Canonical cortical graph neural networks and its application for speech enhancement in audio-visual hearing aids ☆

Leandro A. Passos [a], João Paulo Papa [b], Amir Hussain [c], Ahsan Adeel [a,d,]*

[a] CMI Lab, School of Engineering and Informatics, University of Wolverhampton, Wolverhampton, United Kingdom
[b] Department of Computing, São Paulo State University, Bauru, Brazil
[c] School of Computing, Edinburgh Napier University, Edinburgh, Scotland, United Kingdom
[d] deepCI.org 20/1 Parkside Terrace, Edinburgh, United Kingdom

ARTICLE INFO

ABSTRACT

Despite the recent success of machine learning algorithms, most models face drawbacks when considering more complex tasks requiring interaction between different sources, such as multimodal input data and logical time sequences. On the other hand, the biological brain is highly sharpened in this sense, empowered to automatically manage and integrate such streams of information. In this context, this work draws inspiration from recent discoveries in brain cortical circuits to propose a more biologically plausible self-supervised machine learning approach. This combines multimodal information using intra-layer modulations together with Canonical Correlation Analysis, and a memory mechanism to keep track of temporal data, the overall approach termed Canonical Cortical Graph Neural networks. This is shown to outperform recent state-of-the-art models in terms of clean audio reconstruction and energy efficiency for a benchmark audio-visual speech dataset. The enhanced performance is demonstrated through a reduced and smother neuron firing rate distribution. suggesting that the proposed model is amenable for speech enhancement in future audio-visual hearing aid devices.

1. Introduction

According to the World Health Organization (WHO), the number of people requiring hearing rehabilitation in the world is estimated to rise from 430 million nowadays up to 700 million until 2050, with nearly 2.5 billion people presenting at least some degree of hearing impairment [22]. Despite the impairment itself, deafness also impacts on social relationships [19,16] and environment perception [21], leading to other psychological and health conditions [15]. Over such circumstances, employing high-end energy-efficient technological approaches to build cross-modal sensory devices, i.e., combining audio and visual inputs to enhance hearing aid devices, seems a plausible way to improve individuals' life quality. In the last decades, machine learning-based techniques have shown themselves as a suitable approach to tackle issues related to virtually any field of science, industry, or even daily life,

ranging from computer vision [34] to medicine [28], and satellite imagery processing [12,27]. Machine learning also has been successfully employed in the context of speech enhancement [33], whose aim is to enhance speech quality and intelligibility when noise degrades them significantly [5]. Nevertheless, such methods may suffer massive performance degradation in the presence of overwhelming noise [7]. Many works address this problem using multimodal audio-visual (AV) information fusion. Combining AV information usually demands more sophisticated approaches, which intrinsically comprises several challenges, like data alignment, finding semantic gaps between low-level features and high-level information [8], and learning coherent and correlated latent patterns on different input modalities.

Combining noisy audio and clean images for clean signal reconstruction is analogous to reading the lips and body movements of a speaker in a boisterous environment, e.g., a pub with loud music, to obtain some additional information and create a context to enhance the information quality of a speech suppressed by the loud sound. Ngiam et at. [20], for instance, proposed a multimodal method capable of improving a target modality feature representation. Further, Adeel et al. [2] provided several improvements to the field, presenting a chaotic model for lip-reading integrating Internet of Things (IoT) and 5G Cloud-Radio Access Network, further

improving AV information transmission for real-time speech reconstruction. The authors proposed a deep learning-based approach that exploits AV cues to estimate clean audio [4], also considering distinct language speakers in similar work [13].

Recently, Passos et al. [24] proposed a multimodal self-supervised Graph Neural Network (GNN) that combines AV data through using Canonical Correlation Analysis Graph Neural Networks (CCA-GNN) [35], also modeling the temporal information in the graph using the so-called prior-frame positional encoding. The method obtained outstanding results considering audio reconstruction and energy efficiency, analyzed in terms of neuronal activation rate.

Despite the advantages presented [24], the model lacks some points in the context of a biologically plausible approach. In this context, Passos et al. [23] proposed a multimodal approach using burst-dependent learning [25], a method inspired by more recent studies on the physiological mechanism of pyramidal neurons that regulates the learning by the frequency of bursts, where the credit assignment problem is addressed by the primary principles of pyramidal neurons suggested by Körding and König [18]. In parallel, the study of canonical cortical circuits [9,14] provides some interesting insights regarding the brain procedure toward multimodal information processing. Canonical cortical circuits model pyramidal neurons to receive different modalities of information, modulated in an excitatory or inhibitory fashion on deeper layers. Moreover, biologically plausible models should not underestimate the importance of memory in the learning process, which performs a fundamental role in information inference, acting as an intrinsic context for novelty comprehension.

The attributes mentioned above motivated the development of the Canonical Cortical Graph Neural Network, a novel self-supervised architecture that remodels and improves the ideas developed in [24] by introducing a more biologically plausible approach to modulating and filtering the multimodal information inside the so-called cortical graph layers. It also introduces a memory concept inside each cortical graph block, composed of a mechanism to "forget" irrelevant facts and update itself considering every consecutive node, which is presented in a logical time-step sequence using prior-frame-based positional encoding. Experiments conducted over the AV ChiME3 [5] dataset show the Canonical Cortical Graph Neural Network obtained state-of-the-art results, outperforming CCA-GNN in the contexts of faster convergence, audio reconstruction, and firing neuron activation rates. Regarding the latter, the model not only obtained lower rates of activation, but also produced smothier firing-rate distributions.

The main contributions of this paper are presented as follows:

1. To propose the Canonical Cortical Graph Neural Network, a biologically plausible model for multimodal information feature extraction.
2. To introduce a novel paradigm for intra-layer information fusion and memory modeling.
3. To provide a self-supervised energy-efficient model for correlated feature extraction for AV-based clean audio data reconstruction.
4. To foster the literature regarding speech enhancement and AV hearing aids.

