GENERACIÓN DE VIDEO MEDIANTE EL EMPAREJAMIENTO DE IMÁGENES TRAS PROCESAMIENTO DEL LENGUAJE NATURAL EN CANCIONES

Johan Manuel Cabrera Chavarro

Juan Manuel Gonzales Silva

Universidad Surcolombiana

Facultad de Ingeniería

Ingeniería de Software

Neiva - Huila

2020

GENERACIÓN DE VIDEO MEDIANTE EL EMPAREJAMIENTO DE IMÁGENES TRAS PROCESAMIENTO DEL LENGUAJE NATURAL EN CANCIONES

JOHAN MANUEL CABRERA CHAVARRO

JUAN MANUEL GONZALES SILVA

Trabajo de grado presentado como requisito para optar al título de Ingenieros de Software

Asesor: FERNANDO ROJAS ROJAS

UNIVERSIDAD SURCOLOMBIANA

FACULTAD DE INGENIERÍA

INGENIERÍA DE SOFTWARE

NEIVA - HUILA

2020

Nota de aceptación

_____

_____

_____

_____

_____

_____

_____

Firma del presidente del jurado

_____

Firma del jurado

_____

Firma del jurado

Neiva, febrero del 2020

Contenido

Lista de Tablas

Lista de Figuras

Resumen

Pese a tratarse de dos tipos de medios diferentes, el ser humano ha tendido a combinar la música con imágenes para producir videos musicales. Este tipo de contenido multimedia se popularizó a partir de la década de 1980 y aunque es realizado en su mayoría por personas expertas, hoy en día prácticamente cualquiera puede realizarlo. Sin embargo, esa labor de emparejar imágenes con la música, en algunas ocasiones de acuerdo con el sentido de la letra de una canción, sigue requiriendo de mucho esfuerzo y tiempo. Como una posible solución al anterior problema, en este documento se describe un sistema para generar videos tras el procesamiento del lenguaje natural en audio y texto usando los respectivos servicios de Google Cloud y aplicando la técnica de extracción de aspectos sobre canciones.

Palabras claves: procesamiento del lenguaje natural, dictado a texto, extracción de aspectos, generación de video

Abstract

Despite being two different types of media, the human being has tended to combine music with images to produce music videos. This type of multimedia content became popular since the 1980s and although it is mostly done by experts, today virtually anyone can do it. However, this task of matching images with music, sometimes according to the meaning of the lyrics of a song, still requires a lot of effort and time. As a possible solution to the previous problem, this document describes a system to generate videos after processing the natural language in audio and text using the respective Google Cloud services and applying the aspect extraction technique about songs.

Keywords: natural language processing, speech to text, aspect extraction, video generation

Introducción

El video musical es un tipo de contenido multimedia, muy popular a partir de la década de 1980, que está conformado por música e imágenes que le aportan una gran representación visual (Cai *et al.*, 2007; Xu *et al.*, 2008). Es utilizado como una forma efectiva de entretenimiento (Funasawa *et al.*, 2010), útil, en algunos casos, para que los artistas lancen sus canciones al estrellato (Foote *et al.*, 2002).

La popularidad de este medio no se debe solo al hecho que transmite a las personas información auditiva y visual de manera más entretenida y divertida (Cai *et al.*, 2007; Xu *et al.*, 2008). Pese a ser dos medios distintos, la música y la imagen son percibidas por los humanos con una fuerte conexión y correlación (Wu, Xu, *et al.*, 2012). En psicología, el efecto sinérgico ocasionado por la mezcla de señales auditivas y visuales es llamado fenómeno de simpatía (Iwamiya, 1992), y consiste en una experiencia musical que evoca emociones, recuerdos o eventos imaginarios (Juslin & Västfjäll, 2008). Por lo tanto, el video (o conjunto de imágenes) y la música a menudo se complementan para mejorar la resonancia emocional (Lin *et al.*, 2015).

Otro factor importante para los videos musicales, principalmente de canciones, es la letra. De ella se puede obtener con mayor facilidad el escenario visual, cuando las características acústicas son complejas (Funasawa *et al.*, 2010). En sí, el contenido visual tiene que capturar el significado semántico de la letra, la inspiración emocional de la música o ambos (Wu, Xu, *et al.*, 2012). Esto denota una conexión semántica entre la música y la imagen, como es demostrado por Wu *et al.* (2012).

En este documento encontrará secciones resumidas del procesamiento del lenguaje natural en audio, texto y la técnica de extracción de aspectos, con las cuales se dará contexto al

sistema de generación de video. Igualmente se describirá la metodología utilizada y se mostrará el análisis del desempeño de las fases del sistema.

Justificación

Aunque en el año de realización de este documento (2020) existan diversos softwares para la creación de videos como: Adobe After Effects, Adobe Premier o Windows MovieMaker, esta labor sigue siendo ardua para las personas, inclusive para las profesionales en la materia. Esta dificultad se debe en primer lugar a que para recopilar las imágenes adecuadas para la música y la letra se debe invertir mucho tiempo (y en ocasiones dinero) (Fan *et al.*, 2016; Funasawa *et al.*, 2010). Además, hacer que las imágenes y la música se ajusten y sincronicen de manera idónea, no es una tarea trivial y requiere de mucho esfuerzo (Funasawa *et al.*, 2010; Lin *et al.*, 2015; Xu *et al.*, 2008; Yoon *et al.*, 2009).

Una solución para este problema es el uso de sistemas computacionales que generen, ya sea con recursos propios o disponibles en la web, videos de manera automática. Esto implicaría un incremento en la eficiencia de producción de este tipo de contenido multimedia (Fan *et al.*, 2016; Lin *et al.*, 2015).

Objetivos

General

Desarrollar una aplicación web que permita la generación de videos de canciones, mediante el emparejamiento de imágenes de acuerdo con aspectos extraídos de la letra tras el procesamiento del lenguaje natural.

Específicos

Desarrollar una página web con una interfaz que sea de fácil uso para los usuarios.

Usar los servicios de Google Cloud para transcribir el discurso presente en un archivo de audio.

Utilizar los servicios de Google Cloud para realizar el análisis sintáctico sobre la transcripción obtenida del audio.

Desarrollar un algoritmo que permita la extracción de aspectos presentes en el texto basándose en el resultado del análisis sintáctico.

Integrar las API de dos repositorios de imágenes que permitan obtener imágenes adecuadas para los aspectos extraídos.

Crear mediante el uso de la librería FFMPEG y con el conjunto de imágenes de los aspectos y el audio dado por el usuario un video en formato MP4.

Marco Teórico

Planteamiento del Problema

¿Cómo crear una aplicación web que permita la generación de videos de canciones, mediante el emparejamiento de imágenes de acuerdo con aspectos extraídos de la letra tras el procesamiento del lenguaje natural?

Procesamiento del Lenguaje Natural (PLN)

El procesamiento del lenguaje natural (PLN) es un área de la informática, la inteligencia artificial y la lingüística que implementa varias técnicas para facilitarle a las máquinas el análisis, comprensión, generación, procesamiento, manipulación e identificación del lenguaje escrito y hablado por los humanos (Chopra *et al.*, 2013; Kambhatla & Zitouni, 2013; Shivakumar *et al.*, 2016).

PLN en Audio

El habla es una necesidad primaria y la forma más conveniente de comunicación. Debido a eso, se han creado formas de facilitar la interacción entre humanos y computadores (Khilari & P., 2015).

