Programmable Linear RAM: A New Flash Memory-based Memristor for Artificial Synapses and Its Application to Speech Recognition System

Shifan Gao[1], Guangjun Yang[2], Xiang Qiu[3], Chun Yang[4], Cheng Zhang[4], Binhan Li[2], Chao Gao[2], Hong Jiang[2], Zhexian Wang[2], Jian Hu[2], Jun Xiao[2], Bo Zhang[2], Choonghyun Lee[1,3], Yi Zhao[1, *], and Weiran Kong[2]

[1]College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou, China
[2]Shanghai Huahong Grace Semiconductor Manufacturing Corporation, Shanghai, China
[3]Flash Billion Semiconductor Co. Ltd., Hangzhou, China
[4]State Key Laboratory of Advanced Optical Communication Systems and Networks, Peking University, Beijing, China
*E-mail: yizhao@zju.edu.cn

***Abstract*—**In this work, a new type of flash memory-based memristor, named *programmable linear random-access memory* (PLRAM), is presented to store analog synaptic weights in a single flash memory cell. A PLRAM cell with a self-calibrating program/erase scheme can provide very stable and repeatable analog memory states up to 7 bits in a single cell, which is suitable for an artificial synapse in the neural network. The physical implementation of a discrete Fourier transform on PLRAM arrays shows a remarkably good agreement with theoretical calculations. By taking nonlinearity effects on both forward propagation and backpropagation into consideration, a highly manufacturable speech recognition system, which consists of a 5-layer fully connected neural network (360k artificial synapses, 1576 neurons, and peripheral circuits) has been successfully built on 200 mm wafers. For the first time, the high accuracy of speech recognition (>90%) on 8 different Chinese speech words is demonstrated.

I. Introduction

Recently non-von Neumann architecture has been attracting much attention for data-centric applications such as real-time image processing and speech recognition because of its massively parallel architecture. A great amount of effort has been put in by various research teams to build artificial synapses with emerging non-volatile memory (NVM) devices [1-6] due to their compact structure, ultralow power consumption, and BEOL-compatible processing temperature. However, despite of these advantages there are still major obstacles (asymmetric conductance change, imperfect device reliability, and resistance variation) to overcome in the development of high-performance and ultralow power neural networks. Though the reliability and variability issues of emerging NVM devices can be avoided by SRAM-based implementations [7-9], the energy consumption and performance on the neural networks must be compromised. These advantages and concerns of artificial synapse elements for neuromorphic computing are summarized in **Table I**.

In this work, a new type of flash memory cell with a self-calibrating program/erase scheme, named programmable linear random-access memory (PLRAM), is presented to store analog synaptic weights in a single Flash memory cell. Based on the superior analog properties of PLRAM arrays, a speech recognition system is fully integrated in a 5-layer fully connected neural network (FCNN) with 360k artificial synapses, 1576 neurons, and peripheral circuitry on 200 mm wafers. With a linearity-aware design of artificial synapses in neural networks, a high accuracy of speech recognition (>90%) on 8 different speech words is demonstrated for the first time.

II. PLRAM as an Artificial Synapse

A promising result for analog memory states using embedded NOR Flash memory has been reported [10], where the input signal is coming through the gate. One concern is that it has to be operated in a subthreshold region to cancel out the exponential components in the subthreshold MOSFET current equation, which might cause the degradation of recognition accuracy. We redesigned a conventional 90 nm floating gate flash memory cell to store analog synaptic weights for an artificial synapse. As shown in **Fig. 1**, in this new flash memory cell, the electron can move from the select gate to the floating gate through the tunneling oxide in between, which is specially designed for the precise control of program/erase functionality, by Fowler-Nordheim (FN) tunneling. By decoupling the gate oxide and tunneling oxide in the conventional Flash memory cell, PLRAM provides the better transistor reliability and more analog synaptic weights in a single Flash memory cell. The linear region of $I_d - V_d$ curve in a PLRAM cell can be used as a memristor in a typical crossbar array (**Fig. 2**). The weight updates of PLRAM cells were examined by a single direction program/erase scheme (**Fig. 3**) and average weight changes per erase/program cycle indicate that a precise weight control in PLRAM cells is limited (**Fig. 4**). A self-calibrating program/erase scheme on PLRAM cell is proposed to overcome the above issues. The conductance of each PLRAM cell is measured before programming and a different control voltage is employed to achieve multiple update rates, enabling the analog memory states up to 7 bits in a single PLRAM cell.

