Subsampling Mismatch Noise Cancellation for High-Speed Continuous-Time DACs

Derui Kong and Ian Galton

***Abstract*— Clock skew and component mismatches in continuous-time DACs introduce two types of error: static error and dynamic error. Both types of error typically limit the performance of practical, high-resolution, and continuous-time DACs, but most prior calibration techniques primarily reduce only static error. An exception is a recently published mismatch noise cancellation (MNC) technique that adaptively measures and cancels both types of error over the DAC's first Nyquist band. However, a disadvantage of the technique is that it requires an oversampling ADC that operates at several times the DAC's Nyquist rate to prevent convergence error that would otherwise be caused by aliasing. This paper presents a sub-sampling version of the MNC technique that avoids this limitation at the expense of a lower calibration convergence rate. As proven in the paper, the subsampling MNC technique allows aliasing to occur, but in such a way that convergence error is avoided.**

***Index Terms*— Dynamic element matching, mismatch noise cancellation, subsampling, calibration.**

I. Introduction

A CONTINUOUS-TIME DAC generates an analog output pulse for each digital input code. Ideally, the output pulse during each clock interval is scaled by the DAC's input code value during that clock interval, and except for this scale-factor it has the same shape as all the other pulses. Unfortunately, non-ideal circuit behavior causes input-dependent deviations of both the scale-factor and shape of each output pulse. Error in a DAC's output waveform from pulse scale-factor deviations is called *static error* and that from pulse shape deviations is called *dynamic error*.

The most significant types of static and dynamic error in practical high-resolution continuous-time DACs are caused by 1) inadvertent but inevitable *clock skew* and *component mismatches*, 2) *inter-symbol interference* (ISI), and 3) *signal-dependent output impedance* [1]–[14]. For DACs implemented in present-day CMOS technology that target signal-to-noise-and-distortion ratios (SNDRs) of greater than about 65 dB, error from clock skew and component mismatches is the most significant limitation. Unlike the other types of error, analog circuit design and layout techniques to reduce error from clock skew and component mismatches below this level are not known.

Yet continuous-time DACs with SNDRs of greater than 65 dB are increasingly necessary in high-performance applications such as 4G and 5G cellular base station transmitters. In such cases, calibration techniques are necessary to suppress error from clock skew and component mismatches. Unfortunately, most prior digital calibration techniques primarily reduce only static error, which leaves dynamic error as a major limitation in high-performance continuous-time DACs [1]–[14].

The difficulty in suppressing dynamic error arises from a property inherent to continuous-time DACs. Each DAC output pulse has a bandwidth that far exceeds the DAC's sample-rate, because its duration is time-limited to one clock period. Therefore, a technique that cancels dynamic error must either have a bandwidth that is wider than the DAC's signal bandwidth, or must perform frequency selective cancellation over a single Nyquist band.

Recently, a mismatch noise cancellation (MNC) technique was developed that addresses this difficulty [15], [16]. It incorporates a feedback loop that measures and cancels both static and dynamic error caused by clock skew and component mismatches over the DAC's first Nyquist band. While the MNC technique solves the dynamic error problem, it requires an oversampling ADC that operates at many times the DAC's Nyquist rate. This ultimately limits the maximum achievable signal bandwidth for a given power consumption.

This paper presents a subsampling version of the MNC technique that avoids the oversampling requirement. The original version of the MNC technique requires oversampling to avoid aliasing that would otherwise cause convergence error in the technique's error cancellation feedback loop. The modified version does not prevent aliasing, but is designed such that the aliasing does not cause convergence error. By avoiding oversampling, the modified MNC technique removes the potential signal bandwidth limitation of the original version at the expense of a modest reduction in the feedback loop's convergence rate. The paper presents a rigorous mathematical analysis of the proposed technique, and demonstrates the results via computer simulations.

II. Background Information: Oversampling MNC

Fig. 1 shows a high-level diagram of the IC presented in [16]. It consists of a 14-bit main DAC enclosed in an oversampling MNC feedback loop that adaptively measures

Manuscript received December 24, 2018; revised March 1, 2019; accepted March 31, 2019. This work was supported by Analog Devices, Inc., and in part by the National Science Foundation under Award 1617545. This paper was recommended by Associate Editor T.-C. Lee. *(Corresponding author: Ian Galton.)*

The authors are with the Electrical and Computer Engineering Department, University of California at San Diego, La Jolla, CA 92093-0407 USA (e-mail: galton@ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2019.2909173

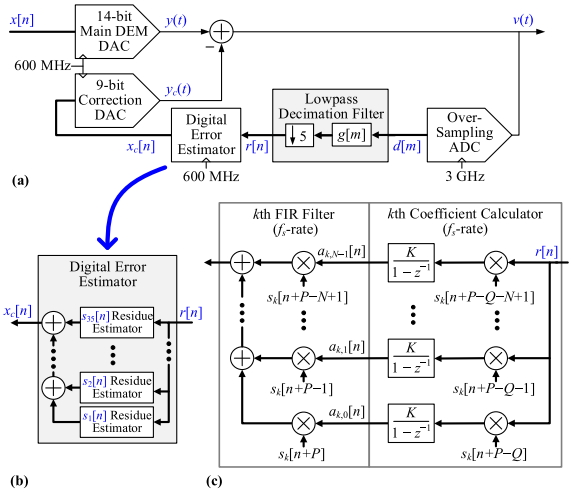

Fig. 1. a) High-level structure of the IC presented in [16], b) high-level structure of the digital error estimator, and c) details of each $s_k[n]$ residue estimator.

and cancels static and dynamic error caused by clock skew and component mismatches within the main DAC over the first Nyquist band. The MNC feedback loop consists of an oversampling ADC, a lowpass decimation filter, a digital error estimator and a correction DAC.

The main DAC incorporates dynamic element matching (DEM) of the type presented in [17]. Its static and dynamic error resulting from clock skew and component mismatches, collectively referred to as mismatch noise in the remainder of this paper, has the form

$$e_{DAC}(t) = \sum_{k=1}^{35} d_k(t) s_k[n_t] \qquad (1)$$

where n_t is the largest integer less than or equal to $f_s t$ with $f_s = 600$ MHz, each $d_k(t)$ is a 600 MHz periodic waveform that depends on clock skew and component mismatches within the main DAC, and each $s_k[n]$ sequence is generated by digital logic within the main DAC's DEM encoder [18]. Specifically, the $s_k[n]$ sequences are pseudo-random 600 MHz sample-rate sequences that take on values of -1, 0 and 1 and are uncorrelated with each other and with the main DAC's input sequence, $x[n]$. Consequently, $e_{DAC}(t)$ is wideband noise that is uncorrelated with $x[n]$ and free of harmonic distortion.

Without DEM, $e_{DAC}(t)$ would still be given by (1), but the $s_k[n]$ sequences would be deterministic, nonlinear functions of $x[n]$, so $e_{DAC}(t)$ would be entirely nonlinear distortion. Hence, DEM eliminates nonlinear distortion that would otherwise be caused by clock skew and component mismatches. However, it does so by converting the nonlinearity into noise, which severely degrades the DAC's signal-to-noise ratio (SNR). The purpose of the MNC feedback loop is to cancel this noise so as to keep the benefit of DEM without the SNR penalty.

The sampling theorem implies that for any $e_{DAC}(t)$ there must exist a correction DAC input sequence, $x_c[n]$, that would cause the correction DAC output waveform, $y_c(t)$, to cancel $e_{DAC}(t)$ over the first Nyquist band up to the accuracy of the correction DAC. As the dynamic range of $e_{DAC}(t)$ is much smaller than that of the main DAC, the resolution and step-size of the correction DAC, and, therefore, the error it introduces, are considerably smaller than those of the main DAC. Consequently, a 9-bit correction DAC with a step-size equal to a quarter that of the main DAC and no DEM or calibration was found to be sufficient in [16] to achieve more than 24 dB of error cancellation.

To make $y_c(t)$ well-approximate $e_{DAC}(t)$ over the first Nyquist band, the MNC feedback loop must measure $e_{DAC}(t)$ over the first Nyquist band. This requires a digitized version of the main DAC's output waveform that has been filtered to include only the first Nyquist band. The oversampling ADC and decimation filter in Fig. 1 perform this operation, so $r[n]$ contains a residual portion of $e_{DAC}(t)$ restricted to the first Nyquist band that is left over from imperfect MNC cancellation. Given that $e_{DAC}(t)$ is correlated with the $s_k[n]$ sequences as indicated by (1) and the decimation filter's impulse response is many 600 MHz samples long, it follows that the residual portion of $e_{DAC}(t)$ in $r[n]$ must be correlated with multiple time-shifted versions of the $s_k[n]$ sequences.