Una de las áreas de aplicación del PLN es el procesamiento del habla, el cual implica el estudio, codificación y decodificación de señales de audio emitidas por humanos para producir una letra, palabra u oración (Khilari & P., 2015; Shivakumar *et al.*, 2016).

El procesamiento del habla es empleado en codificación de voz, síntesis de voz, reconocimiento de voz y tecnologías de reconocimiento de locutor. De las anteriores, la aplicación más importante es el reconocimiento de voz (Khilari & P., 2015).

El principal propósito del reconocimiento de voz es darle la capacidad a una máquina de "escuchar", "entender" y "actuar" a información hablada, procesando una señal acústica obtenida de un micrófono o teléfono, para generar un conjunto o secuencia de palabras por medio de un algoritmo implementado como un programa de computadora (Gaikwad *et al.*, 2010; Khilari & P., 2015). Los sistemas de reconocimiento de voz se intentaron por primera vez a principios de la década de 1950 en los Laboratorios Bell (Khilari & P., 2015).

Estos sistemas pueden ser separados en distintas clases de acuerdo con el tipo de expresión que pueden reconocer: palabra aislada o expresión aislada, palabra conectada, habla continuo y habla espontáneo (Das, 2012). También pueden ser clasificados como sistemas dependientes e independientes del hablante (Allawadi, 2012; Das, 2012).

El sistema de reconocimiento de voz puede constar de las siguientes etapas: análisis, extracción de características, modelado y pruebas (Khilari & P., 2015).

En la etapa de análisis, se tienen en cuenta la estructura física, la dimensión del tracto vocal, la fuente de excitación y la función de comportamiento del locutor, pues son características que inciden en la calidad del habla emitido y por lo tanto en el proceso de reconocimiento de voz (Khilari & P., 2015).

La etapa de extracción de características reduce el tamaño de los datos de la señal de voz antes de la clasificación o reconocimiento de patrones (Mon & Tun, 2015) y es la más importante, ya que juega un papel significativo para separar un hablante de otro (Khilari & P., 2015), de acuerdo con diferentes características individuales incrustadas en los enunciados (Gaikwad *et al.*, 2010).

Las características son los fonemas, es decir, los numerosos sonidos y diferentes sílabas que combinadas forman palabras y/u oraciones (Ramachandra & Sharma, 2018). Estas características deben: ser fácil de medir, no ser susceptible a la imitación, mostrar poca fluctuación de un entorno de habla a otro, ser estable en el tiempo, ocurrir frecuente y naturalmente en el habla (Khilari & P., 2015).

Por lo tanto, un sistema reconocedor de voz que transcriba audio a texto puede incluir: una base de datos de fonemas personalizados, una base de datos de palabras genéricas, una base de datos de palabras de dominio específico, una base de datos de contexto, una base de datos de conversaciones, uno o más modelos estadísticos (por ejemplo, un modelo de lenguaje, modelo acústico, un entrenador de modelos, etc.), varios parámetros del modelo, varias reglas, (Ramachandra & Sharma, 2018) etc.

La base de datos de fonemas personalizados puede ser creada para cada usuario (inclusive manejando variación de acento) durante una fase de entrenamiento, al solicitarle que pronuncie

oraciones predefinidas que cubran todos los fonemas para un determinado idioma (Ramachandra & Sharma, 2018).

Los modelos de lenguaje extraen el patrón de lenguaje y el orden de llegada de las secuencias de palabras o del habla y las clasifican en función de su probabilidad de ocurrencia en tiempo real. Estas estimaciones probabilísticas producen resultados más precisos probando con datos desconocidos o secuencia de voz (Shivakumar *et al.*, 2016).

Algunos modelos estadísticos usados con éxito son: Grapheme-to-Phoneme (G2P), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Cepstral Analysis and Mel-frequency cepstral (MFCCs), el enfoque basado en extracción basada en Kernal, la transformación Wavelet y sustracción espectral (Bhabad & Kharate, 2013), Vector Quantization (VQ), Artificial Neural Network (ANN), Dynamic Time Warping (DTW), GMM, Fuzzy logic o Hidden Markov Model (HMM) (Khilari & P., 2015; Mon & Tun, 2015). De ellos, el más exitoso, flexible y dominante es HMM (Khilari & P., 2015; Mon & Tun, 2015).

La base de datos de fonemas y palabras, en conjunto con el modelo estadístico y demás componentes del sistema, permitirán que de los fragmentos del audio la secuencia de fonemas sea codificada, decodificada y/o mapeada a su forma textual (Ramachandra & Sharma, 2018; Shivakumar *et al.*, 2016).

Cabe advertir, que, dependiendo de la saturación del entorno, los dispositivos de grabación (como el micrófono) pueden capturar mucho ruido no deseado de fondo. Esto provoca pérdida y/o malinterpretación de palabras que empeoran y disminuyen la tasa de precisión y fiabilidad del reconocimiento de voz (Mon & Tun, 2015; Ramachandra & Sharma, 2018). Pese a

que este problema puede superarse mediante el método de detección de punto final (Mon & Tun, 2015), incluso los algoritmos de conversión de voz a texto más sofisticados solo pueden alcanzar una precisión de hasta el 80 por ciento (Ramachandra & Sharma, 2018).

A modo de ejemplo, Carlini & Wagner (2018) representan el audio como un vector x de n dimensiones, donde cada elemento x_i es una señal con valor de 16 bits, muestreado a 16KHz. Luego utilizan el sistema de voz a texto DeepSpeech, el cual primero utiliza la transformación Mel-Frequency Cepstrum (MFC) para reducir la dimensionalidad de entrada, pues divide la forma de onda en 50 cuadros por segundo y asigna cada cuadro al dominio de frecuencia. Luego una Recurrent Neural Network (RNN) que usa LSTM mapea una forma de onda de audio a una secuencia de distribuciones de probabilidad sobre caracteres individuales, en lugar de sobre frases completas. Esta RNN es una función que mantiene un vector de estado s con $s_0 = 0$ y $(s_{i+1}, y^i) = f(s_i, x_i)$, donde la entrada x_i es un cuadro de entrada, y cada salida y^i es una distribución de probabilidad sobre cuál carácter fue pronunciado durante ese cuadro.

PLN en Texto

En los sistemas de PLN para extraer y procesar características gramaticales / lingüísticas de una oración de texto en lenguaje natural según las reglas que definen la gramática del idioma de destino, generalmente hay tres fases: análisis morfológico y léxico, análisis sintáctico y análisis semántico (Brun *et al.*, 2010; Kambhatla & Zitouni, 2013; Chopra *et al.*, 2013).

En el análisis léxico se realiza una tarea llamada tokenizacion donde la entrada lingüística se divide en párrafos, palabras, oraciones, expresiones e inclusive en signos de puntuación (Brun *et al.*, 2010; Chopra *et al.*, 2013; Riefer *et al.*, 2016). Cada uno de esos elementos generalmente se denomina token y se asocian con un léxico que contiene la correspondiente información

morfosintáctica, semántica y parte asociada del discurso (Brun *et al.*, 2010). Dicho análisis, identificación, descripción y asociación de tokens en la etapa léxica se conoce como análisis morfológico (Brun *et al.*, 2010; Mungi & Mustafi, 2016); Chopra *et al.*, 2013). La capa léxica generalmente opera en tokens individualmente, ocasionando a menudo una ambigüedad sustancial al final. Por ejemplo, el token "volar" en inglés podría representar un sustantivo indicativo de un insecto, o un verbo indicativo de movimiento aéreo (Brun *et al.*, 2010).

