Indonesian Journal of Electrical Engineering and Computer Science
Vol. 35, No. 2, August 2024, pp. 1013~1021
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v35.i2.pp1013-1021

AMSVT: audio Mel-spectrogram vision transformer for spoken Arabic digit recognition

Omayma Mahmoudi, Naoufal El Allali, Mouncef Filali Bouami
Laboratory of Applied Mathematics and Information Systems (MASI), Multidisciplinary Faculty of Nador,
Mohammed Premier University, Oujda, Morocco

Article Info

Article history:

Received Feb 16, 2024
Revised Mar 21, 2024
Accepted Apr 28, 2024

Keywords:

AMSVT
Mel-spectrogram
RNNs
Speech digit recognition
Vision transformer

ABSTRACT

This work presents a novel model to recognize spoken digits in the Arabic language. Due to the transformer-based models' tremendous success in natural language processing (NLP), several attempts have been made to extend transformer-based designs to other domains, such as vision and audio. However, our approach consists of extracting and inputting Mel-spectrogram features into our model of the proposed audio Mel-spectrogram vision transformer (AMSVT) for training. The signal processing community has been interested in these models due to the successful use of vision transformers (ViT) in several computer vision applications. This is because signals are frequently recorded as spectrograms (using the Mel-spectrogram, for example), which may be given directly as input to vision transformers. Our model outperformed a group of models in terms of accuracy and time, such as convolutional neural network (CNN)-based and recurrent neural network (RNN)-based.

Corresponding Author:

Omayma Mahmoudi
Laboratory of Applied Mathematics and Information Systems (MASI)
Multidisciplinary Faculty of Nador, Mohammed Premier University
Oujda, Morocco
Email: mahmoudi.omayma@ump.ac.ma

1. INTRODUCTION

Vision transformers (ViT) [1]-[4] has quickly emerged as the most popular area of research in computer vision to date, demonstrating promising results for a variety of applications, including recognizing objects [5], [6], detection [7], [8], and generation [9], [10] in addition to segmentation and production of medical images. These models have also caught the interest of the signal-processing community as a result of the undisputed performance of ViT in handling a variety of computer vision applications [11], [12]. For instance, Gong *et al.* [12] used ViT to process signals by simply feeding the model with spectrograms, which are essentially visual representations of signals. Transformers use mechanisms for self-attention (SA) [13] to efficiently collect long-range relationships and extract contextual data from the input sequence. As a result, ViT is better suited for jobs that call for the modeling of intricate dependencies and interactions between the input characteristics.

In this investigation, we delve into the converse task, specifically scrutinizing the suitability of techniques originally formulated for 2D image analysis to a 1D signal domain. Our focus is on evaluating the suitability of transformer-based approaches designed for image classification to tackle the challenges of audio classification. Despite audio being inherently a 1D signal, working with its spectral representation partially bridges the gap between 1D and 2D domains. Nevertheless, the frequency/time characteristics of a spectrogram differ significantly from those of a conventional image, where pixel relationships and their physical significance maintain consistency in any arbitrary direction. This discrepancy is notable in a spectrogram, where the axes of the image signify different physical occurrences.

Additionally, within a spectrogram, classes intersect, while in a scene image with multiple objects, the classes are adjacent. These distinctions prompt concerns regarding the direct transferability of image processing techniques for the audio classification challenge. While uncertainty has been resolved regarding convolutional neural networks (CNNs), the suitability of attention-based models remains a matter that remains unanswered. Several studies [14]-[16] substantiate the potential applicability of our model, audio Mel-spectrogram vision transformer (AMSVT), to audio classification.

We chose a database consisting of audio recordings, which are numbers spoken in Arabic from 0 to 9, to test the efficiency of our proposed model. This choice is justified by simplicity and clarity reasons: spoken digits are finite and well-defined, making them easier to study and analyze. Practical applications: Many real-world applications, such as automated telephone systems or voice-controlled devices, often require the recognition of spoken digits. Focusing on this area can directly contribute to improving these technologies.

Transformer used in audio processing is the primary topic of this article. By presenting a novel straightforward yet efficient technique for training transformers using spectrograms, where we address current audio transformers' computational complexity and memory needs. To summarize, the following are the primary contributions of our work:

- We suggest AMSVT, which considerably lowers the memory and computing requirements for training transformers in the audio domain.
- We decompose the positional encoding of the transformer [17] into time and frequency positional encoding, enabling simple inference on audio clips of varying duration without the requirement for fine-tuning or interpolating positional encodings.
- An evaluation of this model across several tasks using a spoking Arabic digit dataset, studying various techniques for simplifying training and showing how they impact performance on proposed audio SET.
- The model AMSVT beat RNN-based regarding training performance, memory needs, and generalization. According to our results presented at the end of this work.

