Reconstruction of Multiple-user Voice Commands Using a Hybrid System Based on Thin-film Electronics and CMOS

L. Huang, J. Sanz-Robinson, T. Moy, Y. Hu, W. Rieutort-Louis,
S. Wagner, J. C. Sturm, N. Verma
Princeton University, Princeton, NJ, USA

Abstract

This paper presents a system consisting of an array of thin-film microphone channels on glass, which can be formed on large substrates. Each microphone channel consists of a polyvinylidene difluoride (PVDF) piezoelectric transducer as well as amplifier and scan circuits based on amorphous-silicon (a-Si) thin-film transistors (TFTs). The scan circuits multiplex signals from multiple channels to a CMOS IC for readout. By spatially distributing the channels on a large substrate, audio signals from multiple simultaneous speakers in a space can be both acquired in closer proximity and separated, enabling a multi-user human-computer interface based on voice commands. To overcome low TFT performance in the scan circuits (which limits channel sampling to below the Nyquist rate), a signal reconstruction algorithm is proposed. An 8-channel system demonstrates acquisition and reconstruction of 2 simultaneous audio signals at 2m distance from the array.

System Overview

Fig. 1 illustrates the hybrid thin-film/CMOS system. To span large spaces, the system can scale to multiple arrays of the thin-film channels. This makes scanning of the channels in each array necessary in order to reduce the interfaces to the CMOS IC. Note, only one set of CLK/CLKb signals from the CMOS IC is needed for controlling the scan circuits of all arrays; this leaves only two differential interface signals per array, as discussed in the next section. In an array, separation of N source signals $S_1 \ldots S_N$ (N=2 in our demo) is achieved using K channels (K=8 in our demo) by exploiting diversity of the transfer functions $A_{1,1 \ldots K} \ldots A_{N,1 \ldots K}$ from each source to each microphone. Typically, reconstruction algorithms (e.g., beam forming) rely on Nyquist-sampled channels [1]. But, inherently low performance of TFTs limits the scan circuit (previously presented in [2]) to ~20kHz. With voice signals having bandwidth BW≈10kHz, this substantially limits the number of channels (to ~2). Thus, an algorithm is presented to enable sampling below the channel Nyquist rate.

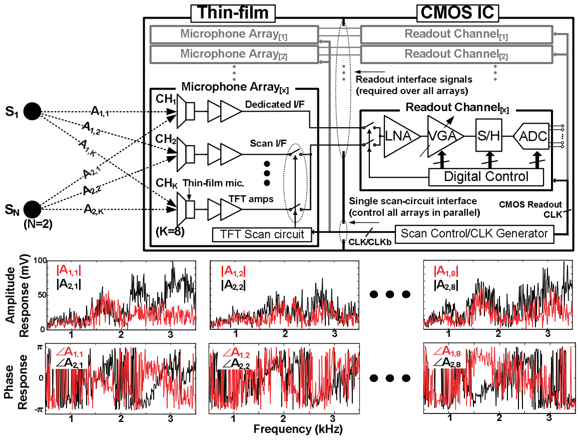


Fig. 1: System architecture with measured transfer functions from each source to each microphone.

Reconstruction Algorithm

Fig. 2 illustrates the reconstruction and separation algorithm. It consists of two phases. First, during calibration, each transfer function is characterized one-at-a-time. This is done using a calibration signal having spectral content at all frequencies of interest, but with precise level at each frequency not critical; in an application, this can be done by prompting users to speak one-by-one (in isolation). The system requires 100ms to characterize each transfer function (corresponding to 0.7s total per user, for all channels). To characterize the transfer functions, Nyquist sampling is necessary. This is achieved as shown in Fig. 1, where the first channel is provided to the CMOS IC via a dedicated interface, and the remaining channels are selected to a shared interface for 100ms each by the TFT scan circuit running at reduced speed (10Hz); the two interfaces are then multiplexed in the IC for readout. This enables amplitude and phase characterization of each channel *with respect to the first channel*.

Second, during reconstruction, interleaved sampling is performed by the scan circuit and IC multiplexor running at full speed; the total speed needed over all channels is N×BW, *i.e., independent of the number of channels*. Thus, for K=8, each channel is sampled at rate (N×BW)/K=2.5kHz, giving the signals $Y_{1 \ldots 8}$, effectively sampled below the Nyquist rate by a factor K/N=4. As seen in Fig. 2, K spectral components must thus be resolved in each frequency bin (K/N aliases from N sources). With adequate diversity in the transfer-function matrix, this is achieved using K channel signals. Having resolved the spectral components, reconstruction is then performed via an inverse modulated-filter-bank formulation [3].

