Event-Driven Predictive Sampling Circuits for Speech Signal Sensing and Processing

Brandon Gresham, Josh Blowers, Steven Sandoval, and Wei Tang
Klipsch School of Electrical and Computer Engineering
New Mexico State University, Las Cruces, New Mexico, 88003, USA

***Abstract*—This paper presents a novel event-driven speech signal sensing, pre-processing, and compression method using Dynamic Predictive Sampling. The Dynamic Predictive Sampling method converts the input analog waveform into a non-uniform sampled event sequence with both amplitude and timestamp data of each event, which is the turning point of the analog signal. The event sequence can be reconstructed without losing the morphology of the input analog signal. Since the selection of turning points is performed during the analog-to-digital conversion process, the circuit generates much less data throughput. This paper studies the trade-off between the compression factor and the performance in speech recognition accuracy of the proposed method. Based on the simulation result, the total data throughput can be reduced by 87% while keeping the quality of the speech signal for speech recognition. An integrated circuit of Dynamic Predictive Sampling has been designed and simulated for the speech sensing task. The proposed method saves computing overhead and data throughput, which is ideal for future low-power embedded voice recognition systems.**

***Index Terms*—event-driven sensing, nonuniform sampling, dynamic predictive sampling, speech signal sensing.**

I. Introduction

Voice recognition is the most natural type of human-computer interaction and has seen rapid progress in recent years [1]. The primary function of speech recognition is to match the input audio analog waveform with one or more words in a vocabulary. Conventionally, this is performed by using the features extracted from the analog waveform, such as timing, frequency, amplitude, and phase. Most of the current industrial solutions, such as Amazon's Alexa, Apple's Siri, Google Assistant, and Microsoft's Cortana all require high computing overhead, which heavily relies on cloud computing and real-time data communication. As the application scenarios expand to broader areas, there is a need for local voice recognition systems. For example, a driver may try to interact with the computer of the car when driving without a reliable communication link. Recently, processing at the edge become a popular research direction because it provides important advantages in terms of energy efficiency, latency, security, privacy, and autonomy compared to cloud computing [2]. Another example of local processing is Keyword Spotting (KWS), which must always remain active to detect the pre-defined keywords in real-time to wake up the entire system, which requires ultra-low power as it usually has a limited battery power supply.

A typical KWS system contains an analog amplifier, analog-to-digital converter, feature extractor, and classifier. Various efforts have been made to reduce the power consumption for the feature extractor [4] and classifiers such as Depthwise Separable Convolutional Neural Network DSCNN [5], Binarized Neural Network (BNN) [6] and Long Short-Term Memory (LSTM) accelerator [7]. Compact RNN methods can achieve acceptable detection accuracy, which is usually above 80% [8]. The power cost of KWS is usually in the level of 10-20 μW [9], sometimes even below 10 μW [10]. CNN-based KWS is also a very popular solution, it usually has a few convolutional (Conv) layers and a few fully connected (FC) layers. Google speech command dataset (GSCD) is widely applied for training and validation [11]. Recurrent Neural Network (RNN) including long short-term memory and gated recurrent unit (GRU) is good at handling sequential data for speech and sound [12]. Error-resilient signal processing such as approximate computing can be applied to achieve power reduction while minimizing accuracy loss [13], [14]. The model is usually trained using Google Speech Command Dataset (GSCD). Typically 12 classes and 10 keywords are in the vocabulary for prediction. They include 10 keywords: "down", "go", "left", "no", "off", "on", "right", "stop", "up", "yes", together with "silence" as well as "unknown" class representing the other 25 keywords in the GSCD dataset [10]. On the hardware sensor side, conventional acoustic signal processing can be categorized into digital-intensive or analog-intensive methods, as shown in Fig. 1 (a) and (b). conventional Nyquist-rate sampling may generate unnecessary data that overload the signal processing devices [15] and increase the necessary system power. Level-crossing sampling may introduce insertion and deletion errors, which introduce drifts in the reconstructed waveform [16]. Recently, a Dynamic Predictive Sampling method has been proposed [17], which is based on predicting the digital value using slopes. Because this method only records the digital value of the unsuccessful