The remainder of this paper is presented as follows. Section 2 provides a theoretical background regarding Graph Neural Networks with Canonical Correlation Analysis and the prior frame-based graph positional encoding, while Section 3 introduces the novel Canonical Cortical Graph Neural Network. Further, Sections 4 and 5 present the methodology and the experimental results, respectively. Finally, Section 6 states conclusions and future works.

2. Theoretical Background

This section provides a brief theoretical background regarding Graph Neural Networks with Canonical Correlation Analysis and the prior frame-based graph positional enconding.

2.1. Graph Neural Networks with Canonical Correlation Analysis

Let $G = (X, A)$ be a graph where $A \in \mathbb{R}^{N \times N}$ represents the adjacency matrix and $X \in \mathbb{R}^{N \times F}$ the input data represented by graph nodes. Additionally, F represents the feature space dimension, and N denotes the number of nodes. The CCA-GNN [35] comprises three main steps, i.e., a random graph generator T, a graph neural network encoder f_θ, where θ denotes the network's learnable weights, and an objective function based on Canonical Correlation Analysis. The graph generator produces two augmented versions of the same graph, which are presented to the graph neural network encoder for further computing and maximizing the canonical correlation between their outputs.

The idea behind such an approach is discarding decorrelated components while preserving correlated ones. In a nutshell, the model tries to keep the more significant information present in both augmented versions and to avoid individual behaviors, such as anomalies and noise. Fig. 1 depicts the Canonical Correlation Analysis Graph Neural Network.

The graph augmentation process comprises the same approach presented in [36,29], which conducts a random feature masking and edge dropping. In this context, each $t_i \sim T$ comprises a distinct view, i.e., a transformed version of G, sampled at each iteration i.

The encoder is composed of a two-layered GNN but can be easily replaceable by any fancier model. The target function is designed to model the learning process as a canonical correlation maximization problem [10] using a self-supervised approach that considers two normalized views, $\mathbf{Z}_A$ and $\mathbf{Z}_B$, produced over randomly augmented versions of the original graph. The objective is to maximize the correlation between these views, formalized as follows:

$$\mathscr{L}(\mathbf{Z}_A, \mathbf{Z}_B) = ||\mathbf{Z}_A - \mathbf{Z}_B||_F^2 + \lambda\left(||\mathbf{Z}_A^T\mathbf{Z}_A - \mathbf{I}||_F^2 + ||\mathbf{Z}_B^T\mathbf{Z}_B - \mathbf{I}||_F^2\right), \quad (1)$$

where λ is a non-negative trading-off hyperparameter, and I is the identity matrix. The left term indicates the invariance term, which is responsible for minimizing the invariance between the two views. In contrast, the term on the right side describes the decorrelation term, which facilitates distinct features to capture different semantics through a regularization procedure.

The terms in Eq. 1 can be decomposed using a variance–covariance perspective [30]. Let s be an augmented version of the graph sampled from an input $\mathbf{x}$, and $\mathbf{Z}_s$ is a view of s obtained from a decoder output. The invariance term is minimized using expectation, described as follows:

$$\mathscr{L}_{inv} = ||\mathbf{Z}_A - \mathbf{Z}_B||_F^2 = \sum_{i=1}^{N}\sum_{k=1}^{D}(z_{i,j}^A - z_{i,j}^B)^2$$
$$\cong \mathbb{E}_{\mathbf{x}}\left[\sum_{k=1}^{D}\mathbb{V}_{s|\mathbf{x}}[\mathbf{Z}_s, k]\right] * 2N, \quad (2)$$

where $\mathbb{V}$ denotes the variance. Similarly, one can formalize the decorrelation term as follows:

$$\mathscr{L}_{dec} = ||\mathbf{Z}_S^T\mathbf{Z}_S - \mathbf{I}||_F^2 = ||\mathbf{Cov}_s[\mathbf{z}] - I||_F^2 \cong \sum_{i \neq j}(\rho_{i,j}^{\mathbf{z}_s})^2, \forall \mathbf{Z}_S \in \{\mathbf{Z}_A, \mathbf{Z}_B\},$$

$$(3)$$

where ρ denotes the Pearson correlation coefficient, and $\mathbf{Cov}$ stands for the covariance matrix.
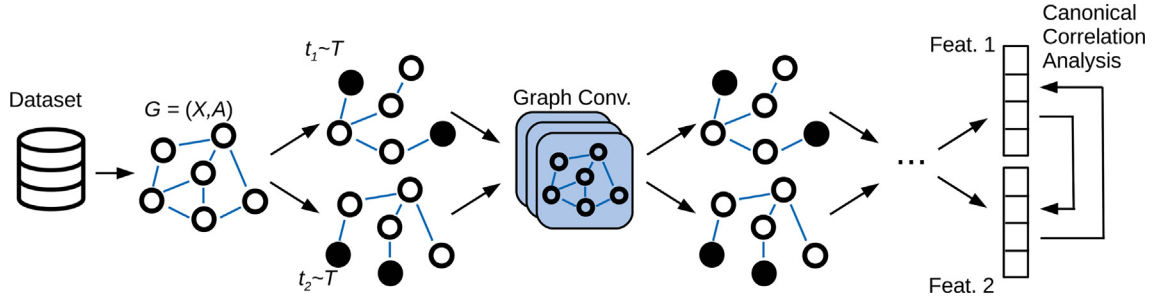
Fig. 1. Canonical Correlation Analysis Graph Neural Network. Each sample represents a node in a graph whose edges describe the relationship between pairs of samples. The random graph generator T produces two augmented versions of this graph, which are employed to feed the GNN model. The output of both versions is compared using the canonical correlation analysis, and the network parameters are adjusted to maximize this metric.

