Heliyon 10 (2024) e26583

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A mobile Deep Sparse Wavelet autoencoder for Arabic acoustic unit modeling and recognition

Sarah A. Alzakari [a], Salima Hassairi [b], Amel Ali Alhussan [a], Ridha Ejbali [b,*]

[a] Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[b] Research Team in Intelligent Machines, National School of Engineers of Gabes, B.P. W 6072, Gabes, Tunisia

ARTICLE INFO

Keywords:
Acoustic units
Mobile architecture
Wavelet networks
Deep learning
Deep sparse wavelet networks
Mel-frequency cepstral coefficients
Perceptual linear predictive
Stacked wavelet autoencoders

ABSTRACT

In this manuscript, we introduce a novel methodology for modeling acoustic units within a mobile architecture, employing a synergistic combination of various motivating techniques, including deep learning, sparse coding, and wavelet networks. The core concept involves constructing a Deep Sparse Wavelet Network (DSWN) through the integration of stacked wavelet autoencoders. The DSWN is designed to classify a specific class and discern it from other classes within a dataset of acoustic units. Mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive (PLP) features are utilized for encoding speech units. This approach is tailored to leverage the computational capabilities of mobile devices by establishing deep networks with minimal connections, thereby immediately reducing computational overhead. The experimental findings demonstrate the efficacy of our system when applied to a segmented corpus of Arabic words. Notwithstanding promising results, we will explore the limitations of our methodology. One limitation concerns the use of a specific dataset of Arabic words. The generalizability of the sparse deep wavelet network (DSWN) to various contexts requires further investigation “We will evaluate the impact of speech variations, such as accents, on the performance of our model, for a nuanced understanding.

1. Introduction

Speech recognition is one of the most important research areas. Several approaches related to this field have been developed. Several techniques have contributed to the evolution of speech recognition systems. The hidden Markov model (HMM), neural network (NN), and wavelet network (WN) are the most speech recognition methods. The WN combines the theories of wavelet and NN and consists of a feed-forward NN. The first WN was applied by Daugmann in image classification [1]. In addition, the concept of WNs has been introduced in the work of Pati [2] and Zhang [3,15]. They constituted by a hidden layer. Other works have been used for recognition systems such as deep learning (DL) [4,29,31,32], which can mirror brain signal processing. The architecture of this new concept has revolutionized machine learning (ML) methods [4,20]. For classification tasks, DL has been classified as representation and supervised learning method since 2006 [5,6]. Bengio and Hinton proposed the first deep learning algorithms for speech recognition. The Restricted Boltzmann Machine (RBM) and Deep Belief Network was proposed by Hinton [23] and autoencoder was

* Corresponding author.
E-mail addresses: saalzakari@pnu.edu.sa (S.A. Alzakari), salima.hassairi.tn@ieee.org (S. Hassairi), aalhussan@pnu.edu.sa (A. Ali Alhussan), ridha_ejbali@ieee.org (R. Ejbali).

https://doi.org/10.1016/j.heliyon.2024.e26583
Received 16 January 2024; Received in revised form 3 February 2024; Accepted 15 February 2024
Available online 19 February 2024
2405-8440/Â© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

introduced by Bengio [2,7,17]. The principle is to consider each two neighboring layers as RBM so that learning approaches a good solution. A back-propagation technique is applied to retune the results [8]. All of these approaches are suitable for desktop application. Unfortunately, these approaches are not suitable for mobiles devices due to low performances of its architectures [28].

In speech recognition systems, many feature extraction techniques are used by the researchers, such as the Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP). MFCCs parameters are considered the natural speech while extracting features [9]. PLP is a combination of spectral analysis and linear prediction analysis. They are based on the psychophysics of hearing to compute a simple auditory spectrum. Therefore, this study contributes to the literature by modeling acoustic units of speech using a hybrid algorithm based on DL, WN and sparse coding (SC). DL and WN are used to extract relevant features. SC is used to reduce connection of DSWN. This reduction decreases time and spaces resources to be suitable for mobile configuration. Our article is organized as follows: In Section II, we present the general context of the proposed approach. Section III demonstrated the proposed approach, the feature extraction techniques that were used. We also, present the ideas of a deep wavelet network (DWN) and a deep sparse wavelet network (DSWN) and how to model acoustic units. In Section VI, we present our experimental results and discussion. We resume our work in section V by a conclusion.

2. General state of the art

The authors of [16] developed a speech recognition approach based on a fast beta wavelet network model. It's a hybrid classifier structure, combining a neural network model as a general representation and wavelets acting as activation functions. The approach was divided into two distinct phases. The first part aimed to model each training signal using fast beta wavelet networks. The second one involved identifying a new test signal by comparing its fast beta wavelet network pattern to those of all training signals. The distances between these models were calculated to allow recognition of the desired signal. On the other hand, the authors of [18] used fuzzy logic approaches to improve the hybridizations of neural networks. While the authors of [27] and has completely changed the wavelet networks with deep networks. They demonstrate the effectiveness of models based on deep learning compared to old models based on neurons networks and its variations such as wavelet network. In the work [29], a survey on deep learning on mobile devices: applications, optimizations, challenges and research opportunities was carried out by the authors to compare a set of artificial intelligence approaches such as convolutional neural network, Long short-term memory and auto-encoders. Author of [31] applied deep neural networs approaches to languages Kazakh. While the author of [32] applied a dynamic time warping optimization algorithm to speech recognition of machine translation to show its effectiveness compared to other algorithms in the literature Table 1.

3. General context of the proposed approach

A. SPARSE CODING

Sparseness is the essential assumption on which sparse coding is based. Empirically, we speak of a sparse signal when it is possible to describe the signal from a small number of elementary components known a priori.

