Sociedad de Ingeniería de Audio

Artículo de Congreso

Congreso Latinoamericano de la AES 2018
24 a 26 de Septiembre de 2018
Montevideo, Uruguay

Análisis Automático de Voz Hablada para Detección de Dificultades en el Aprendizaje de la Lectura

Gabriel De Cola,[1] Guzmán Chalupa,[1] Martín Rocamora[1] y Pablo Cancela[1]

[1] UDELAR FING, Instituto de Ingeniería Eléctrica
Montevideo, Montevideo, 11300, Uruguay

gabriel.de.cola, guzman.chalupa, rocamora, pcancela @fing.edu.uy

RESUMEN

Se propone un sistema para el análisis de un test de dificultad lectora en niños: la nominación automatizada rápida (RAN), que consiste en la denominación secuencial de un conjunto de 5 símbolos dispuestos en una matriz de 5 filas. La solución se basa en que los elementos del test se repiten. Se segmenta la señal de audio identificando inicio y fin de cada palabra, y se extraen características para compararlas, teniendo en cuenta deformaciones temporales. Las palabras se agrupan automáticamente y se compara la matriz original y la secuencia identificada, lo que permite detectar los errores cometidos.

0. INTRODUCCIÓN

El test RAN (del inglés, *Rapid Automatized Naming*) es un predictor de dificultades del lenguaje muy utilizado en niños [1], en el que se debe nombrar en voz alta y de forma secuencial una serie de elementos dispuestos en una matriz. Los elementos de la matriz provienen de un alfabeto de 5 símbolos, pudiendo ser letras, números, colores u objetos. La evaluación del test mide el tiempo en realizarlo y los errores cometidos.

En este trabajo se analizan de forma automática grabaciones de tests RAN, evaluando los errores de omisión, sustitución o inserción de elementos en la lectura de la secuencia. Se construyó una base de datos de archivos de audio bajo ciertas hipótesis: contener sólo la voz de quien realiza el test, con pausas entre palabras, y buena relación señal a ruido.

1. SOLUCIÓN PROPUESTA

La solución propuesta aprovecha la repetición de elementos en la matriz, es decir, se basa en la autosimilitud de la señal de audio. El sistema implementado segmenta la señal de audio para identificar el inicio y el fin de cada palabra usando un detector de actividad vocal (VAD). Luego extrae características que permiten comparar las palabras detectadas. Existe una etapa de alineamiento entre los segmentos de audio correspondientes a las palabras detectadas, que busca establecer su similitud, teniendo en cuenta deformaciones temporales entre realizaciones diferentes de una mis-

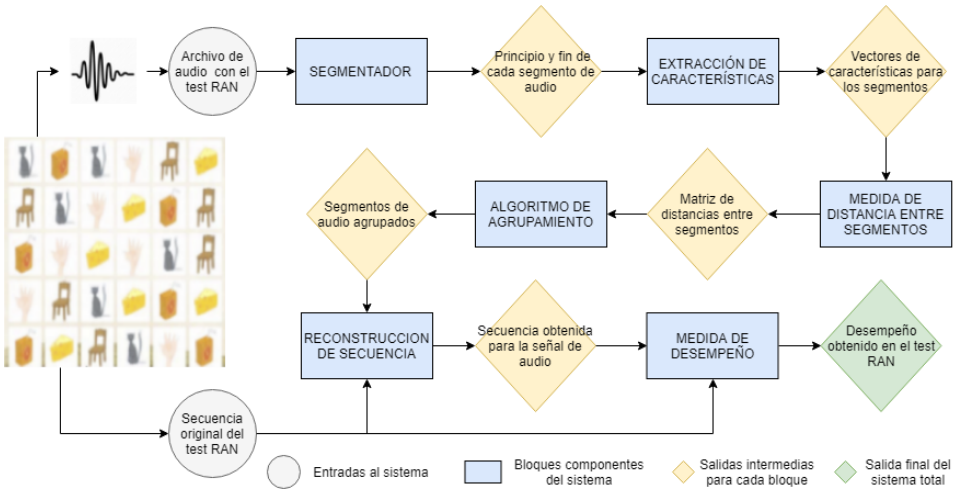

Figura 1: Diagrama de bloques del sistema propuesto.

ma palabra. Luego, los segmentos de audio se agrupan automáticamente para establecer cuáles corresponden a una misma palabra. Por último, se realiza una comparación entre la matriz original del test RAN y la secuencia de palabras identificadas, que permite detectar los errores cometidos y asignar un valor de desempeño. La Fig. 1 muestra un diagrama del sistema implementado.

El segmentador de palabras usa la energía en tiempo corto de cada trama de audio, y solo si supera cierto umbral se considera voz hablada [2]. Luego, se extraen los Coeficientes Cepstrales de Frecuencia Mel (MFCC) que representan características tímbricas de la señal de audio. Para comparar los segmentos de voz identificados se emplea Dynamic Time Warping (DTW), ya que permite tener en cuenta deformaciones temporales entre realizaciones diferentes de una misma palabra. La implementación de DTW usa la distancia cepstral ponderada por seno elevado entre los MFCCs de las tramas de audio. El camino de alineamiento óptimo entre dos palabras se obtiene calculando la distancia acumulada mínima entre las tramas de audio correspondientes.

Las palabras se agrupan usando Spectral Clustering de modo de obtener un grupo por cada tipo de símbolo del test. Para ello se construye una matriz de similitud entre cada pareja de segmentos de voz usando la distancia obtenida con DTW. Además, se detectan y eliminan valores atípicos comparando la distancia de cada elemento con el resto de los elementos del grupo.

Por último, para recrear la secuencia vocalizada se consideran todas las posibles correspondencias entre grupos y símbolos. Usando distancia de edición se comparan las secuencias resultantes con la secuencia RAN de referencia. El puntaje resultante se basa en los errores cometidos para la secuencia con menor distancia.

2. Base de datos

Se generó una base de datos de 120 grabaciones de audio, para una misma versión del test RAN cuyos símbolos son objetos, con etiquetas de inicio y fin de palabras, en la que participaron 20 hablantes adultos (hombres y mujeres). Para cada hablante hay dos grabaciones donde se lee correctamente la secuencia, y cuatro donde se introducen algunos errores de interés.

3. Experimentos y resultados

La base de datos creada permite evaluar el sistema propuesto. Se realizaron experimentos para entrenar y evaluar el desempeño del segmentador y el resto del sistema (denominado clasificador) de forma independiente, usando validación cruzada con una partición de los datos por hablante. Para la evaluación del sistema total se utilizaron los parámetros óptimos obtenidos de entrenar cada módulo independientemente. El desempeño del segmentador es de 92.87 %. La clasificación con segmentación manual alcanza 97.75 % de acierto, mientras que el sistema total con segmentación automática obtiene 90.44 % de acierto.

4. Conclusiones

Se propuso un sistema para el análisis automático de la grabación de un test RAN basándose en la autosimilitud de la señal. Se creó una base de datos con adultos que arroja buenos resultados. Se puede esperar que al evaluar en niños el desempeño disminuya, no obstante, los resultados obtenidos son prometedores, validan el enfoque y sugieren lineas de trabajo futuro.

Agradecimientos

Trabajo parcialmente financiado por ICT4V.

Referencias

[1] Carreiras M. Valle Lisboa J. Zugarramurdi C., Lallier M., “Diseño de una evaluación digitalizada de predictores de las dificultades lectoras,” 2016.

[2] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 2002.