

Benefits of Stochastic Computing in Hearing Aid Filterbank Design

Timothy J. Baker, Yiqiu Sun and John P. Hayes
 Department of Electrical Engineering and Computer Science
 University of Michigan
 Ann Arbor, MI, 48109 USA
 {bakertim, sunsusan, jhayes}@umich.edu

Abstract—Designing low-cost filterbanks is important due to severe resource limitations imposed by hearing aid size. Here, we develop a novel FIR filterbank employing stochastic computing (SC). SC-based filters use (pseudo)-random bit-streams to efficiently perform the core filtering operation. We demonstrate that SC is well-suited to low-cost filterbank design and compare our SC filterbank to a conventional sequential binary (SB) design. We show that the SC design achieves the same accuracy and latency as the SB one, with an exceptionally large 70% reduction in chip area. The power consumption of our proposed SC filterbank is 38-96% that of the SB design.

Keywords—filterbanks, hearing aids, stochastic computing, FIR filters

I. INTRODUCTION

The World Health Organization estimates that 430 million people have hearing loss that affects their quality of life [1]. Hearing aids provide a key solution to this problem. A major component of a hearing aid is a filterbank; see Fig. 1a. The filterbank decomposes the input sound into frequency bands that are selectively amplified to match a specific pattern of hearing loss or audiogram. For example, the audiogram in Fig. 1b indicates that the patient has severe high frequency hearing loss and requires more amplification at the upper end of the audio spectrum.

Filterbanks present many design challenges [2][3]. Not only must their frequency response accurately match patient audiograms, but they must also meet stringent constraints on physical size, response time, and power consumption. Most prior work [2][3] aims to reduce the filterbank's computational cost while meeting other hearing aid requirements such as low power to enable long battery life [2]. One important class of filterbanks is composed of finite impulse response (FIR) filters. FIR filters have linear phase response which often makes them preferred over alternatives like infinite impulse response (IIR) filters. However, FIR filters are usually larger than IIR filters and rely on the weighted addition operation which involves many costly multiplications and complex design trade-offs. The efficient implementation of FIR filterbanks using conventional (non-stochastic) technologies has been studied for decades [2][3].

This work explores the role of an unconventional circuit technology known as stochastic computing (SC) in filterbank design. SC encodes data in randomized bit-streams called

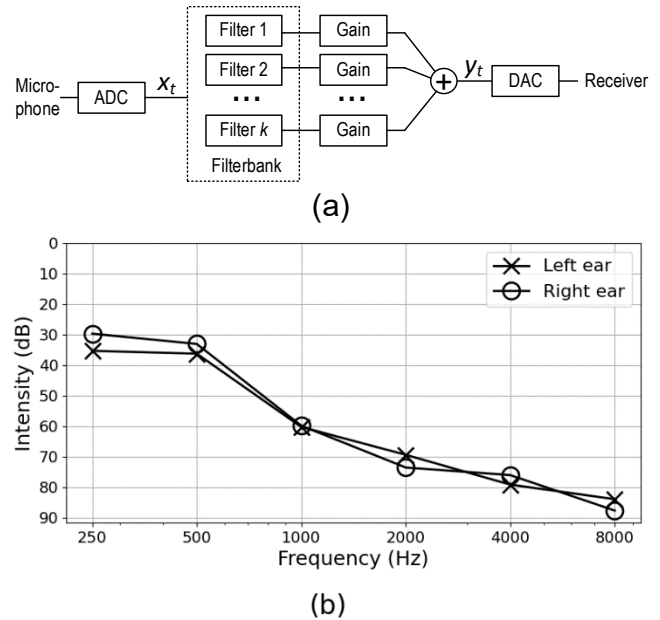


Fig. 1. (a) Basic structure of a filterbank for a digital hearing aid.; (b) audiogram. Each datapoint in the audiogram indicates the least intense (faintest) sound that a patient can hear at the given frequency.

stochastic numbers [4]. This encoding enables arithmetic operations to be implemented with tiny logic circuits, e.g., a single AND gate can perform multiplication. Such simple elements lead to very low area which makes SC an appealing candidate for filterbank design. Further, SC has other advantages like high fault tolerance. However, SC has relatively low accuracy due to its unusual number representation.