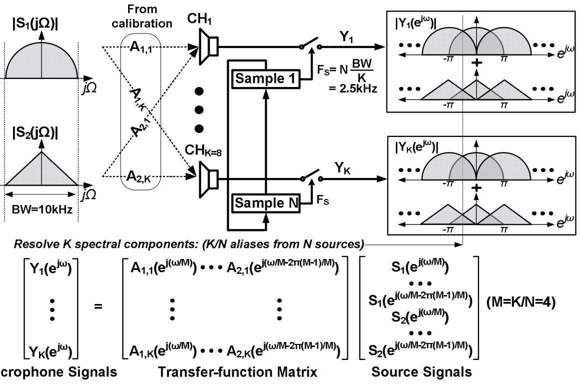


Fig. 2: Reconstruction and separation during interleaved sampling.

Thin-film Microphone Channels

Fig. 3 shows the microphone channel. The microphone is a 1.5cm(w)×1.0cm(l)×28μm(t) transparent PVDF diaphragm beam. Transparent electrodes are applied by spray-coating silver nanowires [4]. The frequency response shown is tuned by diaphragm sizing to match human speech, concentrated in 0.5-3kHz. For typical speech at a distance of 2m, the average sensitivity of 5mV/Pa yields a microphone signal of ~40μV.

Each channel consists of a two-stage differential amplifier, formed from a-Si TFTs with W/L=3600/6μm. The first stage is a gain stage, while the second is a buffer stage to drive long interconnects on the substrate. The amplifier chain has measured gain of 20dB, with pass band 0.3-3kHz and CMRR of 50dB at 100Hz. The small amplitudes and low frequencies of the microphone signals make 1/f noise in the amplifiers an important concern. On the other hand the amplifiers mitigate the effect of stray coupling (e.g., 60Hz) from extrinsic sources on the long interconnects. Fig. 3 analyzes this tradeoff based on measurements of the amplifier noise (PSD shown) and extrinsic noise. Without stray noise, the amplifier results in 4× higher input-referred noise, compared to direct acquisition of the microphone by the IC ($16\mu V_{RMS}$ vs. $4\mu V_{RMS}$). However, adding just $160mV_{RMS}$ of stray noise on the interconnect (60Hz applied at V_{B3}) leads to reduced input-referred noise thanks to gain provided by the amplifier. After the long interconnects (~1m), signals are provided to the CMOS IC through the TFT scan circuit previously reported [2], which operates at 20kHz from a 35V supply. Placing the TFT switches of the scan circuit after the interconnects mitigates the capacitance that must be driven due to the step response during scanning.

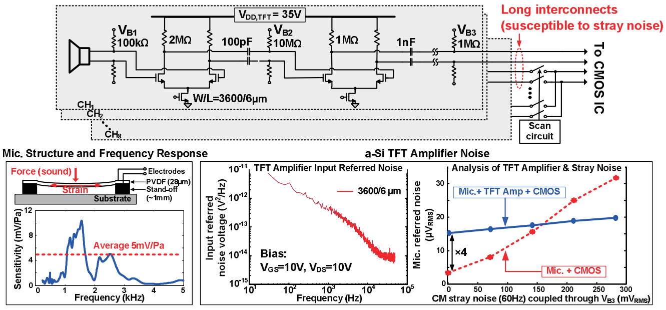


Fig. 3: Thin-film microphone channel.

CMOS Readout Circuitry

Fig. 4 shows the CMOS readout circuit, which performs sample acquisition at the end of the scan window. Following the low-noise amplifier (LNA) with gain of 16dB, a variable-gain amplifier (VGA) in folded-cascode topology provides 6-27dB gain (programmable by load resistor). Folded cascode addresses the large magnitudes possible, and gain programmability addresses the large magnitude variations possible in the microphone-channel transfer functions (Fig. 1) seen following calibration. After this, two S/H's alternatingly acquire samples for digitization by an 11-b dual-slope ADC. Two S/H's are employed with programmable capacitors to accommodate increased scanning rates if the number of sources N is increased beyond 2.

