Universitat Politècnica de València

Departamento de Sistemas
Informáticos y Computación

Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Trabajo Final de Máster

LECTURA DE LABIOS MEDIANTE TÉCNICAS DE MACHINE LEARNING

Autor: David Gimeno Gómez
Tutor: Carlos David Martínez Hinarejos

Curso 2019/2020

Resumen

Durante una conversación nuestro cerebro se encarga de combinar información procedente de múltiples sentidos con el objetivo de mejorar nuestra capacidad a la hora de comprender el mensaje que estamos percibiendo. Diferentes estudios han demostrado la importancia que presenta la información visual en estas situaciones, así como su relación con los sonidos producidos. Como bien sabemos, la lectura de labios es una tarea compleja cuyo objetivo es interpretar el habla cuando el audio no se encuentra disponible. Al prescindir de un sentido tan crucial como es el oído, dado que esta señal presenta una mayor cantidad de información respecto al reconocimiento del habla, será necesario ser conscientes de los desafíos que presenta dicha carencia.

El propósito de este proyecto consiste en construir las bases de un sistema capaz de imitar la habilidad humana de interpretar el habla leyendo los labios del interlocutor. Desde un punto de vista más amplio, nuestra tarea no se distingue sustancialmente de otras como pueden ser el reconocimiento automático del habla a partir del audio o el reconocimiento de texto manuscrito. En otras palabras, nos encontramos bajo el marco de las tecnologías del lenguaje. Por ello, nuestra experimentación se fundamenta en torno a los avances realizados en este ámbito, explorando, en nuestro caso, un sistema basado en Modelos Ocultos de Markov Continuos. No obstante, el núcleo central sobre el que se apoya nuestro proyecto es el estudio y análisis de las diferentes características visuales que pueden representar la naturaleza de los movimientos labiales, por lo que se ha requerido el empleo de técnicas relacionadas con la Visión por Computador. Además, para poder llevar a cabo el proyecto ha sido necesario estudiar la literatura al respecto, así como obtener un conjunto de datos propicio, los cuales pertenecen a un subconjunto del corpus RTVE empleado en las evaluaciones Albayzín de Tecnologías del Habla.

Las aplicaciones de este tipo de sistemas abarcan desde la transcripción de películas mudas de la época (tanto enfocadas al entretenimiento como la documentación histórica), proporcionar apoyo al reconocimiento del habla acústica cuando la calidad del audio se encuentra deteriorada o corrupta, así como el empleo de contraseñas visuales silenciosas o incluso dar soporte a la síntesis de voz para personas con dificultades en el habla dependiendo de sus movimientos labiales.

Palabras Clave: Lectura de labios, *Machine Learning*, Tecnologías del Lenguaje, Visión por Computador, Modelos Ocultos de Markov, *Deep Learning*

Abstract

During a conversation, our brain is responsible for combining information obtained from multiple senses in order to improve our ability to understand the message we are perceiving. Different studies have shown the importance of presenting visual information in these situations, as well as its relationship with the sounds produced. As we know, lip reading is a complex task whose objective is to interpret speech when audio is not available. By dispensing with a sense as crucial as hearing, since this signal presents a greater amount of information regarding speech recognition, it will be necessary to be aware of the challenge that this lack presents.

The purpose of this project is to build the foundations of a system capable of imitating the human ability to interpret speech by reading the lips of the interlocutor. From a broader point of view, our task is not substantially different from others, such as automatic speech recognition from audio or handwritten text recognition. In other words, we are under the framework of language technologies. Therefore, our experimentation is based on the advances made in this area, exploring, in our case, a system based on Continuous Hidden Markov Models. However, the central core on which our project is based is the study and analysis of the different visual characteristics that may represent the nature of lip movements. Consequently, the use of techniques related to Computer Vision has been required. In addition, in order to carry out the project, it has been necessary to study the literature on this topic, as well as to obtain a suitable data set, which belongs to a subset of the RTVE corpus, used in the Albayzín evaluations of Speech Technologies.

The applications of this type of systems range from the transcription of ancient silent films (both focused on entertainment and historical documentation), to provide support for acoustic speech recognition when audio quality is impaired or corrupted, apart from the use of silent visual passwords or even support speech synthesis for people with speech difficulties depending on their lip movements.

Keywords: Lipreading, Machine Learning, Speech Technologies, Computer Vision, Hidden Markov Models, Deep Learning

Resum

Durant una conversació el nostre cervell s'encarrega de combinar informació procedent de múltiples sentits amb l'objectiu de millorar la nostra capacitat a l'hora de comprendre el missatge que estem percebent. Diferents estudis han demostrat la importància que presenta la informació visual en aquestes situacions, així com la seua relació amb els sons produïts. Com bé sabem, la lectura de llavis és una tasca complexa on l'objectiu és interpretar la parla quan l'àudio no està disponible. Al prescindir d'un sentit tan crucial com és l'oïda, ja que aquest senyal presenta una major quantitat d'informació respecte al reconeixement de la parla, caldrà ser conscients dels reptes que presenta aquesta carència.

El propòsit d'aquest projecte consisteix a construir les bases d'un sistema capaç d'imitar l'habilitat humana d'interpretar la parla llegint els llavis de l'interlocutor. Des d'un punt de vista més ampli, la nostra tasca no es distingeix substancialment d'altres com poden ser el reconeixement automàtic de la parla a partir de l'àudio o el reconeixement de text manuscrit. En altres paraules, ens trobem davall el marc de les tecnologies del llenguatge. Per això, la nostra experimentació es fonamenta al voltant dels avanços realitzats en aquest àmbit, explorant, en el nostre cas, un sistema basat en Models Ocults de Markov Continus. No obstant això, el nucli central sobre el qual es recolza el nostre projecte és l'estudi i anàlisi de les diferents característiques visuals que poden representar la naturalesa dels moviments labials, cosa per la qual s'ha requerit l'ús de tècniques relacionades amb la Visió per Computador. A més, per poder dur a terme el projecte ha sigut necessari estudiar la literatura al respecte, així com obtenir un conjunt de dades propici, els quals pertanyen a un subconjunt del corpus RTVE, emprat en les avaluacions Albayzín de Tecnologies de la Parla.

Les aplicacions d'aquest tipus de sistemes abasten des de la transcripció de pel·lícules mudes de l'època (tant enfocades a l'entreteniment com la documentació històrica), donar suport al reconeixement de la parla acústica quan la qualitat de l'àudio es troba deteriorada o corrupta, a més de l'ús de contrasenyes visuals silencioses o fins i tot donar suport a la síntesi de veu per a persones amb dificultats en la parla depenent dels seus moviments labials.

Paraules Clau: Lectura de llavis, *Machine Learning*, Tecnologies del Llenguatge, Visió per Computador, Models Ocults de Markov, *Deep Learning*

Índice de Contenido

Índice de Figuras

Índice de Tablas

1 | Introducción

Durante una conversación nuestro cerebro combina la información percibida desde múltiples sentidos con el propósito de mejorar nuestra capacidad a la hora de comprender el mensaje que está emitiendo el interlocutor. De hecho, diferentes estudios han demostrado la influencia que presenta la señal visual sobre la percepción del habla. Principalmente, hablamos de los estudios realizados por McGurk y McDonald [56], donde demostraron que si la expresión bucal no se correspondía con el audio emitido, el oyente se confundía, percibiendo un sonido diferente al que realmente fue. No obstante, como bien sabemos, la lectura de labios es una tarea compleja, cuyo objetivo es interpretar el habla cuando el audio no se encuentra disponible, es decir, considerando únicamente la información procedente del canal visual. Al prescindir de un sentido tan crucial como es el oído, será necesario ser conscientes de los desafíos que presenta dicha carencia.

El propósito ideal de este proyecto consiste en construir las bases de un sistema capaz de imitar la habilidad humana de interpretar el habla natural leyendo los labios del interlocutor. Para ello, emplearemos únicamente la información contenida en la señal visual, donde podemos encontrar diferentes aspectos que pueden facilitar la percepción y comprensión de lo que el o la hablante está diciendo. Estamos hablando de un proceso en el que el o la oyente debe prestar atención tanto al movimiento de los labios como a la postura que adoptan la lengua y los dientes. Incluso existen otros factores de gran relevancia, como la gesticulación del emisor o, sin duda, el contexto en el que se desenvuelve la conversación. A pesar de ello, tal y como se ha sugerido previamente, esta tarea plantea un gran desafío.

No obstante, antes de abordar estos detalles debemos conocer dos conceptos fundamentales: el fonema y el visema [82, 28]. El primero de ellos se define como la unidad mínima de sonido con la que podemos distinguir una palabra de otra dentro de un lenguaje, mientras que el visema se asocia a la representación visual de un fonema, en concreto, mediante las expresiones faciales. Desafortunadamente, no existe una correspondencia directa entre ambos. Es por ello que, al tratar de resolver esta tarea, nos enfrentamos a distintos problemas que entorpecen la comprensión del discurso. Entre ellos, destacamos la ambigüedad visual, ya que en multitud de ocasiones podemos observar cómo diferentes sonidos se corresponden con un mismo visema, como puede ser el caso de los fonemas /p/, /b/ y /m/. Por otro lado, además, identificamos fonemas que son producidos desde la garganta o que implican movientos de la lengua que no pueden ser vistos e interpretados por el receptor del mensaje. Sumado a esto, habría que tener en cuenta las condiciones en las que fue grabada la locución, así como los cambios en iluminación o la variedad que presentan las diferentes personas respecto a la fisionomía bucal.

Desde un punto de vista más amplio, nuestra tarea no se distingue sustancialmente de otras como pueden ser el reconocimiento automático del habla a partir del audio [30] o el reconocimiento de texto manuscrito [69]. En otras palabras, nos encontramos bajo el marco de las Tecnologías del Lenguaje. Por ello, nuestra experimentación se fundamenta en torno a los avances realizados en este ámbito. Concretamente, centraremos nuestra experimentación sobre una aproximación tradicional, es decir, sobre un sistema basado en Modelos Ocultos de Markov combinados con Modelos de Mixuturas de Gaussianas (GMM-HMM, por sus siglas en inglés) [30]. Esta decisión supone un estudio de gran interés, puesto que a lo largo de la literatura la interpretación del habla continua y espóntanea se ha tratado, en reducidas ocasiones, desde el paradigma tradicional.

Por otro lado, para poder hacer frente al tratamiento del habla natural en el idioma español, recopilaremos nuestro propio conjunto de datos. Estos datos tendrán su origen en el *corpus* RTVE[1], donde disponemos de una gran cantidad de material audiovisual procedente de múltiples programas emitidos por la cadena pública. Se trata de un *corpus* de reconocido prestigio, ya que ha sido empleado en las evaluaciones Albayzín.

