ISSN:1533-9211
SEYBOLD REPORT
www.seybold-report.com

A Comprehensive Analysis on Speech to Image Translation using Deep Learning Models

Chandan Rani S R

Research Scholar,

Department of Computer Science and Engineering,

RNS Institute of Technology,

Bangalore, India

Dr. N P Kavya

Professor,

Department of Computer science and Engineering,

RNS Institute of Technology,

Bangalore, India

ISSN: 1533 - 9211

CORRESPONDING AUTHOR:

Chandan Rani S R
Chandanrani.phd@gmail.com

KEYWORDS:

Automated Speech Recognition, Deep learning, Machine learning, Speech-to-image translation, Visual Information.

Received: 06 January 2024
Accepted: 22 January 2024
Published: 29 January 2024

TO CITE THIS ARTICLE:

Rani, C. R. S., & Kavya, N. P. (2024). A Comprehensive Analysis on Speech to Image Translation using Deep Learning Models. *Seybold Report Journal*, *19*(1), 165-182. DOI: 10.5110/77. 1103

Abstract

In the past, the primary means of conveying visual information relied on text-based explanations or visual cues. However, these approaches are not accessible to individuals with visual impairments or those who do not understand the language used for visual information. To address this communication barrier, the concept of speech-to-image translation was introduced, leveraging machine learning and deep learning techniques. This study focuses on the introduction of a speech-to-image translation method using machine learning and deep learning. A dataset was collected containing speech samples paired with corresponding images. The speech data was converted into text using Automatic Speech Recognition (ASR) techniques. Globally, a comprehensive analysis of classification-based deep learning models for image translation is developed and conducted by various researchers. This analysis covered their performance, datasets used, and feature extraction methods. The utilization of deep learning for speech-to-image translation presents several challenges, which are also discussed. Finally, potential areas for future research in this field are identified and presented alongside the listed challenges. In summary, this abstract highlight the introduction of a speech-to-image translation method using machine learning and deep learning. It outlines the dataset collection process, conversion of speech to text, and the analysis of existing deep learning models for image translation. The challenges associated with deep learning-based speech-to-image translation are acknowledged, and potential future research directions are identified.

Introduction

Attention related encoder decoder designing is a real and forceful pattern for speech texts like Automatic Speech Recognition (ASR) and Speech Translation (ST) and that has created consequential progress. Recent advancements in the field have demonstrated the potential for developing end-to-end methods thatdirectly translate source speech to text without the need for intermediate linguistic transcriptions. However,these methods still require paired data consisting of source audio and target text translations for effective end-to-end training [1, 2]. Another area of research focus involves the productive work related to generativeadversarial networks, which have achieved promising outcomes in image synthesis tasks. Recently, there has been growing interest in photo-realistic image synthesis due to its potential applications in computer graphics and photo retouching [3]. Some tasks have developed direct speech translation in each method utilizing deep learning to avoid an issue in a former speech translation [4]. Using the Dual Generator Attentional GAN method, text contexts are regularly encoded into semantic vectors and collected betweenthe generator and differentiator as the provision that displays splendid results in managing entire text containing images [5]. However, these methods established several constraints and modules to produce naturalistic images. These constraints significantly increase the number of parameters and floating-point operations per second required by the methods [6]. The majority of accessible web data is unlabeled, and obtaining human explanations for it can be costly. The requirement for annotated data has been a challenge in machine translation. However, unsupervised machine translation has shown impressive results, even withlimited resources in low-resource linguistics, and it executes remarkably well even for languages that are linguistically distant from each other [7].

The utilization of spectral standardization as a normalization strategy, along with forceful calculational and statistical functions, has achieved several advancements in DNN adversarial training andnormalization [8]. The main goal of this research is to produce naturalistic images that follow the content of provided texts utilizing TextControlGAN and that estimate the method utilizing a quantitative approachand conventional GAN metrics [9]. With the purpose of enhanced GAN, to create the design assembled superior and employed spectral standardization to the GAN structure as well as established the impact and effectiveness of spectral standardization over different empirical. Then, this method can mitigate the loss and should retrieve faster outcomes [10]. As in the text-to-speech method, the suggested method is the same as the vector quantized autoencoder but easily follows the heuristic vector quantized autoencoder training criteria and the empirical English speech database displayed promising outcomes [11]. Through the execution of proposed attention GAN on well-known CelebA, AR Face, and Sefie2Anime datasets obtainedbetter results. The results exhibit an ability to generate images, that possess a heightened sense of realism with enhanced clarity and richer information.

