IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ENERGY HARVESTING TECHNOLOGIES FOR WEARABLE AND IMPLANTABLE DEVICES

Received February 17, 2020, accepted February 29, 2020, date of publication March 10, 2020, date of current version March 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2979799

MSP-MFCC: Energy-Efficient MFCC Feature Extraction Method With Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications

QIN LI[1,2], YUZE YANG[1,2], TIANXIANG LAN[1,3], HUIFENG ZHU[1,2], (Student Member, IEEE), QI WEI[1,2], FEI QIAO[1,2], (Member, IEEE), XINJUN LIU[4], (Member, IEEE), AND HUAZHONG YANG[1,2], (Fellow, IEEE)

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China
[3]School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[4]Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China

Corresponding author: Fei Qiao (qiaofei@tsinghua.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 91648116, and in part by the Beijing Innovation Center for Future Chips, Tsinghua University.

ABSTRACT Feature extraction is an essential part of automatic speech recognition (ASR) to compress raw speech data and enhance features, where conventional implementation methods based on the digital domain have encountered energy consumption and processing speed bottlenecks. Thus, we propose a Mixed-Signal Processing (MSP) architecture to efficiently extract Mel-Frequency Cepstrum Coefficients (MFCC) features. We design MSP-MFCC to pre-process speech signals in the analog domain, which significantly reduces the cost of the analog-to-digital converter (ADC), as well as the computational complexity of the digital back-end. Moreover, MSP-MFCC eliminates the time-consuming Fourier transform in the conventional digital realization by improving processing flow. We fabricated the analog part based on 180nm CMOS mixed-signal technology, then measured the chip. The measured results show the energy consumption of MSP-MFCC is 0.72 μJ/frame, and the processing speed is up to 45.79 μs/frame. MSP-MFCC achieves 95% energy saving and about 6.4$\times$ speedup than state of the art. Further, by using the features extracted by MSP-MFCC, speech recognition simulation reaches the accuracy of 98.2%, which also keeps the leading performance to its current counterparts. The proposed MFCC extractor is competitive for integration in the ultra-low-power always-on wearable speech recognition applications.

INDEX TERMS Mixed signal processing architecture, energy-efficient feature extraction, mel-frequency cepstrum coefficients (MFCC), wearable speech recognition application.

I. INTRODUCTION

Speech interaction has become an essential way of human-machine interaction [1], [2], in which, automatic speech recognition (ASR) plays a vital role in perceiving speech signals. In scenarios such as energy-constrained, network restricted wearable devices, energy-efficient speech recognition is important for the working and standby time of the devices. However, always-on ultra-low-power wearable speech recognition is still a challenge for these devices. Thus, the energy-efficient fast-processing ASR system has always been widely concerned [3]–[7]. As shown in Fig. 1(a), the current ASR system is composed of feature extraction and recognition modules. The feature extraction costs the most energy consumption in specific tasks [3], [8], [9] and determines the recognition performance even in the end-to-end speech recognition system [10]. In addition, the feature extraction part is always-on in the recognition or wake-up tasks, which is energy-consuming. Thus, we focus on the feature extraction of the wearable speech recognition system in this work.

Inspired by the human hearing mechanisms, Mel-Frequency Cepstrum Coefficients (MFCC) feature is presented and becomes the most widely used feature [4] due to its high accuracy in this field. However, in the ASR tasks for mobile devices, the entire MFCC feature extraction

The associate editor coordinating the review of this manuscript and approving it for publication was Hadi Heidari.

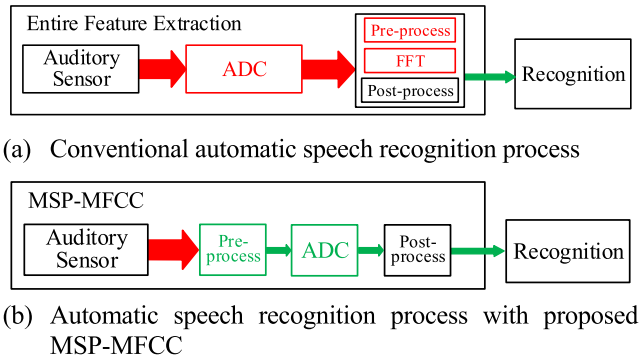

FIGURE 1. **(a) Conventional feature extraction process incurs significant workload on the Analog-to-Digital Converter (ADC) and Fast Fourier Transform (FFT), whereas (b) MSP-MFCC alleviates the workload of ADC and eliminates FFT.**

process accounts for nearly 32% to 93% of system power consumption [3]–[5]. Therefore, a lot of works have been continuously proposed to increase the efficiency of extracting MFCC feature. Fully considering the arithmetic property, Jo *et al.* [4] proposed an energy-efficient floating-point MFCC extraction architecture based on field-programmable gate array (FPGA) with the improvement of frequency transformation and optimization of bit-width. Some other works [11], [12] about efficient MFCC extraction are also proposed based on FPGA for low-cost speech recognition systems. In addition, efficient parallel implementation of MFCC feature extraction on graphics processing units (GPU) [13] and digital signal processor (DSP) [14] are presented showing faster extraction than CPU implementation. It has been reported that if the front-end acoustic algorithm executed on dedicated custom hardware, the energy consumption can be reduced prohibitively, and then the battery life can be significantly extended [15]. Therefore, there are also some works focusing on digital application specific integrated circuit (ASIC) [5], [6] realization to achieve higher energy-efficiency and faster processing.

However, all the previous works are implemented in the digital signal domain, where analog-to-digital converter (ADC) would consume much energy to process a large amount of redundant raw data from the microphone [3]–[8]. Besides, the indispensable Fast Fourier Transform (FFT) in conventional digital signal processing realization costs most processing time [13]. Some work [16]–[18] proposed the analog feature extraction method to avoid A/D conversion. Nevertheless, in their work, the simple features extracted in the analog domain are only suitable for simple tasks such as voice activity detection. These simple features also lead to poor recognition accuracy when it comes to the automatic speech recognition application. In summary, ADC and the FFT operation have been the energy consumption and processing speed bottleneck of the entire MFCC feature extraction process.

To achieve more energy-efficient and faster feature extraction for wearable automatic speech recognition, a novel mixed-signal processing architecture to extract MFCC features (MSP-MFCC) is proposed here. As shown in Fig. 1(b), we maintain that it is more natural and faster to remove unnecessary frequency domain transform operations in acoustic features extraction. Moreover, without sampling and quantization, the acoustic features extracted in the analog domain are lossless and free of sampling noise. In this paper, MSP-MFCC is investigated, improved and implemented from the disciplines of architecture, algorithm, and silicon proven:

1) Architecture Techniques: Proposed mixed-signal processing architecture achieves higher efficiency and faster speed than state of the art. Moreover, the ADC bottleneck problem that has been neglected by conventional works is investigated and eliminated in this architecture.
2) Algorithm Techniques: The processing flow of conventional MFCC realization is revised. The proposed time-domain energy distribution extraction method avoids time-consuming and energy-hungry FFT operation.
3) Silicon Verifications: These techniques include the area-saving stair-stepping high-pass filter operation and the framing operation designed for mixed-signal realization. We further study the performance and improve the flexibility of the analog processing circuit according to various real applications. The analog processing parts of MSP-MFCC are fabricated and measured to evaluate the feasibility. According to the experiment results, MSP-MFCC achieves the best performance so far, with 95% energy saving and about 6.4× speedup than state of the art, as well as the comparable recognition accuracy.