The MNC feedback loop measures the residual portion of $e_{DAC}(t)$ by correlating $r[n]$ with time-shifted versions of the 35 $s_k[n]$ sequences, and uses the measurement results to generate the correction DAC input sequence. Each of the 35 $s_k[n]$ residue estimators in the digital error estimator consists of a coefficient calculator block and an FIR filter with input $s_k[n+P]$ as shown in Fig. 1c.[1] The coefficient calculator correlates $r[n]$ with $N = 9$ time-shifted versions of $s_k[n]$. Each correlation is performed by multiplying $r[n]$ by a time-shifted version of $s_k[n]$ (which is -1, 0, or 1 during each 600 MHz clock period), and the result is scaled by $K = 8 \cdot 10^{-6}$ and accumulated. The accumulator outputs, $\alpha_{k,0}[n], \alpha_{k,0}[n], \ldots, \alpha_{k,8}[n]$, form the impulse response of the FIR filter, so each $s_k[n]$ residue estimator operates as an adaptive FIR filter. The 35 adaptive filters converge as necessary for $y_c(t)$ to well-approximate $e_{DAC}(t)$ over the first Nyquist band as proven in [15].

The MNC technique can operate either as a foreground or background calibration technique. While $e_{DAC}(t)$ is a broadband $x[n]$-dependent waveform, the $d_k(t)$ waveforms and the digital error estimator's target FIR filter coefficients depend primarily on component mismatches, clock skew, and other parameters that do not change significantly over time. Hence, the IC in [16] runs the MNC feedback loop during foreground calibration, and subsequently freezes the FIR filter coefficients and disables the ADC during normal DAC operation.

III. Subsampling MNC

As explained in [15], the accuracy required of the oversampling MNC technique's ADC is modest, e.g., in the IC presented in [16] the ADC's SNDR is less than 30 dB while the post-calibration signal-band SNDR of the DAC is over 77 dB. Yet the oversampling requirement poses a practical problem for DAC samples-rates above a few GHz. For instance, modifying the IC presented in [16] to have a

[1] In the IC presented in [16] P, Q, and N are set to 3, 21, and 9, respectively.

DAC sample-rate of 6 GHz, would require an ADC with a sample-rate of about 30 GHz. While low-SNDR ADCs at such high sample-rates are not impossible, a modified MNC technique that allows for an ADC sample-rate closer to that of the DAC would be preferable in terms of reducing power consumption, all other things being the same.

A. MNC Convergence Accuracy in the Presence of Aliasing

If the oversampling ADC and decimation filter in Fig. 1 were replaced by a Nyquist-rate ADC sampled at the same rate as the main DAC, the ADC output would contain all of the content of the main DAC's Nyquist bands aliased down onto its signal band. As each of the main DAC's Nyquist bands contains components correlated to the $s_k[n]$ sequences, the digital error estimator would adaptively cancel the sum of the error from all the aliased bands simultaneously, but it would fail to cancel error in any one of the Nyquist bands individually. This problem could be solved by inserting an anti-aliasing filter prior to the ADC, but this is not a practical option given the wide bandwidth and narrow transition band required of the filter.

Although it is necessary to avoid aliasing in the oversampling version of the MNC technique to measure the necessary MNC FIR filter coefficients, the following line of reasoning implies that it is at least mathematically possible to measure the necessary MNC FIR filter coefficients in the presence of aliasing. The output of the correction DAC in Fig. 1a has the form $y_c(t) = \alpha_c(t)x_c[n_t]$ where $\alpha_c(t)$ is a 600 MHz periodic waveform [18]. As shown in [15], the MNC feedback loop causes the impulse response of the kth FIR filter in Fig. 1c to converge such that the filter's transfer function well-approximates

$$H_k\left(e^{j\omega T_s}\right) = e^{-j\omega P T_s}\frac{D_{p-k}(j\omega)}{A_{p-c}(j\omega)} \quad \text{for } |\omega| \le \pi f_s \qquad (2)$$

where $f_s = 600$ MHz, $T_s = 1/f_s$, and $D_{p-k}(j\omega)$ and $A_{p-c}(j\omega)$ are the continuous-time Fourier transforms of one period of the T_s-periodic waveforms $d_k(t)$ and $\alpha_c(t)$, respectively.

It follows from (2) that the FIR filter coefficients could be calculated directly from one period of $\alpha_c(t)$ and one period of each $d_k(t)$ for $k = 1, 2, \ldots, 35$, and they could be calculated approximately from sampled versions of these 35 one-period waveforms. Moreover, the samples could be measured directly from the main DAC and correction DAC outputs. For example, to measure five samples of $\alpha_c(t)$ over one f_s-rate clock period the input to the correction DAC could be set to a non-zero constant value, and the five samples could be measured at its output over one clock period. Although more complicated, each of the $d_k(t)$ waveforms could be isolated by appropriately manipulating the DEM encoder and then similarly sampled.

This procedure would still require oversampling, but it can be further modified to avoid oversampling by recognizing that the measurements described above could be spread over five clock periods rather than over a single clock period. As depicted in Fig. 2, the f_s-rate periodicity of the $\alpha_c(t)$ and $d_k(t)$ waveforms ensures that an ADC sampled at a rate of $5f_s/6$ would collect the same information over a duration of $6T_s$ as an ADC sampled at a rate of $5f_s$ would collect over a duration of T_s, where $T_s = 1/f_s$. Hence, oversampling can be avoided at the expense of a longer data collection duration.

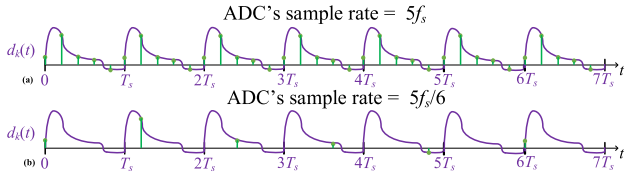

Fig. 2. a) Oversampling $d_k(t)$, and b) subsampling $d_k(t)$.

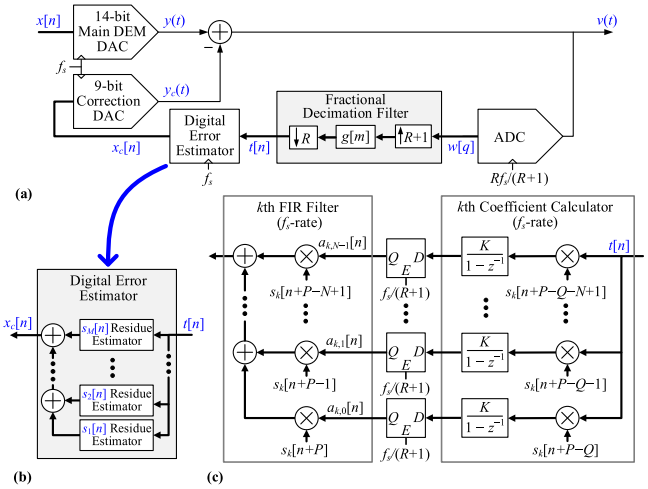

Fig. 3. a) High-level structure of the subsampling MNC technique, b) high-level structure of the digital error estimator, and c) details of each $s_k[n]$ residue estimator.

The argument above is the outline of a proof-by-construction that subsampling MNC is mathematically possible. However, the constructed procedure would only work as a foreground calibration technique, whereas the oversampling MNC technique works as either a foreground or background calibration technique, and it would be computationally expensive.

B. The Subsampling MNC Technique

A more practical way of exploiting the effect described above is the proposed subsampling MNC (SMNC) technique shown in Fig. 3. It differs from the oversampling MNC technique in three ways: an $Rf_s/(R+1)$-rate subsampling ADC is used in place of the Rf_s-rate oversampling ADC, where R is an integer greater than 1 (Fig. 1 is drawn for the specific case of $R = 5$), a fractional decimation filter is used in place of the lowpass decimation filter, and a bank of latches updated at times $n = 0, (R+1), 2(R+1), \ldots$ separate each coefficient calculator and FIR filter. The fractional decimation filter is equivalent to the cascade of an $R+1$-fold up-sampler, a digital filter with impulse response $g[m]$, and an R-fold down-sampler, but it can be implemented as the polyphase structure shown in Fig. 4 such that all its components are clocked at a rate of f_s [19]. Therefore, the highest clock-rate in the system is f_s.

The ADC sample-rate is slightly lower than f_s whereas the DAC output spectra are non-zero over several $f_s/2$-wide

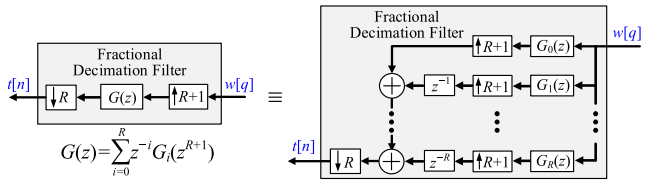

Fig. 4. Polyphase structure for fractional decimation filter.

Nyquist bands. Therefore, the ADC output, $w[q]$, contains significant aliasing. However, as explained shortly, the subsampling effect depicted in Fig. 2 (for the specific case of $R = 5$) prevents the aliasing from causing MNC convergence error. In particular, as proven in the remainder of the paper the subsampling MNC technique converges to the same set of FIR filter coefficients as the original oversampling MNC technique, but with a lower convergence rate.

