UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

MEJORAS DE UN SISTEMA DE BÚSQUEDAS EN VOZ Y APLICACIÓN A DETECCIÓN DE MENCIONES EN MEDIOS DE COMUNICACIÓN

María Pilar Fernández Gallego
Tutor: Doroteo Torre Toledano

Mayo 2016

MEJORAS DE UN SISTEMA DE BÚSQUEDAS EN VOZ Y APLICACIÓN A DETECCIÓN DE MENCIONES EN MEDIOS DE COMUNICACIÓN

AUTOR: María Pilar Fernández Gallego
TUTOR: Doroteo Torre Toledano

Área de Tratamiento de Voz y Señales (ATVS)
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Mayo de 2016

Resumen (castellano)

Las menciones publicitarias son contenidos publicitarios no pregrabados que habitualmente dicen los locutores de radio o TV para promocionar un producto o empresa. La dificultad de la detección de menciones publicitarias consiste en que el audio no se repite igual cada vez, como ocurre con los anuncios publicitarios convencionales, donde se pueden emplear técnicas más efectivas como las de audio fingerprinting. Este Trabajo Fin de Grado propone la utilización de un sistema de búsqueda de palabras clave en castellano para la detección de menciones publicitarias.

En principio el TFG tenía como objetivo mejorar un sistema de búsqueda de palabras claves en castellano para aplicarlo a la detección de menciones publicitarias. Finalmente se ha construido un sistema nuevo prácticamente desde cero. Para ello ha sido necesario en primer lugar entrenar y evaluar un nuevo reconocedor de voz en español empleando la herramienta Kaldi y las bases de datos Fisher Spanish y Callhome Spanish. Con este proceso se ha conseguido reducir la tasa de error de palabra del 49.88% que se obtenía con el anterior reconocedor en español del grupo ATVS al 41.10% sobre voz conversacional telefónica.

Para la evaluación de detección de menciones a través de palabras clave se ha creado, también como parte de este TFG y en colaboración con otros estudiantes de TFG del grupo, una base de datos en castellano, que hemos denominado ATVS-Radio, que contiene unas 300 horas de audio, de las cuales 25 horas han sido etiquetadas con varias informaciones. En particular para este trabajo se han etiquetado las 62 menciones publicitarias que aparecían. Para la detección de menciones se ha modificado el léxico del reconocedor para incluir 51 palabras clave a detectar en las menciones, y se ha aplicado el reconocedor a todas las menciones publicitarias, consiguiendo detectar cerca de un 74% de las mismas. Este resultado todavía podría mejorase sustancialmente porque es posible realizar una mejor adaptación del reconocedor a la tarea, en particular adaptando el modelo de lenguaje que no ha sido modificado para la detección de palabras clave.

Abstract (English)

The advertising mentions are contents that are not prerecorded, usually are said by radio or TV broadcasters to publicize a product or a company. The difficulty on detecting advertising mentions is that the audio is not exactly repeated every time, as happens with conventional advertising where more efficient techniques such as audio fingerprinting can be used. This Bachelor Thesis proposes the use of a keyword search system in Spanish for the detection of advertising mentions.

At first the Bachelor Thesis was aimed to improve a keyword search in Spanish and apply it to the detection of advertising mentions. Finally, it has been built a new system from scratch. For that, firstly, it has been necessary to train and evaluate a new speech recognizer in Spanish using the Kaldi tool and databases Fisher Spanish and Callhome Spanish. This process has reduced the word error rate from 49.88% that was obtained with the previous ATVS recognizer in Spanish to 41.10%, both on conversational telephone speech.

For the evaluation of detection of mentions it has been created, also being part of the Bachelor Thesis, and in collaboration with other Bachelor Thesis students from the group, a database in Spanish which has been called ATVS-Radio, containing 300 hours of audio, 25 of which have been tagged with different informations. In particular, to do this work we have labeled the 62 advertising mentions appearing in the 25 hours. For the detection of mentions, the lexicon of the recognizer has been modified to include 51 keywords and the recognizer has been applied to all advertising mentions, achieving a detection rate of about 74%. This result could improve even more because it's possible to do a better adjustment of the recognizer to the task, specifically by adapting the language model that hasn't been modified in this work for keyword detection.

Palabras clave (castellano)

Reconocedor de voz, detección de menciones, detección de palabras clave, Fisher Spanish, Callhome Spanish, ATVS-Radio.

Keywords (inglés)

Speech recognizer, mention detection, keyword detection, Fisher Spanish, Callhome Spanish, ATVS-Radio.

.

Agradecimientos

Ha llegado el momento de cerrar una de las etapas más importantes de mi vida, a pesar de haber sido difícil y que no todo han sido buenos momentos, todo el esfuerzo ha valido la pena. Por ello no puedo olvidar a la gente que ha estado a mi lado en todo momento.

En primer lugar quería dar las gracias a mi tutor Doroteo Torre Toledano por darme la oportunidad de realizar este proyecto, además de su ayuda y paciencia durante estos meses.
También se lo quería agradecer al resto del grupo de ATVS por su ayuda y amabilidad desde el primer día.

Quería dar las gracias de una forma especial a mis padres y mi hermana que desde siempre han estado apoyándome y dándome ánimos para seguir adelante, sin vosotros no hubiese conseguido nada. Gracias a vosotros soy lo que soy.

A los amigos que he conocido durante estos cuatro años, por todos los momentos vividos. Especialmente a Alberto, amigo y compañero de viaje, muchas gracias por haber estado ahí siempre, juntos el camino ha sido más fácil.

Gracias a todos, sois increíbles, me siento orgullosa de haber podido compartir una etapa tan especial de mi vida a vuestro lado.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 Introducción

1.1 Motivación

En la sociedad contemporánea se otorga una gran relevancia a los contenidos multimedia siendo estos cada vez más abundantes. Se pueden observar algunos ejemplos de ello en los medios de comunicación o en Internet.

A pesar de que se ha conseguido un alto grado de madurez, precisión y rapidez en las búsquedas en contenido textuales, no ha ocurrido lo mismo en la búsqueda de contenido multimedia (audio y video fundamentalmente) ya que actualmente se encuentran todavía bastantes limitaciones en las búsquedas en estos tipos de contenido.

Para hacer frente a esta situación, en los últimos años se ha venido haciendo un esfuerzo considerable para emplear técnicas basadas en reconocimiento de voz para facilitar las búsquedas sobre contenidos multimedia que incluyen voz. Algunos ejemplos de ello son las evaluaciones que se realizan periódicamente, como pueden ser: NIST OpenKWS [1], MediaEval Query-by-Example [2] y ALBAYZIN Search-on-Speech [3].

Las búsquedas en voz tienen múltiples aplicaciones, desde recuperación de información de repositorios multimedia, a aplicaciones de seguridad. Otra de las aplicaciones que tiene puede ser el análisis de medios de comunicación (TV y radio), siendo una de éstas la detección de menciones de empresas o personajes públicos. También se puede emplear para controlar que en los medios de comunicación los locutores realicen las menciones publicitarias que se habían contratado, que es el objetivo de este trabajo.

Para que estas aplicaciones sean de utilidad las búsquedas que se realizan deben ser rápidas y efectivas, consiguiendo unos resultados precisos.

Pero esto no ocurre de igual forma en todas las lenguas ya que para el inglés, al ser el idioma más hablado del mundo, la investigación ha sido mucho mayor, por lo que se ha conseguido sistemas mucho más eficientes que en otras lenguas.

Este TFG se centra en el español, que aunque también es un idioma con bastantes recursos lingüísticos, no ha recibido tanta atención en cuanto al desarrollo del reconocimiento de voz como el inglés, por lo que los resultados son bastante más limitados.

Cabe decir que este TFG parte de un trabajo previo [4] que se empleó en la evaluación ALBAYZIN Search-on-Speech 2014, anteriormente comentada. Nuestro objetivo será mejorar este sistema y emplearlo en un contexto diferente, en este caso hemos elegido la detección de menciones en medios de comunicación.

1.2 Objetivos

El objetivo principal de este proyecto es la mejora de un sistema de reconocimiento de palabras clave en español y su aplicación a la detección de menciones en los medios de comunicación.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

- **Capítulo 1: Introducción**

  Contiene la motivación y los objetivos de este proyecto.

- **Capítulo 2: Estado del arte**

  En este capítulo podemos encontrar una explicación de qué hay que tener en cuenta a la hora de desarrollar un reconocedor automático de voz. Así como las técnicas que se han utilizado y las que utilizaremos en este proyecto de forma detallada tanto para el reconocimiento de voz como para el reconocimiento de palabras clave.

- **Capítulo 3: Diseño**

  En este capítulo se explica las herramientas utilizadas para la realización del sistema, además se detallan las bases de datos utilizadas.

- **Capítulo 4: Desarrollo**

  En esta sección se explica la implementación y las distintas técnicas de entrenamiento que se han utilizado en el sistema, además se incluye como se ha realizado la evaluación del sistema desarrollado.

- **Capítulo 5: Integración, Pruebas y Resultados**

  En esta sección se explica las pruebas realizadas para comprobar el rendimiento del sistema y los resultados obtenidos con el sistema implementado.

- **Capítulo 6: Conclusiones y Trabajo futuro**

Este capítulo contiene las conclusiones del trabajo realizado, además de las posibles líneas para mejorar el sistema desarrollado.

2 Estado del arte

2.1 Reconocimiento de voz

Un sistema de reconocimiento de voz es una herramienta computacional capaz de procesar la señal de voz emitida por el ser humano y reconocer el mensaje que se está transmitiendo convirtiéndolo habitualmente en texto.