A diferencia de la fase anterior, en el análisis sintáctico mediante reglas gramaticales e información contextual se considera el valor o significado de cada token dentro de expresiones idiomáticas, frases, oraciones o cualquier otra combinación ordenada (Brun *et al.*, 2010; Kambhatla & Zitouni, 2013; Mungi & Mustafi, 2016; Chopra *et al.*, 2013; El Maarouf *et al.*, 2014). A veces se divide en un nivel de desambiguación donde se determina inequívocamente las partes del discurso de algunos tokens que eran ambiguos o no identificados a nivel léxico (Brun *et al.*, 2010). Por lo general en esta fase se realiza el etiquetamiento Part-Of-Speech (POS) y se generan los árboles de dependencia (Kambhatla & Zitouni, 2013; Kao & Poteet, 2007; Riefer *et al.*, 2016).

El etiquetamiento POS o etiquetamiento gramatical es el proceso donde se marca cada token con su correspondiente parte del discurso (sustantivo, verbo, adjetivo, etc.), basándose en su definición y contexto (Mungi & Mustafi, 2016; Riefer *et al.*, 2016; Shah & Jinwala, 2015). Existen muchos métodos diferentes para el etiquetado de POS que, en promedio alcanzan a marcar correctamente los tokens con una precisión cercana al 97% (Manning, 2011).

Los árboles de dependencia son representaciones donde cada nodo hoja es un token de la oración de entrada, el nodo raíz (S) es la oración completa y los nodos de nivel intermedio como

NP (frase nominal), VP (frase verbal), PP (frase preposicional), etc., entre los nodos raíz y hojas, están organizados jerárquicamente y conectados según las reglas de sintaxis de la gramática (Kambhatla & Zitouni, 2013; Mungi & Mustafi, 2016; Riefer *et al.*, 2016). Son más fáciles de entender que los clasificadores estadísticos (Mungi & Mustafi, 2016; Pons *et al.*, 2016) y han sido usados en los sistemas para vincular la fuente y el objetivo de la opinión (Brun, 2011). En general, las dependencias son tres: nombre de la relación, gobernador y dependiente (Mungi & Mustafi, 2016).

El análisis semántico es un proceso donde se utilizan léxicos u ontologías para asignar un significado al tipo de token usado e identificar los roles semánticos (Chopra *et al.*, 2013; Kambhatla & Zitouni, 2013; Pons *et al.*, 2016). La capacidad de reconocer y etiquetar estos roles o argumentos es una tarea clave para responder el “Quién”, "Cuándo", "Qué", "Dónde", "Por qué", etc., en aplicaciones de traducción automática, extracción de información, generación de lenguaje natural, respuesta a preguntas o resumen de texto (Kambhatla & Zitouni, 2013).

Como se pudo notar, la gramática es fundamental en las tres fases anteriores y su función se puede resumir en que define reglas que gobiernan la estructura de las palabras (morfología), reglas que gobiernan la estructura de las oraciones (sintaxis) y reglas que gobiernan el significado de las palabras y oraciones (semántica) (Kambhatla & Zitouni, 2013).

Cabe resaltar que, herramientas como Stanford CoreNLP, Natural Language ToolKit (NLTK), Apache OpenNLP, Gate NLP library o Spacy, facilitan la realización de los procesos anteriormente descritos en lenguajes de programación como PHP, Python o Java.

Extracción de Aspectos

Con esta técnica se detectan y extraen de manera detallada las entidades (con sus respectivos aspectos) referenciadas en cada una de las frases del texto, para asignarles la polaridad adecuada y luego clasificarlas según la fuerza de tal valor (Guzman & Miranda, 2016; Hu & Liu, 2004a; Marrese-Taylor & Matsuo, 2017; Quan & Ren, 2014; Ravi & Ravi, 2015; Schouten & Frasincar, 2016; Sun *et al.*, 2017; Titov & Mcdonald, 2008). Es decir, el objetivo es descubrir la entidad, el aspecto y el sentimiento (Sun *et al.*, 2017).

Un aspecto, característica, atributo u objetivo de la opinión es un concepto que usualmente corresponde a un tema arbitrario expresado y que es considerado representativo (Marrese-Taylor & Matsuo, 2017; Poria & Gelbukh, 2016). Igualmente, puede ser explícito cuando es mencionado directamente, e implícito cuando se presenta el caso contrario (Hu & Liu, 2004b). Además, cumple un rol importante en la minería de opinión al permitir a usuarios tomar decisiones (Caputo *et al.*, 2017) mientras comparan las características de diferentes elementos de la misma categoría (L. Qiu *et al.*, 2016), aunque no siempre sea igual a la preferencia e interés de los usuarios (Caputo *et al.*, 2017).

Al momento de extraer aspectos, es útil guiarse por las relaciones de dependencia presentes entre los términos de las características y las palabras de sentimiento (G. Qiu *et al.*, 2011). Por lo tanto, un sustantivo o una frase de sustantivos, si es acompañado por uno o más adjetivos, es considerado un aspecto (Eirinaki *et al.*, 2012) porque las personas tienden a usar adjetivos para expresar la inclinación de sus sentimientos hacia específicas características de un producto (Wei *et al.*, 2010).

Estado del Arte

En esta sección se mencionan investigaciones que están relacionadas con sistemas de generación automática de videos o diapositivas, ya sea utilizando el procesamiento de lenguaje natural u otras técnicas afines. Estas investigaciones, hasta el momento de la realización de este documento, han optado principalmente por analizar el sentido de la letra, el flujo de escenas de un conjunto de videos y el ritmo de la música. Algunas utilizan como fuente de las imágenes el internet, una base de datos de videos, y otras se deciden por usar fotos personales.

Una de esas investigaciones es llevada por Xu *et al.* (2008), donde se presenta un sistema algorítmico para la generación automática de diapositivas para una pieza musical, basándose en el ritmo, la letra y usando fotos personales. El principal reto con el que lidian es la nula anotación de las fotos, lo cual complica su procesamiento. Para solucionar dicho problema, introducen un algoritmo que infiere la relevancia de las fotos personales con respecto a la letra de la canción basándose en imágenes similares disponibles en la web.

Cai *et al.* (2007) ofrecen otro enfoque donde proponen una estrategia para seleccionar las palabras o frases más importantes de la letra, y usarlas para buscar imágenes en la web. La calidad de la imagen es evaluada con algunas técnicas como la detección de rostros y clasificación de paisajes. Además, mantienen las imágenes con un estilo similar, filtrándolas para que sus colores estén de acuerdo con las emociones que refleja una canción. Finalmente, con ayuda del ritmo y el tempo de la canción las imágenes son alineadas y convertidas en un video.

Funasawa *et al.* (2010) proponen un sistema que genera una diapositiva utilizando imágenes obtenidas de la web basándose en palabras importantes derivadas de las letras de

canciones. Las imágenes son filtradas basándose en el sentimiento que refleja por completo la letra de la canción.

Para producir el contenido visual de una canción, Wu, Xu, *et al.* (2012) obtienen un conjunto de imágenes desde el internet basándose en la letra. También utilizan un conjunto de reglas para filtrar las imágenes y dejar las que mejor coincidan con la música. Igualmente permiten que los usuarios den fotos personales para personalizar la generación de los videos.

Otra alternativa es dada por Lin *et al.* (2015), donde consideran la estructura de la expresión emocional para la generación de videos musicales basados en las emociones. Para esto aplican un modelo de curso temporal emocional acústico (o visual) (ETCM) que predice la emoción dominante en un segmento de la música o del video. Luego alinean la música y los segmentos del video siguiendo la coincidencia en la emoción.