These empirical findings strongly suggest that the attention GAN model holds significant potential for generating highly naturalistic images that accurately capture intricate details [12]. During the inference process, no auxiliary works were employed, and the method operated solely with an individual encoder to input the source linguistics' speech and an individual decoderto produce the target linguistics' speech [13]. This study aims to translate the speech description to a visualimage that is appropriate to that specific audio description. Analyzing and evaluating model performance utilizing various assessment criteria.

Literature Survey

Xinsheng Wang et *al.* [14] implemented the generating images from the spoken description. They proposeda speech-to-image generation architecture that utilized speech representations to generate natural images, without relying on explicit textual information. This approach enabled the inclusion of undeclared languages, providing benefits in scenarios where text-based methods would be limited. The utilization of this approach resulted in high-quality natural image generation and achieved state-of-the-art execution because of the attention GAN model that enhanced visual fidelity and outperformed the previous approaches. However, it would be valuable to evaluate the suggested method on a genuinely undeclared language, rather than relying solely on well-established languages like English Liming wang et *al* [15] implemented more efficacious and precise spoken word discovery for speech-to-image extraction. Despite the use of self-attention layers in embedding learned, TDNN has the best results for speech-to-image extraction. This approach improved the retrieval method's ability to capturemeaningful variations in visual words and proved to be crucial for enhancing word discovery methods. However, the study found that the combination of Sequence Matching Theory (SMT) with self-attention hurts word discovery, resulting in decreased performance.

Jiguo Li et *al.* [16] implemented the direct speech to image translation. The study focused on translating image signals directly from speech signals without the need for transcription. We trained a speech encoder alongside a pre-trained image encoder to illustrate the given speech signals as embedding features. A stacked generative adversarial network was utilized to combine maximum-worth images basedon embedding features. The proposed method achieved excellent performance, and it was believed that generating images from speech signals without relying on text transcription offered a new perspective for extracting semantic details from the speech signals. Nevertheless,

maximum resolution raised the complexity of the approach and made it unstable.

Danny Merkx et *al* [17] implemented language learning utilized the speech to image retrieval. The utilization of associated multiple-layer GRU, salience sampling, cyclic learning rates, and vectorial self- attention led to a significant improvement in image caption extraction execution. Subsequently, the layers in the approach were trained to identify words in the input. When word encoding is present, the identification of deeper network layers is superior. Enhances language learning by providing visual context and real-world associations to aid comprehension and vocabulary acquisition. While images can be helpfulin specific situations, they often lack the depth and richness of real-life contexts.

Xinsheng Wang et *al* [18] suggested speech to image production through adversarial learning. The proposed architecture translated speech descriptions into photo-realistic images without relying on explicitdetails and enabled the inclusion of undeclared languages. An embedding network grasped speech embedding under the supervision of relevant visual information. The suggested approach enhanced in the accuracy and efficiency as well as enhanced the high standard of the generated images because of the suggested S2IGAN assures where the outcome images closely resemble the target or ground truth images and also produce maximum quality images within a reasonable amount of time. S2IGAN heavily relied onthe quality of the input speech to produce accurate and visually clear images. However, the presence of noisy speech could potentially affect the robustness and reliability of the model.