Dado que son sistemas que transforman voz en texto a menudo se los conoce como sistemas *Speech-To-Text, STT.* Para realizar esta transformación, el reconocedor de voz debe integrar múltiples fuentes de conocimiento, desde el conocimiento acústico-fonético (i.e. cómo suena cada fonema) hasta el conocimiento sintáctico y gramatical (representado habitualmente en modelos estadísticos de cada idioma), pasando por el conocimiento léxico (palabras en un idioma y su correspondiente transcripción fonética).

2.1.1 Marco histórico

El reconocimiento de voz ha evolucionado a lo largo de los años gracias a los avances del procesamiento de señal, algoritmos, hardware, etc. En esta sección comentaremos algunas de las técnicas más destacadas que se han empleado a lo largo de la historia.

- Durante los años 70 se comenzó a usar **Dynamic Time Warping (DTW)**. Es un algoritmo que mide la similitud entre dos secuencias temporales que pueden variar en duración y velocidad. Para ello hace uso del alineamiento temporal consiguiendo así alinear los vectores de características más similares entre sí. El fruto del alineamiento es el camino de alineamiento óptimo, el cual debe tener el menor coste posible [5].

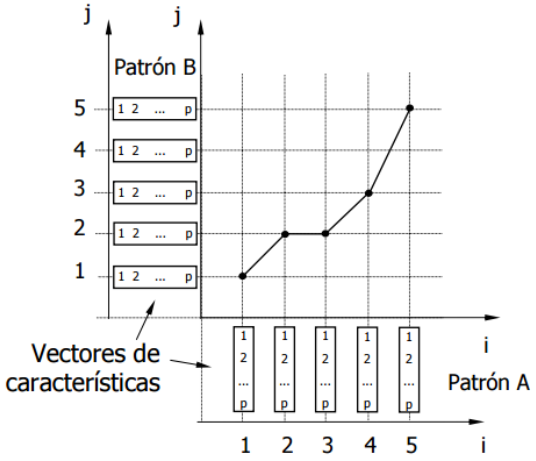

Figura 2-1: Técnica de DTW [5]

- En los años 90 comenzó el uso de modelos estadísticos, los Modelos Ocultos de Marcov **(Hidden Markov Model, HMM).** Los HMMs son modelos estadísticos cuyo objetivo es determinar los parámetros desconocidos a partir de parámetros observables los cuales no se corresponden de forma determinista sino probabilística con los estados posibles del sistema [7]. Estos algoritmos han constituido el estado del arte en reconocimiento de voz durante al menos 3 décadas.
- En la década actual los modelos basados en las Redes Neuronales Profundas (**Deep Neural Netwoks, DNN)** han sustituido o al menos complementado a los Modelos Ocultos de Markov**.** Son modelos matemáticos inicialmente basados en las capacidades de procesamiento de la información del cerebro, aunque no tratan de simular más que de una forma básica el funcionamiento de las neuronas. Para ello se realiza un entrenamiento discriminativo sin hacer suposiciones respecto a los estadísticos de las características. Las redes neuronales (a veces denominadas redes neuronales artificiales para remarcar que sus diferencias con las redes neuronales biológicas) son en realidad un modelo bien conocido desde hace décadas. Su resurgir actual se debe a que debido a las mejoras de los sistemas computacionales y al incremento de los datos disponibles, es posible trabajar con estructuras multicapa que permiten ir aprendiendo niveles secuencialmente más abstractos en la representación de los datos.

Las técnicas de HMM y DNN las explicaremos más delante con más detalle.

2.1.2 Modelos acústico-fonéticos, léxicos y de lenguaje

Para el desarrollo de un sistema de reconocimiento de voz es necesario desarrollar modelos acústicos-fonéticos, léxicos y de lenguaje.

El modelo de lenguaje es un componente que trata de predecir la palabra que seguirá a una secuencia de palabras reconocida previamente. Los humanos utilizamos algún tipo de modelo de lenguaje en nuestro procesamiento del habla. Es lo que nos permite completar una frase cuando nos falta una o varias palabras. En el caso del reconocimiento de voz se emplean habitualmente modelos estadísticos basados en n-gramas que modelan la probabilidad de aparición de secuencias de n palabras. Estos modelos se entrenan habitualmente con una gran cantidad de texto o transcripciones de una gran cantidad de voz. Con ello se intenta predecir la siguiente palabra en una secuencia del lenguaje. El modelo de lenguaje es una pieza crucial del reconocedor de voz, ya que permite adaptar el reconocedor a la tarea. Sin un adecuado modelo de lenguaje la tasa de error de

reconocimiento (medida habitualmente como la tasa de error de palabra, Word Error Rate, WER) tiende a elevarse considerablemente.

Otro elemento clave del reconocedor de voz es el modelo del léxico. En este caso se trata del conjunto de palabras que el reconocedor conoce, y deben tener asociadas sus transcripciones fonéticas. El léxico es un elemento clave porque el reconocedor no puede reconocer una palabra que no esté en su léxico. En este punto pueden resultar importantes variaciones en la pronunciación debidas a variedades dialectales, que pueden hacer interesante incluir pronunciaciones alternativas en el léxico.
Pero, sin duda el modelo más complejo de un reconocedor de voz es el modelo acústico-fonético, que modela cómo suena cada uno de los distintos fonemas (o alófonos) de una lengua. Para ello se emplean modelos estadísticos o de otro tipo (por ejemplo discriminativos como las redes neuronales) para unidades de habla (habitualmente fonemas), a partir de los cuales construir modelos de palabras y por último evaluar las probabilidades del modelo acústico a través de los métodos de concatenación.

Los modelos acústico-fonéticos no suelen modelar directamente la señal de voz, sino que operan sobre unas secuencias de vectores de características que se extraen de la señal de voz en un proceso de extracción de características. Por ello, antes de pasar a describir los modelos acústico-fonéticos más empleados vamos a describir las técnicas de extracción de características más comunes, ejemplificadas en la etapa de extracción de características que hemos utilizado.

2.1.3 Extracción de características

Lo primero que se debe realizar, tanto para la creación de los modelos fonéticos como para el reconocimiento, es la extracción de características acústicas de las grabaciones. Dicha extracción la realizaremos en el dominio de la frecuencia. En particular usaremos **Mel Frequency Cepstral Coefficients (MFCC)** que es la extracción de características más utilizada [6]. Habitualmente no se emplean ya directamente los MFCC, sino que se emplean técnicas de transformación de características que también comentaremos.

2.1.3.1 Coeficientes MFCC

Son coeficientes que representan la voz y que se basan en algunas características básicas de la percepción auditiva humana. Las bandas de frecuencia que se toman son en una escala perceptual (no lineal en frecuencia como la respuesta del oído humano) denominada escala Mel. Por otro lado, los parámetros dependen sólo del logaritmo del módulo de la amplitud de la señal acústica, debido a que se conoce

que el oído humano es bastante insensible a la fase de la señal acústica y que su respuesta a la amplitud es más aproximada a una respuesta logarítmica que lineal [6].

De forma más detallada, para obtener los coeficientes MFCC se siguen los siguientes pasos:

- **Enventanar:** Se segmenta la señal con ventanas solapadas. En nuestro caso se emplean ventanas de Hamming de 25ms, con un desplazamiento de 10 ms entre ventanas.
- **Pasar al dominio de la frecuencia** a través de la FFT (Fast Fourier Transform).
- **Quedarnos sólo con la respuesta en amplitud,** tomando el módulo de la FFT
- **Calcular la energía** en cada banda de frecuencias del espectro haciendo uso de un banco de filtros Mel. El banco de filtros en escala Mel trata de simular el comportamiento en frecuencia de la respuesta de la cóclea del oído interno humano.
- **Hacer el logaritmo** de la energía en cada banda de frecuencia (ya que la respuesta del oído humano a la intensidad sonora es aproximadamente logarítmica).
- **Hacer la DCT (Discrete Cosine Transform)** de las amplitudes resultantes (sólo los primeros coeficientes) serán los coeficientes MFCC. Habitualmente se emplean los 13 primeros coeficientes en reconocimiento de voz. Esta operación implica un suavizado del espectro que permite reducir el efecto de la vibración de las cuerdas vocales.

2.1.3.2 Transformación de características

Las características obtenidas tienen distorsiones debido a los efectos introducidos por el canal y el ruido, lo cuales pueden ser lineales y no lineales.
Para disminuir esos efectos existen varias técnicas como pueden ser Cepstral Mean Subtraction (CMS), Feature warping o **Cepstral Mean and Variance Normalization (CMVN);** nos centraremos en esta última [6]**.**
Por otro lado, un vector de coeficientes MFCC representa sólo un contexto temporal de 25 ms, por lo que se puede realizar una ampliación del contexto combinando vectores MFCC anteriores y posteriores en un único vector de parámetros, de forma que el vector de parámetros represente un contexto temporal más amplio.

- ***CMVN:***

Consiste en la normalización de la distribución de los datos de entrada con los factores comentados anteriormente forzando a tener media 0 y varianza 1. Con esto se consigue aliviar el problema variabilidad de sesión que hemos comentado.

- ***Expansión del contexto:***

Este paso consiste primero en tomar ±5 frames respecto al de entrada (de este modo pasaríamos de tener un vector MFCC de 13 coeficientes a tener uno de 11x13=143 coeficientes) y a continuación realizar una reducción de la dimensionalidad de los vectores de características obtenidos (se suele realizar esta reducción con **Linear Discriminant Analysis, LDA**, para pasar a vectores de en torno a 40 dimensiones).

2.1.4 Reconocimiento

2.1.4.1 Modelos Ocultos de Markov (HMM)

Un modelo de Markov es un modelo estadísticos donde se asume que el sistema que se va a modelar es un proceso de Markov, correspondiendo cada estado de forma determinista a un evento observable.
En reconocimiento de voz se usan modelos de Markov de primer orden, es decir que solo dependan de la salida inmediatamente anterior [7].

