Large-Area Microphone Array for Audio Source Separation Based on a Hybrid Architecture Exploiting Thin-Film Electronics and CMOS

Josue Sanz-Robinson, Liechao Huang, *Student Member, IEEE*, Tiffany Moy, *Student Member, IEEE*, Warren Rieutort-Louis, *Member, IEEE*, Yingzhe Hu, Sigurd Wagner, *Life Fellow, IEEE*, James C. Sturm, *Fellow, IEEE*, and Naveen Verma, *Member, IEEE*

***Abstract*—We present a system for reconstructing-independent voice commands from two simultaneous speakers, based on an array of spatially distributed microphones. It adopts a hybrid architecture, combining large-area electronics (LAE), which enables a physically expansive array ($>$1 m width), and a CMOS IC, which provides superior transistors for readout and signal processing. The array enables us to: 1) select microphones closest to the speakers to receive the highest SNR signal; 2) use multiple spatially diverse microphones to enhance robustness to variations due to microphones and sound propagation in a practical room. Each channel consists of a thin-film transducer formed from polyvinylidene fluoride (PVDF), a piezopolymer, and a localized amplifier composed of amorphous silicon (a-Si) thin-film transistors (TFTs). Each channel is sequentially sampled by a TFT scanning circuit, to reduce the number of interfaces between the large-area electronics (LAE) and CMOS IC. A reconstruction algorithm is proposed, which exploits the measured transfer function between each speaker and microphone, to separate two simultaneous speakers. The algorithm overcomes 1) sampling-rate limitations of the scanning circuits and 2) sensitivities to microphone placement and directionality. An entire system with eight channels is demonstrated, acquiring and reconstructing two simultaneous audio signals at 2 m distance from the array achieving a signal-to-interferer (SIR) ratio improvement of $\sim$12 dB.**

***Index Terms*—Amorphous silicon (a-Si), critically sampled, flexible electronics, large area electronics, microphone array, source separation, thin-film, thin-film transistors (TFT).**

I. Introduction

As ELECTRONICS becomes ever more pervasive in our daily lives, it will no longer be confined to our phones and tablets, but rather will be seamlessly integrated into the environment in which we live, work, and play. In such a form factor, there is an opportunity for systems that foster collaborative spaces and enhance interpersonal interactions. With this motivation, we present a system that enables voices signals from multiple simultaneous speakers to be separated and reconstructed, ultimately to be fed to a voice-command recognition engine for controlling electronic systems. The cornerstone of the system is a spatially distributed microphone array, which exploits the diversity of the audio signal received by different microphones to separate two simultaneous sound sources. To create such an array, we take advantage of large area electronics (LAE).

LAE is based on thin-film semiconductors and insulators deposited at low temperatures, which enables compatibility with a wide range of materials. This has led to the development of diverse transducers, including strain, light [1], gas [2], and pressure sensors [3], integrated on substrates such as glass or plastic, which can be large ($\sim$m^2), thin ($<$10 μm), and conformal. LAE can also be used to create thin-film transistors (TFTs) for providing circuit functionality. We base our system on amorphous silicon (a-Si) TFTs, since industrially this is the most widely used TFT technology for fabricating backplanes within flat panel displays [4]. However, low-temperature processing results in a-Si TFT performance that is substantially worse than that of silicon CMOS transistors available in VLSI technologies. For example, n-channel a-Si TFTs have electron mobility of $\mu_e \sim 1\ \mathrm{cm}^2/\mathrm{Vs}$ and unity-gain cutoff frequency of $f_T \sim 1$ MHz [5], while CMOS has corresponding values of $\mu_e \sim 500\ \mathrm{cm}^2/\mathrm{Vs}$ and $f_T \sim 300$ GHz. Alternate technologies emerging for TFTs (e.g., metal oxides) offer higher levels of performance; however, their performance remains much lower than silicon CMOS [6].

Thus, to enable a high level of circuit functionality alongside the sensing capabilities, we adopt a hybrid system architecture [6], which combines LAE and CMOS ICs. In the LAE domain, we create distributed microphone channels, comprising thin-film piezoelectric microphones and localized TFT amplifiers, as well as TFT scanning circuits for sequentially sampling the microphone channels, so as to reduce the number of analog interface wires to the CMOS IC. In the CMOS domain, we perform audio signal readout, sampling control, and ultimately signal processing using a source reconstruction algorithm we propose.

This paper is organized as follows. Section II describes system-level design considerations, including motivation for the array toward overcoming nonidealities in the thin-film microphones and algorithmic approaches for overcoming sampling rate limitations imposed by the TFT circuits. Section III focuses on the design and implementation details of the system,

Manuscript received September 03, 2015; revised October 26, 2015; accepted November 05, 2015. Date of publication December 21, 2015; date of current version March 29, 2016. This paper was approved by Guest Editor Masato Motomura. This work was supported in part by the Qualcomm Innovation Fellowship, and in part by the National Science Foundation (NSF) under Grant ECCS-1202168 and Grant CCF-1218206.

The authors are with Princeton University, Princeton, NJ 08544 USA (e-mail: jsanz@princeton.edu; liechaoh@princeton.edu; tmoy@princeton.edu; rieutort@princeton.edu; yingzheh@princeton.edu; wagner@princeton.edu; sturm@princeton.edu; nverma@princeton.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/JSSC.2015.2501426

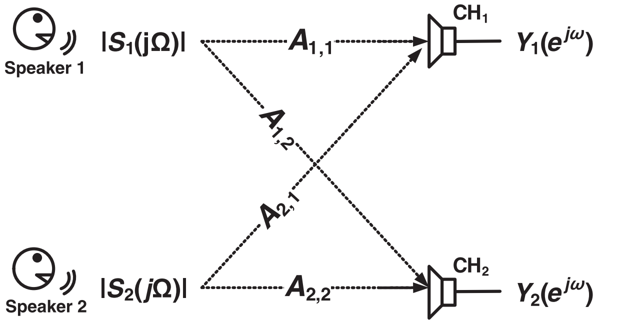


Fig. 1. System of equations for separating two simultaneous sources recorded with two microphones using previously measured transfer functions.

starting with the speech separation algorithm and then the LAE and CMOS circuit blocks. Section IV presents the prototype and its measured performance. Finally, Section V presents conclusion.

II. System Design Approach

The system focuses on separating two sound sources that are speaking simultaneously. This section first describes the challenges raised by practical microphones in a practical room, and then describes how these challenges can be overcome through the use of LAE. A widely used approach for source separation is to carry out time delay beamforming; however, this has the disadvantage of requiring a relatively large number of microphone channels [7], [8]. On the other hand, the problem can be approached from the perspective of a linear time invariant (LTI) system, where the propagation of sound between every speaker and every microphone is described by a linear transfer function. As shown in Fig. 1, the contributions from multiple sources received at a given microphone can thus be modeled as a convolutional mixture [9]. Restated in the frequency domain, the frequency components of the received signals $[Y_1(e^{j\omega}),\ Y_2(e^{j\omega})]$ can be related to the source signals $[S_1(e^{j\omega}),\ S_2(e^{j\omega})]$ by measuring the transfer functions $[A_{1,1}(e^{j\omega}),\ A_{2,1}(e^{j\omega}), A_{1,2}(e^{j\omega}),\ A_{2,2}(e^{j\omega})]$

$$\underbrace{\begin{bmatrix} Y_1(e^{j\omega}) \\ Y_2(e^{j\omega}) \end{bmatrix}}_{\text{microphone signals}} = \underbrace{\begin{bmatrix} A_{1,1}(e^{j\omega}) & A_{2,1}(e^{j\omega}) \\ A_{1,2}(e^{j\omega}) & A_{2,2}(e^{j\omega}) \end{bmatrix}}_{\text{transfer-function matrix}} \underbrace{\begin{bmatrix} S_1(e^{j\omega}) \\ S_2(e^{j\omega}) \end{bmatrix}}_{\text{source signals}} \tag{1}$$

Through this linear system of equations, the source signals can in principle be resolved using as few as two microphone channels.

However, in practice, the ability to resolve the source signals in this way is degraded by uncertainty in the transfer-function measurements, leading to severe dependence of the reconstruction quality on the precise location and response of the microphones. This is particularly relevant for thin-film microphones fabricated on a large flexible sheet. They experience substantial variations in their frequency response, due to the following reasons.

1) *Sound propagation:* In addition to $1/r$ pressure and amplitude attenuation, sound traveling in a room experiences reverberations and reflections due to the surfaces of the room. This can be simulated using the image method [10]. Fig. 2 shows how for a simulated room, this causes the transfer function for spatially distributed microphones to vary greatly, even when using perfectly uniform microphones and loudspeakers as sources.
2) *Intrinsic microphone variations:* During fabrication and deployment, important microphone parameters, such as membrane tension and air volume, are subject to variation. Fig. 3 shows measured data from an anechoic chamber, of thin-film microphones fabricated to be nominally identical. As seen, the actual frequency response varies substantially in our experiments. Although refining fabrication methods can reduce this variation, experience with fabrication over large areas and on flexible substrates shows that significant variations are likely to remain.
3) *Microphone directionality*: The microphone structure employed in this work is shown in Fig. 4, consisting of a double clamped membrane composed of the piezopolymer material, polyvinylidene fluoride (PVDF). Standoffs mount the membrane approximately 1 mm from the large-area sheet. Sound acts on both faces of the membrane, leading to substantial directionality variation in the measured transfer function shown. The details of the PVDF microphone used in this work are given in Section III-B.