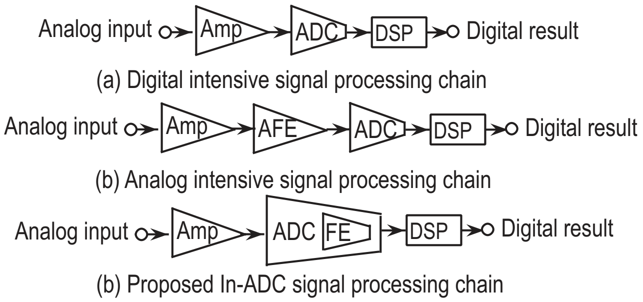
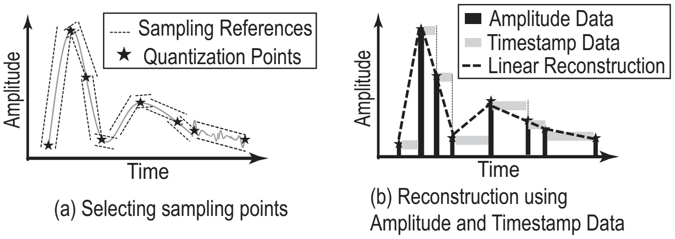

Analog input
Amp
ADC
DSP
Digital result
(a) Digital intensive signal processing chain
Analog input
Amp
AFE
ADC
DSP
Digital result
(b) Analog intensive signal processing chain
Analog input
Amp
ADC
FE
DSP
Digital result
(b) Proposed In-ADC signal processing chain

Fig. 1. Comparing the conventional processing chain (a),(b) modified from [3] with the proposed method (c) in the speech signal sensing systems.

predictions, which are the turning points in the analog signal, the Dynamic Predictive Sampling system can greatly reduce the amount of output data. The feature extraction (FE) can be partially performed in the ADC, as shown in Fig. 1 (c). This paper applies the Dynamic Predictive Sampling method in speech signal sensing and studies the trade-off between data throughput and performance of the reconstructed signal.



(a) Selecting sampling points

(b) Reconstruction using Amplitude and Timestamp Data

Fig. 2. Dynamic Predictive Sampling uses the prediction with an error threshold as the sampling reference to select the turning points for digitization (Left), the output of Dynamic Predictive Sampling contains both the amplitude and timestamp data of the selected sampling points, which can be used for reconstruction of the analog waveform (right).

II. Circuit and System Design

Dynamic Predictive Sampling processes the analog waveform by using prior digitized sampling points to generate a prediction of current sampling values. It then decides whether to perform digitization of the current sampling point based on the comparison result between the current sampling and the prediction of the current sampling. By doing so, Dynamic Predictive Sampling only selects the turning point in the analog waveform to perform digitization. A digital prediction error is introduced as $Delta$, which is added to and subtracted from the predicted digital value to create the upper threshold and lower threshold for comparison. These two digital thresholds are then converted into analog values to compare with the current analog sampling value.

If the prediction is successful, the predicted digital value is used for the next linear prediction until an analog sampling value is higher than the upper threshold or lower than the lower threshold. In such a case, the prediction is considered as failed and digitization is necessary for the sampling, which is marked as the selected sampling point, as shown in Fig. 2 (a). The timing information between the selected sampling points is recorded as a timestamp. These selected sampling points are usually turning points in the input analog waveform. The output of the Dynamic Predictive Sampling contains both the amplitude of the turning points and the timestamp between the turning points. Using the amplitude and timestamp data, the input analog waveform can be reconstructed using piecewise linear reconstruction or other advanced reconstruction methods, as shown in Fig. 2 (b).

Fig. 3 presents the circuit block diagram of a Dynamic Predictive Sampling System. The circuit contains an analog comparator, a digital-to-analog converter (DAC), a prediction logic, a decision logic, and a digital timer. The circuit operates similarly with a successive approximation register (SAR) ADC. The analog portion of the circuit (Comparator and DAC) is controlled by a sampling clock, which is 8kHz for speech signal sensing while the digital portion of the circuit is operated by a processing clock running at 100kHz, which is much faster than the sampling clock in the case that a SAR digitization is necessary. The analog input signal is always compared with the analog reference value generated from the DAC. The comparison result is processed by the prediction logic and decision logic to generate the digital prediction for the DAC. The prediction logic also uses the digital error $Delta$ as the input to calculate the upper and lower threshold of the prediction. If the prediction is not successful, the prediction and decision logic will be switched to a 10-bit SAR logic to perform digitization. At this moment the digitized sampling becomes the data output. The timer is also reset to zero and starts counting the timestamp between the selected sampling point and the next selected sampling point. The prior timestamp is stored in the timestamp register and sent to the Timestamp Output.