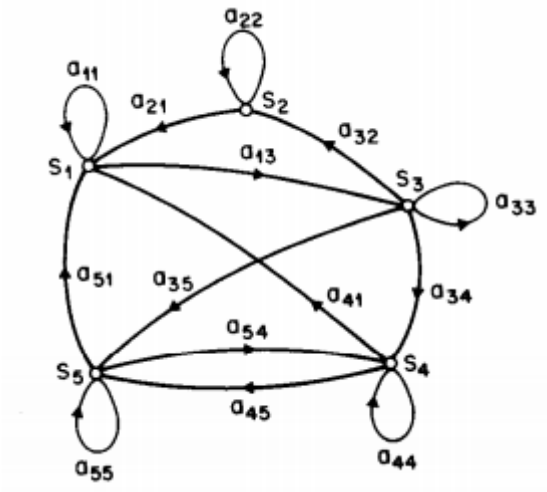

Figura 2-2: Cadena de Markov con cinco estados [7]

Un modelo oculto de Markov es un modelo de Markov (discreto y de primer orden) en que las salidas observables se corresponden de forma probabilística con los estados del sistema. En reconocimiento de voz los estados se corresponden con fonemas (en realidad con la parte inicial, media y final de un fonema) y el objetivo es determinar la secuencia de estados ocultos más probable (secuencia de fonemas más probable) para una secuencia de vectores de parámetros observables.

Los elementos por los que está caracterizado un HMM es:

- *N*, número de estados del modelo, aunque los estados son ocultos, para muchas aplicaciones algún significado físico está conectado a los estados o grupo de estados. En el caso del reconocimiento de voz, como hemos dicho los estados tienen una relación directa con los fonemas.
- *M*, el número de símbolos observables por estado, en caso de un modelo de Markov discreto. En reconocimiento de voz la observación es un vector de características continuo, con lo que este parámetro no se utiliza (existe un infinito número de observaciones posibles por estado).
- *La distribución de probabilidad entre transiciones de estados.* Los estados son en todo caso discretos, de forma que la probabilidad de transición entre estados es una matriz de probabilidades. Habitualmente de cada estado sólo es posible transitar a uno o varios estados, de modo que esta distribución de probabilidad será muy dispersa (con muchos ceros).
- *La distribución de probabilidad de un símbolo observable*, en el caso de un modelo de Markov discreto. Siendo más habitual en voz los modelos de Markov continuos, esta distribución de probabilidad se sustituye por una función densidad de probabilidad, que normalmente se modela como una mezcla de Gaussianas.
- *Estado inicial.*

Por lo tanto un modelo oculto de Markov contiene dos procesos estocásticos:

- *Una cadena oculta de Markov:* La cual modela la secuencia de fonemas teniendo en cuenta la variabilidad temporal.
- *Un proceso observable:* Que tiene en cuenta la variabilidad espectral, el cual va tomando valores dependiendo de las características de la secuencia de fonemas.

Con ambos procesos se puede llegar a reconocer la secuencia de estados ocultos (es decir la secuencia de fonemas) a partir de la secuencia de observaciones (la secuencia de vectores de parámetros). Pero para ello es necesario previamente estimar los parámetros del modelo oculto de Markov en un proceso de entrenamiento. Los modelos ocultos de Markov permiten resolver tres problemas básicos que nos permiten por un lado entrenar y por otro lado emplear los HMMs para reconocimiento de voz.

Problema 1: Problema de puntuación, es decir, dada una secuencia de observaciones (vectores de parámetros) y un modelo (entrenado previamente)

como calcular la probabilidad de la secuencia vista, dado el modelo. Este problema se resuelve con el algoritmo de Forward-Backward.

Problema 2: Problema de reconocimiento de estados, es decir, dada una secuencia de observaciones (vectores de parámetros) y un modelo, elegir cual es la secuencia de estados (y por tanto de fonemas) que mejor explica esas observaciones. Actualmente este problema que no es otro que el problema del reconocimiento se resuelve con el algoritmo de Viterbi o algoritmos más complejos que permiten obtener más de una hipótesis de reconocimiento.

Problema 3: Problema de entrenamiento, es decir, dado un conjunto de observaciones (vectores de parámetros) de entrenamiento para los que se conocen las secuencias de estados ocultos (secuencias de fonemas) correspondientes, saber cómo ajustar los parámetros del modelo (esencialmente las funciones densidad de probabilidad asociadas a cada estado y las probabilidades de transición) para maximizar la probabilidad de observar el conjunto de entrenamiento dado el modelo. Este problema se resuelve con el algoritmo de Baum-Welch,

Por último, es necesario decir que en el entrenamiento no se ajusta la topología del modelo, que viene determinada por el número de estados y la interconexión que hay entre ellos mediante las probabilidades de transición no nulas. En el caso del reconocimiento de voz, las topologías que se emplean son habitualmente lo que se conoce como **Modelo de Bakis** (o topología de izquierda a derecha) que ha demostrado ser adecuado para modelar señales de voz.

Figura 2-3 Ejemplo Modelo de Bakis

2.1.4.2 Redes Neuronales Profundas (DNN).

Las técnicas de aprendizaje automático (*machine learning*) convencionales solían basarse en una etapa de extracción de características normalmente diseñada en función del conocimiento humano del problema, y una etapa de reconocimiento de patrones sobre esas características, que típicamente trataba de obtener la clase a la que pertenecía un patrón de entrada de forma más o menos directa. Esta aproximación ha sido muy exitosa ya que permitía trabajar con relativamente poca capacidad de cálculo y datos.

En los últimos años esta limitación en capacidad de cálculo y datos está siendo cada vez menor, lo que ha propiciado que cada vez se obtengan mejores resultados con lo que se ha venido en denominar el aprendizaje profundo (*Deep learning*), que consiste en un aprendizaje con múltiples niveles (a veces partiendo de datos en

crudo, como el audio directamente), teniendo cada nivel una representación de los datos más abstracta que el nivel anterior. Con esto se logró conseguir poder procesar con una mayor facilidad y eficiencia una gran cantidad de datos. Unas de las técnicas de aprendizaje profundo más importantes en la actualidad son las DNN [8].

Los pasos (básicos y muy simplicados) a seguir para la creación de una DNN:

1. Elegir una arquitectura (conectividad entre neuronas):
   - Número de unidades de entrada: Dimensión de las características.
   - Número de unidades de salida: Número de clases.
   - Número de capas ocultas y número de unidades en las mismas: Depende del problema que se esté tratando de resolver, cuanto mayor es la complejidad del problema mayor es el número de capas ocultas y unidades en las capas ocultas que necesitaremos.
2. Inicialización de los pesos de forma aleatoria.
3. Realizar la propagación forward.
4. Calcular la función de coste.
5. Por último realizar la propagación backward.
6. Actualizar los pesos de la red en función de los resultados del paso anterior.

Los pasos 3 a 5 se deben repetir por cada ejemplo o conjunto de ejemplos. El paso 6 se puede realizar por cada ejemplo o por cada conjunto de ejemplos. Típicamente hay que repetir el proceso durante varias épocas (epochs) consistentes en presentar todos los datos del conjunto de entrenamiento a la red. Una de las formas más habituales de determinar cuándo parar con este proceso es emplear la técnica de validación cruzada, que consiste en parar cuando los resultados que consigue la red en un conjunto de datos (denominado de validación cruzada) que no se emplea en el entrenamiento dejan de mejorar o empiezan a empeorar. De este modo evitamos que la red se sobre-adapte a los datos de entrenamiento.

En reconocimiento de voz las redes neuronales profundas se emplean fundamentalmente para sustituir los modelos de mezclas de Gaussianas como estimadores de la probabilidad de un estado para un vector de características dado, con lo que se genera un sistema combinado HMM-DNN (el HMM sigue funcionando igual con la única diferencia de que la estimación de la probabilidad de un estado dado un vector de observación la realiza la DNN en lugar de un modelo de mezcla de Gaussianas). En los últimos años también han aparecido sistemas de reconocimiento de voz que eliminan totalmente los HMMs, sustituyéndolos por redes profundas recurrentes como las **Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNNs),** y consiguiendo por tanto un reconocedor de voz que emplea exclusivamente DNNs .

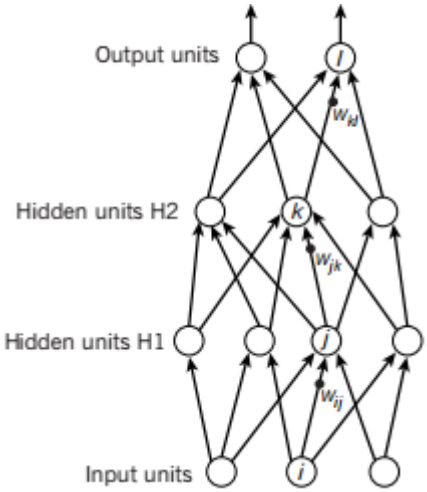

Figura 2-4 Ejemplo de DNN [8]

2.2 Reconocimiento de palabras clave

El reconocimiento de palabras clave es un problema muy relacionado con el reconocimiento de voz que consiste en la identificación de palabras específicas dentro de una locución. Es un algoritmo de búsqueda de palabras clave en archivos de audio de forma mecanizada, pudiendo así localizar el archivo donde se encuentran dichas palabras clave y dónde se encuentran dentro de ese archivo. Esto supone un ahorro importante de tiempo frente a la búsqueda manual.
Uno de los problemas más complejos del reconocimiento de palabras clave es ser capaces de encontrar las palabras fuera de vocabulario (i.e. palabras que no están en el vocabulario del reconocedor y que por tanto es imposible reconocer) como pueden ser los nombres propios, extranjerismos, acrónimos…ya que no suelen estar en el vocabulario de los reconocedores de habla y suelen ser los términos usados en las consultas.