To characterize the effect of these variations, for separating two speech sources, we calculate the signal-to-interferer (SIR) ratio, as given by [11]

$$\text{SIR} = 10\ \log_{10}\left(\frac{\|S_{\text{Target}}(t)\|^2}{\|E_{\text{Interferer}}(t)\|^2}\right). \tag{2}$$

$S_{\text{Target}}(t)$ is the original sound source we wish to recover, while $E_{\text{Interferer}}(t)$ is the remaining component from the second source, which has not been fully removed by the separation algorithm. Fig. 5(a) shows a simulation in an ideal anechoic room, wherein room reverberations, microphone variations, and microphone directionality are not considered. The room parameters used for simulations throughout this paper are shown in Fig. 6. In this simulation, eight microphones are incorporated in a linear array with spacing of 15 cm (array width = 105 cm), but only two are selected for source separation using the approach in (1). Each of the eight-choose-two microphone permutations (56 possible pairs) are examined. A 10 s speech segment is used as the sound emitted by each simulated source. Each segment consists of three sentences from male A and female B speaker from the TSP speech database [12]. It is processed by concatenating 100 ms windows, as outlined in Section III-A2. The results show that nearly uniform SIR improvement (24 dB, relative to the unprocessed input signal) is achieved regardless of the two microphones selected. On the other hand, Fig. 5(b) shows a simulation considering practical levels of room reverberations, microphone variations, and microphone directionality. In this case, the SIR improvement varies greatly (from 6 to 20 dB) due to the intrinsic and positional variations of the microphones. To mitigate this variation, we propose an approach that takes advantage of LAE in two ways.

1) By having multiple spatially distributed microphones, we can select a subarray that is closest to the two speakers, as illustrated in Fig. 7. This allows us to receive the

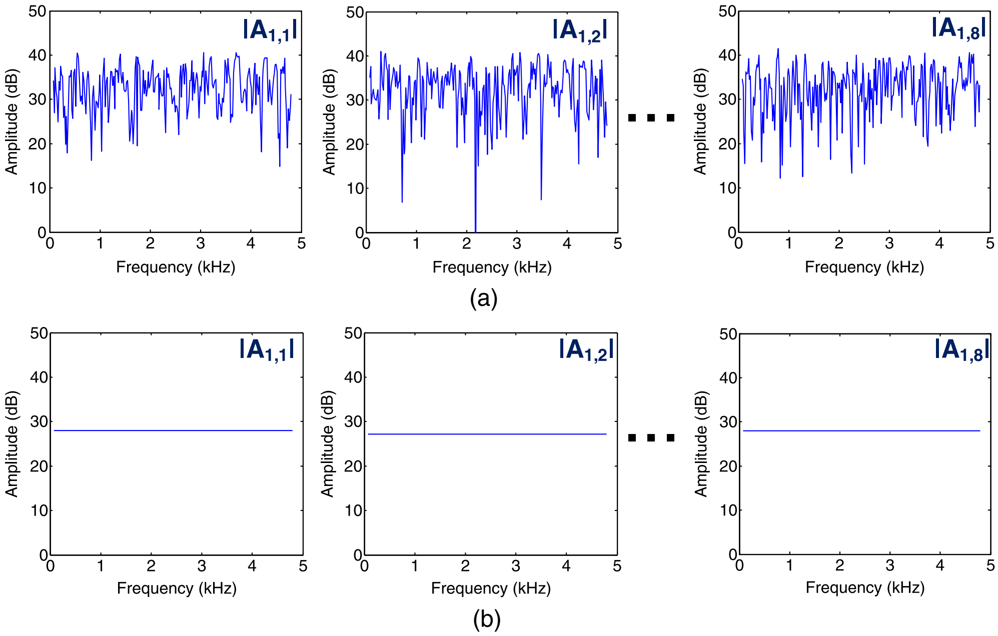


Fig. 2. Simulated frequency response of perfectly uniform, omnidirectional microphones, and speakers in (a) reverberant room and (b) nonreverberant room.

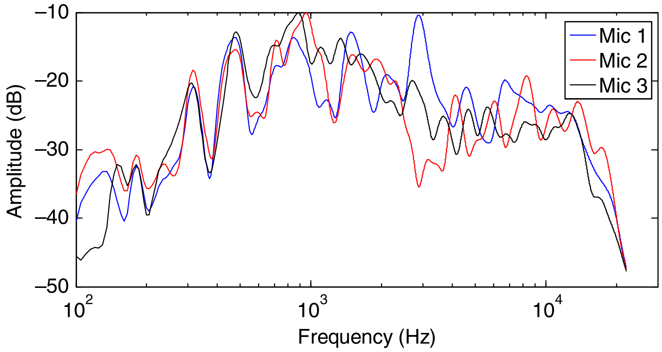


Fig. 3. Frequency response of piezopolymer, PVDF, microphones measured in an anechoic chamber at an angle of 0° (directly facing the source).

highest SNR signal, enabling higher quality microphone recordings and improved transfer-function estimates.

2) Each subarray is composed of eight microphone channels. Section III-A2 describes the algorithm that carries out signal separation using the microphone inputs from the subarray. This approach enhances robustness to the microphone variations (as quantified below).

However, using multiple subarrays each with eight microphones raises the problem that a large number of interfaces would be needed between the LAE and CMOS domain. This is costly and limits the scalability of the system. To address this, the eight channels from each subarray are sequentially sampled using a TFT scanning circuit. With this configuration, as shown in Fig. 8, we reduce the number of interfaces between LAE and CMOS.

One of the challenges of sampling in the LAE domain is that, using a-Si TFT scanning circuits, the scanning frequency is limited to 20 kHz (described further in Section III-D). This means that each channel can no longer be sampled at the Nyquist rate. Instead each channel of the subarray is critically sampled. Namely, over the eight-channel subarray, each channel is sampled at 2.5 kHz; since for high intelligibility we can bandpass filter human speech between 300 Hz and 5 kHz [13], this results in four aliases from each source, giving a total of eight aliases for the two sources. Section III-A2 describes the algorithm for separating these aliases using signals acquired from the eight microphone channels. Fig. 9 illustrates the benefit, comparing the simulated performance of the critically sampled system with eight microphones, to the best, median, and worst performance from 8-choose-2 microphone combinations shown in Fig. 5(b). As seen, the proposed critically sampled system (with eight microphones) performs at the same level as the median combination (with two microphones). The precise performance ultimately required depends on the further processing needed in specific applications (e.g., speech recognition), and is the subject of on-going investigation as various applications are being explored. However, what we see of critical importance is that, exploiting the ability to form a spatially distributed array with several microphone channels, the proposed approach overcomes the severe sensitivity to microphone placement that is experienced when using just two microphones, which would otherwise limit performance in a practical room with practical microphones.

III. System Design Details

Fig. 8 shows the eight-channel subarray hybrid system, which combines LAE and CMOS [14]. In the LAE domain there are eight microphone channels, each consisting of a PVDF microphone and a localized amplifier based on a-Si TFTs. The first of eight channels directly feeds the CMOS IC, forming a dedicated analog interface, required as described below for calibration. The remaining seven channels are connected to a large-area scanning circuit, which sequentially samples the channels in an interleaved manner; thus reduced to a single additional

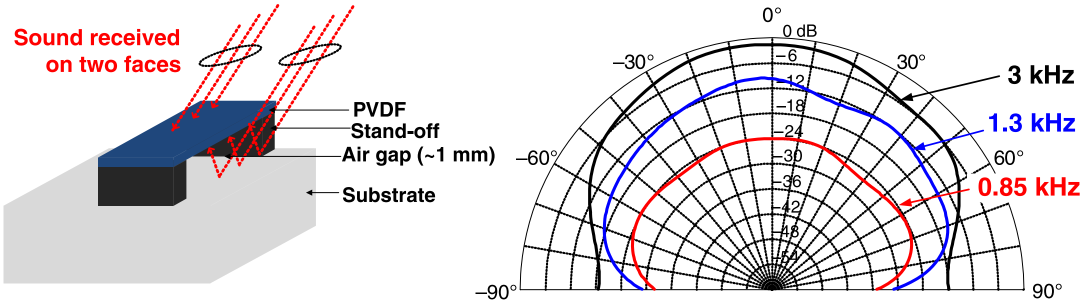


Fig. 4. Polar diagram measured in an anechoic chamber of a PVDF microphone.