2.2. Prior Frame-based Graph Positional Enconding

A time-sequence based approach for graph positional encoding was recently proposed by Passos et al. [24]. The method computes the node neighborhood by connecting them to their k prior frame nodes in time sequence and attributing connection weights to the edges according to their distances in this time-based space. Fig. 2 depicts this idea.

The calculation of the edge weight w_{ij} that connects a node i to a node j is performed as follows:

$$w_{ij} = k + 1 - d_{ij}, \tag{4}$$

where d_{ij} denotes the distance from node i to node j in a frame-step space. Those values are stored in a distance matrix used to compute the positional encoding of the nodes.

3. Canonical Cortical Graph Neural Network

This paper presents a novel self-supervised approach for training multi-modal graph neural networks in a more biologically plausible way. In this context, the architecture combines several concepts inspired on cortical circuits observed in the brain [14,9] to model memory and multimodal information fusion with canonical correlation analysis [10], and to maximize the correlation of the information extracted from different inputs. Fig. 3 presents a general overview of the proposed model, which suggests the more interesting procedures of the model are implemented at layer level. Therefore, Fig. 4 illustrates this in-layer process, also depicting the behavior of each operation.

A formal description of the procedures depicted in Fig. 4 and performed inside each cortical layer is provided bellow. Firstly, one should compute the audio f_a, visual f_v, memory f_m, and modulation f_w filters as follow:

$$f_a = \sigma(\mathbf{W}_a h_a + b_a), \tag{5}$$
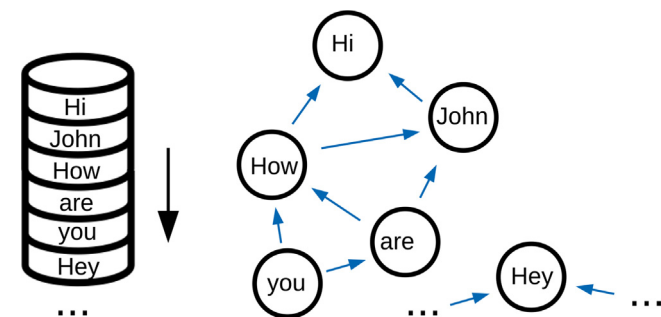
$$f_v = \sigma(\mathbf{W}_v h_v + b_v), \tag{6}$$

$$f_m = \sigma(\mathbf{W}_m[h_a, h_v] + b_m), \tag{7}$$

and

$$f_w = \sigma(\mathbf{W}_w[h_a, h_v] + b_w), \tag{8}$$

where σ stands for the Sigmoid function, $\mathbf{W}_a, \mathbf{W}_v, \mathbf{W}_m$, and $\mathbf{W}_w$, are the weight matrices for the audio, visual, memory, and modulation filters, respectively, and b_a, b_v, b_m, and b_w are the biases for the audio, visual, memory, and modulation filters, respectively, while $[h_a, h_v]$ denotes the concatenation of the audio h_a and the visual h_v graph convolution outputs. Further, the pre-modulation ρ is computed as follows:

$$\rho = tanh(\mathbf{W}_\rho[h_a, h_v] + b_\rho), \tag{9}$$

where $\mathbf{W}_\rho$ and b_ρ are the pre-modulation weight matrix and bias, respectively. Finally, the modulation ω is computed as follows:

$$\omega = f_w \otimes \rho, \tag{10}$$

where $\otimes$ stands for the dot product.

The memory μ is the more tricky updating since it considers both "forgetting" irrelevant memories using the memory filter f_m and introducing new experiences presented in the modulation ω. Moreover, the operation is performed individually for each node $n \in \{0, \ldots, N\}$ since they are connected in a logical temporal sequence established by the prior frame positional encoding [24], described as follows:

$$\mu^n = \omega^n + (f_m^n \otimes \mu^{n-1}), \tag{11}$$

where μ^n, ω^n, and f_m^n are the node's n memory, modulation, and memory filter. In the sequence, the memory is updated as follows:

$$\mu = tanh(\mathbf{W}_\mu \mu + b_\mu), \tag{12}$$

where $\mathbf{W}_\mu$ and b_μ are the memory weight matrix and bias, respectively. Finally, the layer output, i.e., the new node representation of the audio and visual graphs h_a' and h_v', respectively, are computed as follows:

$$h_a' = \mu \otimes f_a \tag{13}$$

and

$$h_v' = \mu \otimes f_v. \tag{14}$$

Finally, h_a' and h_v' become the node representations of the subsequent layer audio and visual graphs, respectively, in case of an intermediate layer. Regarding the output layer, the model performs the canonical correlation analysis between h_a' and h_v' using Eq. 1 and backpropagates this value to optimize the network parameters.

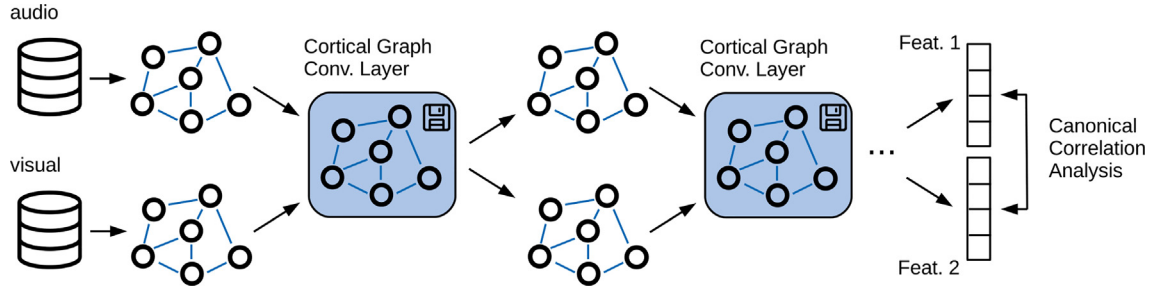**Fig. 2.** Prior frame-based node neighborhood modeling considering 2 prior frames.