No obstante, el núcleo central sobre el que gravita nuestro proyecto es el estudio y análisis de las características visuales, donde evaluaremos características relacionadas tanto con la geometría como con la apariencia bucal, explorando técnicas clásicas [54] así como aproximaciones *Deep Learning* [29]. De este modo, estudiando cómo se comportan indivualmente o combinadas entre sí, podremos determinar qué características logran reprensentar con mayor o menor éxito la naturaleza de los movimientos labiales. Para poder llevar a cabo este análisis es necesario tratar con las imágenes directamente. Por ello, este proyecto integra técnicas de *Machine Learning* pertencientes a dos grandes ámbitos en la materia: las Tecnologías del Lenguaje y la Visión por Computador.

Por último, las aplicaciones de este tipo de sistemas abarcan desde la transcripción de películas mudas de la época, tanto enfocadas al entretenimiento como a la documentación histórica, como a cimentar el reconocimiento del habla acústica cuando la calidad del audio se encuentra deteriorada o corrupta, al empleo de contraseñas visuales silenciosas o incluso a dar soporte a la síntesis de voz para personas con dificultades en el habla dependiendo de sus movimientos labiales.

1.1. Motivación

El motivo principal de este trabajo fin de máster es la creciente importancia que está adquiriendo la inteligencia artificial, tanto a nivel laboral como social. Además, resulta de gran interés contemplar cómo los fundamentos probabilísticos en los que se basan estas técnicas de *Machine Learning* pretenden simular, en ciertas ocasiones, el comportamiento del cerebro humano. Por otro lado, este ámbito conlleva la participación en proyectos que involucran equipos de trabajo multidisciplinares. Es por ello que la inteligencia artificial abarca un amplio espectro de aplicaciones, y todo apunta a que va a ser partícipe del desarrollo e innovación de nuestra sociedad.

Centrando la atención sobre la lectura de labios automática, una de las razones esenciales que han impulsado este proyecto es que la tarea continúe siendo un problema abierto a la investigación. No obstante, en igualdad de condiciones, se presenta la ayuda que supondría un sistema capaz de leer los labios a aquellas personas que padezcan problemas de audición. Aparte, este tipo de sistemas, o más bien la investigación respecto a las características visuales, puede ser útil cuando tratamos de reconocer el habla en situaciones donde el audio se encuentra deteriorado. De este modo, respaldamos la construcción de sistemas capaces de integrar varias fuentes de información.

[1] http://catedrartve.unizar.es/rtvedatabase.html

1.2. Objetivos

Nuestro objetivo ideal sería alcanzar, mediante un sistema tradicional y teniendo en cuenta el actual estado del arte, resultados aceptables en cuanto al reconocimiento automático de la lectura de labios. En otras palabras, un sistema preliminar con el que, posteriormente, poder continuar la investigación en futuros proyectos. No obstante, para poder lograr este propósito, es necesario conseguir otros objetivos que permitan guiar el desarrollo del proyecto en su totalidad. Entre ellos destacamos los siguientes:

- Adquirir experiencia con todo el proceso que conlleva la construcción completa de un sistema basado en técnicas de *Machine Learning*.
- Construir un *corpus*, de considerable tamaño, enfocado a la interpretación del habla continua para el español.
- Estudiar y analizar diferentes aproximaciones a la hora de extraer las características visuales.

1.3. Estructura

Respecto a la estructura sobre la que se apoya nuestro proyecto, cabe destacar que, tras la introducción presente en este capítulo, nos adentraremos en el estado del arte en cuanto a la lectura de labios automática, describiendo ya no sólo las distintas aproximaciones a la hora de construir el sistema final, sino también cuáles son los conjuntos de datos audiovisuales más relevantes. Todo estos detalles quedan encapsulados en el **Capítulo 2**, a partir del cual podremos guiar diferentes vertientes del desarrollo de nuestro trabajo. Posteriormente, el **Capítulo 3** engloba los aspectos relacionados con el *corpus* recopilado, especialmente para cumplir los propósitos planteados de cara a la interpretación del habla espóntanea. Por otro lado, el proceso que ha permitido detectar la región bucal a partir de las imágenes de vídeo es descrito en el **Capítulo 4**. En cuanto al **Capítulo 5**, se incluyen las distintas alternativas, respecto a la extracción de características visuales, que se estudian a lo largo del proyecto. Tras comprender los diferentes tipos de características, el **Capítulo 6** abarca los fundamentos teóricos en los que se basa la construcción del sistema automático. Con todo esto, encauzamos la redacción hacia los experimentos realizados, cuyos análisis y reflexiones quedan expuestos en el **Capítulo 7**. De este modo, en último lugar, el **Capítulo 8** muestra las conclusiones extraídas a nivel general, donde, además, se exponen los problemas identificados respecto al proyecto elaborado, así como las vías para futuros desarrollos.

2 | Estado del arte

Antes de abordar el proyecto, es necesario tener una cierta noción del contexto histórico en el que se han ido desarrollando los Sistemas Automáticos para la Lectura de Labios, tomando como fuente principal el artículo publicado por Fernández-López y M. Sukno [27]. No obstante, con el objetivo de comprender mejor los distintos aspectos presentes en este y sucesivos capítulos de la memoria, comenzaremos con una breve introducción al Aprendizaje Automático, también conocido como *Machine Learning*. Tras esta sección de carácter general, nos centraremos en describir la evolución que han sufrido los sistemas mencionados previamente, ya sea comentando los diversos conjuntos de datos empleados, las diferentes técnicas exploradas a la hora de extraer características visuales o las distintas aproximaciones adoptadas en la construcción del sistema, detallando las prestaciones de éstas en base al *Word Recognition Rate* (WRR). De esta forma, revisando la literatura al respecto, podremos comprender mejor el entorno o estado en el que nos encontramos, así como justificar las decisiones que hemos tomado a lo largo del proyecto.

2.1. Introducción al Aprendizaje Automático

La Inteligencia Artificial es una de las principales ramas que pertenecen al ámbito de las Ciencias de la Computación. Presenta sus inicios alrededor de los años cincuenta, consolidando sus fundamentos ya no solo en torno a los avances desarrollados en materias como la estadística, la matemática lógica y el procesamiento de datos, sino también a los avances relacionados con el *hardware* que han permitido el cómputo eficiente de estos algoritmos. Dentro de esta rama podemos distinguir numerosas disciplinas, como la Planificación Inteligente [32], los Sistemas Multi-Agente [87] y el Reconocimiento de Formas [8]. Todas ellas proporcionan el soporte de un amplio abanico de aplicaciones con un gran impacto tanto a nivel económico como social. Este tipo de técnicas, ya sea en combinación o de forma aislada, se conocen hoy en día como *Machine Learning* [8]. Su objetivo consiste en desarrollar técnicas que permitan a los computadores aprender, tomar decisiones o identificar patrones a partir de un conjunto de datos y con la mínima intervención humana posible.

En esta sección nos centraremos en el campo del Reconocimiento de Formas, puesto que éste se ocupa de aspectos más perceptivos como puede ser el reconocimiento facial, la detección de objetos o la interpretación del habla. En términos generales, el esquema de este tipo de sistemas quedaría reflejado en la Figura 2.1, donde se plasman los diferentes módulos o etapas que lo componen. En el primer módulo debemos capturar, mediante un sensor adecuado, el objeto principal que integre la tarea que deseamos resolver (en nuestro caso, el objeto capturado sería el plano del locutor y el sensor una cámara de vídeo). Tras esto, obtenemos una representación que pueda ser procesada por compu-

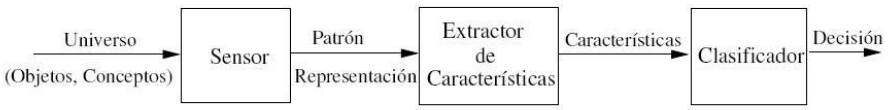

Figura 2.1: Esquema General de un Sistema de Reconocimiento de Formas

tadores. Sin embargo, con el objetivo de facilitar el aprendizaje y mejorar las predicciones del clasificador, en ocasiones es necesario extraer una serie de características a partir de esta representación, logrando así aumentar su capacidad discriminativa. Es entonces, con los datos apropiadamente procesados, cuando nos disponemos a emplear nuestro clasificador. Este último módulo será el encargado de proporcionar, de un modo automático, una solución a la tarea que estemos tratando.

Existen numerosos formalismos en los que podemos basar la construcción de nuestro clasificador, y la decisión de cuál debemos emplear dependerá de la tarea que deseemos resolver, así como de la naturaleza de nuestros datos. Normalmente, estos clasificadores o modelos se fundamentan en la teoría de la decisión estadística [8], siendo al final un conjunto de parámetros y probabilidades que se van estimando a medida que se procesa el conjunto de datos. Esta etapa se conoce como entrenamiento o aprendizaje, el cual está dirigido por una función objetivo (directamente relacionada con los propósitos de la tarea en cuestión) que le indica cómo corregir los errores que esté cometiendo y, de esta forma, devolver predicciones de mayor calidad. Se trata, entonces, de un aprendizaje inductivo, puesto que al inicio el sistema dispone de un escaso conocimiento respecto al problema a resolver y es mediante la observación de ejemplos como logra aprender y generalizar la tarea. Es necesario destacar que dicho sistema no es programado para realizar dicha tarea, tal y como se ha podido deducir previamente. Principalmente, se distinguen dos tipos de aprendizaje dependiendo de cómo se se encuentre construido el conjunto de datos:

- Aprendizaje supervisado: se relaciona con aquellos casos en los que el conjunto de entrenamiento dispone de la información completa, es decir, tanto de los datos de entrada como los de salida (resultado que debería predecir el sistema). Por ejemplo, podría entrenarse un sistema para que pudiera discernir si un correo electrónico es *spam* o no. Para ello, cada muestra de aprendizaje consistiría en una tupla, donde la entrada sería el cuerpo del correo y la salida un indicador de si el correo en cuestión es o no *spam*.

- Aprendizaje no supervisado: cuando los datos de entrenamiento solo disponen de la información de entrada. A diferencia del caso anterior, el sistema no conoce qué información es satisfactoria o no para el objetivo del aprendizaje. Debido a esto, deberá extraer patrones que le permitan agrupar los datos en función de sus atributos. Sin embargo, este tipo de aprendizaje requiere de la interpretación de un ser humano para darle utilidad. Un ejemplo sería proporcionar imágenes para que el sistema aprendiese a clasificarlas según si es un coche, un perro o un balón.