For the inference operation, each individual cell should be able to reach the target state. A high accuracy of PLRAM conductance for different programming targets is achieved by establishing a feedback control loop, which can hold the conductance at the target state (**Fig. 5**). To verify the feasibility of PLRAM cells for practical applications, the weight retention test was performed on PLRAM arrays at $250\,^{\circ}C$. The worst case of stored weight changes is about 4%, which meets the retention specification for a reliable artificial synapse design (**Table II**).

III. Desing of Speech Recognition Systems

A. Feature extraction with PLRAM

For a speech recognition process, the first step is to extract features for identifying the linguistic content and discarding the

unnecessary information like background noise and emotion. A fast Fourier transform (FFT) is a typical algorithm that samples a speech signal over a period of time and transfers it into the frequency domain. Although the FFT can bring the computational complexity down to $O(nlog^n)$, it is still far from the hardware-friendly implementation $O(n)$ and takes a long time and computational resources. In this sense, a discrete Fourier transform (DFT) is beneficial for speech signal processing tasks through the matrix multiplication. The physical implementation of DFT on PLRAM arrays was performed and its programming error exhibits a high correlation coefficient with the theoretical calculations (**Fig. 6**). Two clear peaks in frequency components are observed in both theoretical calculations and PLRAM-based neural networks and their correlation loss is negligible (**Fig. 7**).

B. Nonlinearity effects on forward propagation

Besides the nonlinearity in the weight updates that plays a key role in the backpropagation of neural network, there is an additional nonlinearity concern with respect to the forward propagation in the inference operation, that has not been properly framed. Since the results ought to be linear with the input, the output characteristics need to be linear. It indicates that the artificial synapse has to exhibit a true resistor-like behaviors, which obey the Ohm's law. By optimizing the process flow to have a longer linear region in the transistor operation, PLRAM cells can act as a better voltage-controlled resistor (**Fig. 8**). It is confirmed that the correlation of theoretical calculations and PLRAM neural networks is improved with higher linearity in the transistor operation (**Fig. 9**).

Since this problem shows up not only in the transistor, but also in the emerging NVM devices like RRAM and PCM, a systematic study of its impact on the correlation coefficient (or accuracy) is carried out. The current in the emerging NVM device increases sharply over the linear region as the voltage increases, while it saturates in the transistor. A second order correction proportional to the weight is used to model two different types of I-V curves with a nonlinearity coefficient, denoted as Δ (**Fig. 10**). As nonlinearity increases the correlation coefficient decreases sharply in emerging NVM device, but it decreases relatively slow in the transistor. By setting the effective region as the correlation coefficient larger than 0.99, the acceptable Δ can be defined (**Fig. 11**). The error caused by the nonlinearity of current or conductance also propagates through the neural networks and it can accumulate (**Fig. 12**). It is worthy to note that the correlation of theoretical calculations and experimental data (corresponding to accuracy of neural network) decreases slightly in each neural network layer, suggesting that having a high correlation coefficient at the first neural network layer is necessary to guarantee the high accuracy of speech recognition. Interestingly, nonlinearity effects in PLRAM cells can be minimized by accelerating drain-induced barrier lowering with a higher drain bias or channel length scaling. This gives a different guiding principle for device optimization opposite to a conventional logic CMOS.