The rest of this paper is organized as follows. We introduce the basic theory of the commonly used MFCC algorithm and hardware implementation analysis in Section 2. The detailed description of the MSP-MFCC is presented in Section 3. Section 4 shows the measured performance of MSP-MFCC and the performance comparison with conventional architecture. Section 4 also shows the fabricating results of the essential components in the MSP-MFCC. The conclusion is drawn in the final part of Section 5.

II. MFCC ALGORITHM INTRODUCTION AND HARDWARE IMPLEMENTATION ANALYSIS

A. CONVENTIONAL MFCC EXTRACTING METHOD

The commonly used MFCC extraction process is shown in Fig. 2 [19], including a microphone in the front-end, analog-to-digital converter, and feature extraction in the back-end. The following sections detail the MFCC algorithm and the conventional implementation process. The working mechanism of the human auditory system resembles a set of filters, which could process the acoustic signal at different frequencies. As a kind of feature, the MFCC takes the concern of that property and could describe the signals' energy distribution in the Mel-frequency domain [20].

1) FRONT-END AND DATA CONVERSION

Conventionally, due to processing in the digital domain, an ADC with at least the sampling rate of 16 kHz and the

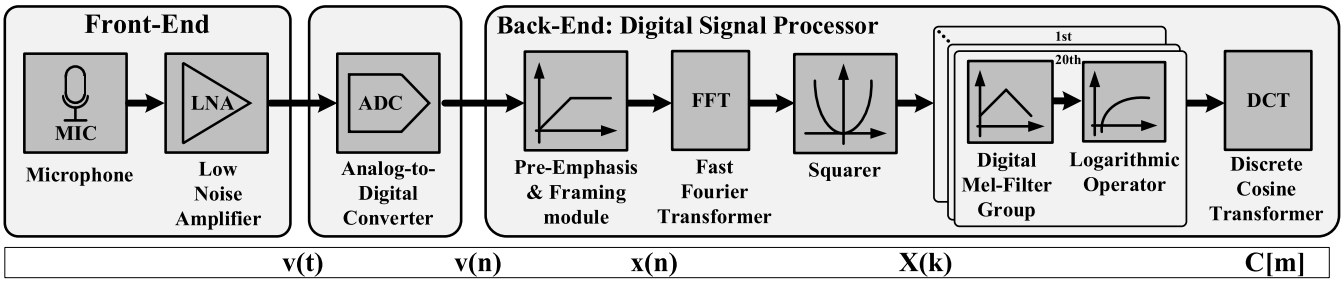

FIGURE 2. Conventional MFCC extraction process [19].

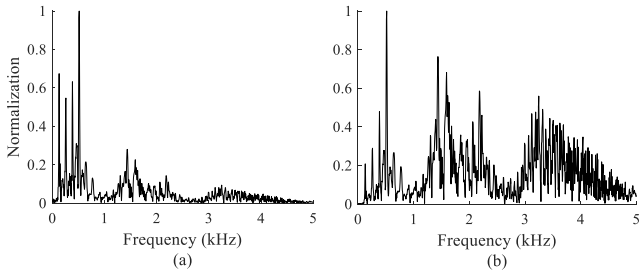

FIGURE 3. Spectrum of speech "6" (a) before (b) after the pre-emphasis.

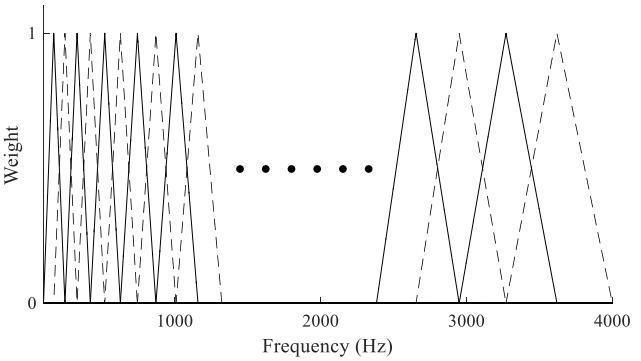

FIGURE 4. Spectrum of Mel-filters with gradually widened bandwidth [19], [20].

precision of 16-bit [3]–[7] [11]–[14] is required to convert the analog speech signals to the digital signals. In the conversion, the continuous input speech signal v(t) is sampled and quantized into the discrete signal v[n].

2) PRE-EMPHASIS AND FRAMING MODULE

In order to compensate the high-frequency damping caused by the blurring effect of the lip, the quantized input voice is then pre-emphasized by a High-Pass Filter (HPF) to balance the amplitudes of low frequency and high frequency. As shown in Fig. 3, the spectral amplitude of the speech signal is well balanced after pre-emphasis. Then the framing operation with the half overlap [4] is performed to keep the invariance of features and the smoothness of the signal in each frame. All subsequent operations are performed frame by frame.

3) FREQUENCY-DOMAIN TRANSFORMATION

In order to extract the energy distribution on the spectrum, FFT is performed to transform the time domain signal into the frequency domain, and square process (1) is applied here to transform the amplitude spectrum of the signal to the energy spectrum.

$$X\,[k] = |FFT\,(x\,[n])\,|^2. \tag{1}$$

4) MEL-FILTERS AND POST-PROCESSING

The spectrum of Mel-filters is shown in Fig. 4. According to the character that the human ear is more sensitive to low-frequency voice than the high one, the bandwidth of Mel-filters widens gradually with the increase of frequency to extract sufficient energy information in the low-frequency bands. The bounds of each filter are calculated by the fixed equation between the frequency and Mel-frequency [21]. The post-processing operations including logarithmic multiplying and Discrete Cosine Transformation (DCT) are performed next to transform the filtered signals to MFCC features. as follows:

$$C\,[m] = \sum_{k=1}^{K} Log(X\,[k])cos(\frac{\pi m\,(k-0.5)}{K}), \tag{2}$$

where m = 1,…, M is the length of MFCC of each frame, X[k] is the output energy of the Kth band, and C[m] is the output feature.

B. HARDWARE IMPLEMENTATION ANALYSIS

The energy-efficient feature extraction process, e.g. the human auditory system, does not have an ADC for speech sampling and quantizing. Obviously, not all raw speech signals from the microphone are equally significant. It is energy-hungry to indiscriminately convert the raw redundant analog speech into a digital signal through a high sampling rate and high precision ADC. More than that, depending on the over-sampling characteristic of the sigma-delta ADC [5], a sampling frequency well above the maximum input frequency is required, which results in more energy consumption.