Por otro lado, existen dos vertientes dentro del *Machine Learning*. Dependiendo del dominio en el que se defina la tarea a resolver, distinguimos sistemas enfocados en:

- Clasificación: los datos de entrada pertenecerán a un dominio arbitrario pero los de salida serán de un conjunto finito, generalmente pequeño, de C elementos denominados clases. De esta manera, una vez el sistema se encuentre entrenado, proveerá al usuario una clasificación de la nueva entrada que le proporcione.

- Regresión: en esta ocasión, tanto los datos de entrada como los de salida pertenecen a dominios arbitrarios. Típicamente, en el dominio de los números reales. A modo de ejemplo, podría entrenarse un sistema proporcionándole un gran conjunto de muestras relacionadas con el nivel de agua de un pantano. De esta forma, conseguimos un sistema capaz de realizar una predicción sobre el nivel de agua que puede alcanzarse.

- Predicción Estructurada: a diferencia de las anteriores categorías, se trataría de un sistema encargado de generar una salida de una mayor complejidad, al estar compuesta por diferentes unidades que presentan, además, cierta relación entre ellas. En este ámbito, encontramos aplicaciones como el etiquetado sintáctico de una oración, la traducción automática o el reconocimiento de texto manuscrito.

Una vez comprendidas las líneas generales en las que se respalda el *Machine Learning*, podemos conocer mejor nuestro problema. Por ello, respecto a la lectura de labios, sabemos que se trataría de un sistema basado en aprendizaje supervisado, puesto que disponemos tanto de la entrada visual codificada como de las transcripciones asociadas al mensaje del locutor. Todos los detalles relacionados con el *corpus* construido se describen en secciones posteriores. Por otro lado, nuestro sistema se englobaría en lo que denominamos Predicción Estructurada, ya que su finalidad es interpretar una serie de características visuales que le permitan determinar la secuencia de palabras correspondientes.

Por último, indicar que esta sección constituye una breve introducción. Si se desea indagar más en el ámbito del Reconocimiento de Formas y el *Machine Learning* se recomienda notablemente el libro publicado por Bishop [8], donde podemos encontrar una explicación más detallada y completa respecto a los fundamentos de este tipo de técnicas.

2.2. Conjuntos de Datos

Como hemos constatado anteriormente, los datos son un elemento imprescindible en cualquier sistema desarrollado en el ámbito del *Machine Learning*. Por ello, antes de describir las diferentes aproximaciones estudiadas en la literatura (en cuanto a la lectura de labios se refiere), daremos a conocer varios *datasets*. Algunos de ellos, son ampliamente conocidos por su aplicación en el Reconomicimiento Automático del Habla (RAH) a partir del audio, puesto que, normalmente, este tipo de *corpus* se construyen en formato audiovisual. Varios de los *corpus* comentados en este apartado pueden observarse en la Figura 2.2.

Los aspectos que diferencian un conjunto de datos de otro van a determinar la complejidad y los detalles a tener en cuenta a la hora de construir nuestro modelo. Entre ellos, destacaremos el número de hablantes o *speakers* presentes, la resolución con la que fue grabado, la cantidad en horas que lo conforman o el tamaño del vocabulario. Por ello, esta sección se divide en diversos apartados con el objetivo de reflejar, de una forma más adecuada, cómo estos conjuntos de datos han ido evolucionando desde el planteamiento de tareas sencillas, como podría ser reconocer dígitos aislados, hasta alcanzar y proporcionar el soporte ante escenarios más realistas, tratando ya el reconocimiento del habla espóntánea. Esta evolución, tal y como podremos comprobar más adelante, encauza el desarrollo e investigación de los sistemas enfocados a la lectura de labios.

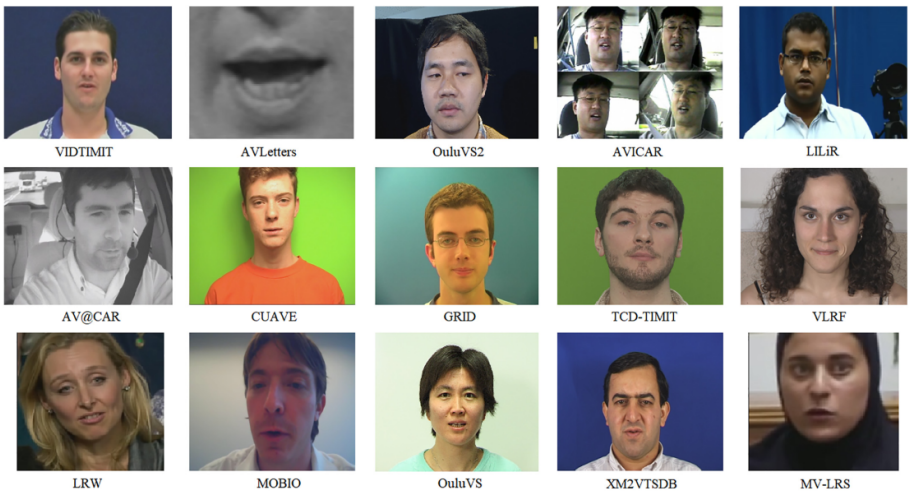

Figura 2.2: *Corpus* Audiovisuales destacados [27]

2.2.1. Reconocimiento de Alfabeto y Dígitos

Los primeros *corpora* surgen en la década de los noventa y fueron diseñados para tareas específicas y simples como es el reconocimiento del alfabeto y dígitos aislados, donde encontramos un vocabulario reducido. Este tipo de *datasets* han sido muy populares, ya que han permitido tratar el recocimiento del habla bajo un entorno controlado, siendo de gran utilidad para analizar la capacidad de los algoritmos en etapas tempranas de su desarrollo. En la Tabla 2.1 se muestra una lista con los *datasets* más populares dentro de este ámbito, detallando ciertos aspectos interesantes.

Tabla 2.1: *Corpus* Audiovisuales para el reconocimiento del alfabeto y dígitos

Nombre	Año	Lengua	*Speakers*	Tarea	Clases	Resolución	Duración
AVLetters [54]	1998	Inglés	10	Alfabeto	26	376x288, 25 fps	13 min
XM2VTS [57]	1999	Inglés	295	Dígitos	10	720x576, 25 fps	59 min
BANCA [4]	2003	Múltiple	208	Dígitos	10	720x576, 25 fps	∼14 h
AVOZES [33]	2004	Inglés	20	Dígitos	10	720x480, 30 fps	∼2 h
AVICAR [49]	2004	Inglés	86	Alfabeto Dígitos	26 13	720x480, 30 fps	∼33 h
AV@CAR [60]	2004	Español	20	Alfabeto Dígitos	26 10	768x576, 25 fps	∼1 h 50 min
CUAVE [66]	2004	Inglés	36	Dígitos	10	720x480, 30 fps	14 min
AVLetters2 [18]	2008	Inglés	5	Alfabeto	26	1920x1080, 50 fps	15 min
AV Digits [68]	2018	Inglés	53	Dígitos	10	1280x780, 30 fps	-

h: horas, min: minutos

En cuanto a los *datasets* más emblemáticos, destacan AVLetters (1998) [54] y CUAVE [66] para el reconocimiento del alfabeto y dígitos, respectivamente. No obstante, tal y como se observa en la Tabla 2.1, otros *corpus* han mejorado ciertas debilidades respecto a éstos como puede ser: mejorar la resolución o aumentar el número de hablantes. Por otro lado, destacamos el *corpus* AV@CAR [60], ya que está enfocado al reconocimiento del alfabeto y los dígitos en español.

2.2.2. Reconocimiento de Palabras y Oraciones

El vocabulario limitado y la poca cantidad de datos que caracteriza los *datasets* comentados anteriormente no permite la construcción de sistemas robustos y enfocados a aplicaciones realistas. En otras palabras, para alcanzar dichos objetivos se requiere alimentar al sistema con un *corpus* de gran envergadura, capaz de representar y capturar una amplia gama de variabilidad que proporcione al sistema la posibilidad de reconocer, en última instancia, el habla espontánea. Esta dinámica, además, habría sido impulsada por el auge de las arquitecturas basadas en *Deep Learning*, puesto que este tipo de aproximaciones necesita generalizar y estimar adecuamente un gran número de parámetros. En consecuencia, tal y como se demuestra más adelante y ha ocurrido en otros muchos ámbitos, estas arquitecturas han proporcionado significativos avances en el campo. De este modo, se produjo el desarrollo de *datasets* ya no sólo dirigidos a la clasificación de palabras aisladas, sino también al reconocimiento de oraciones, tal y como refleja la Tabla 2.2. Es necesario indicar que algunos de ellos, como AV@CAR [60] y AV Digits [68], ya estaban presentes en el reconocimiento del alfabeto o dígitos.

Tabla 2.2: *Corpus* Audiovisuales para el reconocimiento a nivel de oración

Nombre	Año	Lengua	*Speakers*	Tarea	Clases	Resolución	Duración
VIDTIMIT [75]	2002	Inglés	43	Oraciones	346	512x384, 25 fps	30 min
AV@CAR [60]	2004	Español	20	Oraciones	250	768x576, 25 fps	∼8 h
AVICAR [49]	2004	Inglés	86	Oraciones	1317*	720x480, 30 fps	∼33 h
GRID [15]	2006	Inglés	34	Frases	51	720x576, 25 fps	∼28 h
OuluVS [89]	2009	Inglés	20	Frases	10	720x576, 25 fps	16 min
MOBIO [55]	2012	Inglés	150	Oraciones	-	640x480, 16 fps	61 h
MODALITY [19]	2015	Inglés	35	Palabras	182	1920x1080, 100 fps	-
RM-3000 [39]	2015	Inglés	1	Oraciones	1000*	360x640, 60 fps	∼4 h
TCD-TIMIT [36]	2015	Inglés	62	Oraciones	5954	1920x1080, 30 fps	∼6 h
LRW [14]	2016	Inglés	>1000	Palabras	500	256x256, 25 fps	∼111h
VLRF [26]	2017	Español	24	Oraciones	1374*	1280x720, 50 fps	∼3 h
LRS [12]	2017	Inglés	>1000	Oraciones	17.428*	160x160, 25 fps	∼33 h

h: horas, min: minutos
* Número de palabras distintas. En otro caso, se refiere al número de oraciones o frases predefinidas

Por otro lado, la mayoría de estos *corpus* tratan con un conjunto de oraciones predefinidas, es decir, se centrarían en lo que se conoce como una clasificación a nivel de oración en lugar de interpretar el habla espontánea. Debido a que este último caso es el objetivo de nuestro proyecto, centramos la atención sobre el *dataset Lip-Reading Sentences in the Wild* (LRS) [12], construido a partir de múltiples programas emitidos entre 2010 y 2016 por la cadena BBC. Estamos hablando de un *corpus* con un vocabulario de 17.428 palabras diferentes, combinadas en 118.116 locuciones con el correspondiente plano frontal del *speaker*. Sin duda, las 33 horas de planos recogidos sin ningún tipo de restricción, consolidan este *dataset* como el mejor candidato para el desarrollo de sistemas de aplicación real. De hecho, tal y como se introdujo al inicio de esta memoria, uno de nuestros objetivos a largo plazo sería elaborar un *corpus* similar para el español. En relación con este último detalle, cabe destacar el *corpus* VLRF [26] para el español, donde disponemos de alrededor de 3 horas de vídeos con 1374 palabras distintas. Sin embargo, no cumple los propósitos mencionados previamente, ya que las escenas fueron tomadas en condiciones controladas y asegurando que los interlocutores se esforzaran por vocalizar de forma adecuada y expresiva.