2.2.1 Tipos de sistema de búsqueda

Los sistemas de búsqueda en voz pueden clasificarse en cuatro tipos:

- **Keyword Spotting (KWS):** En este reconocedor se conocen los términos de la consulta antes del procesado del audio siendo la entrada del sistema dichos términos.
- **Spoken Term Detection (STD):** En este caso el audio es procesado sin un conocimiento previo de los términos de la consulta.
- **Query by Example Spoken Term Detection (QbE STD):** Es un sistema de Spoken Term Detection en el que la entrada del sistema es una consulta acústica (una grabación de voz) en lugar de la transcripción de un término.

- **Query by Example Spoken Document Retrieval (QbE SDR):** Se trata de un sistema en el que también la consulta es acústica, pero en el que lo que se pretende es recuperar los documentos que contengan dicha consulta acústica, no siendo importante la localización precisa del término dentro del documento.

Todos estos sistemas miden su rendimiento en términos porcentaje de pérdida (no detectar términos que aparecen) y de falsas alarmas (detecciones erróneas de términos).

2.2.2 Tipos de sistemas según el método de funcionamiento

En la actualidad los tipos de sistemas según su funcionamiento son: reconocedores de habla continua de gran vocabulario, modelos de relleno y reconocedores de voz de sub-unidades de palabras. Para la solución del problema de detección de palabras fuera de vocabulario se puede emplear modelos combinados y proxy words [9].

2.2.2.1 LVCSR

Los reconocedores de habla continúa de gran vocabulario **(Large Vocabulary Continuous Speech Recognition, LVCSR)**, se caracterizan por perseguir el reconocimiento de palabras pronunciada de manera natural y cubriendo un vocabulario extenso (un vocabulario de decenas de miles de palabra que cubran las más frecuentes de un idioma).
Este sistema es el que mejor funciona si todas las palabras que consulta el usuario están dentro del vocabulario del sistema. Sin embargo este fenómeno no siempre sucede siendo imposible reconocer una palabra que este fuera de vocabulario (out-of-vocabulary , OOV).

2.2.2.2 Sistemas basados en Modelos de relleno

A diferencia de los LVCSR, los sistemas basados en modelos de relleno contienen en su diccionario únicamente las palabras clave que se desean extraer de audio junto con unos modelos especiales de relleno que tratan de modelar el resto de las palabras del idioma (e incluso otros eventos acústicos).
Estos sistemas son los segundos con mejores resultados pero tienen el inconveniente de que si se cambia alguna palabra a buscar debe de volver a procesar todo el audio por lo que no son útiles en las aplicaciones donde el vocabulario cambia constantemente además de que tampoco son capaces de identificar las palabras fuera de vocabulario.

2.2.2.3 Reconocedores de voz de sub-unidades de palabra

Estos sistemas son los que peores resultados obtienen debido a que no consideran ningún modelo de palabra durante el proceso de reconocimiento, pero al contrario del sistema basado en modelos de relleno, no necesita reprocesar el audio cuando el vocabulario de la aplicación cambia, por lo que permite una mayor flexibilidad. Además también tiene una mayor velocidad de procesamiento del audio.
Esto es debido en que se basan en sub-unidades de palabra es decir fonemas, silabas, etc como unidad de trabajo a la hora de entrenar y evaluar, por lo que las combinaciones de dichas unidades permiten la formación de cualquier palabra. Esto hace posible la detección de las palabras fuera de vocabulario (de hecho estos sistemas no manejan el vocabulario en el reconocimiento, por lo que no existe el problema de las palabras fuera del vocabulario).

2.2.2.4 Modelos combinados

Este tipo de sistemas emplean un LVCSR para palabras del vocabulario y un reconocedor de sub-unidades para las palabras fuera del vocabulario (OOV). Este sistema trata de aunar las ventajas de ambos sistemas (mejores resultados de un sistema LVCSR para palabras del vocabulario y posibilidad de detectar palabras fuera del vocabulario con el sistema basado en subunidades de palabras).

2.2.2.5 Proxy words

Otra aproximación para las palabras fuera del vocabulario es el método de las "proxy words", es decir buscar palabras del vocabulario acústicamente similares a las palabras a buscar y que no están en el vocabulario. Esta es la aproximación que utiliza la herramienta Kaldi que hemos usado en este TFG.

3 Diseño

3.1 Herramientas utilizadas: Kaldi

Kaldi es una herramienta para el reconocimiento de voz escrito en C++. Está bajo la licencia de Apache, la cual es una de las menos restrictivas posibles. Kaldi es la herramienta de reconocimiento de voz que más se utiliza actualmente en investigación, ofreciendo unos resultados de reconocimiento en el estado del arte, y en la que se integran rápidamente los últimos avances en reconocimiento de voz [10] [11].

Esta herramienta utiliza dos conjuntos de librerías externas:

- OpenFST: Librería de transductores de estados finitos para modelos las transiciones entre los estados.
- BLAS/LAPACK: Algebra numérica para cálculos matemáticos.

La arquitectura de esta herramienta se puede observar en la siguiente figura:

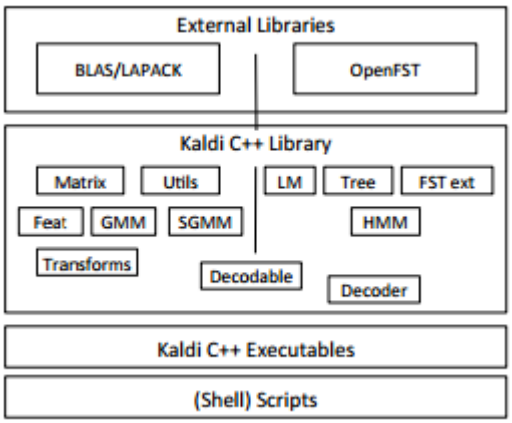

Figura 3-1 Componentes de Kaldi [11]

Este TFG se realizará totalmente empleando Kaldi, utilizando las funciones y scripts, que ya están implementados de la forma más genérica posible.

3.2 Bases de Datos

Este TFG parte de varias bases de datos, Fisher Spanish y Callhome Spanish son bases de datos del Linguistic Data Consortium (LDC) [12], ambas contienen conversaciones telefónicas espontáneas en español (grabadas en Estados Unidos). Por otro lado, se utilizará una base de datos capturada y etiquetada también como parte de este proyecto: la base de datos ATVS-Radio que contiene programas de radio de diferentes emisoras obtenidos a través de Internet y etiquetados con distintas informaciones.

Base de datos	Horas	Número Grabaciones
Fisher Spanish	163	819
Callhome Spanish	60	120
ATVS-Radio	300	300

Tabla 3-1 Resumen bases de datos

3.2.1 Fisher Spanish

Esta base de datos ha sido realizada por el Consorcio de Datos Lingüísticos en Estados Unidos, contiene aproximadamente 163 horas de conversaciones telefónicas, en las cuales participaban 136 locutores diferentes. Las grabaciones se realizaron en Estados Unidos, por lo que el español es principalmente español americano con diferentes variedades dialectales [13].

Consta de 819 grabaciones de 10 ó 12 minutos cada una, almacenadas en ficheros con formato NIST Sphere. Las conversaciones constan de dos canales y una tasa de muestreo de 8000 muestras por segundo.

Además también se dispone de la transcripción de dichas grabaciones almacenadas en ficheros de texto plano. Estas transcripciones de referencia son esenciales para realizar el entrenamiento de los modelos acústicos, léxicos y de lenguaje.

3.2.2 Callhome Spanish

Esta base de datos ha sido realizada por el Consorcio de Datos Lingüísticos en Estados Unidos al igual que Fisher Spanish, la cual consta de 120 conversaciones telefónicas sin guión entre hablantes nativos de español [14].
Todas las llamadas tienen una duración de 30 minutos, éstas se originaron primero en América del Norte y luego en otros destinos internacionales. Los participantes mayoritariamente eran familiares o amigos cercanos.

Al igual que Fisher Spanish los ficheros están en formato NIST Sphere, tienen dos canales y una tasa de muestreo de 8000 muestras por segundo, además de proporcionarnos la transcripción de las grabaciones en ficheros de texto plano. Estas transcripciones de referencia son esenciales para realizar el entrenamiento de los modelos acústicos, léxicos y de lenguaje.

3.2.3 Base de datos ATVS de radio

Esta base de datos ha sido realizada como parte del trabajo de este TFG, y de forma coordinada con cuatro compañeros más del grupo ATVS-UAM.

La base de datos consta de 300 horas de audio, de las cuales 25 horas están etiquetadas y 275 horas sin etiquetar.

Las especificaciones de las grabaciones acordadas son:

- 16000 muestras por segundo (Para este TFG se pasó a 8000 muestras por segundo),
- 16 bits/muestra.
- Un canal de grabación.
- Formato wav.
- Fechas de podcast para el etiquetado: 25-05-2015 al 05-06-2015.
- Fechas de podcast para el no etiquetado: 08-06-2015 al 17-08-2015.

Cabe destacar que en todos los programas elegidos los locutores son nativos de español.

