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

Josue Sanz-Robinson, Liechao Huang,Tiffany Moy, Warren Rieutort-Louis,
Yingzhe Hu, Sigurd Wagner, James C. Sturm, and Naveen Verma
Princeton University
Princeton University, Engineering Quadrangle
Olden Street, Princeton, NJ 08544
609-858-3750
{jsanz, liechaoh, tmoy, rieutort, yingzheh,wagner, sturm, nverma} @princeton.edu

August 2015

Abstract–This paper presents a system for reconstructing independent voice commands from two simultaneous speakers, based on an array formed from spatially-distributed microphones. It adopts a hybrid architecture, which combines large-area electronics (LAE), a technology well-suited for creating a physically expansive sensor array (> 1 m width) and a CMOS IC, which provides superior transistors for read-out and signal processing. We take advantage of the LAE array in two ways: (1) select microphones that are in closest proximity to the speakers to receive the highest SNR signal; (2) use multiple spatially-diverse microphones to enhance robustness to microphone variation. In the LAE domain each microphone channel consists of a thin-film transducer formed from PVDF, a piezoelectric polymer, and a localized amplifier composed of amorphous silicon (a-Si) thin-film transistors (TFTs). Each channel is sequentially sampled by an a-Si TFT scanning circuit, to reduce the number of interfaces between the 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 entire system with eight channels is demonstrated, acquiring and reconstructing two simultaneous audio signals at 2 m distance from the array with a signal-to-interferer ratio improvement of ∼12 dB.

1 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 in order 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 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 cm^2/Vs and unity-gain cutoff frequency of $f_T \sim$ 1 MHz, while CMOS has corresponding values of $\mu_e \sim$ 500 cm^2/Vs and $f_T \sim$ 300 GHz.

Thus, to enable a high-level of circuit functionality alongside the sensing capabilities, we adopt a hybrid system architecture [5], 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.

The paper is organized as follows. Section 2 describes system-level design considerations, including motivation for the array towards overcoming non-idealities in the thin-film microphones and algorithmic approaches for overcoming sampling rate limitations imposed by the TFT circuits. Section 3 focuses on the design and implementation details of the system, starting with the speech separation algorithm and then the LAE and CMOS circuit blocks. Section 4 presents the prototype and its measured performance. Finally, Section 5 presents conclusions.

2 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 [6] [7]. 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 Figure 1, the contributions from multiple sources received at a given microphone can thus be modelled as a convolutional mixture [8]. 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 $[A_{1,1}(e^{j\omega}),\ A_{2,1}(e^{j\omega}), A_{1,2}(e^{j\omega}),\ A_{2,2}(e^{j\omega})]$:

$$\underbrace{\begin{bmatrix} Y_1(e^{j\omega}) \\ Y_2(e^{j\omega}) \end{bmatrix}}_{Microphone\ Signals} = \underbrace{\begin{bmatrix} A_{1,1}(e^{j\omega}) & A_{2,1}(e^{j\omega}) \\ A_{1,2}(e^{j\omega}) & A_{2,2}(e^{j\omega}) \end{bmatrix}}_{Transfer-function\ Matrix} \underbrace{\begin{bmatrix} S_1(e^{j\omega}) \\ S_2(e^{j\omega}) \end{bmatrix}}_{Source\ Signals} \tag{1}$$

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. This is particularly relevant for thin-film microphones fabricated on a large flexible sheet. They experience substantial variations in their frequency re-

sponse, 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 [9]. Figure 3 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. **Microphone Variations:** During fabrication and deployment, important microphone parameters, such as membrane tension and air volume, are subject to variation. Figure 4 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 Figure 5, consisting of a double clamped membrane composed of the piezopolymer material, 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 3.2.

To characterize the effect of these variations, for separating two speech sources, we calculate the signal-to-interferer (SIR) ratio, as given by [10]:

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

$S_{Target}(t)$ is the original sound source we wish to recover, while $E_{Interferer}(t)$ is the remaining component from the second source, which has not been fully removed by the separation algorithm. Figure 6(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 Figure 2. In this simulation, 8 microphones are incorporated in a linear array with spacing of 15 cm (array width = 105 cm), but only 2 are selected for source separation using the approach in Equation (1). Each of the 8-choose-2 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 [11]. It is processed by concatenating 100 ms windows, as outlined in Section 3.1.2. The results show that nearly uniform SIR improvement (24 dB, relative to the unprocessed input signal) is achieved regardless of the 2 microphones selected. On the other hand, Figure 6(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 dB to 20 dB). 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 sub-array that is in the closest proximity to the two speakers, as illustrated in Figure 7. This allows us to receive the highest SNR signal, enabling higher quality microphone recordings and improved transfer function estimates.

2. Each sub-array is composed of 8 microphone channels. Section 3.1.2 describes the algorithm that carries out signal separation using the microphone inputs from the sub-array. This approach enhances robustness to the microphone variations (as quantified below).

However, using multiple sub-arrays each with 8 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 8 channels from each sub-array are sequentially sampled using a TFT scanning circuit. With this configuration, as shown in Figure 10, 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 3.4). This means that each channel can no longer be sampled at the Nyquist rate. Instead each channel of the sub-array is critically sampled. Namely, over the 8-channel sub-array, each channel is sampled at 2.5 kHz; since for high intelligibility we can bandpass filter human speech between 300 Hz and 5 kHz [12], this results in four aliases from each source, giving a total of eight aliases for the two sources. Section 3.1.2 describes the algorithm for separating these aliases using signals acquired from the 8 microphone channels. Figure 8 illustrates the benefit, comparing the simulated performance of the critically sampled system with 8 microphones, to the best, median, and worst performance

from 8-choose-2 microphone combinations shown in Figure 6(b). As seen, the proposed critically sampled system achieves performance at the level of the median combination, overcoming the severe sensitivity to microphone placement that would otherwise limit performance in a practical room with practical microphones.