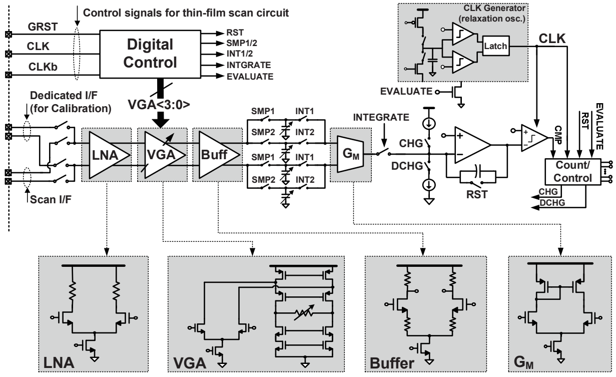


Fig. 4: CMOS readout circuits.

Measurement Results

The system prototype is based on PVDF film microphones and a-Si TFT circuits fabricated in house on glass at 180°C, as well as an IC fabricated in 130nm CMOS (shown in Fig. 5). Fig. 6 shows key measurements of the TFT amplifiers and CMOS readout circuits, and also provides a summary of the component-level measurements.

Fig. 7 shows the system demonstration setup and results. Two speakers separated by an angle of 120° are placed at a radial distance of 2.5m from the center of the microphone array. The entire array spans a width of 60cm. Calibration is performed using a white-noise signal (0.5-3.5kHz) played one-by-one through each speaker. Following this, source signals S_1 and S_2, sampled at 10kHz and intentionally synthesized to have DFTs with the distinct wedge-shaped magnitudes shown, are played simultaneously through the two speakers. For illustration, DFTs from three microphone channels (Y_1, Y_2, Y_8) sampled at 2.5kHz are shown, exhibiting source superposition and aliasing. As shown, the reconstruction algorithm, using the acquired 2.5kHz signals, successfully recovers samples of the original signals at 10kHz, showing the correct wedge-shaped magnitudes.

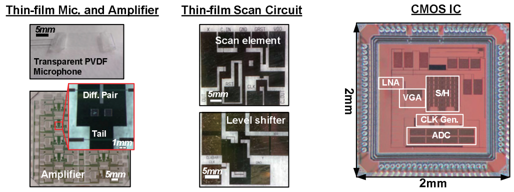
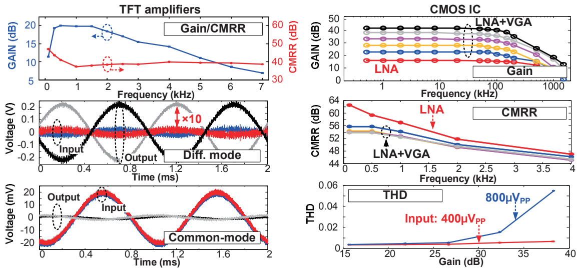


Fig. 5: Micrographs of components.



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	22μA @ 3.6V
Power	100μA @ 35V		Readout	450μA @ 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	360μA @ 35V	Input Referred Noise		$4\mu V_{rms}$

Fig. 6: Component-level measurements and summary.

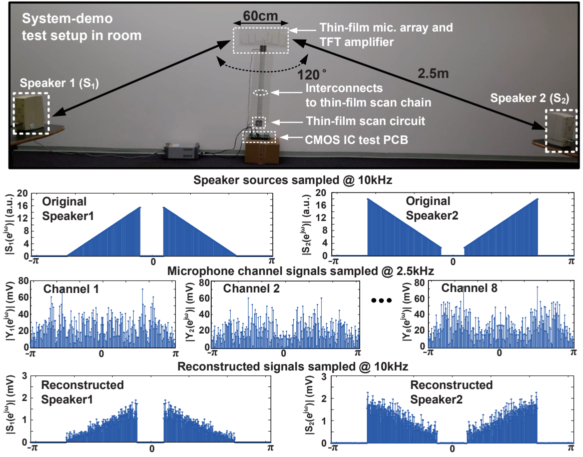


Fig. 7: Demonstration setup and results of two-source separation & reconstruction.

Reference:

[1] B. D. Van Veen, et al., *ASSP Mag.*, pp. 4 - 24, Apr. 1988.
[2] T. Moy, et al., *DRC*, pp. 271-272, June 2014.
[3] P. Sommen, et al., *IEEE TSP*, vol. 56, no.10, Oct. 2008.
[4] J. A. Spechler, et al., *Appl. Phys. A*, vol. 108, pp. 25-28, May 2012.