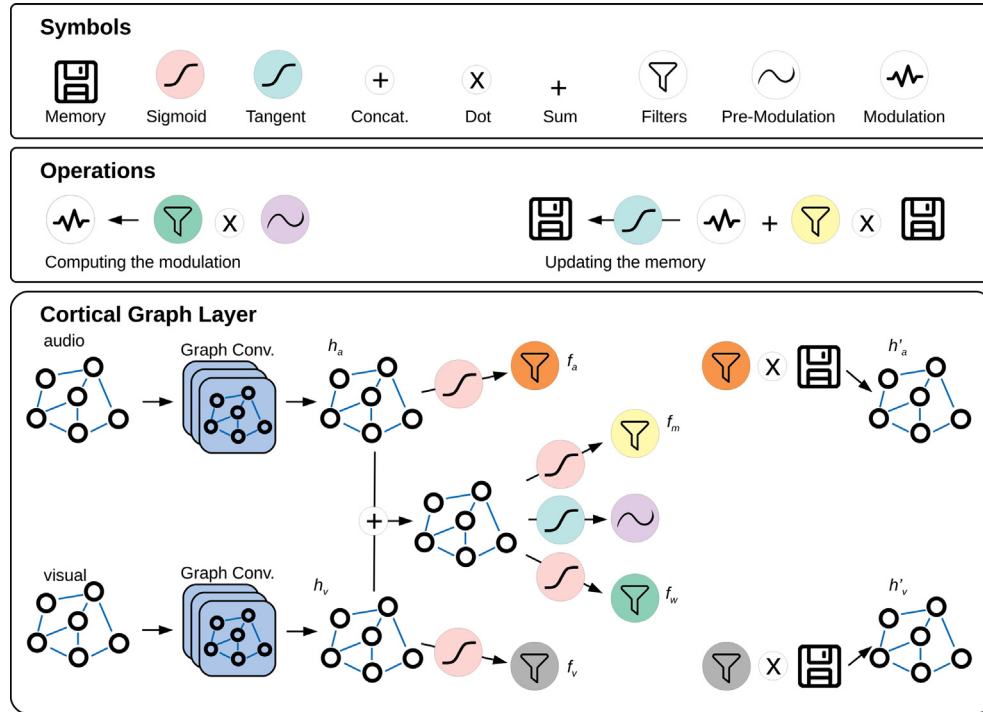**Fig. 3.** Canonical Cortical Graph Neural Network.

Fig. 4. Cortical graph layer. The top and middle frames describe the symbols and operations employed in the cortical graph block, respectively. The bottom block depicts the layer's pipeline.

4. Methodology

This section describes the dataset and configuration employed during the experiments.

4.1. AV ChiME3 Dataset

The dataset used in this paper aims to combine environmental information, i.e., audio and visuals, to train an efficient model for enhancing and amplifying clean audio signals considering multi-modal data. The dataset comprises triples composed of image, clean audio, and noisy audio signals, from which the image and noisy audio denote the model's input while the clean signal stands for the desired output, i.e., the instance target in the context of supervised learning. The videos are extracted from Grid [11] data-set, in which different classes of noises (public transport, pedes-trian area, street junction, cafe) with signal to noise ratios (SNR) ranging from −12 to 12 dB extracted from ChiME3 [6] are intro-duced, composing the AV ChiME3 [5] dataset. Further, the samples are preprocessed to improve the sentence alignment and incorpo-rate multiple visual frames to include temporal data. In total, the dataset contains 989 sequences from 5 different speakers, described as one black male, two white males, and two white females. Each sequence comprises 48 frames, summing up to a total of 47,472 synchronized triples of samples.

4.1.1. Audio feature extraction

Log-FB vectors were employed to extract both clean and noisy audio features. The technique samples the audio signal at 22,050kHz for further segmenting it into M 16 ms frames with 800 samples per frame and 62.5% increment rate. Furthermore, it uses a hamming window and Fourier transformations to produce a 2048-bin power spectrum. Moreover, it employs a logarithmic compression to obtain the 22-dimensional log-FB signals.

4.1.2. Visual feature extraction

The samples extracted to generate the visual set of features were obtained using an encoder-decoder architecture over the Grid Corpus dataset. Lip-regions were detected using Viola-Jones [31] algorithm, for further tracking the frame sequence using the method proposed in [26]. Additionally, a manual effort was employed to randomly inspect the sentences, ensuring good lip tracking [1]. The encoder-decoder approach is then used to create vectors of pixel intensities, in which the 50 first components are

vectorized in a zigzag order and then interpolated to match the equivalent audio sequence.

4.2. Experimental Setup

The experiments conducted in this paper compare the proposed Canonical Cortical Graph Neural Networks against the recent state-of-the-art model CCA-GNN for the task of multimodal clean audio reconstruction based on noisy audio and visual features, extracted using logFB and an encoder-decoder approach, respectively, as described in Section 4.1. Both models comprise a similar architecture composed of two hidden layers, the first comprising 512 and the second 256 neurons. The parameters were selected empirically. The learning is conducted to maximize the canonical correlation analysis for coherent feature extraction during 200 epochs using the Adam optimizer with a learning rate of 10^{-3} and a trading-off parameter of $\lambda = 0.0001$ (Eq. 1). At the same time, the hyperparameters of the CCA-GNN follow the configuration employed in [24]. Finally, the graphs are generated considering eight distinct prior-frame scenarios, i.e., $k \in [3, 5, 7, 10, 15, 20, 25, 30]$. Notice that the plots provided in experimental results comprise only $k \in [3, 10, 30]$ to illustrate the difference between a low, medium, and a high number of neighbors/prior frames.