2.2.3. Reconocimiento Multi-Vista

A pesar de que los *datasets* audivisuales han sido, principalemente, grabando al locutor desde un plano frontal, un sistema dedicado a la lectura de labios automática podría beneficiarse si dispone de conjuntos de datos donde cada una de las muestras ha sido filmada desde distintos puntos de vista. Esto proporcionaría robustez frente a múltiples escenarios, ya que no podemos asegurar que el locutor se encuentre siempre frontal a la cámara una vez despleguemos nuestro sistema. Además, se han realizado estudios donde se demuestra que si el o la hablante se encuentra en un ángulo ligeramente distante al de un plano frontal, la identificación del relieve y las protuberancias labiales resulta más sencillo, facilitando de esta forma el reconocimiento [47]. Los autores de esta publicación, emplearon en sus experimentos el *corpus* LILiR [48], puesto que provee ya no sólo un amplio rango de perspectivas, sino de una clasificación a nivel de oración.

Tabla 2.3: *Corpus* Audiovisuales Multi-Vista

Nombre	Año	Lengua	*Speakers*	Tarea	Clases	Resolución	Duración	Perspectiva (°)
CUAVE [66]	2004	Inglés	36	Dígitos	346	512x384, 25 fps	14 min	-90, 0, 90
AVICAR [49]	2004	Inglés	86	Oraciones	1317*	720x480, 30 fps	∼33 h	Variable (4 vistas)
CMU AVPFV [45]	2007	Inglés	10	Palabras	150	640x480, 30 fps	-	0, 90
QuLips [65]	2010	Inglés	2	Dígitos	10	720x576, 25 fps	-	0, 10, 20, ..., 90
LILiR [48]	2010	Inglés	12	Oraciones	200	720x576, 25 fps	-	0, 30, 45, 60, 90
LTS5 [25]	2011	Francés	20	Dígitos	10	1920x1080, 25 fps	-	0, 30, 60, 90
OuluVS2 [2]	2015	Inglés	53	Frases	540	1920x1080, 30 fps	∼1 h	0, 30, 45, 60, 90
TCD-TIMIT [36]	2015	Inglés	62	Oraciones	5954	1920x1080, 30 fps	∼6 h	0, 30
MV-LRS [13]	2017	Inglés	>1000	Oraciones	14.960*	160x1060, 25 fps	∼20 h	Entre 0 y 90

h: horas, min: minutos
* Número de palabras distintas. En otro caso, se refiere al número de oraciones o frases predefinidas

Por lo tanto, y aunque este aspecto haya sido solventado en cierta medida por *corpora* recogidos sin ningún tipo de restricción (como pueden ser LRW [14] o LRS [12]), se han desarrollado numerosos esfuerzos al respecto. De entre ellos, destacamos aquellos *corpus* reflejados en la Tabla 2.3. De hecho, los mismos autores de los *datasets* que acabamos de mencionar son los creadores del *corpus* MV-LRS [13], el cual presenta la misma filosofía. Debido a esto, en él podemos encontrar cualquier perspectiva del hablante, ya que, a diferencia de sus predecesores, engloba programas de debate donde intervienen múltiples interlocutores.

2.3. Extracción de Características Visuales

Una vez hayamos obtenido los datos, el siguiente paso es determinar cómo vamos a representar el movimiento labial producido en cada muestra de nuestro *corpus*. Esta representación debe ser robusta frente a diferentes aspectos que pueden distorsionar el posterior aprendizaje de nuestro modelo. Estamos hablando de aspectos ya mencionados en la introducción de la memoria, como pueden ser los cambios en la iluminación o en la postura que adopte el interlocutor. Además, la secuencia de características que obtengamos debe reflejar una connotación temporal, de forma que el mensaje decodificado a partir de esta representación sea coherente.

Al sumergirnos en la literatura, podemos contemplar un amplio abanico de técnicas que han sido empleadas en la extracción de estas características visuales.

No obstante, integrando las apreciaciones que realizan numerosos autores [54, 51], podemos englobar estas técnicas en las siguientes categorías:

- Características Geométricas: en esta aproximación solamente se consideraría la forma que adopta la boca del interlocutor a lo largo del discurso. En definitiva, consiste en extraer una serie de puntos de referencia (*landmarks*) que delimiten el contorno exterior e interior de los labios. Estas características permiten calcular métricas de gran interés, como pueden ser la altura, anchura y área bucal en un determinado instante de tiempo. Esto ha sido posible gracias a la contribución de modelos estadísticos conocidos como *Active Shape Models* (ASM) [17].

- Basadas en la Apariencia: en este caso, las características son extraídas aprovechando la información visual contenida en los píxeles de la región bucal. Diversos autores han estudiado la aplicación de técnicas conocidas en el ámbito de la Visión por Computador, como pueden ser *Principal Component Analysis* (PCA) [86], *Discrete Cosine Transform* (DCT) [1] o *Scale-Invariant Feature Transform* (SIFT) [52]. Entre ellas, cabe destacar los conocidos *Active Appearance Models* (AAM) [16] que podrían definirse como una versión mejorada de los ASM en la que además de identificar los *landmarks*, se proporcionaría información respecto a la apariencia.

- Basadas en el Movimiento: hace referencia a la técnica *Optical Flow* [78], capaz de capturar los movimientos realizados entre dos *frames* consecutivos en función de los cambios en la intensidad que se hayan producido. Podemos hacernos una idea a partir de la Figura 2.3, donde los autores del artículo previamente citado lograban capturar los movimietos labiales.

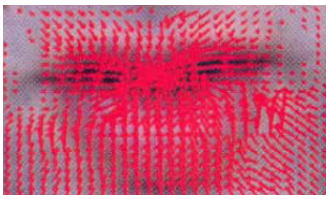

Figura 2.3: Características Basadas en el Movimiento [11]

- Basadas en *Deep Learning*: este tipo de aproximaciones consiste en delegar, normalmente sobre Redes Neronales Convolucionales [44], la responsabilidad de extraer características robustas y de un alto nivel de abstracción. Es común en diversos artículos el empleo de *Autoencoders* [6], arquitectura que se detalla en secciones posteriores de la memoria.

- Híbridas: aquellas características resultantes de la combinación de varias de las técnicas anteriormente descritas. Un caso típico es unir las características geométricas con las de apariencia.

Por otro lado, dentro del ámbito *Deep Learning*, esta etapa de extracción de características no existe como tal, sino que se encontraría integrada en las denominadas arquitecturas *end-to-end*, donde un único sistema trata de predecir la transcripción a partir de las imágenes directamente. Normalmente, tal y

como se observa en el apartado 2.4.2, en estas aproximaciones se emplean Redes Neuronales Convolucionales.

En último lugar, cabe destacar que este proyecto centra gran parte de su atención en el estudio de diversas técnicas para la extracción de características. Es por ello que este apartado se halla focalizado en esbozar, sin entrar en detalles, las distintas vertientes presentes en la literatura. Por estas razones, en el Capítulo 5 se describen, de una forma más completa, aquellas técnicas escogidas para la experimentación de este proyecto.

2.4. Sistemas de Lectura de Labios Automática

En sus orígenes, los Sistemas de Reconocimiento Automático del Habla (SRAH) se basaban únicamente en la información acústica, dado que las señales de audio contienen más información que las de vídeo en cuanto a la distinción de los fonemas. Hoy en día, este tipo de modelos son poderosos sistemas capaces de entender el lenguaje hablado con índices muy altos de reconocimiento cuando la señal acústica no está corrupta [10]. Sin embargo, en las situaciones en las que esta señal se encuentra degradada decaen las prestaciones de este tipo de sistemas. Es entonces cuando aparece la necesidad de confiar en la información proveniente del canal visual, tomando de base el proceso subconsciente que realiza nuestro cerebro, sobre todo en ambientes o entornos ruidosos, combinando la información de ambos canales para comprender mejor a su interlocutor. Este aspecto propulsó la investigación de los denominados Sistemas Audiovisuales de Reconocimiento Automático del Habla [24, 70, 12]. Estos sistemas intentan equilibrar la contribución del audio y del vídeo para desarrollar modelos que sean robustos ante adversidades acústicas, aportando una mejora significativa de las prestaciones en dichas condiciones.

Por otro lado, en las últimas décadas ha habido un incremento en el interés de decodificar el habla usando exclusivamente la información procedente del canal visual, imitando la capacidad humana de leer los labios [27]. Como ya conocemos, esta rama de los SRAH se enfrenta a numerosos desafíos, ya descritos en otras secciones de la memoria. En ella, el objetivo es construir un sistema encargado de modelar las características visuales de forma que éste sea capaz de leer los labios a partir de imágenes de vídeo. Sin necesidad de introducir ningún tipo de información acústica. Siguiendo el esquema propuesto por Fernández-López y M. Sukno en su revisión [27], dividimos los desarrollos presentes en la literatura en dos ramas dependiendo de si el modelo construido se basa en formalismos tradicionales o, por el contrario, en arquitecturas *Deep Learning*, una evolución típica en numerosos ámbitos del *Machine Learning* que ha supuesto, por norma general, mejores resultados frente al paradigma clásico. En este apartado centramos nuestra atención sobre aquellas publicaciones que resultan de interés para el desarrollo de nuestro proyecto, ya sea porque emplean Modelos Ocultos de Markov (HMM, por sus siglas en inglés) o plantean el modelado del habla espóntanea, así como aquellas que describan características visuales llamativas. No obstante, se recomienda leer el *review* anteriormente citado en caso de que se desee una revisión exhaustiva de la literatura al respecto.

