2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) | 978-1-6654-9772-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/VLSITechnologyandCir46769.2022.9830455

A 31-Feature, 80nW, 0.53mm^2 Audio Analog Feature Extractor based on Time-Mode Analog Filterbank Interpolation and Time-Mode Analog Rectification

Subhajit Ray, Peter R. Kinget
Columbia University, New York, NY, USA, email: subhajit.ray@columbia.edu

Abstract: To alleviate the feature extraction bottleneck in always-on, on-device Keyword Spotting (KWS), we propose two novel analog circuit techniques that are combined into an efficient analog feature extraction architecture: 1) *Time-Mode Analog Filterbank Interpolation* (TM-AFI) uses digital XOR gates to double the number of outputs of an analog filterbank, 2) *Time-Mode Analog Rectification* (TM-AR) uses a single digital XOR gate as an analog full-wave rectifier. Among other analog feature extractor chips using a software classifier for a KWS demo, the 31-feature, 80nW, 0.53mm^2 prototype is 18× more power-efficient and 3.3× more area-efficient than the most area- and power-efficient published works, respectively, while maintaining competitive >90% accuracy on 10 keywords.

Motivation: Always-on, on-device KWS demands ultralow power and area. Fully-integrated KWS chips suffer from the feature extractor consuming 2.7×/2.2× more power/area than the backend classifier [1], motivating research into efficient feature extraction. Compared to digital feature extraction [2], analog [3] can be 14.6× more power-efficient, but is 10.7× less area-efficient. Simultaneously achieving power- and area-efficiency in analog feature extraction while maintaining accuracy through a backend classifier remains a challenge.

Advantages of TM-AFI: The essential architecture of analog feature extractors [3-5] has remained the same for nearly 40 years [6]: *N* bandpass filters followed by *N* rectifiers to produce *N* bandpass responses. By using XORs to compute subtractions between adjacent filters, TM-AFI breaks the *N*-to-*N* limitation by requiring only *N*/2 filters to produce *N* responses. This reduces filterbank power/area by 2×/2× and total power/area by 1.6×/1.7×.

Advantages of TM-AR: The conventional diode bridge full-wave rectifier is both power- and area-hungry [7]. TM-AR replaces this with a single XOR that computes an absolute value, equivalent to full-wave rectification. This reduces rectifier power/area by 6.6×/2.8× (including overhead of conversion to time-mode) and total power/area by 3.7×/1.9×.

Time-Mode Analog Signal Representation: Fig. 1 shows the architecture. After the analog audio is analyzed by a filterbank, each frequency component is compared against a triangle wave by a continuous-time comparator to convert from voltage-mode to time-mode, where an analog number is represented as a continuous-valued pulse width (**Fig. 2**). In this representation, *an XOR gate is an efficient implementation of subtraction merged with an absolute value* (**Fig. 3**). This property is the basis of TM-AFI and TM-AR, and, interestingly, can be linked to stochastic bitstream-based computing from the 1960s [8].

Principle of TM-AFI (Fig. 4): Each green XOR computes an absolute-valued subtraction between adjacent time-mode frequency components, merging two functions: 1) the creation of a new, interpolated component, as the subtraction between two bandpass signals is a bandpass signal, and 2) the full-wave rectification of the interpolated component. This produces 2*N*-1=31 bandpass responses from *N*=16 bandpass filters.

Principle of TM-AR (Fig. 5): Each black XOR gate computes an absolute-valued subtraction between a "native" time-mode frequency component and a 50% duty cycle square wave. Because the latter represents a time-mode zero, the absolute-valued subtraction reduces to just an absolute value.

Integration & Conversion to Digital (Fig. 5): Next, the output duty cycle of each XOR is converted into a proportional spike rate by an integrate-and-fire (IAF) [7], and a counter integrates this spike rate by counting the number of spikes as a digital number, merging integration with conversion to digital [7]. Integration windows 20ms long with 10ms overlap are implemented using a pair of counters.

Circuit Design (Fig. 6): Selected for its power efficiency, the filter is a super source-follower (SSF) biquad [9]. To further optimize power, the output-referred noise is set equal to the input-referred offset of the comparator. With the capacitances kept the same across the bank, the center frequencies (g_m/C) are scaled geometrically from 100Hz to 4kHz via the bias current, owing to the constant g_m/I_d in deep weak inversion. Since the original SSF biquad [9] has limited Q, a cross-coupled pair is introduced to boost it to 3 because this choice results in Q=3 for the interpolated responses. The comparator is a differential difference open-loop OTA. As its delay limits the min/max pulse widths, it's set to 10% of the PWM period. To efficiently reduce offset, 1 of 8 sub-comparators is selected.

Measurements (Fig. 7): The 65nm LP CMOS prototype consumes 80nW from 0.6V/0.4V analog/digital supplies and 0.53mm^2 active area. The center left confirms that TM-AFI interpolates new responses from the native ones in the upper left, while the lower left shows that the interpolated center frequencies are between the native ones. Mismatch-induced variations are tolerated because for KWS, they can be learned by the classifier, a property exploited to reduce area. Using the Google Speech Commands (GSC) dataset [10], the chip's digital output is fed directly into a standard small-footprint software classifier [11] to achieve 91.5% accuracy on 10 keywords. Even though the max/min in a typical spectrogram is 36dB, competitive accuracy is maintained, demonstrating that KWS does not require high dynamic range from the feature extractor, a property exploited to reduce power.