It has been suggested that SC is insufficiently accurate for digital filtering [5], but recent work [6][7] and this paper show that this is not necessarily the case. And, as we also show here, SC's usual need for costly binary-stochastic data conversion circuits is greatly reduced by the sharing of stochastic circuits possible among filters in an SC filterbank.

The main contributions of this paper are:

1. The successful application of SC to hearing-aid FIR filterbanks leading to a low-cost, accurate and flexible design.

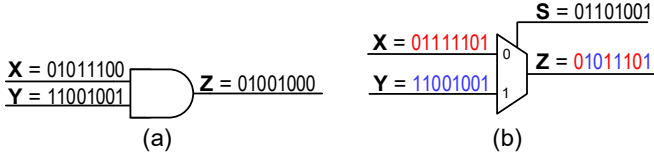


Fig. 2. Key SC elements. (a) AND gate acting as an SN multiplier where $\mu_X = \mu_Y = 0.5$ and $\mu_Z = \mu_X \mu_Y = 0.25$. (b) Multiplexer (MUX) performing scaled addition where $\mu_X = 0.75$, $\mu_Y = 0.5$, $\mu_S = 0.5$, and $\mu_Z = 1/2(\mu_X + \mu_Y) = 0.625$.

2. Cost and performance comparisons of our SC filterbank design with a conventional non-stochastic design.

3. Demonstration of accurate audiogram matching with SC FIR filterbanks on a representative audiogram.

The remainder of this paper is organized as follows. First, Sec. II reviews SC FIR filter design basics and prior work. Next, Sec. III introduces our SC FIR filterbank design which is then evaluated in Sec. IV. Lastly, Sec. V summarizes the main contributions and concludes the paper.

II. BACKGROUND

First, we review the basics of SC in relation to FIR filter design.

A. Stochastic Computing

In SC, data is represented by a pseudo-random stream of bits called a stochastic number (SN). An SN $\mathbf{X} = X_1 X_2 \dots X_N$ has a defining parameter $P_X = \mathbb{P}(X_i = 1)$ which is the probability that an arbitrary bit X_i of \mathbf{X} takes value 1. \mathbf{X} 's length N is application-dependent; its numeric value μ_X is derived from P_X and depends on the SN format used. Generally, the accuracy of μ_X improves as N is increased.

The two basic formats for SNs are unipolar where $\mu_X = P_X$, and bipolar where $\mu_X = 2P_X - 1$. The bipolar format allows for negative-valued SNs. For example, with $N = 8$, SN $\mathbf{X} = 00100001$ has an estimated unipolar value of $+0.25$ and bipolar value of -0.5 . Scaling can be used to accommodate numbers outside the $[0,1]$ and $[-1,1]$ intervals.

Representing data probabilistically with SNs leads to interesting and computationally efficient arithmetic circuits. For instance, consider an AND gate with unipolar SN inputs \mathbf{X} and \mathbf{Y} and output \mathbf{Z} . The output bit-stream's numerical value μ_Z is $P_Z = \mathbb{P}(X_i \wedge Y_i = 1)$ which, assuming \mathbf{X} 's and \mathbf{Y} 's bits are statistically independent or uncorrelated, yields $\mathbb{P}(X_i = 1)\mathbb{P}(Y_i = 1) = \mu_X \mu_Y$. Thus, $\mu_Z = \mu_X \mu_Y$ implying that an AND gate is a unipolar multiplier.

Fig. 2a illustrates unipolar SC multiplication where inputs \mathbf{X} and \mathbf{Y} with $\mu_X = \mu_Y = 0.5$ yield output \mathbf{Z} with $\mu_Z = 0.25$. Most SC circuits require uncorrelated SNs but, as we will see in Sec. III, correlation can sometimes be exploited to enhance SC by introducing new operations or increasing accuracy [6][7][8].