In order to increase the critical capacity, it is possible to reduce the density of a signal to be learned through sparse coding as identified in the cerebral cortex [24]. Parsimonious coding is a minimization of the number of neurons mobilized to represent information. As a result, the network better distinguishes between two items within its memory, thus tolerating additional message learning. In other words, parsimony encourages diversity. In signal processing, the notion of parsimony corresponds to maximizing the signal while minimizing the number of variables needed, grouped in a dictionary, to represent the signal as a linear combination [25]. Some neural networks such as the parsimonious auto-encoder [26] create these dictionaries by grouping the features necessary for parsimonious encoding of information.

In order to reduce the complexity of GWNs, we have thought of sparse coding. At the end, we will have optimal GWNs in terms of the number of wavelets in the hidden layer and the number of connections. The auto-encoder, resulting from the optimal GWN, will contain the minimum number of nodes and connections in the hidden layers.

B. GENERAL APPROACH

We identify the best wavelets that were used in these WNs. Knowing that we will do the classification of a class from the dataset

Table 1
List of study used differents studies for speech recognition.

Ref#	Year	coefficients	modeling	Classifier	Dataset	Classification Accuracy
[16]	2015	MFCC	Bêta wavelet network	Linear classifier	Recording of Arabic sentences from a text corpus	97%
[18]	2015	MFCC	FWN-FDSS	Fazzy approach	Recording of Arabic sentences from a text corpus	98,33%
[27]	2016	MFCC, PLP	Deep wavelet network	Linear classifier	Recording of Arabic sentences from a text corpus	95%
[29]	2022	MFCC	CNN	non-linear activation	Gurbani hymn dataset	89,5%
[31]	2024	MFCC	DNN	Linear	transcribed Kazakh speech	94%
[32]	2023	MFCC	DTW	Linear	Not indicated	92.8%
			DNN			91.5%
			ABC			92.3%

versus all the other classes, we will consider that the dataset contains two classes C1 and C2. C1 is subject to classification and C2 contains the rest of classes of the dataset. Therefore, we display in the form of a table all WNs for both classes (C1 and C2) [1,7,19,30].

The wavelets are then sorted according to their coefficient. The coefficient reflects the pertinence of the wavelet in the approximation of the signals of C1.

The idea is to favor the wavelets that are most used to approximate C1 signals and to penalize those that aren't. The result is a GWN that is a good approximator of C1 signals and very sensitive to other ones [1,19].

Coefficient scores calculated, Fig. 1, for each wavelet are used to choose pertinent wavelets to construct a GWN that approximates only in the best way every signal of C1 [19].

As shown in Fig. 2, the coefficients whose contributions are minimal or zero are eliminated during the reconstruction of the signals. This process is justified by the theory of sparse coding. This idea reduces the calculation time, the memory space and does not affect the quality of the reconstructed signals [7]. GSWN will be suitable for mobile resources.

4. Proposed approach

A. SYSTEM OVERVIEW

Fig. 3 illustrates the proposed method for modeling acoustic units of speech using DSWN.

According to this Fig. 3, the proposed system consists of three parts: features extraction, training phase, and recognition phase.

B. FEATURES EXTRACTION

The objective of features extraction is to extract form the speech signal a set of perceptually significant features [14]. Several signal analysis techniques can be used in this process. The proposed architecture is based on PLP and MFCCs [10]. We used a window of 32 m s to sweep the speech signals, with an overlap of 10 m s.

After windowing the signal, the fast Fourier transformation is calculated to extract the frequency components for each frame. Then, the logarithmic Mel-scaled filter bank is applied to each Fourier transform frame as follows:

$$frequency\ (Mel\ Scaled) = \left[2595 \Big/ log\left(1 + \frac{f(Hz)}{700}\right)\right] \tag{1}$$

As previously mentioned, the Mel scale is linear up to 1 kHz [9] and logarithmic at frequencies greater than 1 kHz. Discrete cosine transformation (DCT) is the computation of the output values of the filter bank. The DCT calculation attempts to arrange the coefficients depending on their significance and excludes unreliable ones. The dynamic coefficients are calculated by applying the first and second derivatives. The PLP technique estimates the parameters of an autoregressive filter and all-pole modeling for a better auditory spectrum. This technique respect the hearing limits and speech production of a human to reduce the number of direct waveform samples to a few that represent the perceived frequency concentrations and widths.

C. TRAINING PHASE

In this section, we described the idea of construction of DWN and DSWN.

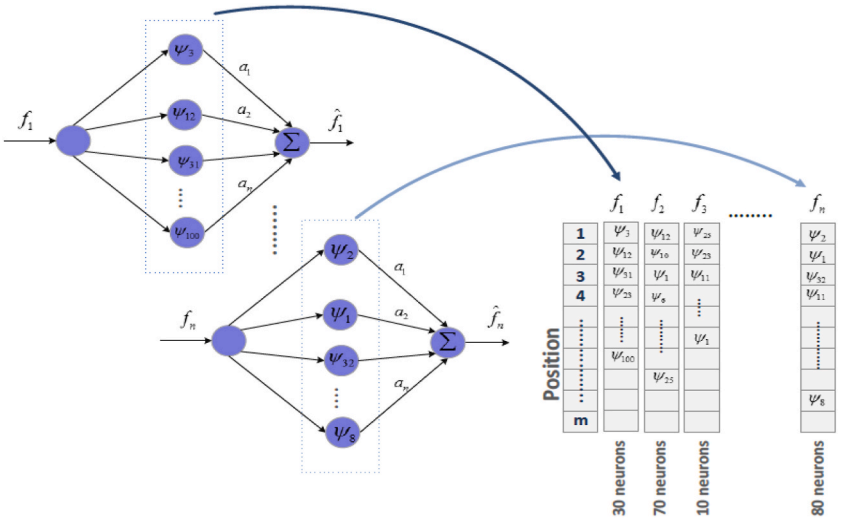

Fig. 1. GWN coefficient's calculation.