To show that the SMNC technique converges to the same FIR filter coefficients as the oversampling MNC technique, it is helpful to first redraw Fig. 3a in an equivalent form that is easier to compare to Fig. 1a. Theorem 1 presented below provides this equivalent form.

Theorem 1: The system shown in Fig. 5 with

$$g^{(l)}[m] = \begin{cases} g[m], & \text{if } (l-m) \mod (R+1) = 0, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

and

$$t[n] = r^{(l)}[n] \quad \text{where } l = (-n) \mod (R+1), \tag{4}$$

(i.e., $t[n]$ is the output of the $R+1$ to 1 multiplexer) generates the same $t[n]$, $x_c[n]$, $y_c(t)$, $y(t)$, and $v(t)$ as that shown in Fig. 3a if both systems start with the same initial conditions and have the same input sequence, $x[n]$.

Proof: It follows from the definition of an up-sampler that the output of the $(R+1)$-fold up-sampler in Fig. 3a can be written as $d[m]p[m]$ where $d[m]$ is the output of an Rf_s sample-rate ADC in Fig. 5 and

$$p[m] = \begin{cases} 1, & \text{if } m \mod (R+1) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

This and the signal processing shown in Fig. 3a imply that

$$t[n] = \sum_{m=-\infty}^{Rn} d[m]p[m]g[Rn-m] \tag{6}$$

in Fig. 4. The signal processing shown in Fig. 5 and (4) imply that

$$t[n] = \sum_{m=-\infty}^{Rn} d[m]g^{((-n) \mod (R+1))}[Rn-m] \tag{7}$$

in Fig. 5. Therefore, it is enough to show that the right sides of (6) and (7) are equal, which is equivalent to showing that

$$g^{((-n) \mod (R+1))}[Rn-m] = p[m]g[Rn-m]. \tag{8}$$

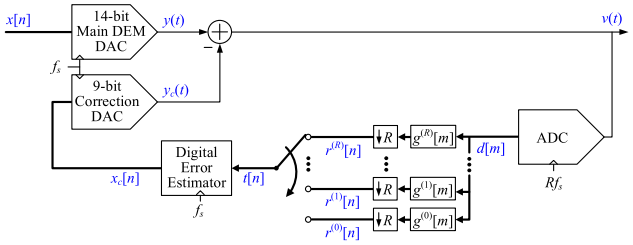

Fig. 5. Modified version of Fig. 3a with equivalent behavior.

Given that $[(-n) \mod (R+1)]-m] \mod (R+1) = -n-m \mod (R+1)$, (3) implies

$$g^{((-n) \mod (R+1))}[m] = \begin{cases} g[m], & \text{if } (-n-m) \mod (R+1) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Given that $[-n-(Rn-m)] \mod (R+1) = m \mod (R+1)$, replacing m with $Rn-m$ in (9) results in

$$g^{((-n) \mod (R+1))}[Rn-m] = \begin{cases} g[Rn-m], & \text{if } m \mod (R+1) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Substituting (5) into the right side of (8) results in the right side of (10), which shows that (8) holds.

The SMNC equivalent system of Fig. 5 is a useful analysis tool because it can be related to the original oversampling MNC technique as follows. Equation (3) implies that

$$\sum_{l=0}^{R} g^{(l)}[m] = g[m] \tag{11}$$

and Fig. 5 implies that

$$r^{(l)}[n] = \sum_{m=-\infty}^{Rn} d[m]g^{(l)}[Rn-m], \tag{12}$$

so

$$\sum_{l=0}^{R} r^{(l)}[m] = \sum_{m=-\infty}^{Rn} d[m]g[Rn-m]. \tag{13}$$

The right side of (13) is equal to $r[n]$ in the oversampling MNC technique shown in Fig. 1 (generalized with 600 MHz replaced by f_s and 3 GHz replaced by Rf_s). Therefore, the output of the oversampling MNC technique's decimation filter can be written as

$$r[n] = \sum_{l=0}^{R} r^{(l)}[m], \tag{14}$$

with $r^{(l)}[n]$ given by (12).

It follows that $t[n]$ in Fig. 3a (which is identical to that in Fig. 5 as implied by Theorem 1) is different from $r[n]$ in Fig. 1, even when the $v(t)$ waveforms in the two systems are equal. In particular, for equal $v(t)$ waveforms in the two systems, $t[n]$ in Fig. 3a for each n is equal to one of the $r^{(l)}[n]$ sequences whereas $r[n]$ in Fig. 1a is equal to the sum of the $r^{(l)}[n]$ sequences. This difference between $t[n]$ and $r[n]$ is the result of aliasing caused by the SMNC technique's subsampling. As explained in Section III-A, the oversampling

ADC is required in Fig. 1a to prevent aliasing that would cause convergence error. However, as proven in the next section, the SMNC technique converges correctly despite the aliasing caused by subsampling.

A qualitative explanation of this paradox is as follows. During foreground calibration, $x[n]$ is chosen such that the statistics of the $s_k[n]$ sequences do not change over time. The latches following each coefficient calculator in Fig. 3c ensure that $R+1$ samples of $t[n]$ are correlated against the shifted versions of the $s_k[n]$ sequences before the FIR filter coefficients are updated, and it follows from (3) that each of the $g^{(l)}[n]$ impulse responses have only one non-zero value for each set of $R+1$ samples. These observations imply that the average change of each coefficient calculator's accumulator during each set of $R+1$ samples is the same as it would be if $t[n]$ were replaced by $r[n]$ as given by (14) and the corresponding coefficient calculator were updated on just the first of every $R+1$ samples. Thus, instead of performing correlations on all $R+1$ of the $r^{(l)}[n]$ sequences simultaneously at each sample time, n, as done by the oversampling MNC technique, the SMNC technique equivalently performs correlations on all $R+1$ of the $r^{(l)}[n]$ sequences sequentially over successive sets of $R+1$ sample times.

C. Extension to Background Operation

With the $s_k[n]$ residue estimators implemented as shown in Fig. 3c, it is necessary for the statistics of the $s_k[n]$ sequences to be time-invariant as described above. This is easy to achieve during foreground calibration by ensuring that the statistics of $x[n]$ do not change over time. During background calibration, though, $x[n]$ is arbitrary, so it cannot be assumed that its statistics are time-invariant.

This problem can be solved by modifying the $s_k[n]$ residue estimators during background calibration as follows. The main DAC's DEM encoder ensures that the probability distribution of each $s_k[n]$ conditioned on $s_k[n] \neq 0$ is constant and independent of $x[n]$ [17]. Therefore, the problem can be solved by applying two changes to Fig. 3c during background calibration. The first change is to only update the bank of latches once every accumulator has been clocked $R+1$ times since the last time the bank of latches was clocked. The second change is to only clock the mth accumulator when $s_k[n+P-Q-m] \neq 0$ and $n \bmod (R+1)$ is distinct from $n' \bmod (R+1)$ for every prior time index n' of $s_k[n'+P-Q-m] \neq 0$ since the last time the bank of latches was updated. These modifications ensure that each accumulator in the kth coefficient calculator is updated with $r^{(l)}[n]$ information once for each value of $l = 0, 1, \ldots, R$ prior to each time the bank of latches is clocked and that the probability distribution of each $s_k[n]$ when the accumulators are updated is time-invariant.

IV. Convergence Analysis

Each $r^{(l)}[n]$ sequence in Fig. 5 can be written as

$$r^{(l)}[n] = r^{(l)}_{ideal}[n] + r_e^{(l)}[n] + r_c^{(l)}[n] \tag{15}$$

where $r^{(l)}_{ideal}[n]$ is what $r^{(l)}[n]$ would have been without the main DAC's mismatch noise and without the SMNC feedback loop, $r_e^{(l)}[n]$ represents error that would have been caused by the main DAC's mismatch noise without the SMNC feedback loop, and $r_c^{(l)}[n]$ represents the effect of the SMNC feedback loop. The correction DAC's error can be neglected, because it is much smaller than that of the main DAC as explained in Section II. Consequently, the relationship between $x_c[n]$ and $r_c^{(l)}[n]$ well-approximates that of a linear time-invariant (LTI) discrete-time system with impulse response $-h_c^{(l)}[n]$ (the negative sign simplifies the subsequent analysis). The system is causal and at least one clock delay is introduced by the ADC, so $h_c^{(l)}[n] = 0$ for all $n < 1$. Therefore,

$$r_c^{(l)}[n] = \sum_{i=1}^{\infty} x_c[n-i]\left(-h_c^{(l)}[i]\right), \tag{16}$$

where, as can be seen from Fig. 3,

$$x_c[n] = \sum_{k=1}^{M}\sum_{m=0}^{N-1} a_{k,m}[n]s_k[n+P-m]. \tag{17}$$

The kth portion of the main DAC's mismatch noise, $d_k(t)s_k[n_t]$ in (1), has the same form as the output of a DAC with input sequence $s_k[n]$ and T_s-periodic pulse shaping waveform, $d_k(t)$. Thus, the relationship between $s_k[n]$ and its contribution to $r_e^{(l)}[n]$ must also be that of a causal LTI discrete-time system with at least one clock delay. Denoting the LTI system's impulse response as $b_k^{(l)}[n]$, it follows from (1) that

$$r_e^{(l)}[n] = \sum_{k=1}^{M}\sum_{i=1}^{\infty} s_k[n-i]b_k^{(l)}[i]. \tag{18}$$