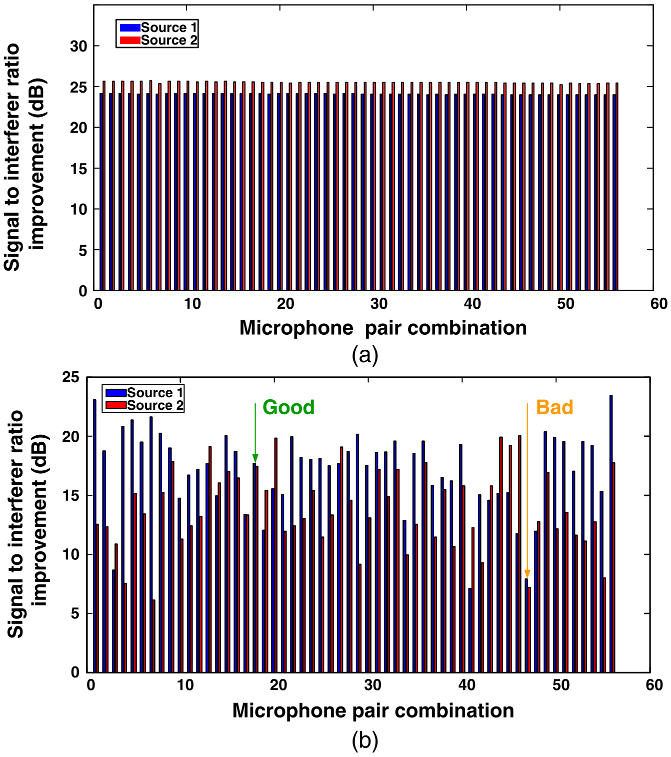


Fig. 5. Reconstruction results for 8-choose-2 pairs of microphones. Simulated in a room (a) without reverberations, directionality, or microphone process variation and (b) with reverberations and directional microphones.

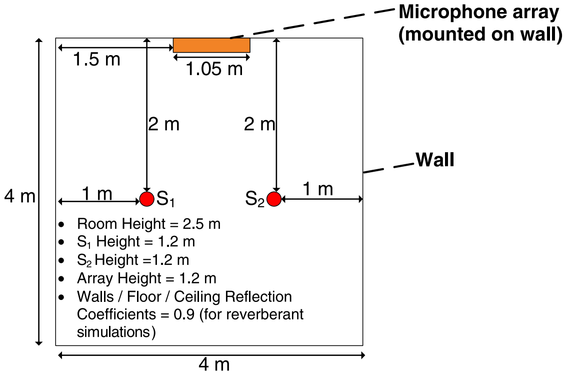

Microphone array (mounted on wall)
1.5 m
1.05 m
2 m
2 m
Wall
4 m
1 m
S1
S2
1 m
Room Height = 2.5 m
S1 Height = 1.2 m
S2 Height =1.2 m
Array Height = 1.2 m
Walls / Floor / Ceiling Reflection Coefficients = 0.9 (for reverberant simulations)
4 m

Fig. 6. Simulation parameters for two simultaneous sound sources in a reverberant room.

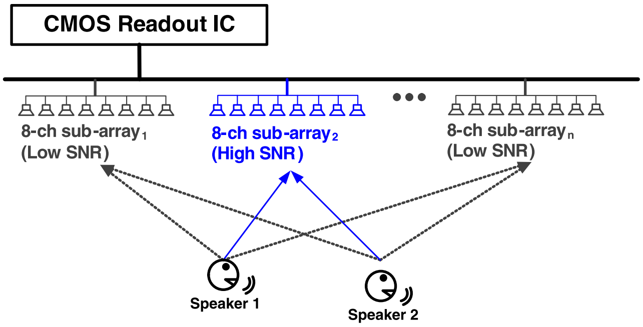

CMOS Readout IC
8-ch sub-array1 (Low SNR)
8-ch sub-array2 (High SNR)
8-ch sub-arrayn (Low SNR)
Speaker 1
Speaker 2

Fig. 7. Proposed structure of the microphone array composed of high SNR subarrays in close proximity to the speaker.

analog interface to CMOS. The CMOS IC includes digital control to multiplex between the two interfaces, to achieve critical sampling over the entire eight-channel subarray. The CMOS IC is primarily used for audio signal readout and digitization. After digitization, the critically sampled signal, consisting of the interleaved samples from the eight microphones, each effectively sampled at 2.5 kHz, is fed to an algorithm for speech separation (currently off-chip).

A. *Speech Separation Algorithm*

The algorithm is divided into two steps. The first step consists of calibration, which involves measuring the transfer functions between each source and each microphone. The second step is reconstruction, which uses the previously measured transfer functions to solve a system of equations and, thus, separate the two speech sources. Our algorithm differs from prior work [9] by enabling us to critically sample our microphone channels. We also expand upon previous work on critically sampled microphone arrays, which assume only a single source is present, so they do not support source separation [15].

1) Calibration: It is used to measure the values of the transfer functions at every frequency component required for reconstruction. This measurement is carried out using a calibration signal, which has spectral content that covers all frequencies of interests. In a practical application, this signal can be obtained by prompting users to speak one-by-one in isolation. For the frequency band of interest measurements of each transfer function can be done with a $\sim$ 100 ms window, since this a suitable window length for estimating the transfer function when using speech [16]. A 7 s speech signal was recorded for calibration. This corresponds to 1 s for each of the seven channels, giving ample signal to identify a 100 ms window having high SNR

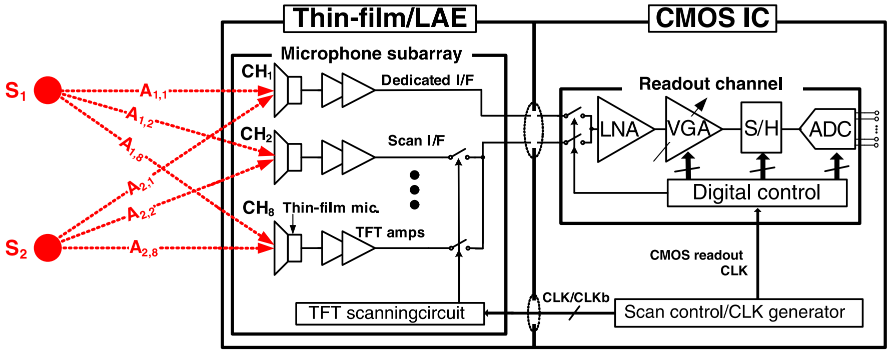


Fig. 8. System architecture, combining CMOS ICs and LAE.

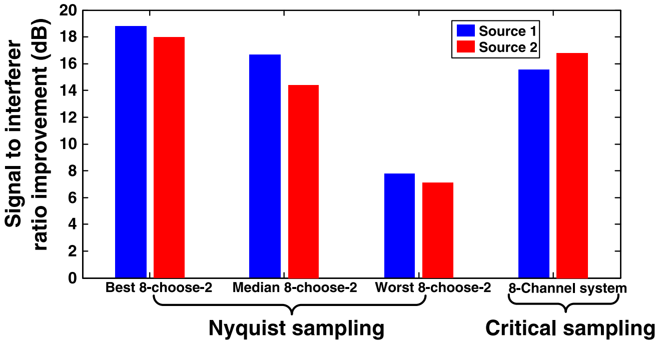


Fig. 9. Simulated reconstruction results for the best, median, and worst 8-choose-2 pairs of microphones, and for the 8-channel critically sampled subarray.

for estimating the transfer function from speech. Additionally, the absolute transfer function with respect to each source is not required; this would be problematic to measure since it would require recording at exactly the location of each source, in order to de-embed the effect of sound propagation in the room. Instead, each transfer function can be measured with respect to a designated reference channel within the array.

When characterizing the transfer functions, Nyquist sampling of the microphones is necessary (so that reconstruction can later be performed for each frequency bin of the Nyquist-sampled source). Fig. 10 shows how this is achieved, along with raw Nyquist samples from three representative channels. The system employs two analog interfaces from LAE to CMOS for each subarray. The reference channel is provided continuously to the CMOS IC via a dedicated interface, while the remaining channels are selected and characterized one at a time. This enables Nyquist-sampled measurement of each channel, allowing each transfer function to be obtained with respect to that of the reference channel.

2) Reconstruction: Having measured the transfer functions, now two users can speak simultaneously while the eight channels are critically sampled at a total rate of 20 kHz (2.5 kHz/channel). The signal processing challenges to perform reconstruction following such critical sampling when there is just a single source and when the transfer-function to each microphone can be expressed as a simple time delay are explored in [15]. However, for the multispeaker microphone system, additional challenges are raised due to the simultaneous sources and due to the complex frequency dependencies of the transfer functions from each source to each microphone. The signal processing employed in this work to overcome these is described as follows. As illustrated in Fig. 11, considering speech limited to a frequency of 5000 Hz, for every frequency bin of reconstruction, this leads to four aliases from each of the two sources. For each frequency bin, the eight unknowns can thus be resolved using the system of equations shown in (3) at the bottom of the page.

Using this approach, the total sampling rate required scales with the number of sources, rather than the number of microphones. For example, when reconstructing $N = 2$ simultaneous sources, assumed to have bandwidth of $\mathrm{BW} = 2 \times 5$ kHz (double sideband), interleaved sampling is carried out over all $(N \times \mathrm{BW})/K = 2.5$ kHz, which means the signals $Y_{1\ldots8}$ are effectively sampled below the Nyquist rate by a factor of $K/N = 4$, where K is the number of microphones and N is the number of sources. This is important because it overcomes significant variations in the reconstruction quality by increasing the diversity in spatial position and response of the microphones (as shown in Fig. 7), while limiting the required sampling rate to a level that can be achieved by the TFT scanning circuit.