3 System Design Details

Figure 9 shows the eight-channel sub-array hybrid system, which combines LAE and CMOS [13]. In the LAE domain there are 8 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 analog interface to CMOS. The CMOS IC includes digital control to multiplex between the two interfaces, to achieve critical sampling over the entire 8-channel sub-array. 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 8 microphones, each effectively sampled at 2.5 kHz, are fed to an algorithm for speech separation (currently off-chip).

3.1 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.

3.1.1 Calibration

Calibration 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 [14]. 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 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). Figure 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 sub-array. 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.

3.1.2 Reconstruction

Having measured the transfer functions, now two users can speak simultaneously while the 8 channels are critically sampled at a total rate of 20 kHz (2.5 kHz per channel). As illustrated in Figure 11, considering speech limited to a frequency of 5000 Hz, for every frequency bin of reconstruction, this leads to 4 aliases from each of the 2 sources. For each frequency bin, the 8 unknowns can thus

be resolved using the following system of equations:

$$\begin{bmatrix} Y_1(e^{j\omega}) \\ \vdots \\ Y_K(e^{j\omega}) \end{bmatrix} = \begin{bmatrix} A_{1,1}(e^{j(\omega/M)}) & \cdots & A_{2,1}(e^{j(\omega/M-2\pi(M-1)/M)}) \\ \vdots & & \vdots \\ A_{1,K}(e^{j(\omega/M)}) & \cdots & A_{2,K}(e^{j(\omega/M-2\pi(M-1)/M)}) \end{bmatrix} \begin{bmatrix} S_1(e^{j(\omega/M)}) \\ \cdots \\ S_1(e^{j(\omega/M-2\pi(M-1)/M)}) \\ S_2(e^{j(\omega/M)}) \\ \cdots \\ S_2(e^{j(\omega/M-2\pi(M-1)/M)}) \end{bmatrix}$$

$(M = K/N = 4)$

$Microphone\ Signals \qquad Transfer-function\ \ Matrix \qquad Source\ Signals$

(3)

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 BW=2×5 kHz (double sideband), interleaved sampling is carried out over all (N×BW)/K=2.5 kHz, which means the signals $Y_{1...8}$ 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 Figure 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 8 channels. Next the individual time samples corresponding to each channel are extracted, resulting in 8 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 Equation (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 [16]. Each frame is overlapped

by 75% with the preceding frame, so as to ensure it meets the constant overlap-add condition for the Hanning windows used in order to mitigate artifacts [17].

3.2 Thin-film piezoelectric microphone

Figure 12 shows the microphone, which is based on a diaphragm formed from 1.5 cm (width) $\times$ 1.0 cm (length) PVDF (Polyvinylidene fluoride), 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 standoff 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 Ω/sq are applied to both faces of the film by spray-coating silver nanowires [18], 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 Figure 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 concentrated from 500 - 3000 Hz [12]. 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 $\sim$ 40 μV.

3.3 TFT Amplifiers

In addition to a PVDF microphone, each channel has its own localized two-stage differential amplifier, formed from a-Si TFTs [19] with W/L=3600 μm/6 μm, as shown in Figure 13. The first stage is a gain stage (with gain of 17 dB), while the second is a buffer stage (with gain of 3 dB) to drive long ($\sim$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).

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. Figure 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 Figure 13(a)). Figure 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 $4\times$ 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. This shows the benefit of using localized TFT amplifiers fabricated over large-areas to interface with the microphones.

3.4 TFT Scanning Circuit and LAE / CMOS Interfaces

For every sub-array, 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 sub-arrays, corresponding to three signals from CMOS to LAE, required for controlling the large-area scanning circuits.

After the long LAE interconnects ($\sim$1 m), signals are provided to the CMOS IC through the TFT scanning circuit previously reported in [20]. The circuit is placed after the long interconnects to minimize the capacitance that must be driven due to the step response during scanning. The circuit is shown in Figure 14(a), consisting of level converter blocks and scan blocks, based only on NMOS devices, since the extremely low mobility of holes in a standard a-Si TFT technology precludes the use of PMOS devices ($\mu_h <$ 0.1 cm^2/Vs) [21] . 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 $CLK_{IC}/CLKb_{IC}$ in order 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 $CLK_{IC}/CLKb_{IC}$ (as shown in Figure 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.

3.5 CMOS IC

The outputs of the scanning circuit are fed directly into the CMOS IC for readout. As shown in Figure 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.

3.5.1 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 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 4, the simulation matches the measured results.

3.5.2 Variable-Gain Amplifier

The variable-gain amplifier is important because the microphone variability and variations in speaker distance from the microphones means that the received 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 3.1.1.

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).

3.5.3 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. Further, the hold capacitors are configurable, implemented as a 4-bit capacitor DAC. 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 opamp with dominant pole compensation for stability.

4 Prototype Measurements and System Demo

Figure 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. The CMOS IC was implemented in a 130 nm technology from IBM. The microphone sub-array spanned a width of 105 cm, and consisted of eight PVDF microphones, linearly spaced by 15 cm.

Table 1 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 sub-array consumes 12 mW. The CMOS readout IC consumes 0.6 mW in total. Figure 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 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 $\times$ 6 m classroom. The testing setup is shown in Figure 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 kHz 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 $\sim$ 50 dB_{SPL}. Figure 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. Figure 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 with a signal-to-interferer ratio improvement of 11 dB.

5 Conclusions

Multi-speaker 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; (3) demonstrate an 8-channel sub-array system, spanning the entire signal chain from the transducer to digitization,

which successfully separates two simultaneous audio sources.

6 Acknowledgements

The work is funded by the Qualcomm Innovation Fellowship, and NSF grants ECCS-1202168 and CCF-1218206. We also 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 Transactions on 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, Sept. 2014.

[3] C. Dagdeviren, Y. Su, P. Joe, R. Yona, Y. Liu, Y. Kim, Y.Huang, A. Damadoran, J. Xia, L. Martin, Y. Huang, and J. Rogers, "Conformable amplified lead zirconate titanate sensors with enhanced piezoelectric response for cutaneous pressure monitoring," *Nature Communications*, vol. 5, no. 4496, pp. 1–10, Aug. 2014.