It follows from (4) that

$$t[n-l] = r^{(l)}[n-l] \quad \text{if } n \bmod (R+1) = 0. \tag{19}$$

As indicated in Fig. 3c, each FIR filter coefficient, $\alpha_{k,m}[n]$, only changes at times $n = 0, R+1, 2(R+1), \ldots$, i.e., when $n \bmod (R+1) = 0$. Therefore, Fig. 3c and (19) imply that for each of these values of n and for each $m = 0, 1, \ldots, N-1$,

$$a_{k,m}[n] = a_{k,m}[n-1] + K\sum_{l=0}^{R} s_k[n-l+P-Q-m]\, r^{(l)}[n-l]. \tag{20}$$

For all other values of n, $\alpha_{k,m}[n] = \alpha_{k,m}[n-1]$. Substituting (16)-(18) into (15), and substituting the result into (20), implies that

$$\begin{aligned} a_{k,m}[n] = a_{k,m}[n-1] &+ \sum_{l=0}^{R}\Bigg\{K\cdot s_k^2[n-l+P-Q-m] \\ &\left(b_k^{(l)}[Q-P+m] - \sum_{q=0}^{N-1} a_{k,q}\right. \\ &\left.[n-l-Q-m+q]\, h_c^{(l)}[Q+m-q]\right) \\ &+ K e_{k,m}^{(l)}[n-l]\Bigg\} \end{aligned} \tag{21}$$

for each n that satisfies $n \bmod (R+1) = 0$ and $m = 0, 1, \ldots, N-1$, where

$$\begin{aligned} & e_{k,m}^{(l)}[n] \\ & = s_k[n+P-Q-m] \\ & \quad \left\{ \sum_{i=1}^{\infty} \sum_{\substack{j=1 \\ j \neq k}}^{M} \left(s_j[n-i]\, b_j^{(l)}[i] - \sum_{q=0}^{N-1} a_{j,q}[n-i]\, s_j \right. \right. \\ & \qquad \left. [n+P-i-q]\, h_c^{(l)}[i] \right) \left(\sum_{\substack{i=1 \\ i \neq Q-P+m}}^{\infty} s_k[n-i]\, b_k^{(l)}[i] \right. \\ & \qquad \left. - \sum_{\substack{i=1 \\ i \neq Q+m-q}}^{\infty} \sum_{q=0}^{N-1} a_{k,q}[n-i]\, s_k[n+P-i-q]\, h_c^{(l)}[i] \right) \\ & \qquad \left. + r_{ideal}^{(l)}[n] \right\}. \end{aligned} \tag{22}$$

Equations (21), for $m = 0, 1, \ldots, N-1$ and each n that satisfies $n \bmod (R+1) = 0$, can be written in matrix form as

$$\begin{aligned} \mathbf{a}_k[n] &= \mathbf{a}_k[n-1] + K \sum_{m=0}^{N-1} \sum_{l=0}^{R} s_k^2[n-l+P-Q-m] \\ &\quad \times \left(\mathbf{b}_{k,m}^{(l)} - \sum_{q=0}^{N-1} \mathbf{H}_{m,q}^{(l)} \mathbf{a}_k[n-l-Q-m+q] \right) \\ &\quad + K \mathbf{e}_k[n] \end{aligned} \tag{23}$$

where

$$\mathbf{a}_k[n] = \left[a_{k,0}[n],\ a_{k,1}[n],\ \cdots,\ a_{k,N-1}[n] \right]^T, \tag{24}$$

$\mathbf{H}_{m,q}^{(l)}$ is an $N \times N$ matrix given by

$$\mathbf{H}_{m,q}^{(l)} = \left[h_{j,k} = \begin{cases} h_c^{(l)}[Q+j-k], & \text{if } j = m, k = q, \\ 0, & \text{otherwise,} \end{cases} \right], \tag{25}$$

$\mathbf{b}_{k,m}^{(l)}$ is an $N \times 1$ vector given by

$$\mathbf{b}_{k,m}^{(l)} = \left[b_j = \begin{cases} b_k^{(l)}[Q-P+j], & \text{if } j = m, \\ 0, & \text{otherwise,} \end{cases} \right], \tag{26}$$

and $\mathbf{e}_k[n]$ is an $N \times 1$ vector given by

$$\mathbf{e}_k[n] = \sum_{l=0}^{R} \left[e_{k,0}^{(l)}[n-l], e_{k,1}^{(l)}[n-l],\ \ldots,\ e_{k,N-1}^{(l)}[n-l] \right]^T. \tag{27}$$

The $\mathbf{a}_k[n]$ vector represents the kth adaptive FIR filter's coefficients at time n. The loop gain, K, is small by design to ensure that the coefficients converge to values with low variances, so (23) implies that $\mathbf{a}_k[n]$ depends only very weakly on any one of the time-shifted $s_k[n]$ sequences. Furthermore, all of the time-shifted $s_k[n]$ sequences are statistically independent. Consequently, $\mathbf{a}_k[n]$ is well-approximated as being statistically independent of each time-shifted $s_k[n]$ sequence. This type of *independence assumption* is widely used in the analysis of adaptive filters wherein slowly updated adaptive filter coefficients are assumed to be approximately independent from the data processed by the system [20]–[22].

Expanding the right side of (22) results in a sum of several products. Of these, $s_k[n+P-Q-m]s_j[n+P-i-q]\alpha_{j,q}[n-i]h_c^{(l)}[i]$ and $s_k[n+P-Q-m]s_k[n+P-i-q]\alpha_{k,q}[n-i]h_c^{(l)}[i]$ are the only products whose means are not exactly zero. However, their means are nearly zero by the independence assumption because $s_k[n+P-Q-m]s_j[n+P-i-q]$ and $s_k[n+P-Q-m]s_k[n+P-i-q]$ are zero mean. This implies that the mean of $\mathbf{e}_k[n]$ is well-approximated as zero, i.e.,

$$\bar{\mathbf{e}}_k[n] = \mathbf{0}. \tag{28}$$

Given that $s_k[n]$ is restricted to values of -1, 0, and 1, and its statistics are time-invariant, the mean of $s_k^2[n]$ is a constant, c_k, between 0 and 1, i.e.,

$$\overline{s_k^2}[n] = c_k \quad \text{for all } n \quad \text{where } 0 < c_k \leq 1. \tag{29}$$

Taking the expectation of (23), and applying (28), (29), and the independence assumption yields

$$\begin{aligned} & \bar{\mathbf{a}}_k[n] \\ & = \bar{\mathbf{a}}_k[n-1] + c_k K \\ & \quad \times \sum_{m=0}^{N-1} \sum_{l=0}^{R} \left(\mathbf{b}_{k,m}^{(l)} - \sum_{q=0}^{N-1} \mathbf{H}_{m,q}^{(l)} \bar{\mathbf{a}}_k[n-l-Q-m+q] \right) \end{aligned} \tag{30}$$

where $\bar{\mathbf{a}}_k[n]$ is the mean $\mathbf{a}_k[n]$ for each n that satisfies $n \bmod (R+1) = 0$. This can be rewritten as

$$\begin{aligned} \bar{\mathbf{a}}_k[n] &= \bar{\mathbf{a}}_k[n-1] - c_k K \sum_{m=0}^{N-1} \sum_{l=0}^{R} \sum_{q=0}^{N-1} \mathbf{H}_{m,q}^{(l)} \bar{\mathbf{a}}_k \\ &\quad [n-l-Q-m+q] + c_k K \mathbf{b}_k, \end{aligned} \tag{31}$$

where

$$\mathbf{b}_k = \sum_{m=0}^{N-1} \sum_{l=0}^{R} \mathbf{b}_{k,m}^{(l)}. \tag{32}$$

A simplification can be made by defining $\mathbf{H}_{\mathbf{c}}^{(J)}$ to be the sum of all $\mathbf{H}_{m,q}^{(l)}$ over $l = 0, 1, \ldots, R$, $m = 0, 1, \ldots, N-1$, and $q = 0, 1, \ldots, N-1$, restricted to values of m, l, and q that satisfy $l + Q + m - q = J$, such that

$$\sum_{m=0}^{N-1} \sum_{l=0}^{R} \sum_{q=0}^{N-1} \mathbf{H}_{m,q}^{(l)} = \sum_{J=1}^{R+Q+N-1} \mathbf{H}_{\mathbf{c}}^{(J)}. \tag{33}$$

The lower limit of J is 1 because (25) implies that $\mathbf{H}_{m,q}^{(l)} = \mathbf{0}$ for $m - q \leq -Q$ given that $h_c^{(l)}[n] = 0$ for all $n \leq 0$. Applying (33) to rearrange the triple sum in (31) and applying

$\overline{\mathbf{a}}_k[n] = \overline{\mathbf{a}}_k[n-1]$ for values of n that satisfy $n \bmod (R+1) \neq 0$ gives

$$\bar{\mathbf{a}}_k[n] = \bar{\mathbf{a}}_k[n-1] - \begin{cases} c_k K \displaystyle\sum_{J=1}^{R+Q+N-1} \mathbf{H}_{\mathbf{c}}^{(J)} \bar{\mathbf{a}}_k[n-J] + c_k K \mathbf{b}_k, \\ \quad \text{if } n \bmod (R+1) = 0, \\ \mathbf{0}, \quad \text{otherwise}, \end{cases} \tag{34}$$

for each integer, n.

