Cochlea-inspired sensor for speech recognition

Paolo Han Beoletto*†[1], Antonio S. Gliozzi[1], Gianluca Milano[2], Carlo Ricciardi[1], and Federico Bosia[1]

[1]Department of Applied Science and Technology [Politecnico di Torino] – Italy
[2]Istituto Nazionale di Ricerca Metrologica – Italy

Abstract

The digital transition is creating an increasing demand for speech recognition to be integrated into everyday devices. The human-machine interface is becoming increasingly reliant on the device's capability to interact with the user and the accurate detection of spoken words plays a central role in achieving this. Traditional deep learning methods are powerful but consume significant energy and computational resources. They are based on computing architectures where sensing, memory, and processing are separated: this approach increases latency and power consumption, especially when dealing with large networks of sensors. To address these challenges, in-sensor computing and near sensor computing are interesting paradigms, integrating computational tasks directly within the sensor material itself. This not only reduces energy consumption and data transfer times but also opens new doors for intelligent devices that can perform low-level computations internally. This work investigates how mechanical metamaterials can offer energy-efficient solutions for processing wave-based information in applications like IoT, artificial intelligence, and beyond. The device presented in this talk integrates a neuromorphic (mimicking brain's neural processing) auditory sensor inspired by the human cochlea with a physical reservoir based on a memristive architecture. The reduced power consumption and training cost, together with the fast encoding of spatio-temporal inputs, make it an interesting option for speech recognition tasks in the field of Internet of Things. The sensor is based on a spiral structure that exhibits tonotopy, i.e., the spatial discrimination of elastic waves depending on their frequency content. Thanks to an optimization routine the parameters that define the geometry of the spiral are selected to maximize the tonotopic effect in the desired frequency range.This sensor aims to be used as an artificial cochlea: with piezoelectric read-out in different locations of the device, we obtain a spatio-temporal map that is called "cochleagram", whose components are equivalent to the signals measured at different locations of the spiral. The parallel reading of the vibration signal in different locations of the device allows the direct separation of its frequency components, converting the auditory signal into recognizable patterns, thus enhancing the overall performance of speech recognition systems that benefit from frequency-related data. This measurement is run on an entire dataset of spoken digits, with 10 classes corresponding to the digits ranging from 0 to 9, for a total of 3000 samples. A linear classifier is trained with the dataset of measured cochleagrams in order to discriminate between the 10 classes, reaching an overall accuracy of 97.66%. Compared to the classification accuracy achieved by state-of-the-art software on the same dataset, the hardware implementation proposed here delivers superior performance while offering the significant advantage of eliminating the need for additional digital processing units. To further reduce the energy consumption of the

*Speaker
†Corresponding author: paolo.beoletto@polito.it

system, it is possible to simplify the classifier by reducing the number of features required for the classification. This can be achieved through the integration of a reservoir, which is a network of coupled nonlinear elements capable of capturing time dependencies in the data. By creating a high-dimensional representation of the input, the reservoir significantly reduces the complexity of the neural network processing. The reservoir in this system is implemented in materia, leveraging memristive connections between silver nanowires. Cochleagrams, representing the spatio-temporal output of the spiral sensor, are fed into the reservoir as time-varying inputs. The reservoir condensates the information contained in the original maps in fewer features, allowing the classifier to operate with fewer timesteps of the cochleagram while maintaining high accuracy. Experimental results show that the inclusion of the reservoir preserves the system's ability to discriminate spoken digits, allowing to train the classifieron lower-dimensional data without losing its classification accuracy. This system is an insteresting low-power, efficient alternative to traditional methods for speech recognition, particularly in resource-constrained IoT applications.