To implement this algorithm, a frame is taken consisting of a total of 2048 samples (102 ms) sampled at 20 kHz in an interleaved manner from the eight channels. Next the individual

$$\underset{\text{microphone signals}}{\underset{(M = K/N = 4)}{\begin{bmatrix} Y_1(e^{j\omega}) \\ \vdots \\ Y_K(e^{j\omega}) \end{bmatrix}}} = \underset{\text{transfer-function}\quad\text{matrix}}{\begin{bmatrix} A_{1,1}(e^{j(\omega/M)}) & \cdots & A_{2,1}(e^{j(\omega/M-2\pi(M-1)/M)}) \\ \vdots & & \vdots \\ A_{1,K}(e^{j(\omega/M)}) & \cdots & A_{2,K}(e^{j(\omega/M-2\pi(M-1)/M)}) \end{bmatrix}} \underset{\text{source signals.}}{\begin{bmatrix} S_1(e^{j(\omega/M)}) \\ \cdots \\ S_1(e^{j(\omega/M-2\pi(M-1)/M)}) S_2(e^{j(\omega/M)}) \\ \cdots \\ S_2(e^{j(\omega/M-2\pi(M-1)/M)}) \end{bmatrix}} \tag{3}$$

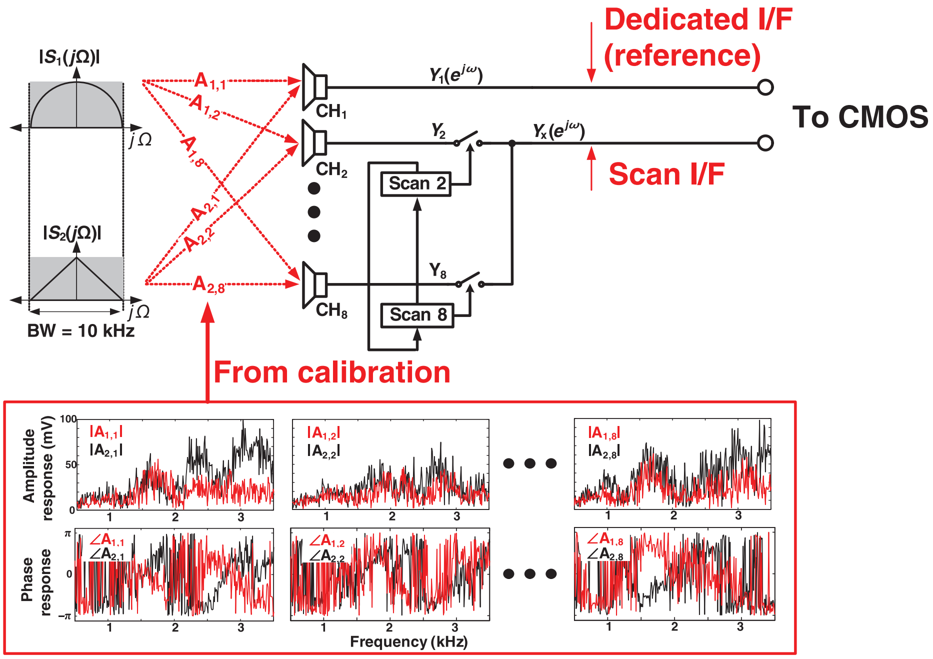


Fig. 10. Calibration procedure used to find the transfer functions between each source and microphone.

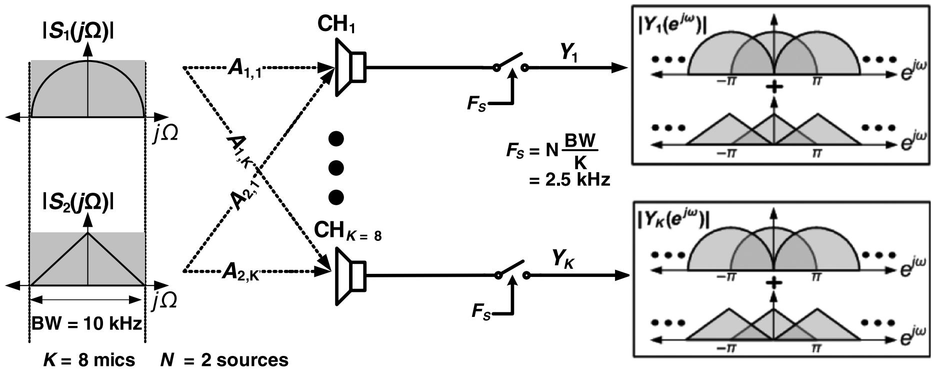


Fig. 11. Algorithm for separating and reconstructing two acoustic sources from under-sampled microphones in a subarray using previously calibrated transfer functions.

time samples corresponding to each channel are extracted, resulting in eight undersampled frames (one per microphone) containing 128 samples at 2.5 kHz. Then, an FFT is applied to each frame to derive the discrete Fourier transform (DFT) components. For each frequency sample of the DFT, the system of equations shown in (3) can now be setup and solved, so as to obtain the four aliased frequency components for each source. Then, using a modulated filter bank formulation, as outlined in [15], the four components can be used to reconstruct the DFT samples of the source signal sampled at 10 kHz (i.e., the Nyquist rate).

Having done this over multiple frames, the time-domain samples of the source signals can be obtained by taking an inverse Fourier transform. To process a long audio signal, the sequential frames are concatenated using the standard overlap-sum technique [17]. Each frame is overlapped by 75% with the preceding frame, so as to ensure it meets the constant overlap-add condition for the Hann windows used to mitigate artifacts [18].

B. *Thin-Film Piezoelectric Microphone*

Fig. 12 shows the microphone, which is based on a diaphragm formed from 1.5 cm (width) × 1.0 cm (length) PVDF, a piezoelectric thin-film polymer. The PVDF is 28 μm thick and is clamped using adhesive (cyanoacrylate glue) on both ends, with a tension of $\sim$0.2 N. It is clamped to acrylic posts, which stand off 1 mm from the sheet. This form factor enables the microphone to be used in a flexible, on-sheet application. To leverage the inherent translucency of the PVDF film, transparent electrodes with a sheet resistance of $\sim 8\ \Omega/\text{sq}$ are applied to both faces of the film by spray-coating silver nanowires [19], resulting in a clear, unobtrusive microphone.

The structure we developed functions primarily in d_{31} mode, where it converts horizontal strain into a vertical potential difference between the electrodes. As shown in Fig. 12, the measured sensitivity versus frequency has numerous resonant peaks arising from the double-clamped structure. We have tuned the tension and dimensions of the PVDF diaphragm to design the resonant peaks to match human speech, which is

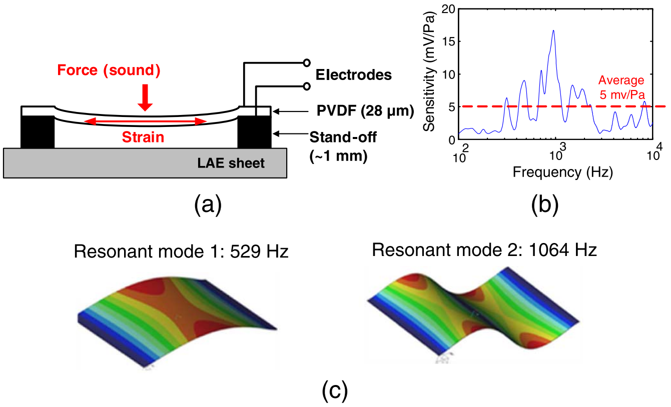


Fig. 12. Thin-film PVDF microphone design, including (a) structure, (b) frequency response (measured in an anechoic chamber), and (c) finite element simulations showing the resonant modes.

concentrated from 500 to 3000 Hz [13]. The sensitivity plot shown is for typical speech at a distance of 2 m. In this case, the average sensitivity of 5 mV/Pa yields a microphone signal of ~40 µV.

C. TFT Amplifiers

In addition to a PVDF microphone, each channel has its own localized two-stage differential amplifier, formed from a-Si TFTs [20] with $W/L = 3600$ µm/6 µm, as shown in Fig. 13. The first stage is a gain stage (with gain of 17 dB), whereas the second is a buffer stage (with gain of 3 dB) to drive long (~1 m) LAE interconnects. The overall amplifier chain has gain of 20 dB, with a passband from 300 Hz to 3 kHz and CMRR of 50 dB at 100 Hz (all measured). For experimentation and testability, we have used surface mounted passives; however, previously we have shown how thin-film resistors and capacitors can be monolithically integrated with the a-Si TFTs without having to modify the process flow [21].

The small amplitudes and low frequencies of the microphone signals raise an important noise tradeoff. Namely, the TFT amplifiers provide gain, which increases the immunity to stray noise coupling, which the long LAE interconnects are susceptible to (e.g., 60 Hz); but they also introduce intrinsic noise themselves. Fig. 13(b) shows the input referred noise power spectral density (PSD) measured from a TFT amplifier. In the frequency band of interest, the dominant noise is $1/f$ noise. To analyze the noise tradeoff, common-mode noise at 60 Hz is intentionally coupled to the differential LAE interconnects preceding the CMOS IC [through the bias node V_{B3}, see Fig. 13(a)]. Fig. 13(c) plots the noise of a channel, measured following digitized readout by the CMOS IC, but referred back to the passive PVDF microphone. Two cases are considered: 1) a case without localized TFT amplifier (i.e., microphone and CMOS readout IC only) and 2) a case with the localized TFT amplifier (i.e., microphone, TFT amplifier, and CMOS readout IC). As seen, with no stray noise coupling, the total input referred noise with the TFT amplifier is worse by 4× due to the intrinsic noise of the amplifier. However, when just 160 mV of stray coupling noise is applied, the localized TFT amplifier leads to lower input referred noise. It should be noted that when experimentally testing the system in a practical room, we typically experienced stray coupling noise significantly greater than 160 mV for certain channels. This shows the benefit of using localized TFT amplifiers fabricated over large areas to interface with the microphones.