2.4.1. Aproximaciones Tradicionales

Durante este período, la mayoría de los sistemas se conformaron a través de los conocidos HMMs o las *Support Vector Machines* (SVM) y, al igual que cuando describimos los *datasets*, podemos constatar una evolución desde tareas sencillas, como el reconocimiento de dígitos, hasta tareas que adquieren un nuevo nivel de dificultad, como es el tratamiento de oraciones. Respecto a la extracción de características, se observa una gran variedad de aproximaciones, barriendo prácticamente todos los tipos descritos en el apartado anterior. No obstante, ha predominado el empleo de AAMs, DCT o combinaciones de éstas con otras transformaciones como, por ejemplo, PCA.

En cuanto al reconocimiento del alfabeto y dígitos, los *corpora* mayormente estudiados han sido CUAVE [66], XM2VTS [57] y AVLetters2 [18]. Respecto al *dataset* CUAVE, diversos artículos evaluaron características extraídas mediante DCT y LDA, alcanzando 53.12 % WRR [53] y 60.00 % WRR [35], respectivamente. Por otro lado, las arquitecturas presentadas por Papandreou junto a otros autores [63], lograron, empleando para ello características obtenidas mediante un modelo AAM, resultados en torno al 83.00 % WRR. El reconocimiento del alfabeto ha seguido una dinámica similar. Algunos autores, como ocurre en el artículo [18], alcanzaban precisiones alrededor del 90 % WRR al analizar diferentes técnicas para la extracción de características como son PCA, AAM y los denominados filtros *Sieve*, además de estudiar una serie de aproximaciones relativas a si la evaluación era dependiente o no del hablante.

Si hablamos del reconocimiento a nivel de palabras u oraciones, aumentan el número de artículos al respecto a medida que se va incrementando la cantidad de datos disponibles, siendo los *corpus* GRID [15], OuluVS [89], OuluVS2 [2] y RM-3000 [39] los más populares. Dado que este tipo de aproximaciones presentan cierta similitud con los propósitos planteados en nuestro proyecto, como son el empleo de HMMs, lidiar con secuencias de palabras o explorar distintas características visuales, destacamos encarecidamente las siguientes publicaciones:

- Sobre el *dataset* GRID, Harvey y Lan con la colaboración de otros autores [46] compararon diferentes características desde DCT, filtros *Sieve*, PCA y AAM mediante la construcción de un sistema basado en HMMs. Debido al reducido vocabulario presente en el *corpus*, pudieron construir un HMM por cada una de las palabras presentes en las locuciones. De esta forma, obtuvieron resultados entre el 40.00 % WRR y 65.00 % WRR, concluyendo como mejor representación la generada mediante los AAM.

- Varias publicaciones [81, 40] trabajaron con arquitecturas muy similares a las abordadas en nuestro proyecto, aunque no fueran llevadas a cabo con la herramienta Kaldi [71]. Se trata de sistemas donde por cada fonema presente en el *corpus* tendremos un HMM dedicado a él, tal y como sería en los SRAH tradicionales. En este escenario, los autores de ambos artículos estudiaron las prestaciones obtenidas tanto con modelos dependientes como independientes del contexto. Lograron resultados en torno al 47.48 % WRR [81] y al 75.58 % WRR [40], empleando, en ambos casos, características basadas en AAM.

- En cuanto al español, encontramos el *corpus* VLRF [26], donde Fernández-López y M. Sukno [26] realizan una experimentación de gran calidad al

involucrar la precisión con la que un humano lee los labios, así como resultados por el sistema automático tanto a nivel de fonema como a nivel de palabra. En este último, caso se demuestra que las prestaciones del sistema basado en HMMs se encontrarían en torno al 20 % WRR.

Estos y otros artículos que se mencionarán a medida que avancemos en nuestro proyecto, han consolidado los raíles sobre los que, en difentes momentos y aspectos cruciales, ha rodado el desarrollo de nuestro sistema.

Por último y en base a la literatura que hemos inspeccionado, cabe destacar que en esta etapa clásica, aunque sí se afrontan tareas en cuanto al reconocimiento de oraciones, en ningún momento se ha tratado el habla espontánea como tal, ya que normalmente trabajan con un número de frases predefinidas o con un vocabulario acotado. Esto no quiere decir que la lectura de estos artículos no sea de interés o no pueda ser provechosa, sino todo lo contrario, puesto que cooperar y compartir ideas es de vital importancia en la investigación. Por lo tanto, debido a las razones expuestas previamente, nuestro proyecto presenta una novedad al pretender un reconocimiento del habla continua mediante una aproximación tradicional, como son los HMMs o su versión híbrida.

2.4.2. Aproximaciones *Deep Learning*

Gracias a los avances producidos en materia de computación, así como a la disponibilidad de grandes bases de datos audiovisuales, ha sido posible el desarrollo de sistemas basados en la tecnología *Deep Learning*. En un principio, las redes neuronales se emplearon como extractores de características [29, 61, 64] pero los aspectos relacionados con este tema se detallan en la Sección 5.3, ya que forma parte de nuestra experimentación. Posteriormente, se consolidaron los denominados sistemas híbridos, donde las redes neuronales eran combinadas con HMMs. Esta fusión permite prescindir de los GMM para, mediante las Redes Neuronales Profundas (DNN, por sus siglas en inglés), modelar las probabilidades de emisión. De este modo, se constituían los conocidos modelos híbridos o DNN-HMMs, capaces de solventar las limitaciones del anterior paradigma al tratar ya no sólo con más de un *frame* de información para sus estimaciones, sino al afrontar, de una forma más adecuada, datos que pertenezcan a espacios de representación no lineales [38]. Más tarde, las redes neuronales recurrentes, concretamente las *Long-Short Term Memory* (LSTM) [31], se presentarían como la arquitectura sucesora de los HMMs, capaces de capturar información contextual tanto de largo como de corto alcance. La gran potencia de estos modelos sería explotada mediante la construcción de sistemas *end-to-end*, los cuales han proporcionado el actual estado del arte.

Centraremos nuestra atención sobre las arquitecturas *end-to-end*, es decir, aquellos sistemas íntegros y compactos encargados de obtener la transcripción a partir de las imágenes directamente. No obstante, cabe destacar dos publicaciones basadas en modelos DNN-HMM y que, además, emplean el *toolkit* Kaldi [71]. El primero de ellos [73], aunque trate de resolver el reconocimiento de dígitos aislados haciendo uso del *corpus* CUAVE, realiza un estudio interesante respecto a las características visuales, evaluando incluso una representación extraída mediante redes neuronales. Además, muestran como, en base a sus resultados, parece que los modelos híbridos superan en prestaciones a los modelos convencionales basados en GMM. Por otro lado, la tesis escrita por Thangthai

[80] expone una amplia gama de experimentos sobre el *corpus* TCD-TIMIT [36], alcanzando cerca de un 62 % WRR. Estos resultados, tal y como se afirma en la tesis, presentan la limitación de evaluarse sobre un conjunto de datos grabado en un entorno controlado, aunque, eso sí, se trata de un gran *corpus* donde encontramos alrededor de 6000 palabras de vocabulario. Otro aspecto relevante que describe el autor es la investigación que plantea respecto al silencio y cómo poder tratarlo adecuadamente a partir de la señal visual.

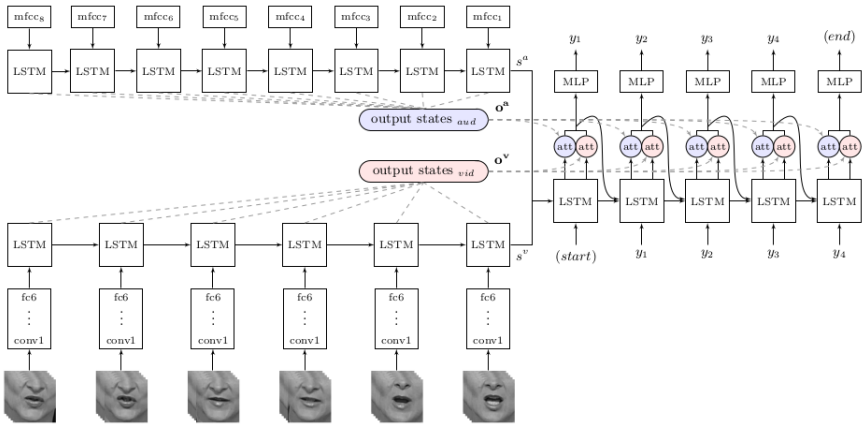

Figura 2.4: Arquitectura estado del arte en lectura de labios [12]

Como ya hemos avanzado, la vanguardia se encuentra en los sistemas *end-to-end* basados en técnicas de *Deep Learning*. Por norma general, a pesar de que numerosos autores hayan explorado distintas alternativas, en este ámbito impera lo que conocemos como arquitecturas *Sequence-to-Sequence*, ampliamente utilizadas en el campo del reconocimiento del habla, el texto manuscrito o la traducción automática. Normalmente, esta arquitectura se fundamenta en la combinación de Redes Neuronales Convolucionales (CNN, por sus siglas en inglés) como un primer bloque encargado de extraer las características visuales, junto con LSTMs dedicadas a identificar las relaciones temporales y proporcionar la transcripción final. Más concretamente, consistiría en una topología *Encoder-Decoder*, donde un primer módulo, denominado *encoder*, procesaría las características visuales obtenidas previamente para obtener una representación abstracta con la que alimentar al decodificador, normalmente también construido mediante LSTMs. Dicha representación abstracta contiene cierta información relacionada con el contexto, lo que aporta robustez de cara a la decodificación.

Sin embargo, este proceso es complejo y suele estar respaldado por lo que en la literatura se conoce como Modelo de Atención (AM, por sus siglas en inglés). El principal propósito de este modelo es guiar e indicar al *decoder* en qué zonas de la representación debe fijarse o centrar su atención de forma que la transcripción sea lo más acertada posible. Es necesario tener en cuenta que la decodificación en este tipo de sistemas se realiza a nivel de caracter. Por lo tanto, al final, esta etapa final dependerá de los caracteres que hayamos decodificado hasta el momento, de la influencia del AM, del decodificador y, además, de un Perceptrón Multicapa (MLP, por sus siglas en inglés) que viene a ser el módulo encargado de predecir el caracter correspondiente en un determinado instante

de tiempo.