C. Fully connected neural network based on PLRAM

Based on the superior analog properties of PLRAM arrays, a speech recognition system was fully integrated in a 5-layer FCNN (one layer for feature extraction, three layers for classification, and one layer for output) with 360k artificial synapses, 1576 neurons, and the peripheral circuitry on 200 mm wafers (**Fig. 13**). As the first step for feature extraction, the input voice audio signal is processed by DFT and triangular filters are applied on a Mel-scale to the power spectrum for extracting frequency bands [11]. The Mel-scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies. After that, the output flows through a 3-layer FCNN, and a winner-takes-all strategy gives the classification results. The top-view optical images of a full chip and enlargement of key blocks including PLRAM-based neural networks (synapses), input/output neurons, and activation function are shown (**Fig. 14**). To evaluate the performance of PLRAM-based speech recognition system in terms of natural language processing, 8 Chinese speech words were randomly chosen. Each step of the speech recognition system shows a remarkably good agreement between FCNN model and output results, where the recognition accuracy of 8 different speech words is higher than 90% (**Fig. 15**). The initial performance results of the speech recognition chip are summarized in **Table III**. With further optimization of peripheral circuits and analog properties of PLRAM cell (artificial synapse), a low power and higher classification accuracy can be achieved for fast and energy-efficient mobile applications.

IV. CONCLUSION

A novel programmable linear random-access memory (PLRAM) was presented to store analog synaptic weights in a single Flash memory cell. The results show that PLRAM with a self-calibrating program/erase scheme can provide very stable and repeatable analog memory states up to 7 bits in a single cell, which is suitable for an artificial synapse in the neural network. By taking nonlinearity effects on both forward propagation and backpropagation into consideration, a highly manufacturable speech recognition system, which consists of a 5-layer fully connected neural network (360k artificial synapses, 1576 neurons, and peripheral circuits) has been successfully built on 200 mm wafers. The high accuracy of speech recognition (>90%) on 8 different speech words was also demonstrated for the first time.

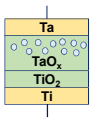
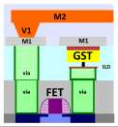
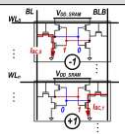
ACKNOWLEDGMENT

This work was performed in collaboration with various research and industry teams. The system architecture and circuit design were performed by Flash Billion Semiconductor and the chips were fabricated by Huahong Grace Semiconductor Manufacturing Corp. The Characterization and analysis were done by Zhejiang University. This work was supported in part by the Key Research and Development Program of Zhejiang Province (2019C01158).

REFERENCES

[1] X. Zhang et al., *EDL,* vol. 39, no. 2, pp. 308-311, 2018. [2] Q. Luo, et. al., *IEDM*, pp. 48-51, 2017. [3] J. Shin, et. al., *IEDM*, pp. 63-66, 2018. [4] G. W. Burr et al., *IEDM*, pp. 697-700, 2014. [5] S. Ambrogio et al., *Nature*, vol. 558, pp.60-67, 2018. [6] H. Tsai et al, *VLSI Tech. Symp.*, T82-T83, 2019. [7] P. A. Merolla et al., *Science*, vol. 345, pp. 668-673, 2014. [8] R. Guo et al., *VLSI Circuit Symp.*, pp. C120-C121, 2019. [9] M. Davies et al., *IEEE Micro*, vol. 38, pp. 82-99, 2018. [10] X. Guo et al., *IEDM*, pp. 151-154, 2017. [11] S. Davis and P. Mermelstein, *IEEE TASSP*, vol. 28, pp. 357–366, 1980.