Atito *et al.* [18], present ASiT, a groundbreaking self-supervised transformer designed for creating versatile audio representations. ASiT incorporates self-distillation and group-masked model learning to adeptly capture both local and global contextual information. The researchers assessed their pre-trained models across a range of audio and speech classification tasks, encompassing audio event classification, keyword detection, and speaker identification. Additionally, they conducted thorough ablation investigations, including analyses of diverse pretraining methods. The proposed ASiT framework outperforms existing techniques, even those leveraging supplementary datasets for pretraining, setting a new standard for performance in five audio and speech classification tasks. Remarkably, it substantially improves performance across all evaluated tasks. The code and pre-trained weights will be shared openly with the scientific community.

Gong *et al.* [19], introduce the audio spectrogram transformer (AST), a novel audio categorization model that relies solely on attention mechanisms, eliminating the need for convolutions. The performance of AST was evaluated on various audio classification benchmarks, achieving remarkable state-of-the-art scores, including a 0.485 mAP on AudioSet, 95.6% accuracy on ESC-50, and 98.1% accuracy on speech commands V2. Ristea *et al.* [20] propose the separable transformer (SepTr), a configuration that employs two transformer blocks consecutively. The initial block attends to tokens within the same time interval, while the second focuses on tokens within the same frequency bin. Through an examination of three benchmark datasets, they illustrated that the separable architecture surpasses both traditional ViT and state-of-the-art techniques. SepTr exhibits a reduced memory footprint compared to conventional transformers by scaling the number of trainable parameters linearly with the input size.

In this study, part 2 will focus on our distinctive contribution, namely the design of the suggested audio Mel-spectrogram vision transformer. Section 3 will next thoroughly assess the dataset chosen and the performance improvements resulting from the AMSVT proposed. This will be followed by a conclusion that summarizes the key results and lessons from the study.

2. THE PROPOSED AUDIO MEL-SPECTROGRAM VISION TRANSFORMER (AMSVT)

The suggested AMSVT framework primarily involves two steps, depicted in Figure 1. Initially, we input the mel-spectrogram extracted from the audio and then pass it through the pre-trained AMSVT to extract frame-level spatial features. These spatial features are concatenated to form a feature vector derived from 30 consecutive frames.

2.1. Features extraction using ViT

In contrast to models based on RNN approaches for image classification tasks, the AMSVT architecture, built entirely upon the conventional transformer design [21], achieved exceptionally high accuracy. This architecture effectively captures long-range dependencies within input sequences through a self-attention mechanism. Notably, ViT represents an endeavor to classify images using the transformer model. It dissects the input image into linearly projected patches, employs trainable positional embeddings to determine the patch sequence, and utilizes a transformer encoder and a multilayer perceptron to achieve the final classification.

Since a normal transformer only accepts a 1D sequence of tokens as input, the input picture of the spectrogram in the first section is split into nonoverlapping patches. A sequence of flattened 2D patches called $x_p \in \mathbb{R}^{N\times(P^2.C)}$ is created from an image called $x \in \mathbb{R}^{H\times W\times C}$ in order to handle 2D images, which are typically in 2D format. Here, (H, W, C) stands for the picture's height, width, and the resolution of each image patch is denoted by (P, P), with the total number of patches represented as $(N = HW/P^2)$. The patch size P is often set to 16×16 or 32×32, with the smaller P size being able to record longer sequences and the larger P size being able to capture shorter sequences. The $16 \times 16\ P$ has been employed in our situation to extract features.