After maximizing the canonical correlation analysis between the two channels, the features extracted by the networks are employed to feed a dense layer responsible for reconstructing the clean signal by minimizing the mean squared error as the cost function. The model is optimized during 2,000 epochs using the Adam optimizer with a learning rate of 0.005 and a weight decay of 0.0004.

The dataset was divided into 20 folds to provide an in-depth statistical analysis. Each fold comprises 50 sequences of 48 frames each, summing up to 2,400 samples per fold. As stated in Section 4.1, the dataset is formed by three subsets: (i) clean audio, (ii) noisy audio, and (iii) clean visual. The noisy audio and the clean visual input are used to feed the multimodal GNNs, while the clean audio is considered the reconstruction target. Finally, each fold is split into train, validation, and test sets, following the proportions of 60%, 20%, and 20%, respectively. The Wilcoxon signed-rank test [32] with 5% of significance was considered for statistical evaluation.

5. Experiments

This section exploits the superiority of the proposed Canonical Cortical Graph Neural Networks over the state-of-the-art CCA-GNN. The experimental results consider the contexts of feature extraction analysis, clean audio signal reconstruction, as well as neuronal activation and energy efficiency.

5.1. Feature Extraction Analysis

This section explores the task of self-supervised feature extraction in terms of canonical correlation analysis. The idea consists of extracting correlated features considering both the audio and visual channels, contributing to better features for clean audio reconstruction. In this context, the Canonical Cortical Graph Neural Networks, presented in Fig. 5 as Cortical, showed a performance 75% higher than the CCA-GNN, on average, considering a small, medium, and high neighborhood, i.e., $k \in [3, 10, 30]$.

5.2. Clean Signal Reconstruction

Fig. 6 provides some insights regarding the convergence of the model regarding the Mean Squared Error (MSE) over the testing
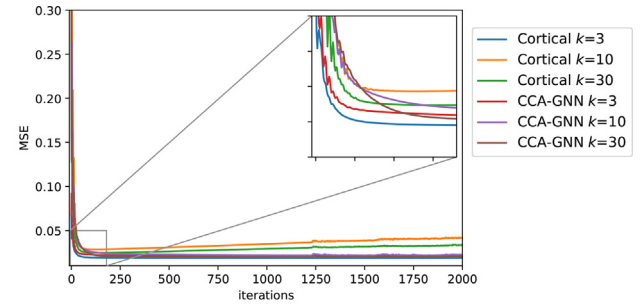

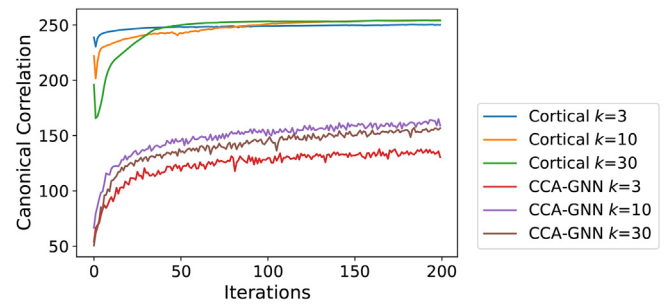**Fig. 5.** Canonical correlation analisys maximization.

Fig. 6. Clean audio reconstruction error convergence considering the proposed model (Cortical) and CCA-GNN architectures based on the noisy audio and clean visuals.

set reconstruction. In this context, one can observe that (i) the proposed model (Cortical) converges a bit faster than CCA-GNN (the blue line when $k = 3$) and remain the best result until completing the 2,000 epochs; (ii) the convergence is faster for small values of k, i.e., $k = 3$, for both Cortical and CCA-GNN; (iii) the proposed model presents an overfitting behavior for larger values of k, described by the orange ($k = 10$) and green ($k = 30$) lines.

Further, Table 1 provides the final MSE values over the testing set considering eight distinct k scenarios, i.e., $k \in [3, 5, 7, 10, 15, 20, 25, 30]$. Notice bold values stand for the best values considering the Wilcoxon signed-rank test [32] with a significance of 5%. Such results show that the proposed model presents a better behavior when exposed to a reduced number of neighbors, i.e., the historical information is reduced, obtaining the lowest MSE overall over this scenario. This result can be explained by the memory implemented in the architecture, i.e., since the memory tries to model the predictions based on past frames, a longer temporal sequence makes this information ambiguous, leading the model to an extra exposure to past instances, thus overfitting. The idea is reinforced by the CCA-GNN approach, where an opposite behavior is observed in most

Table 1
Average Mean Squared Error and standard deviation over Canonical Cortical GNN and CCA-GNN considering clean audio reconstruction given noisy audio and visual inputs over two block layers network architecture.

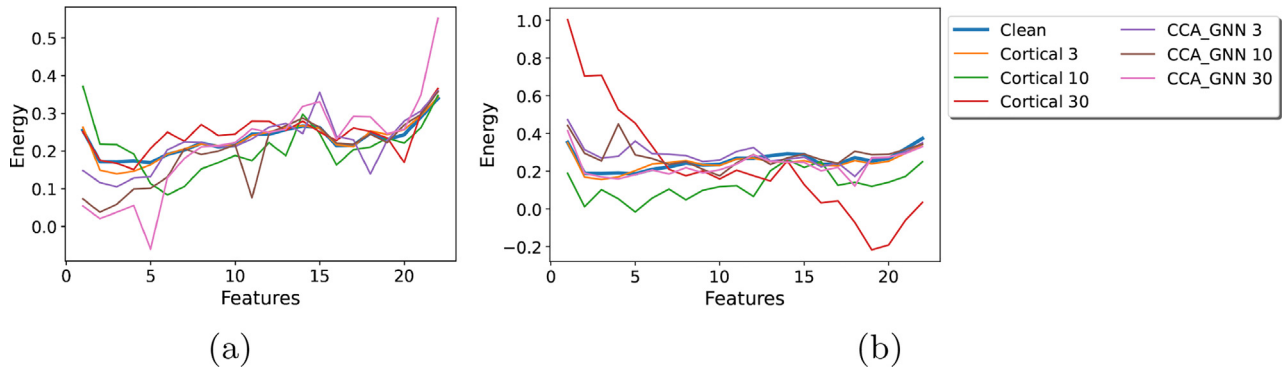| Neighbors	CCA GNN	Cortical
3	0.0204 ± 0.0048	0.0188 ± 0.0041
5	0.0208 ± 0.0049	0.0200 ± 0.0057
7	0.0214 ± 0.0052	0.0243 ± 0.0067
10	0.0216 ± 0.0052	0.0282 ± 0.0059
15	0.0216 ± 0.0048	0.0278 ± 0.0049
20	0.0205 ± 0.0045	0.0266 ± 0.0048
25	0.0200 ± 0.0045	0.0262 ± 0.0056
30	0.0199 ± 0.0045	0.0246 ± 0.0058

