Γ goes too low.

- Akin to a learning rate, we have a pruning rate, q, which sets the maximum increase of pruning levels in one iteration. Also, since L_i is kept small (to reduce the effort in characterization), the increase in error for every iteration is large. To prevent large jumps in error, we only allow changes in pruning levels of the $q(<s)$ components which have the q largest gradient values by using a 'rank' function, rank_q that assigns values $1, 1/2, \ldots, 1/q$, to the selected ones while assigning a value of 0 to others. Hence, the final update equation is given by:

$$\vec{L}(n+1) = \vec{L}(n) + q \times \text{rank}_q(Is(n)\nabla c)_q \tag{4}$$

where I_s is the identity matrix with all rows set to zero other than 's' randomly chosen ones.

This whole method is described graphically in Figure 2.

2.3. Genetic Algorithm

The greedy algorithm presented in the previous subsection has an inherent drawback. It often end up with less optimal solutions. This is because in each of its step, it has knowledge of only the information about next step but not the entire system. To overcome this drawback, we explore the genetic algorithm based optimization approaches which do a parallel search in the population. This helps in from getting trapped by local minima. They also work on chromosomes which contain encoded version of parameters of potential solutions. In this work, we chose an efficient category of genetic algorithm called the Non-dominated Sorting Genetic Algorithm[27] to realize the optimal solution for our inexact pruning. This approach is faster than other genetic algorithms because it has logarithmic complexity and also it maintains elitism over successive generations. Below, we describe the procedure of fitting our optimization problem into a genetic algorithm scenario.

The DSP block of a hearing aid has a number of filter banks, say m in number. In our case, each of these filter banks have various number of multipliers which can be pruned by our inexact multiplier library. In order to do a time-efficient optimal pruning we choose to assign a particular inexact level to each filter bank, instead of each component. In essence, all the components (multipliers in this case) in a given filter bank i will have a particular

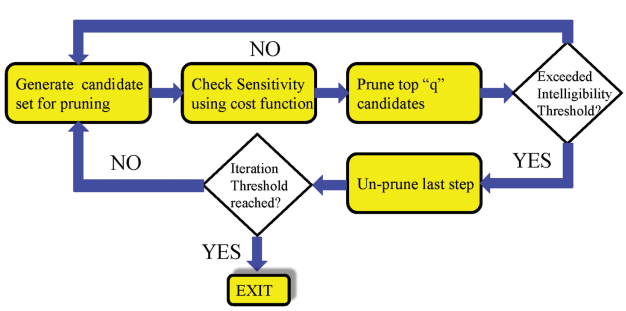

Fig. 2. A flowchart describing the proposed greedy algorithm.

have inexact level $l_i \leq L$. Here, L is the number of levels present in our inexact library. Our chromosome χ for optimization looks like below

$$\chi = l_1 l_2 l_3 \ldots l_m \quad l_i \geq L \tag{5}$$

where in each l_i is represented by a k bit binary value, where $k = d \log_2(L)$. We formulated a genetic algorithm based approach which tries to get the "good solution $\chi_{(\text{opt})}$" for our FIR bank such that the intelligibility (I) is within the threshold. The genetic algorithm tries to find the best choice for each filter such that the overall power savings P is maximum and drop in intelligibility is minimum. We re-define the problem of finding Optimum Filter Configuration (OFC) (mentioned in (2)) as follows:

$$\begin{aligned} \text{OFC}\{L_{\text{opt}}\} &= \textit{minimize}[P = f_p(L)]; \\ &\textit{maximize } [I = f_{\text{int}}(L)]; \\ &\textit{subject to } L_u \in L_{\text{opt}} \in L_{\max}; \\ &\textit{and } P \leq P_0; \;\; I \geq \Gamma_{\text{TH}} \\ &\textit{for some constants } P_0; \Gamma_{\text{TH}} \end{aligned}$$

The functions f_p, f_{int} give the power and Intelligibility of a particular configuration (L). The value P_0 to be obtained from the specification and the metric Γ_{TH} is defined in (2). The Non-dominated Sorting Genetic Algorithm (NSGA-II)[27] which can handle multi-objective optimization is used to solve the above optimization problem. Here, power (P) and Intelligibility (I) are the two objectives. At the end we are left with a set of pareto points (filter bank configurations, in this context) which are equally good based on the objective values.

3. HEARING AID ARCHITECTURE

To demonstrate the operation of this algorithm, we choose a digital hearing aid, which interacts with our auditory sense, as the platform.

The basic architecture of a hearing aid, shown in Figure 3 has an analog front-end amplifier followed by a wide dynamic range analog to digital converter (ADC) with sigma-delta ADC being the most popular choice.[29, 30] This is followed by the digital processor, which typically has two main parts:[31] a filter bank to decompose the speech signal into different sub-bands and a wide dynamic range compressor (WDRC) that compresses the input speech to fit the reduced dynamic range of an impaired ear. The main task of a hearing aid is, therefore, to selectively amplify sounds of a particular frequency range in order to fit the limited hearing range of the impaired ear. The block diagram of the architecture of the digital processor showing the auditory compensation scheme is displayed in Figure 4. After frequency decomposition by the filter bank, prescribed insertion gains are applied to each of the bands according to the NAL-NL1 formula[31] in order to raise the

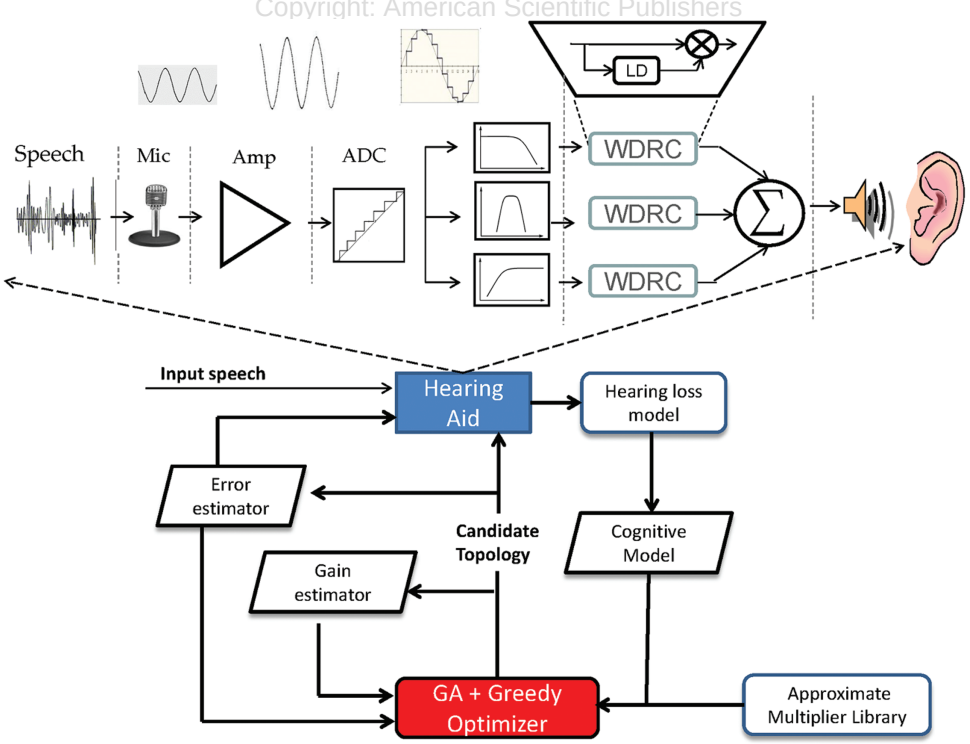

Fig. 3. Framework showing various module of the considered hearing aid architecture and evaluation of cost using the inexactness introduced in the design. The optimization loop uses a library of pre-characterized inexact VLSI components to quickly achieve a near optimal solution which is then evaluated rigorously through detailed synthesis procedures. The gain estimator and error estimator blocks help in predicting the performance gains in terms of power/area and error made due to introduction on inexactness in the design respectively.