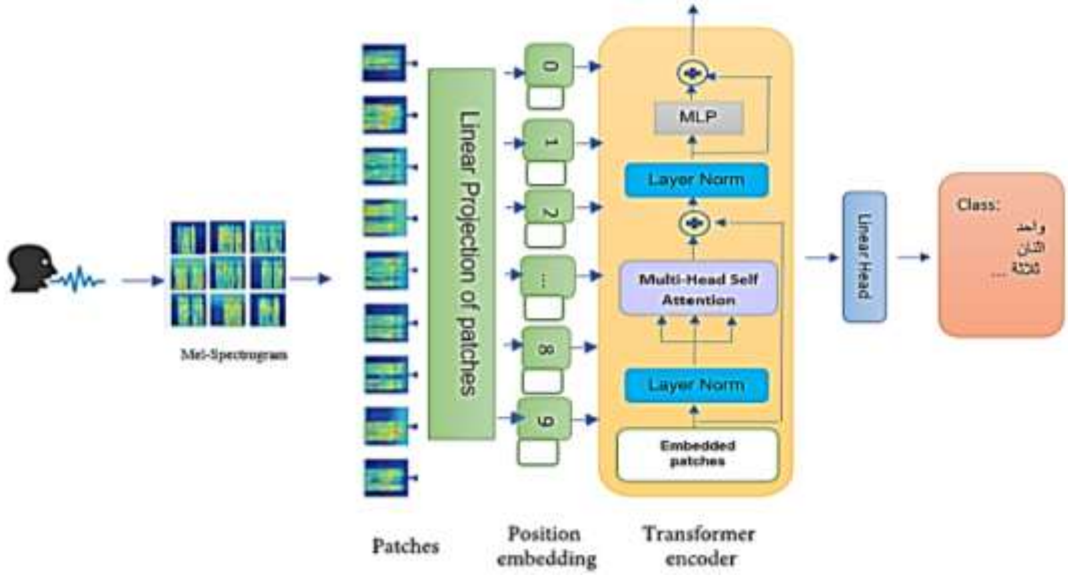


Figure 1. The proposed framework for arabic digit recognition using AMSVT

2.2. Linear embedding layer

The linear projection of sequence patches into a vector of dimension d is accomplished through a learned embedding matrix E. Subsequently, these embedded representations are combined with a trainable classification token v_{class} . Since the embedded patches lack a specific order, positional information E_{pos} is employed to rearrange the spatial information to match that of the original image. The mathematical representation of the embedded patches with token Z_0 is presented in (1).

2.3. Mel-spectrogram ViT encoder

The transformer encoder layer comprises L identical layers, as depicted in Figure 2. receives the embedded Z_0 patches as a consequence of (1). Each layer also consists of two parts, such as a multilayer perceptron (MLP) and a multihead self-attention (MSA) block. Two thick layers make up the last block of MLP. The MSA and MLP are respectively represented mathematically in (2) and (3).

$$Z_0 = [v_{\text{class}}\,; x_1E; x_2E; .. x_nE] + E_{pos}, E \in \mathbb{R}^{(P^2.C)\times d}, E \in \mathbb{R}^{(n+1)\times d} \tag{1}$$

$$z_l' = \text{MSA}\big(LN(z_{l-1})\big) + z_{l-1}, l = 1 \dots L \tag{2}$$

$$z_l = \text{MLP}\big(LN(z_l')\big) + z_l', l = 1 \dots L \tag{3}$$

The initial element in the sequence, z_L^0, is supplied to the external head classifier in the encoder's final layer so it can forecast the class label.

$$y = LN(z_L^0). \tag{4}$$

Figure 2(a) present the architecture general of our proposed model, and Figure 2(b) determines the composition of the transformer encoder layer in detail. The pivotal component of the transformer model is the MSA, responsible for identifying the most and least significant patches and excluding the latter from the input sequence. As illustrated in Figure 2(c), the MSA is segmented into four layers, including linear, self-attention, and concatenation layers, to consolidate the output from various heads. Essentially, the attention process can be visualized using attention weights, calculated by summing all values in a sequence z. By multiplying components (Q, K) with the three learning matrices $\mathrm{U_{QKV}}$, three values Q (query), K (key), and V (value) are obtained from the input sequence. A single self-attention (SA) is visually depicted in Figure 2(d), and its mathematical formulation is provided in (5).

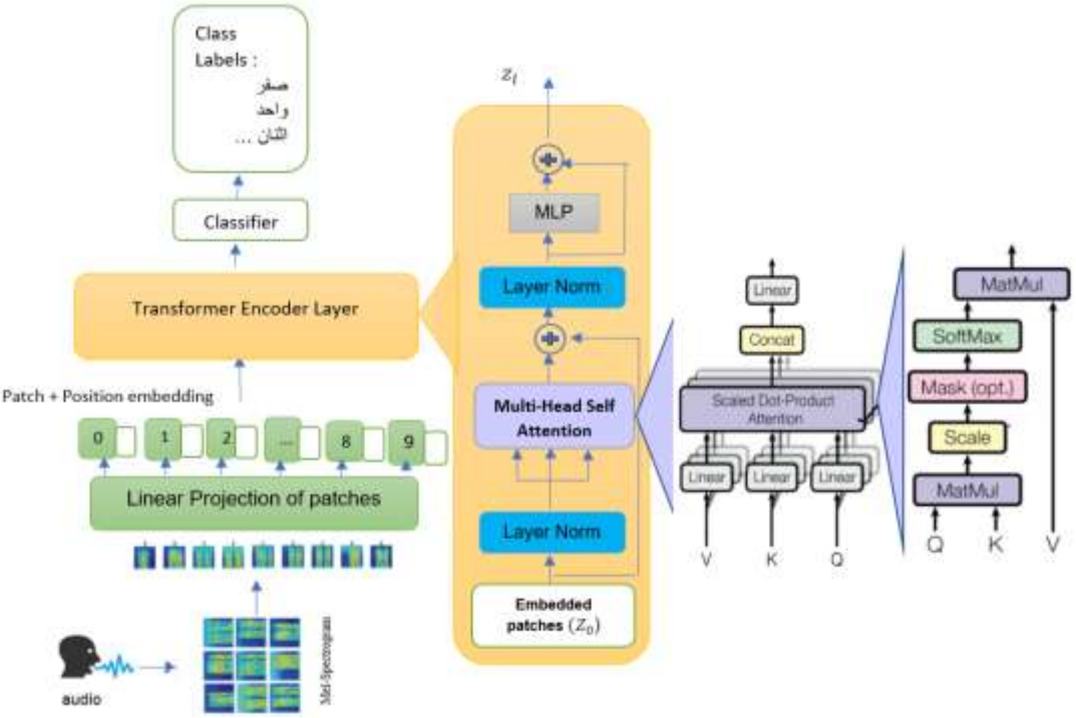
$$[Q, K, V] = zU_{QKV}, U_{\mathrm{QKV}} \in \mathbb{R}^{d\times 3D_K}. \tag{5}$$