Esta arquitectura, la cual podemos observar en la Figura 2.4, fue la que presentaron los autores que constituyen el actual estado del arte en la lectura de labios continua y sin ningún tipo de restricción, alcanzando en torno al 50 % WRR sobre el *corpus* LRS [12]. Lo primero que nos llama la atención es que no sólo emplean la señal de video, sino que adicionalmente introducen información acústica durante la etapa de entrenamiento. Es después, en el período de decodificación, cuando hacen pruebas prescindiendo de un canal u otro. No hay lugar a dudas de que los autores estuvieron influenciados por los avances en el reconocimiento automático del habla acústica, ya que existe un parecido razonable con la arquitectura *Listen, Attend and Spell* [10]. De ahí que ésta fuera bautizada como *Watch, Listen, Attend and Spell*.

Entre una multitud de publicaciones, destacamos el artículo realizado por Assael y el resto de sus compañeros [3], donde se enfrentan a una tarea de clasificación a nivel de oración. Construyeron un sistema con capas convolucionales capaces de lidiar con información espacio-temporal. La salida proporcionada por este módulo óptico sería suministrada a una red neuronal recurrente bidireccional, es decir, procesando las características en ambos sentidos con el objetivo de extraer información que pueda ser de utilidad. Por último, la decodificación correría a cargo de los conocidos *Connectionist Temporal Classification* (CTC) [34]. Por otro lado, introdujo aspectos interesantes respecto al *data augmentation* que puede realizarse en este tipo de tareas. Uno de ellos era aplicar el efecto espejo u *horizontal flip* sobre una locución entera, mientras que el otro sugería duplicar, con un cierto factor aleatorio, varios frames en una locución con el objetivo de enseñar al sistema que las mismas palabras pueden pronunciarse con una velocidad u otra.

Para finalizar, nombrar que existen aproximaciones similares a las descritas en este apartado pero también otras que han implementado arquitecturas alternativas no carentes de interés. Por ejemplo, Stafylakis y Tzimiropoulos [76] propusieron una arquitectura *end-to-end* que combinaba LSTMs con un bloque de *Residual Neural Networks* [37] para hacer frente a una clasificación a nivel de palabra. Otro aspecto a destacar, que ya fue mencionado en la Sección 2.3, es que, por norma general, se ha cedido la extracción de características a las CNNs, ya que la calidad de los resultados puede beneficiarse si el módulo encargado de obtener la representación visual se encuentra guiado por los fallos a la hora de generar las transcripciones finales y no se emplea de forma aislada como sucedía en los sistemas tradicionales.

2.5. Sumario

Llegados a este punto, somos conscientes del progreso que han sufrido los sistemas enfocados al reconocimiento visual del habla. Partiendo desde las aproximaciones clásicas, hemos visto cómo al principio se planteban tareas sencillas hasta que, a medida que aumentaba el interés y la recopilación de *corpus* audivisuales, se pudo hacer frente ya al reconocimiento de tareas que presentasen una mayor complejidad. Por norma general, los HMM han imperado en este tipo de aproximaciones, debido a su gran potencial cuando se trata de procesar datos con un cierto carácter temporal. En esta etapa, los esfuerzos se centraron esencialmente en encontrar una representación de calidad, explorando numero-

sas alternativas, tal y como se indicó en el apartado 2.3. Sin embargo, no se alcanzó ningún consenso a este respecto. Posteriormente, como ha ocurrido en otros ámbitos, la investigación gravitó hacia las tecnologías *Deep Learning*, concretamente, hacia las arquitecturas *end-to-end*, donde se construyen sistemas compactos dedicados a un aprendizaje directo a partir de las imágenes. De esta forma, todo parámetro que constituye el modelo es estimado conforme al objetivo final: obtener la transcripción escrita. En este caso, la combinación de CNNs y LSTMs es la arquitectura más empleada.

En cuanto a los resultados, en muchas ocasiones es díficil realizar una comparación debido a la multitud de *datasets* y aproximaciones exploradas. No obstante, en términos generales, se ha observado que las aproximaciones basadas en *Deep Learning* han superado notablemente las prestaciones obtenidas en la etapa tradicional. Además, en base a la literatura consultada, se ha observado que el reconocimiento del habla espóntanea como tal solamente ha sido tratado en breves ocasiones y con aproximaciones *Deep Learning*, mientras que en el resto de escenarios, encontramos modelos enfocados al procesamiento de palabras u oraciones predefinidas que no pueden extrapolarse al reconocimiento del habla natural.

Con todo esto, el estado del arte actual respecto a la lectura de labios automática dentro del habla espontánea se situa en torno al 50 % WRR [12]. A pesar de que estos resultados puedan parecer modestos si los comparamos con el reconocimiento acústico del habla, son, sin lugar a dudas, un progreso significativo en la materia. Más aún si tenemos en cuenta los desafíos que plantea la carencia de esa información auditiva. Por otro lado, tal y como sugieren los autores del artículo [27], el modelado temporal de secuencias es un factor crucial en este tipo de sistemas, siendo de vital importancia su capacidad para capturar información contextual que pueda solventar las ambigüedades visuales, ya comentadas a lo largo de la memoria. Por ello, la lectura de labios automática continúa siendo un campo abierto de investigación.

3 | Construcción del *Corpus*

Como se ha sugerido a lo largo de la memoria, los datos constituyen el pilar fundamental de todo sistema basado en técnicas de *Machine Learning*, por lo que podemos encontrar referencias a una multitud de datos audiovisuales en la literatura. No obstante, debido a nuestros objetivos, más que emplear uno de los *datasets* mencionados en el estado del arte se ha optado por construir nuestro propio *corpus*. Un *corpus* cuyo propósito sea integrar un conjunto de datos recopilado a partir de un entorno no controlado y sin restricciones, como puede ser un programa de televisión. De esta forma, pretendemos dotar al sistema con datos robustos frente a escenarios realistas, donde el interlocutor no tiene por qué permanecer en una postura estática, sino que podrá realizar movimientos o rotaciones de cabeza durante la emisión de su discurso, entre otros detalles que veremos más adelante. Otro motivo que ha impulsado la recopilación de un *corpus* propio ha sido que la mayoría de las bases de datos audiovisuales que cumplen estas cualidades no están dedicadas al español, tal y como se pudo observar en el Sección 2.2. Por otro lado, la única restricción que establecemos es que, en cada uno de los vídeos que recopilemos, solo podrá aparecer una persona, el o la *speaker* propiamente dicho, ya que si no fuera así podrían surgir problemas en etapas sucesivas (concretamente, en la detección labial). Sí que es verdad que si aparecen en escena otras personas pero se encuentran lo suficiente lejos para no interferir en posteriores procesos, podemos considerar el vídeo como muestra del *corpus* perfectamente.

Figura 3.1: Extractos del *corpus* audivisual recopilado

Con todo esto, decidimos recopilar nuestro *corpus* a partir de un subconjunto de la base de datos audiovisual RTVE [50]. En un *dataset* que disfruta de gran prestigio por su empleo en las evaluaciones Albayzín, conocidas por su implicación en las Tecnologías del Habla. Se trata de un *corpus* constituido por numerosos programas emitidos entre 2015 y 2018 por Radio Televisión Española, lo que cumple con los objetivos planteados anteriormente. Concretamente, debido a limitaciones temporales, para nuestro proyecto se han recogido datos comprendidos entre diciembre de 2017 y enero de 2018, abarcando un gran abanico de *speakers* y situaciones, tal y como se sugiere en la Figura 3.1, presentes en el telediario conocido como 20H. Por último, indicar que cada muestra de nuestro *corpus* estará formada por el vídeo que contiene la locución, así como por la transcripción, totalmente supervisada, que le corresponde.

3.1. Detalles del Corpus

En esta sección exponemos detalles concretos respecto al *corpus* recopilado. En primer lugar, es necesario explicar dos aspectos relevantes a tener en cuenta:

- Una vez tuvimos una cantidad considerable de segundos recopilados, por cada *speaker* se realizó una especie de *Data Augmentation*, de modo que incrementásemos la duración del *dataset* pero sin aumentar el vocabulario. Este proceso consistía en segmentar aquellas locuciones con una duración relativamente larga y donde el *speaker*, al realizar pausas, nos permitiese extraer porciones. Entonces, de una sola muestra podríamos obtener varios extractos con la intención de facilitar el alineamiento del sistema.

- En cuanto a las transcripciones, hay ciertos factores que pueden dificultar el reconocimiento y de los cuales podemos prescindir. Debido a ello, han sufrido un preproceso por el cual han sido eliminados todos los acentos y signos de puntuación, además de convertir todo carácter a su versión minúscula.

Por otro lado, como bien hemos avanzado, este conjunto de datos se encuentra conformado por un gran número de *speakers*. Sin embargo, no todos ellos presentan la misma participación, puesto que algunos son reporteros temporales o aparecen en pocas ocasiones, mientras que otros son presentadores del telediario. En consecuencia, nuestro *corpus* presenta la distribución de segundos que se muestra en la Figura 3.2. Llama la atención que uno de los *speakers*, concretamente el primero de ellos, ocupa, prácticamente, alrededor de una hora de locuciones, ya que es el presentador principal del telediario. Este desequilibrio, tal y como se comenta en el Capítulo 7, puede suponer problemas a la hora de estimar los parámetros del modelo, puesto que éstos pueden ceñirse en exceso a la fisionomía o naturaleza de este hablante. No obstante, en el supuesto de que prescindamos de este *speaker*, continua existiendo cierto desequilibrio, aunque en menor medida. Por lo tanto, gracias a esta información somos conscientes de que debemos establecer una partición adecuada antes de adentrarnos en el entrenamiento del sistema, tal y como explicamos en la Sección 7.1.