Los programas de radio elegidos y las franjas horarias de emisión, grabación y etiquetado quedan recogidos en la siguiente tabla:

Cadena	Programa	Emisión	Grabación	Etiquetado
COPE	La mañana	6:00-12:00	10:00-11:00	10:00-10:30
CADENA SER	Hoy por hoy	6:00-12:00	9:00-10:00	9:30-10:00
ONDA CERO	Más de uno	6:00-12:30	8:30-9:30	8:30-9:00
ONDA CERO	Julia en la onda	16:00-19:00	18:00-19:00	18:00-18:30
ROCK FM	El pirata y su banda	6:00-10:00	6:00-7:00	6:00-6:30

Tabla 3-2 Programas base de datos

3.2.3.1 Etiquetado

La herramienta utilizada para el etiquetado fue *Wavesurfer*, el etiquetado se distribuye en cuatro niveles:

1. VOZ/ VOZ TELEFONICA/ NO VOZ.

2. MUSICA/NO MUSICA.

3. El tercer nivel se corresponde con la publicidad existiendo tres posibilidades:
    a. *Anuncios publicitarios:* AN_<MARCA>_<PRODUCTOLIBRE>. Por anuncio se entiende un segmento de audio pregrabado que se reproduce siempre igual cada vez que aparece el anuncio.
    b. *Menciones publicitarias:* ME_<MARCA>_<PRODUCTOLIBRE>. Por mención publicitaria se entiende un segmento de audio en el que el locutor que conduce el programa u otra persona hace una intervención no pregrabada (muchas veces en directo) para promocionar un producto o marca. Precisamente este es el objetivo principal de este TFG, ser capaz de detectar menciones publicitarias en medios de comunicación (para los anuncios se pueden utilizar técnicas de fingerprinting más sencillas y efectivas).
    c. NO PUBLICIDAD.

4. Este nivel corresponde con los locutores, distinguiendo entre varios tipos:
    a. Locutores habituales: <NNNAAA> es decir las tres primeras letras del nombre y las tres primeras letras del apellido.
    b. Locutores ocasionales: L1-L9.
    c. Voz solapada: SOLAP (tramos mayores o igual de 1s).
    d. NO.

Un ejemplo de etiquetado lo podemos ver en la siguiente figura:

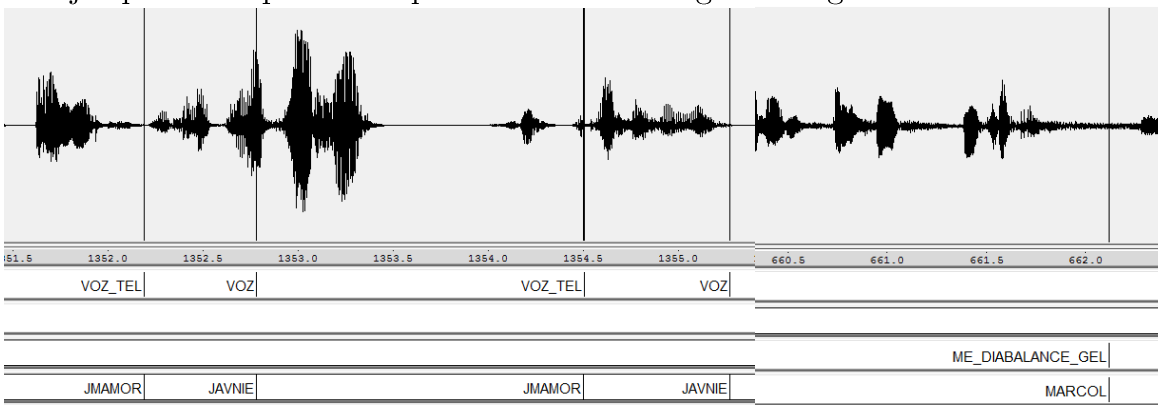

Figura 3-2 Ejemplo Etiquetado

Para la realización de este TFG se ha realizado una extracción de las menciones que aparecen en la base de datos. La base de datos contiene en total de 62 menciones distribuidas en los programas de radio de la siguiente forma:

- La mañana: 22 menciones

- Más de uno: 28 menciones
- Hoy por hoy: 12 menciones
- El pirata y su banda / Julia en la onda: 0 menciones.

3.3 Tareas y experimentos

A partir de las herramientas y bases de datos explicadas, entrenaremos un reconocedor de voz en español con objetivo de mejorar el que se usaba en un proyecto de Fin de Carrera previo [4]. Para ello, vamos a utilizar las bases de datos Fisher Spanish y Callhome Spanish en lugar de sólo Fisher Spanish, con esto rediseñaremos el léxico, los modelos de lenguaje y los modelos acústicos.
Una vez realizado dicho entrenamiento del reconocedor de voz, se realizará una modificación para tratar de detectar menciones publicitarias. Comprobaremos su funcionamiento a través de la base de datos ATVS-Radio.

4 Desarrollo

El primer paso en el desarrollo de este TFG consiste en preparar los datos para las fases posteriores del desarrollo. En primer lugar será necesario entrenar un reconocedor en español. Para ello es necesario preparar los datos de entrenamiento de los modelos acústicos y de lenguaje, así como obtención del léxico. El reconocedor de español será evaluado para comprobar su grado de precisión, por lo que también será necesario preparar los datos de evaluación. Finalmente, se empleará el reconocedor para hacer las pruebas de detección de menciones publicitarias, y de nuevo será necesario preparar los datos para hacer las pruebas finales y obtener una evaluación de la precisión en la detección de menciones.

4.1 Preparación de los datos de audio

En primer lugar hay que crear los siguientes ficheros a partir de las bases de datos comentadas para preparar los datos de audio de las grabaciones obtenidas:

- El fichero 'Text' asocia cada locución con su transcripción.
- El fichero 'Segments' define para cada locución el fichero de audio en el que está, así como su inicio y final.
- El fichero 'Wav.scp' relaciona cada grabación con su identificación.
- El fichero 'Reco2file_and_channel' asigna cada grabación con el canal que le corresponde.
- El fichero 'Utt2spk' identifica cada locución con su locutor.
- El fichero 'Spk2gender' asocia cada locutor con su género.
- El fichero 'Spk2utt' relaciona cada locutor con las locuciones que le corresponden.

Fichero	**Estructura**
Text	*<utterance-id> <text>*
Segments	*<utterance-id> <recording-id> <segment-begin> <segment-end>*
Wav.scp	*<recording-id> <extended-filename>*
Reco2file_and_channel	*<recording-id> <filename> <recording-side (A or B)>*
Utt2spk	*<utterance-id> <speaker-id>*
Spk2gender	*<speaker-id> <gender>*

Spk2utt	*<speaker-id><utterance-id_1><utterance-id_2>…*

Tabla 4-1 Estructura de los ficheros preparados para los datos de audio

La función principal de estos ficheros es relacionar la información obtenida de las locuciones, como puede ser la estructura de las grabaciones, la segmentación temporal de las locuciones o la información de los locutores. Todo ello para que sea posible emplear estos datos tanto para entrenar como para evaluar el reconocedor de voz.

Estos ficheros se han generado para el conjunto de entrenamiento y evaluación con las bases de datos Fisher Spanish y Callhome Spanish. Finalmente, también se han tenido que generar para el subcorpus de menciones extraído de la base de datos ATVS-Radio.

4.2 Preparación de los datos textuales

El siguiente paso es la elaboración de los datos textuales para ello se debe crear los siguientes ficheros:

Fichero	Contenido
Extra_questions.txt	*Vacío*
Nonsilence_phones.txt	*Fonemas convencionales*
Optional_silence.txt	*SIL*
Silence_phones.txt	*SIL, NSN, LAU y SPN*
Lexicon.txt	*Palabras transcritas y además <UNK> SPN, !SIL SIL*

Tabla 4-2 Contenido de los ficheros de datos textuales

El fichero lexicon.txt contiene el léxico es decir las palabras que conoce el reconocedor y sus trascripciones fonética. Dichas palabras deben estar en orden alfabético y sin duplicaciones, aunque si pueden tener una misma palabra varias transcripciones fonéticas.

Para su creación se ha analizado las transcripciones textuales proporcionadas por las bases de datos, en éstas podemos distinguir diferentes tipos de datos:

- Palabras en castellano.
- Palabras en inglés.

- Siglas.
- Interjecciones.
- Palabras erróneas y palabras parciales.

4.2.1 Procedimiento de generación del léxico

Para conseguir el fichero de Lexicon.txt se ha debido realizar los siguientes pasos:

1. Extracción del texto, cada línea del fichero resultante contiene un segmento de la grabación.
2. Cada segmento se compone de cadenas de caracteres separadas por espacios en blanco. Por lo tanto se divide dichas cadenas en palabras aisladas consiguiendo un fichero que contenga una palabra por línea.
3. Se eliminan los signos de puntación de las palabras obtenidas.
4. Se ordenan alfabéticamente.
5. Se eliminan las palabras duplicadas.

Una vez obtenido la lista de palabras, se realiza la transcripción fonética a través de un transcriptor fonético automático de español proporcionado por el grupo ATVS. Éste se encuentra realizado en lenguaje Perl y tiene como entrada un fichero de texto con las palabras del léxico correspondientes y a la salida se obtiene sus transcripciones fonéticas, cada fonema está separado por un espacio en blanco, dicha transcripción se realizara según la siguiente tabla de reglas, a la que aplicarán algunas excepciones contenidas en otro fichero:

Letra	Fonema	Comentarios	Letra	Fonema	Comentarios
a	a		**o**	o	
á	a		**ó**	o	
b	b		**p**	p	
c	T		**q**	k	
c	k	Delante de a,o,á,ó,u,ú	**r**	r	
d	d		**r**	R	Comienzo de palabra
e	e		**rr**	R	
é	e		**s**	s	
f	f		**t**	t	
g	g		**u**	-	Detrás de q o de g.(-)
g	x	Delante de e,é,i,í	**u**	u	
h	-		**ú**	u	

i	i	**v**	b	
í	i	**w**	u	
j	x	**x**	s	
k	k	**y**	y	Delante o detrás de a,e,i,o,u,á,é,í, ú
l	l	**y**	i	
ll	y	**z**	T	
m	m	**ch**	C	
n	n	**ñ**	N	

Tabla 4-3 Correspondencia letras y fonemas

Se debe de destacar que las siglas y las interjecciones tanto en inglés como en castellano se han transcrito de forma manual, al igual que algunas palabras en inglés, haciendo una translación a fonemas en castellano, (según la siguiente tabla):

Fonema inglés	Fonema castellano	Fonema inglés	Fonema castellano
aa	a	**iy**	i
ae	a	**jh**	y
ah	a	**k**	k
ao	o	**ng**	ng
aw	au	**ow**	ou
ay	ai	**oy**	oi
dh	d	**p**	o
eh	e	**sh**	s
er	er	**th**	th
ey	ei	**uh**	u
gg	g	**uw**	u
hh	x	**z**	s
ih	i	**zh**	s

Tabla 4-4 fonemas inglés correspondencia castellano

Una vez realizado estos pasos ya se obtiene las transcripciones fonéticas y uniéndolas con el fichero de palabras, se obtiene el fichero lexicon.txt.
Finalmente este fichero consta de 40328 palabras. Algunas de estas palabras están duplicadas como puede ser algunas siglas, que se encuentran transcritas fonéticamente en inglés y en castellano para tener en cuenta las dos posibilidades.