The Q vector's value undergoes multiplication by the dot product with the K vectors in a specific input sequence, determining the relative significance of each element about others. The resulting output is then scaled and passed through the SoftMax activation function to ascertain the significance of the patch with the highest attention score, as formally expressed in (6).

$$A = \mathrm{SoftMax}\left(\frac{QK^T}{\sqrt{D_K}}\right), A \in \mathbb{R}^{n\times n}. \tag{6}$$

Instead of using only one value for Q, K, and V, the MSA combines the numerous attention heads h. As shown in (7), the outputs from each SA are combined to choose robust and optimum features, which are then projected to the required dimensions using a feedforward layer with learnable weights W.

$$MSA(z) = \mathrm{Concat}\,\big(SA_1(z); SA_2(z); \dots SA_h(z)\big)W, W \in \mathbb{R}^{h\cdot D_K\times D} \tag{7}$$



(a) (b) (c) (d)

Figure 2. The AMSVT architecture: (a) the model's primary architecture, (b) the encoder layer of the transformer model, (c) the MSA head, and (d) the SA head

3. METHOD

Performance metrics. In each of our trials, we use classification accuracy as a metric for assessment. This is done by calculating the accuracy and loss for each model of RNNs-based like long short term memory network (LSTM), Bidirectional-LSTM (BiLSTM), RNN, and gated recurrent unit (GRU) and compared with our model AMSVT data preprocessing. For our dataset, we apply the Mel-spectrogram with Nx=1,024, R=32, and a window size of 200. The magnitude's square root is then calculated for each Mel-spectrogram, and the results are then mapped to 128 Mel bins. A single-channel output matrix is created by converting the result to a logarithmic scale (decibels) and normalizing it to the range [0, 1].

Training details and hyperparameter tuning, we discovered common ideal hyperparameters when tuning the AMSVT and RNN-based hyperparameters utilizing the validation sets. As a result, Adam optimizes each model while employing the cross-entropy loss function. After every 10 epochs, we apply a decay factor of 0.5 and begin with an initial learning rate of 10-4. On batches of 64, we train each model for 100 iterations. For AMSVT, we configured the number of blocks to L=3 and the token size to d=2,048, AMSVT was created with 8 attention heads. For the RNN-based architecture, the accuracy and training duration of the suggested model might be impacted by the learning rate and training iterations. To attain optimal performance, both parameters were changed to different values. Given that the learning rate ought to be regarded as the most important hyperparameter, it can be critical to comprehend how to alter it correctly to get the best results. The network weight changes are governed by the learning rate. The model initially learned at a rate of 10−3.

The training iteration follows, with a starting value of=1,000 iters. To create the training steps, the training iterations were multiplied by the epoch size. Between 1,000 and 2,000 training steps were used, with 10 epochs of batch size 64/64. The training steps were augmented, and this resulted in high accuracy.

The networks were trained using the softmax activation function and the cross-entropy loss. With a 10−2 initial learning rate, the model learned swiftly but eventually began to overfit. When the model was overfitted, it was seen that the accuracy decreased. The model was trained slowly and the network accuracy increased by changing the learning rate to 10−3.

Instead of using a multi-core central processing unit (CPU), the proposed architecture was implemented on a single machine's graphics processing unit (GPU). The decision to utilize a GPU was made due to the GPU's ease of implementation and speedy debugging. Performance evaluation: Python (version 3.6) and the Jupyter integrated development environment have been used to accomplish the suggested solution. Where Table 1 summarizes the environment used in this study.

Table 1. Deep-learning system environment

	DL Toolkit	PyTorch 12
Edge Computing	Language	Python 3.6
	OS	Windows
	RAM	32 GB
	GPU	One NVIDIA GTX 3080, 11GB
	CPU	AMD Ryzen 9 5000 Quad-Core Processor

4. RESULTS AND DISCUSSION

4.1. Data sets

The Arabic SC Dataset (v1.0) [22] is comprised of spoken words and is designed for training and evaluating keyword-spotting systems. The dataset includes 12,000 1-second recordings featuring 40 common speech commands. Each audio file is one second in duration, sampled at 16 kHz, with contributions from 30 participants, each recording 10 utterances for each keyword. Consequently, there are 300 audio files for each keyword.

For our study, we specifically extracted, for the reasons mentioned earlier, recordings that articulate the digits from 0 to 9. Resulting in the utilization of 10 audio files that encompass digits in total (10×10×40=4,000), as shown in Table 2. The division of these audio samples involves allocating 70% for training, 15% for validation, and 15% for testing.