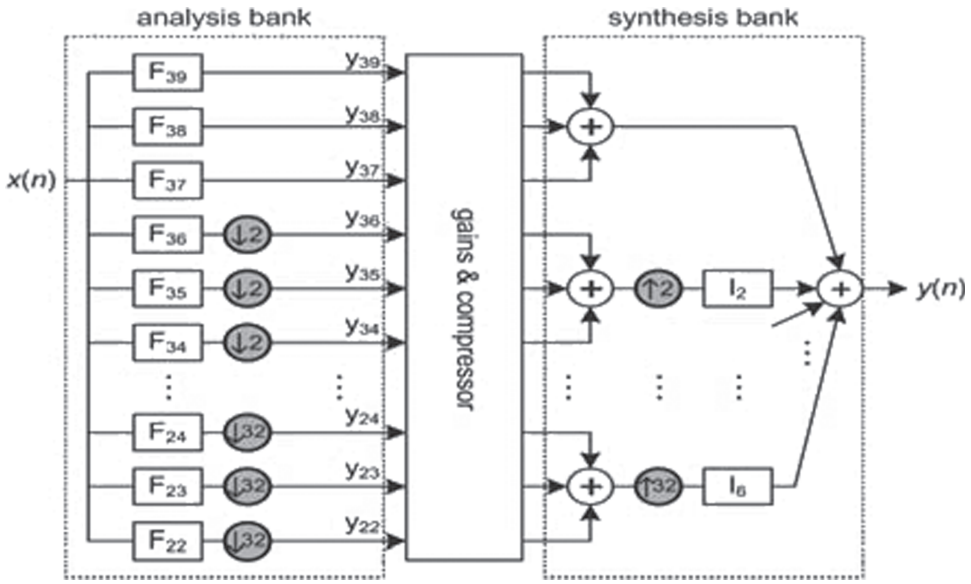

Fig. 4. Block diagram of the multi-rate implementation of the 18-band ANSI S1.11 1/3-octave filter bank for auditory compensation.

hearing threshold and overcome hearing loss. In this work, we have focused on the VLSI implementation of the filter bank which is an important part of this architecture and has been the focus of significant research.[28, 32] The WDRC is implemented as a software module in MATLAB.

Due to the good match with the frequency characteristics of the human ear, ANSI S1.11 1/3-octave filter bank specifications[33] are used to guide the design of the FIR filter bank in this work. ANSI S1.11 standard defines $431 = 3$ octave bands covering frequencies 0–20 kHz. For our application, we have chosen the bands 22–39 in the ANSI standard that covers the normal speech frequency range of 250 Hz to 8 kHz. The ANSI S1.11 standard defines three types of filters namely, class-0, class-1 and class-2. The difference in the parameters of these filters are based on the relative performance requirements of the filters with respect to stopband attenuation, operating range, environmental considerations[28] etc. Keeping the stop-band attenuation comparable to other hearing aid filter banks as in Ref. [28], class-2 filters have been used for our application. Figure 5 depicts the ANSI S1.11 class-2 filter

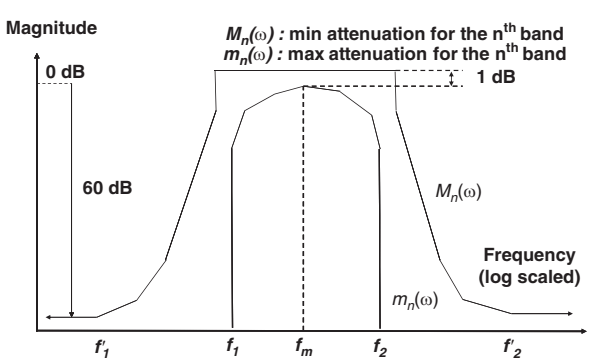

Fig. 5. Magnitude specification for ANSI S1.11 class-2 filters.[28]

specification on the nth 1/3 octave band.[28] Here $M_n(\omega)$ and $m_n(\omega)$ are the limits on the minimum and maximum attenuation on the nth filter band, respectively.

In Figure 4, x is the input speech signal which is decomposed into 18 frequency bands by the filter block. Straightforward implementation of the 18 bands of the ANSI specification would involve design of very high order filters since the bandwidths of the bands 22–27 are very low. Hence, we use the multi-rate architecture similar to Refs. [28, 32]. The bands 37, 38 and 39 constitute one octave; hence the 18 bands cover 6 octaves. Each of these bands are specified by its mid-band frequency f_m and its bandwidth Δf. For the nth band the mid band frequency is defined by:

$$f_m(n) = 2^{(n-30)/3} * f_r \tag{6}$$

Here f_r is the reference frequency which is set to 1 kHz according to ANSI standard.[28] For example the mid-band frequency of the 39th band is $f_m(39) = 8$ kHz. From the mid band frequency we can determine the band edge frequencies $f_1(n)$ and $f_2(n)$ for each of the n bands as:

$$f_1(n) = f_m(n) * 2^{(-1/6)}, \quad \text{and} \quad f_2(n) = f_m(n) * 2^{(1/6)} \tag{7}$$

Therefore the bandwidth of the nth band is given by: $\Delta f(n) = f_2(n) - f_1(n)$.

Using these specifications, we can implement the FIR filters for highest octave (bands 37–39), and from these we can recreate the other bands since the bandwidth of band F_n is exactly half of that of F_{n+3}. We can achieve the ideal frequency characteristics for each octave by reducing the sampling rate of each octave by a factor 2 as shown in Figure 4. Finally the outputs are up-sampled and then combined by the compressor to produce the desired output speech. During this, we use another filter denoted by I, to suppress the imaging distortion caused by up sampling.

This architecture, apart from making it easier to implement narrow bandwidth filters, reduces the computational complexity by minimizing the sampling rate for band limited channels.

4. SIMULATION FRAMEWORK: HEARING LOSS MODEL, COGNITIVE MODEL AND GAIN/ERROR ESTIMATOR

In this Section, we describe the various models we used to realize the optimum approximate version of the Hearing Aid architecture which we introduced in previous Section. Our entire simulation framework is implemented in MATLAB which includes the hearing aid architecture, hearing loss model, cognitive model and gain/error estimator is depicted in Figure 3. We shall describe each of them in detail in the following sub-sections.

4.1. Hearing Loss Model

The parameters of the WDRC in the hearing aid needs to be tuned according to the chosen hearing loss model which is patient specific. For this work, 'Presbycusis' is chosen as the hearing loss model since this is one of the most common sensorineural hearing loss problem. The audiograms of a person suffering from 'Presbycusis,' one with normal hearing and regular speech are shown in Figures 6(a) and (b) respectively for vowels and consonants. It can be seen that the hearing threshold, defined as the softest or lowest intensity of sounds that one can hear, is lower than the intensity of regular speech for persons with normal hearing. However, in the case of hearing disorders, the hearing threshold for certain frequencies of sound are more than the intensity of normal speech. People with such disorders fail to hear sounds in this frequency range. Figures 6(a) and (b) depict the raised hearing threshold at high frequencies for people suffering from 'presbycusis.'[32] It is evident from this figure that people with this hearing disorder will have difficulty in hearing since their hearing threshold is above the intensity of regular speech at high frequencies.

4.2. Cognitive Model

Due to human cognitive abilities, the signal to noise ratio (SNR) of a speech sample is not necessarily proportional to its 'intelligibility' as perceived by a human listener. To estimate intelligibility, we have developed a two stage model where the first stage uses a standard method to estimate speech quality while a second custom module is developed on the basis of behavioral experiments to transform the quality metric to intelligibility. This is depicted in Figure 7(a) and we shall next describe these modules one by one.