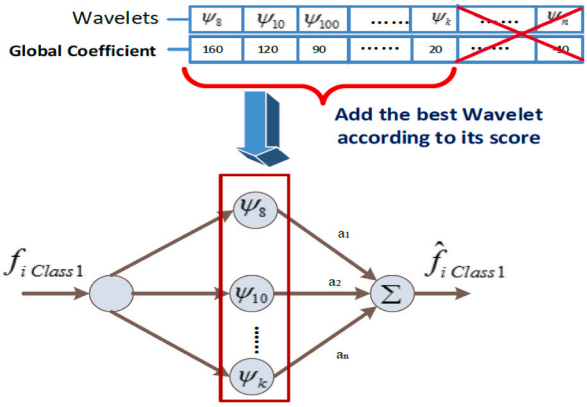

Fig. 2. GSWN construction.

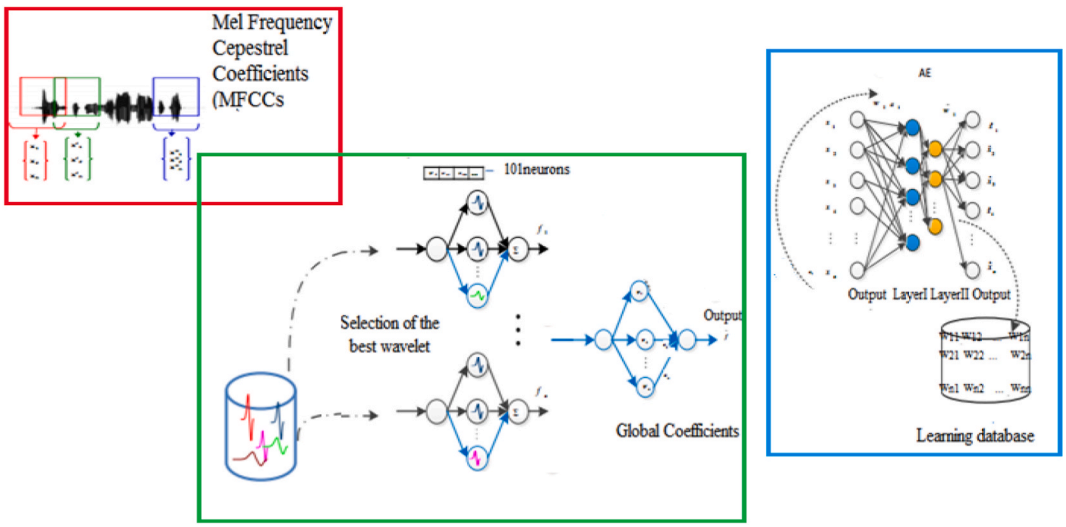

Fig. 3. Speech recognition system based on the DSWN [27].

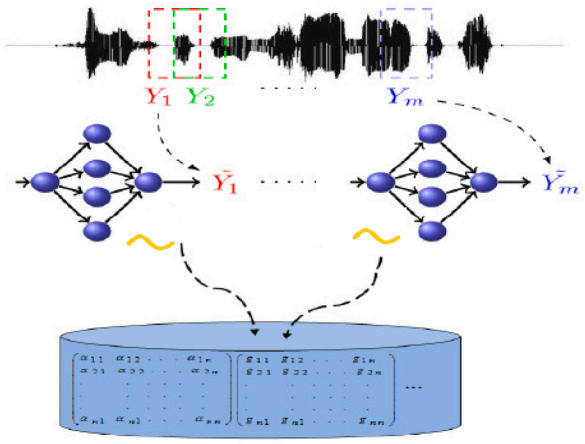

Fig. 4. Wavelet network modeling of an acoustic unit [27].

1) BEST CONTRIBUTION ALGORITHM

As first step, we generate a WN for each acoustic unit of classes of the speech training dataset. We used the best contribution algorithm [11,12] to build the WN. In the training phase, we construct a one-dimensional WN using the Fast Wavelet Transform (FWT) [13,16,18,20].

To evaluate our modeling approach, Fig. 4, of acoustic unit, we used the peak-signal-to-noise ratio (PSNR).

- Calculation of the wavelet scores

As a second step, we reckon the wavelet scores to create a global wavelet network (GWN). The GWN will model the signals of one class in the database (denoted as C1). We present the obtained WNs into two tables [19]. The first one is composed of the WNs of signals of C1, as illustrated in Table 2. The second one is composed of the WNs of the other signals of the dataset (denoted as C2).

Thereafter, we count the number of all the wavelets used in the C1 for each wavelet and for each position. The GWN will be built based on the best wavelets (based on the number of occurrences). The best wavelet is the one with the maximum score values in each position (Table 3).

As presented in Table 4, we create a new GWN of one class after identifying the wavelet scores.

As shown in Fig. 5, we used the best coefficient to identify the wavelets composed GWN. In this step, each class of the dataset is modeled by a GWN. The efficiency of each network is checked using PSNR.

- Building of Wavelet AE

After construction a GWN for each class of the dataset, we convert all GWN to a WAE. Fig. 6 illustrates the creation of an AE from a GWN [19].

Therefore, the decoded signal issued form the wavelet AE is calculated as follows:

$$
\begin{aligned}
&if\ i \neq j \begin{cases} < \psi_i, \widetilde{\psi}_j \geq = 0 \\ else \quad < \psi_i, \widetilde{\psi}_j \geq = 1 \end{cases} \\
&f = x_1, x_2, \ldots, x_n \\
&\widehat{f} = \widehat{x_1}, \widehat{x_2}, \ldots, \widehat{x_n} \\
&\psi_i = w_{1i}, w_{2i}, \ldots, w_{ni} \\
&\widehat{\psi_i} = \widehat{wx_{1i}}, \widehat{w_{2i}}, \ldots, \widehat{w_{ni}}
\end{aligned}
\tag{2}
$$

2) CREATION OF DWN

Fig. 7 present a two hidden layers DWN construction. This architecture is resulted from r a series of stacked autoencoders.