Fig. 7. Two examples of clean audio signal reconstruction considering the proposed approach (Cortical) and the baseline, i.e., CCA-GNN.

cases, i.e., since CCA-GNN does not implement a memory, higher numbers of neighbors usually lead to lower prediction errors.

Finally, Fig. 7 depicts some examples of clean audio reconstruction for both models. Notice that the Canonical Cortical Graph Neural Networks with a reduced number of past frames, i.e., $k = 3$, obtained almost perfect reconstructions of the clean signal in both cases, which reinforces the idea that the memory implementation replaces the necessity of longer temporal information, described by a higher number of neighbors. Fig. 6(b) shows that CCA-GNN is also capable of producing good representations when then the number of past frames is big enough, i.e., $k = 30$, even though the reconstruction is not as good as the proposed method.

5.3. State-of-the-art Comparison

Table 2 provides a comparison of the Canonical Cortical Graph Neural Network against the state-of-the-art results reported in recent works using the AV ChiME3 dataset and multimodal

Table 2
Comparison of the Canonical Cortical Graph Neural Network against state-of-the-art results reported in recent approaches for audio-visual speech reconstruction in terms of ean squared error.

Technique	Best value
Cortical (ours)	0.0188
CCA-GNN	0.0199
CCA-MLP [24]	0.0189
MLP [5]	0.0200
LSTM [3]	0.0780
CC-STOI DL [17]	1.2770

approaches for audio-visual speech enhancement, i.e., CCA-GNN and CCA Multilayer Perceptron (CCA-MLP) [24], a Long-Short Term Memory (LSTM) [3] and a Multilayer Perceptron (MLP) [5] based approaches, as well as a canonical correlation-based short-time objective intelligibility deep learning (CC-STOI DL) [17] method. Notice the proposed approach provided the most accurate results, outperformed all the compared tecniques.

5.4. Neuronal Activation Analysis

The neuronal activation rate analysis is fundamental for future audio-visual hearing aids since the metric reflects the model's energy efficiency, a critical feature considering energy-constrained environments like hearing aids and embedded devices. In this context, reducing the firing rate directly implies a reduction in energy consumption. Fig. 8 depicts the neuron activation rate over the intermediate block layers concerning the audio and visual channels. Notice that both audio and video channels obtained pretty similar values concerning the Cortical model, suggesting that the two tracks contribute almost equally to the results.

A numerical representation is presented in Table 3, which comprises a more complete set of neighbors (past frames), i.e., $k = [3, 5, 7, 10, 15, 20, 25]$. The results reinforce the idea that the proposed model relies almost equally upon both noisy audio and visual modalities for the final output since their areas under the curve are practically the same for all cases, except for $k = 10$, with an irrelevant difference, i.e., 256 and 255 for noisy audio and visual, respectively. The results also show that the proposed model outperforms the baseline in the context of neuronal activation and, consequently, energy consumption in every possible scenario, showing itself as a suitable approach for speech enhancement in
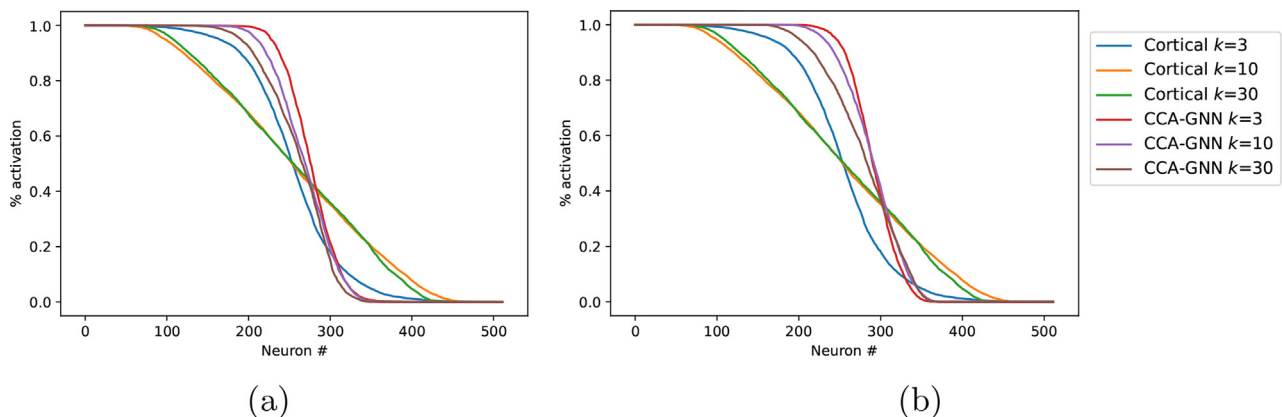

Fig. 8. Neuron activation rate considers the multimodal architecture's hidden block over (a) noisy audio and (b) visual channels.