While it is easy to determine the SNR, it is more difficult to measure the subjective quality of speech.

Perceptual Evaluation of Speech Quality (PESQ) is a family of standards (ITU-T recommendation P.862)[20] comprising a test methodology for the automated assessment of the speech quality as experienced by a user of a telephony system. It uses a sensory model to compare the original, unprocessed signal with the degraded signal and develops an objective score as a mark of comparison. In our case, we use the output of the inexact hearing aid as the degraded speech input to PESQ. The details of the algorithm can be obtained from Ref. [20] and references therein; here, we describe the important steps for the sake of completeness. First, PESQ aligns the power levels of the two speech signals under test followed by aligning the two signals in the time domain to compensate for any delays. Next, both the original and degraded speech are windowed and converted to the frequency domain using Fourier transform. The frequency bins are then mapped to the pitch scale and the intensities are warped to map to perceived 'loudness' levels. These are the two steps where the specific details of the psycho-acoustics of human hearing are taken into account. These processed signals are now subtracted to provide an estimate of perceptual quality. In our case, this quality is not necessarily a direct measure of intelligibility—however, the pre-processing performed by PESQ is still relevant for us. Hence, we use this quality metric as a feature that can be mapped to intelligibility as described next. To obtain the intelligibility of

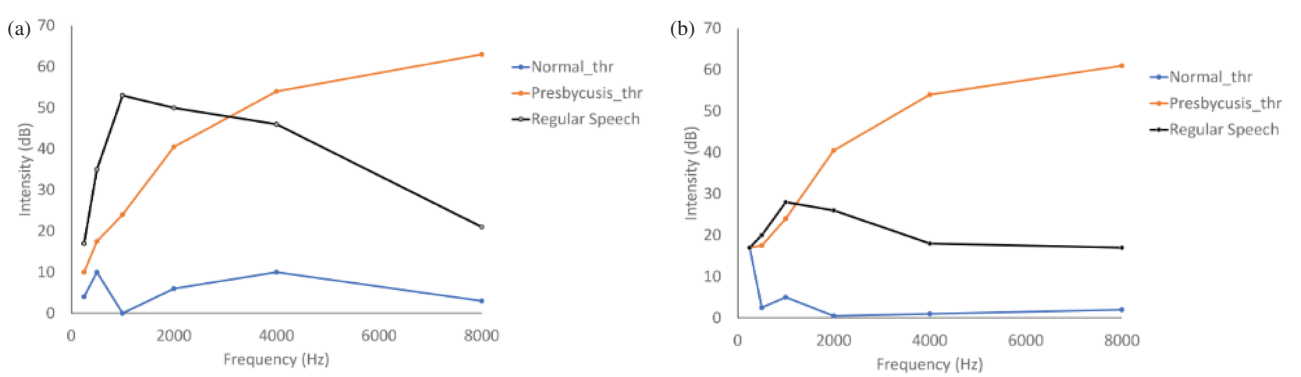

Fig. 6. Comparison of threshold intensity levels for persons with normal hearing, 'Presbycusis' with intensity of regular speech for (a) vowels (b) consonants.

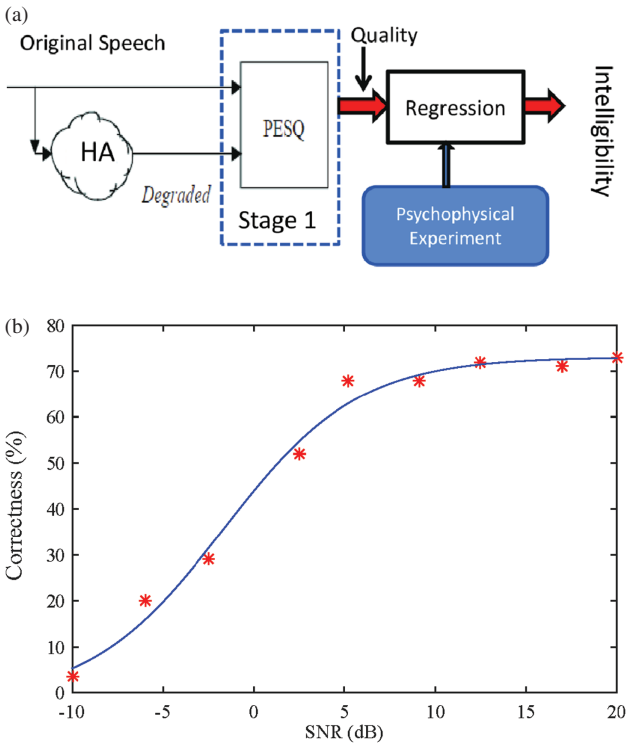

Fig. 7. (a) The cognitive model of intelligibility we use has two stages based on PESQ and psychophysics experiments respectively. (b) Result of the psychophysical experiment demonstrating the relationship between SNR and intelligibility.

degraded speech, a database of speech samples were created using a corpus of four hundred two-syllable words drawn from the Celex lexical database for English.[34] These words were now corrupted with white noise and cocktail party noise of different intensities so that the SNR of each sample varied between -10 and 20 dB. This study was approved by the Internal Review Board (IRB) at Nanyang Technological University, Singapore. Fourteen subjects with normal hearing were now chosen for the psychophysics experiment in which they were instructed to listen to a randomly selected set of 100 words with varying levels of SNR and type the word they thought they heard. The subjects chosen are from the same age group and belong to different gender and ethnicity to ensure a balance. These results were processed manually to correct for homophones and spelling errors. The average percentage correctness achieved in these experiments is plotted against the SNR in Figure 7(b) along with a sigmoidal function fit $f(x) = b/(1 + e^{k(x-a)}) + c$ where $a = -1.66$, $b = 74.81$, $c = -1.75$ and $k = -0.27$. The values are obtained using MATLAB curve fitting tool. As can be expected, the correctness of results indicating intelligibility does not change much as long as SNR is high enough (5 dB). However, below a certain threshold (SNR $\approx$ 2.5 dB), there is a sharp drop in intelligibility. Interestingly, even at high SNR, the intelligibility is not 100%—the reason for this is traced back to the fact that there will be some words which the subject has never heard and hence cannot comprehend even if the quality of speech is good. To correct for this effect, we normalize this curve by the maximum obtained percentage score. For our final model, we process the same speech samples for different SNR levels used in the test through PESQ to get a perceptual quality metric. Then a polynomial regression method is used to convert this quality score into the percentage correctness or intelligibility obtained in the behavioral experiment. We have observed that the nature of noise introduced by pruning is similar to white noise for small levels of pruning but departs from this assumption at higher levels of pruning. Improving the cognitive model to take this into account is a topic of further research.

4.3. Gain/Error Estimator

For every step of the iterative optimization process, we need to estimate the gains achieved in power and area due to pruning and the corresponding error introduced (shown in Fig. 3). Performing a synthesis of the pruned circuit at every stage is extremely time consuming. Hence, to accelerate the simulations, we perform a coarse estimate of the gain and error at each pruning step. This is done by creating a library of different pruned multipliers and adders which are used in the design. For each of these, a detailed characterization is done to obtain the area and power benefits for each design over the unpruned structure. Table I demonstrates the result for such a characterization of pruned multipliers in 65 nm CMOS process. These results are obtained after performing synthesis using Cadence RTL compiler, Place and Route using Cadence SoC Encounter and finally simulating the extracted post-layout spice netlist (with parasitic) using the Mentor Graphics ADiT fast spice simulator. This table is used to estimate the power and area benefits obtained for the entire FIR filter bank.