3) CREATION OF A DSWN

SC techniques are employed for feature extraction in recent years and there have been various algorithms which attempt to encode signals' sparse representations by using optimization techniques. SC is also integrated into deep neural networks that use an unsupervised learning algorithm for feature extraction by imposing a sparsity constraint on the hidden units.

After maintaining the configuration of the network, a sparsity component that animates the encoding vector to a sparse representation is added. In the literature, there are many ways to accomplish this. We select k-sparse AEs [22] for their simplicity. K-sparse AE obtains the *k* highest activations in the encoding *vector* and zeros out the rest. In our method, we obtain a sparse WAE from the GSWN, and the weight matrix is formed using wavelets from the GWN and its coefficient. Fig. 8 illustrates the creation phase of the sparse WAE.

Table 2
WAVELET COMPOSING THE NETWORKS OF C1.

Signal	Wavelet	Number of neurons
f_1	$\Psi_3 - \Psi_{15} - \Psi_{100} - \Psi_1 - \ldots$	123
f_2	$\Psi_1 - \Psi_{60} - \Psi_{15} - \Psi_{23} - \ldots$	30
.	.	.
.	.	.
.	.	.
f_n	$\Psi_{15} - \Psi_{100} - \Psi_2 - \Psi_4 - \ldots$	199

Table 3
Number of occurrences of each wavelet for C1 and C2

Class1 Wavelet	1	2	3	…	N
ψ_1	10	27	39	…	6
ψ_2	100	25	30	…	8
…	…	…	…	…	…
ψ_n	4	12	7	…	99

Table 4
Best coefficients.

ψ_1	ψ_3	ψ_{60}		ψ_n
100	72	50	…	99

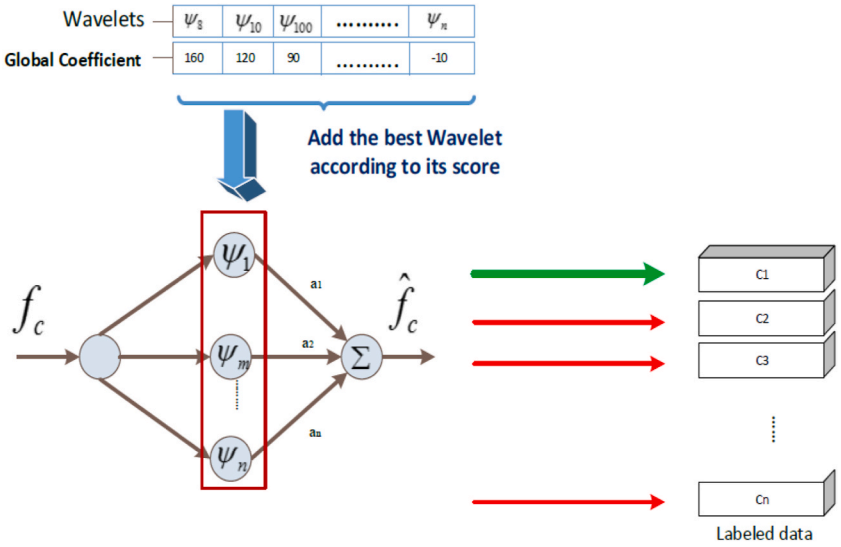

Fig. 5. Creation of GWN.

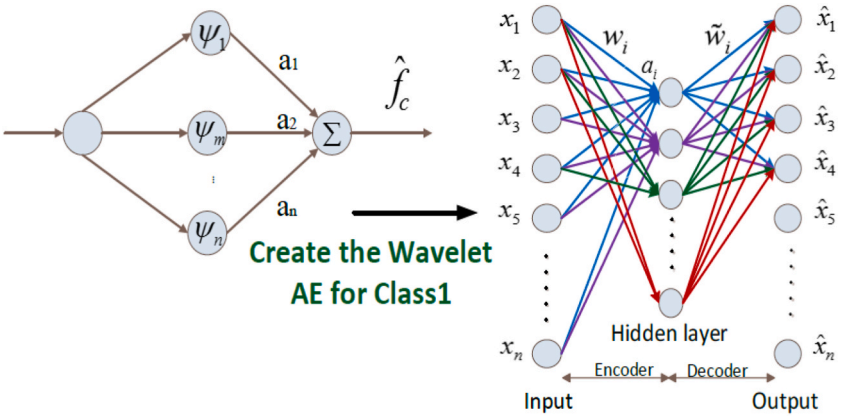

Fig. 6. A class modeling using a wavelet autoencoder.