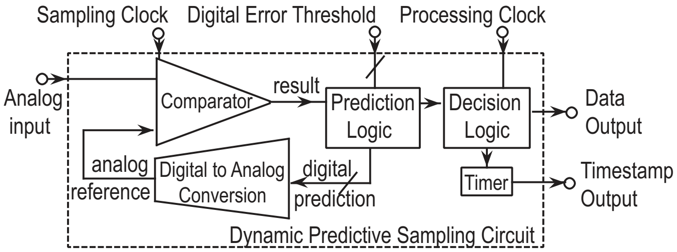
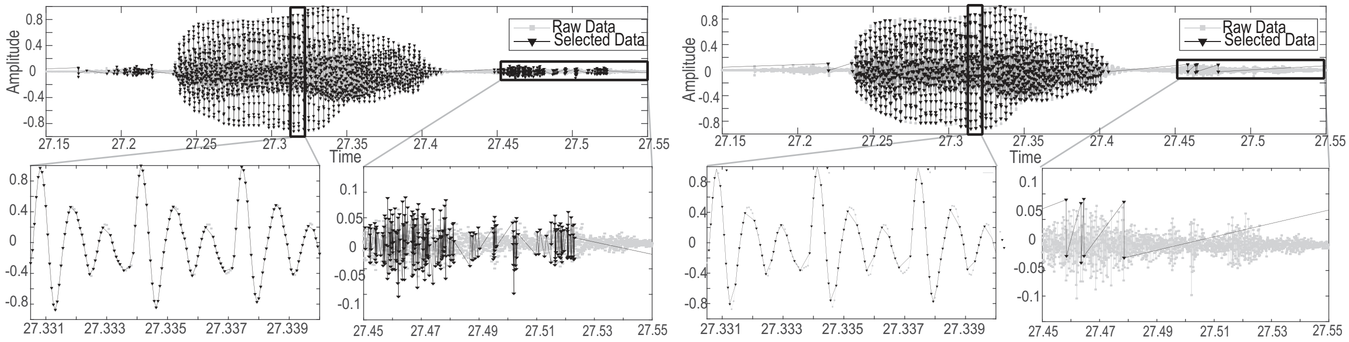


Fig. 3. Block diagram of the Dynamic Predictive Sensing Circuit.

The Dynamic Predictive Sampling method combines the advantages of Nyquist sampling and Level Crossing sampling. Compared with the conventional level-crossing sampling, the Dynamic Predictive Sampling method records multi-bit digital data of the amplitude and digital data of the time-stamp between the sampling events, which removes the potential insertion/deletion of pulses and drift errors. Compared with the conventional Nyquist sampling method using a fixed clock, the dynamic predictive sensing method separates the sampling and quantization processes and only performs quantization at the turning points of the input analog signal, which saves much data throughput when the input signal is sparse or contains a large portion of linear structure. The additional digital error input provides a trade-off between data throughput and the accuracy of the reconstructed signal. It also saves power consumption since many quantization steps can be skipped when the prediction is successful. Therefore, dynamic prediction sampling is ideal for speech signal sensing applications.

III. Simulation and Implementation Results

Applying Dynamic Predictive Sampling in speech signal sensing saves much power and data throughput. The speech signal contains the active voice portion and the pause/noise portion. The voice portion is in the frequency band of 300-3400 Hz with a higher amplitude than the pause/noise portion. The pause/noise portion contains a wider bandwidth with a lower amplitude. In a conventional speech sensing system, the sampling rate is 8k Sample/second or 16 kSample/second. Each sampling is digitized to 8-16 bits. Therefore, the data throughput is between 64kb/second and 256kb/second. Such a high data throughput requires a non-trivial power for the



(a) Selected Sampling Points when error threshold is 2% of amplitude

(b) Selected Sampling Points when error threshold is 5% of amplitude

Fig. 4. Time-domain simulation results of selected sampling points with different error thresholds.