Table I. Artificial synapses for neuromorphic computing

	RRAM [1-3]	PCM [4-6]	SRAM-based [7-9]
Device architecture			
Advantages	• Compact MIM structure • BEOL-compatible (< 400 ℃) • Ultralow energy consumption	• Large dynamic range • BEOL-compatible (< 400 ℃) • Cell size scalability	• Robust and reliable • Existing technology • Fast clock frequency
Concerns	• Abrupt nature of filament formation • Higher resistance variation • Imperfect device reliability	• Resistance drift effect • Abrupt RESET process • Asymmetric conductance change	• Larger cell size • Power and cost constraints

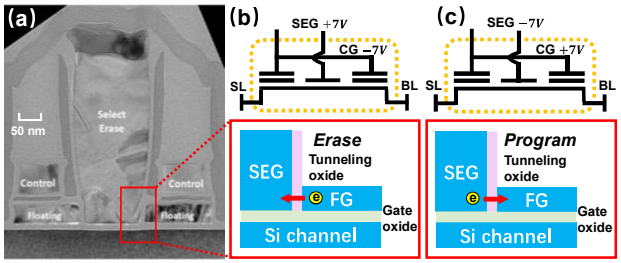


Fig. 1 (a) TEM image of programmable linear random-access memory. **(b)** Erase mechanism of PLRAM cell. **(c)** Program mechanism of PLRAM cell. By decoupling the gate and tunneling oxide, PLRAM provides the better transistor reliability and more analog synaptic weights.

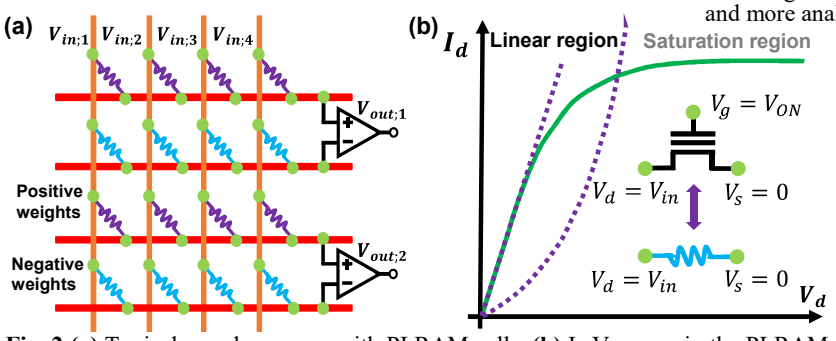


Fig. 2 (a) Typical crossbar arrays with PLRAM cells. **(b)** I_d-V_d curve in the PLRAM cells. The linear region in the transistor operation can be used as a memristor.

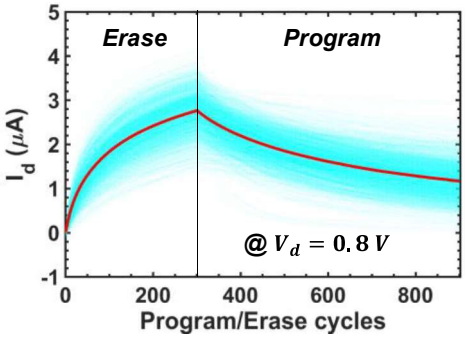


Fig. 3 Weight updates of PLRAM cells with a single direction program/erase scheme. The transparent blue background shows the program and erase of the weights in 897 cells, and the red line shows the average result.

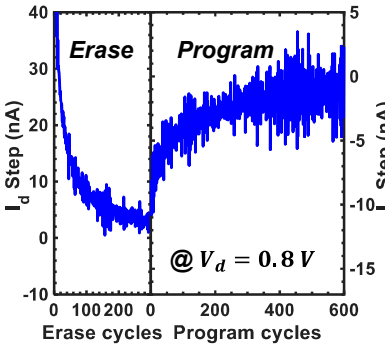


Fig. 4 Change of weights in a single step erase/program, averaged between 897 cells. The small step ensures a precise control of the weight.