Conventional work rarely pays attention to the energy consumption of the microphone and ADC front end. However, an ADC with the high sampling rate and the high precision introduces not only significant energy consumption but also redundant data [19]. As shown in Fig. 5(a), it is the comparison of power consumption among modules in the

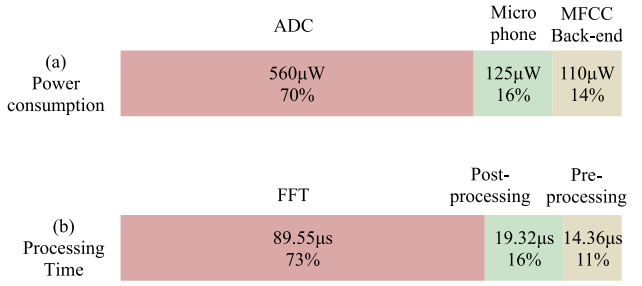

FIGURE 5. (a) Power consumption of the entire MFCC extraction including Microphone [9], ADC [8] and MFCC back-end [5]; (b) Processing time of each part in conventional MFCC extraction back-end [13].

entire feature extraction process, where the values of each module refer to state of the art [5], [8], [9]. It can be seen that the ADC occupies most of the power consumption in the conventional feature extraction process. Unfortunately, as the process advances and the CMOS size shrinks, the Sigma-delta ADC will consume more energy to achieve the same performance [22], [23]. In other words, the ADC bottleneck problem will become increasingly serious.

Besides, as shown in Fig. 5(b), computationally expensive FFT costs almost 73% of the total processing time in MFCC extraction back-end [13]. The physical meaning of MFCC is the energy distribution of the input signal in different frequency bands. For conventional systems, this means performing FFT to convert the time-domain signal to the frequency domain and then squaring the spectrum amplitude. Therefore, the time-consuming FFT is indispensable for conventional realization. That is, in the back end of the traditional MFCC extraction process, FFT transformation becomes the bottleneck. In summary, the key to achieving energy-efficient and fast MFCC feature extraction is eliminating the FFT operation while reducing the processing cost of the ADC.

III. MSP-MFCC ARCHITECTURE AND COMPUTATION UNITS

A. ALGORITHM TECHNIQUES

Actually, instead of FFT, energy distribution can also be extracted directly by filtering, squaring and integrating the input signal in the time domain using a set of band-pass filters and squarers. The FFT is necessary for traditional signal processing considering the convenient signal analysis in the frequency domain. However, the frequency band of the MFCC is fixed [4], [20], the time domain filter configuration does not need to be analyzed and changed after design. Thus, it is reasonable and equally convenient to obtain the energy distribution by filtering, squaring and integrating in the time domain. Moreover, processing the input signal in the time domain has the same result as conventional digital realization, which is explained as follows:

For each filtered frame F_i, its signal $x_i(t)$ and its FFT result $X_i(\omega)$ satisfy the Parseval's theorem [24]:

$$E_i = \int_{-\infty}^{+\infty} |x_i(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |X_i(\omega)|^2 d\omega \qquad (3)$$

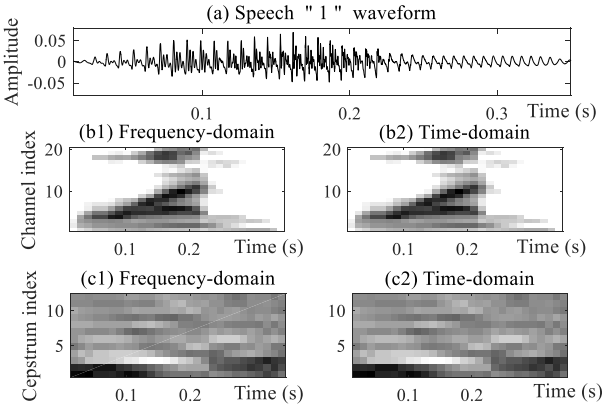

FIGURE 6. (a) Original signal, (b) energy distribution, and (c) MFCC extraction results by frequency-domain (left) and time-domain integration (right).

whereE_i is the energy of frame F_i. That is, integration in the frequency domain is 2π times as in the time domain, which means the energy distribution of the input signal can be calculated by integration in the time domain without frequency transformation. As an example shown in Fig. 6, we extract the energy distribution for the input speech "1" in the time domain and the frequency domain respectively. The extraction results obtained by the two extraction methods are identical. The constant 2π is uninfluential to the MFCC features because DCT operation could transform the constant into the direct-current component of coefficients.

B. ARCHITECTURE TECHNIQUES

The FFT bottleneck has been eliminated by time-domain energy distribution extraction. In order to solve ADC bottleneck problems at the same time, we propose the mixed-signal processing architecture for MFCC feature extraction. The detailed processing flow of MSP-MFCC is shown in Fig. 7, where the operations, including Mel-filters, squaring, and low-pass filter, are put into analog front-end. The principle of architecture is as follows.

1) PRE-EMPHASIS MODULE

After the low noise amplifier, pre-emphasis is necessary to enhance the energy in high frequency. Conventional works perform the pre-emphasis by passing through a high-pass filter, the spectrum of which is shown in Fig. 8(a). As shown in Fig. 8(b), in order to simplify the filtering operation, the stair-stepping high-pass filter operation with increasing gain from low frequency to high frequency is considered here. This modification results in almost no degradation of the recognition accuracy, but greatly simplifies the pre-emphasis implementation in the analog domain.

2) ANALOG MEL-FILTERS MODULE

According to the conventional realization method [20], twenty passbands are chosen here to filter the signal at different frequency bands. That is, twenty band-pass filters (BPFs) should be performed next. Five BPFs with the same gain

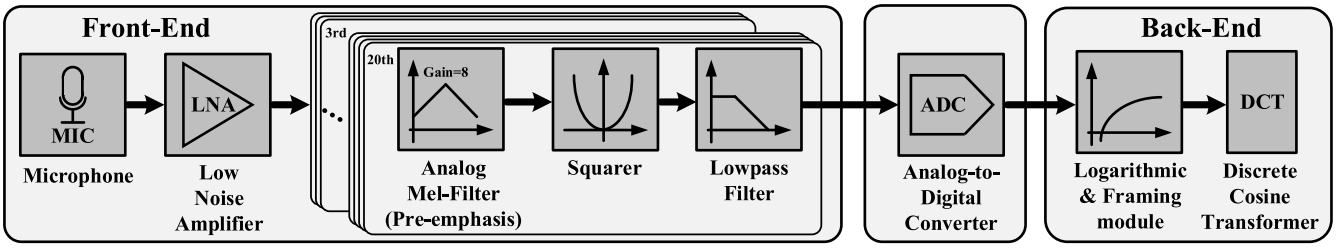

FIGURE 7. Proposed MSP-MFCC (Mixed-signal processing architecture for MFCC feature extraction).

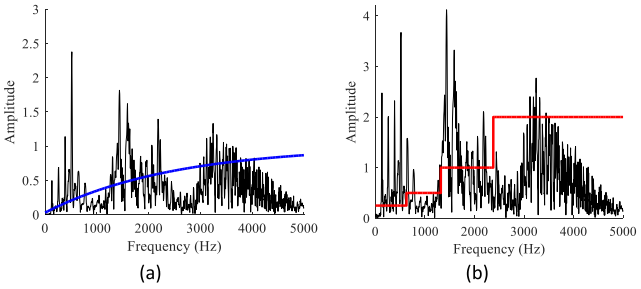

FIGURE 8. Frequency Spectrum of speech "6" (Black line) after (a) conventional smoothing pre-emphasis (Blue line); (b) stair-stepping pre-emphasis (Red line).