Addition in SC is scaled since SN values are derived from probabilities confined to the $[0,1]$ interval. To implement scaled addition, a simple multiplexer (MUX) can be used, as in Fig. 2b. Here, \mathbf{X} with value $\mu_X = 0.75$ and \mathbf{Y} with $\mu_Y =$

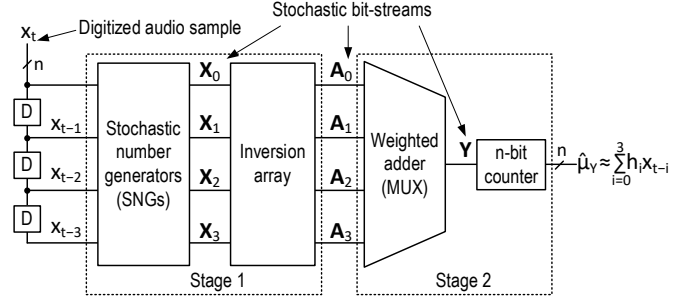


Fig. 3. Representative SC circuit performing weighted addition. The blocks labeled "D" are register delay units (D flip-flops). Stage 1 is a preprocessing step that prepares suitable SNs \mathbf{A}_0 to \mathbf{A}_3 for weighted addition. Stage 2 performs weighted addition using a MUX and then estimates the output \mathbf{Y} 's value using a counter.

0.5 are being added with the aid of a control SN \mathbf{S} with $\mu_S = 0.5$. In this configuration, both \mathbf{X} and \mathbf{Y} have a 50% chance of being selected each clock cycle implying that half of \mathbf{Z} 's bits are expected to be propagated from \mathbf{X} and the rest from \mathbf{Y} . Consequently, \mathbf{Z} 's value is an evenly weighted sum of μ_X and μ_Y , namely, $\mu_Z = 1/2(\mu_X + \mu_Y) = 0.625$. By adjusting μ_S , other scaled (weighted) sums can be implemented by a mux.

While encoding data into SNs enables low-cost arithmetic processing, generating the input SNs can be costly. A stochastic number generator (SNG) is needed to convert an n -bit binary integer B to an SN \mathbf{X} with $P_X = B/2^n$. An SNG is commonly built around a comparator and a pseudo-random number source (RNS) such as a linear feedback shift register (LFSR) [4]. SC designs may need large numbers of SNGs, which make them a major contributor to overall hardware cost. In this work, we follow the recent preference for low-discrepancy SNGs, which typically lead to more accurate outputs than LFSRs [9].

B. SC FIR Filter Design

An M -tap FIR filter implements the operation

$$y_t = \sum_{i=0}^{M-1} h_i x_{t-i} \quad (1)$$

where the $\{h_i\}$ are the constant filter coefficients, x_t is the input signal, e.g., a digitized audio stream, and y_t is the filtered output signal. The $\{h_i\}$ are the key filter design parameters and are computed from the filter's frequency response specification with the aid of software tools like MATLAB. Filters with more taps are larger, slower and more costly, but tend to do better filtering. Unlike conventional filters, SC-based FIR filters have been the topic of only a few studies such as [5][7][10][11]. The SC approach we present here is novel in that it applies is recent correlation-based accuracy optimizations suggested in [6][7] to digital filterbanks.

SC FIR filters are best explained with an example. Fig. 3 shows a 4-tap SC FIR filter that operates as follows. First, SNGs convert the four inputs $\{x_t, x_{t-1}, x_{t-2}, x_{t-3}\}$ to bipolar SNs \mathbf{X}_0 to \mathbf{X}_3 where the SN values are set to $\mu_{X_i} = x_{t-i}$. Then, if h_i is negative, \mathbf{X}_i is negated by the inverter array, otherwise \mathbf{X}_i is left unchanged. Because inverting a bipolar SN flips the sign of the SN's value, this step accounts for the sign of h_i .