[4] Y. Kuo, "Thin film transistor technology-past, present, and future," *Electrochemical Society Interface*, vol. 22, no. 1, pp. 55–61, 2013.

[5] N. Verma, Y. Hu, L. Huang, W. Rieutort-Louis, J. Sanz-Robinson, T. Moy, B. Glisic, S. Wagner, and J. C. Sturm, "Enabling scalable hybrid systems: architectures for exploiting large-area electronics in applications," *Proceedings of IEEE*, vol. 103, no. 4, pp. 690–712, April 2015.

[6] E. Weinstein, E. Steele, K. Agarwal, and J. Glass, "A 1020-node modular microphone array and beamformer for intelligent computing spaces," MIT, M1T/LCS Technical Memo MIT-LCS-TM-642, Tech. Rep., 2004.

[7] H. V. Trees, *Optimum Array Processing: Part IV Detection, Estimation, and Modulation Theory.* New York: John Wiley and Sons, pp. 66, 2002.

[8] K. Kokkinakis and P. Loizou, *Advances in Modern Blind Signal Separation Algorithms: Theory and Applications.* Morgan and Claypool, pp. 13-19, 2010.

[9] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[10] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[11] P. Kabal, "Tsp speech database," Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada, Tech. Rep., September 2002.

[12] R. L. Freeman, *Fundamentals of Telecommunications.* John Wiley and Sons, pp. 90-91, 2005.

[13] L. Huang, J. Sanz-Robinson, T. Moy, Y. Hu, W. Rieutort-Louis, S. Wagner, J. C. Sturm, and N. Verma, "Reconstruction of multiple-user voice commands using a hybrid system based on thin-film electronics and cmos," *VLSI Symposium on Circuits (VLSIC)*, no. JFS4-4, 2015.

[14] 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 Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

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

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

[17] C. Roads, *The Computer Music Tutorial.* Massachusetts Institute of Technology: MIT Press, pp. 553-555, 1996.

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

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

[20] T. Moy, W. Rieutort-Louis, Y. Hu, L. Huang, J. Sanz-Robinson, J. C. Sturm, S. Wagner, and N. Verma, "Thin-film circuits for scalable interfacing between large-area electronics and cmos ics," *Device Research Conference*, pp. 271–272, June 2014.

[21] R. Street, *Hydrogenated Amorphous Silicon.* Cambridge University Press, pp. 237-243, 1991.

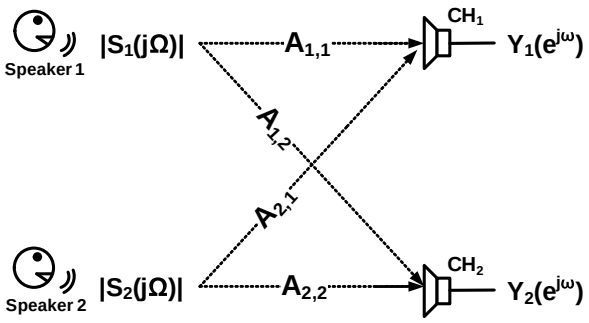


Figure 1: System of equations for separating two simultaneous sources recorded with two microphones using previously measured transfer functions.

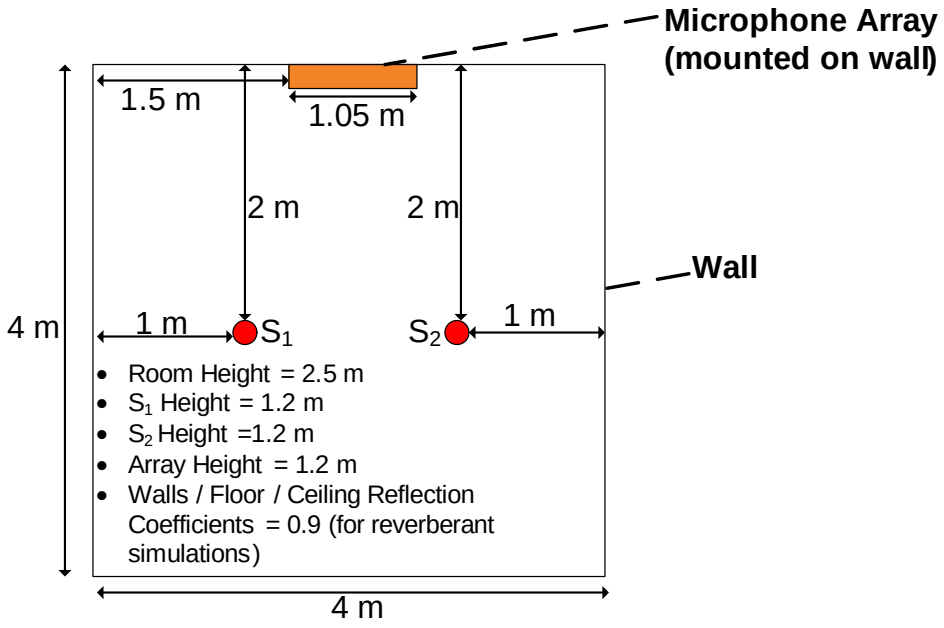
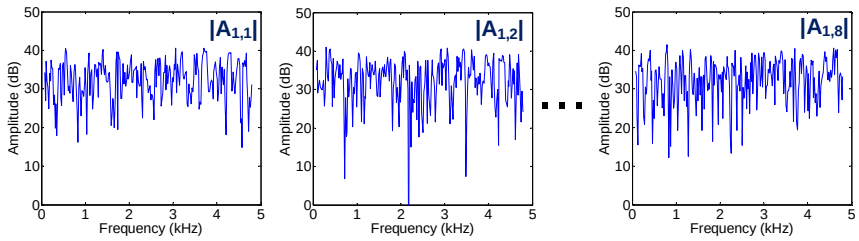
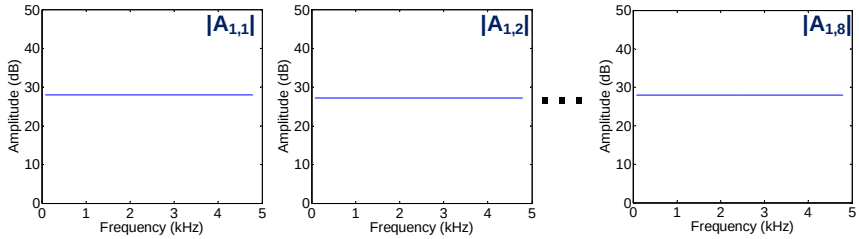


Figure 2: Simulation parameters for two simultaneous sound sources in a reverberant room.