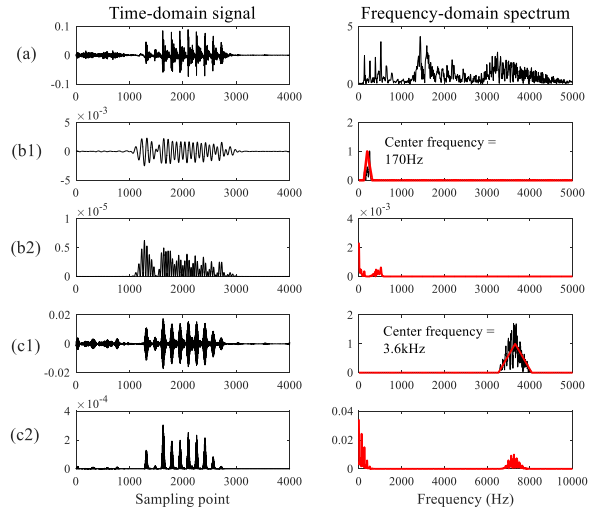

FIGURE 9. Time-domain signals (left) and their spectrums (right). (a) Original signal after pre-emphasis; (b1) Passing through band-pass filters with center frequency of 170Hz then (b2) through squarer; (c1) Passing through band-pass filters with center frequency of 3.6kHz then (c2) through squarer; After the squaring operation, the average energy of the signal is transferred to DC value.

are used here to form a group. Such the four groups with different gain can perform both the band-pass filtering and the stair-stepping high-pass filtering. Thus, no extra circuits are needed in this stair-stepping filter, which will be explained in Section III.C.

3) ENERGY DISTRIBUTION EXTRACTION

As shown in Fig. 9(b1), 9(c1), after passing through the band-pass filters with the center frequency of 170Hz and 3.6kHz, the signals in the corresponding frequency bands are filtered out. For each passband, the filtered signal is then sent to squarer to extract the energy density. Any signal can be decomposed into the sum of multiple single-frequency signals [24]. Therefore, in order to explain the principle of squarer clearly, we analyze the squaring process of a single-frequency signal $Asin(\omega t)$, where A and ω are the amplitude and frequency of the signal. The squaring process is defined as:

$$(Asin(\omega t))^2 = A^2(1 - cos\,(2\omega t))/2 \tag{4}$$

where $A^2/2$ represent the energy density of the signal. The value of DC is the energy density of the signal. That is, the energy density of the signal is extracted by squaring the signal in the form of direct-current (DC) value$A^2/2$. In this way, the average energy value of the signal in this frequency can be extracted simply by low-pass filtering. As shown in Fig. 9(b2), 9(c2), after the squaring operation, original spectrums are moved to the DC and the double frequency positions.

4) DATA CONVERSION

After identical signal processing, twenty energy values of different bands for each frame are extracted without time-consuming FFT operation. In order to realize real-time speech recognition, it is necessary to extract the energy value in a half-frame time. So that the outputs' rate of change is 80Hz because half of the frame duration is 12.5ms. Thus, an ultra-low sampling rate ADC could be adopted here with lower energy consumption compared with the conventional 16 kHz, 16-bit ADC. Besides, compared with conventional 400 samples per frame, the data size here is only 20 samples per frame, which means a significant reduction for the latency of ADC in MSP-MFCC. In summary, extracting the energy distribution on the time-domain signal in the analog domain is a natural, low-cost, and fast way to process the voice signal.

5) POST-PROCESSING

As described in Section II, the input voice signal is pre-emphasized and framed first before being sent into filters. There is half overlap between every two frames, whereas framing operation is hard to do in the analog domain for the lack of analog memory to store continuous overlap. Therefore, a new framing method designed specifically for mixed-signal MFCC feature extraction is proposed. By extracting the energy distribution of the half-frame signal and splicing the distributions of the adjacent half frame, the framing

operation is implemented in the digital domain. For the back-end, Log and DCT operations that are more appropriate for discrete processing are still calculated in the digital domain. The implementation of these operations is the same as the conventional works.

In summary, the preprocessing in the analog domain reduces the dimensionality of the data, which greatly reduces the processing cost and energy consumption of the ADC. At the same time, the analog filter and the squarer are performed to extract the energy spectrum distribution of the signal directly in the time domain, avoiding the time-consuming FFT operation. Thus, the energy distribution extraction in the analog domain is a more natural and efficient way with higher speed and lower cost for its absence of data conversion and frequency-domain transformation. Operations that are difficult to implement in the analog domain, such as framing, logarithmic, and DCT, still run in the digital domain. The MSP-MFCC solves bottleneck problems including energy-hungry ADC and time-consuming FFT in the conventional architecture, and achieves more energy-efficient feature extraction. The proposed architecture is more advantageous in resource-constrained scenarios such as small mobile devices. The detailed implementation of the analog part will be introduced as follows. Besides, another advantage of this architecture is that any part of the frequency band can be turned off according to the actual scene to save energy for processing these frequency bands. It demonstrates flexibility while further reducing energy for a given assignment.

In summary, the preprocessing in the analog domain reduces the dimensionality of the data, which greatly reduces the processing cost and energy consumption of the ADC. At the same time, the analog filter and the squarer are performed to extract the energy spectrum distribution of the signal directly in the time domain, avoiding the time-consuming FFT operation. Thus, the energy distribution extraction in the analog domain is a more natural and efficient way with higher speed and lower cost for its absence of data conversion and frequency-domain transformation. Operations that are difficult to implement in the analog domain, such as framing, logarithmic, and DCT, still run in the digital domain. The MSP-MFCC solves bottleneck problems including energy-hungry ADC and time-consuming FFT in the conventional architecture, and achieves more energy-efficient feature extraction. The proposed architecture is more advantageous in resource-constrained scenarios such as small mobile devices. The detailed implementation of the analog part will be introduced as follows.

C. ANALOG MEL-FILTER GROUP

The structure of the analog Mel-filter group element, which is a compact, energy-efficient, electronically tunable continuous-time BPF, is based on the Capacitively Coupled Current Conveyor (C^4) [25]. All the transistors operate in the sub-threshold region with low-voltage power supply and high current efficiency. As shown in Fig. 10(a), twenty

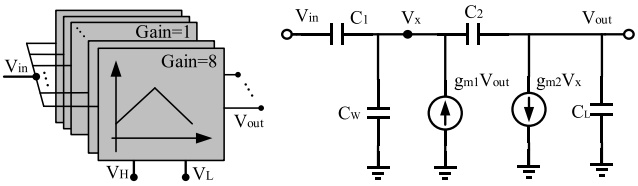

FIGURE 10. (a) The schematic diagram of the Mel-filter group with stair-stepping gain. The center frequency can be electronically tuned by the bias voltage of V_H and V_L. [25]; (b) The simplified small-signal model. From the transfer function of the filter element, the voltage gain is depended on C1 while the center frequency is depended on g_{m1} and g_{m2}.