Estimating the error incurred at the final output is a bit more difficult due to the cascaded structure of the sub-blocks in the filter. To get good estimates of the error, we first generate a probability distribution of error at the output of an individual pruned block by comparing the results

Table I. Characteristics of pruned multipliers in 65 nm CMOS.

Pruning level	Power (normalized)	Area (normalized)	Mean error	St. dev. of error
1	1	1	0	0
2	0.804	0.875	0.122	0.72
3	0.692	0.753	0.373	1.06
4	0.519	0.733	−0.486	1.53
5	0.472	0.711	−0.038	2.62
6	0.458	0.570	−0.540	3.68
7	0.363	0.476	−2.540	7.54
8	0.302	0.472	−3.010	11.50
9	0.245	0.385	−6.490	14.89
10	0.220	0.417	−6.200	22.50
11	0.184	0.323	−14.33	29.81

of a pruned an unpruned circuit in simulations. The mean and estimates of these errors are also reported in Table I. This was done by observing the error for 10,000 trials for each pruned topology and generating a histogram with 1000 bins as mentioned in Ref. [21]. The histograms in Ref. [21] were obtained by keeping both inputs random for each pruned block and thus spanning the entire input range of the multipliers—hence, they are termed global histograms. Contrary to Ref. [21], in this work, we follow a different approach in obtaining the error histograms for each pruned block (belonging to one of the L inexact levels). It can be noted that the weights used by each filter are kept constant throughout the design process. By keeping one of the pruned block's input as constant (corresponding to a particular filter coefficient) and varying the other input randomly, we obtain a new set of histograms. We term these histograms as Regional histograms since one input is fixed at a certain region of input space. Obviously, the errors obtained in these simulations reflect the actual set of errors incurred by the multiplier when it is used in the FIR filter. However, the trade-off is that we have to do many more characterization steps (increase by a factor equal to the number of coefficients of the FIR filter) than the method in Ref. [21]. To assess whether or not, this is a worthwhile effort, we compare the error histograms of a pruned multiplier (for inexact level 10 and filter coefficient 23 (in case of Regional Histograms)) using the earlier and proposed methods in Figures 8(a) and (b) respectively.

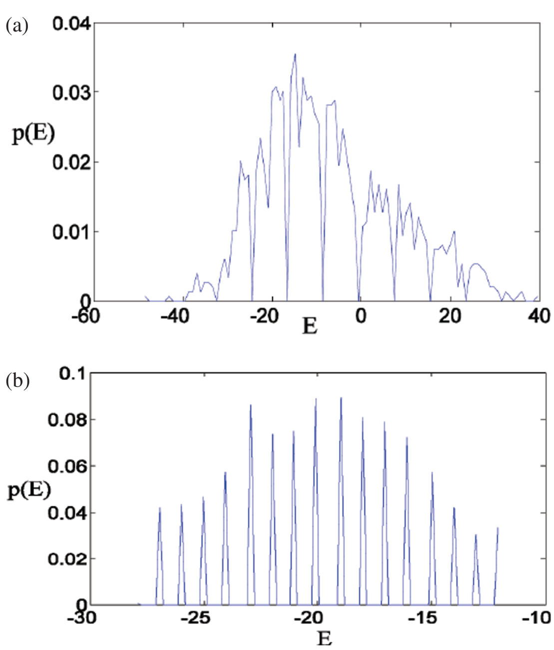

Fig. 8. Error histograms (a) global[21] and (b) regional, showing the probability of error $p(E)$ and the obtained error (difference of exact and inexact outputs for given input). It can be seen that the regional histograms give more accurate error than wide spread error given by global histograms.

The predicted error value (E) is obtained as the difference between exact (Y_{inexact}) and inexact outputs (Y_{exact}).

$$E = Y_{\text{inexact}} - Y_{\text{exact}} \tag{8}$$

The predicted error (E) and the probability of its occurrence ($p(E)$) for both global and regional cases are shown in these plots. It can be inferred from the plots that spread of error in case of regional error (from -27 to -12) is only 17% of the spread of error from global histograms (-42 to 40) for this particular coefficient value.

Further, we define a metric Error Spread Ratio (ESR) to quantify the over-estimation effect of global histograms of Ref. [21]. We measure the spread of error as difference between maximum and minimum error ($E_{\max} - E_{\min}$) for both Regional Histograms (RH) and global histograms (GH). The ratio of spread of regional to global is defined as ESR as:

$$\text{ESR} = \frac{E_{\max}(RH) - E_{\min}(RH)}{E_{\max}(GH) - E_{\min}(GH)} \tag{9}$$

For twenty six different coefficients (chosen arbitrarily out of the ones used in this design) of FIR, the ESR values are shown in Figure 9. It can be clearly observed that all these ratios are less than unit value, emphasizing better accuracy of Regional Histograms. In other words, using the earlier method of global histogram based prediction would drastically over-predict the actual error. Our proposed error prediction can now enable us to continuing the pruning and increasing inexactness for more power and area savings. Hence, we generate Regional histograms for multipliers of all inexact levels and for each coefficient of FIR. All these are stored in separate database, so that a selected histogram for a given inexact level and filter coefficient can be used for predicting the error during runtime.

We now generate a random number in MATLAB according to the desired distribution (by mapping from a uniform random distribution to the desired one through the cumulative distribution function) and add it to the output of each block according to its own error profile. Since the

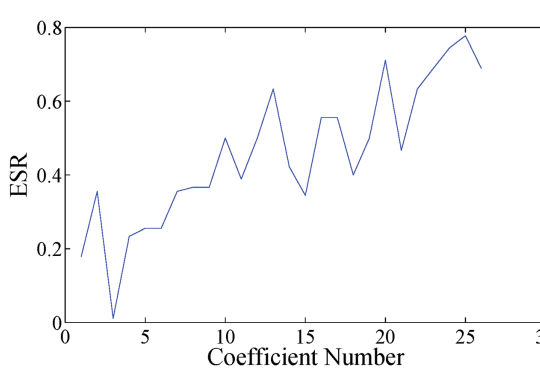

Fig. 9. The error spread ratio (ESR) for various coefficients of FIR. Plot suggests that good number of coefficients gave half of the error spread compared to the global histogram.

number of bins and the number of samples used for calculating the error are reasonably high, the model gives a good estimate of the actual error. However, as mentioned earlier, this additive error model does not hold true for very large pruning levels. We have not encountered such pruning levels for the permitted range of intelligibility; however, this might be needed for other applications and is a point for future research.

5. RESULTS

In this section, we shall describe the results obtained by pruning the FIR filters in the hearing aid using previously mentioned optimization algorithms. These algorithms were described in a generic way in previous Section 2. Here we shall first mention the specific values of the parameters used in our design.

5.1. Choosing Parameters

The circuit under consideration is an FIR filter. Hence, the library has $N_E = 2$ elemental circuits: an adder and a multiplier. However, closer inspection reveals that the area and energy consumed by an array multiplier is ≈10 times more than that of a ripple carry adder. Hence, it is expected that pruning multipliers would provide higher system level gains than pruning adders. Moreover, reducing the number of possible nodes to be pruned would result in a smaller search space for the optimization algorithm potentially speeding up simulations. Therefore, finally we have $N_E = 1$ since multipliers were the only element chosen for pruning. It can be seen from Table I that the maximum level of pruning for a multiplier is $L_1 = 11$. The filter used in the hearing aid has 18 bands covering 6 octaves as described before. In all 510 ripple carry adders and array multipliers each make up the arithmetic architecture of the FIR filter implying $N_G = 510$. The performance of the circuit is measured in terms of the Intelligibility (I) which is denoted as Q_p in previous Section 2. The final performance metric Γ is obtained by averaging the intelligibility over $\nu = 3$ sample words. For our simulations, we chose a low value of $\Gamma_{TH} = 50\%$ for the threshold in order to obtain a large trade-off curve from which we can choose a desired operating point according to application requirements.