DJ-MVP es el nombre dado por Fan *et al.* (2016) al productor automático de videos musicales que usa el método de síntesis concatenativo para generar mezclas audiovisuales al estilo del video musical creado sobre videos musicales existentes. Este sistema permite que los usuarios seleccionen videos musicales de una base de datos. Esos videos musicales son posteriormente separados en la parte visual y la auditiva. Luego, tanto la música como el video son segmentados de acuerdo con una función que detecta el ritmo de las melodías. Cuando el usuario otorgue la canción de la cual desea que se genere un video, el sistema la segmentará también de acuerdo con el ritmo. Después, los segmentos de video son organizados de acuerdo con una clasificación de similitud de audio, clasificación de similitud de video y otras reglas heurísticas. Finalmente, la pista de audio que el usuario otorgó es presentada junto a los segmentos de los videos tratados como un nuevo video.

Yoon *et al.* (2009) presentan un método de conservación de información para la generación de videos musicales que utiliza la segmentación en múltiples niveles del video y el audio. Este sistema se divide en un módulo de análisis y un módulo de coincidencia. En el módulo de análisis se segmenta la música y el video usando la información de flujo, para luego analizar la velocidad y brillo de cada segmento. En el módulo de coincidencia se usan esas dos últimas características para asignar los segmentos coincidentes a una música. Si este módulo no encuentra una coincidencia satisfactoria, se aplica una segmentación multinivel que subdivide aún más el segmento y repite el cálculo hasta encontrar una mejor coincidencia.

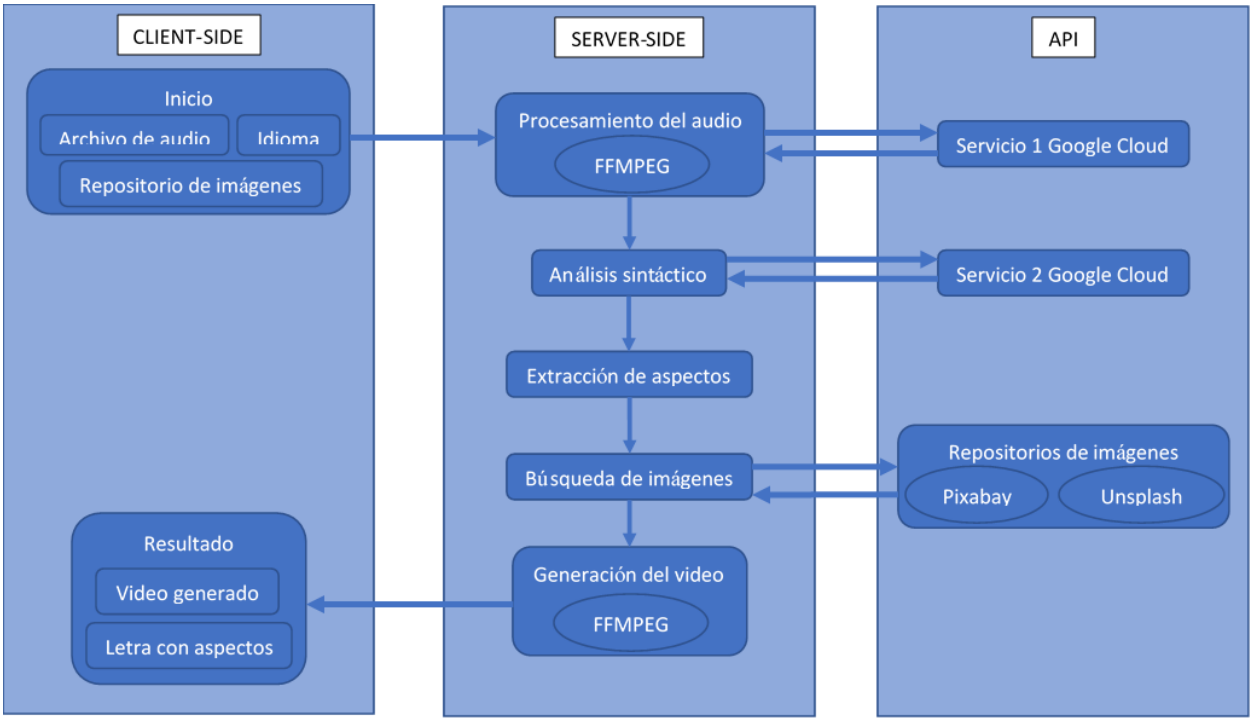
TextAlive Online es una aplicación web presentada por Kato *et al.* (2015) que le permite a los usuarios crear videos con tipografía cinética, es decir, muestra animadas las letras de las canciones de manera interactiva y sincronizada con el audio.

Metodología

El flujo general de funcionamiento del sistema presentado en este documento consta de las siguientes etapas: interacción del usuario, procesamiento del audio, extracción de aspectos, búsqueda de imágenes y creación del video.

Figura 1

Arquitectura General del Sistema Propuesto. Fuente: Elaboración propia



Interacción del Usuario

Esta fase es donde el usuario le otorga al sistema la información necesaria para generar el video. Para eso contará con una aplicación web que le mostrará inicialmente un formulario. Allí podrá utilizar un campo para subir el archivo de audio, el cual solo admitirá archivos con extensión MP3, WAV o FLAC. También tendrá una lista desplegable para seleccionar entre Unsplash o Pixabay como repositorio de imágenes y otra lista desplegable para indicar el idioma hablado en el audio, el cual solo podrá ser inglés o español. Finalmente, toda la información anteriormente descrita será enviada al servidor usando el protocolo POST.

Procesamiento del Audio

Tras recibir el archivo de audio, en esta fase primero se verifica el códec que usa, con la intención de cumplir con todas las condiciones que exige el servicio de Google Cloud. Si el sistema detecta que el códec del audio es diferente a FLAC, automáticamente realiza su cambio, pues códecs como el MP3 son con pérdida que puede reducir la exactitud de la transcripción. Para cumplir con el cambio de códec se utiliza FFMPEG, una colección de software libre para convertir, grabar y transmitir audio y video. Luego se envía la debida solicitud al servicio de Google Cloud y este retorna una respuesta que el sistema itera y concatena hasta conseguir la transcripción completa con signos de puntuación y sus debidas marcas de tiempo, es decir, el tiempo inicial y final de cada palabra hablada reconocida en el audio.

Extracción de Aspectos

Esta fase inicia enviando la transcripción al servicio de Google Cloud que realiza el análisis sintáctico. La respuesta del servicio será el texto con su debido etiquetamiento POS y con los árboles o relaciones de dependencia. Para que una o más palabras sean consideradas como aspectos, deben cumplir con algunas de las siguientes reglas o patrones que se basarán tanto en el etiquetamiento POS como en la relación de dependencia:

- Ser uno o más sustantivos que son modificados por un adjetivo dentro de una oración.
- Ser uno o más sustantivos que no son modificados por un adjetivo dentro de una oración.

Así por ejemplo en la primera estrofa de la canción La camisa negra de Juanes que dice:

Tengo la camisa negra.

Hoy mi amor está de luto.

Hoy tengo en el alma una pena

y es por culpa de tu embrujo.

Hoy sé que tú ya no me quieres

y eso es lo que más me hiere.

Que tengo la camisa negra

y una pena que me duele.

el algoritmo debería extraer como aspectos las palabras: camisa negra, amor, luto, alma, pena, culpa, embrujo, camisa negra, pena.