Equation (34) is an N-dimensional matrix difference equation that converges if and only if $\overline{\mathbf{a}_k}[n] \to \mathbf{a}'_k$ as $n \to \infty$, where $\mathbf{a}'_k$ is a constant vector. Taking the limit of (34) as $n \to \infty$ implies that if the system converges then

$$\mathbf{a}'_k = \mathbf{a}'_k - c_k K \mathbf{H}_{\mathbf{c}} \mathbf{a}'_k + c_k K \mathbf{b}_k \tag{35}$$

where

$$\mathbf{H}_{\mathbf{c}} = \sum_{J=1}^{R+Q+N-1} \mathbf{H}_{\mathbf{c}}^{(J)}. \tag{36}$$

It follows from (35) that if the system converges, then

$$\mathbf{a}'_k = \mathbf{H}_{\mathbf{c}}^{-1} \mathbf{b}_k. \tag{37}$$

Equations (25), (33) and (36) imply that

$$\mathbf{H}_{\mathbf{c}} = \left[h_{j,k} = h_c[Q + j - k]\right], \quad \text{where } h_c[n] = \sum_{l=0}^{R} h_c^{(l)}[n]. \tag{38}$$

Given that $-h_c^{(l)}[n]$ is the impulse response of the transfer function between $x_c[n]$ and $r_c^{(l)}[n]$, it follows from (14) and Theorem 1 in Section III that $h_c[n]$ is the impulse response of the transfer function between $x_c[n]$ and $r[n]$ in the oversampling version of the MNC technique shown in Fig. 1. As proven in [15], the FIR filter coefficients in the oversampling MNC technique converge to values that satisfy (37) with $\mathbf{H}_{\mathbf{c}} = [h_{j,k} = h_c[Q+j-k]]$. Therefore, provided the FIR filter coefficients in the subsampling version of the MNC technique converge, they must converge to the same values as those of the oversampling version of the MNC technique.

It remains to show that the subsampling MNC technique's coefficients converge, i.e., that $\overline{\mathbf{a}}_k[n]$ always converges to $\mathbf{a}'_k$ as $n \to \infty$ for each k. This is done by showing that $\mathbf{z}_k[n]$ converges to $\mathbf{0}$ as $n \to \infty$, where

$$\mathbf{z}_k[n] = \bar{\mathbf{a}}_k[n] - \mathbf{a}'_k, \tag{39}$$

and, as implied by (34) and (35),

$$\mathbf{z}_k[n] = \mathbf{z}_k[n-1] - \begin{cases} c_k K \displaystyle\sum_{J=1}^{R+Q+N-1} \mathbf{H}_{\mathbf{c}}^{(J)} \mathbf{z}_k[n-J] \\ \quad \text{if } n \mod (R+1) = 0 \\ \mathbf{0}, \quad \text{otherwise}. \end{cases} \tag{40}$$

The analysis makes use of vector and matrix norms. For any N-dimensional vector $\mathbf{v} = [v_j]$ and $N \times N$ matrix $\mathbf{H} = [h_{j,k}]$, the *vector norm* of $\mathbf{v}$ and the *matrix norm* of $\mathbf{H}$are defined as

$$\|\mathbf{v}\| = \max_{1 \le m \le N} |v_m| \quad \text{and} \quad \|\mathbf{H}\|_1 = \max_{1 \le m \le N} \sum_{n=1}^{N} |h_{m,n}|. \tag{41}$$

Theorem 2 presented below, and proven in the Appendix, shows that $\mathbf{z}_k[n]$ converges to $\mathbf{0}$ as $n \to \infty$ for each k if $h_c[n]$, Q, and K satisfy certain conditions. It does so by showing that $\|\mathbf{z}_k[n]\| \to 0$ as $n \to \infty$. To simplify the notation, the system's initial conditions are taken to be zero, i.e., $\overline{\mathbf{a}_k}[n] = 0$ for all $n < 0$, so (39) implies that $\mathbf{z}_k[n] = -\mathbf{a}'_k$ for all $n < 0$.

Theorem 2: Suppose $0 \le r < 1$, $0 < g < 1$, $0 < 2Kh_c[Q] < 1$, and $\mathbf{z}_k[n] = -\mathbf{a}'_k$ for all $n < 0$, where

$$r = \frac{1}{h_c[Q]} \sum_{\substack{m=Q-(N-1) \\ m \neq Q}}^{Q+(N-1)} |h_c[m]| \tag{42}$$

$$g = \sum_{J_1=1}^{R+Q+N-1} \sum_{J_2=1}^{R+Q+N-1} \frac{\left\| \mathbf{H}_{\mathbf{c}}^{(J_1)} \mathbf{H}_{\mathbf{c}}^{(J_2)} \right\|_1 \left(1 - (1 - 2Kh_c[Q])^{J_1 - 1}\right)}{2h_c^2[Q](1-r)(1-2Kh_c[Q])^{J_1+J_2-2}}. \tag{43}$$

Then

$$\|\mathbf{z}_k[n]\| \le \left\|\mathbf{a}'_k\right\| (1 - c_k K (1-r)(1-g) h_c[Q])^{\lfloor n/(R+1) \rfloor + 1} \tag{44}$$

for all $n \ge 0$, where $\lfloor n/(R+1) \rfloor$ is the largest integer less or equal to $n/(R+1)$.

Inequality (44) implies that $||\mathbf{z}_k[n]||$ converges to 0 following an exponential-like trajectory for each k. This and (39) imply $\mathbf{a}_k[n] \to \mathbf{a}'_k$ for each k. Therefore, the conditions in the hypothesis of the theorem are sufficient to guarantee the convergence of SMNC.

The theorem's hypothesis places certain requirements on the values of $h_c[n]$, Q, and K. The $0 \le r < 1$ requirement and the definition of r in (42) imply that the $h_c[Q]$ must be positive and larger than the sum of multiple adjacent samples of the impulse response. As explained in [15], $0 \le r < 1$ is also a necessary condition for the convergence of the oversampling version of the MNC technique and can be easily satisfied in practice. The requirement that $0 \le g < 1$ and $0 < 2Kh_c[Q] < 1$ sets an upper bound on K.

Theorem 2 also provides insight into the convergence rate. It indicates that increasing K increases the convergence rate. It also implies that reducing the probability of $s_k[n] = 0$ over time, which increases the value of c_k in (29), leads to faster convergence.

Theorem 2 predicts how the expected value of each filter coefficient evolves over time, but it does not provide insight into the variance of the noise component of each filter coefficient. Intuitive reasoning similar to that in [15] and extensive simulations indicate that the noise variance can be made arbitrarily small by reducing K. Therefore, K represents a tradeoff between convergence accuracy and convergence speed.

V. Simulation Results

Three sets of computer simulation results are presented in this section. The first set demonstrates that oversampling is indeed required for the original version of the MNC technique presented in [15] to work properly. The second set

demonstrates the effectiveness and of the SMNC technique. The third set demonstrates the transient convergence behavior of the SMNC technique and compares it to that predicted by Theorem 2 presented in the previous section.

All simulations implement the same main DAC and correction DAC architectures, the same DAC clock-rate of $f_s = 3$ GHz, and the same MNC design parameters P, Q, N and K of 3, 21, 9 and $8 \cdot 10^{-6}$, respectively. As in [16], the main DAC consists of the DEM encoder presented in [17] followed by 36 1-bit DACs. The DEM encoder converts the 14-bit main DAC input sequence, $x[n]$, into 36 1-bit sequences, each of which drives a 1-bit DAC with weight K_i. For $i = 1, 2, \ldots, 20$ the values of K_i are $1, 1, 2, 2, 4, 4, \ldots, 512, 512$, respectively, and for $i = 21, 22, \ldots, 36$, each K_i has a value of 1024. Each 1-bit DAC implements a 25% return-to-zero (RZ) phase to avoid ISI. Also as in [16], the correction DAC is implemented without DEM or calibration and its minimum step-size is $\Delta/4$, where Δ is the main DAC's minimum step-size.