Figure 3 present the extraction of features from the audio, which depicts the primary steps in the Mel-spectrogram extraction of features method. We used a window of 25 ms length with 10 ms length in the framing step. Then, we extracted 13 Mel-frequency cepstral coefficients (MFCC) features using the discrete cosine transform (DCT) on these values. The first feature was then deleted because it had no useful information, leaving only the next 12 features.

Table 2. The chosen keywords with their translations

Translation	Keyword
Zero	صفر
One	واحد
Two	اثنان
Three	ثلاثة
Four	اربعة
Five	خمسة
Six	ستة
Seven	سبعة
Eight	ثمانية
Nine	تسعة

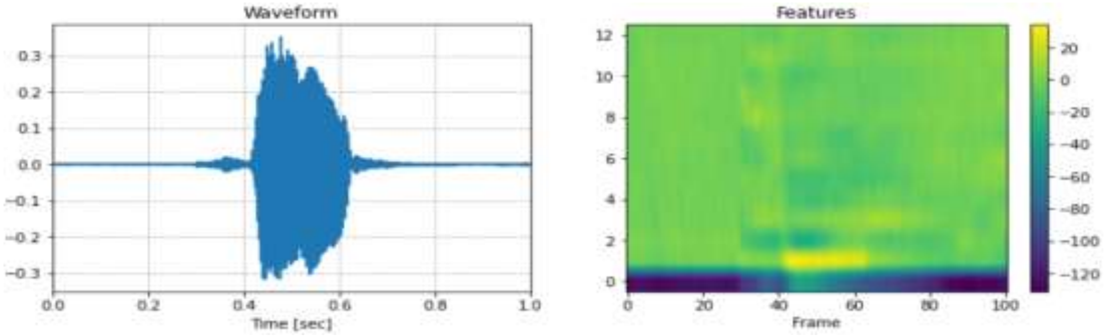


Figure 3. Presentation of waveforms and features using a (zero) instance from our dataset

4.2. Results

The tests we carried out led to the accuracies results that we obtained for the data set related to numbers spoken in Arabic [23]. Where we calculated the accuracy and training time for each of the following model's LSTM [24], GRU [25], RNN [26], BiLSTM [27], and compared them to our proposed model AMSVT as shown in Table 3. As our model was ahead of the previous models by a difference of 12%. As far as we know, no research has been conducted on Arabic-spoken digits using such methods.

Table 3. Speech digit recognition accuracy result

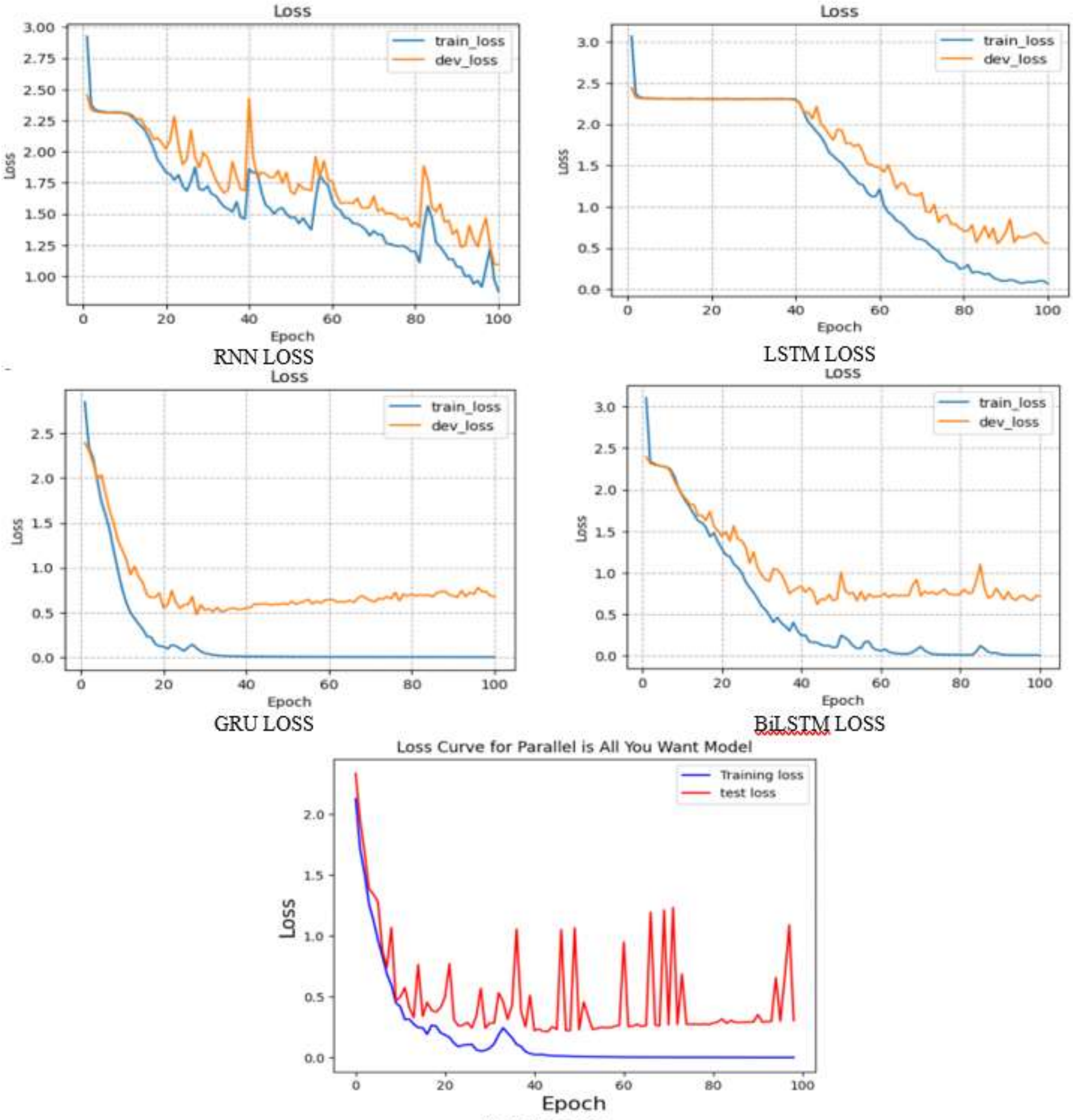
Method	Accuracy	Taken time
RNN	0.57	864 s
LSTM	0.82	1,063 s
GRU	0.84	1,135 s
BiLSTM	0.77	1,064 s
AMSVT	0.95	250 s

