Speech Recognition with Cochlea-Inspired In-Sensor Computing

ADVANCED INTELLIGENT SYSTEMS
Open Access
www.advintellsyst.com

Speech Recognition with Cochlea-Inspired In-Sensor Computing

*Paolo H. Beoletto, Gianluca Milano, Carlo Ricciardi, Federico Bosia, and Antonio S. Gliozzi**

Traditional speech recognition methods rely on software-based feature extraction that introduces latency and high energy costs, making them unsuitable for low-power devices. A proof-of-concept demonstration is provided of a bioinspired tonotopic sensor for speech recognition that mimics the human cochlea, using a spiral-shaped elastic metamaterial. The measured modal response of the structure at different frequencies generates a spatially distributed signal, providing a spatiotemporal map of the input named "tonogram". The device acts as an in-sensor physical reservoir computing system, working simultaneously as a sensor and as a computing unit, capable of extracting features of spoken words relevant to speech recognition. Results indicate that this can serve as a valid alternative to traditional software-based digital preprocessing, ensuring high accuracy in terms of classification, while reducing computational requirements. This work demonstrates the potential of bioinspired metamaterials for energy-efficient auditory sensing and, beyond speech recognition, for applications such as IoT devices and edge computing artificial intelligence systems.

1. Introduction

In today's rapidly evolving digital landscape, the increase of connected devices through the Internet of Things (IoT) and artificial intelligence has led to an explosion in the number of sensor nodes,[1–4] expected to reach 30 billion by 2030. In the same scenario, machine learning (ML) workloads have rapidly grown, raising concerns about their carbon footprint.[5–7] Traditional computing architectures, where sensing, memory, and processing are separated, struggle to meet the growing energy demands. 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,[8–13] integrating computational tasks directly within the sensing architecture itself. This not only reduces energy consumption and data transfer times but also opens new possibilities for intelligent devices that compute on the edge. Bioinspired and neuromorphic computing systems have gained significant attention as efficient alternatives to traditional computing architectures. Human sensory systems have inspired several in-sensor or near-sensor applications; for example, image sensors mimicking the human visual system can combine in-sensor architectures for low-level processing inspired by the retina,[14,15] with neuromorphic near-sensor readout systems that process the signal in a similar way to the visual cortex.[16–19]

One of the most critical applications that can be tackled with bioinspired computing is speech recognition, a task that is essential for human–machine interaction, artificial intelligence, and assistive technologies.[20] Speech recognition requires the conversion of acoustic energy into neural representations, a process that in biological systems is performed by the cochlea. The cochlea is a spiral-shaped, fluid-filled sensory organ located in the inner ear.[21] It plays a crucial role in the process of hearing by converting sound waves into electrical signals that the brain can interpret. Its properties allow mammals to perceive sounds in wide amplitude and frequency ranges, spanning nearly 10 octaves and 120 dB,[22,23] making it an interesting candidate as a biological system that could inspire functional structures for applications in the field of sound sensing and processing. The main feature of interest of the cochlea is its tonotopy, i.e., the spatial discrimination of elastic waves depending on their frequency content. The variation in thickness and stiffness along the spiral creates a gradient that alters the wave amplitude as it propagates, with the peak amplitude location determined by the stimulus frequency.[24,25] Sound-induced pressure fluctuations induce vibrations in the basilar membrane, causing the Organ of Corti to stimulate hair cells, which convert mechanical motion into electrical signals, generating action potentials for auditory processing.[26–28] The cochlea does not simply act as a passive converter of sound into electrical signals; rather, it plays

P. H. Beoletto, C. Ricciardi, F. Bosia, A. S. Gliozzi
Department of Applied Science and Technology
Politecnico di Torino
C.so Duca degli Abruzzi, 24, 10129 Torino, Italy
E-mail: antonio.gliozzi@polito.it

G. Milano
Advanced Materials Metrology and Life Science Division
INRiM (Istituto Nazionale di Ricerca Metrologica)
Strada delle Cacce 91, 10135 Torino, Italy

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202500526.

DOI: 10.1002/aisy.202500526

an active role in processing and extracting relevant features, organizing incoming sounds into a pattern of activity in the auditory nerve.

Several studies have tried to understand and replicate this functionality in digital systems,[29] leading to the development of software models that mimic cochlear sound processing. Just as the basilar membrane separates frequencies along the cochlear spiral, software-based models like Mel-Frequency Cepstral Coefficients (MFCCs)[30] and Lyon's cochlear model (cochleagram)[31] apply a series of filters to break down an audio signal into its spectral components. MFCCs approximate the cochlea's logarithmic frequency scaling, capturing perceptually relevant features, while Lyon's model simulates the traveling wave dynamics of the cochlea, incorporating aspects of biological auditory processing. These processes of feature extraction are currently necessary preprocessing steps to implement speech recognition tasks in emerging hardware technologies. Indeed, this data processing can be the most important step for extracting relevant features of the input signal, strongly contributing to or even determining the ultimate performance of the entire computing system for speech recognition.[32] While these software-based approaches have proven effective, they rely on conventional digital processing, which introduces energy costs and latency. A hardware mimicking cochlear signal treatment and feature extraction could be a promising alternative for sensing and processing sound on the same physical substrate, directly at the matter level. Adopting an in-sensor processing approach can be particularly useful for IoT applications, where devices often need to operate in environments with limited resources in terms of power and memory: processing the acoustic signal *in materia* could reduce energy consumption, crucial for battery-powered devices, and minimize the memory required for processing and storing data.

In the field of bioinspired metamaterials, several designs based on locally resonant structures have drawn inspiration from cochlear tonotopy.[33–35] Since local resonance phenomena typically occur within a narrow frequency band, achieving effects over a broad frequency range requires combining a large number of elements with different resonant frequencies. As human vocalization extends over a frequency range that exceeds two decades, i.e., from 100 Hz to 20 kHz, and fundamental frequencies for speech intelligibility reach 5 kHz,[36] it is crucial to overcome this limit for applications related to speech recognition. Coiled structures have proved capable of performing rainbow trapping over wide frequency ranges,[37] so that a spiral-shaped tonotopic resonator inspired by the shape of the cochlea itself[38] provides an effective solution to span the whole speech spectrum using a close-packed structure that is not reliant on resonating elements. In the field of psychoacoustics, similar structures[39] have attracted attention as potential functional bionic devices.

In this work, we present a bioinspired sensor inspired by the cochlea, based on a spiral-shaped elastic metamaterial. Its tonotopic frequency response, coupled with the piezoelectric readout of 16 independent channels placed in different locations, allows to process *in materia* time-domain of elastic signals associated with spoken words. The combination of the outputs of the channels results in a spatiotemporal map containing frequency-related content, similar to the outputs of software-based processing algorithms, such as Lyon's cochleagrams or MFCCs. This map is given the name "tonogram," as its working principle is based on the tonotopy of the structure, which allows to spatially discriminate the frequency content of the processed signal. The potential of this sensor in speech recognition is evaluated by analyzing the tonograms associated with 3000 spoken digit samples from the audioMNIST (Modified National Institute of Standards and Technology database), the audio counterpart of the MNIST.[40] First, its feature extraction capability is visualized using *t*-distributed stochastic neighbor embedding (*t*-SNE), and then tonograms are employed to train a linear classifier for digit and speaker recognition tasks. Results demonstrate that this innovative low-power hardware technique matches the performance of state-of-the-art software-based methods, positioning itself as a promising solution for efficient speech recognition hardware technologies.