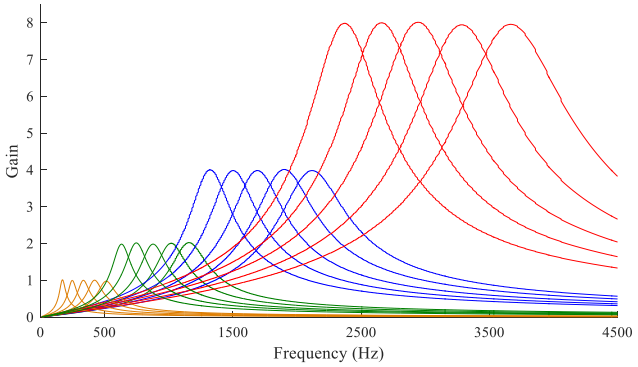

FIGURE 11. The spectrums of 20 tunable filter elements, whose center frequency are exponentially distributed from 170 Hz to 3.6 kHz. The filter elements are divided into 4 groups with gradually increasing gain to simulate the behavior of a stair-stepping high-pass filter.

BPFs with stair-stepping gain extract the voice signal in different bands. Due to the compact structure, both the high and low center frequencies, can be easily tuned by the bias voltages of V_{H} and V_{L}. The simplified small-signal model of this filter element is showed in Fig. 10(b), and the resulting transfer function is set by

$$\frac{V_{out}}{V_{in}} = -\frac{C_1}{C_2}\frac{s\frac{C_2}{g_{m1}}(1 - s\frac{C_2}{g_{m2}})}{s^2\frac{C_oC_T}{g_{m1}g_{m2}} + s\left(\frac{C_2}{g_{m1}} + \frac{C_o}{g_{m2}}\right) + 1} \tag{5}$$

where C_2 is the equivalent capacitance between node V_X and $\mathrm{V}_{\mathrm{out}}$; C_T is defined as total capacitance with $C_T = C_1 + C_w + C_2$; C_o is defined as the output capacitance with $C_o = C_2 + C_L$; g_{m1} and g_{m2} are transconductance parameters. In normal cases, the time constant of positive zero $\tau_f = C_2/g_{m2}$ is quite small and can be neglected in the transfer function.

The passband voltage gain A_v is set by

$$A_v = -\frac{C_1}{C_2}\frac{1}{1 + \frac{g_{m1}}{g_{m2}}\frac{C_o}{C_2}} \tag{6}$$

The center frequency f_c is set by

$$f_c = \frac{\sqrt{g_{m1}g_{m2}}}{2\pi\sqrt{C_oC_T}} \tag{7}$$

By scaling up or down g_{m1} and g_{m2} proportionally while keeping C_o and C_T same and by scaling up or down C_1 while keeping C_T same, the f_c and the A_v of each filter can be tuned respectively without changing the quality factor.

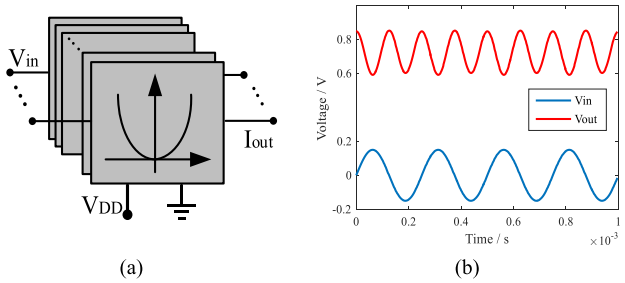

FIGURE 12. (a) The schematic diagram of the square. Within the input range of the squarer, the squaring relationship between input and output has nothing to do with V_{DD} [26]; (b) Simulation results of squarer. Vin is a 4kHz 0.4Vpp sine signal. The DC value and amplitude of the output signal is proportional to the square of the input signal amplitude.

In this paper, the Mel-filter group has 20 filter elements that are divided into 4 groups. The f_c of each element is exponentially distributed from 170 Hz to 3.6k Hz. Each group has a different A_v, e.g., $\times 1$, $\times 2$, $\times 4$ and $\times 8$, to imitate the behavior of stair-stepping high-pass filter and conduct pre-emphasis on the acoustic signals of different frequencies. Fig. 11 shows the simulated frequency response of the proposed Mel-filter group. The BPFs implemented by analog circuit are difficult to achieve ideal cutoff characteristic under the low power consumption constraints. For a specific BPF, this will cause the attenuated signals in other frequency bands to be extracted. However, the center frequency skewing of BPFs can be alleviated in this case. Even if the center frequency is shifted, the undesired cutoff characteristic causes the original band signal to be extracted with attenuation. The influences to recognition accuracy of this effect will be discussed in Section 4.1.

D. ANALOG SQUARE OPERATION

The structure of the analog square circuit is referred to [26]. The schematic diagram of the squarer element is shown in Fig. 12(a), where the output current I_{out} is in proportion to the square of V_{in} as:

$$I_{out} = 2K_N \left(\frac{V_{in}}{2}\right)^2 = \frac{K_N}{2}(V_{in})^2 \tag{8}$$

where K_N is the transconductance parameter of transistor. Within the input range of the squarer, the squaring relationship between input and output has nothing to do with the supply voltage V_{DD} [26]. Thus, energy consumption can be slashed by reducing the V_{DD} while having a negligible effect on squaring accuracy. Besides, the structure is free of body effect, and the channel length modulation effect can be degraded by using long channel devices. The squaring result is shown in Fig. 12(b) where, as an example, a 4k Hz sinusoidal signal $A\ sin\ \omega t$ is sent to the squaring circuit and the output signal is equal to:

$$\frac{K_N}{2}(A\ sin\ \omega t)^2 = \frac{K_N A^2}{4}(1 - cos\,(2\omega t)) \tag{9}$$

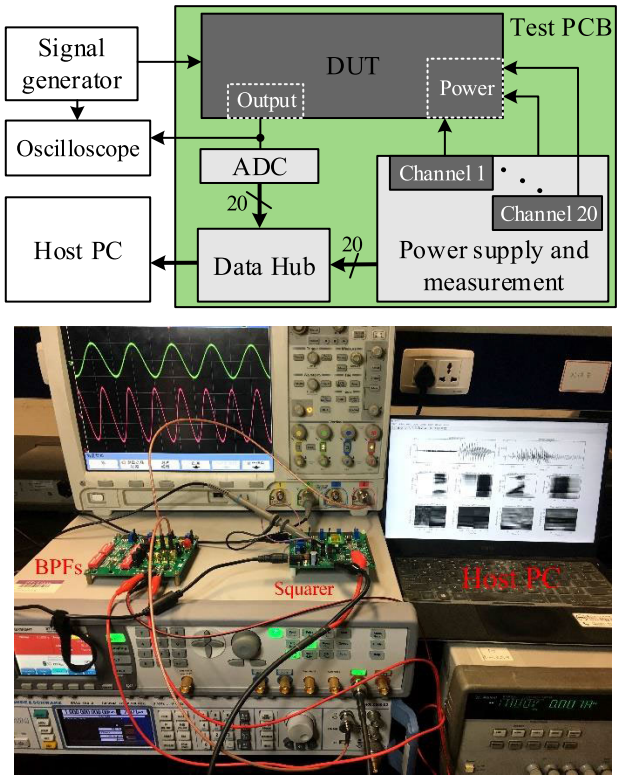

FIGURE 13. Block diagram of test PCB (top) and photograph of testing environment (bottom).

where $\omega = 8000\pi\, rad/s$ in this example. That is, the average value of output, $K_N A^2/4$, which is also equivalent to the average energy of input, is proportional to the input amplitude. The output current is then filtered by a current-mode low-pass filter [27] to extract the value of average energy. It is converted to the voltage by the resistor for the convenience of measurement. The constant $K_N/4$ would be transformed into the constant addition in the Log phase and eliminated in the DCT operation.