(a)



(b)

Figure 3: Simulated frequency response of perfectly uniform, omnidirectional microphones and speakers in a (a) reverberant room, and (b) non-reverberant room.

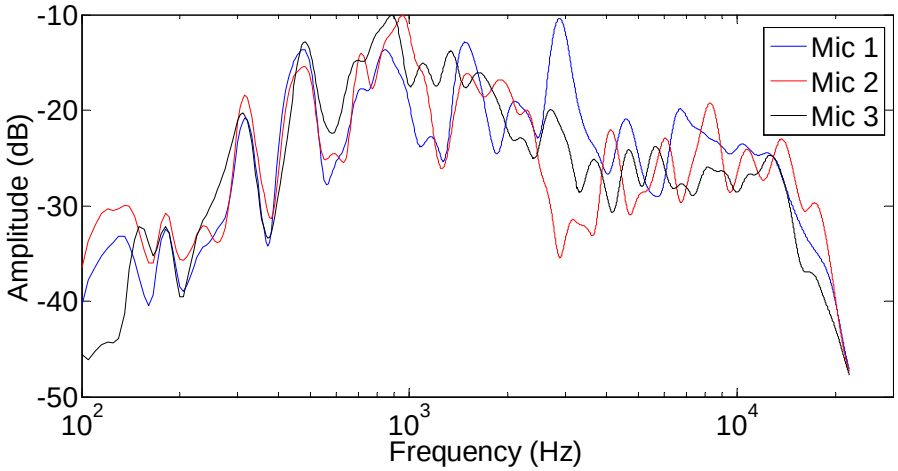


Figure 4: Frequency response of piezopolymer, PVDF, microphones measured in an anechoic chamber at an angle of 0° (directly facing the source).

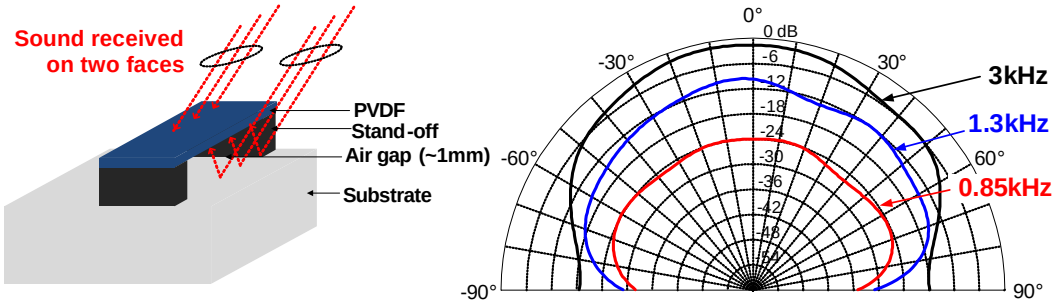
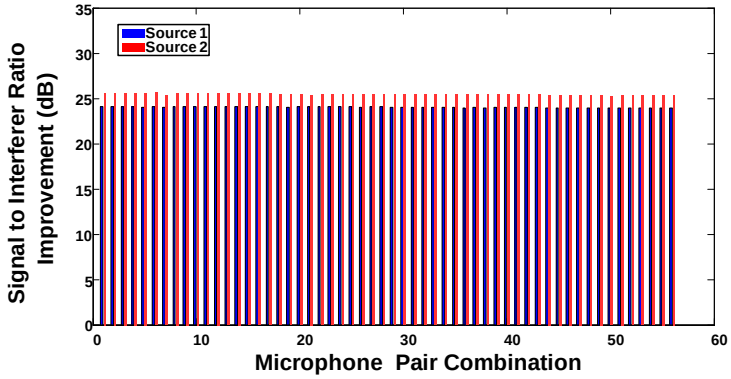
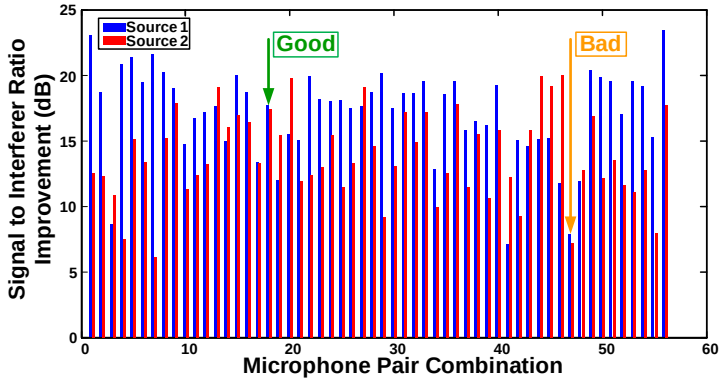


Figure 5: Polar diagram measured in an anechoic chamber of a PVDF microphone.



(a)



(b)

Figure 6: Reconstruction results for 8-choose-2 pairs of microphones. (a) Simulated in a room without reverberations, directionality, or microphone process variation; (b) with reverberations and directional microphones.