The same set of mismatch noise parameters was used for each simulation. Dynamic mismatch noise was simulated by inserting a random Gaussian delay with a standard deviation of 0.6 ps on each 1-bit DAC clock time. Static mismatch error was simulated by introducing 1-bit DAC step-size errors. The step-size error for each of the 1024-weight 1-bit DACs was chose as a Gaussian random variable with a standard deviation of 0.15% of the 1-bit DAC's step size, 1024Δ. That of each of the other 1-bit DACs, including those in the correction DAC, were chosen similarly, except that the standard deviation was divided by the square root of the 1-bit DAC's step-size divided by 1024Δ.

Each simulation includes a 5-bit VCO-based ADC of the type implemented in the IC presented in [16]. Aside from its noise and distortion, the VCO-based ADC is equivalent to a sinc lowpass filter followed by a first-order $\Delta\Sigma$ modulator ADC with 5-bit quantization [23]. No ADC calibration was applied, so the ADC's nonlinearity is high: with a full-scale sinusoidal input waveform, the second and third harmonic distortion terms are -26 dBc and -47 dBc, respectively.

Fig. 6 shows simulated output spectra from the system with the original version of the MNC technique and a -1 dBFS sinusoidal input signal, with and without oversampling the ADC. Fig. 6a shows the output spectrum with MNC disabled and Fig. 6b shows the output spectrum with MNC enabled for an oversampling ratio of $R = 5$. This oversampling ratio in conjunction with the sinc lowpass filtering inherent to the VCO-based ADC is sufficiently high for the aliasing error to be negligible over the DAC's 0 to $0.42 f_s$ signal band.[2] In this case, MNC improves the SNDR by 18 dB over the DAC's signal band. Fig. 6c shows the output spectrum for MNC enabled but without oversampling, i.e., with the ADC sampled at f_s. Some SNDR improvement still occurs in this case relative to the case with MNC disabled, because aliasing does not prevent MNC from canceling a low-frequency portion of the mismatch noise. However, the aliasing prevents cancellation of higher-frequency mismatch noise and, therefore, prevents significant SNDR improvement.

[2]The decimation filter's non-ideal transition bandwidth causes aliasing at frequencies between $0.42 f_s$ and $0.5 f_s$, which limits MNC accuracy over this band. As explained in [16], this exclusion band can be reduced by increasing the digital filter's complexity.

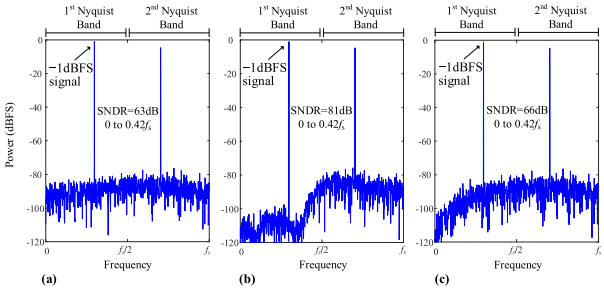

Fig. 6. Representative simulated output spectra with a) MNC off, b) MNC on, and c) MNC on but without oversampling.

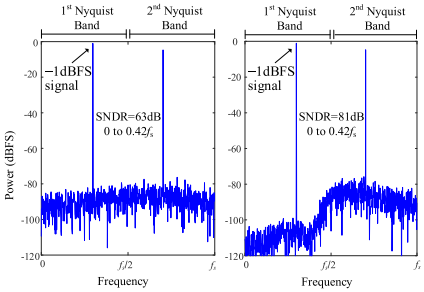

Fig. 7. Representative simulated output spectra without/with SMNC.

Fig. 7 shows the simulated output spectrum from the system with the SMNC technique and a -1 dBFS sinusoidal input signal for an ADC sample-rate of $5 f_s/6$, i.e., $R = 5$. Compared to the case without MNC shown in Fig. 6a, the SMNC technique improves the SNDR by 18 dB. This result supports the paper's assertion that the SMNC technique provides roughly the same SNDR improvement as the original MNC technique despite aliasing from not oversampling.

In the simulations described above, the adaptive FIR filter coefficients were obtained during foreground calibration mode and then frozen for use during normal DAC mode. During foreground calibration, $x[n]$ was chosen to toggle pseudo-randomly between -2389.5Δ and -2388.5Δ. In principle, any $x[n]$ sequence with time-invariant statistics as required by the foreground mode version of the SMNC technique would work, but this choice of $x[n]$ is attractive because of its small dynamic range, which simplifies the ADC, and it results in $sk[n]$ sequences with a low percentage of zero values, which is beneficial for rapid convergence.

The SMNC technique's foreground calibration convergence time for the simulation results shown in Fig. 7 was about 3 ms. This is approximately $R = 5$ times longer than that of the original MNC technique, as expected. Much as in the case of the original MNC technique as explained in [15], the convergence time of the SMNC technique can be decreased by increasing K, but this comes at the expense of increased noise variance of each adaptive FIR filter's coefficients. A practical way to reduce the convergence time without a noise penalty is to use a relatively large value of K during an initial portion of foreground calibration mode so the conversion rate is relatively high while the adaptive FIR filter coefficients get close to their final values, and then reduce K during the

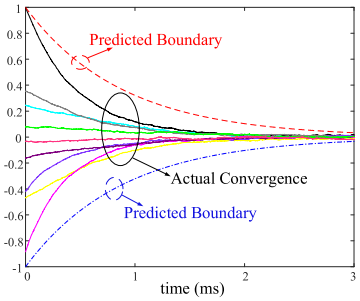

Fig. 8. Transient convergence behavior of the SMNC technique's adaptive filter coefficients for representative value of k.

final portion of foreground calibration mode to reduce the coefficient variances.

Fig. 8 shows the transient convergence behavior of the SMNC technique's adaptive FIR filter coefficients for a representative value of k and a constant value of K, i.e., $K = 8 \cdot 10^{-6}$. The solid curves represent the differences between the instantaneous values of the coefficients, $\alpha_{k,m}[n]$, and their ideal values for $m = 0, 1, \ldots, N-1$ and a representative value of k. The definition of $\mathbf{z}_k[n]$ in (39) implies that the mean of each curve must be bounded by $-||\mathbf{z}_k[n]||$ and $||\mathbf{z}_k[n]||$. These upper and lower bounds, as predicted in Theorem 1, are plotted as dashed curves in the figure. The simulation results show that although the noise in the system causes the filter coefficients to fluctuate around their mean values, they are still mostly within the predicted upper and lower mean bounds.

APPENDIX

The proof uses the following well-known matrix theory results [24]. For any $N \times 1$ vectors $\mathbf{v}$ and $\mathbf{w}$, and any $N \times N$ matrix $\mathbf{H}$, the vector and matrix norms defined in (41) are such that

$$\|\mathbf{H}\mathbf{v}\| \leq \|\mathbf{H}\|_1 \|\mathbf{v}\| \tag{45}$$

and

$$\|\mathbf{v}\| - \|\mathbf{w}\| \leq \|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|. \tag{46}$$

A. Proof of Theorem 2

If $\mathbf{a}'_k = 0$, then (40) and the initial condition of $\mathbf{z}_k[n] = -\mathbf{a}'_k$ for all $n < 0$ imply that $\mathbf{z}_k[n] = 0$ for all $n \geq 0$ and (44) holds. The rest of the proof considers the case of $\mathbf{a}'_k \neq 0$.

The proof applies mathematical induction. The *inductive step*, which is proven shortly, is: for any integer $n \geq 0$, if

$$\frac{\|\mathbf{z}_k[i]\|}{\|\mathbf{z}_k[i-1]\|} \geq 1 - 2c_k K h_c[Q], \tag{47}$$

for all $i < n$, then the theorem's hypothesis ensures that (47) also holds for $i = n$ and

$$\frac{\|\mathbf{z}_k[n]\|}{\|\mathbf{z}_k[n-1]\|} \leq \begin{cases} 1 - c_k K (1-r)(1-g) h_c[Q], \\ \quad \text{if } n \bmod (R+1) = 0, \\ 1, \quad \text{otherwise.} \end{cases} \tag{48}$$

The induction *base step* requires that (47) hold for all $i < 0$. The proof of the base step follows from the initial condition of $\mathbf{z}_k[n] = -\mathbf{a}'_k$ for all $n < 0$ and (41). Hence, if the inductive step is true, it follows from induction that (47) and (48) must hold for all $n \geq 0$. In addition, applying (48) for $n \geq 0$ with the initial condition of $\mathbf{z}_k[-1] = -\mathbf{a}'_k$ leads to (44).

It remains to show that the inductive step is true. This is shown in the remainder of the proof.