2. Results

2.1. The Metasensor

The structure of the cochlea-inspired metasensor is optimized following the method presented in the study by Dal Poggetto et al.[38] A logarithmic spiral defines the central line of the geometry as follows:

$$r_C(\theta) = r_0 e^{k_r \theta/\theta_{max}}, \theta \in [0, \theta_{max}] \quad (1)$$

where r_0 is the radius at $\theta = 0$, k_r is the polar slope of the spiral, and $\theta_{max} = 2\pi n_T$ is defined by the number of turns n_T. The cross-section at an angle θ is defined as the rectangle $b(\theta) \times h(\theta)$, whose width and height vary following the equations

$$b(\theta) = b_0 e^{k_b \theta/\theta_{max}}, h(\theta) = h_0 e^{k_h \theta/\theta_{max}} \quad (2)$$

Here, b_0 and k_0 represent the width and height of the metasensor at $\theta = 0$, while k_b and k_h define the variation of the width and height along the spiral, respectively.

The aim of the design procedure is to obtain the structure with the optimal tonotopy for a given volume and width, in the frequency range of interest, i.e., the frequency band relevant for speech recognition from 100 Hz to 10 kHz. This corresponds to a distribution of the maxima of the out-of-plane displacement that follows a linear relation between the logarithm of the frequency and the spatial variable θ (the angle describing the path along the central line). This results in the ability of unambiguously spatially resolving different frequency components of an elastic wave that contains speech-related information. As shown in Dal Poggetto et al.[38] the parameters (k_b, k_r, k_h) can be optimized using an active set algorithm[41] implemented in MATLAB that compares the results of the eigenfrequency finite element simulation of the structure under study with the objective function representing the required tonotopic distribution and updates the geometry in an iterative procedure. The aim is to minimize the metric $\Gamma = \sum_i d_i^2$, where $d_i = |s_i - \overline{\omega}|/\sqrt{2}$ is the distance between the maximum of displacement of the ith out-of-plane mode and the line $\overline{\omega} = s$, that considers the normalized spatial coordinate s_i and the normalized frequency $\overline{\omega}_i = \log(\omega_i/\omega_{min})/\log(\omega_{max}/\omega_{min})$. Depending on the frequency band of interest, different values can be selected for

$\omega_{max} = 2\pi f_{max}$ and $\omega_{min} = 2\pi f_{min}$ to cover the desired frequency range with a uniform distribution in space. For this application, the values are restricted to the range $[f_{min}, f_{max}] = [10^2, 10^4]$ Hz. Numerical simulations are performed using the material properties of the commercial polymer used for the fabrication, Solflex Tech Tough polymeric resin, characterized by a Young's modulus $E = 2.5$ GPa, a Poisson's ratio $\nu = 0.33$, and a mass density $\rho = 1150\,\mathrm{kg\,m^{-3}}$. The structure is considered clamped at the external end of the spiral and free at the internal one. Volume, number of turns, initial thickness, and tolerance $(V, n_T, h_0, \Delta_r^{min})$ are kept fixed throughout the process. The optimization routine results in a structure with clearly distinct maxima for different frequencies, as shown in **Figure 1**a: the higher the frequency of the eigenmode, the closer the maximum amplitude of oscillation is located to the inner end of the structure. Figure 1b,c shows the numerically calculated modal shape of two out-of-plane eigenmodes at 190 and 880 Hz, respectively; the locations of the two maxima of oscillation can be clearly distinguished spatially.

To test the structure experimentally, a piezoelectric transducer with a diameter of 20 mm applied at the external end of the spiral is used to drive the oscillations, while a laser vibrometer (Polytec, OFV-505) is used to detect the out-of-plane velocities on the sample surface at different spatial locations. The surface is scanned by translating the structure by means of two perpendicular linear stages (see Figure 1d), creating a 40×40 mm^2 map with steps of 1 mm. In each position, the oscillation of the structure in response to a frequency sweep ranging from 100 Hz to 5 kHz is measured. Studying the Fourier transform in all the points allows us to obtain the full field response of the spiral at each frequency to verify the predicted tonotopic behavior. Figure 1e,f shows the experimental counterpart of Figure 1b,c, displaying the measured out-of-plane oscillation amplitude at 190 and 880 Hz, respectively. The numerical modal shape is reproduced well by experimental results at both frequencies, confirming the expected tonotopic behavior.

The tonotopy of the resonator is used to spatially discriminate the frequency content of an elastic signal. By measuring the oscillation amplitude of the resonator in different positions, it is possible to obtain output signals that contain information about different frequency bands. A piezoelectric readout system is used to enable the parallel measurement of 16 channels (see **Figure 2**c). The channels are limited to 16 in order not to exceed the number of modes supported by the resonator in the speech intelligibility frequency range. The piezoelectric patches are placed along the central line of the spiral, following the order displayed in Figure 2a, so that the oscillation amplitude read by each patch corresponds to different frequency spectra. The response to a frequency sweep in the locations of the 16 channels is studied: all channels display an enhanced signal response around 3.8 kHz due to the resonance characteristics of the input piezoelectric disk. However, this does not mask the trend of the

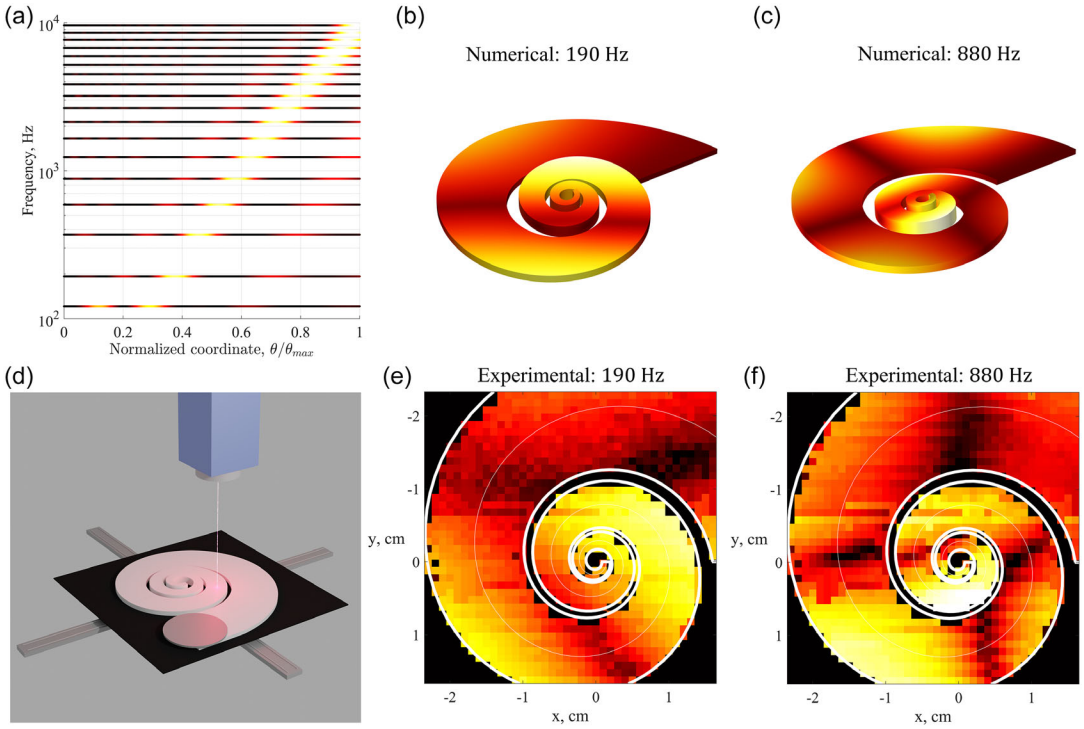

Figure 1. Tonotopic behavior of the cochlea-inspired spiral resonator. a) Spatial distribution of the oscillation amplitude of the eigenmodes of the resonator along the central line of the spiral. This numerical result demonstrates an optimized tonotopy, where the logarithm of the frequency of maxima of oscillation amplitude for each eigenmode depends linearly on the spatial coordinate θ. b–c) Numerical modal shapes of the resonator at 190 and 880 Hz, respectively. d) Schematics of the measurement setup: a 20 mm piezoelectric transducer excites the spiral's oscillations, while a laser vibrometer measures out-of-plane velocities at various points, scanning the surface via two perpendicular linear stages. e–f) Fourier components of the experimental response of the resonator at 190 and 880 Hz.