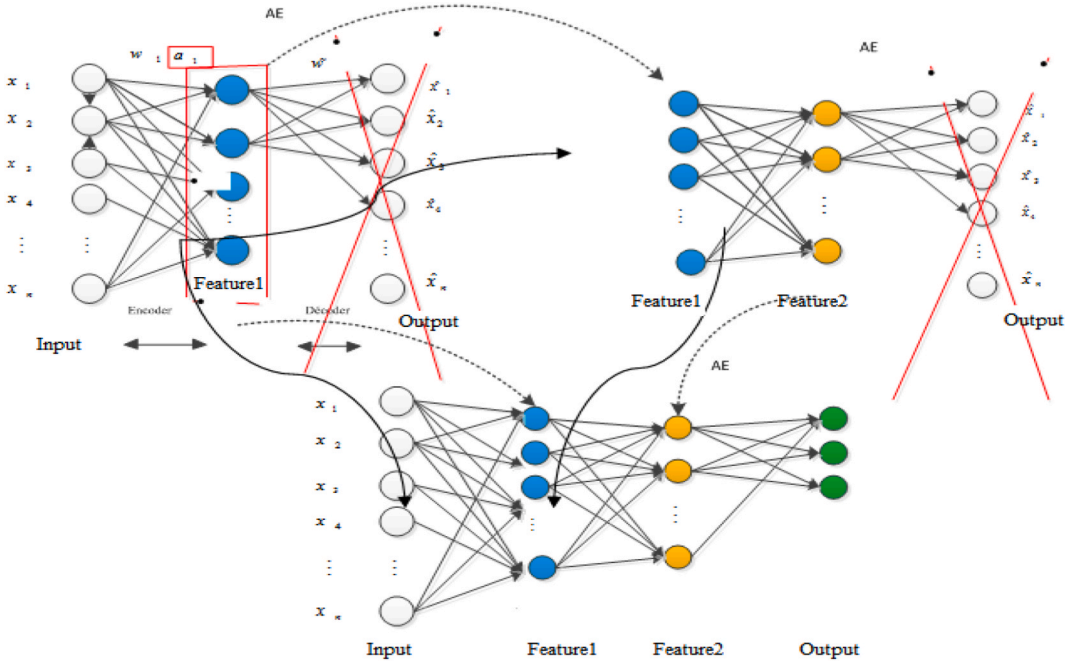

Fig. 7. A deep neural network with two hidden layers.

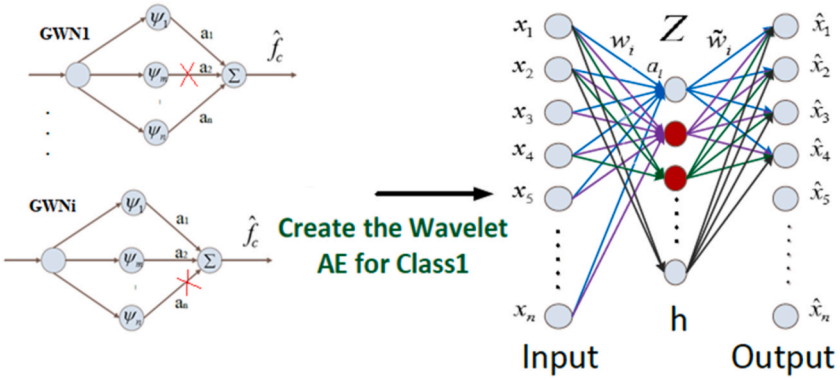

Fig. 8. Sparse WAE

The objective of using sparse coding is to reduce the number of connections between the nodes of the autoencoder without losing the quality of the modeling. The reduction of the connections will generate the reduction of the calculation time during the recognition phase (test phase) and the memory space necessary for the storage of the models of the acoustic units. These considerations will facilitate the implementation of this proposal on mobile architectures reduced in terms of memory space and computing time.

D. Recognition phase

The scenario of the recognition phase is similar to the learning phase. Fig. 9 illustrates the sequence of the recognition process.

The recognition process begins with the signal windowing phase. This phase is followed by feature extraction. These characteristics will then be projected onto the DSWNs created during the learning phase. To identify similar signals, the Euclidean distance is used to compare the input and output of each DSWN.

$$\boldsymbol{D} = \left\| \boldsymbol{a}_{i=1}^{n} \boldsymbol{V}_i \boldsymbol{Y}_i - \boldsymbol{a}_{j=1}^{n} \boldsymbol{V}_j \boldsymbol{Y}_j \right\| \tag{3}$$

Where $\boldsymbol{V}_i \boldsymbol{Y}_i$ and $\boldsymbol{V}_j \boldsymbol{Y}_j$ are two networks. The Euclidean distance between two networks using the same wavelet functions can be expressed as follows:

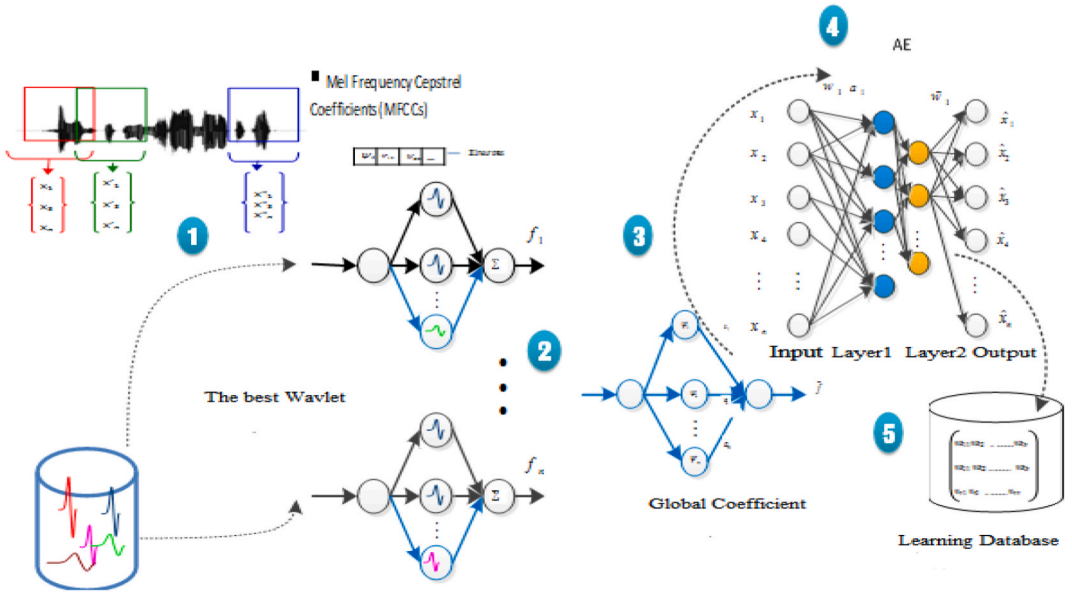

Fig. 9. Recognition phase [27].

$$\boldsymbol{D} = \left(\boldsymbol{D}^{t}\left(\boldsymbol{Y}_{ij}\right)\boldsymbol{D}^{1/2}\right. \tag{4}$$

Where $\boldsymbol{D} = \boldsymbol{d_1}\ldots\boldsymbol{d_n^t}$ and $\boldsymbol{Y}_{ij} = <\boldsymbol{Y}_i, \boldsymbol{Y}_j>$.