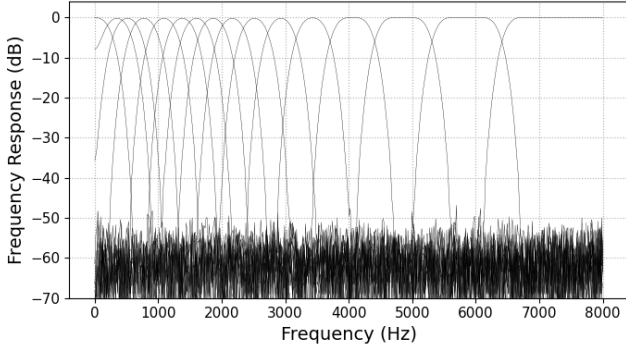


Fig. 4. Frequency response of our proposed 12-bit precision SC filterbank. As in [3], the subbands are spaced non-uniformly over the 0 to 8,000 Hz audio spectrum. The noise at the bottom is due to many stopband frequencies.

Consequently, the inverter array's output is \mathbf{A}_0 to \mathbf{A}_3 with $\mu_{A_i} = \text{sign}(h_i)x_{t-i}$. Finally, a 4-input mux whose select inputs' values are determined by the $|h_i|$'s performs scaled weighted addition on \mathbf{A}_0 to \mathbf{A}_3 . The mux's output is \mathbf{Y} with

$$\mu_Y = \frac{1}{\sum_{i=0}^{M-1} |h_i|} \sum_{i=0}^{M-1} |h_i| \mu_{A_i} = \frac{1}{\sum_{i=0}^{M-1} |h_i|} \sum_{i=0}^{M-1} h_i x_{t-i} \quad (2)$$

the latter being a suitably scaled version of the FIR filtering equation (1). The scale factor $1/\sum |h_i|$ denotes a gain which is accounted for later during audiogram matching. It is needed since the output SN \mathbf{Y} 's value is confined to the $[-1, 1]$ interval.

After addition, the output SN \mathbf{Y} must be converted back to a conventional binary number. Since \mathbf{Y} is a bipolar SN, an up-down counter is used which increments when bit-stream \mathbf{Y} outputs a 1 bit and decrements when \mathbf{Y} outputs a 0. The counter's output is $\hat{\mu}_Y$, an estimate of \mathbf{Y} 's value $\mu_Y = 2P_Y - 1$. The difference between the estimated output value $\hat{\mu}_Y$ and exact output value μ_Y is the error of the stochastic circuit. SC errors fluctuate randomly and typically diminish with longer SNs. Thus, there is a fundamental accuracy-latency trade-off in SC.

III. STOCHASTIC COMPUTING FILTERBANK

Here, we first give the specifications of our non-uniform filterbank. We then describe its proposed SC design.

A. Filterbank Specification

We consider a high-performance 16-channel FIR filterbank like that of [3] whose frequency response (Fig. 4) is based on the Bark scale [12]. Each of the 16 filters has 119 taps and the coefficients are determined using MATLAB. The filterbank's nonuniform subband spacing has the advantage of matching characteristics of human hearing, such as the fact that humans can differentiate low frequency sounds better than high frequency sounds [12]. Note there is little consensus on the best spacing of the subbands, and the proposed SC design can be flexibly applied to other subband spacings such as symmetric spacing [2].

B. Stochastic Computing Filterbank Design

Fig. 5 shows our proposed SC filterbank. The input logic for SN generation (which normally accounts for 90% of each

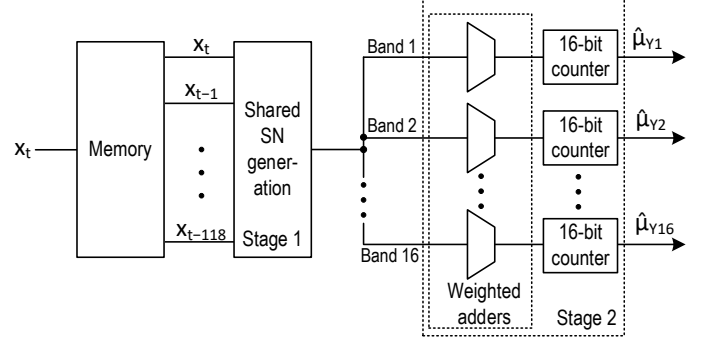


Fig. 5. Proposed SC filterbank structure. The 16 filters share Stage 1 containing SNGs and the inverter array, while each filter has its own Stage 2 containing an SN MUX weighted adder and counter. The filterbank core comprises all components except the memory block.

filter's area) is referred to as Stage 1 and is shared amongst all 16 filters thus saving considerable area and power. As seen in Fig. 5, the filters' processing and output logic called Stage 2 is not shared but is tailored to each filter's individual coefficients.