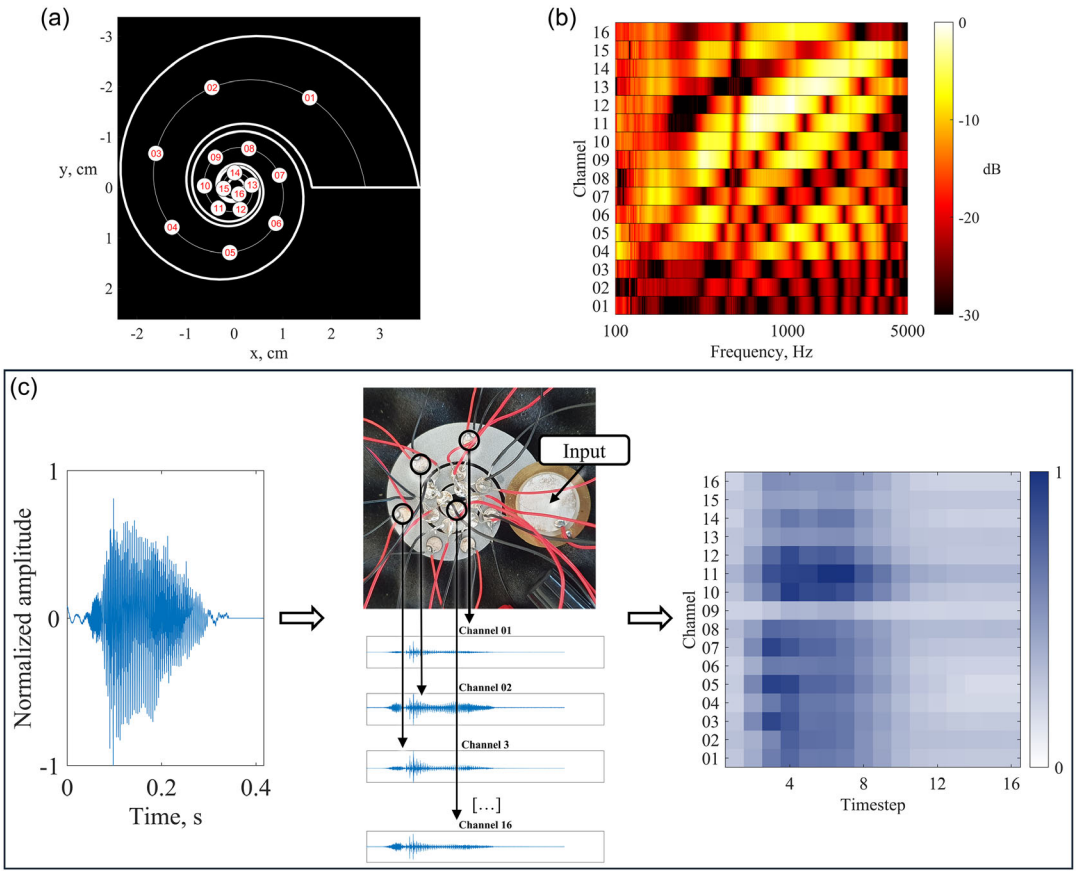

Figure 2. Description of the device. a) Schematic of the device with 16 channels located throughout the central line of the resonator. b) Response of the device in the location of each channel to a frequency sweep ranging from 100 Hz to 5 kHz. c) Processing mechanism of the device with picture: the time-domain input of a spoken word (here "zero") is used as input voltage on the excitation piezoelectric disk; each channel reads the oscillation amplitude at its location with piezoelectric read-out, then the signal power for each channel in a discrete set of timesteps is computed resulting in a 16×16 spatio-temporal map named tonogram.

response from lower to higher frequencies when ranging from outer to inner channels. The resonant behavior is not related to the device itself, as it is caused by the excitation system used for the experiments. We thus remove the bias introduced by the piezoelectric transducer by subtracting the average response over the entire spiral. The response in the region from 100 Hz to 5 kHz is reported in Figure 2b. The frequency content of the channels partially overlaps, but the "center of mass" of the response shifts following the tonotopic behavior predicted by the simulations and verified by the laser measurements.

2.2. Spoken Digits

The device is designed to generate a spatiotemporal map of a time-domain signal of a spoken word with a low-latency parallel readout of the 16 channels. Figure 2c shows the steps of this process, in the case of the word "zero": the time-domain signal is generated by the piezoelectric transducer on the outer end of the cochlea; this excites the whole structure, and the 16 channels read the oscillation at the specific locations where they are placed. The data are reduced to a 16 × 16 map, by calculating the signal power of each channel for a specific timestep out of 16: this map is henceforth referred to as "tonogram" and corresponds to a spatiotemporal representation of the time-domain audio signal with a limited number of features. The term tonogram is inspired by the fact that the frequency decomposition is based on the tonotopy of the resonator. Representations of this kind are typically less complex to interpret and discriminate for neural networks trained for speech recognition tasks, as they benefit from information related to the frequency content of the spoken word more than time-domain data.

2.3. Feature Extraction

In order to verify the effectiveness of the analog processing performed by this device, a test is performed on a spoken digit recognition task. As done in similar hardware validation studies,[13,42,43] a single constrained corpus is used to benchmark the system in a controlled and interpretable setting. The input data are taken from the audioMNIST dataset,[40] which consists of a repository of 30 000 samples of digits from 0 to 9 spoken by 60 speakers with 50 repetitions each. This dataset is designed as a

preliminary "sandbox" for experimenting with novel architectures and interpretation algorithms, in the style of the MNIST dataset of handwritten digits for image recognition,[44] and is used in benchmark tasks for edge computing applications,[45,46] explainable artificial intelligence (XAI)[47,48] and beyond. Since the process of transforming the auditory signal into a tonogram is extremely time-consuming (requiring a real-time acquisition for each spoken word), the dataset was limited to the first six speakers, resulting in a total of 3000 samples. To evaluate the full system beyond this subset, classification performances on the full AudioMNIST dataset were estimated through simulations, as detailed in the Appendix. Each of the samples consists in an audio file with a sampling frequency of 48 kHz, whose raw vector is padded with zeros in order to match the length of the longest vector. For every sample, the process is similar to that described in the previous section and shown in Figure 2d: the audio file is converted into a time-domain signal and fed into the input piezoelectric disk, so that a 16×16 tonogram is obtained from the channel readout. Repeating the measurement for the entire database allows us to create a set of tonograms. The aim of this study is to evaluate the feature extraction capability of the device. The t-SNE technique[49] allows to visualize how data separation occurs. The results are compared to those obtained with three conventional preprocessing algorithms commonly used by the speech recognition community,[32] i.e., Lyon's ear cochleagram, the MFCC filter, and a linear spectrogram. These are methods used to decompose a time-domain signal into frequency channels, which differ in terms of nonlinear filters (**Figure 3**a). Each of them is used on the dataset to obtain spatiotemporal maps with the same dimensionality (i.e., 16×16), for the comparison between them to be unbiased in terms of complexity of the system.

