API DE RECONOCIMIENTO DE VOZ PARA LENGUAJE CASTELLANO USANDO REDES NEURONALES

SERGIO FLORES LABRA
INGENIERO CIVIL EN COMPUTACIÓN

RESUMEN

En los últimos años, el Reconocimiento de Voz ha jugado un papel importante dentro del mundo tecnológico. Esta tecnología permite, entre otras aplicaciones, la transcripción de audio a texto, sistemas de subtitulado en tiempo real, traducción de frases audibles y asistentes virtuales para dispositivos móviles, etc. Existen muchos mecanismos para la solución de este tipo de problemas, tales como los clasificadores, los filtros de señales y las redes neuronales; de los cuales este último ha logrado llevarse los elogios por el porcentaje de exactitud de res puesta. En los últimos 17 años, con la incorporación de Deep Learning, se ha alcanzado exactitudes superiores a un 98% según lo declarado por Google. El modelo de Red Neuronal que ha permitido estos grandes avances es el modelo LSTM, el cual consiste en una red neuronal que posee una memoria a largo plazo y en base a su retroalimentación y un par de mejoras, permite entregar una respuesta más precisa al calcular la probabilidad de la palabra que se está detectando. Las soluciones existentes requieren del acceso al centro de procesamiento del proveedor, afectando a los clientes que no cuentan con conexión a internet. Por lo anterior, en esta memoria se construye una API que incluye funcionalidades para los trabajos de reconocimiento de voz en el lenguaje castellano y que pueda trabajar desconectada de la internet. El ideal es poder contar con una herramienta que no solo reconozca la frase, sino que también pueda corregir la sintaxis de la respuesta por medio de un diccionario de palabras. Para evaluar si la API logra cumplir con las expectativas planteadas, se utilizan pruebas de exactitud del modelo y tiempo de respuesta. El corrector de palabras basado en diccionario responde en 0.5 segundos con una exactitud de un 93.8%. Cosa distinta se da en el modelo de reconocimiento, ya que los resultados no pueden ser medidos de la misma forma por no contar con una base de datos de entrenamiento razonablemente grande. Esto se debe a que los modelos neurales para reconocimiento de voz basados en Deep Learning necesitan mucha información para su entrenamiento, la cual es muy difícil de conseguir, ya que estamos

hablando del orden de los cientos de miles de datos. Por otra parte el hecho de trabajar con lenguaje Castellano dificulta aún más el problema, ya que los caracteres con tildes y diéresis causan dificultades en el procesamiento.

ABSTRACT

In recent years, Voice Recognition has played an important role in the technological world. This technology allows, among other applications, audio-to-text transcription, real-time subtitling systems, translation of audible phrases and virtual assistants for mobile devices, etc. There are many mechanisms for solving this type of problem, such as classifiers, signal filters and neural networks, of which the latter has been praised for the percentage of response accuracy. In the last 17 years, with the incorporation of Deep Learning, accuracy has been achieved above 98% as declared by Google. The Neuronal Network model that has allowed these breakthroughs is the LSTM model, which consists of a neural network that has a long term memory and based on its feedback and a couple of improvements, allows to provide a more precise response when calculating the probability of the word being detected. Existing solutions require access to the provider's processing center, afecting customers without an Internet connection. Therefore, an API is built in this memory that includes functionalities for speech recognition works in Spanish language and that can work disconnected from the Internet. The ideal is to have a tool that not only recognizes the phrase, but can also correct the syntax of the answer by means of a word dictionary. Tests of model accuracy and response time are used to assess whether the API is meeting expectations. The dictionary-based word corrector responds in 0.5 seconds with an accuracy of 93.8%. This is diferent from the recognition model, as results cannot be measured in the same way because there is no reasonably large training database. This is because the neural models for speech recognition based on Deep Learning need a lot of information for their training, which is very difficult to achieve, since we are talking about the order of hundreds of thousands of data. On the other hand, the fact of working with Spanish language makes the problem even more difficult, since the characters with accents and dieresis cause difficulties in the processing.