Búsqueda de Imágenes

Dependiendo del repositorio de imágenes elegido por el usuario, el sistema en esta fase construye las consultas adecuadas para cada uno de los aspectos extraídos en la etapa anterior. La respuesta a estas peticiones contendrá un conjunto de máximo diez imágenes por aspecto. De ese grupo el sistema elegirá la más popular dentro del repositorio.

Creación del Video

En esta fase se utiliza FFMPEG para unir en un video el audio aportado por el usuario y las imágenes conseguidas del repositorio seleccionado. En este proceso las marcas de tiempo obtenidas en la etapa de análisis del audio son usadas para sincronizar cada imagen con el contenido del audio de manera adecuada. Esta sincronización se realiza tomando como tiempo de inicio el lapso donde la primera palabra de la oración empieza a pronunciarse. En ese caso la duración de la primera imagen será marcada por el tiempo final de pronunciación del primer aspecto. Para la imagen del segundo aspecto dentro de la misma oración, su duración será dada tomando como tiempo de inicio el tiempo final de la imagen inmediatamente anterior y como tiempo final el lapso final de pronunciación de su respectivo aspecto o el lapso de pronunciación de la última palabra de la oración, si es que no hay más aspectos en ella. Por lo tanto, si se

supone que cada palabra es pronunciada en un segundo, para los dos primeros versos de la primera estrofa de La camisa negra de Juanes la duración de cada imagen sería:

Tengo la camisa negra.

Hoy mi amor está de luto.

- Camisa negra => tiempo inicio = 0, tiempo final = 4, duración = 4 segundos.
- Amor => tiempo inicio = 4, tiempo final = 7, duración = 3 segundos.
- Luto => tiempo inicio = 7, tiempo final = 10, duración = 3 segundos.

Finalmente, se mostrará en una nueva interfaz el video resultante y la letra con los aspectos extraídos resaltados en negrilla.

Análisis de Pruebas y Resultados

En esta sección se muestra el cómo se probó y se evaluó el rendimiento del sistema en las fases de transcripción de audio, extracción de aspectos y búsqueda de imágenes, usando las siguientes canciones: La camisa negra de Juanes, It's my life de Bon Jovi, Hijo de la luna de Mecano. Para evaluar la efectividad de las etapas de transcripción del audio y de extracción de aspectos con una canción, se decidió usar *Precision* y *Recall*. Cuando se trató con las tres canciones se adoptó el macro y micro promedio utilizados por El Maarouf *et al.* (2014) y Wei *et al.* (2010).

$$Precision_i = \frac{TP}{(TP + FP)} = \frac{c_i}{e_i}$$

$$Recall_i = \frac{TP}{(TP + FN)} = \frac{c_i}{m_i}$$

$$Micro\ promedio\ de\ Precision = \sum_{i=1}^{n}\left(\frac{e_i * Precision_i}{\sum_{j=1}^{n} e_j}\right) = \sum_{i=1}^{n}\left(\frac{e_i * \frac{c_i}{e_i}}{\sum_{j=1}^{n} e_j}\right) = \frac{\sum_{i=1}^{n} c_i}{\sum_{j=1}^{n} e_j}$$

$$Micro\ promedio\ de\ Recall = \sum_{i=1}^{n}\left(\frac{m_i * Recall_i}{\sum_{j=1}^{n} m_j}\right) = \sum_{i=1}^{n}\left(\frac{m_i * \frac{c_i}{m_i}}{\sum_{j=1}^{n} m_j}\right) = \frac{\sum_{i=1}^{n} c_i}{\sum_{j=1}^{n} m_j}$$

$$Macro\ promedio\ de\ Precision = \frac{\sum_{i=1}^{n} Precision_i}{n}$$

$$Macro\ promedio\ de\ Recall = \frac{\sum_{i=1}^{n} Recall_i}{n}$$

Donde:

- TP = True Positive
- FP = False Positive
- FN = False Negative
- $c_i =$ En la fase de transcripción del audio es el número de palabras correctamente identificadas por el servicio de Google Cloud. En la fase de extracción de aspectos es el número de aspectos extraídos correctamente mediante la técnica bajo examen.
- $e_i =$ En la fase de transcripción del audio es la cantidad de palabras identificadas por el servicio de Google Cloud. En la fase de extracción de aspectos es la cantidad de aspectos extraídos mediante la técnica bajo examen.
- $m_i =$ En la fase de transcripción del audio es la cantidad de palabras reales en la letra de la canción. En la fase de extracción de aspectos es el número de aspectos reconocidos manualmente en la letra de la canción.
- $n =$ Cantidad de canciones usadas.

Transcripción del Audio

Las pruebas se realizaron siguiendo la metodología descrita anteriormente para las tres canciones. Las siguientes son imágenes que muestran la comparación entre el texto transcrito del audio y el contenido real de la letra de la canción:

Figura 2

Texto Transcrito de la Canción La Camisa Negra

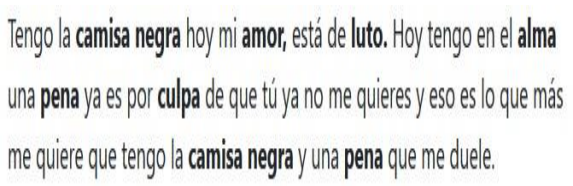


Figura 3

Texto Real de la Canción La Camisa Negra

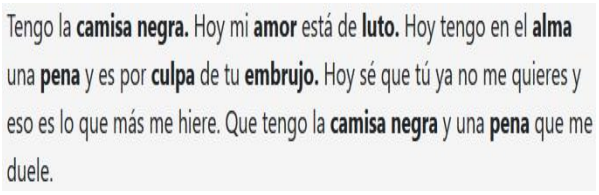


Figura 4

Texto Transcrito de la Canción Hijo de la Luna

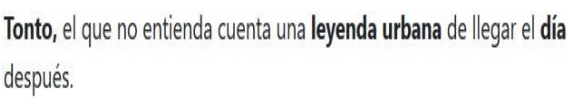


Figura 5

Texto Real de la Canción Hijo de la Luna

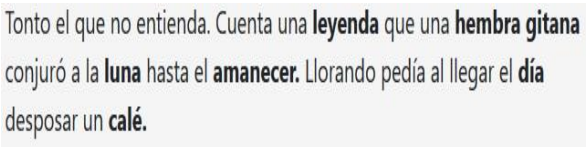
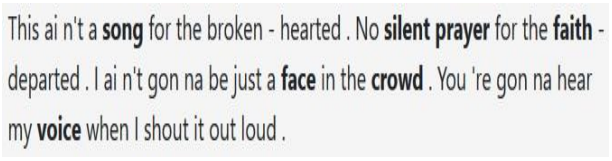


Figura 6

Texto Transcrito de la Canción It's my Life

Figura 7

Texto Real de la Canción It's my Life



Se observa que la canción con la que mejor se desempeñó fue con La camisa negra de Juanes, lo cual se pudo constatar calculando sus respectivos *Precision* y *Recall*.