D. TFT Scanning Circuit and LAE / CMOS Interfaces

For every subarray, there are two analog interfaces to CMOS, corresponding to the signals from the reference and scanned microphone channels. There is also a digital interface shared across all subarrays, corresponding to three signals from CMOS to LAE, required for controlling the large-area scanning circuits.

After the long LAE interconnects (~1 m), signals are provided to the CMOS IC through the TFT scanning circuit previously reported in [22]. The circuit is placed after the long interconnects to minimize the capacitance that must be driven due to the step response during scanning. The circuit is shown in Fig. 14(a), consisting of level converter blocks and scan blocks, based only on NMOS devices, since the extremely low mobility of holes in a standard a-Si TFT technology precludes the use of PMOS devices ($\mu_h < 0.1$ cm^2/Vs) [5]. The overall scanning circuit operates at 20 kHz from a 35 V supply. As shown, it takes two-phase control signals from the CMOS IC $\mathrm{CLK_{IC}}/\mathrm{CLKb_{IC}}$ to generate signals (EN $< i >$) to sequentially enable the microphone channels one at a time. In addition, a third reset signal is required to reset the whole system. Proper control of $\mathrm{CLK_{IC}}$/ $\mathrm{CLKb_{IC}}$ [as shown in Fig. 14(b)] enables readout from the seven channels, as well as multiplexing of the dedicated channel within the CMOS IC for readout over all eight channels.

The CMOS control signals are fed to the TFT level converter blocks, which convert 3.6 V CMOS levels to roughly 10 V. Scanning speed is limited by a critical time constant within the scan blocks, set by the load resistor R_L and the output capacitor C_{int}. R_L must be large enough so that the intermediate node X can be pulled down by the TFT. C_{int} needs to be large enough to drive the capacitance of subsequent TFTs. Thus, the resulting time constant is ultimately set by the TFTs, limiting the scanning speed to 20 kHz.

E. CMOS IC

The outputs of the scanning circuit are fed directly into the CMOS IC for readout. As shown in Fig. 15, the CMOS IC consists of a low-noise amplifier for signal acquisition, a variable-gain amplifier (VGA) to accommodate large variations in the audio signals, a sample-and-hold (S/H), and an ADC. In particular, the use of a VGA is critical within the large-area microphone system.

1) Low-Noise Amplifier: The LNA is implemented as a resistively loaded differential amplifier. In order to achieve the low-noise performance, a relatively large-sized input transistor (96 µm/12 µm) is employed to reduce the $1/f$ noise. Moreover, a large current (100 µA) is consumed to further reduce the noise floor. As a result, in simulation, the LNA is

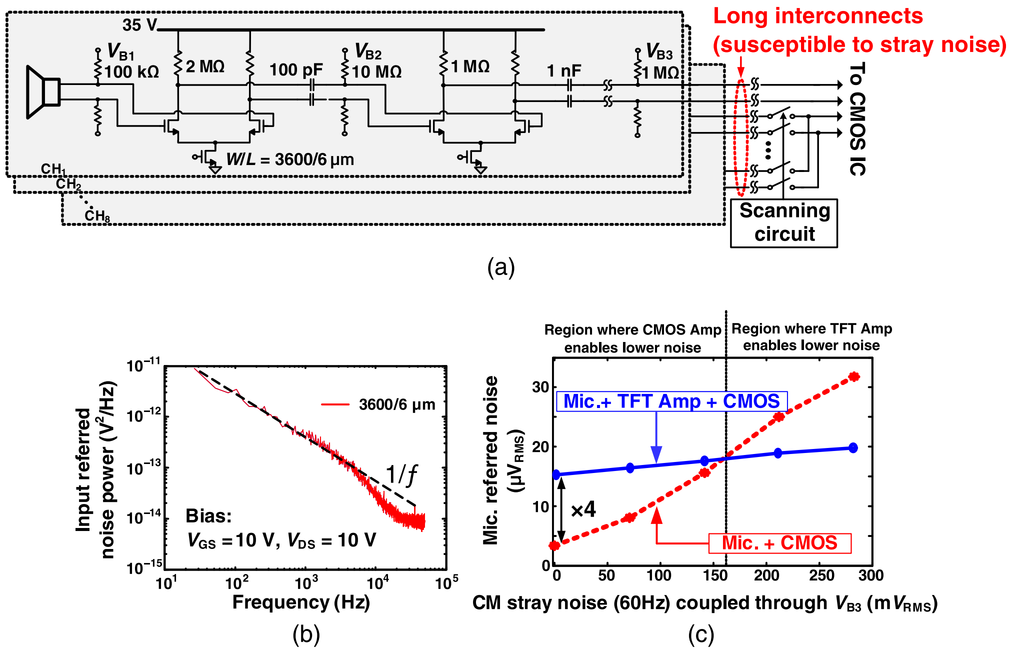

35 V
V_B1
100 kΩ
2 MΩ
100 pF
V_B2
10 MΩ
1 MΩ
1 nF
V_B3
1 MΩ
W/L = 3600/6 μm
CH1
CH2
CH8
Long interconnects (susceptible to stray noise)
To CMOS IC
Scanning circuit
(a)
Input referred noise power (V²/Hz)
3600/6 μm
1/f
Bias: V_GS = 10 V, V_DS = 10 V
Frequency (Hz)
(b)
Region where CMOS Amp enables lower noise
Region where TFT Amp enables lower noise
Mic.+ TFT Amp + CMOS
Mic. + CMOS
×4
Mic. referred noise (μV_RMS)
CM stray noise (60Hz) coupled through V_B3 (mV_RMS)
(c)

Fig. 13. (a) Schematic of a two stage TFT amplifier. (b) Measured noise characteristics of an a-Si TFT. (c) Tradeoff between a localized TFT amplifier and CMOS.

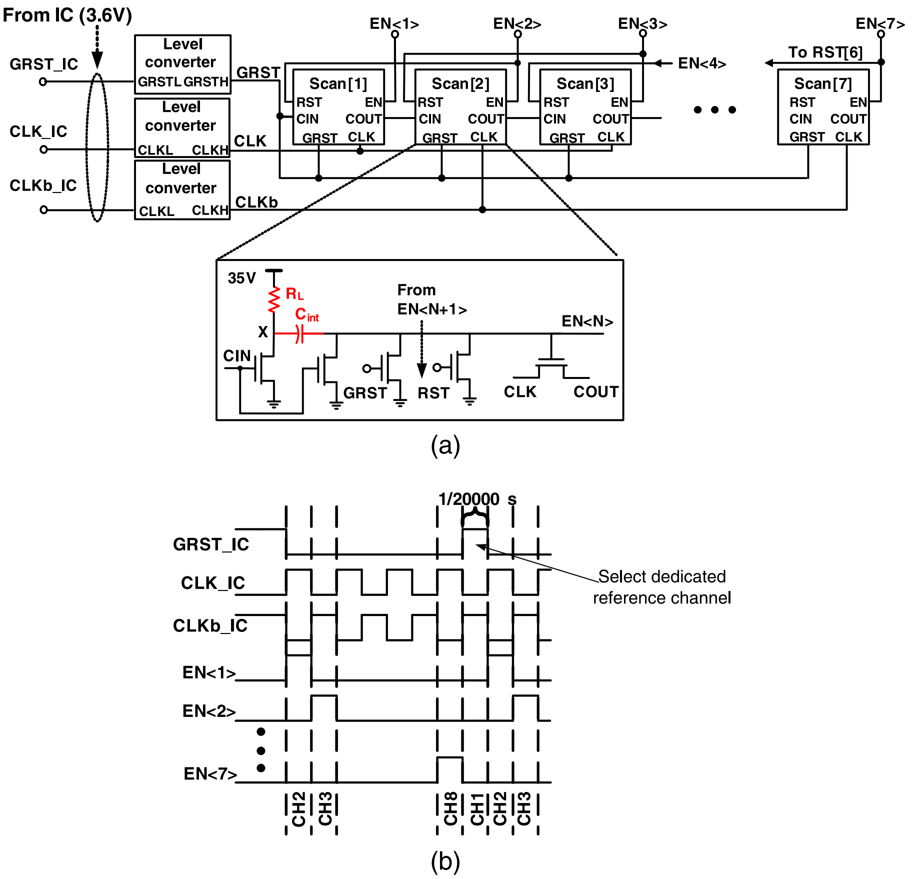


Fig. 14. TFT scanning circuit (a) schematic and (b) timing diagram [22].

designed to have a gain of 16 dB with 2.6 μV_{RMS} integrated noise and 100 Hz $1/f$ corner. As shown in Section IV, the simulation matches the measured results.