The circuit metric we want to optimize is calculated in terms of the power (P). It is to be minimized keeping $\Gamma > \Gamma_{thresh}$. In order to give more weight to the cost savings in terms of power, a metric $M = \exp(P)$ is proposed. The hence formed cost function is sensitive to the changes in power due to the use of the expansive nonlinear function. M is maximum for the unpruned topology Lu of our circuit. In every iteration s (15 in our case) out of the NG nodes of the circuit G are sensitized. We can now define our cost metric as:

$$C = \frac{e^P}{\varepsilon + I}\lambda u(I - ITH) \tag{10}$$

For each of the s sensitized nodes, the cost is calculated and the nodes are then ranked in increasing order of cost. The top $q = 3$ nodes with the least cost values are chosen for actual pruning in one iteration. We chose the values for regularization constant (ϵ) and penalty factor (λ) as 0.05 and 2000 respectively. The detailed operations of the algorithm were earlier detailed in Section 2 and shown in Figure 2.

5.2. Finding Heuristics

The plot shown in Figure 10 demonstrates the rapid fall of SNR (much before Intelligibility) with decrease in Power. This plot supports our claim of choosing Intelligibility as judging parameter compared to SNR. The results shown in Figure 10 are from Ref. [21] and further will be used for comparative study.

Experiments were performed to evaluate the effect of each inexact level individually on the entire filter bank. Starting from completely exact level (L_1) up to highly inexact level (L_{11}), each are used in realizing the entire filter bank with their mean square error (MSE) compared to exact output being measured. The variation of MSE with Inexact level is illustrated in Figure 11. Level 11, which had a large MSE is excluded from the optimization procedure, in order to improve the performance of the trade-off curves. We term this as level heuristic. We can observe from the figure that the (MSE) keeps degrading with the increase in level of inexactness. An exception can be seen at level 10, where the (MSE) is better than at level 9. This can be justified as the mean error mentioned in Table I follow a similar trend.

5.3. Optimized Pruning

In addition to the greedy optimization approach, we explored a genetic algorithm based optimization as mentioned in the previous Section. The objective of this optimization, as before, is to minimize (Power) of the filter

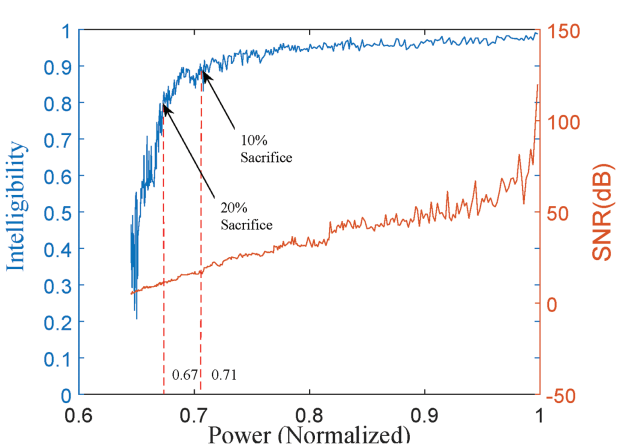

Fig. 10. Plot showing the trade-off between intelligibility and SNR with the power (normalized). The fall in intelligibility is significantly smaller compared to drop in SNR with decrease in power.[21]

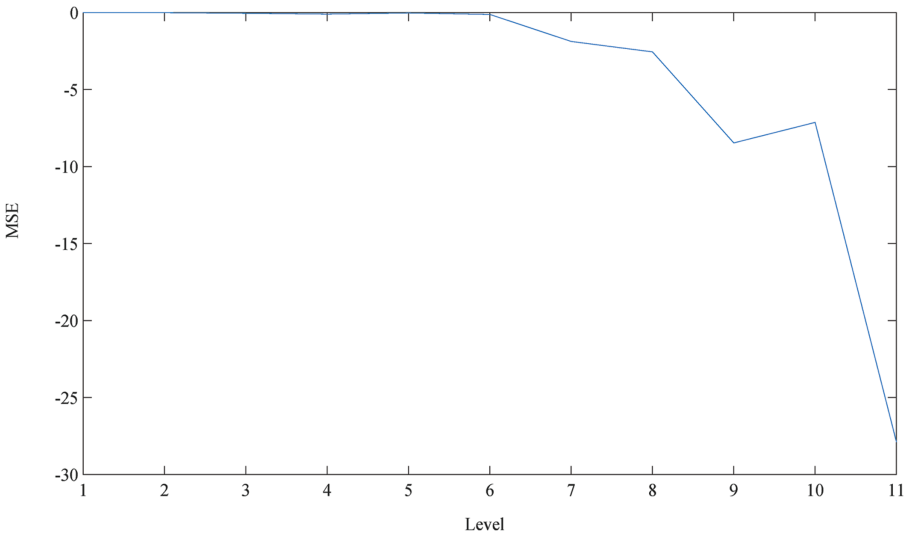

Fig. 11. The level heuristic is obtained by assigning all the components of the configuration a particular level of inexactness. The level which has large MSE is not chosen initially to obtain better trade-off curves.

bank while keeping up its performance (Intelligibility) as much as possible. The input speech is divided into 18 bands according to the specification mentioned in previous section.[28] Following the level heuristic (mentioned earlier), a particular level of inexactness is assigned to the bands. We have used the MATLAB Optimization tool Box for running the NSGA to find out the L_{opt}. Initial population is considered in binary format. Two point cross over function is used with a cross over fraction of 0.8. Pareto front population fraction is chosen to be standard value of 0.35.[27] We ran the solver for approximately 30 hours to obtain the pareto curve. A plot of the trade off curve between Intelligibility and Power, for this GA approach is shown in Figure 12 (green colour). From a separate set

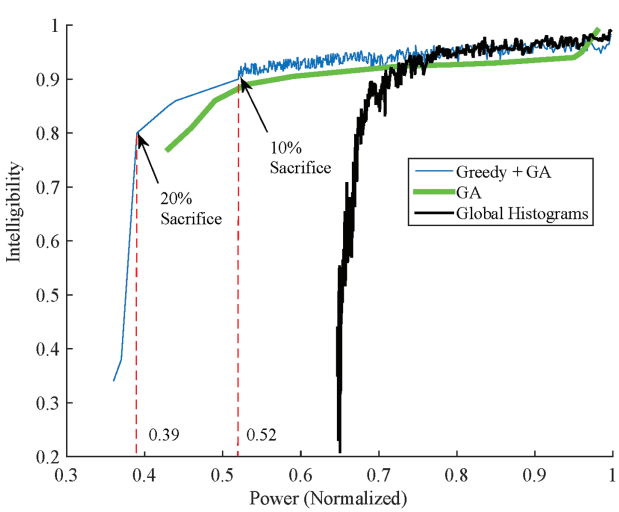

Fig. 12. Plot showing performance comparison of only GA and Greedy + GA approaches. There is a huge improvement in Power savings with use of regional histograms (blue and green curves) as against global histogram approach of Ref. [21] (black curve).

of experimental result, we have observed that employing only greedy technique gives a better spread of solution compared to only GA. However it becomes exceedingly slow after certain point (Int = 0.9, Power = 0.52)—it took more than 120 hours for the greedy optimization to move from point (Power = 0.52 to Power = 0.39). The GA solver however converges to the solution faster.