Tabla 1

Cálculo de Medidas de Rendimiento en la Fase de Transcripción del Audio

Medida	Canciones		
	La camisa negra	Hijo de la luna	It's my life
Precision	$45/47 = 0.96$	$11/14 = 0.78$	$0/6 = 0$
Recall	$45/51 = 0.88$	$11/28 = 0.39$	$0/36 = 0$
Micro p. Precision	$(45 + 11 + 0)/(47 + 14 + 6) = 56/67 = 0.83$		
Micro p. Recall	$(45 + 11 + 0)/(51 + 28 + 36) = 56/115 = 0.49$		
Macro p. Precision	$(0.96 + 0.78 + 0)/3 = 0.58$		
Macro p. Recall	$(0.88 + 0.39 + 0)/3 = 0.42$		

Con los promedios obtenidos se puede notar que hubo un regular desempeño al transcribir el discurso inmerso en el audio. Esto se debió al tipo de música que acompañaba la voz humana. Canciones como Hijo de la luna y It's my life tienen más arreglos acústicos que acompañan la voz del cantante si se compara con La camisa negra. Esto además de hacer más agradable la canción, por desgracia también hace de ruido que limita la eficacia del sistema en esta etapa. Debido a lo anterior, para trabajos futuros se considera necesario incluir un paso intermedio entre esta etapa y la de extracción de aspectos, que le permita al usuario revisar la transcripción del audio. Igualmente se podría recomendar al usuario usar el sistema con archivos de audio que contengan poca o nula música.

Extracción de Aspectos

En esta etapa se decidió trabajar tanto con la transcripción obtenida del audio como con el texto real para verificar el desempeño del módulo extractor de aspectos perteneciente al sistema. Los aspectos extraídos de los datos anteriormente mencionados se pueden observar en la Figura 2, Figura 3, Figura 4, Figura 5, Figura 6 y Figura 7 en negrilla.

Tabla 2

Cálculo de Medidas de Rendimiento en la Fase de Extracción de Aspectos del Texto Transcrito

Medida	Canciones		
	La camisa negra	Hijo de la luna	It's my life
Precision	$8/8 = 1$	$3/3 = 1$	$2/2 = 1$
Recall	$8/8 = 1$	$3/3 = 1$	$2/2 = 1$
Micro p. Precision	$(8 + 3 + 2)/(8 + 3 + 2) = 13/13 = 1$		
Micro p. Recall	$(8 + 3 + 2)/(8 + 3 + 2) = 13/13 = 1$		
Macro p. Precision	$(1 + 1 + 1)/3 = 1$		
Macro p. Recall	$(1 + 1 + 1)/3 = 1$		

Tabla 3

Cálculo de Medidas de Rendimiento en la Fase de Extracción de Aspectos del Texto Real

Medida	Canciones		
	La camisa negra	Hijo de la luna	It's my life
Precision	$9/9 = 1$	$6/6 = 1$	$5/6 = 0.83$
Recall	$9/9 = 1$	$6/6 = 1$	$5/7 = 0.71$
Micro p. Precision	$(9 + 6 + 5)/(9 + 6 + 6) = 20/21 = 0.95$		
Micro p. Recall	$(9 + 6 + 5)/(9 + 6 + 7) = 20/22 = 0.91$		
Macro p. Precision	$(1 + 1 + 0.83)/3 = 0.94$		
Macro p. Recall	$(1 + 1 + 0.71)/3 = 0.90$		

Como se puede observar, los resultados en esta etapa son bastantes satisfactorios. Sin embargo, hay tres detalles que ameritan ser comentados.

El primero es sobre la palabra "Tonto" presente en la transcripción y en el contenido real de Hijo de la luna, que en el primer escenario es tomado como aspecto y en el segundo no. Esto no es por un error en el algoritmo que extrae los aspectos, sino que se debe a la estructura gramatical presente en las oraciones. En la transcripción esta palabra es seguida por una coma que le cambia el sentido, pasando de ser un adjetivo a ser un sustantivo.

El segundo detalle es sobre el desempeño con contenido en el idioma inglés manejado en la canción It's my life. Aunque el resultado es aceptable, se debe reconocer que el algoritmo falló en detectar y extraer dos aspectos presentes en la letra real. Estos aspectos fueron "broken-

hearted" y "faith-departed", los cuales son expresiones que podrían traducirse como "con el corazón roto" y "difunto", respectivamente. Esto hace percibir la necesidad de manejar reglas específicas para diferentes idiomas de acuerdo con la gramática correspondiente.

El tercer asunto está presente en la letra real de la canción La camisa negra. En este texto se puede observar la oración "Hoy sé que tú ya no me quieres y eso es lo que más me hiere" donde no hay palabras que cumplan con una de las dos reglas para ser aspecto. Esto al momento de generar el video podría representar por lo menos 2 segundos donde la imagen mostrada no concuerde con lo hablado en el audio. Por tal razón, para generar un video con contenido visual más adecuado se decidió adicionar dos nuevas reglas:

- Ser un verbo en una oración sin sustantivos.
- Ser un verbo en una oración sin sustantivos y modificado por una partícula de negación.

Con estas dos nuevas reglas el sistema extraerá de la anterior oración las palabras "sé", "no quieres" y "hiere" como aspectos.

Además de los anteriores asuntos, se reconoce que el algoritmo descrito en este documento es incapaz de extraer aspectos implícitos. También es un algoritmo que no detecta el sentido figurado de las palabras, es decir, no diferencia lo que es una metáfora, eufemismo, analogía o cualquier otro recurso retórico. Por lo tanto, para la frase "Ella tiene ojos de cielo" el sistema en vez de seleccionar como aspecto "hermosos ojos", buscará imágenes relacionadas con "ojos" y "cielo".

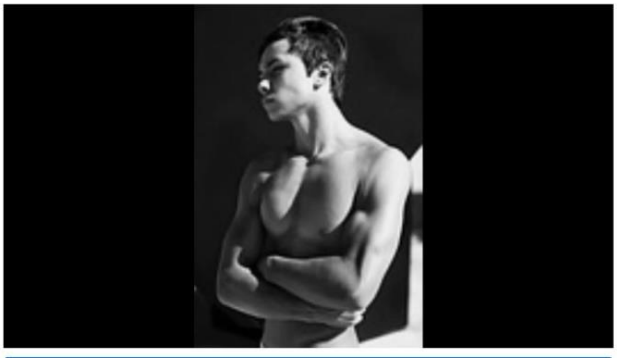
Búsqueda de Imágenes

Aunque las pruebas se realizaron con las tres canciones mencionadas, para el análisis de los resultados se hará mayor enfoque en la canción La camisa negra, por ser la de mejor

rendimiento. Para esta canción se seleccionó Pixabay como repositorio y se obtuvieron las siguientes imágenes:

Figura 8

Imagen Obtenida para el Primer Aspecto Extraído



Tengo la **camisa negra** hoy mi **amor,** está de **luto.** Hoy tengo en el **alma** una **pena** ya es por **culpa** de que tú ya no me quieres y eso es lo que más me quiere que tengo la **camisa negra** y una **pena** que me duele.

Figura 9

Imagen Obtenida para el Segundo Aspecto Extraído

Tengo la **camisa negra** hoy mi **amor,** está de **luto.** Hoy tengo en el **alma** una **pena** ya es por **culpa** de que tú ya no me quieres y eso es lo que más me quiere que tengo la **camisa negra** y una **pena** que me duele.

Figura 10

Imagen Obtenida para el Tercer Aspecto Extraído

Tengo la **camisa negra** hoy mi **amor,** está de **luto.** Hoy tengo en el **alma** una **pena** ya es por **culpa** de que tú ya no me quieres y eso es lo que más me quiere que tengo la **camisa negra** y una **pena** que me duele.

Figura 11

Imagen Obtenida para el Cuarto Aspecto Extraído

Tengo la **camisa negra** hoy mi **amor,** está de **luto.** Hoy tengo en el **alma** una **pena** ya es por **culpa** de que tú ya no me quieres y eso es lo que más me quiere que tengo la **camisa negra** y una **pena** que me duele.