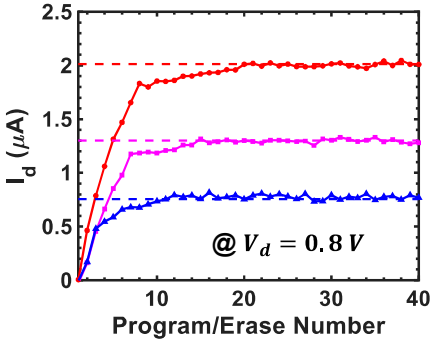


Fig. 5 High accuracy of PLRAM conductance for different programming targets. Current oscillation at the end shows a precise control of the conductance.

Table II. Chip level PLRAM current deviation

	Max. increase			Max. decrease		
Chip #	1	2	3	1	2	3
ΔI_d@ $V_d = 0.8\ V$ (%)	3.8	4	3.6	-0.8	-1.8	-2
	Ave.			Med.			Std.		
Chip #	1	2	3	1	2	3	1	2	3
ΔI_d@ $V_d = 0.8\ V$ (%)	1.6	1.2	0.6	1.6	1	0.4	0.4	0.4	

*Baking was performed in 250℃ for 72 hours with dynamics range 5 μA.

(a) DFT matrix:

$$W_{\cos} = \left(\frac{\cos(-2\pi jk/N)}{\sqrt{N}}\right)_{j,k=0,\ldots N-1}$$

$$W_{sin} = \left(\frac{\sin(-2\pi jk/N)}{\sqrt{N}}\right)_{j,k=0,\ldots N-1}$$

Example:

$$W_{cos} = \frac{1}{\sqrt{N}}\begin{bmatrix} \cos 0 & \cos 0 & \cos 0 & \cdots \\ \cos 0 & \cos(-\frac{\pi}{2}) & \cos(-\pi) & \cdots \\ \cos 0 & \cos(-\pi) & & \\ \vdots & \vdots & \ddots & \end{bmatrix}$$

$$|\text{Freq. domain}|^2 = |W_{cos} \times \text{Time domain}|^2 + |W_{sin} \times \text{Time domain}|^2$$

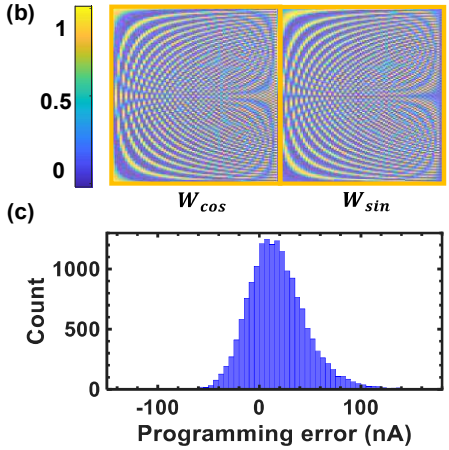


Fig. 6 (a) Discreate Fourier transform (DFT) matrix. **(b)** The positive part of the weight of cosine part and sine part in DFT matrix. Only half of DFT matrix is shown due to symmetry. **(c)** Histogram of programming error for the positive part, compared with the theoretical calculations.

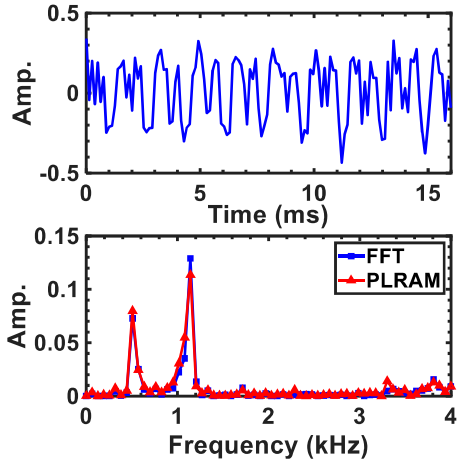


Fig. 7 (a) Input audio signal. **(b)** Comparison of DFT implementation on PLRAM arrays and theoretical calculations (FFT).