Taxonomy

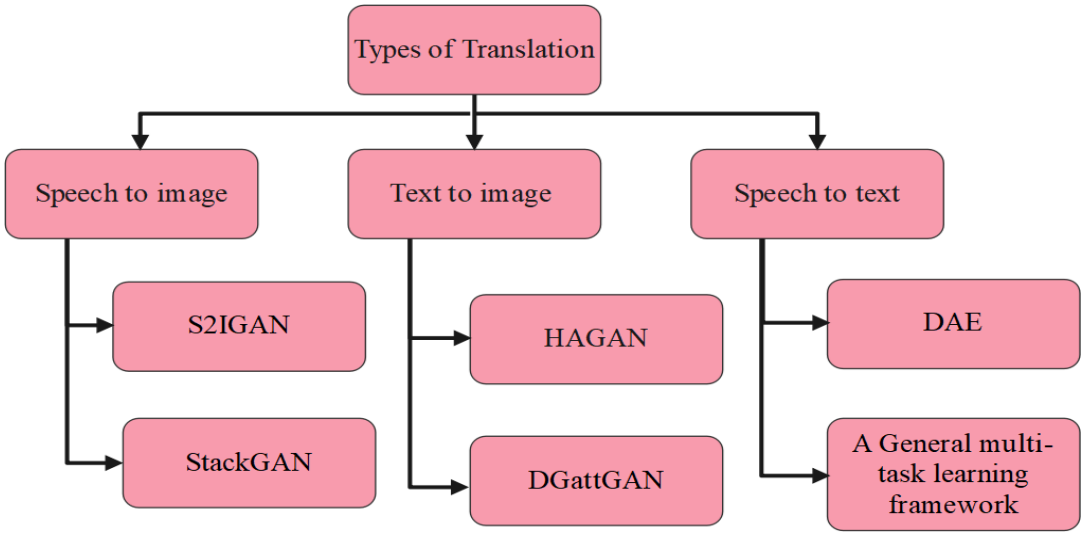

Figure 1: Taxonomy of Speech to image translation using Deep Learning

Speech to Image

Speech to Image Generative Adversarial Network (S2IGAN)

The main aim of the method is to produce modified images that are semantically aligned with a given spoken statement. To achieve this goal, the researchers propose the use of S2IGAN, a framework consistingof two key modules: the Speech Embedding Network (SEN) and the Relation-supervised Densely-stacked Generative (RDG) module. The advantage of S2IGAN enables the generation of visually realistic images from speech, bridging the gap between auditory and visual modalities, and Facilitates applications in multimedia content creation, virtual environments, and accessibility by providing a direct pathway for converting speech into visual representations [14].

StackGAN

StackGAN utilizes a two-stage generative model, consisting of a text-conditional generator and a refinement network, to generate high-resolution images from textual descriptions. In the first stage, a low-resolution image is generated based on the text input and in the subsequent stage, a refinement network upscales the low-resolution image to a higher resolution with finer details. The advantage of StackGAN lies in its ability to generate more realistic and visually appealing images by incorporating both global andlocal conditioning information from the text, resulting in improved image quality and better alignment with textual descriptions. [16].

Text to image

Hybrid Attentional Generative Adversarial Networks (HAGAN)

HAGAN employ an attention mechanism to focus on specific regions of an image and generate high-quality, realistic images based on textual descriptions. The advantage of HAGAN lies in their ability to generate images with improved visual details and better alignment with textual descriptions by selectively attending to relevant regions, resulting in more accurate and visually pleasing image synthesis.[3].

Dual Generator Attentional GAN (DGattGAN)

Dual Generator Attentional GAN employs two generators, each focusing on different levels of image details, and an attention mechanism to generate high-quality images conditioned on textual descriptions. Its ability to generate images with enhanced visual details and better alignment with textual descriptions by leveraging dual generators and attention mechanisms, results in improved image synthesisquality [5].

Speech-to-text

Unsupervised Speech Processing and Denoising Autoencoder (DAE)

The architecture of this method further developed an unsupervised and Machine translation method of speech processing. At initial derive an ST method and integrate a linguistic method into an architecture, then process the translated outcomes utilizing DAE. The unsupervised speech segmentation approach is initially utilized to produce speech segments correlated to the spoken words. Enhancing the quality of the translation, employing the DAE to precise the translated outcomes [2].

A General multi-task learning framework

This approach suggested a typical multi-task architecture to use text data for ASR and ST works. The machine translation work is selected as an auxiliary work to be combined and trained alongside of theST work. In the Automatic Speech Recognition (ASR) work, the auxiliary work points the given phonemeconsequential to the correlated sub word tokens derivative from statement [1].

Table 1: Comparative analysis

Author	Methodology	Advantages	Limitations
Xinsheng Wang et al. [14]	Speech to image Generation (S2IG)	The utilization of this approach resulted in high-quality natural image generation and achieved state-of-the-art execution because of the attention GAN model that enhanced visual fidelity and outperforms the previous approaches	However, it required testing the suggested method on a truly undeclared linguistic in preference to the well-furnished English language.