T-SNE is a dimensionality reduction technique commonly used for projecting high-dimensional data in a lower-dimensional space; in this case, it reads a space of 16×16 dimensions in a 2D plane. As the probability of two points being adjacent is conserved in the transformation, this method allows to obtain a visual interpretation of complex data structures. By assigning a different color to the samples belonging to each class, it is possible to evaluate the effectiveness of the frequency decomposition method of choice: the formation of clusters of dots of the same color denotes the ability of a method to spatially discriminate different digits, while a random distribution is connected to poor performance in feature extraction. Figure 3b–e shows the distributions obtained with t-SNE applied to the datasets associated with the four methods: while the spectrograms display a scarce capability of resolving some of the classes (Figure 3e), better results in terms of clustering are obtained with the remaining filters. In particular, the most promising results are obtained from tonograms in Figure 3b and MFCCs in Figure 3c, which display well-formed, distinct clusters. This suggests that tonograms effectively encode frequency-dependent information in

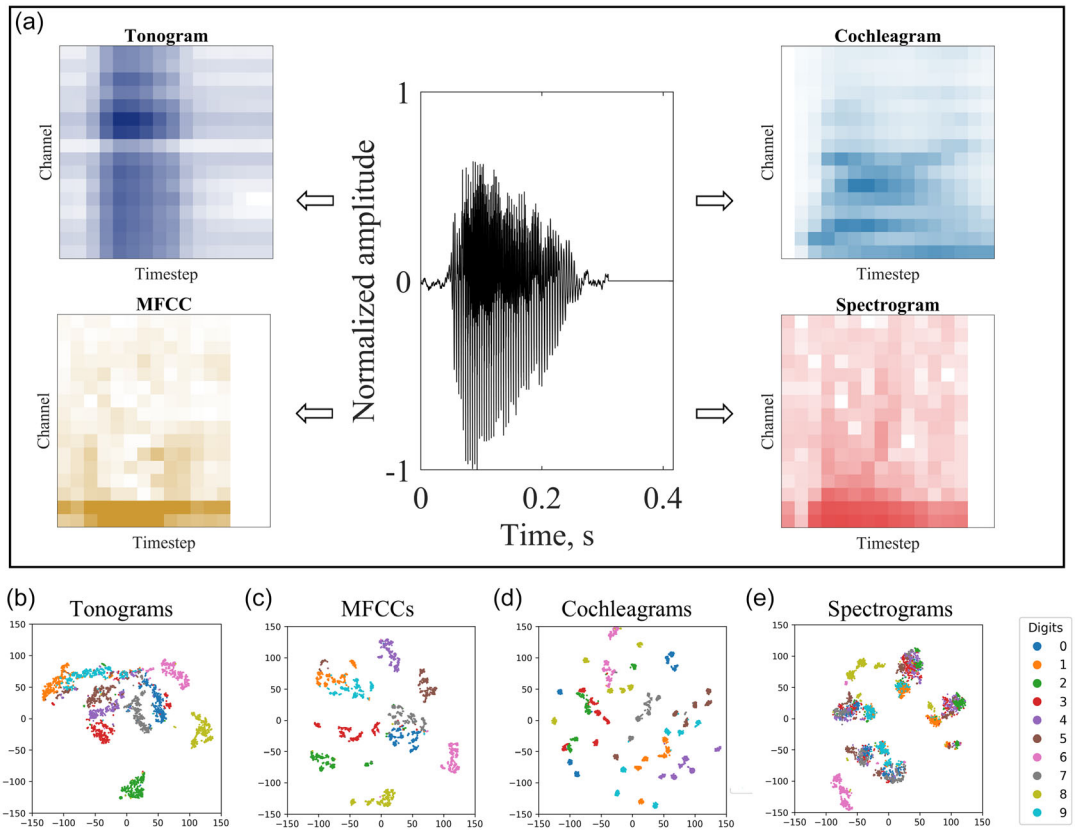

Figure 3. a) Comparison between different preprocessing methods. The time-domain signal of the word "nine" is converted experimentally into a tonogram, and the corresponding MFCCs, Lyon's cochleagram, and spectrogram are computed via software. b–e) T-SNE visualization of the dataset processed with tonograms, MFCCs, Lyon's cochleagrams, and spectrograms.

a structured manner, enhancing classification performance. The clustering quality is quantified with silhouette scores, a metric that evaluates the cohesion within and separation between clusters, with values ranging from –1 to 1. This choice is motivated by the fact that commonly used classification metrics (such as precision, recall, and F1-score) are tailored to assess the performance of complete recognition systems, while here we focus on evaluating the effectiveness of the feature extraction stage, independently of any particular classifier. A higher score indicates that points are well-matched to their own cluster and well-separated from others, while a lower or negative score suggests poor clustering. The value obtained from the spectrograms is 0.01, which is the lowest result. The metrics of MFCCs (Figure 3c) and Lyon's cochleagrams (Figure 3d) are, respectively, 0.44 and 0.11, suggesting that the first method projects the data over a space that separates better the different classes. The results obtained with tonograms have a silhouette score of 0.31, which indicates a better performance with respect to cochleagrams. The comparison with state-of-the-art methods suggests that this device's clustering performance is highly promising: as a hardware-based approach, it achieves results comparable to (or even exceeding) software-based methods, demonstrating the effectiveness of *in-materia* computing.

2.4. Digit Classification

The system's dynamics are exploited to perform a computational task in the context of physical reservoir computing, separating the process into two main components: the first is related to feature extraction, which identifies the main characteristics of the input signal, and the second is classification.[50] The feature extraction, which is performed with a hardware approach in the case of tonograms, allows to reduce the classification task to a simple read-out. Since the aim of this work is to evaluate the performance of hardware-based feature extraction in comparison with state-of-the-art software methods, a linear classifier is chosen to highlight the effect of the preprocessing stage.[32] For this, a pseudoinverse approach is employed: the Moore–Penrose pseudoinverse of the training data matrix is computed and used to derive the optimal weight matrix, which projects input features onto the target space in a least-squares sense.[51] K-fold crossvalidation is employed to evaluate the model's performance. Specifically, the dataset is partitioned into 10-fold, ensuring that the samples corresponding to each speaker are evenly distributed across these folds. In each iteration, onefold is designated as the test set, while the remaining ninefolds serve as the training set. This process is repeated 10 times, with each fold acting as the test set once. The word success rate (WSR) for each iteration is computed, after which the results are averaged across all folds.

To evaluate the classification performance, a confusion matrix is computed and normalized across true classes to express recognition accuracy in percentage form, providing insights into the model's ability to discriminate between different digit classes. This algorithm is repeated over the data obtained with the four methods, obtaining the confusion matrices reported in **Figure 4**a. The values of the WSR of the four methods are displayed in Figure 4b, with a mean value and an error bar related to the distribution of values obtained with the K-fold approach. The spectrograms exhibit the worst performance with WSR of $75.5 \pm 1.9\%$. The remaining filters display very similar recognition rates; in particular, MFCCs stand at $98.6 \pm 0.5\%$, Lyon's cochleagrams reach $98.4 \pm 0.6\%$, while the best result is achieved with the tonograms at $98.7 \pm 0.6\%$. This result confirms the potential of this device to be implemented *in materia* analogic feature extraction. The tonotopy of the metasensor is therefore shown to capture the core traits of the digit frequency content in an efficient way.

To better understand the differences between these feature extraction models, the relevance of the features in each map is analyzed using the analysis of variance (ANOVA) F-test.[52] This statistical test evaluates whether the mean values of the features differ significantly across different classes, allowing to determine which features contribute the most to distinguishing between them. The heat maps shown in Figure 4c represent the F-value for each entry of the spatiotemporal maps associated with the four methods. This metric quantifies the ratio between the variance of the feature across the classes and the variance within the classes: a feature associated with a high F-value has more relevance in discriminating between different digits. Tonograms distribute the information across all channels, with variations more pronounced over time. A similar pattern is observed in spectrograms, whereas both MFCCs and Lyon's cochleagrams carry less information related to the time domain. Figure 4d presents the average F-value of the features belonging to each channel. All four methods tend to have lower relevance in the higher channels, though cochleagrams and MFCCs exhibit more peaked distributions. By exploiting this information, it is possible to study speech recognition performance with a reduced number of features: for each method, the linear classifier is trained with $m \in [1, 15]$ channels, keeping each time the ones with the highest F-value for the specific method, so that there is no bias associated with this selection: for example, the first channel to be removed is 9 for tonograms, 15 from MFCCs, 1 for cochleagrams, and 16 for spectrograms, following the order given by the information reported in Figure 4d. The WSR is evaluated, and the mean and standard deviation associated with the k-fold approach are reported in Figure 4e. Tonograms maintain an advantage over cochleagrams and spectrograms for a smaller number of channels, despite being outperformed by MFCCs. This test provides additional evidence of the good performance of the feature extraction via a hardware approach, in line with the standards of the state-of-the-art software-based methods. The main advantage of the hardware approach lies in the reduction of computational cost and latency. Although classification is still performed digitally, replacing the software-based feature extraction with a physical preprocessing stage eliminates over 99% of the total inference time, as shown in the Appendix. This results in lower latency, since signal processing occurs in real time during acquisition and improves efficiency by removing the need for the majority of digital operations typically required in conventional pipelines.