Figura 12

Imagen Obtenida para el Quinto Aspecto Extraído

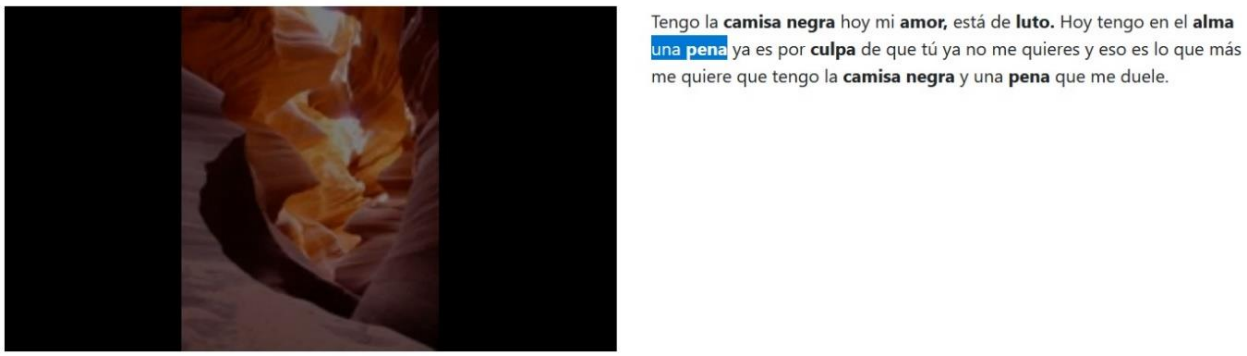


Figura 13

Imagen Obtenida para el Sexto Aspecto Extraído

Figura 14

Imagen Obtenida para el Séptimo Aspecto Extraído

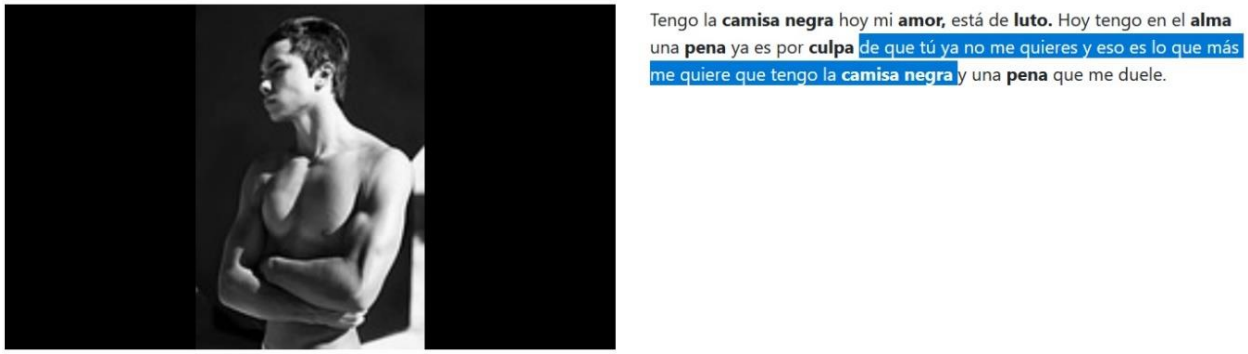


Figura 15

Imagen Obtenida para el Octavo Aspecto Extraído

Lo más notable de este conjunto de imágenes es que hay dos que no concuerdan mucho con el aspecto relacionado. Una de esas imágenes es la del aspecto “camisa negra”, de la cual se esperaba que mostrara una prenda negra de vestir el torso, pero no fue así. La otra imagen es la del aspecto “pena”, que, en vez de mostrar una persona triste, muestra una formación geológica, lo cual queda fuera del contexto de la canción. Esto hace ver la necesidad para trabajos futuros de usar un algoritmo que verifique objetos en las imágenes, lo cual mejoraría la precisión y concordancia imagen-aspecto.

Otra situación resaltable fue el no retorno de imágenes para un aspecto. Esto pudo deberse a la insuficiente variedad de imágenes presentes en los repositorios elegidos para este sistema, lo cual se solucionaría en futuras versiones eligiendo repositorios con mayor variedad o buscando imágenes por toda la internet. Por ejemplo, el aspecto “hembra gitana” presente en la canción Hijo de la luna no retorna ningún resultado usando el repositorio Pixabay. En ese caso el sistema separa el aspecto en las dos palabras que lo conforman y buscó dos imágenes. Sin embargo, cuando el sistema definitivamente no obtiene ningún resultado, opta por omitir el respectivo aspecto y darle su tiempo de duración al siguiente.

Conclusiones

Este documento muestra un sistema para la generación de video mediante el emparejamiento de imágenes tras PLN usando servicios de Google Cloud y extracción de aspectos en canciones.

Los resultados experimentales con tres canciones, dos en español y una en inglés, mostraron que en la transcripción del audio la métrica *Precision* tuvo un micro promedio de 0.83 y un macro promedio de 0.58, mientras que para *Recall* se obtuvo un micro promedio de 0.49 y un macro promedio de 0.42. Estos resultados hacen ver la necesidad de limitar la subida solo de audios con poco ruido de fondo, o que para futuras versiones se le habilite al usuario supervisar y mejorar la transcripción del audio antes de realizar la extracción de aspectos, y/o añadir un algoritmo que separe el canto de la melodía.

Para la extracción de aspectos el rendimiento fue mucho mejor con promedios mayores al 0.90. Sin embargo, se deben manejar diferentes reglas para las diferentes gramáticas de cada uno de los idiomas que se quieran habilitar.

En cuanto a la calidad de las imágenes seleccionadas por el sistema, se notó que es un factor que se puede mejorar incluyendo un módulo de detección de objetos que certifique que la pareja entidad-aspecto esté realmente presente en la imagen.

Bibliografía

Allawadi, N. (2012). *Speech-to-Text System for Phonebook Automation*. Thapar Institute of Engineering & Technology. Retrieved from http://tudr.thapar.edu:8080/jspui/handle/10266/1748

Bhabad, S. S., & Kharate, G. K. (2013). An Overview of Technical Progress in Speech Recognition. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(3), 488–497.

Brun, C. (2011). Detecting opinions using deep syntactic analysis. In *International Conference Recent Advances in Natural Language Processing, RANLP* (pp. 392–398).

Brun, C., Hagège, C., & Roux, C. (2010). U.S. Patent No. 7,822,597. Washington, DC: U.S. Patent and Trademark Office.

Cai, R., Zhang, L., Jing, F., Lai, W., & Ma, W.-Y. (2007). Automated Music Video Generation using WEB Image Resource. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07* (pp. II-737-II–740). IEEE. https://doi.org/10.1109/ICASSP.2007.366341

Caputo, A., Basile, P., de Gemmis, M., Lops, P., Semeraro, G., & Rossiello, G. (2017). SABRE: A Sentiment Aspect-Based Retrieval Engine. In *Information Filtering and Retrieval* (pp. 63–78). Springer, Cham. https://doi.org/10.1007/978-3-319-46135-9_4

Carlini, N., & Wagner, D. (2018). Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In *2018 IEEE Security and Privacy Workshops (SPW)* (pp. 1–7). IEEE. https://doi.org/10.1109/SPW.2018.00009

Chopra, A., Prashar, A., & Sain, C. (2013). Natural Language Processing. *INTERNATIONAL JOURNAL OF TECHNOLOGY ENHANCEMENTS AND EMERGING ENGINEERING RESEARCH*, *1*(4), 131–134. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.407.6907

Das, S. (2012). Speech Recognition Technique: A Review. *International Journal of Engineering Research and Applications*, *2*(3), 2071–2087.