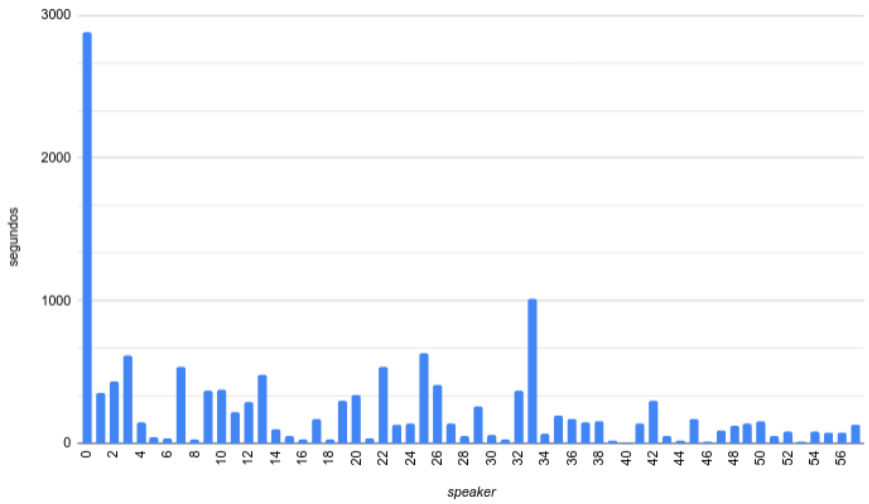

Figura 3.2: Distribución de segundos entre los diferentes *speakers* del *corpus*

En último lugar, se debe hacer mención de ciertos detalles respecto al *corpus* que permitan conocerlo de un modo similar a como hemos expuesto los *datasets* del estado del arte, añadiendo alguna información más concreta o que pueda resultar de interés, tal y como refleja la Tabla 3.1. Destacamos, en primer lugar, la resolución de cada una de las muestras (480x270 píxeles), la cual no se corresponde con el tamaño que presenta la región bucal o la cara del *speaker*, sino con el plano entero que fue tomado para la emisión del telediario. Este hecho va a suponer que, al final, trataremos con imágenes de una reducida dimensión si nos comparamos con otros *corpus*. Por otro lado, podemos observar como nuestro conjunto de datos exhibe una gran variabilidad en cuanto al número de palabras o fonemas que aparecen en una locución. Sin embargo, subrayamos un aspecto importante: el número de fonemas por locución. Podemos observar que, en promedio, hablamos de 49 fonemas en cada una de las muestras, aunque en ocasiones podemos alcanzar un máximo de 270 fonemas; y estamos hablando de vídeos que no suelen superar más de 10 segundos de duración. Este detalle, junto al reducido número de *frames* al que nos limita la señal visual (30 *frames*/segundo), puede ocasionar problemas en el alineamiento de los datos, tal y como mencionaremos en la Sección 8.1.

Tabla 3.1: Detalles y estadísticas respecto al *corpus* audiovisual recopilado

Aspecto	**Detalles relacionados**		
Lengua	Español		
Resolución	480x270 píxeles, 30 frames/segundo		
Speakers	58	**Hombres:** 18	**Mujeres:** 40
Duración	∼4 horas		
Locuciones	3287		
Vocabulario	3754		
Fonemas†	23		
Palabras por locución	**Mediana:** 10	**Máximo:** 62	**Mínimo:** 1
Fonemas por locución†	**Mediana:** 49	**Máximo:** 270	**Mínimo:** 4

† estas estadísticas pueden variar en función de los fonemas acordados

3.2. Observaciones respecto a la complejidad de la tarea

Una de las ventajas de construir nuestro propio *corpus* es poder observar, durante un largo período de tiempo, los distintos aspectos relacionados con la tarea en cuestión, es decir, la lectura de labios automática. Hemos contemplado, cuestionado y examinado numerosas situaciones que nos han permitido conocer al detalle aspectos a tener en cuenta en futuras etapas. No obstante, también nos ha permitido identificar una serie de dificultades que podrían surgir cuando tratemos de modelar el habla visual. Entre ellas, destacamos:

- Dificultad a la hora de reconocer el silencio. Por una parte, la boca no siempre tiene por qué estar cerrada en estas situaciones, puesto que una persona puede estar con la boca abierta sin pronunciar sonido alguno. Por otro lado, podemos pensar que el silencio puede identificarse cuando no haya movimientos labiales. Esta aproximación podría ser más coherente. Sin embargo, en ocasiones podemos estar emitiendo un sonido relativamente uniforme y estar, durante varios *frames* del vídeo, con la boca aparentemente estática.

- Observamos que el contexto presenta una influencia de gran calibre. De hecho, dependiendo del fonema anterior o posterior, la deformación bucal que está articulando el *speaker* puede llegar a ser distinta aun cuando se pronuncia el mismo fonema. Un buen ejemplo suele ser cuando el siguiente fonema involucra la vocal 'o', ya que percibimos como la boca se va cerrando instantes previos a su pronunciación.

- Otro aspecto de gran importancia, bien conocido en la literatura, es la ambigüedad existente cuando se pretende reconocer el habla a partir del canal visual. Estamos hablando de los homo-visemas, concepto que haría referencia a aquellos fonemas cuya representación visual a partir de los labios sea idéntica o de gran parecido, pero su transcripción sea diferente. Por ejemplo, este suceso lo encontraríamos a la hora de distinguir entre los fonemas /b/, /p/ y /m/.

- Debido a la naturaleza con la que se han recopilado los datos, a menudo observamos que la velocidad con la que hablan los *speakers* puede variar de una locución a otra. Este detalle podría suponer problemas a la hora de que el sistema estimase sus parámetros, ya que estas variaciones en velocidad dificultan el alineamiento.

- Por último, y debido a la misma razón que exponíamos en el punto anterior, identificamos ciertos detalles que, en el caso de aparecer con excesiva frecuencia, podrían afectar sobre las prestaciones del sistema. Se trata de cambios de iluminación, sobre todo si es un reportero a pie de calle, poca vocalización, errores o confusiones durante el discurso para después rectificar y retomar el hilo, humedecerse los labios o, incluso, bajar la cabeza para leer de sus apuntes.

4 | Detección labial en imágenes

Antes de poder abordar la extracción de características visuales que representen los movimientos labiales a lo largo de un discurso, debemos ser capaces de extraer nuestra Región de Interés (ROI, por sus siglas en inglés). En nuestro caso, tal y como podemos intuir, se trata de la boca del emisor. Esta etapa la conocemos como detección labial, y será aplicada sobre cada uno de los *frames* que conforman todas las locuciones presentes en el *corpus*. Para ello, haremos uso de las librerías OpenCV [9] y Dlib [42], ampliamente conocidas en el ámbito del *Machine Learning* y la Visión por Computador. Centraremos nuestra atención sobre el *software* proporcionado por la herramienta Dlib que nos ha permitido identificar los *landmarks* o puntos de referencia faciales, tal y como sugiere la Figura 4.1. En ella, podemos observar claramente dos etapas, las cuales constituyen el núcleo central de este capítulo. El proceso en sí consistiría en detectar la cara del interlocutor para después, a partir de este paso previo, identificar los *landmarks* con los que podremos obtener la región bucal deseada.

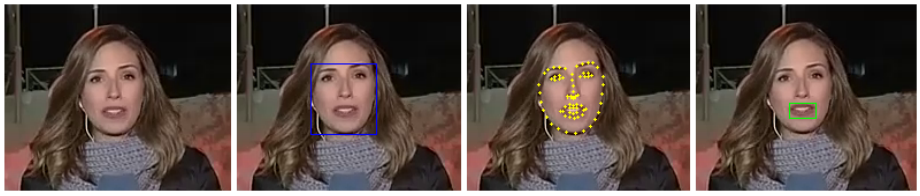

Figura 4.1: Proceso por el que se detecta la boca del interlocutor en imágenes

4.1. Identificación facial

La primera etapa, tal y como hemos avanzado anteriormente, consiste en detectar la cara del *speaker* en un *frame* concreto. De esta forma, facilitamos la tarea al predictor encargado de identificar los *landmarks* faciales, ya que estaríamos limitando la zona donde debe localizar estos puntos de referencia.

En nuestro caso, el *software* escogido emplea una de las técnicas clásicas en Visión por Computador: *Histogram of Oriented Gradients* (HOG), técnica respaldada en uno de los artículos publicados por Dalal y Triggs [20], donde se ha demostrado que la combinación de características HOG con clasificadores lineales, como pueden ser las SVM, proporcionan una detección y reconocimiento de objetos de gran calidad. A fin de cuentas, se trata de un descriptor visual, cuya base se fundamenta en el cómputo de los gradientes, tanto en la dirección horizontal como vertical, tal y como se muestra en la Figura 4.2. Entonces, mediante el uso de una ventana deslizante, va recorriendo la imagen con el objetivo de proporcionar una representación robusta que permita distinguir los objetos presentes en ella. Para ello, en cada ventana se calcula el histograma de estos gradientes, teniendo en cuenta su orientación y magnitud. Con todo esto y un conjunto de datos apropiadamente etiquetado ya podríamos entrenar nuestro clasificador lineal.

No obstante, hay que hacer notar que existen diversas alternativas. Por ejemplo, el uso de arquitecturas *Deep Learning* o los métodos basados en el empleo de clasificadores en cascada junto a las denominadas características Haar, como puede ser la popular aproximación propuesta por Viola y Jones [84].

Figura 4.2: Gradientes obtenidos para el cómputo del descriptor HOG. Izquierda: gradiente horizontal. Derecha: gradiente vertical.

4.2. Localización de los *landmarks*

Tras detectar la cara del hablante, estamos delimitando la zona o región donde deben localizarse los *landmarks* faciales y, por lo tanto, facilitando la complejidad de esta tarea, al tratarse de forma indirecta. La herramienta que utilizamos en nuestro proyecto ha creado un módulo encargado de obtener estos puntos de referencia, tomando como principal referencia el artículo escrito por Kazemi y Sullivan [41]. A diferencia de éstos, que detectan cerca de 200 *landmarks*, el *toolkit* Dlib proporciona los 68 *landmarks* que refleja la Figura 4.3. En ella, observamos como, a partir de esta información, seremos capaces de extraer la zona donde se ubica la boca del interlocutor, o bien una serie de métricas con gran detalle, tal y como se explica en la Sección 5.1. No obstante, para poder realizar este alineamiento facial, apelativo que recibe en el artículo anteriormente citado, se ha hecho uso del *corpus* iBUG 300-W [74], perfectamente anotado en base a las 68 *landmarks* comentados y recogido bajo la filosofía *in the wild*.

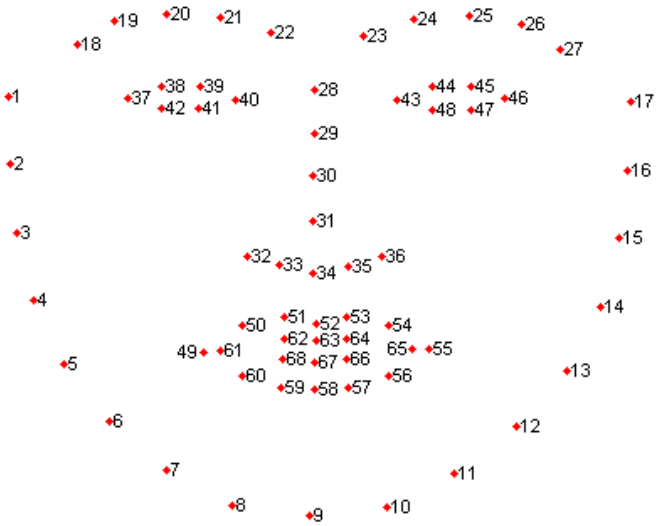

Figura 4.3: *Landmarks* faciales proporcionados por el *software* Dlib.