2.5. Speaker Classification

In the context of speaker recognition, feature extraction plays a crucial role in distinguishing between different speakers based

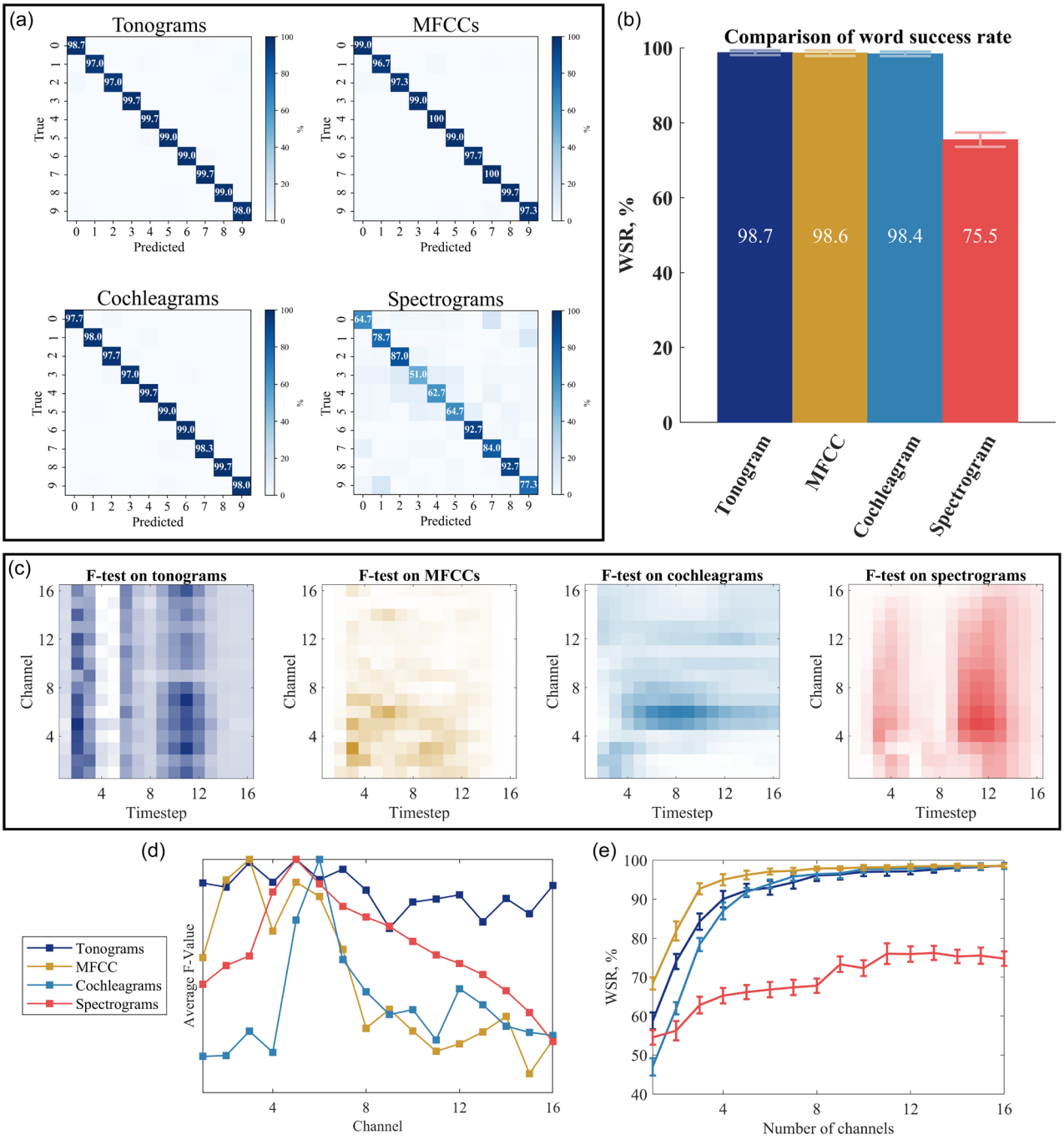

Figure 4. Results associated with digit classification tasks. a) Confusion matrices associated with the linear classifier trained with tonograms, MFCCs, Lyon's cochleagrams, and spectrograms. b) Mean value and standard deviation of the WSR for the classification task associated with each method. c) Heatmaps obtained with the ANOVA F-test on the dataset processed with the four methods. The maps display the F-value of each feature normalized to the highest value from brighter to darker colors. d) Average F-value of the features belonging to each channel for the four distinct methods. e) WSR obtained by training the linear classifier with a limited number of channels, each time removingthe most relevant one in terms of F-value. The channel to be removed is different for each method, as the average F-values have different distributions.

on their vocal characteristics. The effectiveness of a feature extraction method directly impacts the classification performance, influencing the system's ability to differentiate between speakers with similar voice patterns. Following the same approach as in the previous section, a linear classifier based on a Moore–Penrose pseudoinverse is trained to discriminate between the six speakers who contribute to the 3000 samples forming the dataset. The task is performed on the data treated

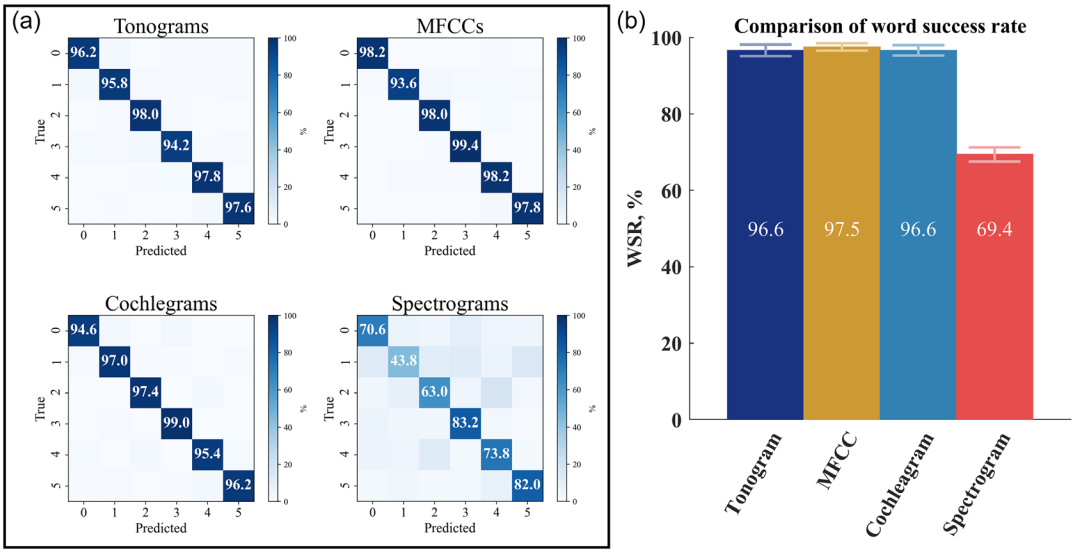

Figure 5. Results associated with speaker classification tasks. a) Confusion matrices associated with the linear classifier trained with tonograms, MFCCs, Lyon's cochleagrams, and spectrograms. b) Mean value and standard deviation of the WSR of the classification task associated with each method.

with the four feature extraction methods (tonograms, MFCCs, cochleagrams, and spectrograms), and the confusion matrices reported in **Figure 5**a are produced. For this task, the best performance is achieved via MFCC with WSR of $97.6 \pm 0.9\%$ (Figure 5b); however, as in the previous case, both tonograms and cochleagrams provide very similar performances, with $96.6 \pm 1.5\%$ and $96.6 \pm 1.4\%$, respectively. The worst result is associated with spectrograms with WSR equal to $69.4 \pm 1.9\%$. Unlike digit recognition, which focuses on phonetic content, speaker recognition relies on features that encode speaker-specific traits, such as pitch, timbre, and speech dynamics. Information related to the temporal evolution of the signal assumes less relevance, and this could be one of the reasons for the slight decrease of accuracy for the hardware approach. However, tonogram performances are still in line with those relative to the software methods, which suggests that they effectively capture speaker-specific characteristics, making them a viable alternative to traditional software-based feature extraction methods.