Hence, in order to get a good spread of solution and reach the optimum solution faster, we have implemented a combination of both GA and greedy approaches. In this method, we use the greedy technique till the point Power = 0.52, based on Regional Histograms. From this point, we further prune the filter configuration using GA approach as mentioned earlier. The trade-off curve for this approach is shown in Figure 12 (blue colour). The result of the approach mentioned in Ref. [21] is also shown in Figure 12 (black curve). This approach used global histograms for predicting the error and did not exclude the level which is having high MSE. It can be clearly observed that the trade-off curve corresponding to Ref. [21] started declining at Power = 0.71 much earlier compared to others. This can be ascribed to the fact that Ref. [21] used global histograms and did not take care of level that had high MSE. The curve representing the greedy and GA including level restriction performed better in terms of solution spread and the fall of intelligibility occurred at Power = 0.39. The only greedy approach to reach Power = 0.39 took more than 5 days, however the GA part of the final curve took only 6 hours. This proves the advantage of combining the greedy and GA optimization techniques without much loss in the spread of solution. We present the time taken by each of the optimization solvers along with the power savings obtained at 20% loss of Intelligibility is given in Table II.

Table II. Summary of runtimes and power savings (at 20% loss of intelligibility) for various optimization approaches.

S. no	Approach	Runtime (in hrs)	Power savings
1	Greedy	120	1.47×
2	GA	30	2.17×
3	Greedy + GA	34	2.56×

5.4. Discussion

While we have demonstrated the gains for inexact design of a digital filter bank, we would like to emphasize that the methodology introduced here is generic and can be used to induce inexactness in any of the components of a modern day digital hearing aid, the primary ones being Analog to Digital Converter (ADC) and DSP followed by peripherals. Nevertheless, we can estimate the system level gains achievable by our approach with the assumption that the filter bank dominates the DSP power budget. According to Ref. [31], the power consumption of ADC is 20% and that of other peripherals is 10% of the overall power consumption. The rest of the power consumption (70% of overall) is due to the DSP block. If we term the power consumption of ADC as P_{ADC}, that of peripherals as P_{Peri} and that of DSP as P_{DSP}, we can define the total power consumption P_{tot} of Hearing Aid as follows:

$$P_{\text{tot}} = P_{\text{ADC}} + P_{\text{Peri}} + P_{\text{DSP}} \tag{11}$$

where $P_{\text{ADC}} = 0.2P_{\text{tot}}$, $P_{\text{Peri}} = 0.1P_{\text{tot}}$ and $P_{\text{DSP}} = 0.7P_{\text{tot}}$. With our optimum approximation approach we aim to reduce only the P_{DSP} term of Eq. (11), keeping other terms constant. From Section 5.3, we obtain the value of P_{DSP} new as $0.52P_{\text{DSP}}$, i.e., $0.364P_{\text{tot}}$ for a 10% reduction in intelligibility. Hence, we rewrite the Eq. (11) as follows

$$P_{\text{tot\_new}} = 0.2P_{\text{tot}} + 0.1P_{\text{tot}} + 0.364P_{\text{tot}} \tag{12}$$

From Eq. (12) we get the new total power of hearing aid $P_{\text{tot\_new}}$ as $0.664P_{\text{tot}}$ which is 33.6% reduction in overall power consumption. For the case of 20% Intelligibility sacrifice, the new total power of hearing aid will be $0.573P_{\text{tot}}$, which is 42.7% reduction in overall power of the Hearing Aid.

In terms of comparisons, most of the recent works in approximate computing have focused on developing architectures for Image or Video Processing applications while the work on hearing aids reduced power dissipation by conventional low power digital design techniques and not inexact design. As shown in Table III, Refs. [3 and 7] focused on using inexact architectures for Image Compression using DCT. Hatfield et al. in his work[5] used reduced precision arithmetic for low power atmospheric modeling. Reference [35] presents an extensive survey of various works which introduce inexact hardware at various design layers like software, architectural and circuit. They build a elemental logic block using logic minimization and build bigger logic blocks using that smaller block selectively. Characterization of approximate components based on the number of approximate input bits is done in work.[36] Such characterized elemental components are used for pruning a large circuit with heuristic based optimization to achiever overall power reduction. Works involving inexact circuit pruning in domain of audio applications are rarely reported in literature. The key contribution of our present work is

Table III. Summary of related works.

S. no.	Reference	Approach	Domain	Comments
1	Liu et al.[3]	Truth table manipulation	Image	No perceptual metric based optimization, no library based scalable approach
2	Almurib et al.[7]	Approximate discrete cosine transform computation	Image	No perceptual metric based optimization, no library based scalable approach
3	Hatfield et al.[5]	Low precision arithmetic	Atmospheric modelling	No perceptual metric based optimization, no library based scalable approach
4	Sengupta et al.[36]	Logic minimization	Audio	No perceptual metric based optimization, no library based scalable approach
5	Shafique et al.[35]	Logic minimization	Image	No perceptual metric based optimization, scalable library approach
6	Wang et al.[37]	Algorithm optimization, cycle reduction	Audio	No approximation techniques used, no perceptual metric
7	Wu et al.[38]	Charge recovery logic	Audio	No approximation techniques used, no perceptual metric
8	Gerlach et al.[39]	Adaptive beam forming algorithm	Audio	No approximation techniques used, no perceptual metric
9	Kadiyala et al.[21]	Approximate library	Audio	Optimization based on perceptual metric, only greedy method used—sub-optimal approximate solution, coarse error models, scalable library approach
10	This work	Greedy + GA optimization, approximate library	Audio	Optimization based on perceptual metric, greedy + genetic algorithm for better solution at similar computation time, fine grained error models, scalable library approach

the introduction of inexactness in a methodical manner, especially for audio applications.

On the other hand works like Refs. [37–39], focused on reducing power dissipation in hearing aids with techniques like optimizing the algorithms, using charge recovery architectures, adaptive beam forming algorithms respectively. These works which focused on audio applications, are yet to tap approximate architecture. We also compare our results with those shown in Figure 10. We can conclude that both the approaches presented in this work performed better than the approach mentioned in Ref. [21]. This can be attributed to the fact that we have used much refined way of predicting the error (Regional histograms). Our final greedy and GA algorithm combined gives improvement of in Power at both 10% and 20% Intelligibility sacrifice compared to the results mentioned in work.[21] A summary of the above works along with their advantages and disadvantages, is presented in Table III.

6. CONCLUSION

We presented a methodical approach for reducing the power associated with a digital circuit (area can be included, if required) by factoring in the neurocognitive processing done in our brains on incoming sensory signals. Earlier efforts in designing inexact circuits used traditional metrics like SNR—however, we can improve the designs further by taking into account the cognitive processing done by the brain. Our model allows one to quickly estimate the effect of inexact design on the user's experience without having to perform costly field studies. To demonstrate our methodology, we chose a digital hearing aid as the platform with circuit pruning to introduce inexactness and used 'intelligibility' of speech as the metric. We introduced a novel combination of greedy heuristic and genetic algorithm, based pruning strategy that allows us to prune very large circuits which fastens the search for optimal solution. Using our methods to prune the filter bank in the hearing-aid, we demonstrate 1.92×, 2.56× improvement in performance in terms of power consumed while producing 10, 20% less intelligible speech compared to the corresponding exact hearing aid.

Acknowledgment: Funding from Singapore MOE through grant ARC 8/13 is acknowledged.