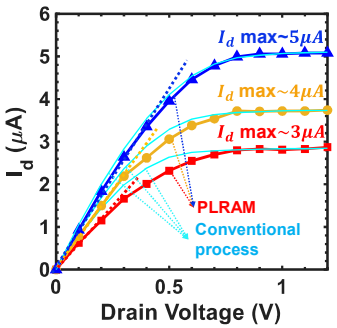


Fig. 8 $I_d - V_d$ curves of the flash memory cells. PLRAM cells show a longer linear region in the transistor operation than the conventional flash memory cells. As I_{dmax} increases, the linearity is improved.

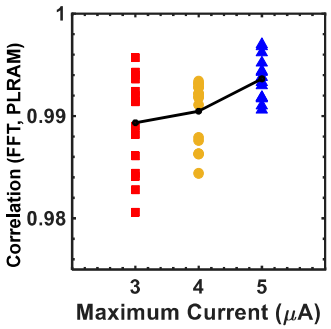


Fig. 9 Correlation of theoretical calculations (FFT) and PLRAM neural networks. The maximum correlation coefficient of 99.5% is achieved with I_{dmax} ~5μA.

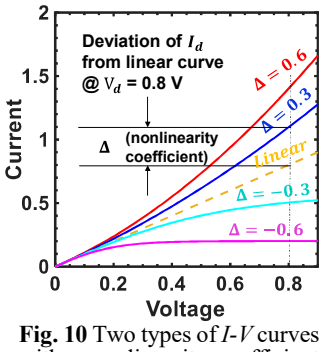


Fig. 10 Two types of *I-V* curves with a nonlinearity coefficient (Δ). The upper curves are from two-terminal emerging NVM devices and the lower ones are from PLRAM devices.

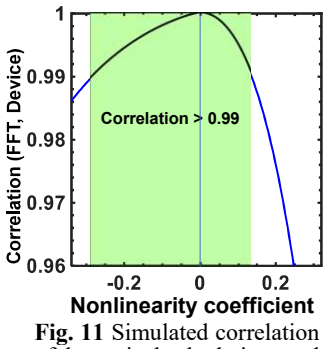


Fig. 11 Simulated correlation of theoretical calculations and PLRAM neural networks as a function of a nonlinearity coefficient. PLRAM device has a higher tolerance for the nonlinearity than emerging NVM devices.

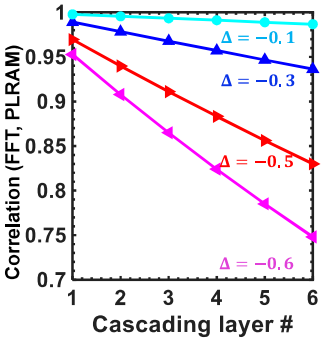


Fig. 12 Estimated correlation coefficient between theoretical calculation and PLRAM neural network as a function of neural network layers.

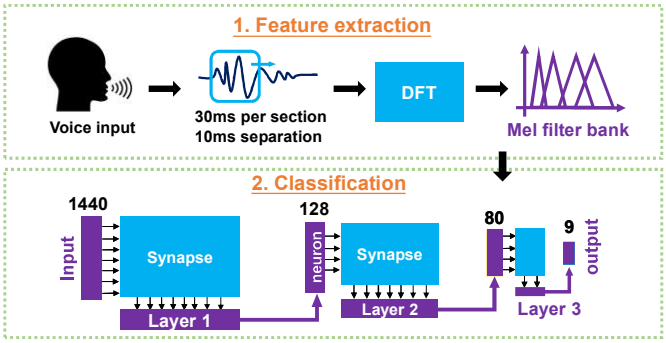


Fig. 13 Speech recognition system based on PLRAM neural networks. The system consists of 360k artificial synapses, 1576 neurons, and the peripheral circuitry. After feature extraction with DFT and Mel filter bank, the classification is done with 3 layers of fully connected neural network.