Therefore, to compare two networks with the same wavelet functions amounts comparing their weights matrices.

5. Experimental results

We have validated our approach on a corpus recorded from the work of Boudraa [21]. We manually segmented the corpus into Arabic words using Praat and chose 18 different words. In our approach, 2/3 of the data was used in the training phase and the rest (1/3) in the testing phase of the dataset. We have chosen the MFCC and PLP coefficients to represent the acoustic data.

PSNR was found to be effective in both classes. Nevertheless, we have rebuilt the global network from C1. C1 and C2 had almost the same wavelets. Thus, we could explain the signal quality performance in both classes. In addition, the overall rate of C1 obtained by PSNR was better than that of C2. Fig. 10 illustrates the approximation of C2 signals by DSWN of C1. We can conclude that the DSWN modeling C1 cannot model the signals of C2.

As shown in Tables 5 and 6, the autoencoder performs well between the reconstructed signal and the original signal. However, the quality of the signals gradually decreases in both classes due to the increase in error and degradation at their levels with increasing numbers of hidden layers. The following figures illustrate the recognition rate given by the DSWN compared to the WN and the intelligent approach. We evaluated our approach on Arabic words using MFCC and PLP coefficients.

To evaluate its performance, the proposed system was compared with other methods in the literature. As shown in Table 7 and Fig. 11, the proposed approach gave the best recognition rates compared to the intelligent approach and WN for MFCCs coefficient extracted from signal of Arabic words.

Table 8 and Fig. 12 show the results issued form the proposed approach, intelligent approach, and WN based on PLP coefficients of signal of Arabic words.

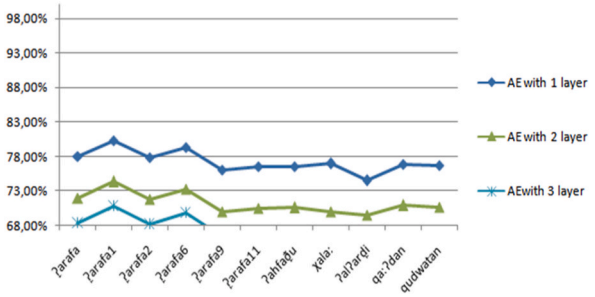

Fig. 10. PSNR values of C1 and C2 with AE.

Table 5
PSNR values of C1 of the AE that approximates C1.

Arabic word (speaker)	ʔˤ arafa (1)	ʔˤ arafa (2)	ʔˤ arafa (3)	ʔˤ arafa (4)	ʔˤ arafa (5)	ʔˤ arafa (6)
AE With one layer	0.8	0.8	0.78	0.8	0.77	0.77
AE With two layer	0.72	0.73	0.72	0.73	0.7	0.71
AE With three layer	0.68	0.7	0.69	0.7	0.66	0.7

Table 6
PSNR values of C2 of the AE that approximates C 1.

Arabic word	xala:	ʔalardˤi	qa: ʔdan	qudwatan	ʔahfa dˤu
AE With one layer	0.76	0.75	0.77	0.77	0.77
AE With two layer	0.7	0.69	0.71	0.71	0.7
AE With three layer	0.66	0.66	0.67	0.67	0.67

Table 7
Speech recognition rates using a DSWN with MFCC coefficients.

Arabic word	ʔˤ arafa	xala:	ʔalardˤi	qa: ʔdan	ʔˤ alkabfa	biθima:riha
WN	0.85	0.97	0.92	0.93	0.98	0.85
Intelligent approach	0.92	0.99	0.94	0.92	0.99	0.86
Proposed approach	0.93	0.98	0.95	0.99	0.99	0.87

Arabic word	minkuma:	sˤa:ʔiman	ʔalmusa:diruna:	ladˤaʔˤathu	zama:nuna:
WN	0.89	0.94	0.83	0.86	0.93
Intelligent approach	0.85	0.94	0.82	0.84	0.95
Proposed approach	0.88	0.95	0.88	0.86	0.96

Arabic word	ba:lana	biqawlin	ɣabatˤa	ka:la	kunta	mina	lahum
WN	0.87	0.89	0.95	0.83	0.98	0.91	0.89
Intelligent approach	0.88	0.9	0.96	0.87	0.98	0.98	0.95
Proposed approach	0.94	0.9	0.97	0.98	0.99	0.99	0.95

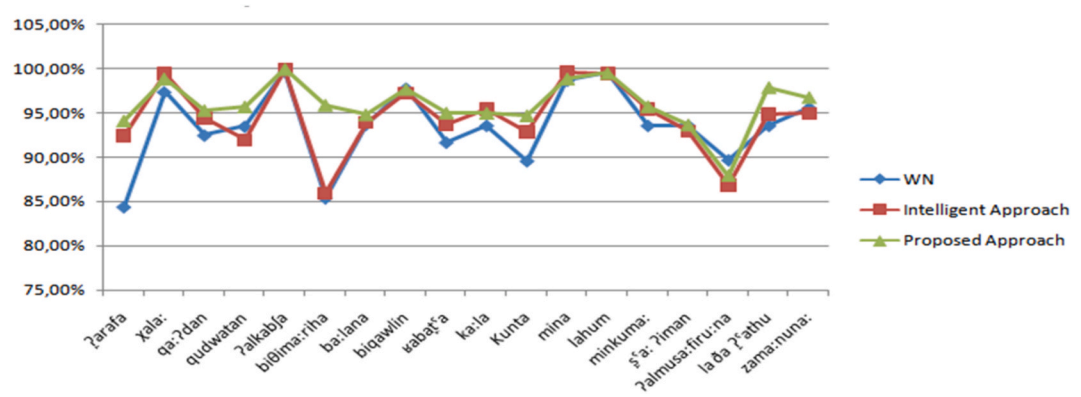

Fig. 11. Recognition system based on using MFCC coefficients.