3. Discussion

A crucial step to implement speech recognition tasks is the transformation of the time-domain signal associated with spoken words into spatiotemporal maps. This work proposes a hardware approach to perform feature extraction in materia, based on a metasensor inspired by the human cochlea that uses the mechanical properties of a physical resonator as a computational unit. The system behaves as an in-sensor reservoir-computing device, simultaneously performing sensing and processing by extracting the main features of the input signal in its physical domain. By exploiting the tonotopic properties of a spiral-shaped elastic metamaterial, the system achieves frequency discrimination in a hardware-based manner. Using the piezoelectric readout of 16 parallel channels, the device produces a tonogram of the spoken words, whose features carry information on the frequency content of the original signal. The direct signal transformation mirrors the way reservoir computing captures temporal dependencies without the need for intermediate digital processing. Experimental validation using the audioMNIST dataset demonstrates that this analog processing approach performs competitively with established digital techniques, such as MFCCs and Lyon's cochleagrams, both in digit and in speaker recognition tasks, while reducing computational complexity as the signal is processed directly *in sensor*. While the approach promises clear advantages in latency and computational efficiency, it is important to acknowledge some limitations related to the current prototype,

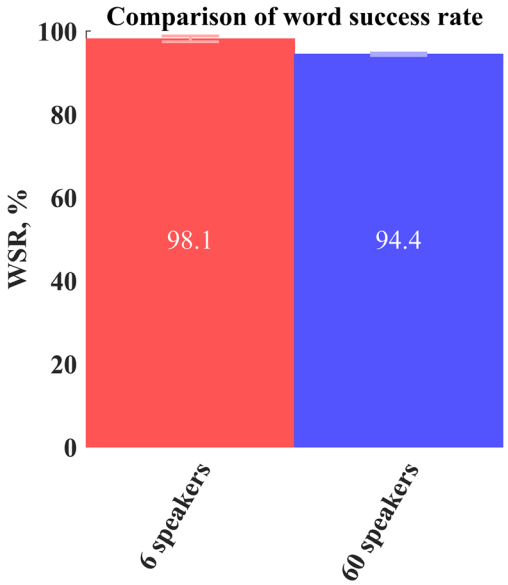

Figure 6. WSR of the digit recognition task performed with the simulated tonograms. This approach allows to evaluate the performance on the full audioMNIST dataset, reaching a WSR of $94.4 \pm 0.3\%$, compared to $98.1 \pm 0.7\%$ related to the subset of 3000 simulated tonograms.

in order to outline paths for future development: first, it relies on a laboratory-scale PXI system for signal acquisition, which is not optimized for energy consumption or integration. Although the metasensor itself is passive and requires no power, a final hardware implementation will also require a custom low-power readout system with an optimized interface with electronics. Second, the classification accuracy achieved with the tonotopic features, although competitive, still slightly lags behind MFCC-based methods in the speaker recognition task. This trade-off is justified by the substantial gains in computational simplicity and latency. Additionally, while the current design is optimized for speech-related tasks, the metasensor geometry can be modified and optimized to address different frequency bands, enabling adaptation to other application domains. Future work will focus on the miniaturization of the structure and the development of custom low-power readout electronics, paving the way toward a fully integrated and efficient hardware implementation. The proposed bioinspired metasensor offers an efficient and scalable solution for auditory processing. Beyond speech recognition, these findings highlight the potential of mechanical metamaterials for low-power hardware-based signal processing in IoT and edge computing applications.

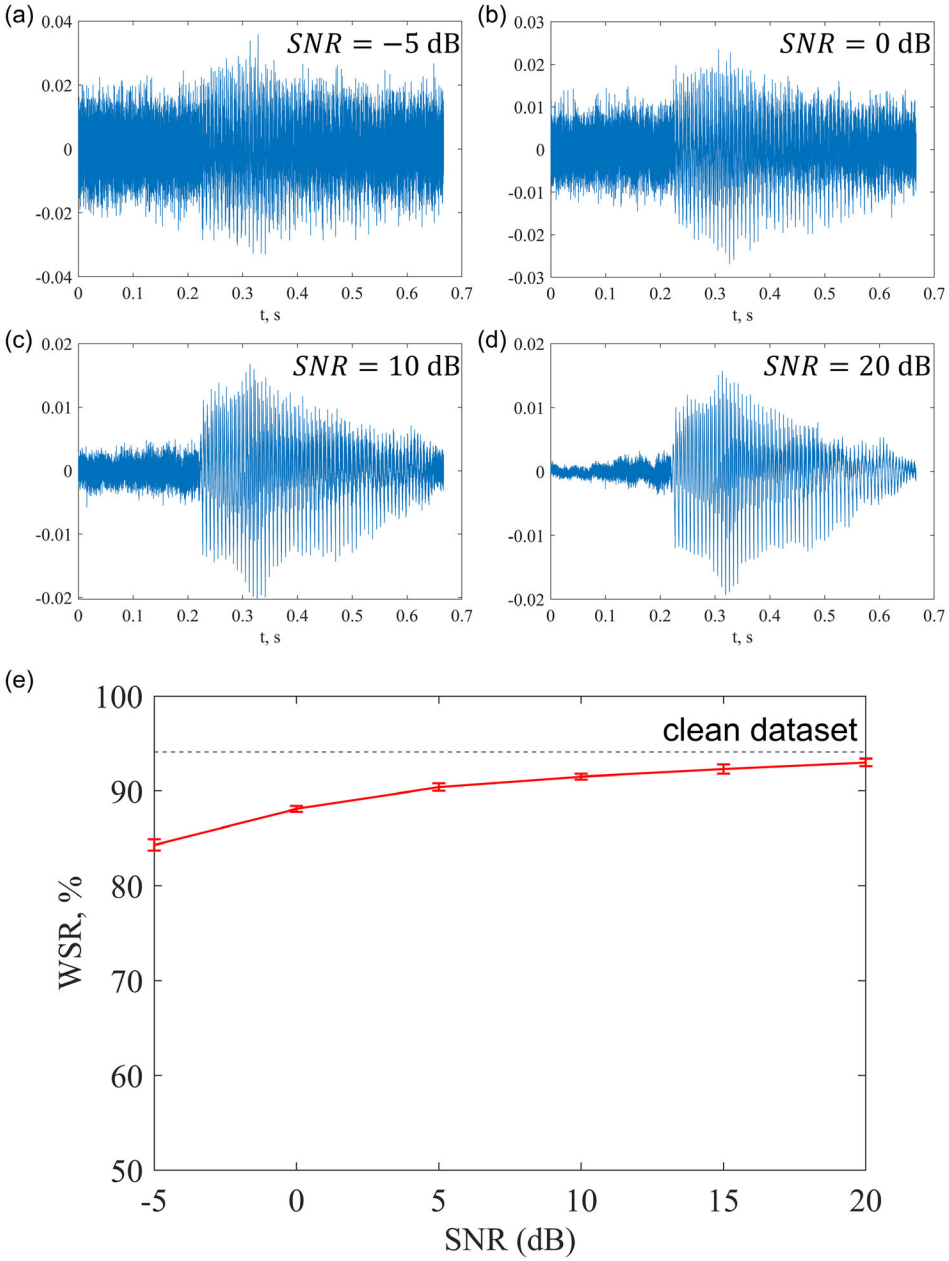

Figure 7. a–d) Time-domain waveforms of the spoken digit "zero" under different SNR levels (−5, 0, 10, 20dB, respectively), illustrating the effect of additive Gaussian noise on the input signals. e) Word success ratio for the digit recognition task for SNR ∈ {−5, 0, 5, 10, 15, 20}dB compared with the baseline of the clean dataset.