References

1. Z. Du, A. Lingamneni, Y. Chen, K. V. Palem, O. Temam, and C. Wu, Leveraging the error resilience of neural networks for designing highly energy efficient accelerators. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34, 1223 **(2015)**.
2. L. Qian, C. Wang, W. Liu, F. Lombardi, and J. Han, Design and evaluation of an approximate wallacebooth multiplier, *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE **(2016)**, pp. 1974–1977.
3. W. Liu, L. Qian, C. Wang, H. Jiang, J. Han, and F. Lombardi, Design of approximate radix-4 booth multipliers for error-tolerant computing. *IEEE Transactions on Computers* 66, 1435 **(2017)**.
4. F. P. Russell, P. D. Düben, X. Niu, W. Luk, and T. N. Palmer, Exploiting the chaotic behaviour of atmospheric models with reconfigurable architectures. *Comput. Phys. Commun.* 221, 160 **(2017)**.
5. S. Hatfield, P. Düben, M. Chantry, K. Kondo, T. Miyoshi, and T. Palmer, Choosing the optimal numerical precision for data assimilation in the presence of model error. *Journal of Advances in Modeling Earth Systems* 10, 2177 **(2018)**.
6. A. Ranjan, A. Raha, S. Venkataramani, K. Roy, and A. Raghunathan, Aslan: Synthesis of approximate sequential circuits, *Proceedings of the Conference on Design, Automation and Test in Europe*, European Design and Automation Association **(2014)**, p. 364.
7. H. A. Almurib, T. N. Kumar, and F. Lombardi, Approximate dct image compression using inexact computing. *IEEE Transactions on Computers* 67, 149 **(2018)**.
8. S. P. Kadiyala, V. K. Pudi, and S.-K. Lam, Approximate compressed sensing for hardware-efficient image compression, *2017 30th IEEE International System-on-Chip Conference (SOCC)*, IEEE **(2017)**, pp. 340–345.
9. Y. Wu, X. Yang, A. Plaza, F. Qiao, L. Gao, B. Zhang, and Y. Cui, Approximate computing of remotely sensed data: Svm hyperspectral image classification as a case study. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9, 5806 **(2016)**.
10. S.-L. Chen, J. F. Villaverde, H.-Y. Lee, D. W.-Y. Chung, T.-L. Lin, C.-H. Tseng, and K.-A. Lo, A power-efficient mixed-signal smart adc design with adaptive resolution and variable sampling rate for low-power applications. *IEEE Sensors Journal* 17, 3461 **(2017)**.
11. S. Liu, Y. Shen, J. Wang, and Z. Zhu, A 10-bit self-clocked sar adc with enhanced energy efficiency for multi-sensor applications. *IEEE Sensors Journal* 18, 4223 **(2018)**.
12. T.-T. Zhang, M.-K. Law, P.-I. Mak, M.-I. Vai, and R. P. Martins, Nano-watt class energy-efficient capacitive sensor interface with on-chip temperature drift compensation. *IEEE Sensors Journal* 18, 2870 **(2018)**.
13. I. Mahbub, S. A. Pullano, H. Wang, S. K. Islam, A. S. Fiorillo, G. To, and M. Mahfouz, A low-power wireless piezoelectric sensor-based respiration monitoring system realized in CMOS process. *IEEE Sensors Journal* 17, 1858 **(2017)**.
14. S.-L. Chen and G.-S. Wu, A cost and power efficient image compressor VLSI design with fuzzy decision and block partition for wireless sensor networks. *IEEE Sensors Journal* 17, 4999 **(2017)**.
15. B. Li, P. Gu, Y. Wang, and H. Yang, Exploring the precision limitation for RRAM-based analog approximate computing. *IEEE Design and Test* 33, 51 **(2016)**.
16. C. Li, W. Luo, S. S. Sapatnekar, and J. Hu, Joint precision optimization and high level synthesis for approximate computing, *Proceedings of the 52nd Annual Design Automation Conference*, ACM **(2015)**, p. 104.
17. F. S. Snigdha, D. Sengupta, J. Hu, and S. S. Sapatnekar, Optimal design of jpeg hardware under the approximate computing paradigm, *Proceedings of the 53rd Annual Design Automation Conference*, ACM **(2016)**, p. 106.
18. S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, Deep learning with limited numerical precision, CoRR, abs/1502.02551 **(2015)**, Vol. 392.
19. M. Imani, A. Rahimi, and T. S. Rosing, Resistive configurable associative memory for approximate computing, *2016 Design, Automation and Test in Europe Conference and Exhibition (DATE)*, IEEE **(2016)**, pp. 1327–1332.
20. J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, Perceptual evaluation of speech quality (PESQ) the new itu standard for end-to-end speech quality assessment part ii: Psychoacoustic model. *Journal of the Audio Engineering Society* 50, 765 **(2002)**.

21. S. P. Kadiyala, A. Sen, S. Mahajan, Q. Wang, A. Lingamneni, J. S. German, X. Hong, A. Banerjee, K. V. Palem, and A. Basu, Perceptually guided inexact dsp design for power, area efficient hearing aid, *Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE*, IEEE **(2015)**, pp. 1–4.

22. A. Lingamneni, C. Enz, J.-L. Nagel, K. Palem, and C. Piguet, Energy parsimonious circuit design through probabilistic pruning. *2011 Design, Automation and Test in Europe*, IEEE **(2011)**, pp. 1–6.

23. A. Lingamneni, K. K. Muntimadugu, C. Enz, R. M. Karp, K. V. Palem, and C. Piguet, Algorithmic methodologies for ultra-efficient inexact architectures for sustaining technology scaling, *Proceedings of the 9th Conference on Computing Frontiers*, ACM **(2012)**, pp. 3–12.

24. A. Lingamneni, A. Basu, C. Enz, K. V. Palem, and C. Piguet, Improving energy gains of inexact dsp hardware through reciprocative error compensation, *Design Automation Conference (DAC), 2013 50th ACM/EDAC/IEEE*, IEEE **(2013)**, pp. 1–8.

25. L. N. Chakrapani, K. K. Muntimadugu, A. Lingamneni, J. George, and K. V. Palem, Highly energy and performance efficient embedded computing through approximately correct arithmetic: A mathematical foundation and preliminary experimental validation, *Proceedings of the 2008 International Conference on Compilers, Architectures and Synthesis for Embedded Systems*, ACM **(2008)**, pp. 187–196.

26. J. George, B. Marr, B. E. Akgul, and K. V. Palem, Probabilistic arithmetic and energy efficient embedded signal processing, *Proceedings of the 2006 International Conference on Compilers, Architecture and Synthesis for Embedded Systems*, ACM **(2006)**, pp. 158–168.

27. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation* 6, 182 **(2002)**.

28. S. ANSI, S1.11-2004: Specification for Octave-Band and Fractional-Octave-Band Analog and Digital Filters. American National Standards Institute, New York **(2004)**.

29. D. G. Gata, W. Sjursen, J. R. Hochschild, J. W. Fattaruso, L. Fang, G. R. Iannelli, Z. Jiang, C. M. Branch, J. A. Holmes, M. L. Skorcz, E. M. Petilli, S. Chen, G. Wakeman, D. A. Preves, and W. A. Severin, A 1.1-v 270-_a mixed-signal hearing aid chip. *IEEE Journal of Solid-State Circuits* 37, 1670 **(2002)**.

30. S. Kim, J.-Y. Lee, S.-J. Song, N. Cho, and H.-J. Yoo, An energy-efficient analog front-end circuit for a sub-1-v digital hearing aid chip. *IEEE Journal of Solid-State Circuits* 41, 876 **(2006)**.

31. J. M. Kates, Digital Hearing Aids, Plural Publishing, San Diego, CA **(2008)**.

32. Y. Lian and Y. Wei, A computationally efficient nonuniform fir digital filter bank for hearing aids. *IEEE Transactions on Circuits and Systems I: Regular Papers* 52, 2754 **(2005)**.

33. Y.-T. Kuo, T.-J. Lin, Y.-T. Li, and C.-W. Liu, Design and implementation of low-power ansi s1. 11 filter bank for digital hearing aids. *IEEE Transactions on Circuits and Systems I: Regular Papers* 57, 1684 **(2010)**.

34. R. H. Baayen, R. Piepenbrock, and L. Gulikers, The Celex Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pa **(1995)**.

35. M. Shafique, R. Hafiz, S. Rehman, W. El-Harouni, and J. Henkel, Cross-layer approximate computing: From logic to architectures, *Proceedings of the 53rd Annual Design Automation Conference*, ACM **(2016)**, p. 99.