The study's key findings demonstrate that the AMSVT model surpasses GRU and other RNN-based models in terms of accuracy and loss. With an accuracy of 95%, AMSVT outperformed all other models, including GRU, by a significant margin. This highlights the effectiveness of the AMSVT architecture for the given dataset. A key piece of supporting evidence is the comparison with the GRU model, which achieved an accuracy of only 84% on the same dataset, emphasizing the substantial improvement offered by AMSVT.

This section elaborates on the experiment results conducted on various models. Figure 4 shows the training loss of the proposed AMSVT, LSTM, RNN, BiLSTM, and GRU models. The GRU and BiLSTM models had a smooth curve for loss whereas for the LSTM model, there was a sudden drop in loss, and it remained almost constant for the rest of the epochs. The loss for the AMSVT and RNN models seemed to vary over each epoch, where multiple vibrations appear, especially at the level of test loss.

In comparison with previous studies, the current research sets itself apart by introducing the AMSVT model and showcasing its superior performance. Unlike previous studies that primarily relied on RNN-based models like GRU, this study explores a transformer-based approach, which has shown to be more effective. The strengths of the study include its comprehensive evaluation of different deep learning models and the thorough analysis of hyperparameter tuning effects. However, limitations may include the lack of exploration of other transformer architectures and the potential bias in the dataset. Unexpectedly, AMSVT not only outperformed GRU but also surpassed all other deep learning models by significant margins, indicating its robustness and efficacy.

The study aims to assess the performance of the AMSVT model in comparison to traditional RNN-based models for a specific dataset. Its importance lies in demonstrating the superiority of transformer-based architectures like AMSVT for certain tasks, shedding light on potential advancements in deep learning for similar applications. However, unanswered questions remain regarding the generalizability of AMSVT to other datasets and its performance under different conditions. Future research could explore the applicability of AMSVT in diverse domains and investigate ways to enhance its performance further.



RNN LOSS

LSTM LOSS

GRU LOSS

BiLSTM LOSS

AMSVT LOSS

Figure 4. Training and testing history plots of loss of the different models

4.3. Comparison study

Our research has been contrasted with other studies that looked into the Arabic language's digit speech recognition. Table 4 displays a variety of findings from earlier research. MFCC+AMSVT achieves the highest accuracy of 95% in 2024. This suggests that this approach is currently the most effective for speech recognition. The accuracy of most approaches has improved over time. For example, MFCC+CNN improved from 82% in 2021 to 86% in 2020. This shows that research in speech recognition is making progress.

Some approaches, such as MFCC+DTW and DHMM, have seen a decrease in accuracy over time. This could be due to several factors, such as the limitations of these approaches or the increasing difficulty of the datasets. It is not clear which approach is best for all tasks and datasets. The best approach may vary depending on the specific requirements of the task.

Table 4. Results of different approaches

Reference	Year	Approach	Accuracy
Our work	2024	MFCC+AMSVT	95%
Asroni *et al.* [28]	2021	MFCC+CNN	82%
Zada and Ullah [29]	2020	MFCC+HMM and CNN	84%
Alasadi *et al.* [30]	2020	MFCC+SVM	86%
Wazir and Chuah [31]	2019	MFCC+RNN	69%
Hachkar *et al.* [32]	2011	MFCC+DTW and DHMM	92%

5. CONCLUSION

By utilizing our suggested AMSVT, this study has effectively built a voice recognition solution for Arabic digits. Mel Spectrograms have been employed for feature extraction from audio recordings, and RNN-Based has been compared performance with our proposal model. We found the length of time needed to train our model was ideal, whereas the training procedure took far longer for RNN-based models. During the training phase, the created model had little loss and attained an accuracy of 95%. The model's accuracy will be examined in the work's final phase using a different test dataset. The testing step uses 60 samples per digit for a total of 600 data points for all the digits. Where the results were encouraging. 96% accuracy is achieved in the recognition of the majority of the digits.