4. Appendix

4.1. Simulated Tomograms

The experimental validation of the cochlea-inspired sensor was performed on a limited subset of spoken digits (i.e., 10% of the audioMNIST dataset) due to the extremely time-consuming nature of the process, as each word is processed in real time during its pronunciation. In order to demonstrate generality, a faster approach is adopted by simulating the effect of the resonator through the frequency response of its channels. In particular, the entry of a tonogram at channel i and timestep j can be calculated as

$$\gamma_{ij} = \int_{f_{min}}^{f_{max}} T_i(f) \cdot \int_{\hat{t}_j}^{\hat{t}_{j+1}} s(t) e^{i2\pi f t}\, dt\, df \tag{3}$$

The time-domain signal $s(t)$ associated with a sample of the dataset is windowed in $M = 16$ timesteps, limited by a set of $\hat{t}_j$. Each window is Fourier-transformed, and then the frequency response T_i of each channel i is applied. Finally, integration over frequency yields the signal power at each timestep for each channel.

This method is first used to verify the performance for the entire 60-speaker set: the tonograms associated with the 30 000 samples are simulated and used to perform the digit classification task, in order to compare the performances with those obtained with 6 speakers only. The WSR associated with the subset of 3000 simulated tonograms is $98.1 \pm 0.7\%$, while the one obtained on the full dataset is $94.4 \pm 0.3\%$, as shown in **Figure 6**.

4.2. Resilience to Noise

The same approach is used to estimate the resilience of the system to noise by simulating the effect of additive Gaussian noise on the input signals. Noisy input signals with signal-to-noise ratio (SNR) values in the range $\mathrm{SNR} \in \{-5, 0, 5, 10, 15, 20\}\mathrm{dB}$ are processed into tonograms. The WSRs of the noisy sets are compared with the noiseless baseline ($\mathrm{SNR} = \infty$) with WSR of $94.4 \pm 0.3\%$.

An example of time-domain waveforms for the digit "zero" under different noise levels is shown in **Figure 7**a–d. The classification performance degrades progressively as noise increases, as shown in Figure 7e: for $SNR = -5$ dB, the WSR reaches the lowest value of $84.3 \pm 0.6\%$. Although experimental evaluation under noise conditions would require a more elaborate setup, these preliminary results provide a reasonable indication of the system's robustness.

4.3. Computational Impact of Feature Extraction

The approach presented in this work replaces the feature extraction stage, typically carried out through software algorithms, with a hardware-based alternative. However, the classification task is still performed digitally using a linear readout. Ideally, the computational benefits of the system could be quantified by comparing the number of multiply accumulate operations (MACs) required in each case. However, software-based feature extraction algorithms are generally implemented using optimized libraries (e.g., for Fourier transforms, filterbanks, or compressions), which makes it difficult to directly estimate MAC counts. For this reason, we instead analyze the inference time of the full pipeline and assess how it is divided between the feature extraction and classification stages. This provides a practical measure of the computational gain enabled by the hardware preprocessing.

Table 1. Distribution of the inference time across feature extraction and classification for software-based approaches.

Method	Feature extraction [ms]	Classification [µs]
MFCC	1.30	0.54
Lyon's	23.80	0.55
Spectro	0.34	0.54

The three software-based feature extraction methods considered in this study exhibit different computational complexities, which is reflected in their respective inference times. As reported in **Table 1**, Lyon's method is significantly more time-consuming, while the spectrogram is the fastest. In all cases, however, the time required for the classification stage is negligible compared to that of feature extraction, that takes more than 99.8% of the total inference time: the linear classifier involves a single matrix-vector multiplication with 256 input features and 10 output classes, resulting in submicrosecond execution times. This confirms that the feature extraction stage is the bottleneck in terms of computational cost for this application.

5. Experimental Section

Device Fabrication: The metasensor was fabricated using a commercial 3D-DLP printer (Solflex SF650, W2P Engineering GmbH), which featured a UV-LED at 385 nm, a power density of $P = 8 \pm 0.5\ \mathrm{mW\,cm^{-2}}$, a nominal lateral resolution of 50 µm, and supported layer thicknesses ranging from 25 to 200 µm. For this design, 50 µm layers were used, each exposed for 1.5 s, delivering a dose of $12\ \mathrm{mJ\,cm^{-2}}$ per layer. Parts were printed along the z-axis on removable supports for improved bottom surface smoothness, cleaned with isopropyl alcohol, and sonicated twice for 5 minutes. They were then postcured for 1 h in a UV-curing unit with illumination wavelengths ranging from 320 to 450 nm.

Tonogram Measurement: The time-domain signal relative to the spoken word was generated by a piezoelectric transducer placed at the outer end of the cochlea-inspired resonator, exciting the entire structure. The transducer was made of a brass plate, which was 2 mm thick, and its nominal resonant frequency was 3.8 kHz. The response was measured at 16 specific locations, where piezoelectric patches (ceramic disk with a diameter of 5 mm, thickness of 0.5 mm, and nominal resonant frequency of 370 kHz) were glued on top of the resonator. These patches served as output channels, capturing oscillations at their respective positions. A PCI eXtensions for Instrumentation (PXI) system (NI PXIe-5172) was used to perform parallel measurements on the output channels. Due to the PXI's 8-channel limit, data acquisition was performed in two sequential sets of measurements, covering the outer and inner eight channels, respectively. The acquired signals were processed by computing the signal power of each channel at a specific timestep, producing a 16×16 data map. To ensure consistency and enhance feature extraction, the resulting tonograms were normalized to their highest value, and compression was applied to reduce the amplitude range ($C_{dB} = \frac{1}{6} A_{dB}$, where A_{dB} and C_{dB} are

the amplitudes of the map entries expressed in decibels, before and after the compression, respectively).

Linear Classifier: Given the training feature matrix X_{train} and the one-hot encoded target matrix Y_{train}, the pseudoinverse of X_{train} was computed as

$$X_{\text{pinv}} = \left(X_{\text{train}}^T X_{\text{train}}\right)^{-1} X_{\text{train}}^T \tag{4}$$

and the optimal weight matrix W was then obtained as

$$W = X_{\text{pinv}} \cdot Y_{\text{train}} \tag{5}$$

The learnt weights were subsequently applied to the test matrix to generate predicted class scores S_i for each sample, from which the final class labels y_i were assigned based on the argmax operation, selecting the most probable category.