ADC and adds the processing burden to the following digital signal processing or data communication circuits. In reality, it is not necessary to sample and digitize the analog signal during speech pause. This can be achieved using Dynamic Predictive Sampling with a higher error threshold. In addition, during an active speech, the sampling and digitization can be reduced if we can use Dynamic Predictive Sampling to identify turning points of the signal and use the piece-wise linear signal to replace the Nyquist sampling sequence. Since digitization consumes much more power than sampling, this method can save both power and data throughput.

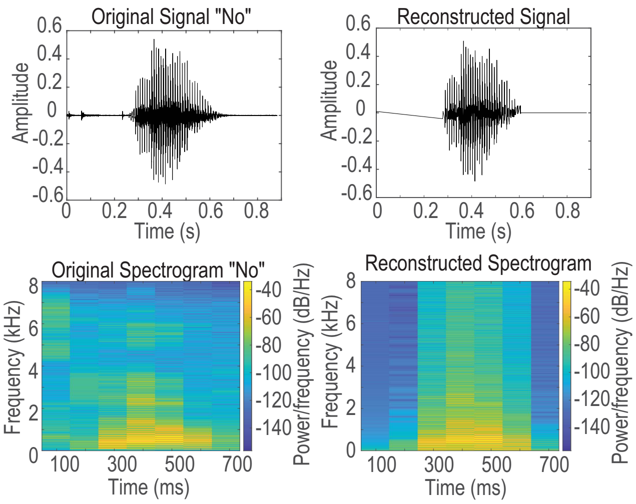


Fig. 5. Frequency-Time analysis for raw speech signal and reconstructed speech signal using Dynamic Predictive Sampling.

A time-domain simulation waveform is shown in Fig. 4. The input waveform is a 30-second speech signal from the TIMIT database [18]. On the left of Fig. 4, the error threshold is set as 2% of the voice amplitude. As shown in the zoomed-in waveform, the selected sampling points show a good-quality reconstructed signal during the speech voice region while only a few sampling points are selected in the pause/noise region. This is because most of the noise amplitude is below the error threshold. The 2% error threshold results in a 4.47 times sampling point reduction. On the right of the figure, the error threshold is set as 5% of the voice amplitude, which results in a 9.8 times sampling point reduction. There are much fewer sampling points in the pause region while the selected sampling points can still maintain the morphology of the waveform in the voice signal region. To evaluate the performance of Dynamic Predictive Sampling, we performed the frequency-time analysis in a typical KWS application. The example result of the word “no” is shown in Fig. 5 with a 3% error threshold. The reconstructed waveform tracks the original waveform well. The spectrogram shows that during the voice active region, the reconstructed waveform is similar to the original waveform, indicating that this method has great potential in KWS.

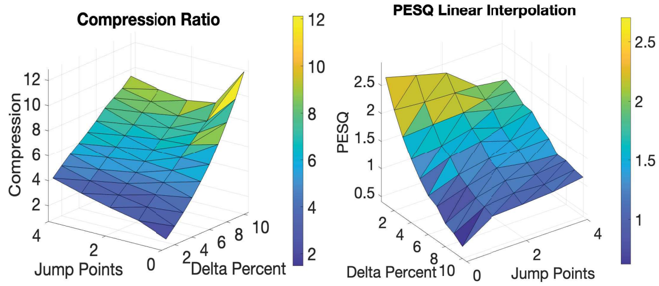
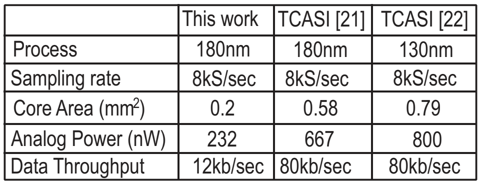


Fig. 6. (left) Compression Ratio with different error threshold and jump points. (right) Reconstructed signal PESQ value with different error thresholds and jump points.

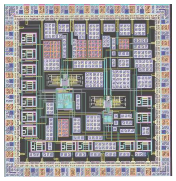
The primary advantage of applying Dynamic Predictive Sampling in speech signal sensing is reducing the data throughput. Although the number of the selected sampling points is much less than the number of the original sampling points, additional timestamp data need to be recorded. The length of the timestamp data depends on the maximum duration between the selected sampling points. In our preliminary estimation, for a speech signal with a sampling rate of 8kS/s, a 4-bit timestamp is enough. The compression ratio between the Dynamic Predictive Sampling and the Nyquist sampling measures the ratio of the total data throughput between the two types of systems. It also depends on the error threshold in the Dynamic Predictive Sampling system. i.e., the $Delta$ value as a percentage of the maximum signal amplitude. Moreover, dynamic prediction can also be performed by skipping a few points. For example, instead of using two consecutive samplings for prediction, the system can allow a few sampling points between the two selected sampling points to perform prediction. More jump points result in a higher compression ratio. The relations between the compression ratio, jump points, and Delta percent are shown in Fig. 6 (left).