Prior studies on SC filters have suggested that extremely long bit-streams are needed to achieve satisfactory accuracy [5][13]. To combat this, our proposed filterbank design is based on correlation-enhanced multiplexer (CeMux) filters [7] which apply accuracy-enhancing correlation-changing techniques from [6] to SC filters. These techniques center around correlating the input SNs during generation as shown in Fig. 6. Here, each SNG shares an RNS which leads to high correlation in the generated SNs.

Normally, such correlation would degrade an SC circuit's accuracy because input correlation biases the output SN's value [8]. However, for mux-based circuits, correlation amongst the input SNs reduces random fluctuations in the output SN without biasing its value, thus improving accuracy [6][7]. Like CeMux filters, our filterbank design also employs a low discrepancy sequence generator as the shared (pseudo) RNS which improves accuracy over using other RNSs [9]. Combined, these design features greatly enhance the filterbank's accuracy.

Besides exploiting correlation, our filterbank differs significantly from the few previously proposed SC filterbanks [10][13]. The design in [10] implements only the multiplications in (2) using SC, whereas our design performs both multiplication and addition with SC, thus achieving much lower area. In [13], the authors propose an infinite impulse response (IIR) filterbank for use in auditory processing. FIR filters have desirable features that IIR filters lack, notably linear phase response. Our SC FIR filterbank is most similar conceptually to the non-SC design in [3] where a shared pre-computational unit similar in function to our shared Stage 1, is used to reduce the computational cost.

IV. DESIGN EVALUATION

In this section, we compare our SC filterbank design to a representative non-stochastic "sequential binary" (SB) design.

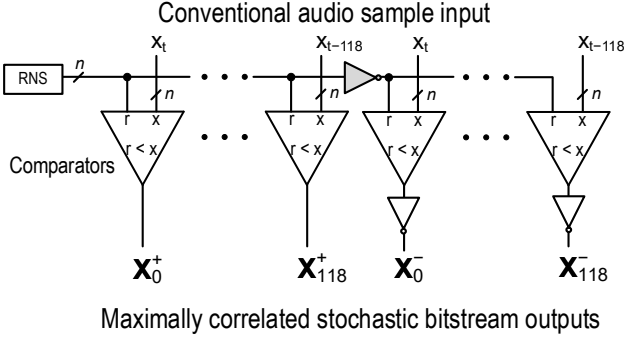


Fig. 6. Details of shared Stage 1 (Fig. 5). Outputs x_i^+ are routed to filters where coefficient $h_i > 0$ while outputs x_i^- are routed to filters where $h_i < 0$. The comparators are core components of SNGs and the shaded n -bit inverter ensures that all outputs $x_0^+, \dots, x_{118}^+, x_0^-, \dots, x_{118}^-$ are maximally correlated to improve accuracy.

A. Design Goals and Assumptions

Our main design and evaluation tools are widely used with non-stochastic digital systems: MATLAB for filter design, Synopsys Design Compiler for logic design, timing, and power analysis, as well as the open-source NanGate 45nm standard cell library for area and power estimation and layout synthesis. Our ability to directly compare with previously proposed filterbank designs [2][3] is severely limited by big differences in the chip technologies used, some of which are proprietary.

The SC filterbank realizes the design described in Sec. III, while the SB filterbank is a sequential implementation of an FIR filter that uses standard optimizations like the exploitation of symmetric coefficients, as in [3]. Each filterbank contains 16 filters (subbands) that implement the FIR filter operation (1) and have the overall passband/stopband frequency response illustrated by Fig. 4. Stopband attenuation is a key filter performance metric; generally, the higher the better.