$$Y_{\text{pred}} = X_{\text{test}} \cdot W = \begin{bmatrix} S_1 \\ S_2 \\ \dots \\ S_{N_{\text{test}}} \end{bmatrix} \tag{6}$$

$$y_{\text{pred}} = \text{argmax}(S_i) \tag{7}$$

Acknowledgements

P.H.B., A.S.G., and F.B. acknowledge funding from the European Union's Horizon 2020 FET Open ("Boheme") under grant agreement no. 863179 and from the "AMPHYBIA" project – funded by the European Union – Next Generation EU, CUP E53D23003060006, within the PRIN 2022 program (D.D. 104 – 02/02/2022 Ministero dell'Università e della Ricerca). Part of this work has been supported by funding from NEURONE, a project funded by the European Union – Next Generation EU, M4C1 CUP I53D23003600006, under the program PRIN 2022 (prj code 20229JRTZA). Part of this work was supported by the European Research Council (ERC) under the European Union's ERC Starting Grant (ERC-2024-STG) agreement "MEMBRAIN" no. 101160604. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Open access publishing facilitated by Politecnico di Torino, as part of the Wiley - CRUI-CARE agreement.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Paolo H. Beoletto, **Antonio S. Gliozzi**, and **Federico Bosia** conceived the idea and coordinated the work. **Gianluca Milano** and **Carlo Ricciardi** contributed to the analysis and interpretation of results. All the authors contributed to the manuscript writing.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

acoustics, bioinspiration, elastic metamaterials, in-sensor computing

Received: May 13, 2025
Revised: July 18, 2025
Published online:

[1] P. Bellini, P. Nesi, G. Pantaleo, *Appl. Sci.* **2022**, *12*, 1607.
[2] W. Ejaz, A. Anpalagan, *Internet of Things for Smart Cities: Technologies, Big Data and Security*, Springer, Germany **2019**.
[3] K. Fizza, A. Banerjee, K. Mitra, P. P. Jayaraman, R. Ranjan, P. Patel, D. Georgakopoulos, *Discover Internet Things* **2021**, *1*, 1.
[4] T. P. Truong, H. T. Le, T. T. Nguyen, *J. Phys.: Conf. Ser.* **2020**, *1432*, 012068.
[5] E. Strubell, A. Ganesh, A. McCallum, *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13693.
[6] D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. R. So, M. Texier, J. Dean, *Computer* **2022**, *55*, 18.
[7] P. Dhar, *Nat. Mach. Intell.* **2020**, *2*, 423.
[8] A. Alu, A. F. Arrieta, et al., *Smart Mater. Struct.* **2025**, *34*, 063501.
[9] F. Zhou, Y. Chai, *Nat. Electron.* **2020**, *3*, 664.
[10] T. Wan, B. Shao, S. Ma, Y. Zhou, Q. Li, Y. Chai, *Adv. Mater.* **2023**, *35*, 2203830.
[11] W. Pan, J. Zheng, L. Wang, Y. Luo, *Engineering* **2022**, *14*, 7797.
[12] H. Jaeger, B. Noheda, W. G. Van Der Wiel, *Nat. Commun.* **2023**, *14*, 4911.
[13] T. Dubček, D. Moreno-Garcia, T. Haag, et al., *Adv. Funct. Mater.* **2024**, *34*, 2311877.
[14] K. Kyuma, E. Lange, J. Ohta, A. Hermanns, B. Banish, M. Oita, *Nature* **1994**, *372*, 197.
[15] G. Zhou, J. Li, Q. Song, et al., *Nat. Commun.* **2023**, *14*, 8489.
[16] Y. Chai, *Nature* **2020**, *579*, 32.
[17] L. Mennel, J. Symonowicz, S. Wachter, D. K. Polyushkin, A. J. Molina-Mendoza, T. Mueller, *Nature* **2020**, *579*, 62.
[18] F. Taherian, D. Asemani, in *2010 IEEE Asia Pacific Conf. on Circuits and Systems*, IEEE, Kuala Lumpur, Malaysia **2010**, pp. 895–898.
[19] G. Milano, G. Pedretti, K. Montano, S. Ricci, S. Hashemkhani, L. Boarino, D. Ielmini, C. Ricciardi, *Nat. Mater.* **2022**, *21*, 195.
[20] M. Anusuya, S. K. Katti, *Int. J. Comput. Sci. Inf. Secur.* **2009**, *6*, 181.
[21] P. Dallos, R. R. Fay, *The Cochlea*, Springer Science & Business Media, New York **2012**.
[22] D. Manoussaki, R. S. Chadwick, D. R. Ketten, J. Arruda, E. K. Dimitriadis, J. T. O'Malley, *Proc. Natl. Acad. Sci.* **2008**, *105*, 6162.
[23] L. Robles, M. A. Ruggero, *Physiol. Rev.* **2001**, *81*, 1305.
[24] G. Von Békésy, *Experiments in Hearing*, Acoustical society of America, New York **1960**.
[25] M. LeMasurier, P. G. Gillespie, *Neuron* **2005**, *48*, 403.
[26] R. Nobili, F. Mammano, J. Ashmore, *Trends Neurosci.* **1998**, *21*, 159.
[27] R. Fettiplace, *Compr. Physiol.* **2011**, *7*, 1197.
[28] I. Russell, P. Sellick, *J Physiol.* **1978**, *284*, 261.
[29] R. F. Lyon, *Human and Machine Hearing*, Cambridge University Press, Cambridge **2017**.
[30] S. Davis, P. Mermelstein, *IEEE Trans. Acoust., Speech, Signal Process.* **1980**, *28*, 357.
[31] R. Lyon, in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. *7*, IEEE, Boston, MA **1982**, pp. 1282–1285.
[32] F. Abreu Araujo, M. Riou, J. Torrejon, et al., *Sci. Rep.* **2020**, *10*, 328.
[33] F. Ma, J. H. Wu, M. Huang, G. Fu, C. Bai, *Appl. Phys. Lett.* **2014**, *105*, 213702.
[34] F. Ma, J. H. Wu, M. Huang, S. Zhang, *Appl. Phys. A* **2016**, *122*, 1.
[35] M. Rupin, G. Lerosey, J. de Rosny, F. Lemoult, *New J. Phys.* **2019**, *21*, 093012.

[36] B. B. Monson, E. J. Hunter, A. J. Lotto, B. H. Story, *Front. Psychol.* **2014**, *5*, 587.
[37] X. Ni, Y. Wu, Z.-G. Chen, L.-Y. Zheng, Y.-L. Xu, P. Nayar, X.-P. Liu, M.-H. Lu, Y.-F. Chen, *Sci. Rep.* **2014**, *4*, 7038.
[38] V. F. Dal Poggetto, F. Bosia, D. Urban, P. H. Beoletto, J. Torgersen, N. M. Pugno, A. S. Gliozzi, *Mater. Des.* **2023**, *227*, 111712.
[39] Y. Liu, L. Wang, J. Chang, F. Ma, *Int. J. Mech. Sci.* **2025**, *287*, 109915.
[40] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, W. Samek, *J. Frank. Inst.* **2024**, *361*, 418.
[41] J. Nocedal, S. J. Wright, *Numerical Optimization*, Springer, New York **2006**.
[42] S. Lilak, W. Woods, K. Scharnhorst, C. Dunham, C. Teuscher, A. Z. Stieg, J. K. Gimzewski, *Front. Nanotechnol.* **2021**, *3*, 675792.
[43] Y. Usami, B. van de Ven, D. G. Mathew, et al., *Adv. Mater.* **2021**, *33*, 2102688.
[44] Y. LeCun, The MNIST database of handwritten digits **1998**, http://yann. lecun. com/exdb/mnist/.
[45] W. Zhang, P. Yao, B. Gao, et al., *Science* **2023**, *381*, 1205.
[46] Q. Yang, W. Jin, Q. Zhang, Y. Wei, Z. Guo, X. Li, Y. Yang, Q. Luo, H. Tian, T.-L. Ren, *Nat. Mach. Intell.* **2023**, *5*, 169.
[47] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, K.-R. Müller, *Proc. IEEE* **2021**, *109*, 247.
[48] M. Mundt, Y. Hong, I. Pliushch, V. Ramesh, *Neural Networks* **2023**, *160*, 306.
[49] L. Van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579.
[50] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, A. Hirose, *Neural Networks* **2019**, *115*, 100.
[51] A. Albert, *Regression and the Moore-Penrose Pseudoinverse*, Academic Press, New York **1972**.
[52] L. St, S. Wold, et al., *Chemom. Intell. Lab. Syst.* **1989**, *6*, 259.