REFERENCES

[1] N. M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M. H. Yang, “Intriguing properties of vision transformers,” *in Advances in Neural Information Processing Systems*, vol. 34, pp. 23296-23308, 2021, doi:10.48550/arXiv.2105.10497.

[2] X. Yue, S. Sun, Z. Kuang, M. Wei, P. H. Torr, W. Zhang, and D. Lin, “Vision transformer with progressive sampling,” *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 387-396, 2021, doi: 10.48550/arXiv.2108.01684.

[3] S. Paul and P. Y. Chen, “Vision transformers are robust learners,” *in Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 2071-2081, June 2022, doi: 10.48550/arXiv.2105.07581.

[4] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, “Semantic segmentation using vision transformers: A survey,” *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106669, 2023, doi: 10.48550/arXiv.2305.03273.

[5] M. Lin, M. Chen, Y. Zhang, C. Shen, R. Ji, and L. Cao, “Supervision transformer,” *International Journal of Computer Vision*, pp. 1-16, 2023, doi: 10.48550/arXiv.2205.11397.

[6] R. Atienza, “Vision transformer for fast and efficient scene text recognition,” *in International Conference on Document Analysis and Recognition*, pp. 319-334, Cham: Springer International Publishing, September 2021, doi: 10.48550/arXiv.2105.08582.

[7] Y. Li, H. Mao, R. Girshick, and K. He, “Exploring plain vision transformer backbones for object detection,” *in European Conference on Computer Vision*, pp. 280-296, Cham: Springer Nature Switzerland, 2022, doi:10.48550/arXiv.2203.16527.

[8] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, “Toward transformer-based object detection,” *arXiv preprint arXiv:2012.09958*, 2020, doi: 10.48550/arXiv.2012.09958.

[9] G. S. Krishna, K. Supriya, and M. Sorgile, “LesionAid: vision transformers-based skin lesion generation and classification,” *arXiv preprint arXiv:2302.01104*, 2023, doi: 10.48550/arXiv.2302.01104.

[10] Y. J. Lee *et al.*, “Demographics prediction and heatmap generation from OCT images of anterior segment of the eye: a vision transformer model study,” *Translational Vision Science and Technology*, vol. 11, no. 11, pp. 7-7, 2022, doi: 10.1167/tvst.11.11.7.

[11] J. Bi, Z. Zhu, and Q. Meng, “Transformer in computer vision,” *in 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pp. 178-188, September 2021, doi: 10.1109/CEI52496.2021.9574462.

[12] D. Gong, J. Lee, M. Kim, S. J. Ha, and M. Cho, “Future transformer for long-term action anticipation,” *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3052-3061, 2022, doi: 10.48550/arXiv.2205.

[13] Y. Chang, F. Li, J. Chen, Y. Liu, and Z. Li, “Efficient temporal flow transformer accompanied with multi-head probsparse self-attention mechanism for remaining useful life prognostics,” *Reliability Engineering and System Safety*, vol. 226, p. 108701, 2022, doi: 10.1016/j.ress.2022.108701.

[14] D. Serdyuk, O. Braga, and O. Siohan, “Audio-visual speech recognition is worth 32328 Voxels,” *in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 796-802, December 2021, doi: 10.48550/arXiv.2109.09536.

[15] A. Berg, M. O'Connor, and M. T. Cruz, “Keyword transformer: a self-attention model for keyword spotting,” *arXiv preprint arXiv:2104.00769*, 2021, doi: 10.48550/arXiv.2104.00769.

[16] Y. Gong, C. I. Lai, Y. A. Chung, and J. Glass, “SSAST: Self-supervised audio spectrogram transformer,” *in Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10699-10709, June 2022, doi: 10.48550/arXiv.2110.09784.

[17] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, “Conditional positional encodings for vision transformers,” *arXiv preprint arXiv:2102.10882*, 2021, doi: 10.48550/arXiv.2102.10882.

[18] S. Atito, M. Awais, W. Wang, M. D. Plumbley, and J. Kittler, “ASiT: audio spectrogram vision transformer for general audio representation,” *arXiv preprint arXiv:2211.13189*, 2022, doi: 10.48550/ARXIV.2211.13189.