36. D. Sengupta, F. S. Snigdha, J. Hu, and S. S. Sapatnekar, Saber: Selection of approximate bits for the design of error tolerant circuits, *in Proceedings of the 54th Annual Design Automation Conference 2017*, ACM **(2017)**, p. 72.

37. P. Wang, B. Fan, Y. Sun, and G. Yang, Optimized realization of wide dynamic range compression based on dsp5535 hearing aid platform, *Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), 2016 IEEE*, IEEE **(2016)**, pp. 1127–1131.

38. H.-S. Wu, Z. Zhang, and M. C. Papaefthymiou, A 13.8_w binaural dual-microphone digital ansi s1. 11 filter bank for hearing aids with zero-short-circuit-current logic in 65nm cmos, *Solid-State Circuits Conference (ISSCC), 2017 IEEE International*, IEEE **(2017)**, pp. 348–349.

39. L. Gerlach, G. Payá-Vayá, S. Liu, M. Weißbrich, H. Blume, D. Marquardt, and S. Doclo, Analyzing the trade-off between power consumption and beamforming algorithm performance using a hearing aid asip, *2017 International Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS)* **(2017)**.

Sai Praveen Kadiyala

Sai Praveen Kadiyala *received the Bachelors and Ph.D. degrees in electrical and computer engineering from Indian Institute of Technology, Kharagpur, India in 2008 and 2015, respectively. His Ph.D. dissertation was on Low Power High Performance Mixed Static Domino Circuit Synthesis. Since 2015, he is working as a Postdoctoral researcher at Nanyang Technological University. His current research interests include developing light weight techniques for anomaly detection in embedded systems, Hardware Security, Adversarial Learning etc.*

Aritra Sen

Aritra Sen *got his B.E from Jadavpur University and worked at NTU, Singapore after that on the topic of approximate hearing aids.*

Shubham Mahajan

Shubham Mahajan *received his B. Tech and M. Tech degrees from IIT Kharagpur. He did a summer internship at NTU, Singapore working on the topic of approximate computing.*

Quingyun Wang

Quingyun Wang*, received a M.Sc from NTU, Singapore and is currently working in the software industry. His interests are in algorithms and signal processing.*

Avinash Lingamneni

Avinash Lingamneni *received the master of science (MS-Thesis) and Ph.D. degrees in electrical and computer engineering from Rice University, Houston, TX, USA, in 2011 and 2014, respectively. His Ph.D. dissertation was on the emerging domain of inexact or approximate computing. His current research interests include developing novel techniques to realize energy parsimonious "inexact" circuits, or circuits which can achieve up to an order of magnitude cost (energy, delay, and/or area) savings in exchange for introducing tolerable amounts of error.*

James Sneed German

James Sneed German *received the Ph.D. in linguistics from Northwestern University in 2009. He then spent two years as a Postdoctoral Researcher at the Laboratoire Parole et Langage (CNRS) in Aix-en-Provence, France. Since 2010, he has been an Assistant Professor in the Division of Linguistics and Multilingual Studies at Nanyang Technological University, Singapore. His research interests cover the cognitive architecture of linguistic sound systems, as well as prosody and the role it plays in signalling both literal and nonliteral meaning.*

Dr. Xu Hong

Dr. Xu Hong *graduated from the University of Chicago with a Ph.D in Psychology in 2007 and a Master's degree in Statistics in 2005. Her thesis project investigated the neural mechanisms of heading and self-motion perception, correlating neural activity with the judgment of heading direction from the optic flow field. She then went to Columbia University for her postdoctoral training to investigate hierarchical information processing for face perception by psychophysics experiments, where she started a line of research on face perception, designing behavioral experiments from an electrophysiological and theoretical neuroscience basis. Since then, her research has encompassed multiple disciplines: mathematical modeling, computer programming, electrical and electronics, behavioral and system neuroscience, visual perception and psychology, civil and mechanical engineering, and design. Dr. Xu Hong continued her research on face and heading perception when she set up her Visual Cognitive Neuroscience (VCN) Lab at Nanyang Technological University in 2009 and collaborated with transport researchers from the School of Civil Engineering, the School of Computer Science and Engineering, the School of Mechanical and Aerospace Engineering, and designers from the School of Art, Design and Media. She is part of the research team at the Transport Research Center (TRC) at NTU. Her research projects on transport include the human centric design for signage and wayfinding in the public transport system, plus speed safety thresholds for personal mobility device (PMD) users and cyclists. Her research team at the VCN lab and TRC investigates information flow in the neural system for vision, wayfinding and transport in the physical environment, and the human factors for infrastructure design, planning and regulation. Dr. Xu Hong is an Assistant Professor in Psychology at Nanyang Technological University, Principal Investigator of the Visual Cognitive Neuroscience Lab, Coordinator of the Cognitive and Neuroscience Cluster at the School of Social Sciences, and Principal Investigator of the Transport Research Center (TRC) at NTU.*

Krishna V. Palem (S'80–M'86–F'04)

Krishna V. Palem (S'80–M'86–F'04) *received the M.S. degree in electrical and computer engineering (biomedical engineering) and the Ph.D. degree from the University of Texas, Austin, TX, USA, in 1981 and 1986, respectively. He is the Kenneth and Audrey Kennedy Professor with Rice University, Houston, TX, USA, with appointments in Computer Science, in Electrical and Computer Engineering and in Statistics. He was a Founder and the Director of the NTU-Rice Institute on Sustainable and Applied Infodynamics. He is a Scholar with the Baker Institute for Public Policy, Rice University. He was a Moore Distinguished Faculty Fellow with Caltech, Pasadena, CA, USA, from 2006 to 2007, and a Schonbrunn Fellow with the Hebrew University of Jerusalem, Jerusalem, Israel, in 1999, where he was recognized for excellence in teaching. In 2002, he pioneered a novel technology entitled probabilistic CMOS (PCMOS) for enabling ultralow-energy computing. Professor Palem was a recipient of the IEEE Computer Society's 2008 W. Wallace McDowell Award. In 2012, Forbes (India) ranked him second on the list of 18 scientists who are some of the finest minds of the Indian origin. He is a fellow of ACM and American Association for the Advancement of Science.*

Arindam Basu (M'10)

Arindam Basu (M'10) *received the B.Tech. and M.Tech. degrees in electronics and electrical communication engineering from the IIT Kharagpur in 2005, and the M.S. degree in mathematics and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, in 2009 and 2010, respectively. He joined Nanyang Technological University in 2010, where he currently holds a tenured associate professor position. He received the Prime Minister of the India Gold Medal in 2005 from IIT Kharagpur. His research interests include bio-inspired neuromorphic circuits, non-linear dynamics in neural systems, low-power analog IC design, and programmable circuits and devices. He was a Distinguished Lecturer of the IEEE Circuits and Systems Society for the 2016–2017 term. He received the Best Student Paper Award from the Ultrasonics symposium in 2006, best live demonstration at ISCAS 2010 and a finalist position in the best student paper contest at ISCAS 2008. He was received the MIT Technology Review's inaugural TR35@Singapore Award in 2012 for being among the top 12 innovators under the age of 35 in Southeast Asia, Australia, and New Zealand. He was a Guest Editor for two special issues in the IEEE Transactions on Biomedical Circuits and Systems for selected papers from ISCAS 2015 and BioCAS 2015. He is serving as a Corresponding Guest Editor for the special issue on low-power, adaptive neuromorphic systems: devices, circuit, architectures and algorithms in the IEEE Journal on Emerging Topics in Circuits and Systems. He is currently an Associate Editor of the IEEE Sensors Journal, the IEEE Transactions on Biomedical Circuits and Systems, and Frontiers in Neuroscience.*