Conclusion (Fig. 6 lower right, Fig. 8): By leveraging time-mode analog signal processing, the demonstrated feature extractor chip simultaneously achieves high power- and area-efficiency while maintaining competitive accuracy through a software classifier, proving the quality of the extracted features.

References: [1]J. Giraldo, VLSI, 2019. [2]W. Shan, ISSCC, 2020. [3]M. Yang, JSSC, 2021. [4]D. Wang, ISSCC, 2021. [5]D. Villamizar, TCASI, 2021. [6]N. Bui, JSSC, 1983. [7]M. Yang, ISSCC, 2018. [8]B. Gaines, AFIPS, 1967. [9]M. Matteis, JSSC, 2015. [10]P. Warden, arXiv:1804.03209, 2018. [11]R. Tang, ICASSP, 2018.

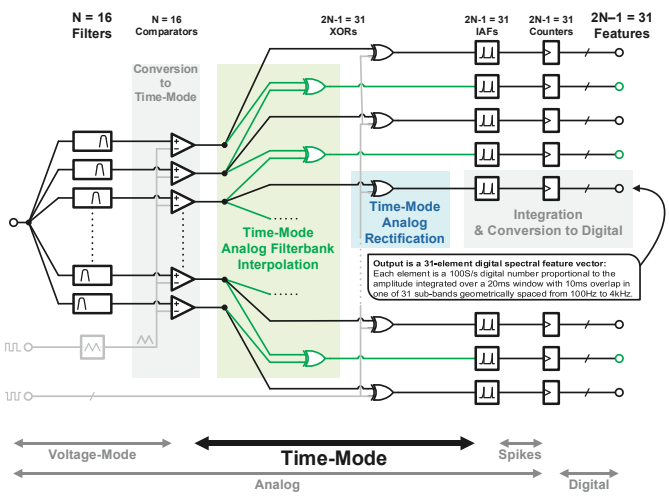

Fig. 1: Chip architecture of the analog feature extractor. Input is an analog audio signal and output is a digital spectral feature vector. Stacking feature vectors over time forms an auditory spectrogram.

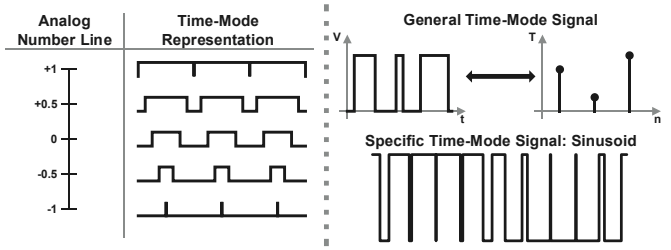

Fig. 2: Time-Mode analog signal representation used in the chip.

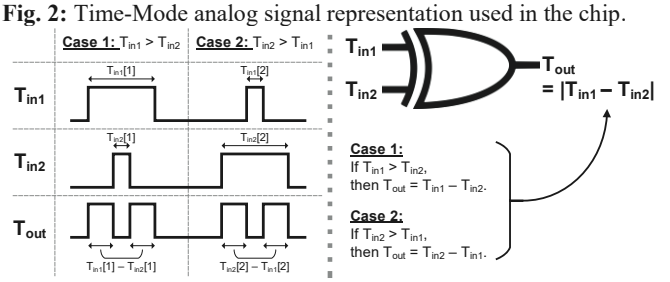

Fig. 3: In Fig. 2's signal representation, an XOR gate is an efficient implementation of subtraction merged with an absolute value.

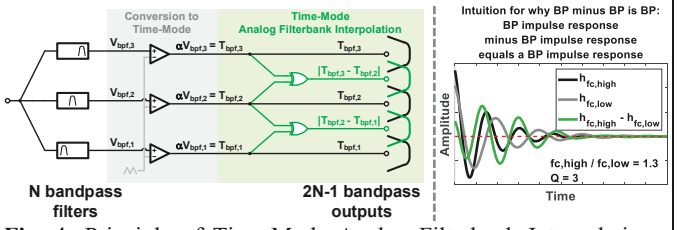

Fig. 4: Principle of Time-Mode Analog Filterbank Interpolation. New, interpolated bandpass responses are created by XOR gates.

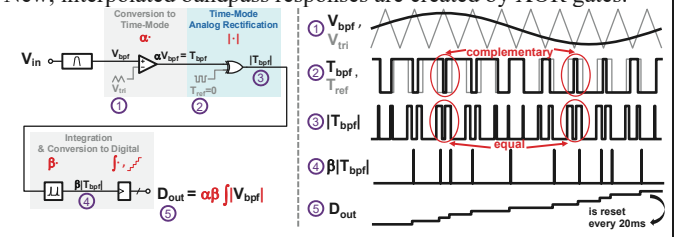

Fig. 5: Principle of Time-Mode Analog Rectification, along with channel waveforms (behaviorally-simulated). Full-wave rectification is performed by a single XOR gate.