[19] Y. Gong, Y. A. Chung, and J. Glass, “AST: audio spectrogram transformer,” *arXiv preprint arXiv:2104.01778*, 2021, doi: 10.48550/arXiv.2104.01778.
[20] N. C. Ristea, R. T. Ionescu, and F. S. Khan, “SEPTR: separable transformer for audio spectrogram processing,” *arXiv preprint arXiv:2203.09581*, 2022, doi: 10.48550/arXiv.2203.09581.
[21] F. Battal, S. Balci, and I. Sefa, “Power electronic transformers: a review,” *Measurement*, vol. 171, p. 108848, 2021.
[22] A. Ghandoura, “Arabic speech commands dataset (v1.0)” *[Data set], Zenodo*, 2021, doi: 10.5281/zenodo.4662481.
[23] O. Mahmoudi, M. F. Bouami, and M. Badri, “Arabic language modeling based on supervised machine learning,” *Revue d'Intelligence Artificielle*, vol. 36, no. 3, p. 467, 2022, doi: 10.18280/ria.360315.
[24] O. Mahmoudi and M. F. Bouami, “Arabic speech emotion recognition using deep neural network,” *in International Conference on Digital Technologies and Applications*, pp. 124-133, Cham: Springer Nature Switzerland, January 2023, doi: 10.1007/978-3-031-29860-8_13.
[25] O. Mahmoudi and M. F. Bouami, “RNN and LSTM models for Arabic speech commands recognition using pytorch and GPU,” *in International Conference on Artificial Intelligence and Industrial Applications*, pp. 462-470, Cham: Springer Nature Switzerland, February 2023, doi: 10.1007/978-3-031-43520-1_39.
[26] O. Mahmoudi and M. F. Bouami, “Arabic speech commands recognition with LSTM and GRU models using CUDA toolkit implementation,” *in 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pp. 1-4, IEEE, May 2023, doi: 10.1109/IRASET57153.2023.10152979.
[27] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “The performance of LSTM and BiLSTM in forecasting time series,” *in 2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019, pp. 3285-3292, doi: 10.1109/BigData47090.2019.9005997.
[28] A. Asroni, K. R. Ku-Mahamud, C. Damarjati, and H. B. Slamat, “Arabic speech classification method based on padding and deep learning neural network,” *Baghdad Science Journal*, vol. 18, no. 2, pp. 0925-0925, 2021, doi: 10.21123/bsj.2021.18.2(Suppl.).0925.
[29] B. Zada and R. Ullah, “Pashto isolated digits recognition using deep convolutional neural network,” *Heliyon*, vol. 6, no. 2, 2020, doi: 10.1016/j.heliyon.2020.e03372.
[30] A. A. Alasadi, T. H. Aldhayni, R. R. Deshmukh, A. H. Alahmadi, and A. S. Alshebami, “Efficient feature extraction algorithms to develop an Arabic speech recognition system,” *Engineering, Technology and Applied Science Research*, vol. 10, no. 2, pp. 5547-5553, 2020, doi: 10.48084/etasr.3465.
[31] A. S. M. B. Wazir and J. H. Chuah, “Spoken Arabic digits recognition using deep learning,” *in 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pp. 339-344, June 2019, doi: 10.1109/I2CACIS.2019.8825004.
[32] Z. Hachkar, A. Farchi, B. Mounir, and J. E. Abbadi, “A comparison of DHMM and DTW for isolated digits recognition system of arabic language,” *International Journal on Computer Science and Engineering (IJCSE)*, vol. 3, no. 3, pp. 1002-1008, Mar. 2011.

BIOGRAPHIES OF AUTHORS

Omayma Mahmoudi (Ph.D. student) received her master’s in data science and intelligent systems from the Poly-Disciplinary Faculty of Nador, University Mohamed Premier of Oujda (Morocco) in 2019. She is now pursuing her Ph.D. in Arabic speech recognition. Her interest is everything about machine learning and deep learning. she has published in internationally reputed journals, books, and conferences. She has also served as a reviewer for scientific journals. She can be contacted at email: mahmoudi.omayma@ump.ac.ma.

Prof. Dr. Naoufal El Allali is a Professor in the Department of Computer Science at the Polydisciplinary Faculty of Nador (FPN), Mohamed First University, Nador, Morocco, affiliated with the MASI Laboratory. He received his Ph.D. degree in Computer Science from FPN in 2022. His current research interests include information retrieval, language representation learning, web mining and text mining, semantic web, web services and web service composition, machine learning, and multi-agent systems. He has published in internationally reputed journals, books, and conferences. He has also served as a reviewer for scientific journals and as a program committee member for several conferences. He can be contacted at email: n.elallali@ump.ac.ma.

Prof. Dr. Mouncef Filali Bouami received an M.Sc. in Electronics from the University of Fez, Morocco in 1998 and a Ph.D. degree from the University of Granada, Spain in 2005 after having defended a doctoral thesis on the modeling of RBF neural networks using T-Norm and T-Conorm operators and weights parameterization. Since 2010 he has been a Senior Lecturer at the Poly-Disciplinary Faculty of Nador, Mohammed premier University, Morocco. His research interest includes machine learning algorithms, text classification and speech recognition methods. He can be contacted at email: m.filalibouami@ump.ac.ma.