If $n \bmod (R+1) \neq 0$, it follows from (40) that $\mathbf{z}_k[n] = \mathbf{z}_k[n-1]$, thus (47) holds for $i = n$ and (48) holds. The rest of analysis considers the case when $n \bmod (R+1) = 0$. In this case, (40) reduces to

$$\mathbf{z}_k[n] = \mathbf{z}_k[n-1] - c_k K \sum_{J=1}^{R+Q+N-1} \mathbf{H}_\mathbf{c}^{(J)} \mathbf{z}_k[n-J]. \tag{49}$$

It follows from (36) that (49) can be rewritten as

$$\begin{aligned} \mathbf{z}_k[n] = \mathbf{z}_k[n-1] &- c_k K \mathbf{H}_\mathbf{c} \mathbf{z}_k[n-1] \\ &- c_k K \sum_{J=1}^{R+Q+N-1} \mathbf{H}_\mathbf{c}^{(J)} (\mathbf{z}_k[n-J] - \mathbf{z}_k[n-1]) \end{aligned} \tag{50}$$

and further rewritten as

$$\begin{aligned} \mathbf{z}_k[n] = (\mathbf{I} - c_k K \mathbf{H}_\mathbf{c}) \mathbf{z}_k[n-1] &- c_k K \\ &\times \sum_{J=1}^{R+Q+N-1} \sum_{m=1}^{J-1} \mathbf{H}_\mathbf{c}^{(J)} (\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m]) \end{aligned} \tag{51}$$

where $\mathbf{I}$ is an $N \times N$ identity matrix. Taking the vector norm on both sides of (51) and applying (46) yields

$$\begin{aligned} \|\mathbf{z}_k[n]\| \leq \|(\mathbf{I} - c_k K \mathbf{H}_\mathbf{c}) \mathbf{z}_k[n-1]\| &+ \sum_{J=1}^{R+Q+N-1} \\ &\times \sum_{m=1}^{J-1} \left\| c_k K \mathbf{H}_\mathbf{c}^{(J)} (\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m]) \right\| \end{aligned} \tag{52}$$

and

$$\begin{aligned} &\|\mathbf{z}_k[n]\| \\ &\geq \|(\mathbf{I} - c_k K \mathbf{H}_\mathbf{c}) \mathbf{z}_k[n-1]\| \\ &\quad - \sum_{J=1}^{R+Q+N-1} \sum_{m=1}^{J-1} \left\| c_k K \mathbf{H}_\mathbf{c}^{(J)} (\mathbf{z}_k[n-m-1] - \mathbf{z}_k[n-m]) \right\|. \end{aligned} \tag{53}$$

The definition of r in (42) and the condition $0 \leq r < 1$ in Theorem 2 imply that $h_c[Q]$ is positive. Therefore, it follows from the definition of $\mathbf{H}_\mathbf{c}$ in (38) and the definition of the matrix norm in (41) that

$$\|h_c[Q]\mathbf{I} - \mathbf{H}_\mathbf{c}\| \leq h_c[Q] r. \tag{54}$$

For any real N-dimensional column vector $\mathbf{v}$, the vector norm of $(\mathbf{I} - cK\mathbf{H}_\mathbf{c})\mathbf{v}$ can be written as

$$\begin{aligned} \|(\mathbf{I} - c_k K \mathbf{H}_\mathbf{c}) \mathbf{v}\| = \|(1 - c_k K h_c[Q]) \mathbf{v} \\ + c_k K (h_c[Q]\mathbf{I} - \mathbf{H}_\mathbf{c}) \mathbf{v}\|. \end{aligned} \tag{55}$$

Applying (45) and (46) yields

$$\|(\mathbf{I}-c_kK\mathbf{H_c})\,\mathbf{v}\| \le (1-c_kKh_c[Q])\,\|\mathbf{v}\| + c_kK\,\|h_c[Q]\,\mathbf{I}-\mathbf{H_c}\|_1\,\|\mathbf{v}\| \quad (56)$$

and

$$\|(\mathbf{I}-c_kK\mathbf{H_c})\,\mathbf{v}\| \ge (1-c_kKh_c[Q])\,\|\mathbf{v}\| - c_kK\,\|h_c[Q]\,\mathbf{I}-\mathbf{H_c}\|_1\,\|\mathbf{v}\|. \quad (57)$$

Applying (54) to (56) and (57) yields

$$\|(\mathbf{I}-c_kK\mathbf{H_c})\,\mathbf{v}\| \le (1-c_kK\,(1-r)\,h_c[Q])\,\|\mathbf{v}\| \quad (58)$$

and

$$\|(\mathbf{I}-c_kK\mathbf{H_c})\,\mathbf{v}\| \ge (1-c_kK\,(1+r)\,h_c[Q])\,\|\mathbf{v}\|. \quad (59)$$

Replacing $\mathbf{v}$by $\mathbf{z}_k[n-1]$ in (58) and (59), and substituting the results into (52) and (53) gives

$$\|\mathbf{z}_k[n]\| \le (1-c_kK\,(1-r)\,h_c[Q])\,\|\mathbf{z}_k[n-1]\| + \sum_{J=1}^{R+Q+N-1}\sum_{m=1}^{J-1}\left\|c_kK\mathbf{H_c}^{(J)}\,(\mathbf{z}_k[n-m-1]-\mathbf{z}_k[n-m])\right\| \quad (60)$$

and

$$\|\mathbf{z}_k[n]\| \ge (1-c_kK\,(1+r)\,h_c[Q])\,\|\mathbf{z}_k[n-1]\| - \sum_{J=1}^{R+Q+N-1}\sum_{m=1}^{J-1}\left\|c_kK\mathbf{H_c}^{(J)}\,(\mathbf{z}_k[n-m-1]-\mathbf{z}_k[n-m])\right\|. \quad (61)$$

Equation (40) with the initial condition $\mathbf{z}_k[n]=-\mathbf{a}_k'$ for $n<0$ implies that each $\mathbf{z}_k[n-m-1]-\mathbf{z}_k[n-m]$ in (60) and (61) is either

$$c_kK\sum_{J=1}^{R+Q+N-1}\mathbf{H_c}^{(J)}\mathbf{z}_k[n-m-J] \text{ or } \mathbf{0}. \quad (62)$$

This observation applied to (60) and (61) results in

$$\|\mathbf{z}_k[n]\| \le (1-c_kK\,(1-r)\,h_c[Q])\,\|\mathbf{z}_k[n-1]\| + \sum_{J_1=1}^{R+Q+N-1}\sum_{m=1}^{J_1-1}\left\|c_k^2K^2\sum_{J_2=1}^{R+Q+N-1}\mathbf{H_c}^{(J_1)}\mathbf{H_c}^{(J_2)}\mathbf{z}_k[n-m-J_2]\right\| \quad (63)$$

and

$$\|\mathbf{z}_k[n]\| \ge (1-c_kK\,(1+r)\,h_c[Q])\,\|\mathbf{z}_k[n-1]\| - \sum_{J_1=1}^{R+Q+N-1}\sum_{m=1}^{J_1-1}\left\|c_k^2K^2\sum_{J_2=1}^{R+Q+N-1}\mathbf{H_c}^{(J_1)}\mathbf{H_c}^{(J_2)}\mathbf{z}_k[n-m-J_2]\right\|. \quad (64)$$

Applying (45) with $\mathbf{H}$ replaced by $c_k^2K^2\mathbf{H_c}^{(J_1)}\mathbf{H_c}^{(J_2)}$, substituting the result into (63) and (64), then applying (46) yields

$$\|\mathbf{z}_k[n]\| \le (1-c_kK\,(1-r)\,h_c[Q])\,\|\mathbf{z}_k[n-1]\| + c_k^2K^2 \times \sum_{J_1=1}^{R+Q+N-1}\sum_{J_2=1}^{R+Q+N-1}\left\|\mathbf{H_c}^{(J_1)}\mathbf{H_c}^{(J_2)}\right\|_1 \times \sum_{m=1}^{J_1-1}\|\mathbf{z}_k[n-m-J_2]\| \quad (65)$$

and

$$\|\mathbf{z}_k[n]\| \ge (1-c_kK\,(1+r)\,h_c[Q])\,\|\mathbf{z}_k[n-1]\| - c_k^2K^2 \times \sum_{J_1=1}^{R+Q+N-1}\sum_{J_2=1}^{R+Q+N-1}\left\|\mathbf{H_c}^{(J_1)}\mathbf{H_c}^{(J_2)}\right\|_1 \times \sum_{m=1}^{J_1-1}\|\mathbf{z}_k[n-m-J_2]\|. \quad (66)$$

It follows from (47), $0<c_k\le 1$ in (29), and Theorem 2's hypothesis of $0<2\ Kh_c[Q]<1$ that

$$\|\mathbf{z}_k[n-i]\| \le \|\mathbf{z}_k[n-1]\|\,(1-2c_kKh_c[Q])^{-i+1} \quad (67)$$

holds for $i=2,3,4,\ldots$. Therefore,

$$\sum_{m=1}^{J_1-1}\|\mathbf{z}_k[n-m-J_2]\| \le \|\mathbf{z}_k[n-1]\|\sum_{m=1}^{J_1-1}(1-2c_kKh_c[Q])^{-m-J_2+1}. \quad (68)$$

The sum in the right side of (68) can be expanded via the geometric series formula as

$$\sum_{m=1}^{J_1-1}(1-2c_kKh_c[Q])^{-m-J_2+1} = \frac{1-(1-2c_kKh_c[Q])^{J_1-1}}{2c_kKh_c[Q]\,(1-2c_kKh_c[Q])^{J_1+J_2-2}}. \quad (69)$$