2) Variable-Gain Amplifier: The VGA is important because the microphone variability and variations in speaker distance from the microphones means that the received

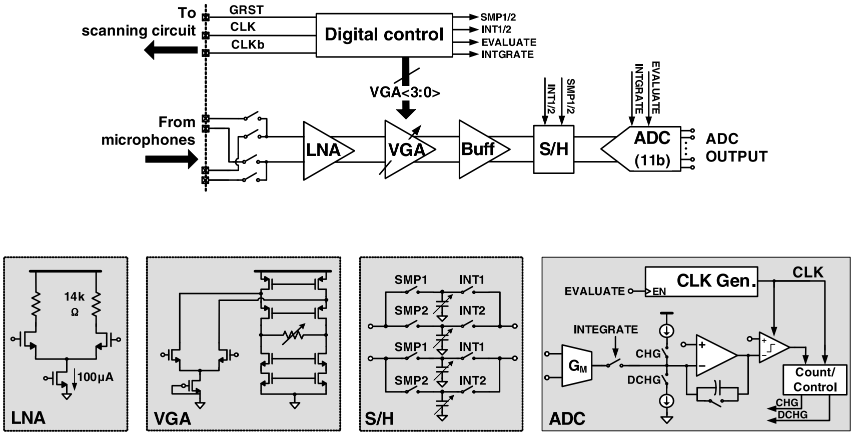


Fig. 15. Schematics of the CMOS IC used for readout and digitization, which incorporates an LNA, VGA, and 11 bit ADC.

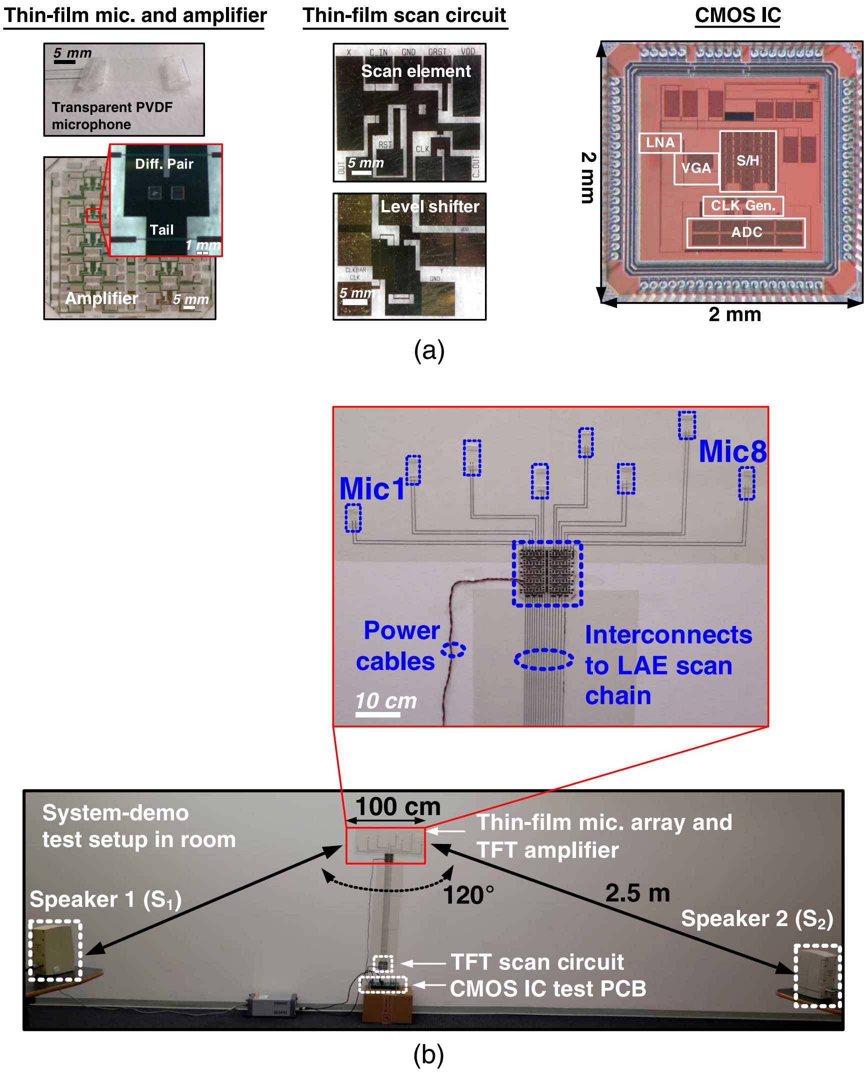


Fig. 16. System Prototype. (a) Micrograph of components: microphone channel (PVDF microphone and a-Si TFT amplifier), a-Si scanning circuit, and CMOS readout IC. (b) Testing setup in classroom for full system demonstration with two simultaneous sources. A microphone array spanning 105 cm is at a radial distance of 2.5 m from two speakers separated by an angle of 120°.

signals can have largely varying amplitude. The VGA thus addresses the dynamic range that would otherwise be required in the readout circuit. The actual gain setting for the VGA is determined for each microphone during the transfer-function calibration described in Section III-A1.

The VGA is implemented as a folded-cascode structure to maximize its output dynamic range over a large span of gain settings within one stage. Gain programmability is achieved via a configurable output resistor, implemented as a 4 bit resistor DAC. The gain provided ranges from 6 to 27 dB (measured).

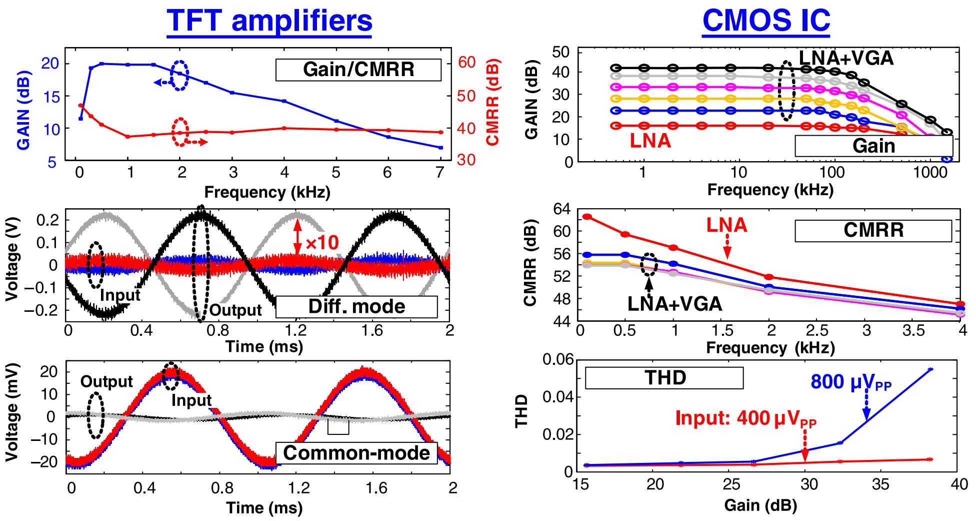


Fig. 17. Component-level measurement.

TABLE I
PERFORMANCE SUMMARY OF THE SYSTEM

Thin-film microphone (PVDF)					
Area		1.5 X 1 cm			
Sensitivity		~5 mV/Pa (1 to 3 kHz)			
Thin-film circuitry (a-Si on glass @ 180℃)		CMOS IC (IBM 0.13 μm)			
Amplifier chain		Power	Scan control	0.08 mW@ 3.6 V	0.62 mW
Power	3.5 mW @ 35 V		Readout	0.54 mW@ 1.2 V	
Gain	20 dB	Gain		16to 43 dB	
Pass-band	0.3 to 3 kHz	Bandwidth		100 kHz	
CMRR (@100 Hz)	49 dB	CMRR	LNA	62 dB	
Input referred noise	$16\,\mu V_{rms}$	(@100 Hz)	LNA+VGA	54 dB	
Scan chain		THD	$400\mu V_{PP}$ input	0.5%	
Scan rate	20 kHz	(Gain: 33 dB)	$800\mu V_{PP}$ input	1.5%	
Power	12 mW @ 35 V	Input referred noise		$4\mu V_{rms}$	

3) Sample-and-Hold and ADC: The S/H is differential and consists of two interleaved samplers. This allows maximal time for step-function transients to settle during scanning of the microphone channels and configuration of the VGA. Furthermore, the hold capacitors are configurable, implemented as 4 bit capacitor DACs. This, along with the VGA, allows the time constant to adapt if increased scanning rates are desired (which would be required to experiment with a number of sources N more than 2), while minimizing in-band noise.

A buffer stage is inserted between VGA and S/H to decouple the VGAs resistive load from the S/H's capacitor, both of which are relatively large and varying. Considering that the input for the buffer is already a relatively large signal after being amplified by the LNA and VGA, the buffer is implemented as a common source amplifier with source degeneration to keep the linearity of the whole system while providing another 7 dB gain.

Following the S/H is an integrating ADC, which digitizes the sample to 11b. A transconductance stage (G_M) generates a current signal, and a low-speed integrating op-amp circuit with switchable input current sources generates the dual slopes required for data conversion via a digital counter. The integrating opamp is implemented as a two-stage op-amp with dominant pole compensation for stability.