Un pequeño ejemplo de este fichero:

```
accident        a k T i d e n t
accidentalmente     a k T i d e n t a l m e n t e
accidentando        a k T i d e n t a n d o
```

4.2.2 Generación del léxico para la detección de menciones

Para la generación de este léxico tenemos que modificar el fichero lexicon.txt que ya teníamos. Para ello los pasos a seguir son:

1. Elegir las palabras clave de cada mención[1].
2. Realizar la transcripción fonética.
3. Añadir las palabras elegidas ya transcritas al fichero lexicon.txt
4. Ordenar alfabéticamente el fichero.

Finalmente el fichero lexicon.txt contiene 40379 palabras. Se han añadido 51 palabras nuevas[2].

4.3 Generación del modelo de lenguaje

Para la creación del modelo de lenguaje se hace uso de la herramienta SRILM que viene integrada con Kaldi.

Este modelo de lenguaje que hemos creado se ha realizado a partir de los textos de Callhome Spanish y Fisher Spanish, siendo ambos en habla conversacional telefónica.

A pesar de que las menciones utilizadas se realizan sobre habla de medios de comunicación, para su detección no hemos adaptado el modelo de lenguaje obtenido en el entrenamiento ni lo hemos entrenado para incluir las palabras que se van a emplear en la detección de menciones, pudiendo ser una posible mejora para un trabajo futuro.

4.4 Extracción de características

En este paso se realiza la extracción de las características acústicas de los ficheros de las grabaciones y con ello la construcción de los modelos fonéticos.

[1] Ver Anexo A: Base de datos menciones.

[2] Ver Anexo B: Palabras clave de las menciones

Este proceso lo realizaremos tanto para el entrenamiento y evaluación del reconocedor de voz como para la detección de menciones.

Para ello la extracción de características la realizaremos en el dominio de la frecuencia y usaremos **Mel Frequency Cepstral Coefficients (MFCC).**

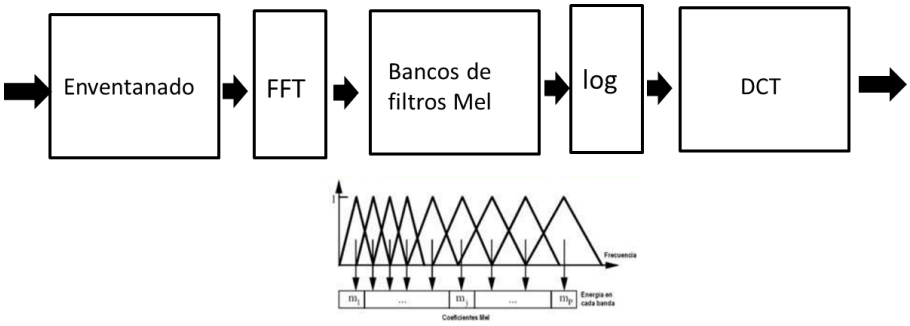

Figura 4-1 Proceso de Extracción MFCC

A continuación se realiza la normalización para la reducción de los efectos introducidos por el canal, para ello se utiliza la técnica de **Cepstral Mean and Variance Normalization (CMVN),** esto se realizara para cada uno de los audios. Por último se realiza una expansión del contexto consiguiendo que los vectores de parámetros representen un contexto temporal más amplio.

4.5 Entrenamiento de los modelos acústicos

El siguiente paso es la construcción de los modelos fonéticos a través de los diferentes tipos de entrenamiento, en este caso hemos realizado 6 tipos de entrenamiento que comentaremos más adelante.
Se debe destacar que el entrenamiento que se lleva acabo es incremental ya que se basan en el modelo anterior entrenado, utilizando siempre los alineamientos temporales entre la voz y la transcripción generados con el modelo anterior entrenado [11].

4.5.1 Modelos básicos MFCC

Consiste en el entrenamiento de un modelo **monofonema** es decir que no dependa del contexto, del cual a continuación del entrenamiento se realiza un alineamiento del modelo para la posterior utilización en el entrenamiento del modelo **trifonema**. En este entrenamiento ya se tiene en cuenta el contexto, tanto el fonema anterior como posterior.
A continuación se realiza de nuevo otro entrenamiento del modelo trifonema, consiguiendo un modelo fonético más eficiente preciso.

4.5.2 Modelos LDA + MLLT

En este caso consiste en aplicar dos técnicas **Linear Discriminant Analysis (LDA)** y **Maximum Likelihood Linear Transform (MLLT)** sobre el alineamiento del modelo de trifonemas.
La técnica de LDA se usa tras realizar la expansión de contexto temporal de las características MFCC para obtener vectores con contexto de ±5 frames, pero de dimensión más reducida..
En cambio MLLT es una técnica de adaptación al entorno y al locutor basada en la transformación lineal de las características obtenidas mediante LDA.

4.5.3 Modelos fMLLR + SAT

Esta técnica se conoce como **Speaker Adaptive Training (SAT)**. En primer lugar se realiza una transformación de las características para su adaptación al locutor mediante **feuture-space Maximum Likelihood Linear Regression (fMLLR)** y a continuación ya si se puede realizar el entrenamiento adaptado al locutor.
Por lo tanto fMLLR es una técnica de adaptación de locutor de igual funcionamiento que MRRL pero mediante la transformada del espacio de características.

4.5.4 Modelos SGMM

Este modelo es el **Subspace Gaussian Mixture Models (SGMM)** se basa en modelar conjuntamente locutor y fonemas empleando técnicas de subespacios y GMMs.

4.5.5 Modelos bMMI + SGMM

Esta técnica es **boosted Maximum Mutual Information (bMMI)** aplicada sobre **subspace GMM (SGMM).**
La técnica bMMI se basa en MMI (Maximum Mutual Information). Se trata de una técnica discriminativa para tratar de hacer que los modelos acústicos no sólo representen correctamente los datos de entrenamiento, sino que se maximice su capacidad de discriminación entre los distintos fonemas.

4.6 Evaluación de los reconocedores

Se realiza una decodificación para cada modelo acústico desarrollado para comprobar su grado de precisión. Ese proceso consistirá en tres pasos:

- ***Construcción del grafo***

Antes de hacer un reconocimiento de voz con Kaldi es necesario construir el grafo de reconocimiento que consiste en combinar en un único grafo el modelo de lenguaje, el léxico, la dependencia del contexto y por último la topología de los HMM [10].

- ***Decodificación del grafo***

Una vez construido el grafo en la decodificación se obtiene la secuencia más probable de palabras integrando todas las fuentes de información del reconocedor.

- ***Score o puntuación***

En este paso se realiza una puntuación sobre las palabras obtenidas en la decodificación. Para ello el sistema realiza una transcripción del audio reconocido según los modelos entrenados y lo compara con las transcripciones de referencia.

4.6.1 Evaluación del reconocedor de voz

Para la evaluación del reconocedor se ha usado las transcripciones que nos facilitaban las bases de datos Fisher Spanish y Callhome Spanish. La evaluación se realiza calculando la tasa de error de palabra o **Word Error Rate (WER).** Éste es el porcentaje de palabras erróneas tras la decodificación calculando el número mínimo de inserciones, borrados y sustituciones de una palabra por otra, respecto al número total de palabras. Es importante hacer notar que la WER puede ser mayor del 100% debido al efecto de las inserciones.

$$WER = \frac{\#\,Inserción + \#\,Borrado + \#\,Sustitución}{\#\,Palabras} * 100$$

4.6.2 Evaluación del reconocedor de palabras clave

En este caso no tenemos la transcripción de las menciones por lo que se ha realizado la evaluación del siguiente modo. Una vez obtenidas las transcripciones del audio reconocido en la etapa de decodificación, si algunas de las palabras clave seleccionadas para cada mención es detectada se considera que la mención se ha detectado, y en caso contrario se considera que la detección no se ha detectado.

5 Integración, pruebas y resultados

5.1 Introducción

En este capítulo se describe las pruebas realizadas para comprobar el funcionamiento de nuestro reconocedor de voz y detector de menciones.

5.2 Resultados de Reconocimiento de voz

Modelo Fonético	% WER
MFCC_1	57.32
MFCC_2	56.76
+LDA + MLLT	50.63
+fMLLR + SAT	46.99
GMM	45.12
+bMMI + SGMM	41.10

Tabla 5-1 Resultados obtenidos del reconocedor de voz

A fin de comparar los resultados obtenidos, se ofrece a continuación una tabla comparativa con los resultados de diferentes reconocedores de voz en español con estas bases de datos:

	Sistema inicial [4]	Sistema desarrollado	Resultado [15]	Mejor resultado de Kaldi
% WER	**49.88**	**41.10**	**36.50**	**29.80**

Tabla 5-2 Comparativa reconocedores de voz

En la siguiente tabla se muestran los mejores resultados obtenidos por la herramienta Kaldi con estas bases de datos. Cabe destacar que el mejor resultado obtenido con Kaldi es con DNNs, y que a fecha de cierre de esta memoria nuestros modelos con DNNs se habían comenzado a entrenar pero el entrenamiento todavía no había terminado.