IV. THE MEASURED PERFORMANCE AND COMPARISON

The front-end processing units of the proposed mixed-signal domain MFCC extractor are fabricated in 180nm CMOS mixed-signal process. To evaluate the performance of extracted features, the speech recognition task with the well-known dataset TI-DIGITS [28] and the Long-Short-Term Memory (LSTM) neural network [29], [30] are adopted. Each utterance of TI-DIGITS is a single spoken number from "zero" to "nine" and an extra "Oh". The speech recognition is realized on Tensorflow R1.0 deep learning development platform.

The test setup of MSP-MFCC's analog part is shown in Fig. 13. The power supply and measurement module can power one set of bandpass filters and squares and measure their respective power consumption. The voice of each frequency band is separately simulated by the signal generator, and the output signal is converted into the digital domain by the on-chip ADC. The output signal and the power consump-

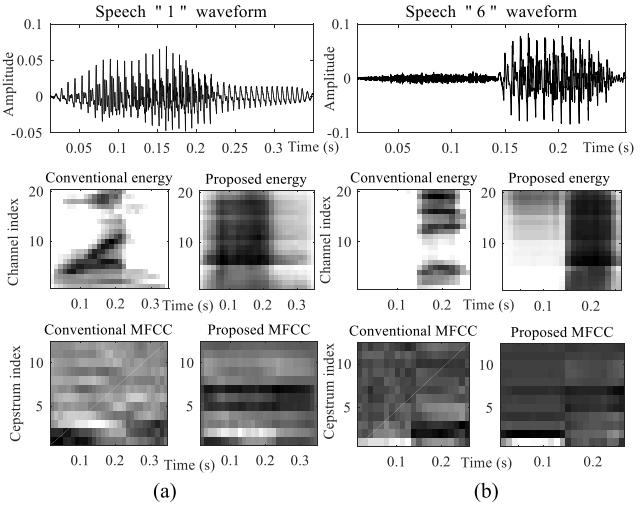

FIGURE 14. Energy distribution and extracted MFCC features for (a) speech "1" and (b) speech "6". The undesired cutoff characteristic of BPFs introduces out-of-band energy and results in the blur of the energy distribution, which have negligible impact on recognition accuracy.

tion value of each module are sent to the host PC for back-end digital processing verification.

A. SYSTEM FUNCTION

The energy distribution and the MFCC features of number "1, 6" that extracted by MSP-MFCC and conventional architecture are shown in Fig. 14. The abscissa is the time while the ordinates are amplitude and the number of frequency passband respectively. Due to the undesired cutoff characteristic of analog BPFs, out-of-band energy is extracted, which introduces the difference between these two results. Nevertheless, the speech recognition system can reach the recognition accuracy of 98.9% (for the 16-bit width) when using the extracted MFCC features. It hardly decreases compared to the conventional implementation and proves that the difference is uninfluential to the recognition. Some conventional works [4], [11], [12] compare their recognition accuracy under different algorithms and even different datasets. It is unreasonable because the increase in accuracy may not be caused by feature improvement, but by better algorithms or datasets. In order to ensure the objectivity of the results, we compare with the work [29], [30] under the same algorithm and dataset. The result shows that MSP-MFCC achieves a comparable recognition accuracy with only 0.1% decrease.

The bit-width of MFCC features influences the recognition accuracy of the speech recognition system, which also determines the precision of ADC. Thus, the relationship between the bit-width of proposed MFCC features and the corresponding recognition accuracy is given here. As shown in Fig. 15, with the increase of bit-width, the recognition accuracy also increases. 8-bit width with the accuracy of 98.2% is a proper point to reach the balance between accuracy and processing cost.

B. DETAILS OF THE FABRICATED CHIP

The main parts of the architecture are the Mel-filter group and the analog square circuits. As the detailed description of the

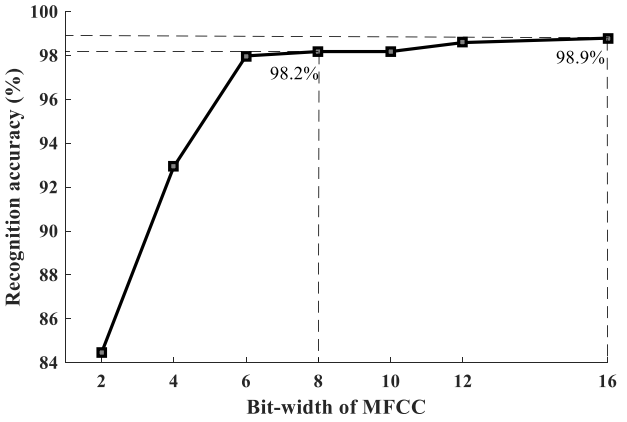

FIGURE 15. Recognition accuracy versus the bit-width of MFCC. 8-bit width is a proper point to balance the accuracy and processing cost.

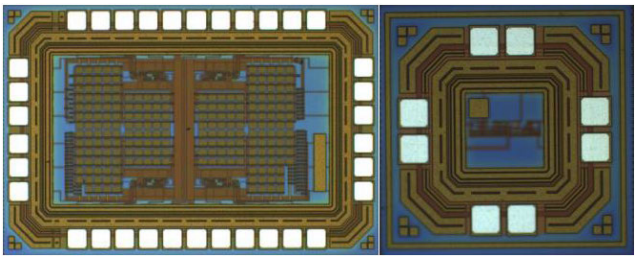

FIGURE 16. Microphotography of the Mel-filters (left) and analog squarer (right).

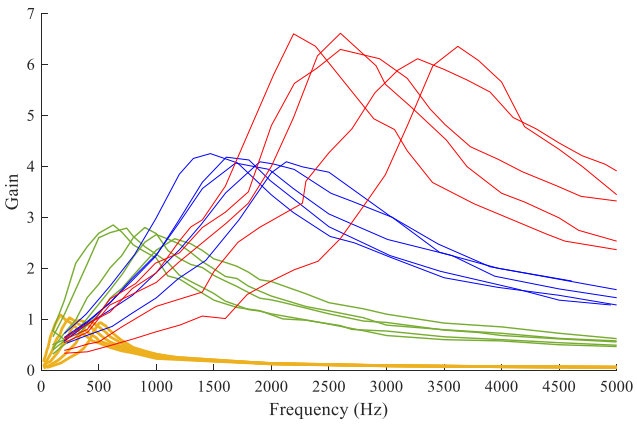

FIGURE 17. The measured results of 20 filter frequency response with their central frequency range from170Hz to 3.6 kHz.

circuits design in Section III, we taped out the circuits design and evaluated the frequency response and power consumption of the Mel-filter group and the analog squarer, the microphotography of which are shown in Fig. 16. The active area of single Mel-filter and squarer are 0.05 mm^2 and 0.001 mm^2 respectively.