Liming wang et *al* [15]	Time Delay Neural Network (TDNN)	This approach improved the retrieval method's ability to capture meaningful variations in visual words and proved to be crucial for enhancing word discovery methods.	However, the study found that the combination of Sequence Matching Theory (SMT) with self-attention harms word discovery, resulting in decreased performance.
Jiguo Li et al [16]	Speech-to-image translation model	The proposed method achieved excellent performance, and it was believed that generating images from speech signals without relying on text transcription offered a new perspective for extracting semantic details	However, maximum resolution raised the complexity of the approach and made it unstable.

		from the speech signals.	
Danny Merkx et al [17]	3 layer Bi-directional GRU	Enhances language learning by providing visual context and real - world associations to aid comprehension and vocabulary acquisition.	While images can be helpful in specific situations, they often lack the depth and richness of real-life contexts.
Xinsheng Wang et al [18]	Speech to image generation (S2IGAN) framework	The suggested approach enhanced the accuracy and efficiency as well as enhance the high standard of the generated images because the suggested S2IGAN assures that the outcome images closely resemble the target	S2IGAN heavily relied on the quality of the input speech to produce accurate and visually clear images. However, the presence of noisy speech could potentially affect the robustness and reliability of the model.

		or ground truth images and also produce maximum quality images within a reasonable amount of time.	
Layne Berry et al [19]	Multilingual SpeechCLIP model	The grasped speech encoders could execute zero shot speech text extraction for English text even when the speech was not in English. These occurred by language versatility where the speech encoders could grasp and retrieve English text even when it was in various languages.	The effectiveness and performance of M-SpeechCLIP heavily depend on the quality, diversity, and representativeness of the pre-trained data. Insufficient diversity or lack of variation in the pre-trained data can lead to biases and hinder the execution of speech-to-image retrieval tasks.

Jianwei Zhu et al. [20]	Phased Bidirectional Generation Network (PBGN) was proposed.	The normalization method made the model training more stable because of the normalization method that helps to stabilize the training process of the model.	PBGN in text-to-image synthesis is a complex training process, which may require significant computational resources and time.

Problem Statement

The problem statement of this work is stated as follows:

- Speech-to-image translation falls short of providing sufficient contextual detail for language learners. At the same time, images can be supportive in particular situations, but they often lack the depth and richness of actual life contexts.
- High-resolution images offer more detailed information, resulting in more naturalistic generated images. However, increasing the maximum resolution raises the complexity of the approach and introduces instability.
- The success of the Speech-to-Image Generative Adversarial Network (S2IGAN) heavily relies on the quality of the input speech to produce accurate and visually clear images. However, the presence of noisy speech can significantly affect the robustness and reliability of the model.
- The effectiveness and performance of M-SpeechCLIP largely depend on the quality, variation, and representativeness of the pre-trained data. Lack of diversity in the pre-trained data can introduce biases or hinder the performance of speech-to-image retrieval works.

Objectives

- To improve the depth and richness of actual life contexts of the speech-to-image translation by using the deep learning algorithm.
- To increase the stability and minimize the consumption time of high-resolution images by using CNN-based algorithms.
- To increase the robustness and reliability of the noisy speech method by using the deep learning algorithm of the Long Short-term Memory algorithm.
- To improve the performance and diversity in the pre-trained data of speech-to-image retrieval by utilizing the Deep Neural Network algorithm.

Conclusion

The concept of speech-to-image translation aims to ease cross-modal interaction, which enhances accessibility and improves the entire grasping and communication with visual detail. It is established to bridge the space among various methods of interaction, especially among spoken language and visual representations. It can have different approaches and advantages as well and establishing the speech-to- image translation for each one can utilize spoken linguistics to explain visual content. The method of speech-to-image translation aids in overcoming the interaction block between spoken language and visual illustrations with machine learning and deep learning techniques. In this study, the speech-to-image translation method was introduced with machine learning and deep learning. Initially, the dataset is collected that contains speech and appropriate image samples. The speech data are converted into text utilizing Automatic Speech Recognition (ASR) method. Along with this, a complete analysis of various former DL related classification models for image translation created by different researchers against the world is covered, involving how well they work, the datasets they utilize, and the feature extraction methods. The utilization of the DL method will further require the performance of speech-to-image translation in the future.

COMPETING INTERESTS

The authors have no compting interest to declare.

Author's Affiliation

Chandan Rani S R

Research Scholar,

Department of Computer Science and Engineering,

RNS Institute of Technology,

Bangalore, India

Dr. N P Kavya

Professor,

Department of Computer science and Engineering,

RNS Institute of Technology,

Bangalore, India

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/ licenses/by/4.0/. *Seybold Report* is a peer-reviewed journal published by Seybold Publications.