IV. PROTOTYPE MEASUREMENTS AND SYSTEM DEMO

Fig. 16 shows the prototype of the whole system, including LAE components and CMOS IC. The PVDF thin-film microphones, and the a-Si TFT amplifiers and scanning circuits deposited at 180 °C on a glass substrate, were all produced in-house. Within a large-area system, these can be thought of as comprising the front plane and back plane, respectively [23], and various methods of fabricating front planes consisting of microphone arrays have been considered [24]. The CMOS IC was implemented in a 130 nm technology from IBM. The microphone subarray spanned a width of 105 cm, and consisted of eight PVDF microphones, linearly spaced by 15 cm. These technologies were selected due to their proved robustness; however, the hybrid LAE-CMOS architecture presented can be readily adapted to different CMOS process nodes and TFT transistor technologies.

Table I provides a measurement summary of all the system components. On the LAE side, each local amplifier channel consumes 3.5 mW and the scanning circuit for each subarray consumes 12 mW. The CMOS readout IC consumes 0.6 mW in total. Fig. 17 shows details from characterization of the TFT amplifier (left) and the CMOS readout circuit (right). The bandwidth of the TFT amplifier is tuned to match human

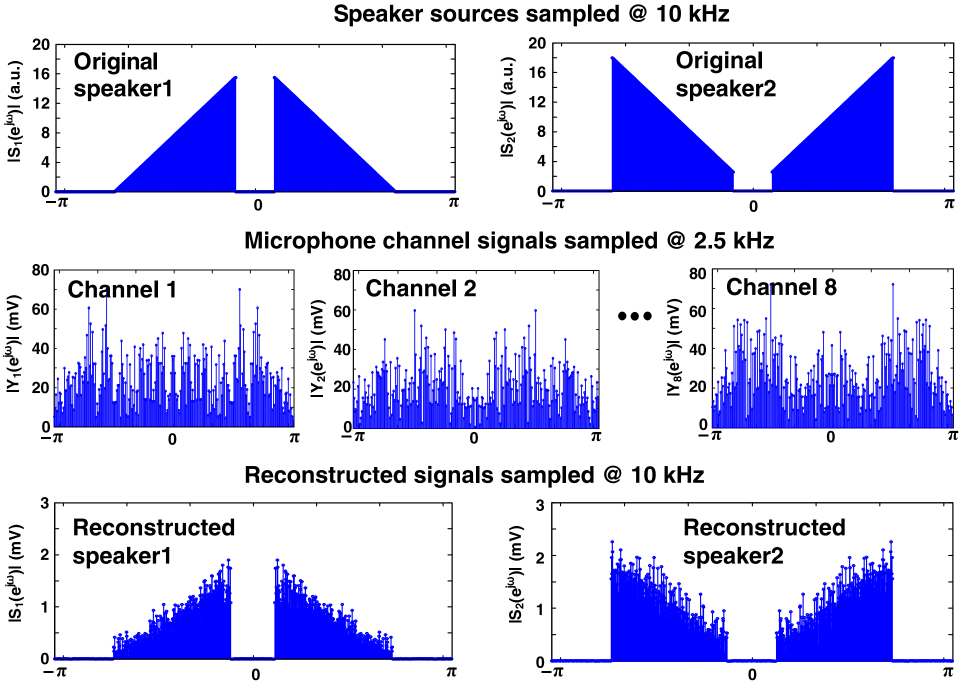


Fig. 18. Demonstration of two-source separation and reconstruction for two simultaneous wedge-shaped inputs.

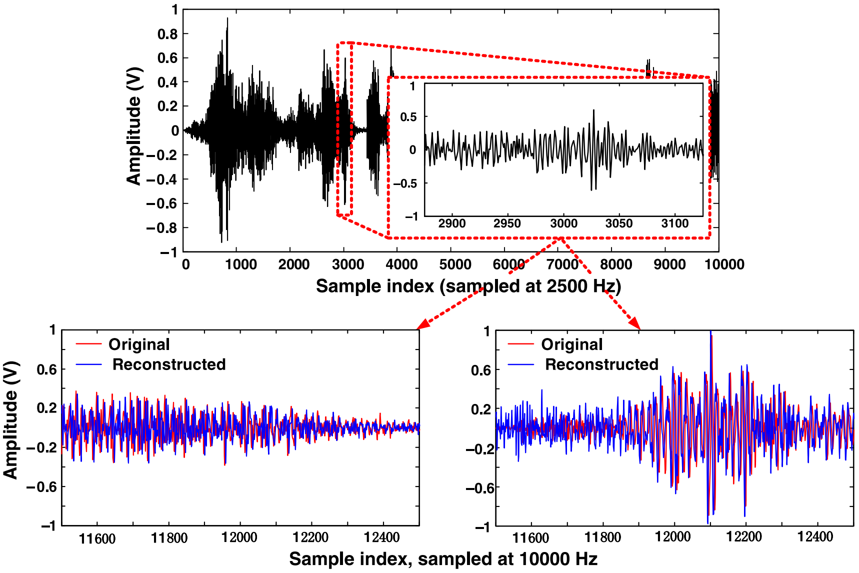


Fig. 19. Demonstration of two-source separation and reconstruction for two simultaneous speech signals.

speech, and filter out-of-band noise. Its CMRR of 49 dB is limited by the mismatch of the TFTs. Nevertheless, as shown in the waveforms, it substantially suppresses stray common-mode noise. The CMOS readout circuit successfully achieves programmable gain from 16 to 43 dB overall. The CMRR and linearity measurements are also shown.

For demonstration, the whole system was tested in a 5 m × 6 m classroom. The testing setup is shown in Fig. 16(b). Two speakers separated by an angle of 120° were placed at a radial distance of 2 m from the center of the microphone array. Calibration was performed using a white-noise signal from 0.5 to 3.5 kHz, which was played one-by-one through each speaker for 7 s to measure the transfer functions. Following the calibration, we played two synthesized source signals S_1 and S_2 simultaneously through the two speakers with a sound pressure level of ~50 dB_{SPL}. Fig. 18 shows the source signals, received signals, and the signals separated by the system. As shown, the two sources were sampled at 10 kHz and intentionally synthesized to have DFTs with distinct wedge-shaped magnitudes. The DFTs of the signals received by three microphone channels (Y_1, Y_2, Y_8) sampled at 2.5 kHz exhibit source superposition and aliasing. Despite this, the reconstruction algorithm, using the acquired 2.5 kHz signals, successfully recovers the wedge-shaped magnitudes at 10 kHz with a signal-to-interferer ratio improvement of 12 dB. To further demonstrate the system, we also played two simultaneous speeches though the two speakers. Fig. 19 shows the time-domain waveforms of the signal received by the first microphone channel and those separated by the system (with the original signal waveforms overlayed). As seen, the two signals are successfully separated at the output, corresponding to a signal-to-interferer improvement of 11 dB. This makes the signal recording clear and intelligible, and prepares them for further speech-recognition processing within applications. The small residual difference between the original and reconstructed signals is primarily due to inaccuracies in transfer-function calibration, leading to errors in the reconstruction of certain signal frequency components.

V. Conclusion

Multispeaker voice separation will enable collaborative control of ambient electronic devices. This paper addresses this application by proposing a hybrid system for speech separation, which is based on combining LAE and a CMOS IC. In this paper, we: 1) develop an LAE microphone array, based on PVDF microphones and a-Si TFT instrumention, which we integrate with a CMOS IC for audio readout; 2) develop an algorithm for source separation, which overcomes the large variability of the PVDF microphones and the sampling rate limitations of the TFT circuits; and 3) demonstrate an eight-channel subarray system, spanning the entire signal chain from the transducer to digitization, which successfully separates two simultaneous audio sources.

Acknowledgment

The authors would like to thank MOSIS for IC fabrication.

References

[1] L. Zhou, S. Jung, E. Brandon, and T. N. Jackson, "Flexible substrate micro-crystalline silicon and gated amorphous silicon strain sensors," *IEEE Trans. Electron Devices*, vol. 53, no. 2, pp. 380–385, Feb. 2006.

[2] H. Wang, L. Chen, J. Wang, Q. Sun, and Y. Zhao, "A micro oxygen sensor based on a nano sol-gel TiO_2 thin film," *Sensors*, vol. 14, no. 9, pp. 16 423–16 433, Sep. 2014.

[3] C. Dagdeviren *et al.*, "Conformable amplified lead zirconate titanate sensors with enhanced piezoelectric response for cutaneous pressure monitoring," *Nat. Commun.*, vol. 5, pp. 1–10, Aug. 2014.

[4] Y. Kuo, "Thin film transistor technology—Past, present, and future," *Electrochem. Soc. Interface*, vol. 22, no. 1, pp. 55–61, 2013.

[5] R. Street, *Hydrogenated Amorphous Silicon*. Cambridge, U.K.: Cambridge Univ. Press, 1991, pp. 237–243.

[6] N. Verma *et al.*, "Enabling scalable hybrid systems: Architectures for exploiting large-area electronics in applications," *Proc. IEEE*, vol. 103, no. 4, pp. 690–712, Apr. 2015.

[7] E. Weinstein, E. Steele, K. Agarwal, and J. Glass, "A 1020-node modular microphone array and beamformer for intelligent computing spaces," MIT/LCS Technical Memo, MIT, Cambridge, MA, USA, Tech. Rep. MIT-LCS-TM-642, 2004.