Table 3
Area under the curve considering the multimodal architecture's hidden block.

Neighbors	Audio		Visual	
	Cortical	CCA-GNN	Cortical	CCA-GNN
3	255	276	255	291
5	253	265	253	288
7	256	266	256	288
10	256	268	255	290
15	255	268	255	288
20	258	267	258	285
25	251	267	251	288
30	255	260	255	280

future audio-visual hearing aid devices considering both accuracy and energy performance.

6. Conclusion

This paper proposed a novel self-supervised method for multimodal correlated feature extraction through canonical correlation analysis maximization. The proposed model comprises a block-based neural network, where each block comprises two Graph Neural Networks layers, i.e., one for noisy audio and the other for the visual features, a memory, and a set of operations to filter, insert, delete, and modulate the input signals. Such operations are inspired in recent discoveries related to cortical cells and their interactions.

Experiments were conducted over the AV ChiME3 dataset, designed for the task of multimodal clean audio reconstruction considering noisy audio and clean visual instances, compared the proposed approach against the CCA-GNN, a similar state-of-the-art method proposed recently for the task. Results show that the proposed Canonical Cortical GNN provides more coherent and better-quality features, reaching higher values of canonical correlation analysis. The proposed approach also obtained more accurate reconstructions, generating cleaner reconstructions. Moreover, the model delivers higher efficiency in terms of energy, evaluated by neurons' firing rate. Finally, it also showed itself to be less dependent on a more extended prior-frame sequence, i.e., high values of k, since the memory can store and track temporal information.

Regarding future work, we aim to extend a similar cortical-based architecture to Convolutional Neural Networks and applications to two-dimensional data. We also aim to implement the model on chips for training and inference for possible future implementation in hearing aid devices for AV speech enhancement.

CRediT authorship contribution statement

Leandro A. Passos: Conceptualization, Methodology, Software, Writing - original draft, Visualization, Investigation, Writing - review & editing. **João Paulo Papa:** Visualization, Writing - review & editing. **Amir Hussain:** Writing - review & editing. **Ahsan Adeel:** Supervision, Data curation, Writing - original draft, Visualization, Writing - review & editing, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Abel, R. Marxer, J. Barker, R. Watt, B. Whitmer, P. Derleth, A. Hussain, A data driven approach to audiovisual speech mapping, in: International Conference on Brain Inspired Cognitive Systems, Springer, 2016, pp. 331–342.

[2] A. Adeel, J. Ahmad, A. Hussain, Real-time lightweight chaotic encryption for 5g iot enabled lip-reading driven secure hearing-aid. arXiv preprint arXiv:1809.04966 (2018).

[3] A. Adeel, J. Ahmad, H. Larijani, A. Hussain, A novel real-time, lightweight chaotic-encryption scheme for next-generation audio-visual hearing aids, Cognitive Computation 12 (3) (2020) 589–601.

[4] A. Adeel, M. Gogate, A. Hussain, Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments, Information Fusion 59 (2020) 163–170.

[5] A. Adeel, M. Gogate, A. Hussain, Whitmer, W.M.: Lip-reading driven deep learning approach for speech enhancement. IEEE Transactions on Emerging Topics, Computational Intelligence (2019).

[6] J. Barker, R. Marxer, E. Vincent, S. Watanabe, The third 'chime'speech separation and recognition challenge: Dataset, task and baselines, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 504–511.

[7] J. Benesty, J. Chen, Habets, E.A.: Speech enhancement in the STFT domain, Springer Science & Business Media, 2011.

[8] M.U. Bokhari, F. Hasan, Multimodal information retrieval: Challenges and future trends, International Journal of Computer Applications 74 (14) (2013).

[9] F. Capone, M. Paolucci, F. Assenza, N. Brunelli, L. Ricci, L. Florio, V. Di Lazzaro, Canonical cortical circuits: current evidence and theoretical implications, Neuroscience and Neuroeconomics 5 (2016) 1–8.

[10] X. Chang, T. Xiang, T.M. Hospedales, Scalable and effective deep cca via soft decorrelation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1488–1497 (2018).

[11] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, The Journal of the Acoustical Society of America 120 (5) (2006) 2421–2424.

[12] K.A. da Costa, J.P. Papa, L.A. Passos, D. Colombo, J. Del Ser, K. Muhammad, V.H. C. de Albuquerque, A critical literature survey and prospects on tampering and anomaly detection in image data, Applied Soft Computing (2020).

[13] M. Gogate, K. Dashtipour, A. Adeel, A. Hussain, Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement, Information Fusion 63 (2020) 273–285.

[14] S. Grossberg, A canonical laminar neocortical circuit whose bottom-up, horizontal, and top-down pathways control attention, learning, and prediction, Frontiers in Systems Neuroscience 15 (2021).

[15] A.S. Helvik, G. Jacobsen, L.R. Hallberg, Psychological well-being of adults with acquired hearing impairment, Disability and rehabilitation 28 (9) (2006) 535–545.

[16] A.R. Huang, J.A. Deal, G.W. Rebok, J.M. Pinto, L. Waite, F.R. Lin, Hearing impairment and loneliness in older adults in the united states, Journal of Applied Gerontology 40 (10) (2021) 1366–1371.

[17] T. Hussain, M. Diyan, M. Gogate, K. Dashtipour, A. Adeel, Y. Tsao, A. Hussain, A novel speech intelligibility enhancement model based on canonicalcorrelation and deep learning. arXiv preprint arXiv:2202.05756 (2022).

[18] K.P. Körding, P. König, Supervised and unsupervised learning with two sites of synaptic integration, Journal of computational neuroscience 11 (3) (2001) 207–215.