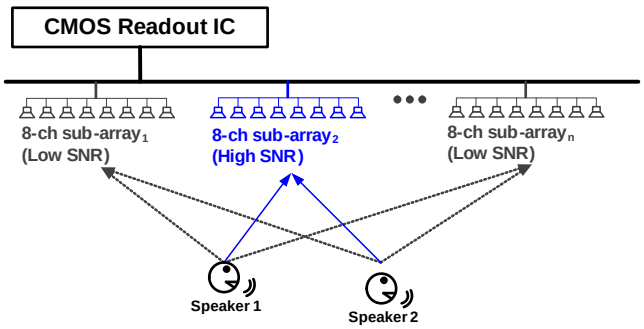


Figure 7: Proposed structure of the microphone array composed of high SNR sub-arrays in close proximity to the speaker.

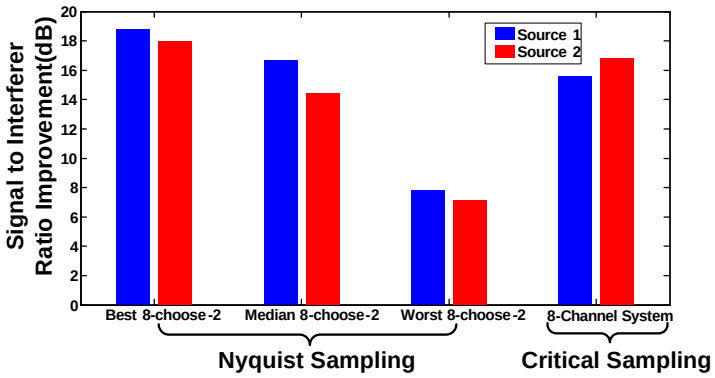


Figure 8: Simulated reconstruction results for the best, median and worst 8-choose-2 pairs of microphones, and for the 8-channel critically-sampled sub-array.

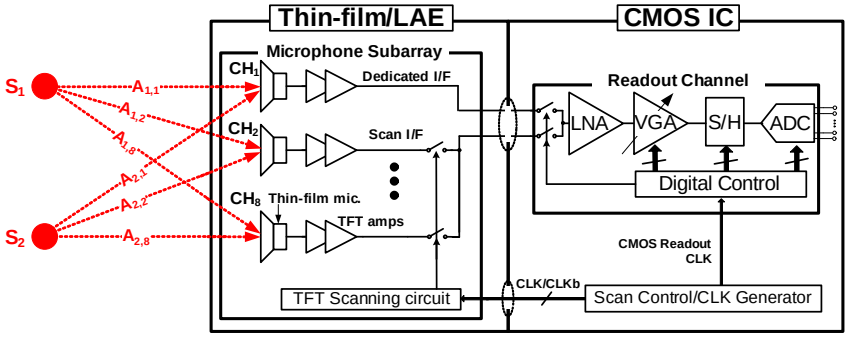


Figure 9: System architecture, combining CMOS ICs and large-area electronics (LAE).

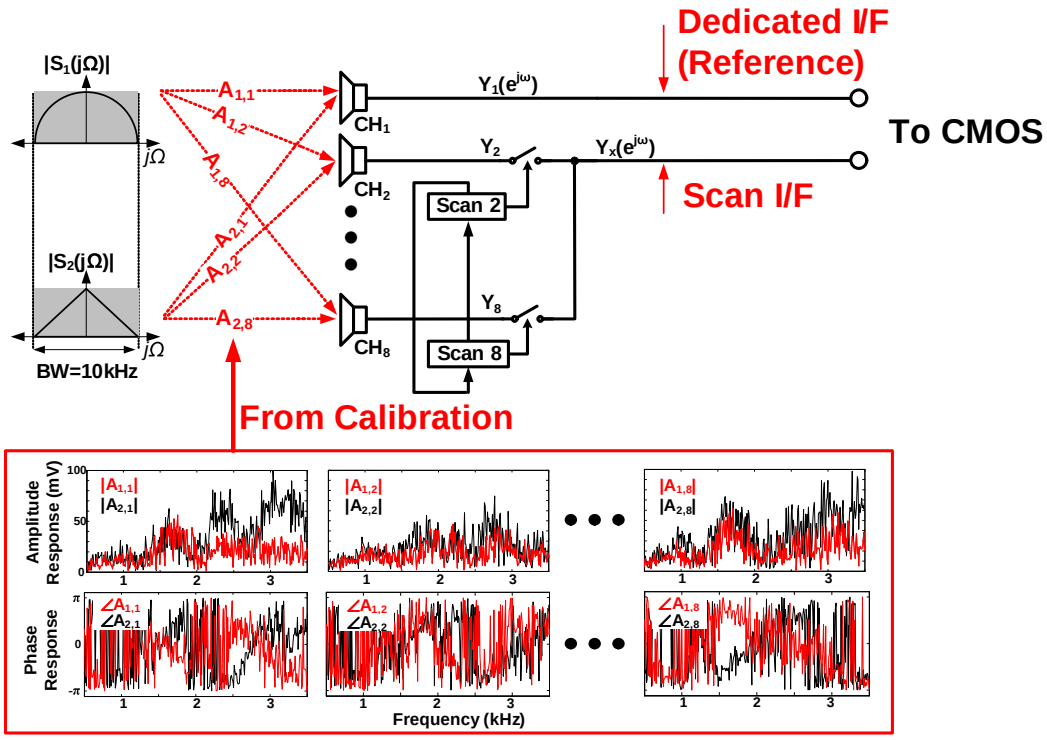


Figure 10: Calibration procedure used to find the transfer functions between each source and microphone.