Fig. 17 shows the measured spectrum of Mel-filters with the frequency ranging from 0 to 5000Hz. The gains and gradually widened bandwidth are not perfectly correspond to the simulation results because of the non-ideal factors in fabrication. Nevertheless, the extracted MFCC achieves the leading performance in recognition accuracy. The measured power consumption of all filters in the different groups

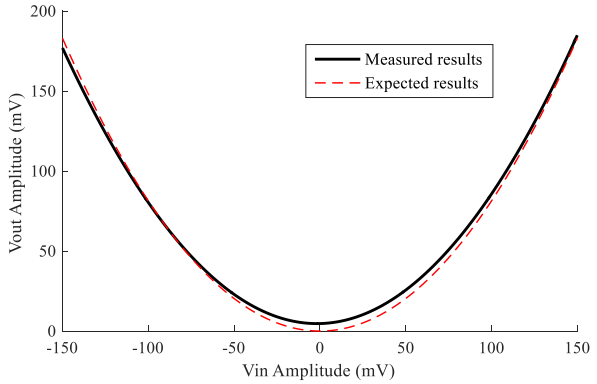

FIGURE 18. The measured and expected transfer characteristic of squarer.

TABLE 1. Comparison with state-of-the-art MFCC extractor with ADC considered.

		This work	[4]	[5]	[12]
Process		**CMOS 180nm mixed signal**	FPGA	Digital ASIC 65nm	DSP
Computation bit-width		**8-bit**	14-bit	16-bit	32-bit
Sampling rate each band		**160 Hz**	16 kHz	16 kHz	8 kHz
Power consumption	ADC	**7.4 μW** [31]	NC*	NC*	NC*
	Core (μW)	**21.4**	188.4	110	51800
Speed# (μs/frame)		**45.79**	292	125	927.78
Accuracy		**98.2%**	96.1%	98.4%	93.3%

NC* Energy consumption of ADC is not considered. # Processing latency of the digital part.

is 3.92 μW. The measured and expected transfer characteristic of analog squarer is shown in Fig. 18. It consumes 0.87 μW with the input range of 300mV.

C. PERFORMANCE SUMMARY AND COMPARISON

Table 1 summarizes the performance of the MSP-MFCC and compares it with conventional architectures. It is notable that energy consumption here includes the data converter and the back-end circuits, from ADC to DCT. According to the computation cost model of conventional MFCC extraction [13], [14], the energy consumption and efficiency of the digital back-end are calculated, which show the improvement of energy efficiency for eliminating expensive frequency-domain transformation. The computation bit-width and sampling rate of each system determine the type of ADC according to the references [8], [31].

The measured results show the 21.4μW power consumption of analog-part MSP-MFCC. The simulated processing latency of the digital part is 45.79 μs/frame. Considering the ADC power consumption of 7.4μW [31], the energy consumption of MSP-MFCC is 0.72 μJ for the frame length of 25ms. Taking into account the 560μW power consumption of the ADC [8], the total energy consumption of the conventional feature extraction system [4] is 14.06μJ/frame. Therefore, MSP-MFCC reaches 95% energy saving and 6.4× speedup than state of the art [4]. To our best knowledge, this is the best performance ever reported for entire MFCC feature extraction.

Besides, the proposed MFCC extraction reaches 98.2% accuracy when used in speech recognition, which also keeps the leading performance to its current digital counterparts. Proposed MSP-MFCC solves bottleneck problems including energy-hungry ADC and time-consuming FFT, and achieves more energy-efficient feature extraction.

V. CONCLUSION

This paper proposes an energy-efficient architecture, MSP-MFCC, using mixed-signal domain information processing to reduce energy consumption and improve the processing speed of MFCC extraction. The bottleneck problems of ADC and FFT in the entire extraction process are solved with architecture and computing paradigm improvement. MSP-MFCC achieves 95% energy saving and about 6.4× speedup than state of the art. To our best knowledge, this is the best performance ever reported for entire MFCC feature extraction. Fabricated results show the proposed MFCC extraction reaches 98.2% recognition accuracy when used in automatic speech recognition tasks. In summary, MSP-MFCC draws techniques from the disciplines of architecture, algorithm, and silicon proven:

A. MIXED-SIGNAL PROCESSING ARCHITECTURE

Proposed mixed-signal processing architecture investigates and solves the ADC bottleneck problem that has been neglected by conventional works. As a result, it achieves significant energy saving than state of the art. This architecture can also be applied to applications where the ADC occupies most of the system's energy consumption.

B. TIME-DOMAIN PROCESSING ALGORITHM

Processing flow of conventional MFCC realization is revised. MSP-MFCC extracts energy distribution in the time-domain without time-consuming FFT operation. This design idea can also be applied to processes where domain transformation takes up most of the processing time.

C. SILICON VERIFICATIONS

These techniques include the area-saving stair-stepping high-pass filter operation and the framing operation designed for mixed-signal realization. We further study the performance and improve the flexibility of the analog processing circuit to real applications.

The proposed architecture not only accelerates MFCC extraction process but also consumes ultra-low energy. The proposed MFCC extractor is suitable for integration in various types of ultra-low-power always-on wearable speech recognition system. The performance is also sufficient for the always-on speech-controlled wearable applications. The mixed-signal processing architecture and design ideas can be extended to other applications where sensing computing

interface is the bottleneck of the system, which should be watched carefully.

ACKNOWLEDGMENT

This article was presented at the IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), 2018.

REFERENCES

[1] B. Liu, H. Qin, Y. Gong, W. Ge, M. Xia, and L. Shi, "EERA-ASR: An energy-efficient reconfigurable architecture for automatic speech recognition with hybrid DNN and approximate computing," *IEEE Access*, vol. 6, pp. 52227–52237, 2018.

[2] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[3] M. Price, J. Glass, and A. P. Chandrakasan, "A low-power speech recognizer and voice activity detector using deep neural networks," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 66–75, Jan. 2018.

[4] J. Jo, H. Yoo, and I.-C. Park, "Energy-efficient floating-point MFCC extraction architecture for speech recognition systems," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 2, pp. 754–758, Feb. 2016.

[5] M. Price, J. Glass, and A. P. Chandrakasan, "A 6 mW, 5,000-word real-time speech recognizer using WFST models," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 102–112, Jan. 2015.

[6] O. Cheng, W. Abdulla, and Z. Salcic, "Hardware–Software codesign of automatic speech recognition system for embedded real-time applications," *IEEE Trans. Ind. Electron.*, vol. 58, no. 3, pp. 850–859, Mar. 2011.

[7] A. P. Vinod, E. M.-K. Lai, P. K. Meher, J. Palicot, and S. Mirabbasi, "Guest editorial: Special issue on embedded signal processing circuits and systems for cognitive radio-based wireless communication devices," *Circuits, Syst., Signal Process.*, vol. 30, no. 4, pp. 683–688, 2011.

[8] F. Cardes, E. Gutierrez, A. Quintero, C. Buffa, A. Wiesbauer, and L. Hernandez, "0.04-mm^2 103-dB-a dynamic range second-order VCO-based audio $\Sigma\Delta$ ADC in 0.13-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 53, no. 6, pp. 1731–1742, Jun. 2018.

[9] Knowles. *Microphones*. Accessed: 2019. [Online]. Available: https://www.knowles.com/subdepartment/dpt-microphones/subdpt-sisonic-surface-mount-mems

[10] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2Letter: An end-to-end ConvNet-based speech recognition system," 2016, *arXiv:1609.03193*. [Online]. Available: http://arxiv.org/abs/1609.03193

[11] N.-V. Vu, J. Whittington, H. Ye, and J. Devlin, "Implementation of the MFCC front-end for low-cost speech recognition systems," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 2334–2337.