[19] S.E. Kramer, T.S. Kapteyn, J.M. Festen, H. Tobi, Factors in subjective hearing disability, Audiology 34 (6) (1995) 311–320.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, In ICML (2011).

[21] W. Noble, Self-assessment of hearing and related function, Wiley-Blackwell, 1998.

[22] W.H. Organization, et al.: Hearing screening: considerations for implementation (2021).

[23] L.A. Passos, A. Khubaib, M. Raza, A. Adeel, Multimodal speech enhancement using burst propagation. arXiv preprint arXiv:2209.03275 (2022).

[24] L.A. Passos, J.P. Papa, J. Del Ser, A. Hussain, A. Adeel, Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement, Information Fusion (2022).

[25] A. Payeur, J. Guerguiev, F. Zenke, B.A. Richards, R. Naud, Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits, Nature neuroscience 24 (7) (2021) 1010–1019.

[26] D.A. Ross, J. Lim, R.S. Lin, M.H. Yang, Incremental learning for robust visual tracking, International journal of computer vision 77 (1) (2008) 125–141.

[27] D.F. Santos, R.G. Pires, L.A. Passos, J.P. Papa, Ddipnet and ddipnet+: Discriminant deep image prior networks for remote sensing image classification. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 2843–2846. IEEE (2021)

[28] R.W. de Souza, D.S. Silva, L.A. Passos, M. Roder, M.C. Santana, P.R. Pinheiro, V.H. C. de Albuquerque, Computer-assisted parkinson's disease diagnosis using fuzzy optimum-path forest and restricted boltzmann machines, Computers in Biology and Medicine 131 (2021).

[29] S. Thakoor, C. Tallec, M.G. Azar, R. Munos, P. Veličković, M. Valko, Bootstrapped representation learning on graphs. arXiv preprint arXiv:2102.06514 (2021).

[30] Y. Tian, X. Chen, S. Ganguli, Understanding self-supervised learning dynamics without contrastive pairs. arXiv preprint arXiv:2102.06810 (2021).

[31] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. vol. 1, pp. I-I. Ieee (2001).

[32] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (6) (1945) 80–83, https://doi.org/10.2307/3001968.

[33] Y. Xu, J. Du, L.R. Dai, C.H. Lee, A regression approach to speech enhancement based on deep neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (1) (2014) 7–19.

[34] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, X. Liu, A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection, IEEE Transactions on Instrumentation and Measurement 71 (2022) 1–14.

[35] H. Zhang, Q. Wu, J. Yan, D. Wipf, S.Y. Philip, From canonical correlation analysis to self-supervised graph neural networks. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021).

[36] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, L. Wang, Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131 (2020).

Leandro A. Passos Received his doctorate in Computer Science at Federal University of São Carlos (UFSCar) in 2018. He is currently a postdoctoral researcher associate at the University of Wolverhampton, UK. Has experience in Machine Learning, acting on subjects related, classification, optimization, brain-inspired neurons, and information fusion, among others.

João Paulo Papa Received his Ph.D. in Computer Science from the University of Campinas, SP, Brazil. During 2008–2009, he had worked as a post-doctorate researcher at the same institute, and during 2014–2015 he worked as a visiting scholar at Harvard University. He has been a Professor at the Computer Science Department, São Paulo, State University, since 2009. He was also the recipient of the Alexander von Humboldt research fellowship in 2017.

Amir Hussain obtained his B.Eng. (with the highest 1st Class Honors) and Ph.D. (in novel neural network architectures and algorithms) from the University of Strathclyde in Glasgow, Scotland, UK, in 1992 and 1997 respectively. Following postdocoral and academic positions at the University of West of Scotland (1996–98), University of Dundee (1998–2000), and University of Stirling (2000–2018) respectively, he joined Edinburgh Napier University, in Scotland, UK, in 2018, as Professor of Computing Science, and founding Director of the Cognitive Big Data and Cybersecurity (CogBiD) Research Laboratory. His research interests are cross-disciplinary and industry focussed, and include secure, explainable and context-aware 5G-IoT driven AI, and multi-modal cognitive computing techniques and real-world applications. He has published more than 400 papers, including over a dozen books and around 150 journal papers. He has led major national, European and international projects and supervised more than 30 PhD students, He is founding Editor-in-Chief of two leading journals: Cognitive Computation (Springer Nature), and BMC Big Data Analytics (BioMed Central); and Chief-Editor of the Springer Book Series on Socio-Affective Computing, and Cognitive Computation Trends. He has been appointed invited Associate Editor of several prestigeous journals, including the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Emerging Topics in Computational Intelligence, (Elsevier) Information Fusion and the IEEE Computational Intelligence Magazine. Amongst other international distinctions, he is General Chair of the IEEE WCCI 2020 (comprising IJCNN, IEEE CEC and FUZZ-IEEE), Vice-Chair of the Emergent Technologies Technical Committee of the IEEE Computational Intelligence Society (CIS), and Chapter Chair of the IEEE UK and RI Industry Applications Society.

Ahsan Adeel holds B. Eng. (Hons), M.Sc. (EEE), and Ph.D. (Cognitive Computing) degrees. Following an EPSRC/MRC fellowship at the University of Stirling (2016–18), he is currently a Lecturer (Assistant Professor) in Artificial Intelligence at the University of Wolverhampton, UK, where he is leading the Conscious Multisensory Integration (CMI) Lab. He has been a Visiting Fellow at MIT Synthetic Intelligence Lab and Computational Neuroscience Lab (University of Oxford). His ongoing multidisciplinary research aims to explore and exploit the power of advanced AI to design unorthodox brain-inspired cognitive computing architectures by integrating suitable deep machine learning, reasoning, and optimization algorithms. His key areas of interest include biophysical and hardware-efficient neural models, explainable artificial intelligence, optimized resource management, multimodal fusion, context-aware decision-making, low power 5G IoT devices, and neuromorphic chips.