[8] H. V. Trees, *Optimum Array Processing: Part IV, Detection, Estimation, and Modulation Theory.* Hoboken, NJ, USA: Wiley, 2002, p. 66.

[9] K. Kokkinakis and P. Loizou, *Advances in Modern Blind Signal Separation Algorithms: Theory and Applications*. San Rafael, CA, USA: Morgan & Claypool, 2010, pp. 13–19.

[10] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[11] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[12] P. Kabal, "TSP speech database," Dept. Electr. Comput. Eng., McGill Univ., Montreal, QC, Canada, Tech. Rep. 1.0, Sep. 2002.

[13] R. L. Freeman, *Fundamentals of Telecommunications*. Hoboken, NJ, USA: Wiley, 2005, pp. 90–91.

[14] L. Huang *et al.*, "Reconstruction of multiple-user voice commands using a hybrid system based on thin-film electronics and CMOS," in *Proc. Symp. VLSI Circuits (VLSIC)*, 2015, no. JFS4-4, pp. 198–199.

[15] P. Sommen and C. Janse, "On the relationship between uniform and recurrent nonuniform discrete-time sampling schemes," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5147–5156, Oct. 2008.

[16] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[17] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, Fl, USA: CRC Press, 2013, pp. 38–40.

[18] C. Roads, *The Computer Music Tutorial*. Cambridge, MA, USA: MIT Press, 1996, pp. 553–555.

[19] J. Spechler and C. Arnold, "Direct-write pulsed laser processed silver nanowire networks for transparent conducting electrodes," *Appl. Phys. A*, vol. 108, no. 1, pp. 25–28, Jul. 2012.

[20] H. Gleskova and S. Wagner, "Amorphous silicon thin-film transistors on compliant polyimide foil substrates," *Electron Device Lett.*, vol. 20, no. 9, pp. 473–475, 1999.

[21] L. Huang *et al.*, "Integrated all-silicon thin-film power electronics on flexible sheets for ubiquitous wireless charging stations based on solar-energy harvesting," in *Proc. Symp. VLSI Circuits (VLSIC)*, Jun. 2012, pp. 198–199.

[22] T. Moy *et al.*, "Thin-film circuits for scalable interfacing between large-area electronics and CMOS ICs," in *Proc. Device Res. Conf.*, Jun. 2014, pp. 271–272.

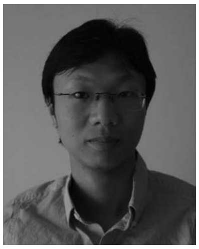
[23] W. Rieutort-Louis *et al.*, "Integrating and interfacing flexible electronics in hybrid large-area systems," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 5, no. 9, pp. 1219–1229, Sep. 2015.

[24] I. Graz *et al.*, "Flexible ferroelectret field-effect transistor for large-area sensor skins and microphones," *Appl. Phys. Lett.*, vol. 89, 073501, 2006.

Josue Sanz-Robinson received the B.Eng. degree (Hons.) from McGill University, Montreal, QC, Canada, in 2010, and the M.A. degree from Princeton University, Princeton, NJ, USA, in 2012. He is currently pursuing the Ph.D. degree at Princeton University, all in electrical engineering.

His research focuses on developing a platform for building hybrid sensing systems, which combine large-area electronics (LAE) and CMOS ICs. His research interests include large-area acoustic systems and microphone arrays to enable novel human–computer interfaces.

Mr. Sanz-Robinson was the recipient of a 2013 Qualcomm Innovation Fellowship.



Liechao Huang (S'12) received the B.S. degree in microelectronics from Fudan University, Shanghai, China, in 2010, and the M.A. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2012, where he is currently pursuing the Ph.D. degree.

His research interests include thin-film circuit design for power, radio, and sensing interfaces, CMOS analog and mixed signal design for sensing interfaces and power management, and hybrid system design combining thin-film circuits and CMOS ICs.

Mr. Huang was the recipient of the Princeton Engineering Fellowship and Gordon Wu Award at Princeton University.

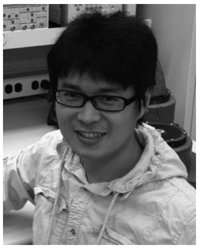
Tiffany Moy (S'14) received the B.S.E. (*magna cum laude*) and M.A. degrees from Princeton University, Princeton, NJ, USA, in 2012 and 2014, respectively. She is currently pursuing the Ph.D. degree at Princeton University, all in electrical engineering.

Her research interests include thin-film circuits and algorithms for hybrid large-area electronics/CMOS system design.

Warren Rieutort-Louis (S'12–M'15) received the B.A. (Hons.) and M.Eng. degrees in electrical and information engineering from Trinity College, Cambridge University, Cambridge, U.K., in 2009, the M.A. and the Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, USA, in 2012 and 2015, respectively.

He was a Graduate Teaching Fellow with Princeton McGraw Center for Teaching and Learning. His research interests include thin-film materials, processes, devices, and circuits for large-area electronic systems.

Dr. Rieutort-Louis was the recipient of the IBM Ph.D. Fellowship, the Andlinger Center Maeder Fellowship in Energy and the Environment, and the Princeton Harold W. Dodds Honorific Fellowship.



Yingzhe Hu received the B.S. degree in both physics and microelectronics from Peking University, Beijing, China, and the M.A. and Ph.D. degrees in electrical engineering from the Princeton University, Princeton, NJ, USA, in 2011 and 2015, respectively.

His research interests include flexible electronics and CMOS IC hybrid sensing system design and capacitive 3-D gesture sensing system design.

Mr. Hu was the recipient of 2013 Qualcomm Innovation Fellowship, Gordon Wu Award at Princeton University, 2013 ISSCC SRP Award, and 2013 VLSI Best Student Paper Award.

Sigurd Wagner (SM'80–F'00–LF'11) received the Ph.D. degree in physical chemistry from the University of Vienna, Vienna, Austria.

Following a Postdoctoral Fellowship at Ohio State University, Columbus, OH, USA, he was with the Bell Telephone Laboratories, Murray Hill, NJ, USA, from 1970 to 1978. He then joined the Solar Energy Research Institute (now NREL), Golden, CO, USA, as the Founding Chief of the Photovoltaic Research Branch. Since 1980, he has been a Professor of Electrical Engineering with Princeton University, Princeton, NJ, USA; in 2015, he became a Professor Emeritus and Senior Scholar. He is a Member of Princeton's Large-Area Systems Group.

Dr. Wagner is a Fellow of the American Physical Society and a Member of the Austrian Academy of Science. He was the recipient of the Nevill Mott Prize for his groundbreaking research, both fundamental and applied, on amorphous semiconductors and chalcopyrites, and the ITC Anniversary Prize for pioneering research on flexible and stretchable large-area electronics, and the comprehensive study of its mechanical behavior.

James C. Sturm (S'81–M'85–SM'95–F'01) was born in Berkeley Heights, NJ, USA, in 1957. He received the B.S.E. degree in electrical engineering and engineering physics from Princeton University, Princeton, NJ, USA, in 1979, and the M.S.E.E. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1981 and 1985, respectively.

In 1979, he joined Intel Corporation, Santa Clara, CA, USA, as a Microprocessor Design Engineer; in 1981, he was a Visiting Engineer at Siemens, Munich, Germany. In 1986, he was joined the Faculty of Princeton University, where he is currently the Stephen R. Forrest Professor of Electrical Engineering. From 1998 to 2015, he was the Director of the Princeton Photonics and Optoelectronic Materials Center (POEM) and its successor, the Princeton Institute for the Science and Technology of Materials (PRISM). From 1994 to 1995, he was a von Humboldt Fellow with the Institut fuer Halbleitertechnik, University of Stuttgart, Stuttgart, Germany. His research interests include silicon-based heterojunctions, thin-film and flexible electronics, photovoltaics, the nano-bio interface, three-dimensional (3-D) integration, and silicon-on-insulator.

Dr. Sturm was a National Science Foundation Presidential Young Investigator. He was the Technical Program Chair and General Chair of the IEEE Device Research Conference, in1996 and 1997, respectively. He served on the Organizing Committee of IEDM (1988 to 1992 and 1998 to 1999), having chaired both the solid-state device and detectors/sensors/displays committees. He has served on the boards of Directors of the Materials Research Society and the Device Research Conference, and co-founded Aegis Lightwave and SpaceTouch. He has won more than 10 awards for teaching excellence

Naveen Verma (M'09) received the B.A.Sc. degree in electrical and computer engineering from the University of British Columbia, Vancouver, BC, Canada, in 2003, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2005 and 2009, respectively.

Since July 2009, he has been with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA, where he is currently an Associate Professor. His research interests include advanced sensing systems, including low-voltage digital logic and SRAMs, low-noise analog instrumentation and data-conversion, large-area sensing systems based on flexible electronics, and low-energy algorithms for embedded inference, especially for medical applications.

Prof. Verma was the recipient or co-recipient of the 2006 DAC/ISSCC Student Design Contest Award, 2008 ISSCC Jack Kilby Paper Award, 2012 Alfred Rheinstein Junior Faculty Award, 2013 NSF CAREER Award, 2013 Intel Early Career Award, 2013 Walter C. Johnson Prize for Teaching Excellence, 2013 VLSI Symposium Best Student Paper Award, and 2014 AFOSR Young Investigator Award.