If the number of jump points and $Delta$ percent is too high, the system can greatly reduce the data throughput, but the quality of the speech signal may also be reduced. The quality of the speech signal is often evaluated using Perceptual Evaluation of Speech Quality (PESQ). PESQ is an automated assessment of speech quality that gives predictions of subjective speech quality through a range of degradation conditions such as background noise, analog filtering, and variable delay. PESQ was originally developed to assess the speech quality of telephone networks and speech codecs [19]. PESQ spans the range 0.5 to 4.5, where 0.5 is indiscernible and 4.5 is distinct and matching the reference audio; A PESQ of 1.5 is the acceptable minimum standard for discernible audio. Due to the lossy nature of the Dynamic Predictive Sampling method, the system needs to balance the compression ratio and the required minimum quality of the recovered audio. Fig. 6 (right) shows that a high $Delta$ percent leads to a high compression ratio but a suboptimal PESQ performance; and a low $Delta$ percent leads to a high PESQ but low compression ratio. To maintain a high quality of the speech signal, the error threshold $Delta$ should be under 3%, which leads to a reasonable compression ratio of 6.67, which is equivalent to 87% data reduction.



	This work	TCASI [21]	TCASI [22]
Process	180nm	180nm	130nm
Sampling rate	8kS/sec	8kS/sec	8kS/sec
Core Area (mm^2)	0.2	0.58	0.79
Analog Power (nW)	232	667	800
Data Throughput	12kb/sec	80kb/sec	80kb/sec

Comparison of Analog Front-end for Speech Processing



Chip Layout

Fig. 7. The integrated circuit was designed with 0.18 μm CMOS process, which can reduce power and data throughput compared to other speech signal sensing systems. The chip layout is 1mm by 1mm with a core area of 0.2mm^2.

An Event Driven Dynamic Predictive Sampling integrated circuit for speech signal sensing is designed with a 0.18 μm CMOS process. The design contains the comparator, the digital-to-analog converter, the prediction logic, the decision logic, and the timer circuit. The test chip layout is 1 mm by 1 mm with a core area of 0.2mm. The sampling rate is 8 kHz while the processing clock is set at 100 kHz. The resolution of the amplitude output is 10-bit and the time output is 4-bit. The output data throughput depends on the input signal sparsity, the error threshold $Delta$ present, and the jump point. The simulation results show that the output data throughput is 12kb/second, which is much less than a standard speech signal sensing data throughput. Due to the power-saving feature in Dynamic Predictive Sampling, the simulated analog power consumption is 232 nW, which is also lower than the typical speech signal sensing system thanks to fewer comparisons in digitization processes. The chip layout and the comparison between recent speech-sensing systems [20], [21] are summarized in Fig. 7.

IV. Conclusion

This paper studies the performance trade-off of Event-Driven Dynamic Predictive Sampling for speech signal sensing. Dynamic Predictive Sampling selects only a few percent of samplings in the analog waveform for digitization. The selection is based on prediction value using the prior sampling points and the user-defined digital error threshold. Such a system can greatly reduce the output data throughput while selecting the key sampling points during the analog-to-digital conversion process. The compression ratio and PESQ value are used to evaluate the performance of speech signal sensing. Simulation results show that with a 3% error threshold in Dynamic Predictive Sampling, the system can achieve an acceptable quality of the speech signal, which refers to 1.5 of the PESQ, and a total of 87% of the data can be saved. An integrated circuit is designed on a 0.18 μm CMOS process, which shows that it can also reduce power consumption compared to standard speech sensing circuits thanks to the sparsity of the speech signal. Dynamic Predictive Sampling has great potential for embedded speech recognition and keyword-spotting applications.

References

[1] Q. Li, Z. Liu, X. Yang, and F. Qiao, "Always-on speech recognition terminals: Designs based on approximate computing methods," *IEEE Nanotechnology Magazine*, vol. 16, no. 1, pp. 57–74, 2022.