Modelo Fonético	% WER
MFCC_1	47.83
MFCC_2	47.35
+LDA + MLLT	42.07
+fMLLR + SAT	37.56
GMM	35.42
+bMMI + SGMM	32.73
DNN	29.80

Tabla 5-3 Resultados de Kaldi

Aunque una WER del 41.10% puede parecer muy elevada, para que podamos hacernos una idea del resultado obtenido por el subsistema desarrollado, a continuación se ofrece algunos ejemplos de frases reconocidas:

Transcrita:
20051009_190753_218_fsp-A-005217-005475 porque nadie le entiende lo que esta diciendo
Reconocida:
20051009_190753_218_fsp-A-005217-005475 porque nadien entiendes lo que esta diciendo

Transcrita:
20051009_190753_218_fsp-A-007982-008700 y siempre andan la gente peleando lo que es
Reconocida:
0051009_190753_218_fsp-A-007982-008700 y siempre anda en la gente peleando de o lo que el

Como hemos podido observar apenas hay diferencias significativas entre ambas, especialmente en lo que a palabras largas (habitualmente las que aportan el mayor significado), a pesar en que difieran en algunas palabras. Esto es debido a que los errores más frecuentes cometidos por cualquier reconocedor son en el reconocimiento de palabras cortas. Este hecho se puede apreciar con claridad en el segundo ejemplo que se ofrece. En este segundo caso, que podría parecer elegido como un caso optimista, hay 2 sustituciones, 3 inserciones y 0 borrados en una frase con 9 palabras de referencia, lo que produce un WER del 5 / 9 x 100 = 55.56%, de modo que es un caso peor que la media de las frases reconocidas.

5.3 Resultados de Detección de Menciones

Los datos utilizados para realizar la evaluación de detección de menciones publicitarias son las menciones publicitarias extraídas de la base de datos ATVS-

Radio. Para realizar esta detección de menciones se ha empleado un sistema de búsqueda de palabras clave basado en el reconocedor de voz de gran vocabulario desarrollado. Para ello hemos seleccionado un total de 51 palabras claves[3] que aparecen en dichas menciones, se han añadido al léxico del reconocedor y se ha decidido que se detectaba una mención cuando se reconocía una de esas palabras clave. De esta forma hemos obtenido los resultados de la siguiente tabla.

5.3.1.1 Resultados de detección de menciones publicitarias

Modelo Fonético	% Acierto
MFCC_1	42.37
MFCC_2	45.76
+LDA + MLLT	74.57
+fMLLR + SAT	64.40
GMM	69.49
+bMMI + SGMM	67.79

Tabla 5-4 Resultados de detección de menciones publicitarias

Como podemos observar el mejor resultado que se ha obtenido con el modelo LDA +MLLT, lo que parece indicar que cuando trabajamos con una base de datos de características distintas no conviene tener un reconocedor de voz muy adaptado a la base de datos de entrenamiento. En vista de estos resultados, hemos tratado de mejorar más ese resultado variando el peso de las puntuaciones acústicas usado en la generación de los lattices, pero el resultado obtenido no ha sido satisfactorio, como podemos ver en la siguiente tabla (el peso de las puntuaciones acústicas inicial era 0.083).

Peso acústico	0.07	0.075	0.083	0.12	0.09
% Acierto	71.18	74.57	74.57	72.88	72.88

Tabla 5-5 Resultados detección de menciones publicitarias en función del peso de la puntuación acústica

[3] Ver Anexo C: Palabras clave asociada a menciones.

6 Conclusiones y trabajo futuro

En este TFG se ha realizado un sistema de búsqueda de palabras claves en castellano y se ha evaluado como aplicación la detección de menciones publicitarias. Por lo tanto ha sido necesario entrenar y evaluar un reconocedor de voz en español, para esto hemos hecho uso de la herramienta Kaldi, de la cual hemos podido comprobar su gran rendimiento para este tipo de desarrollos.

Haciendo uso de las bases de datos Fisher Spanish y Callhome Spanish hemos conseguido reducir la tasa de error de palabra aproximadamente en un 9% respecto al resultado obtenido por el grupo hasta el momento, consiguiendo un 41.10% sobre voz telefónica conversacional. A pesar de haber conseguido unos buenos resultados, todavía hay una diferencia de aproximadamente un 10% con los mejores resultados de Kaldi, esto puede ser debido a las bases de datos utilizadas, ya que al haber sido grabadas en EEUU contienen muchas palabras en inglés, de las cuales no tenemos una correcta transcripción fonética, mientras que los mejores resultados de Kaldi se han obtenido con un léxico revisado manualmente.

Por lo tanto una posible mejora podría ser la realización de una mejor transcripción fonética o directamente eliminar las palabras ingles del léxico, con esto conseguiremos un modelo más limpio ya que una mala transcripción puede generar distorsiones en los modelos acústicos. Y como hemos dicho, esto es muy significativo en este caso por el tipo de bases de datos utilizadas.

También se ha podido comprobar por los mejores resultados de Kaldi que la utilización de redes neuronales produce una mejora en el resultado obtenido, por lo tanto también podría ser una línea de trabajo futuro, ya que a pesar de haber comenzado a entrenar dicho modelo fonético por falta de tiempo no ha sido posible obtener un resultado.

Por ultimo para la evaluación de detección de menciones a través de palabras clave se ha usado la base datos creada en colaboración con otros estudiantes del grupo, ATVS-Radio. De ésta se han podido extraer 62 menciones publicitarias y usando un total de 51 palabras clave para detectar dichas menciones se ha conseguido detectar aproximadamente el 74% de menciones. Hay que destacar que este resultado se ha obtenido en fases muy finales del trabajo y que apenas ha habido tiempo de hacer ajustes en este sistema, por lo que esperamos que con algunos ajustes adicionales se consiga mejorar significativamente este resultado. También

por falta de tiempo han quedado fuera de la evaluación otro aspecto importante: evaluar la tasa de falsos positivos en la detección de menciones publicitarias.

Una posibilidad evidente de mejorar los resultados de detección sería realizando una adaptación del modelo de lenguaje, ya que con tan solo adaptar el léxico no es suficiente para obtener el mejor porcentaje de acierto posible.

Referencias

[1] NIST OpenKWS, website http://www.nist.gov/itl/iad/mig/openkws.cfm (accedida el día 8/05/16)

[2] MediaEval Query-by-Example, website http://www.multimediaeval.org/mediaeval2015/ (accedida el día 8/05/16)

[3] ALBAYZIN Search-on-Speech, webside http://iberspeech2014.ulpgc.es/index.php/albayzin/search-on-speech-evaluation (accedida el día 8/05/16)

[4] Juanchen Xu, "Adaptación de un sistema de búsqueda de palabras clave al castellano", Proyecto de Fin de Carrera, 2014, webside http://hdl.handle.net/10486/662527 (accedida el día 24/05/16)

[5] Müller, Meinard. "Dynamic time warping." Information retrieval for music and motion, Springer ,pp.69-84,2007

[6] J.Ortega García, "Apuntes de clase, Asignatura: Tratamiento de Señales de Voz y Audio: Sistema Auditivo, Sensación Sonora y Parametrización Perceptual", pp.54-58,2016.

[7] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice-Hall International,Inc ,pp.321-350, , 1993.

[8] Yann LeCun, Yoshua Bengio y Geoffrey Hinton,"Deep Learning", Nature , Vol 521,pp.436-444 ,Mayo 2015.

[9] Javier Tejedor, Doroteo Torre, José Colás, "Estado del arte en Wordspotting aplicado a los sistemas de extracción de información en contenidos de voz", en Proceedings del I Congreso Español de Recuperación de Información (CERI). pp.101-109 ,2010.

[10] Webside http://kaldi-asr.org/ (accedida el día 16/05/16)

[11] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. EPFL-CONF-192584. IEEE Signal Processing Society,pp.1-2, 2011.

[12] Linguistic Data Consortium, webside https://www.ldc.upenn.edu/ (accedida el día 18/05/16)

[13] Corpus Switchboard. Disponible en Linguistic Data Consortium con referencia LDC2010S01, webside https://catalog.ldc.upenn.edu/LDC2010S01 (accedida el día 16/05/16)

[14] Corpus Switchboard. Disponible en Linguistic Data Consortium con referencia LDC96S35, webside https://catalog.ldc.upenn.edu/LDC96S35 (accedida el día 16/05/16)

[15] M. Post, G. Kumar, A. López, D. Karakos, C. Callison-Burch, y S. Khudanpur, "Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus", International Workshop on Spoken Language Translation, pp.4,Diciembre 2013.

