UNIVERSIDAD NACIONAL "HERMILIO VALDIZÁN"

FACULTAD DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

CARRERA PROFESIONAL DE INGENIERÍA DE SISTEMAS

DESARROLLO DE UNA SOLUCIÓN MEDIANTE EL USO DE SPEECH RECOGNITION Y GENERATIVE ADVERSARIAL NETWORKS PARA LA GENERACIÓN AUTOMÁTICA DEL RETRATO HABLADO

TESIS PARA OPTAR EL TÍTULO DE INGENIERO DE SISTEMAS

TESISTA:
Bach. BRAD AXL FIGUEROA CARLOS

ASESORA:
Mg. HEIDY VELSY RIVERA VIDAL

HUÁNUCO – PERÚ
2021

DEDICATORIA

A Dios
A mi madre
A mi abuela

AGRADECIMIENTO

A todas las personas involucradas en el desarrollo de esta investigación.

RESUMEN

A través de la presente investigación se desarrolla y propone una solución, haciendo uso de tecnologías de vanguardia relacionadas al campo de la Inteligencia Artificial como Speech Recognition y Generative Adversarial Network, con el objetivo de generar una imagen sintética de un rostro de manera automática partiendo desde la descripción del rostro considerado este como un retrato hablado. El desarrollo fue dividido en dos procesos principales: conversión de audio a texto y generación del rostro. En el primero se hizo uso de la tecnología Speech Recognition y en el segundo Generative Adversarial Network (GAN). En el primero se aplicó una variación de la técnica Transfer Learning conocido como Cross-Language Transfer Learning a través del toolkit NeMo con una arquitectura QuartzNet 15x5 y se reentrenó dicho modelo con 3 datasets distintos en español, después de 21 experimentos se escogió tres de ellos para la obtención del modelo final totalmente enfocado a la tarea que se requiere. En el segundo se entrenó el modelo desde cero reutilizando y adaptando un proyecto existente, después de 19 experimentos se eligieron los cinco modelos que dieron mejores resultados.

Con respecto a la naturaleza de la investigación, es no experimental de corte transversal, de tipo correlacional-causal

debido a que se estudió la correlación entre las variables de estudio sin su manipulación deliberada y obteniendo los datos requeridos en un tiempo y momento determinado, con un nivel explicativo (explicación de la relación de las variables) y aplicado ya que se propone la aplicación de una solución con el fin de mejorar el fenómeno descrito.

Los resultados obtenidos son alentadores para una futura investigación, ya que el mejor modelo elegido para el primer proceso principal obtuvo un indicador WER de 0.13 en el entrenamiento, 0.35 en la validación y 0.59 en el testeo, siendo estos buenos valores dentro del ámbito del Speech Recognition.

Por otro lado, los resultados obtenidos con respecto a la función de pérdida o loss de los mejores modelos elegidos están en el rango de 0.08 y 1.4 para la red discriminadora y entre 0.53 y 5.2 para la red generadora, estos valores son totalmente explicables y se sustentan en el hecho de que no se ha podido invertir muchos recursos como hardware de gama alta para su entrenamiento.

Por último, los resultados obtenidos por ambos procesos principales en conjunto demuestran que existen características de las transcripciones que son bien representadas en función a los modelos utilizados para la generación automática del retrato hablado.

Palabras clave: Propuesta, solución, Speech Recognition, Generative Adversarial Network, NeMo, GAN.

SUMMARY

Through this research, a solution is developed and proposed, making use of cutting-edge technologies related to the field of Artificial Intelligence such as Speech Recognition and Generative Adversarial Network, with the aim of automatically generating a synthetic image of a face starting from the description of the face considered this as a spoken portrait. The development was divided into two main processes: audio to text conversion and face generation. Speech Recognition technology was used in the first and Generative Adversarial Network (GAN) in the second. In the first, a variation of the Transfer Learning technique known as Cross-Language Transfer Learning was applied through the NeMo toolkit with a QuartzNet 15x5 architecture and said model was retrained with 3 different datasets in Spanish, after 21 experiments three of them were chosen to obtain the final model totally focused on the task that is required. In the second, the model was trained from scratch, reusing and adapting an existing project. After 19 experiments, the five models that gave the best results were chosen.

Regarding the nature of the research, it is non-experimental of cross-sectional, correlational-causal type because the correlation between the study variables was studied without their deliberate manipulation and obtaining the required data in a given time and

moment, with an explanatory level (explanation of the relationship of the variables) and applied since the application of a solution is proposed in order to improve the phenomenon described.

The results obtained are encouraging for future research, since the best model chosen for the first main process obtained a WER indicator of 0.13 in training, 0.35 in validation and 0.59 in testing, these being good values within the scope of Speech. Recognition.

On the other hand, the results obtained with respect to the loss or loss function of the best-chosen models are in the range of 0.08 and 1.4 for the discriminating network and between 0.53 and 5.2 for the generating network, these values are fully explainable and can be They are based on the fact that it has not been possible to invest many resources such as high-end hardware for their training.

Finally, the results obtained by both main processes together show that there are characteristics of the transcripts that are well represented based on the models used for the automatic generation of the spoken portrait.

Keywords: Proposal, solution, Speech Recognition, Generative Adversarial Network, NeMo, GAN.