Each design operates in real time and processes one audio sample every 0.0625 ms, corresponding to a sampling frequency of 16 kHz. The area, power and stopband attenuation of the SC and SB designs are determined by the precision (word length) n of each design. Here, n is varied from 8 to 12 bits. The SN length N is set to 2^{n+2} bits which is the shortest length that ensures that the SC filterbank's stopband attenuation is at least as high as the corresponding SB design's stopband attenuation. N is made a power-of-two to maximize hardware efficiency.

B. Experimental Results

As in prior studies like [3], our results apply to the filterbank's "core" and do not include the memory cost associated with storing past audio samples (see also Fig. 5). The memory would likely be implemented with a dual port RAM [3] and would be roughly the same size for both the SC and SB designs. Instead, we focus on where the two designs differ to highlight the computational performance and cost of the SC filterbank. Table I summarizes our experimental results, with accuracy represented by the lowest stopband attenuation of the 16 bands.

TABLE I. FILTERBANK DESIGN COSTS

Precision (bits)	Proposed (SC)			Conventional (SB)		
	Stopband attenuation (dB)	Area (μm^2)	Power (μW)	Stopband attenuation (dB)	Area (μm^2)	Power (μW)
8	27.1	7,928	220	25.0	28,255	631
9	30.7	9,538	287	28.5	34,473	854
10	37.4	10,978	440	33.7	41,055	1,020
11	42.9	12,515	655	40.2	44,973	1,107
12	47.0	14,037	1,161	46.1	49,414	1,206

One major conclusion is that SC can meet the accuracy requirements of hearing aid filterbanks with much shorter bit-streams than previously reported. In [5], it was concluded that SNs of length 2^{2n+1} are required for an SC filter to achieve the same performance as a precision-level n SB design, where n is the binary word length. In contrast, here we show that with 2^{n+2} -bit SNs, the SC filters in our filterbank achieve similar performance to conventional n -bit SB filters. This significant decrease in required bit-stream length is due to the correlation techniques employed in the SC filters and the use of the low discrepancy RNS. Ultimately, both the SC and SB filterbanks can meet essentially the same frequency response targets in terms of filter order, subband spacing, and stopband attenuation, as represented by Fig. 4 and Table I.

The accuracy of the SC filterbank is further indicated by its ability to match a patient's audiogram, which reflects such factors such as subband spacing, number of subbands, and stopband attenuation. Fig. 7 demonstrates this for the 12-bit SC filterbank using a representative member of a standard audiogram set [14]. The maximum matching error (MME) is 0.85 dB. Note that the normal target for MME is 3 dB or less [2] which, as in this example, is fully met by the SC design.

A second major conclusion is that the SC design's area is consistently 70% lower than that of the SB design. This great area efficiency is mainly due to the SC filterbank's use of cheap but accurate MUX-based weighted adders in place of costly conventional multipliers and adders. Importantly, the low area is also due to our proposed sharing of SN generators illustrated in Fig. 5. If the SNGs were not shared, the area of the SC design would be significantly higher.

Table 1 reveals that the SC filterbank's power consumption, is 38-96% that of the SB design for $n \leq 12$. However, power consumption grows steadily with the precision n due to the increase in SN length $N = 2^{n+2}$. The SC filterbank is always configured to process one audio sample every 0.0625 ms, so longer SNs require a faster digital clock, and therefore more power, to meet this constraint. The SB design's power grows more slowly with precision because its power dissipation is only due to increasing circuit area. It is unusual for an SC design to have similar power dissipation to a conventional design, but it occurs here because both designs are constrained to operate in real time with a latency of 0.0625 ms. Hence, a potential limitation of SC filterbank design is that the power consumption will continue to grow if SN length is further increased when more accurate outputs are desired.