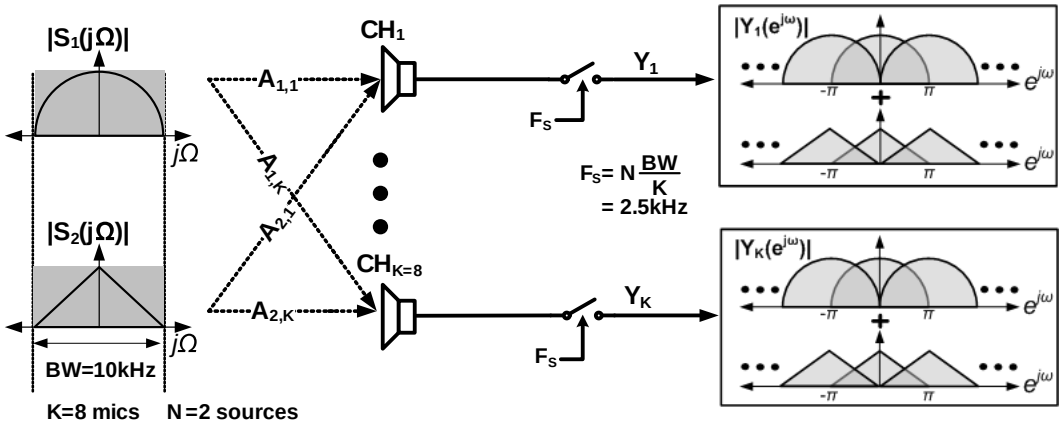


Figure 11: Algorithm for separating and reconstructing two acoustic sources from under-sampled microphones in a sub-array using previously calibrated transfer functions.

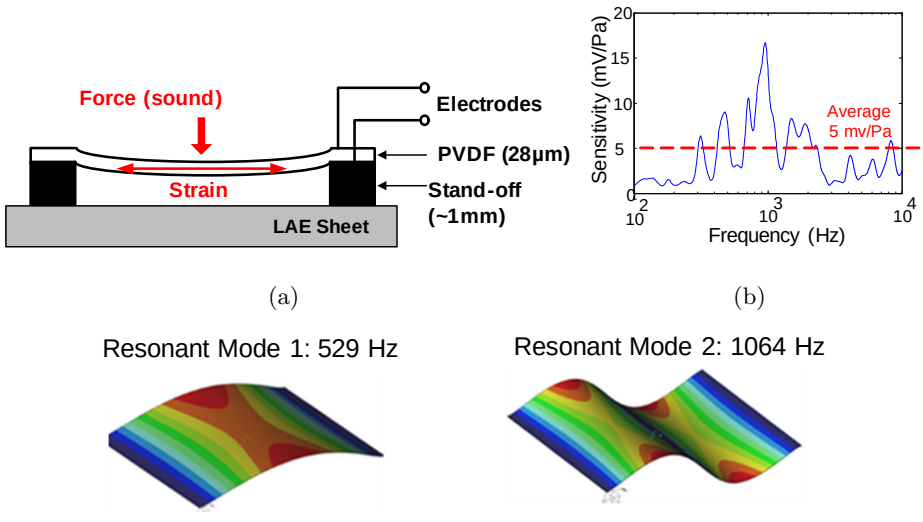


Figure 12: Thin-film PVDF microphone design, including (a) structure, (b) frequency response (measured in an anechoic chamber), and (c) finite element simulations showing the resonant modes.

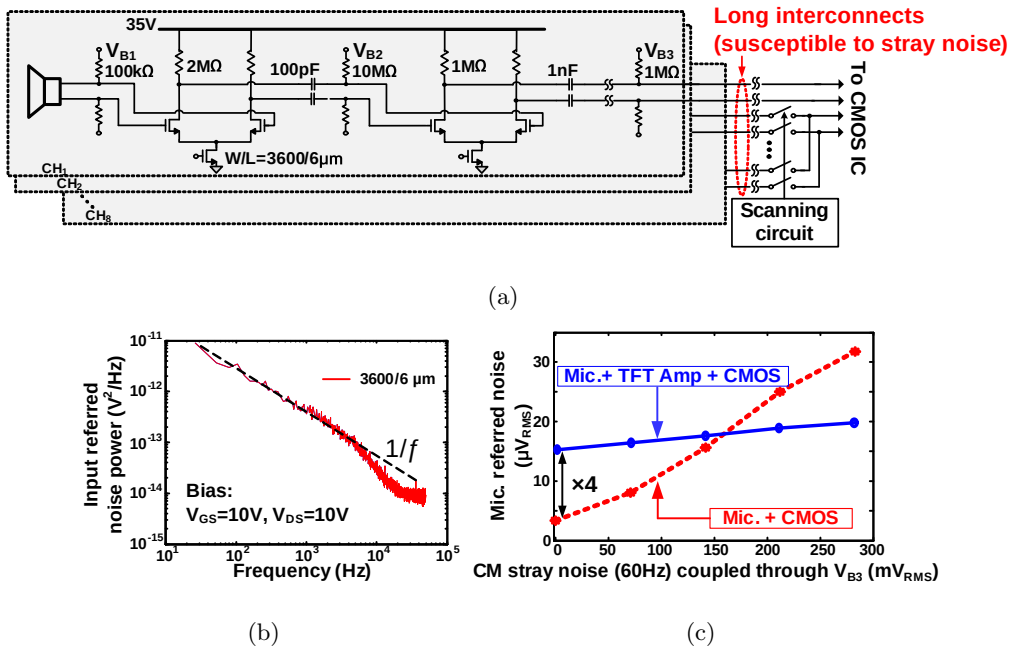


Figure 13: (a) Schematic of a two stage TFT amplifier, including (b) measured noise characteristics of an a-Si TFT , and (c) the tradeoff between a localized TFT amplifier and CMOS.