It follows from (29) that

$$\sum_{m=1}^{J_1-1}(1-2c_kKh_c[Q])^{-m-J_2+1} \le \frac{1-(1-2Kh_c[Q])^{J_1-1}}{2c_kKh_c[Q]\,(1-2Kh_c[Q])^{J_1+J_2-2}}. \quad (70)$$

Substituting (70) into (68) and substituting the result into (65) and (66) yields

$$\frac{\|\mathbf{z}_k[n]\|}{\|\mathbf{z}_k[n-1]\|} \le 1-c_kK\,(1-r)\,h_c[Q]+\sum_{J_1=1}^{R+Q+N-1}\sum_{J_2=1}^{R+Q+N-1}\frac{c_kK\left\|\mathbf{H_c}^{(J_1)}\mathbf{H_c}^{(J_2)}\right\|_1\left(1-(1-2Kh_c[Q])^{J_1-1}\right)}{2h_c[Q]\,(1-2Kh_c[Q])^{J_1+J_2-2}} \quad (71)$$

and

$$\frac{\|\mathbf{z}_k[n]\|}{\|\mathbf{z}_k[n-1]\|} \geq 1 - c_k K\,(1+r)\,h_c\,[Q] - \sum_{J_1=1}^{R+Q+N-1}\sum_{J_2=1}^{R+Q+N-1} \frac{c_k K \left\|\mathbf{H}_{\mathbf{c}}^{(J_1)}\mathbf{H}_{\mathbf{c}}^{(J_2)}\right\|_1 \left(1-(1-2Kh_c\,[Q])^{J_1-1}\right)}{2h_c\,[Q]\,(1-2Kh_c\,[Q])^{J_1+J_2-2}}. \quad (72)$$

Substituting (43) into (71) yields (48) for $n \mod (R+1) = 0$, and substituting (43) into (72) yields

$$\frac{\|\mathbf{z}_k[n]\|}{\|\mathbf{z}_k[n-1]\|} \geq 1 - c_k K\,(1+r)\,h_c\,[Q] - c_k K g\,(1-r)\,h_c\,[Q] = 1 - c_k K\,(2-(1-r)\,(1-g))\,h_c\,[Q]. \quad (73)$$

This implies that (47) holds for $i = n$ for any values of r and g that satisfy $0 \leq r < 1$ and $0 < g < 1$.

REFERENCES

[1] W. Schofield, D. Mercer, and L. S. Onge, "A 16 b 400 MS/s DAC with <−80 dBc IMD to 300 MHz and <−160 dBm/Hz noise power spectral density," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 1, Feb. 2003, pp. 126–482.

[2] Q. Huang, P. A. Francese, C. Martelli, and J. Nielsen, "A 200 Ms/s 14 b 97 mW DAC in 0.18 μm CMOS," in *IEEE Int. Solid State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 1, Feb. 2004, pp. 364–532.

[3] H.-H. Chen, J. Lee, J. Weiner, Y.-K. Chen, and J.-T. Chen, "A 14-bit 150 MS/s CMOS DAC with digital background calibration," in *Proc. Symp. VLSI Circuits*, Jun. 2006, pp. 51–52.

[4] M. Clara, W. Klatzer, B. Seger, A. D. Giandomenico, and L. Gori, "A 1.5 V 200 MS/s 13 b 25 mW DAC with randomized nested background calibration in 0.13 μm CMOS," in *IEEE Int. Solid State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 250–600.

[5] M. Clara, W. Klatzer, D. Gruber, A. Marak, B. Seger, and W. Pribyl, "A 1.5 V 13 bit 130-300 MS/s self-calibrated DAC with active output stage and 50 MHz signal bandwidth in 0.13 μm CMOS," in *Proc. 34th Eur. Solid-State Circuits Conf.*, Sep. 2008, pp. 262–265.

[6] B. Catteau, P. Rombouts, J. Raman, and L. Weyten, "An on-line calibration technique for mismatch errors in high-speed DACs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 7, pp. 1873–1883, Aug. 2008.

[7] C.-H. Lin *et al.*, "A 12 bit 2.9 GS/s DAC With IM3 ≪−60 dBc beyond 1 GHz in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 12, pp. 3285–3293, Dec. 2009.

[8] Y. Tang *et al.*, "A 14 bit 200 MS/s DAC With SFDR >78 dBc, IM3 <−83 dBc and NSD <−163 dBm/Hz across the whole Nyquist band enabled by dynamic-mismatch mapping," *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1371–1381, Jun. 2011.

[9] S. Spiridon *et al.*, "A 375 mW multimode DAC-based transmitter with 2.2 GHz signal bandwidth and in-band IM3 <−58 dBc in 40 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 7, pp. 1595–1604, Jul. 2013.

[10] W.-T. Lin, H.-Y. Huang, and T.-H. Kuo, "A 12-bit 40 nm DAC achieving SFDR >70 dB at 1.6 GS/s and IMD <−61 dB at 2.8 GS/s with DEMDRZ technique," *IEEE J. Solid-State Circuits*, vol. 49, no. 3, pp. 708–717, Mar. 2014.

[11] S. M. Lee *et al.*, "A 14 b 750 MS/s DAC in 20 nm CMOS with <−168 dBm/Hz noise floor beyond Nyquist and 79 dBc SFDR utilizing a low glitch-noise hybrid R-2R architecture," in *Proc. Symp. VLSI Circuits*, Jun. 2015, pp. C164–C165.

[12] G. Engel, M. Clara, H. Zhu, and P. Wilkins, "A 16-bit 10 Gsps current steering RF DAC in 65 nm CMOS achieving 65 dBc ACLR multi-carrier performance at 4.5 GHz Fout," in *Proc. Symp. VLSI Circuits* Jun. 2015, pp. C166–C167.

[13] S. Su and M. S.-W. Chen, "A 12-bit 2 GS/s dual-rate hybrid dac with pulse-error pre-distortion and in-band noise cancellation achieving >74 dBc SFDR and <−80 dBc IM3 up to 1 GHz in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 51, no. 12, pp. 2963–2978, Dec. 2016.

[14] C.-H. Lin *et al.*, "A 16 b 6 GS/s Nyquist DAC with IMD <−90 dBc up to 1.9 GHz in 16 nm CMOS," in *IEEE Int. Solid State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 360–362.

[15] D. Kong and I. Galton, "Adaptive cancellation of static and dynamic mismatch error in continuous-time DACs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 2, pp. 421–433, Feb. 2018.

[16] D. Kong, K. Rivas-Rivera, and I. Galton, "A 600 MS/s DAC with over 87 dB SFDR and 77 dB peak SNDR enabled by adaptive cancellation of static and dynamic mismatch error," *IEEE J. Solid-State Circuits*, to be published. [Online]. Available: http://ispg.ucsd.edu/unpublished-paper/

[17] K. L. Chan, J. Zhu, and I. Galton, "Dynamic element matching to prevent nonlinear distortion from pulse-shape mismatches in high-resolution DACs," *IEEE J. Solid-State Circuits*, vol. 43, no. 9, pp. 2067–2078, Sep. 2008.

[18] J. Remple and I. Galton, "The effects of inter-symbol interference in dynamic element matching DACs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 1, pp. 14–23, Jan. 2017.

[19] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Hoboken, NJ, USA: Prentice-Hall, 1993.

[20] J. E. Mazo, "On the independence theory of equalizer convergence," *Bell Syst. Tech. J.*, vol. 58, no. 5, pp. 963–993, May/Jun. 1979.

[21] B. Widrow, "Adaptive filters," in *Aspects of Network and System Theory*, R. E. Kalman and N. DeClaris, Eds. New York, NY, USA: Holt, Rinehart and Winston, 1971, pp. 563–586.

[22] W. A. Gardner, "Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique," *Signal Process.*, vol. 6, no. 2, pp. 113–133, Apr. 1984.

[23] G. Taylor and I. Galton, "A mostly-digital variable-rate continuous-time delta-sigma modulator ADC," *IEEE J. Solid-State Circuits*, vol. 45, no. 12, pp. 2634–2646, Dec. 2010.

[24] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

Derui Kong received the B.S. degree in microelectronics from Fudan University, Shanghai, China, in 2007, and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2009. From 2009 to 2016, he was with Qualcomm Technologies, Inc., San Diego, CA, USA, where he designed data converters for cellular applications. He is currently pursuing the Ph.D. degree with the University of California, San Diego. His research interests are in the analysis and design of mixed-signal integrated circuits and systems.

Ian Galton received the Sc.B. degree from Brown University in 1984, and the M.S. and Ph.D. degrees from the California Institute of Technology in 1989 and 1992, respectively, all in electrical engineering. Since 1996, he has been a Professor of electrical engineering with the University of California, San Diego, CA, USA, where he teaches and conducts research in the field of mixed-signal integrated circuits and systems for communications. His research involves the invention, analysis, and integrated circuit implementation of critical communication system blocks, such as data converters and phase-locked loops.