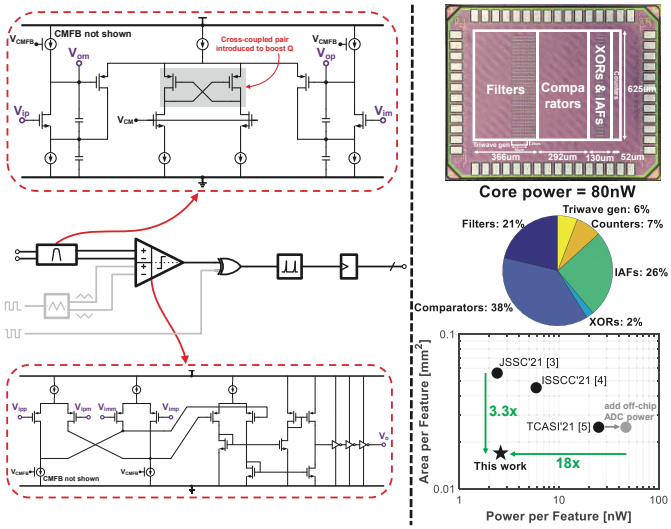

Fig. 6: Schematics of key circuit blocks—filter, comparator—in a channel. Die photo, power breakdown, and performance plot.

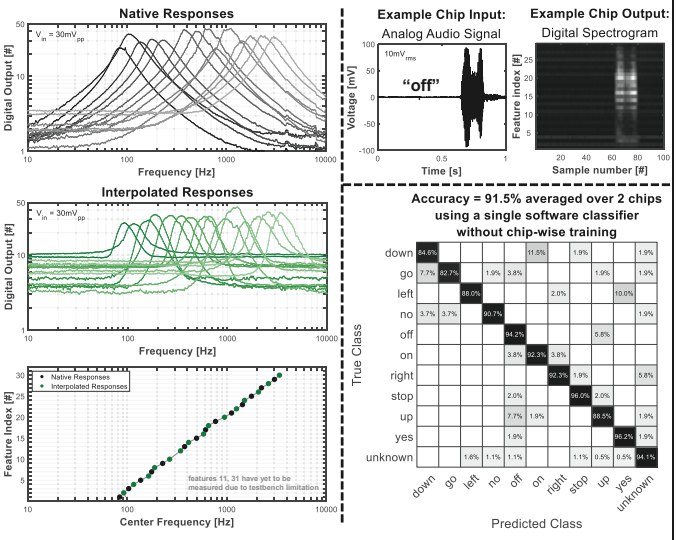

Fig. 7: Sinewave (left) and audio (right) measurement results.

Audio Analog Feature Extractor Chips

		10-keyword demo		< 10-keyword demo	
		This work	TCASI'21 [5]	JSSC'21 [3]	ISSCC'21 [4]
Signal processing technique	[x]	Time-Mode	Switched-Capacitor	Nonlinear	Nonlinear + Normalization
Technology	[nm]	65	130	65	65
Input/output format	[x]	analog/digital	analog/analog	analog/events	analog/events
Off-chip ADC required?	[x]	no	yes, to interface with classifier	no	no
Num of bandpass features	[#]	31	32	16	16
Power*	[nW]	80	800 (w/o off-chip ADC) 1500 (w/ off-chip ADC[b])	38	94
Area	[mm²]	0.53	0.79[b]	0.90	0.72
Power per feature	[nW]	2.6	25.0 (w/o off-chip ADC) 46.9 (w/ off-chip ADC)	2.4	5.9
Area per feature	[mm²]	0.017	0.025[b]	0.056	0.045
Frequency Range	[Hz - Hz]	100 - 4k	30 - 8k	100 - 5k	100 - 5k
Bandwidth-normalized power per feature†	[nW]	9.5	66.2 (w/o off-chip ADC) 124.1 (w/ off-chip ADC)	7.0	17.5

Keyword Spotting Application Demonstration

Accuracy on GSC	[%]	91.5% on 10 words	82.4 - 87.4%[c] on 10 words	94.2% on 1 word	90.2% on 4 words
Keywords used	[x]	yes, no, up, down, left, right, on, off, stop, go	yes, no, up, down, left, right, on, off, stop, go	four	N/A
Classifier	[x]	off-chip software CNN	off-chip software RNN	off-chip software CNN	off-chip hardware SNN

*for fair comparison, power of mic pre-amp subtracted from [3], [4]

†calculated according to eqn used in [5]: $P_{norm} = P_{total} \cdot \frac{1-r}{1-r^N} \cdot \frac{4kHz}{f_{max}}$, $r = \left(\frac{f_{min}}{f_{max}}\right)^{\frac{1}{N-1}}$

[a] estimate of 700nW for off-chip ADC is from Fig. 15 in [5]. [b] does not include area of necessary off-chip ADC

[c] includes the 5-10% degradation from excluding the off-chip logarithm, as stated in sec. IV-B of [5]

Fig. 8: Comparison table of audio analog feature extractor chips using off-chip classifiers for a keyword spotting application demo.