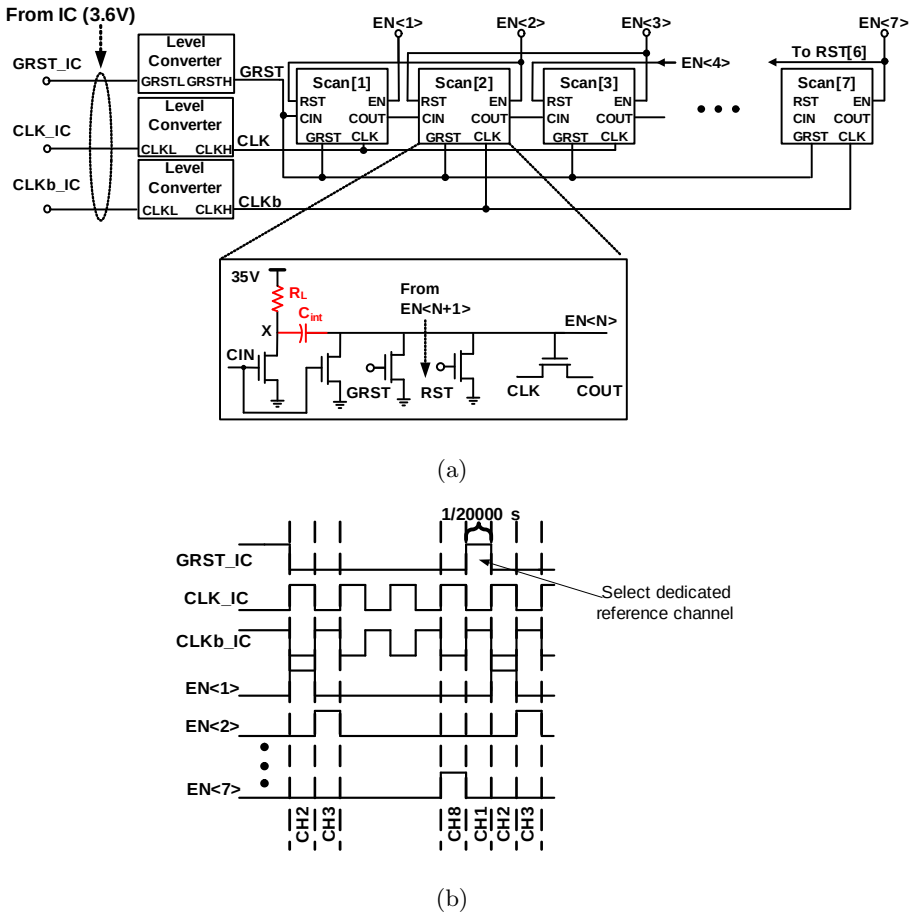


Figure 14: TFT scanning circuit (a) schematic and (b) timing diagram [20].

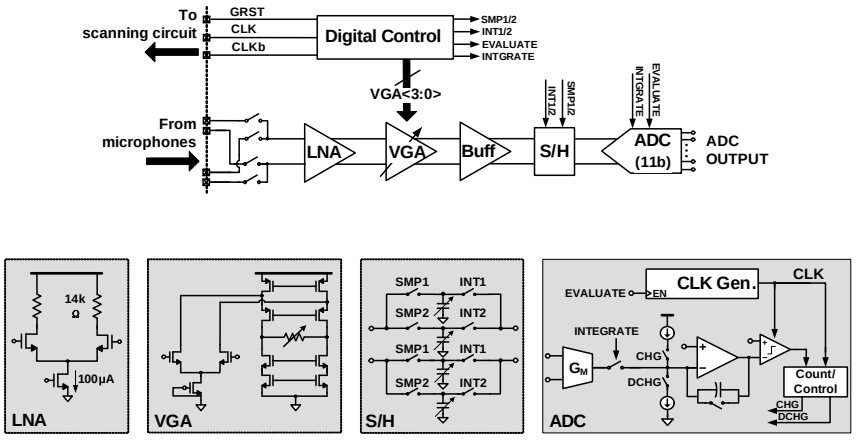
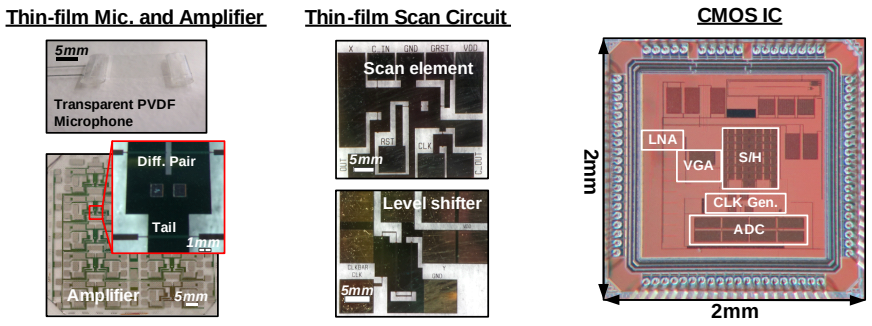
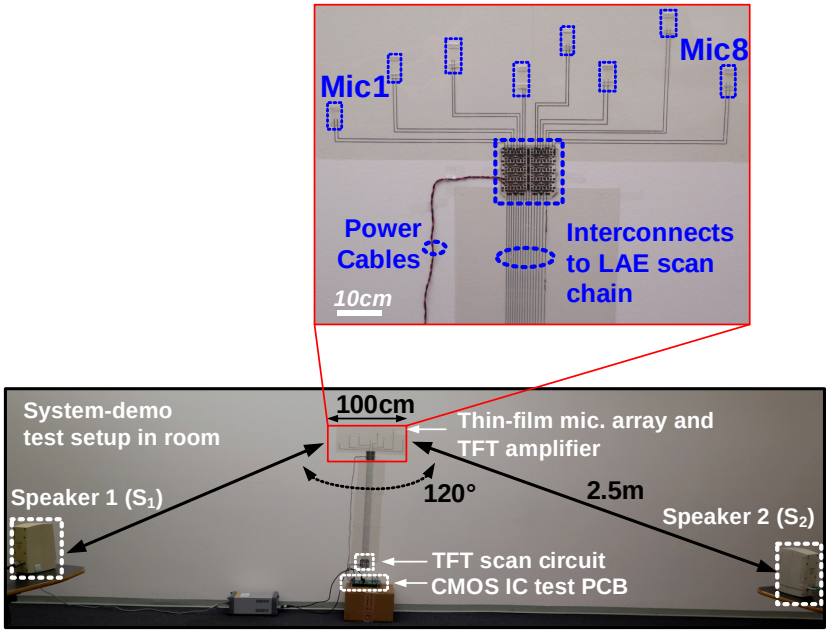


Figure 15: Schematics of the CMOS IC used for readout and digitization, which incorporates a LNA, VGA and 11-bit ADC.