Table 8
Speech recognition rates using a DSWN with PLP coefficients.

Arabic word	ʔahfadˤu	ʔalardˤi	ʔajna	saju? dˤihim	wa:lijan	jastamtiʔu
WN	0.97	0.98	0.89	0.97	0.96	0.98
Intelligent approach	0.98	0.99	0.94	0.95	0.97	0.97
Proposed approach	0.98	0.99	0.96	0.96	0.98	0.97
Arabic word	biθima:riha	ʔˤ arafa	xala:	qa: ʔdan	qudwatan	ʔˤ alkabfa
WN	0.96	0.9	0.96	0.93	0.98	0.99
Intelligent approach	0.95	0.94	0.98	0.95	0.98	0.99
Proposed approach	0.95	0.94	0.98	0.96	0.99	0.99

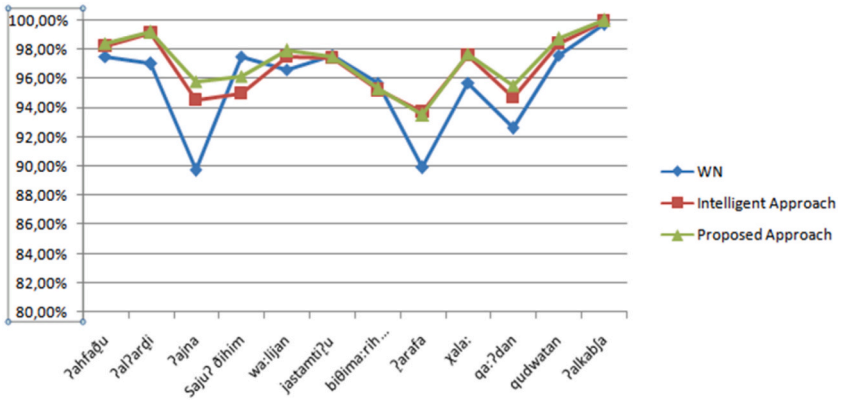

Fig. 12. Word recognition system based on PLP coefficients.

Based on all the figures, the proposed approach is better than the intelligent algorithm and the WN. Given the capacity of deep learning algorithms namely our proposal, the recognition rates can be improved by increasing the size of the training base or by improving the quality of the recordings.

6. Conclusion

This paper presents the different phases of a mobile speech recognition system. Each class of the dataset [21] is modeled by a DSWN. Each DSWN recognize all the signals of the belonging class. The DSWN consist of applying a hybridization of three theories in the recognition system: wavelet network, deep autoencoder, and sparse coding. WN and deep autoencoder is used to extract relevant features. SC is used to reduce the architecture of DWN by elimination of sparse connection. Among the DL algorithms used in the proposed approach, stacked AE and WN are the best contribution algorithms using the FWT. This novel architecture reveals that the DSWN shows enhanced performance with the PLP and MFCC. Despite promising results, we will analyze the potential limits of our methodology. One limitation identified concerns the use of a specific dataset of Arabic words. The generalizability of the sparse deep wavelet network (DSWN) to various linguistic contexts and acoustic environments remains an area requiring further research. We will also examine the impact of variations in speech characteristics, such as accents and dialects, on the performance of our model. These considerations aim to provide a more nuanced understanding of the applicability and robustness of our methodology in a broader spectrum of real-world scenarios.

CRediT authorship contribution statement

Sarah A. Alzakari: W. **Salima Hassairi:** Writing – review & editing, Writing – original draft. **Amel Ali Alhussan:** Writing – review & editing, Writing – original draft. **Ridha Ejbali:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology.

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.

Acknowledgment

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R716), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] R. Ejbali, M. Zaied, A dyadic multi-resolution deep convolutional neural wavelet network for image classification, Multimed. Tool. Appl. 77 (2018) 6149–6163, https://doi.org/10.1007/s11042-017-4523-2.
[2] R. Singh, R. Mehta, N. Rajpal, Efficient wavelet families for ECG classification using neural classifiers, Proc. Comput. Sci. 132 (2018) 11–21.
[3] A. ElAdel, R. Ejbali, M. Zaied, C. Ben Amar, Fast deep neural network based on intelligent dropout and layer skipping, in: 2017 International Joint Conference on Neural Networks, IJCNN, 2017, pp. 897–902, https://doi.org/10.1109/IJCNN.2017.7965947.
[4] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, S.S. Iyengar, A survey on deep learning: algorithms, techniques, and applications, 5, Article 92 (September 2019), ACM Comput. Surv. 51 (2018) 36, https://doi.org/10.1145/3234150. pages.
[5] Yuqi Si, Jingcheng Du, Li Zhao, Xiaoqian Jiang, Timothy Miller, Fei Wang, W. Jim Zheng, Kirk Roberts, Deep representation learning of patient data from Electronic Health Records (EHR): a systematic review, J. Biomed. Inf. 115 (2021) 103671, https://doi.org/10.1016/j.jbi.2020.103671. ISSN 1532-0464.

[6] D. Palaz, M. Magimai-Doss, R. Collobert, End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition, Speech Commun. 108 (2019) 15–32.

[7] R. Ben Ali, R. Ejbali, M. Zaied, Classification of medical images based on deep stacked patched auto-encoders, Multimed. Tool. Appl. 79 (2020) 25237–25257, https://doi.org/10.1007/s11042-020-09056-5.

[8] Vikas Chauhan, Aruna Tiwari, Randomized neural networks for multilabel classification, Appl. Soft Comput. 115 (2022) 108184, https://doi.org/10.1016/j.asoc.2021.108184. ISSN 1568-4946.