There are several possibilities for improving the power efficiency of the SC filterbank. First, since each SN bit is

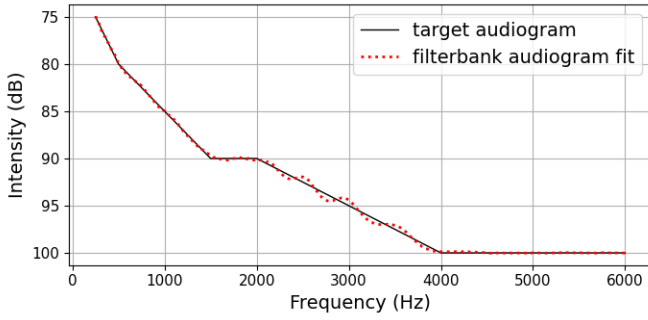


Fig. 7. Audiogram matching results for 12-bit precision SC filterbank.

equally weighted, SC circuits are very resilient to bit-flip errors, so techniques like voltage overscaling could be employed to reduce power consumption [15]. Alternatively, techniques like dynamic scaling [13] could increase the SC filterbank's accuracy for a given bit-stream length. Finally, some authors have proposed using analog memory with SC circuits to greatly mitigate the area and energy cost of memory and SNGs in SC systems [16]. These power-saving possibilities are worthy of further study due to the huge area savings offered by the SC approach to filterbank design.

V. CONCLUSION

In this work, we successfully applied SC to the design of hearing aid filterbanks. SC is unique in its reliance on stochastic bit-stream number representation. We found that the proposed SC design has 70% lower area than conventional SB design, while achieving comparable accuracy and the same latency. The SC filterbank's power is also much lower than the SB design's power for lower precision levels. Further, our proposed SC design is flexible in that changing the subband number and spacing does not significantly change the design's cost. Overall, we find that SC is an exciting and promising new direction for hearing aid filterbank design.

ACKNOWLEDGEMENTS

This research was supported by the U.S. National Science Foundation under grant CCF-2006704.

REFERENCES

- [1] World report on hearing. Geneva: World Health Organization, 2021.
- [2] Wei, Y. et al. "The design of low-power 16-band nonuniform filter bank for hearing aids," *IEEE Trans. Biomed. Ccts & Sys.*, **13**, 112-123, 2019.
- [3] Chong, K. et al. "A 16-channel low-power nonuniform spaced filter bank core for digital hearing aids." *IEEE Trans. Ccts & Sys. II*, **53**, 853-857, 2006.
- [4] Gaines, B.R. "Stochastic computing systems." *Advances in Information Systems Science*, **2**, 37-172, 1969.
- [5] Wang, R. et al. "Design, evaluation and fault-tolerance analysis of stochastic FIR filters." *Microelectron. Reliability*, **57**, 111-127, 2016.
- [6] Baker, T.J. and J.P. Hayes. "The hypergeometric distribution as a more accurate model for stochastic computing." *Proc. DATE*, 592-597, 2020.
- [7] Baker, T.J. and J.P. Hayes. "CeMux: maximizing the accuracy of stochastic mux adders and an application to filter design," 2021, *arXiv:2108.12326*.
- [8] Alaghi, A. and J.P. Hayes. "Exploiting correlation in stochastic circuit design," *Proc. ICCD*, 39-46, 2013.
- [9] Najafi, M.H. et al. "Deterministic methods for stochastic computing using low-discrepancy sequences." *Proc. ICCAD*, **51**, 1-8, 2018.
- [10] Mahesh, V.V. and T.K. Shahana, "Design and synthesis of FIR filter banks using area and power efficient stochastic computing," *Proc. WorldS4*, 662-666, 2020.
- [11] Zhong, K. et al. "Optimizing stochastic computing-based FIR filters." *Proc. DSP*, 1-5, 2018.
- [12] Zwicker, E. "Subdivision of the audible frequency range into critical bands," *Jour. Acoust. Soc. America*, **33**, 248, 1961.
- [13] Onizawa, N. et al. "Area/energy-efficient gammatone filters based on stochastic computation," *IEEE Trans. VLSI Sys.*, **25**, 2724-2735, 2017.
- [14] Bisgaard, N. et al. "Standard audiograms for the IEC 60118-15 measurement procedure." *Trends in Amplification* **14**, 2, 2010.
- [15] Lee, V.T. et al. "Architecture considerations for stochastic computing accelerators," *IEEE Trans. CAD*, **37**, 2277-2289, 2018.
- [16] Khatamifard, S.K. et al. "On memory system design for stochastic computing," *IEEE Computer Architecture Letters*, **17**, 2, 117-121, 2018.