(a)



(b)

Figure 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°.

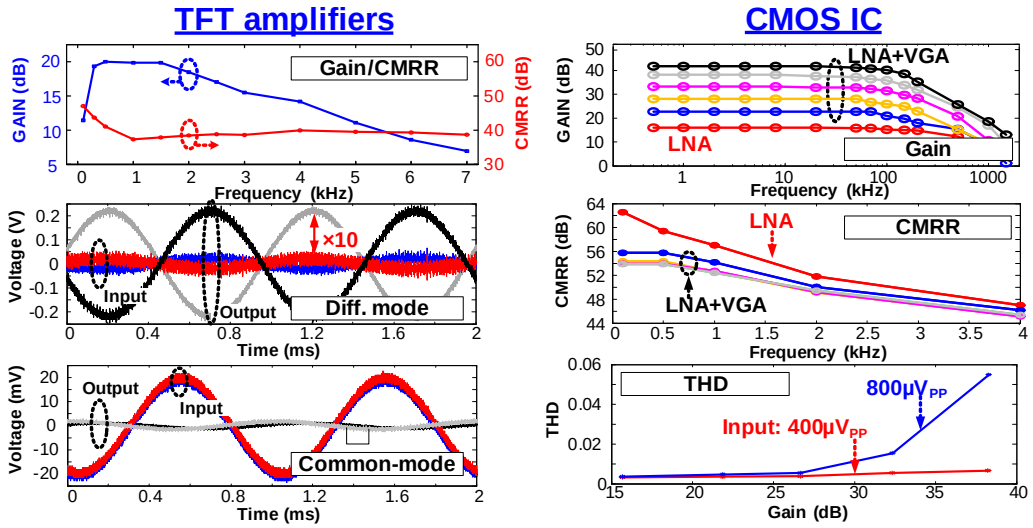


Figure 17: Component-level measurement.

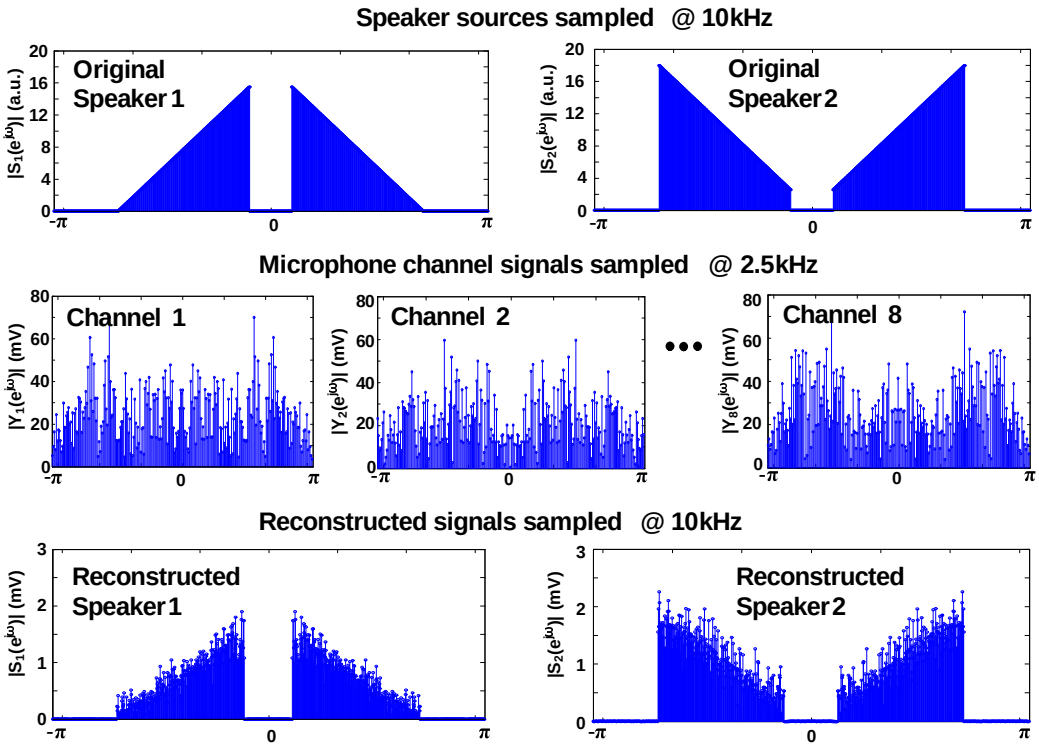


Figure 18: Demonstration of two-source separation and reconstruction for two simultaneous wedge-shaped inputs.

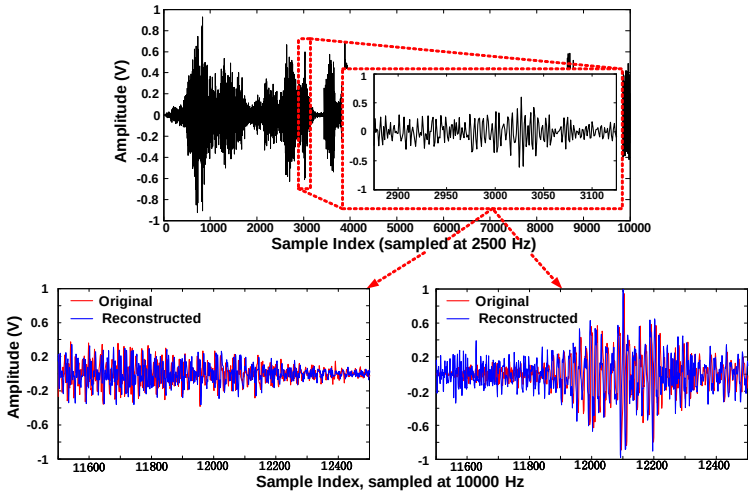


Figure 19: Demonstration of two-source separation and reconstruction for two simultaneous speech signals.

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

Table 1: Performance summary of the system.