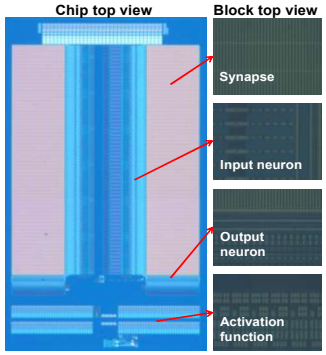


Fig. 14 Top-view optical images of a full chip and key blocks (artificial synapses, input and output neurons, activation function).

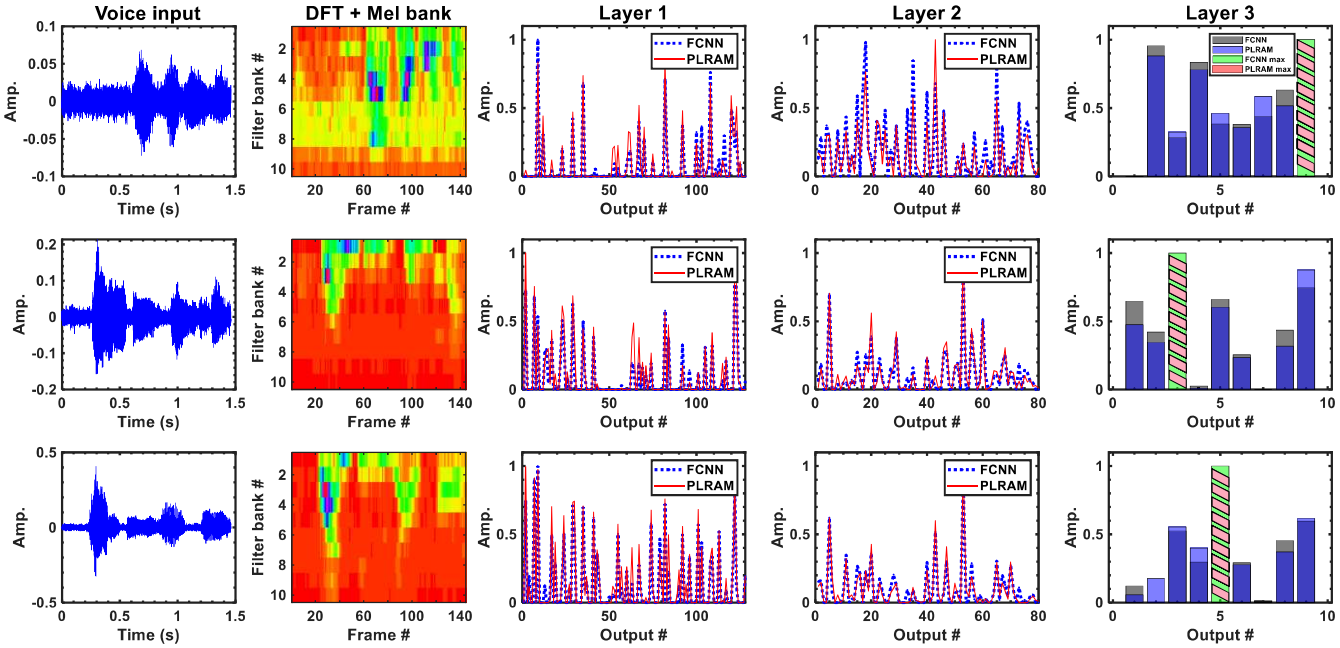


Fig. 15 Three voice examples for the classification. They show a remarkably good agreement between FCNN model and chip results.

Table III. Summary of the system architecture and speech recognition chip performance

Metric	Technology	Synapse #	layer #	DFT corr.	FCL 1 corr.	FCL 2 corr.	FCL 3 corr.	Array area	Cell Area	Energy/ oper.	Throughout	keyword #	PLRAM acc.	Model acc.
This work	90 nm	360448	5	0.993	0.957	0.943	0.931	0.4 mm^2	30 F^2	91 fJ	11 TOPS/W	8	90%	94%