HOW TO CITE THIS ARTICLE:

Rani, C. R. S., & Kavya, N. P. (2024). A Comprehensive Analysis on Speech to Image Translation using Deep Learning Models. *Seybold Report Journal*, *19*(1), 165-182. DOI: 10.5110/77. 1103

REFERENCES

1. Tang, Yun, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. "A general multi-task learning framework to leverage text data for speech to text tasks." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6209-6213. IEEE, 2021.
2. Chung, Yu-An, Wei-Hung Weng, Schrasing Tong, and James Glass. "Towards unsupervised speech-to-text translation." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7170-7174. IEEE, 2019.
3. Chung, Yu-An, Wei-Hung Weng, Schrasing Tong, and James Glass. "Towards unsupervised speech-to-text translation." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7170-7174. IEEE, 2019.
4. Kano, Takatomo, Sakriani Sakti, and Satoshi Nakamura. "Transformer-based direct speech-to-speech translation with transcoder." In *2021 IEEE Spoken Language Technology Workshop (SLT)*,pp. 958-965. IEEE, 2021.
5. Zhang, Han, Hongqing Zhu, Suyi Yang, and Wenhao Li. "DGattGAN: Cooperative up-sampling based dual generator attentional GAN on text-to-image synthesis." *IEEE Access* 9 (2021): 29584- 29598.
6. Ma, Ruixin, and Junying Lou. "CPGAN: An efficient architecture designing for text-to-image generative adversarial networks based on canonical polyadic decomposition." *Scientific Programming* 2021 (2021): 1-9.
7. Das, Anindya Sundar, and Sriparna Saha. "Self-supervised image-to-text and text-to-image synthesis." In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*, pp. 415-426. Springer International Publishing, 2021.

8. Gafni, Oran, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. "Make-a-scene: Scene-based text-to-image generation with human priors." In *European Conference on Computer Vision*, pp. 89-106. Cham: Springer Nature Switzerland, 2022.

9. Ku, Hyeeun, and Minhyeok Lee. "TextControlGAN: Text-to-Image Synthesis with Controllable Generative Adversarial Networks." *Applied Sciences* 13, no. 8 (2023): 5098.

10. Cruz, Anunshiya Pascal, and Jitendra Jaiswal. "Text-to-image classification using attngan with densenet architecture." In *Proceedings of International Conference on Innovations in Software Architecture and Computational Systems: ISACS 2021*, pp. 1-17. Springer Singapore, 2021.

11. Yasuda, Yusuke, Xin Wang, and Junichi Yamagishd. "End-to-end text-to-speech using latent duration based on VQ-VAE." In *ICASSP 2021-2021 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP)*, pp. 5694-5698. IEEE, 2021.

12. Tang, Hao, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. "Attentiongan: Unpaired image-to- image translation using attention-guided generative adversarial networks." *IEEE transactions on neural networks and learning systems* (2021).

13. Kano, Takatomo, Sakriani Sakti, and Satoshi Nakamura. "Transformer-based direct speech-to-speech translation with transcoder." In *2021 IEEE Spoken Language Technology Workshop (SLT)*,pp. 958-965. IEEE, 2021.

14. Wang, Xinsheng, Tingting Qiao, Jihua Zhu, Alan Hanjalic, and Odette Scharenborg. "S2IGAN: Speech-to-image generation via adversarial learning." *arXiv preprint arXiv:2005.06968* (2020).

15. Wang, Liming, Xinsheng Wang, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. "Align or attend? toward more efficient and accurate spoken word discovery using speech-to-imageretrieval." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*, pp. 7603-7607. IEEE, 2021.

16. Li, Jiguo, Xinfeng Zhang, Chuanmin Jia, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, and

Wen Gao. "Direct speech-to-image translation." *IEEE Journal of Selected Topics in Signal Processing* 14, no. 3 (2020): 517-529.

17. Merkx, Danny, Stefan L. Frank, and Mirjam Ernestus. "Language learning using speech to image retrieval." *arXiv preprint arXiv:1909.03795* (2019).
18. Wang, Xinsheng, Tingting Qiao, Jihua Zhu, Alan Hanjalic, and Odette Scharenborg. "Generating images from spoken descriptions." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 850-865.
19. Berry, Layne, Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Hung-yi Lee, and David Harwath. "M-SpeechCLIP: Leveraging Large-Scale, Pre-Trained Models for Multilingual Speech to Image Retrieval." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*, pp. 1-5. IEEE, 2023.
20. Zhu, Jianwei, Zhixin Li, Jiahui Wei, and Huifang Ma. "PBGN: Phased bidirectional generation network in text-to-image synthesis." *Neural Processing Letters* 54, no. 6 (2022): 5371-5391.