Glosario

bMMI	boosted Maximum Mutual Information
CMS	Cepstral Mean Subtraction
CMVN	Cepstral Mean and Variance Normalization
DCT	Discrete Cosine Transform
DNN	Deep Neural Network
DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
fMLLR	future-space Maximum Likelihood Linear Regression
GMM	Gaussian Mixture Models
HMM	Hidden Markov Model
KWS	Keyword Spotting
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
LVCSR	Large-Vocabulary Continuous Speech Recognition
MFCC	Mel Frequency Cepstral Coefficients
MLLT	Maximum Likelihood Linear Transform
MMI	Maximum Mutual Information
OOV	Out-Of-Vocabulary
QbE SDR	Query by Example Spoken Document Retrieval

QbE STD	Query by Example Spoken Term Detection
RNNs	Recurrent Neural Networks
SAT	Speaker Adaptive Training
SDT	Spoken Term Detection
SGMM	subspace Gaussian Mixture Models
SRILM	SRI Language Modeling
STT	Speech-To-Text
WER	Word Error Rate

Anexos

A Base de datos menciones

1. ME_ELCORTEINGLES_TECNOLOGIA_2015-05-25 (73 segundos)
2. ME_ELCORTEINGLES_SUPERMERCADOYBODAMAS_2015-05-26 (66 segundos)
3. ME_ELCORTEINGLES_SUMMERTIME_2015-05-27 (62 segundos)
4. ME_ELCORTEINGLES_SUMMERTIME_2015-05-28 (64 segundos)
5. ME_DIABALANCE_GELESGLUCOSA_2015-05-28 (39 segundos)
6. ME_ELCORTEINGLES_SUPERMERCADO_2015-05-29 (56 segundos)
7. ME_ELCORTEINGLES_SUPERMERCADO_2015-06-01 (51 segundos)
8. ME_ELCORTEINGLES_SUPERMERCADO_2015-06-02 (51 segundos)
9. ME_ELCORTEINGLES_OPERACIONVERANO_2015-06-03 (50 segundos)
10. ME_DIABALANCE_GELESGLUCOSA_2015-06-03 (29 segundos)
11. ME_ELCORTEINGLES_COCINAYBAÑOYBODAMAS_2015-06-04 (66 segundos)
12. ME_ELCORTEINGLES_OPERACIONVERANO_2015-06-05 (57 segundos)
13. ME_SEGURITASDIRECT_ALARMA_2015-05-25 (27 segundos)
14. ME_REVITAL_AMPOLLAS_2015-05-25 (29 segundos)
15. ME_DIABALANCE_GEL_2015-05-26 (36 segundos)
16. ME_RASTREATOR_COMPARADOR_2015-05-26 (23 segundos)
17. ME_SEGURITASDIRECT_ALARMA_2015-05-27 (27 segundos)
18. ME_DIABALANCE_GEL_2015-05-27 (32 segundos)
19. ME_ELCORTEINGLES_VIAJES_2015-05-28 (44 segundos)
20. ME_RASTRATOR_BUSCADOR_2015-05-28 (28 segundos)
21. ME_CORTEINGLESVIAJES_GRANCANARIA_2015-05-29 (39 segundos)
22. ME_DIABALANCE_GEL_2015-05-29 (31 segundos)
23. ME_SEGURITASDIRECT_ALARMA_2015-06-01 (24 segundos)
24. ME_REPSOL_2015-06-01 (3 segundos)
25. ME_DIABALANCE_GEL_2015-06-01 (27 segundos)
26. ME_ORANGE_JAZZTEL_2015-06-01 (29 segundos)
27. ME_LINEADIRECTA_SEGURODEHOGAR_2015-06-02 (29 segundos)
28. ME_RASTREATOR_COMPARADOR_2015-06-02 (27 segundos)
29. ME_SEGURITASDIRECT_ALARMA_2015-06-03 (22 segundos)
30. ME_DIABALANCE_GEL_2015-06-03 (32 segundos)
31. ME_LINEADIRECTA_SEGURODEHOGAR_2015-06-04 (29 segundos)
32. ME_SEGURITASDIRECT_ALARMA_2015-06-04 (27 segundos)
33. ME_BBVA_FINANCIACION_2015-06-05 (24 segundos)
34. ME_DIABALANCE_GEL_2015-06-05 (29 segundos)
35. ME_BIO3_2015-05-25 (4 segundos)
36. ME_BARCELOVIAJES_2015-05-25 (45 segundos)
37. ME_BIO3_2015-05-27 (5 segundos)
38. ME_BARCELO_2015-05-27 (30 segundos)
39. ME_BIO3_2015-05-28 (5 segundos)
40. ME_BARCELO_2015-05-28 (46 segundos)

41. ME_IBERDROLA_2015-05-28 (34 segundos)
42. ME_BIO3_2015-05-29 (4 segundos)
43. ME_BARCELO_2015-05-29 (39 segundos)
44. ME_ENDESA_2015-05-29 (3 segundos)
45. ME_BIO3_2015-06-01 (5 segundos)
46. ME_BARCELO_2015-06-01 (5 segundos)
47. ME_JAZZTEL_2015-06-01 (28 segundos)
48. ME_BIO3_2015-06-02 (4 segundos)
49. ME_BBVA_2015-06-02 (24 segundos)
50. ME_BBTRAVEL_2015-06-02 (31 segundos)
51. ME_BIO3_2015-06-03 (4 segundos)
52. ME_ORANGE_2015-06-03 (18 segundos)
53. ME_BBTRAVELBRAND_2015-06-03 (21 segundos)
54. ME_BIO3_2015-06-04 (6 segundos)
55. ME_BARCELO_2015-06-04 (29 segundos)
56. ME_ENDESA_2015-06-04 (3 segundos)
57. ME_BIO3_2015-06-05 (5 segundos)
58. ME_BARCELO_2015-06-05 (29 segundos)
59. ME_ORANGE_2015-06-05 (31 segundos)
60. ME_BIO3_2015-06-08 (5 segundos)
61. ME_BARCELO_2015-06-08 (43 segundos)
62. ME_ALFONSOX_2015-06-08 (17 segundos)

B Palabras clave de las menciones

1. BIO3
2. LABORATORIOS_BIO3
3. DIABALANCE
4. BAJON_DE_AZUCAR
5. GELES_DE_GLUCOSA
6. ELIGE_ESTAR_TRANQUILO
7. GLUCOSA
8. DIABETES
9. DIABETICO
10. BAJADA_DE_AZUCAR
11. ENDESA
12. IBERDROLA
13. JAZZTEL
14. ORANGE
15. ACCIONES
16. ACCIONISTA
17. BBVA
18. FINANCIACION
19. LINEA_DIRECTA
20. SEGURO_DE_HOGAR
21. COMPAÑÍA_BANKINDER
22. BARCELO_VIAJES
23. BB_TRAVEL_BRAND
24. VIAJERO
25. GEN_QUE_NOS_IMPULSA_A_VIAJAR
26. RASTREATOR
27. BUSCADOR
28. COMPARADOR_TOTAL
29. AHORRAR_TIEMPO_Y_DINERO
30. REPSOL
31. REVITAL_AMPOLLAS
32. SEGURITAS_DIRECT
33. ALARMA
34. ROBAR
35. AVISO_A_POLICIA
36. ROBO
37. VIAJES_EL_CORTEINGLES
38. GRAN_CANARIA
39. ALIMENTACION
40. YA_ES_VERANO
41. EL_CORTEINGLES
42. OPERACIÓN_VERANO

43. CAMPAMENTO_DE_VERANO
44. VERANO
45. SUMMERTIME
46. SUPERMERCADOS_EL_CORTEINGLES
47. BODA_MAS
48. COCINA
49. BAÑO
50. TECNOLOGIAS
51. ALFONSO_X

C Palabras clave asociada a menciones

Mención	Palabras Clave
ME_ELCORTEINGLES_TECNOLOGIA	TECNOLOGIAS EL_CORTEINGLES
ME_ELCORTEINGLES_SUPERMERCADOYBODAMAS	EL_CORTEINGLES SUPERMERCADOS_EL_CORTEINGLES BODA_MAS
ME_ELCORTEINGLES_SUMMERTIME	VERANO SUMMERTIME YA_ES_VERANO EL_CORTEINGLES
ME_DIABALANCE_GELESGLUCOSA	BAJON_DE_AZUCAR GELES_DE_GLUCOSA ELIGE_ESTAR_TRANQUILO GLUCOSA DIABETES DIABETICO BAJADA_DE_AZUCAR
ME_ELCORTEINGLES_SUPERMERCADO	EL_CORTEINGLES SUPERMERCADOS_EL_CORTEINGLES
ME_ELCORTEINGLES_OPERACIONVERANO	OPERACIÓN_VERANO CAMPAMENTO_DE_VERANO VERANO
ME_ELCORTEINGLES_COCINAYBAÑOYBODAMAS	EL_CORTEINGLES COCINA BAÑO BODA_MAS
ME_SEGURITASDIRECT_ALARMA	SEGURITAS_DIRECT ALARMA ROBAR AVISO_A_POLICIA ROBO
ME_REVITAL_AMPOLLAS	REVITAL_AMPOLLAS
ME_RASTREATOR	RASTREATOR BUSCADOR COMPARADOR_TOTAL AHORRAR_TIEMPO_Y_DINERO
ME_ELCORTEINGLES_VIAJES	EL_CORTEINGLES VIAJES_EL_CORTEINGLES
ME_CORTEINGLESVIAJES_GRANCANARIA	EL_CORTEINGLES VIAJES_EL_CORTEINGLES GRAN_CANARIA
ME_BBVA_FINANCIACION	BBVA FINANCIACION
ME_ORANGE_JAZZTEL **ME_JAZZTEL** **ME_ORANGE**	JAZZTEL ORANGE ACCIONES ACCIONISTA
ME_REPSOL	REPSOL
ME_ENDESA	ENDESA

LINEADIRECTA_SEGURODEHOGAR	LINEA_DIRECTA SEGURO_DE_HOGAR COMPAÑÍA_BANKINDER
ME_IBERDROLA	IBERDROLA
ME_BIO3	BIO3 LABORATORIOS_BIO3
ME_BARCELO **ME_BBTRAVEL** **ME_BARCELOVIAJES** **ME_BBTRAVELBRAND**	BARCELO_VIAJES BB_TRAVEL_BRAND GEN_QUE_NOS_IMPULSA_A_VIAJAR VIAJERO
ME_ALFONSOX	ALFONSO_X

Tabla Anexo C-1 Menciones con sus palabras clave