[9] L. Ashok Kumar, D. Karthika Renuka, M.C. Shunmuga Priya, Analysis of audio visual feature extraction techniques for AVSR system, in: ICCAP 2021, 2021, pp. 7–8. December 2021, Chennai, India.

[10] A. Kumar, S. Verma, H. Mangla, A survey of deep learning techniques in speech recognition, in: 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN), 2018, pp. 179–185, https://doi.org/10.1109/ICACCCN.2018.8748399.

[11] I. Teyeb, A. Snoun, O. Jemai, M. Zaied, Fuzzy logic decision support system for hypovigilance detection based on CNN feature extractor and WN classifier, J. Comput. Sci. 14 (11) (2018) 1546–1564, https://doi.org/10.3844/jcssp.2018.1546.1564.

[12] Siwar Yahia, Salwa Said, Mourad Zaied, Wavelet extreme learning machine and deep learning for data classification, Neurocomputing 470 (2022) 280–289, https://doi.org/10.1016/j.neucom.2020.04.158. ISSN 0925-2312.

[13] M. Sakkari, M. Hamdi, H. Elmannai, et al., Feature extraction-based deep self-organizing map, Circ. Syst. Signal Process. 41 (2022) 2802–2824, https://doi.org/10.1007/s00034-021-01914-3.

[14] Y. LeCun, Learning Invariant Feature Hierarchies, Computer Vision-ECCV, 2012.

[15] P. Li, P. Hua, D. Gui, et al., A comparative analysis of artificial neural networks and wavelet hybrid approaches to long-term toxic heavy metal prediction, Sci. Rep. 10 (2020) 13439, https://doi.org/10.1038/s41598-020-70438-8.

[16] R. Ejbali, O. Jemai, M. Zaied, C. Ben Amar, A speech recognition system using fast learning algorithm and beta wavelet network, in: 2015 15th International Conference on Intelligent Systems Design and Applications, ISDA, 2015, pp. 14–18, https://doi.org/10.1109/ISDA.2015.7489241.

[17] Marco Gori, in: Chapter 5 - Deep Architectures, Marco Gori, Machine Learning, Morgan Kaufmann, 2018, pp. 236–338, https://doi.org/10.1016/B978-0-08-100659-7.00005-1. ISBN 9780081006597.

[18] Olfa Jemai, Ridha Ejbali, Mourad Zaied, Chokri Ben Amar, A speech recognition system based on hybrid wavelet network including a fuzzy decision support system, ICMV (2014) 944503.

[19] S. Hassairi, R. Ejbali, M. Zaied, Supervised image classification using deep convolutional wavelets network, in: IEEE 27th International Conference on Tools with Artifcial Intelligence, 2015.

[20] A. ElAdel, R. Ejbali, M. Zaied, C. Ben Amar, Dyadic MultiResolution Analysis-Based Deep Learning for Arabic Handwritten Character Classification" 13th International Conference on Document Analysis and Recognition, ICDAR, 2015.

[21] M. Boudraa, B. Boudraa, Twenty list of ten Arabic sentences for assessment, ACUSTICA acta acoustica 86 (43) (1998), 71, pp. 870–882L. Kozachenko and N. Leonenko. On statistical estimation of entropy of random vector. Problems Infor. Transmiss., 23(2):95–101, 1987.

[22] Alireza Makhzani, J. Frey Brendan, "k-Sparse Autoencoders." CoRR Abs/1312, 2014, p. 5663 (n. pag).

[23] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.

[24] Peter FÖLDIÁK et Malcom P YOUNG, Sparse coding in the primate cortex, The handbook of brain theory and neural networks 1 (1995) 1064–1068.

[25] B. Li, M. Salucci, P. Rocca, W. Ke, W. Tang, The sparsity and incoherence in compressive sensing as applied to field reconstruction, in: 2020 14th European Conference on Antennas and Propagation, EuCAP, 2020, pp. 1–3, https://doi.org/10.23919/EuCAP48036.2020.9135834.

[26] N.G. Andrew, Sparse autoencoder, CS294A Lecture notes 72 (2011) 1–19.

[27] A. Bouallégue, S. Hassairi, R. Ejbali, M. Zaied, Learning deep wavelet networks for recognition system of Arabic words, in: M. Graña, J. López-Guede, O. Etxaniz, Á. Herrero, H. Quintián, E. Corchado (Eds.), International Joint Conference SOCO'16-CISIS'16-ICEUTE'16. SOCO 2016, CISIS 2016, ICEUTE 2016, Advances in Intelligent Systems and Computing, vol. 527, Springer, Cham, 2017, https://doi.org/10.1007/978-3-319-47364-2_48.

[28] T. Zhao, et al., A survey of deep learning on mobile devices: applications, optimizations, challenges, and research opportunities, Proc. IEEE 110 (3) (March 2022) 334–354, https://doi.org/10.1109/JPROC.2022.3153408.

[29] S. Dua, S.S. Kumar, Y. Albagory, R. Ramalingam, A. Dumka, R. Singh, M. Rashid, A. Gehlot, S.S. Alshamrani, A.S. AlGhamdi, Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network, Appl. Sci. 12 (2022) 6223, https://doi.org/10.3390/app12126223.

[30] P.S. Hossain, A. Chakrabarty, K. Kim, M.J. Piran, Multi-label extreme learning machine (MLELMs) for bangla regional speech recognition, Appl. Sci. 12 (2022) 5463, https://doi.org/10.3390/app12115463.

[31] Galym Kapyshev, Marat Nurtas, Aizhan Altaibek, Speech recognition for Kazakh language: a research paper, Proc. Comput. Sci. 231 (2024) 369–372, https://doi.org/10.1016/j.procs.2023.12.219. ISSN 1877-0509.

[32] Shaohua Jiang, Zheng Chen, Application of dynamic time warping optimization algorithm in speech recognition of machine translation, Heliyon 9 (11) (2023) e21625, https://doi.org/10.1016/j.heliyon.2023.e21625. ISSN 2405-8440.