[12] P. EhKan, T. Allen, and S. F. Quigley, "FPGA implementation for GMM-based speaker identification," *Int. J. Reconfigurable Comput.*, vol. 2011, pp. 1–8, 2011.

[13] H. Kou, W. Shang, I. Lane, and J. Chong, "Optimized MFCC feature extraction on GPU," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7130–7134.

[14] J. Manikandan, B. Venkataramani, K. Girish, H. Karthic, and V. Siddharth, "Hardware implementation of real-time speech recognition system using TMS320C6713 DSP," in *Proc. 24th Internatioal Conf. VLSI Design*, Jan. 2011, pp. 250–255.

[15] A. Raychowdhury, C. Tokunaga, W. Beltman, M. Deisher, J. W. Tschanz, and V. De, "A 2.3nJ/frame voice activity detector based audio front-end for context-aware system-on-chip applications in 32 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1963–1969, Aug. 2013.

[16] M. Cho, S. Oh, Z. Shi, J. Lim, Y. Kim, S. Jeong, Y. Chen, D. Blaauw, H.-S. Kim, and D. Sylvester, "17.2 a 142nW voice and acoustic activity detection chip for mm-scale sensor nodes using time-interleaved mixer-based frequency scanning," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 278–280.

[17] M. Yang, C.-H. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "A 1μW voice activity detector using analog feature extraction and digital deep neural network," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 346–348.

[18] K. M. H. Badami, S. Lauwereins, W. Meert, and M. Verhelst, "A 90 nm CMOS, 6 μW power-proportional acoustic sensing frontend for voice activity detection," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 291–302, Jan. 2016.

[19] Q. Li, H. Zhu, F. Qiao, Q. Wei, X. Liu, and H. Yang, "Energy-efficient MFCC extraction architecture in mixed-signal domain for automatic speech recognition," in *Proc. 14th IEEE/ACM Int. Symp. Nanosc. Archit. (NANOARCH)*, Jul. 2018, pp. 1–3.

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[21] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proc. SPECOM*, vol. 1, 2005, pp. 191–194.

[22] T.-C. Wang, Y.-H. Lin, and C.-C. Liu, "A 0.022 mm^2 98.5 dB SNDR hybrid audio $\Delta\Sigma$ modulator with digital ELD compensation in 28 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 11, pp. 2655–2664, Nov. 2015.

[23] Y. H. Leow, H. Tang, Z. C. Sun, and L. Siek, "A 1 V 103 dB 3rd-order audio continuous-time $\Delta\Sigma$ ADC with enhanced noise shaping in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 51, no. 11, pp. 2625–2638, Nov. 2016.

[24] G. B. Arfken and H. J. Weber, "Mathematical methods for physicists," *Amer. J. Phys.*, vol. 67, p. 165, 1999.

[25] D. W. Graham, P. E. Hasler, R. Chawla, and P. D. Smith, "A low-power programmable bandpass filter section for higher order filter applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 6, pp. 1165–1176, Jun. 2007.

[26] I. Chaisayun, S. Piangprantong, and K. Dejhan, "Versatile analog squarer and multiplier free from body effect," *Anal. Integr. Circuits Signal Process.*, vol. 71, no. 3, pp. 539–547, Jun. 2012.

[27] C. Bartolozzi, S. Mitra, and G. Indiveri, "An ultra low power current-mode filter for neuromorphic systems and biomedical signal processing," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, Nov. 2006, pp. 130–133.

[28] R. G. Leonard and G. R. Doddington. *TIDIGITS LDC93S10. Web Download*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[29] A. Graves, N. Beringer, and J. Schmidhuber, "Rapid retraining on speech data with LSTM recurrent networks," IDSIA, Manno, Switzerland, Tech. Rep. IDSIA-09-05, 2005.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*. New York, NY, USA: ACM, 2006, pp. 369–376.

[31] C. Weltin-Wu and Y. Tsividis, "An event-driven, alias-free ADC with signal-dependent resolution," in *Proc. Symp. VLSI Circuits (VLSIC)*, Jun. 2012, pp. 28–29.

QIN LI received the B.S. degree from Tsinghua University, China, in 2016, where he is currently pursuing the Ph.D. degree with the Nanoscale Integrated Circuits and Systems (NICS) Laboratory. His current research interests include mixed-signal IC design for energy-efficient computing and acoustic signal processing.

YUZE YANG received the B.S. degree from Nanjing Tech University, China, in 2018. She is currently pursuing the M.Sc. degree in IC design engineering with The Hong Kong University of Science and Technology. She has been a Research Assistant with the Nanoscale Integrated Circuits and Systems Laboratory, Tsinghua University, China, since January 2018, where she finished this work. Her research interests include mixed-signal IC design, high-performance CMOS, and VLSI IC design for ALU.

TIANXIANG LAN is currently pursuing the bachelor's degree with the School of Information and Electronics, Beijing Institute of Technology, China. He has been a Research Assistant with the Nanoscale Integrated Circuits and Systems Laboratory, Tsinghua University, China, since May 2018. His main research interests include analog IC design and acoustic signal processing.

HUIFENG ZHU (Student Member, IEEE) received the B.S. degree (engineering) in information engineering (instrumentation science and optoelectronic engineering) from Beihang University, China, in July 2017. He is currently pursuing the Ph.D. degree with the Department of Electrical and System Engineering, Washington University in St. Louis, USA. He has been an undergraduate Research Assistant with the Nanoscale Integrated Circuits and Systems Laboratory, Tsinghua University, China, since September 2015.

QI WEI received the Ph.D. degree from Tsinghua University, Beijing, China, in 2010. He is currently an Assistant Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include MEMS inertial sensors, application-specific integrated circuit (ASIC) design, and high-performance data converters.

FEI QIAO (Member, IEEE) received the bachelor's degree from Lanzhou University, China, in 2000, and the Ph.D. degree from Tsinghua University, China, in 2006. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. He has authored around 90 articles. He holds over 30 invented patents. His research interests include low-power circuits design, and energy-efficient integrated perception circuits and systems for intelligent robots, wearables, and the IoT devices.

XINJUN LIU (Member, IEEE) received the Ph.D. degree in mechanical engineering from Yanshan University, Qinhuangdao, China, in 1999. He is currently a Professor with the Department of Mechanical Engineering, Tsinghua University, China. He was an Alexander von Humboldt Research Fellow with the University of Stuttgart, Germany, from 2004 to 2005. His research interests include robotics, parallel mechanisms, and parallel kinematic machines.

HUAZHONG YANG (Fellow, IEEE) received the B.S. degree in microelectronics and the M.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, in 1989, 1993, and 1998, respectively. In 1993, he joined the Department of Electronic Engineering, Tsinghua University, where he has been a Full Professor, since 1998. He is currently a Specially Appointed Professor with the Cheung Kong Scholars Program. He has authored or coauthored over 400 technical articles. He holds over 100 granted patents. His current research interests include wireless sensor networks, data converters, nonvolatile processors, and energy-harvesting circuits.