Eirinaki, M., Pisal, S., & Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, *78*(4), 1175–1184. https://doi.org/10.1016/J.JCSS.2011.10.007

El Maarouf, I., Baisa, V., Bradbury, J., & Hanks, P. (2014). Disambiguating verbs by collocation: Corpus lexicography meets natural language processing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014* (pp. 1001–1006).

Fan, J., Li, W., Bizzocchi, J., Bizzocchi, J., & Pasquier, P. (2016). DJ-MVP. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology - ACE2016* (pp. 1–8). New York, New York, USA: ACM Press. https://doi.org/10.1145/3001773.3001782

Foote, J., Cooper, M., & Girgensohn, A. (2002). Creating music videos using automatic media

analysis. In *Proceedings of the tenth ACM international conference on Multimedia - MULTIMEDIA '02* (p. 553). New York, New York, USA: ACM Press. https://doi.org/10.1145/641007.641119

Funasawa, S., Ishizaki, H., Hoashi, K., Takishima, Y., & Katto, J. (2010). Automated music slideshow generation using web images based on lyrics. In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010* (pp. 63–68).

Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A Review on Speech Recognition Technique. *International Journal of Computer Applications*, *10*(3), 16–24. https://doi.org/10.5120/1462-1976

Guzman, J., & Miranda, C. H. (2016). A review of Sentiment Analysis in Spanish. *TECCIENCIA*, *12*(22), 35–48. https://doi.org/10.18180/tecciencia.2017.22.5

Hu, M., & Liu, B. (2004a). Mining and summarizing customer reviews. In *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Hu, M., & Liu, B. (2004b). Mining opinion features in customer reviews. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 755–760).

Iwamiya, S. (1992). The interaction between auditory and visual processing when listening to music via audio-visual media. *The Journal of the Acoustical Society of Japan*. https://doi.org/10.20697/jasj.48.3_146

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(5), 559–575.

https://doi.org/10.1017/S0140525X08005293

Kambhatla, N., & Zitouni, I. (2013). U.S. Patent No. 8,527,262. Washington, DC: U.S. Patent and Trademark Office.

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. Springer. Retrieved from https://books.google.com.co/books?id=CVtxFWbKT7wC&dq=%22Natural+language+processing+and+text+mining%22&lr=&hl=es&source=gbs_navlinks_s

Kato, J., Nakano, T., & Goto, M. (2015). TextAlive: Integrated Design Environment for Kinetic Typography. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (ACM CHI 2015)* (pp. 3403–3412).

Khilari, P., & P., B. V. (2015). A Review on Speech To Text Conversion Methods. *International Journal of Advanced Research in Computer Engineering & Technology*, *4*(7), 3067–3072. Retrieved from https://pdfs.semanticscholar.org/0924/9025b64a06cd3fda5e830b4c216dd9eb1c57.pdf

Lin, J.-C., Wei, W.-L., & Wang, H.-M. (2015). EMV-matchmaker. In *Proceedings of the 23rd ACM international conference on Multimedia - MM '15* (pp. 899–902). New York, New York, USA: ACM Press. https://doi.org/10.1145/2733373.2806359

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-19400-9_14

Marrese-Taylor, E., & Matsuo, Y. (2017). Replication issues in syntax-based aspect extraction for opinion mining. Retrieved from http://arxiv.org/abs/1701.01565

Mon, S. M., & Tun, H. M. (2015). Speech-To-Text Conversion STT System Using Hidden Markov Model HMM. *International Journal of Scientific & Technology Research*, *4*(6), 349–352. Retrieved from https://www.ijstr.org/final-print/june2015/Speech-to-text-Conversion-stt-System-Using-Hidden-Markov-Model-hmm.pdf

Mungi, A., & Mustafi, J. (2016). U.S. Patent No. 9,495,355. Washington, DC: U.S. Patent and Trademark Office.

Pons, E., Braun, L. M. M., Hunink, M. G. M., & Kors, J. A. (2016). Natural language processing in radiology: A systematic review. *Radiology*. https://doi.org/10.1148/radiol.16142770

Poria, S., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, *108*, 42–49. https://doi.org/10.1016/J.KNOSYS.2016.06.009

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, *37*(1), 9–27. https://doi.org/10.1162/coli_a_00034

Qiu, L., Gao, S., Cheng, W., & Guo, J. (2016). Aspect-based latent factor model by integrating ratings and reviews for recommender system. *Knowledge-Based Systems*, *110*, 233–243. https://doi.org/10.1016/J.KNOSYS.2016.07.033

Quan, C., & Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*, *272*, 16–28. https://doi.org/10.1016/j.ins.2014.02.063

Ramachandra, M., & Sharma, P. (2018). U.S. Patent No. 9,940,932. Washington, DC: U.S. Patent and Trademark Office.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*(November), 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

Riefer, M., Ternis, S. F., & Thaler, T. (2016). Mining Process Models from Natural Language Text: A State-of-the-Art Analysis. *Tagungsband Der Multikonferenz Wirtschaftsinformatik. Multikonferenz Wirtschaftsinformatik (MKWI-16), March 9-11, Illmenau, Germany*.

Schouten, K., & Frasincar, F. (2016). Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 813–830. https://doi.org/10.1109/TKDE.2015.2485209

Shah, U. S., & Jinwala, D. C. (2015). Resolving Ambiguities in Natural Language Software Requirements. *ACM SIGSOFT Software Engineering Notes*, *40*(5), 1–7. https://doi.org/10.1145/2815021.2815032

Shivakumar, K. ., Aravind, K. G., Anoop, T. V., & Gupta, D. (2016). Kannada speech to text conversion using CMU Sphinx. In *2016 International Conference on Inventive Computation Technologies (ICICT)* (pp. 1–6). Coimbatore, India: IEEE. https://doi.org/10.1109/INVENTIVE.2016.7830119

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, *36*, 10–25. https://doi.org/10.1016/J.INFFUS.2016.10.004

Titov, I., & Mcdonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*.

Wei, C.-P., Chen, Y.-M., Yang, C.-S., & Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and E-Business Management*, *8*(2), 149–167. https://doi.org/10.1007/s10257-009-0113-9

Wu, X., Qiao, Y., Wang, X., & Tang, X. (2012). Cross matching of music and image. In *Proceedings of the 20th ACM international conference on Multimedia - MM '12* (p. 837). New York, New York, USA: ACM Press. https://doi.org/10.1145/2393347.2396325

Wu, X., Xu, B., Qiao, Y., & Tang, X. (2012). Automatic music video generation. In *Proceedings of the 20th ACM international conference on Multimedia - MM '12* (p. 1381). New York, New York, USA: ACM Press. https://doi.org/10.1145/2393347.2396495

Xu, S., Jin, T., & Lau, F. C. M. (2008). Automatic Generation of Music Slide Show Using Personal Photos. In *2008 Tenth IEEE International Symposium on Multimedia* (pp. 214–219). IEEE. https://doi.org/10.1109/ISM.2008.39

Yoon, J.-C., Lee, I.-K., & Byun, S. (2009). Automated music video generation using multi-level feature-based segmentation. *Multimedia Tools and Applications*, *41*(2), 197–214. https://doi.org/10.1007/s11042-008-0225-0