[2] T. Tambe, E.-Y. Yang, G. G. Ko, Y. Chai, C. Hooper, M. Donato, P. N. Whatmough, A. M. Rush, D. Brooks, and G.-Y. Wei, "A 16-nm soc for noise-robust speech and nlp edge ai inference with bayesian sound source separation and attention-based dnns," *IEEE Journal of Solid-State Circuits*, vol. 58, no. 2, pp. 569–581, 2023.

[3] M. Yang, H. Liu, W. Shan, J. Zhang, I. Kiselev, S. J. Kim, C. Enz, and M. Seok, "Nanowatt acoustic inference sensing exploiting nonlinear analog feature extraction," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 10, pp. 3123–3133, 2021.

[4] B. Yu, M. Luo, D. Wang, X. Wang, and S. Qiao, "A low-power and low-latency speech feature extractor based on time-domain filter bank," *IEEE Transactions on Circuits and Systems II: Express Briefs*, pp. 1–1, 2023.

[5] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge:keyword spotting on microcontrollers," *arXiv:1711.07128.*, 2017.

[6] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1," *arXiv:1602.02830*, 2016.

[7] J. Giraldo and M. Verhelst, "Laika: A 5uw programmable lstm accelerator for always-on keyword spotting in 65nm cmos," in *ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, pp. 166–169, 2018.

[8] M. Sun, A. Raju, G. Tucker, S. Panchapagesan, G. Fu, A. Mandal, S. Matsoukas, N. Strom, and S. Vitaladevuni, "Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 474–480, 2016.

[9] J. S. P. Giraldo, S. Lauwereins, K. Badami, and M. Verhelst, "Vocell: A 65-nm speech-triggered wake-up soc for 10- μ w keyword spotting and speaker verification," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 868–878, 2020.

[10] Y. S. Chong, W. L. Goh, V. P. Nambiar, and A. T. Do, "A 2.5 μw kws engine with pruned lstm and embedded mfcc for iot applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 3, pp. 1662–1666, 2022.

[11] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv:1804.03209*, 2018.

[12] J. Ott, Z. Lin, Y. Zhang, S.-C. Liu, and Y. Bengio, "Recurrent neural networks with limited numerical precision,," *arXiv:1608.06902*, 2018.

[13] B. Liu, A. Xue, Z. Wang, N. Xie, X. Wang, Z. Wang, and H. Cai, "A reconfigurable approximate computing architecture with dual-vdd for low-power binarized weight network deployment," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, no. 1, pp. 291–295, 2023.

[14] B. Liu, H. Cai, Z. Wang, Y. Sun, Z. Shen, W. Zhu, Y. Li, Y. Gong, W. Ge, J. Yang, and L. Shi, "A 22nm, 10.8 μw/15.1 μw dual computing modes high power-performance-area efficiency domained background noise aware keyword- spotting processor," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 12, pp. 4733–4746, 2020.

[15] M. Saeed, Q. Wang, O. Märtens, B. Larras, A. Frappé, B. Cardiff, and D. John, "Evaluation of Level-Crossing ADCs for Event-Driven ECG Classification," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 6, pp. 1129–1139, 2021.

[16] W. Tang, A. Osman, D. Kim, B. Goldstein, C. Huang, B. Martini, V. A. Pieribone, and E. Culurciello, "Continuous Time Level Crossing Sampling ADC for Bio-Potential Recording Systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 6, pp. 1407–1418, 2013.

[17] M. Renteria-Pinon, X. Tang, and W. Tang, "Real-Time In-Sensor Slope Level-Crossing Sampling for Key Sampling Points Selection for Wearable and IoT Devices," *IEEE Sensors Journal*, vol. 23, no. 6, pp. 6233–6242, 2023.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[19] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752 vol.2, 2001.

[20] S. Zhang, F. Su, Y. Wang, S. Mai, K. P. Pun, and X. Tang, "A low-power keyword spotting system with high-order passive switched-capacitor bandpass filters for analog-mfcc feature extraction," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 11, pp. 4235–4248, 2023.

[21] D. A. Villamizar, D. G. Muratore, J. B. Wieser, and B. Murmann, "An 800 nw switched-capacitor feature extraction filterbank for sound classification," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 4, pp. 1578–1588, 2021.