Entonces, la implementación de este módulo se basa en un algoritmo, cuyo funcionamiento se encuentra consolidado sobre un conglomerado de funciones de regresión. Estas funciones conforman el clasificador final, el cual será entrenado

siguiendo la metodología de un clasificador en cascada o mediante la filosofía *gradient boosting*, tal y como se realiza en el artículo que constituye su principal referencia. A parte de esto, replicando dicho artículo, se incorporan multitud de detalles enfocados esencialmente en acelerar, de forma significativa, el cómputo de este algoritmo hasta alcanzar tiempos en la escala de los milisegundos. Eso sí, sin deteriorar la calidad de sus predicciones. Mientras estas estrategias se detallan en el artículo, describiremos a grandes rasgos el cuerpo central del algoritmo implementado. Se trata de un proceso iterativo, tal y como sugiere la Ecuación 4.1. Respecto a la nomenclatura, los autores denotan con el símbolo $\hat{S}$ la forma o *shape* estimada en un instante de tiempo concreto t, mientras la referencia, la verdadera *shape*, se daría a conocer como S. Este concepto de *shape* no es más que el conjunto de *landmarks* a localizar sobre la imagen, representada mediante el símbolo I. Por último, tenemos la función de regresión dedicada a un instante de tiempo concreto r_t. Este regresor, en función de la imagen I y la actual predicción de forma $\hat{S}^{(t)}$, estimará un vector que nos permitirá obtener la nueva predicción. De este modo, de una forma progresiva, se consigue refinar la predicción de nuestro sistema hasta que se alcanza una condición que indique que el algoritmo ha convergido, tal y como expone la Figura 4.4.

$$\hat{S}^{(t+1)} = \hat{S}^{(t)} + r_t(I, \hat{S}^{(t)}) \tag{4.1}$$

Sin duda, el algoritmo descrito hasta el momento requiere de una inspección más cuidadosa si se desea comprender de un modo más apropiado. Por ello, se recomienda la lectura del artículo citado a lo largo de esta sección.

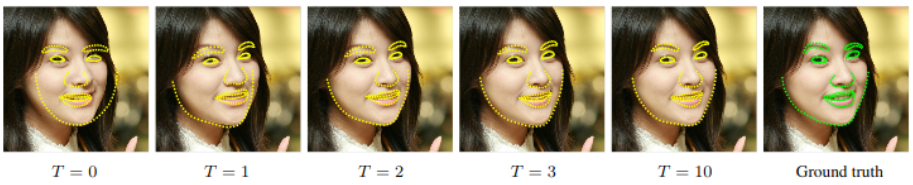

Figura 4.4: Estimación progresiva de los *landmarks* faciales [41]

5 | Extracción de Características Visuales

Una vez ya tenemos nuestro *corpus* adecuadamente recopilado es hora de prepararlo para que pueda ser interpretado por nuestro sistema final. Estamos hablando de la etapa de extracción de características, donde debemos obtener una representación visual robusta que nos permita distinguir, en la medida de lo posible, las diferentes unidades fonéticas de forma que determinemos el mensaje emitido por el *speaker*. Sin embargo, tal y como hemos ido avanzando a lo largo de la memoria, existen numerosas ambigüedades visuales, entre otros desafíos, cuando tratamos de leer los labios. Es por ello que esta etapa adquiere gran importancia, siendo un factor crucial en este tipo de sistemas. De hecho, en la literatura se han estudiado un amplio abanico de alternativas a la hora de representar los movimientos labiales, tal y como describíamos en la Sección 2.3. No obstante, no se ha alcanzado todavía ningún consenso en cuanto a qué características llegan a ser las más representativas o robustas. Debido a estas razones, hemos optado por explorar múltiples aproximaciones en esta etapa, pasando desde características centradas únicamente en la forma o apariencia que adopta la boca del emisor hasta características determinadas por arquitecturas *Deep Learning*. De esta manera, seremos capaces de realizar un análisis con el objetivo de determinar qué tipo de características serían las mejores para nuestro caso de estudio, siendo este estudio el núcleo central de nuestro proyecto. Por lo tanto, este capítulo trata de describir las diferentes características que, en base a la literatura consultada, se han decidido estudiar. Será en sucesivos capítulos, concretamente el Capítulo 7, tras comprender mejor el sistema empleado, donde llevaremos a cabo el análisis mencionado. Por último, antes de adentrarnos en las distintas aproximaciones es necesario destacar que todas tienen en común la metodología a la hora de extraer la secuencia de características, es decir, todas van a definir un vector de características (de talla fija) por cada *frame* que componga el vídeo de la locución. Por tanto, cada muestra del corpus puede presentar una longitud distinta al resto. No es como otro tipo de aproximaciones, normalmente en clasificación de palabras u oraciones predefinidas, donde se establecen una serie de criterios para normalizar dicha longitud variable.

5.1. Características Geómetricas

La primera representación está relacionada con la fisionomía que adopta la boca del *speaker* a lo largo del discurso. Concretamente, este tipo de características consiste en definir, para cada instante de tiempo, un conjunto de métricas acordes con las deformaciones que afectan principalmente a los labios del locutor, sin llegar a contener ningún detalle respecto a la apariencia, como podrían ser características vinculadas a la lengua y dientes. En nuestro caso, vamos a constituir el vector de características a partir de 18 distancias que reflejen la altura y anchura adoptadas. Para llevar a cabo su extracción aprovechamos la información proporcionada por el *software* que describíamos en el Capítulo 4, tal y como refleja la Figura 5.1. En otras palabras, gracias a los *landmarks* expuestos en la Figura 4.3, podremos extraer estas métricas bucales con cierta facilidad. No obstante, con el propósito de obtener unas características robustas frente a la gran variabilidad que presenta la tarea, se han tenido en cuenta los siguientes detalles:

- Las métricas mencionadas son tomadas haciendo uso de la conocida Distancia Euclídea (Ecuación 5.1), ya que, a fin de cuentas, los *landmarks* no dejan de ser puntos situados en el espacio de representación que constituye la imagen. Una de las razones más importantes por las que se ha escogido este tipo de distancia es por su invariabilidad en caso de que el locutor incline la cabeza.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \tag{5.1}$$

- Otro aspecto a tener en cuenta es que el *speaker* puede estar hablando ante la cámara desde distintos planos, es decir, a mayor o menor distancia. Esto supone graves problemas a la hora de modelar las características, ya que una misma posición de la boca puede proporcionar métricas con magnitudes muy dispares en función de la distancia a la que se encuentre la cámara. Para poder solventar o al menos diluir los efectos adversos que puede ocasionar este tipo de situaciones, se ha optado por realizar una normalización al respecto. Usando la Figura 5.1 como apoyo, vemos que en cada instante de tiempo computamos una región más amplia, que abarca lo que identificaríamos como el mentón (recuadro azul) junto a una región ya más ceñida a lo que sería la boca del interlocutor (recuadro verde). En el interior de ambas regiones se encuentran los *landmarks* bucales con los que calcularemos las características finales. Dado que dicho mentón es una zona mucho más estable y uniforme que la boca, podremos normalizar las distancias que tomemos de una forma apropiada. Para ello, si ponemos el ejemplo de medir la anchura bucal comprendida entre las comisuras labiales, bastaría con dividir dicho valor entre la anchura que presenta el mentón. Con el resto de métricas, ya sean alturas o el área ocupada, procederíamos de un modo similar, consiguiendo una representación robusta frente a la distancia con la que fue grabado el *speaker* durante su discurso.

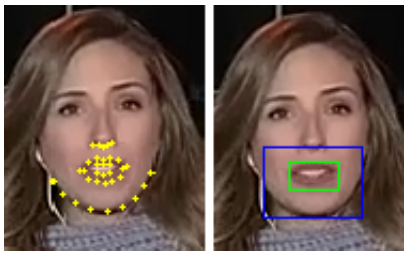

Figura 5.1: Detalles respecto a la extracción de las características geométricas

Una vez resueltos estos detalles, ya podemos estudiar qué métricas pueden resultar más provechosas de cara al reconocimiento visual del habla. A fin de cuentas, decidimos establecer numerosas métricas que definiesen distintos aspectos como pueden ser la anchura y altura, la opertura de la cavidad bucal, la distancia con puntos de referencia como la nariz o barbilla, así como el área que ocupa la boca del locutor. Todas ellas normalizadas siguiendo el esquema que se ha mencionado previamente. Para comprender mejor cómo se han extraído las 18 componentes del vector de características, se ha construido la Tabla 5.1. En ella se refleja a partir de qué *landmarks* (ver Figura 4.3) son calculadas cada una

de las métricas, teniendo en cuenta la excepción que constituye el área bucal, ya que, tal y como se detalla en la tabla, no se trata de una distancia como en el resto de casos. Además, es necesario saber que con *distNariz* nos referimos a distancias desde el contorno de labios externo hasta las fosas nasales, mientras que *distBarbilla* sería similar pero midiendo la distancia hasta la barbilla.

Tabla 5.1: Métricas definidas en las características geométricas

Métricas ([*landmark*, *landmark*])	**Etiqueta**
[49, 50, 51, ..., 68]†	*areas*
[51, 33], [52, 34], [53, 35]	*distNariz*
[68, 8], [58, 9], [56, 10]	*distBarbilla*
[61,65], [60,56], [59,57]	*anchuras*
[50,60], [51,59], [52,58], [53,57], [54,56]	*alturas*
[62,68], [63,67], [64,66]	*operturas*

† Se trata de un caso excepcional al no ser una distancia, sino al calcularse el área del rectángulo que englobe todos los *landmarks* bucales

Por otro lado, en la Figura 5.2 podemos observar la secuencia de características completa de una muestra del corpus en la que se pronuncia la palabra "completamente". Sobre esta secuencia se ha aplicado una normalización z-score, puesto que, tal y como se detalla en el Capítulo 7, proporciona mejores resultados. Concretamente, en este caso, la secuencia que mostramos es el resultado de aplicar dicha normalización *per speaker*, tal y como sugiere Lan junto a otros autores en una de sus publicaciones [48]. No obstante, podemos comprobar que, aunque existan zonas distintivas, no logra capturar suficiente información, ya que se observa cierto carácter homogéneo y se supone que en esta representación se deberían diferenciar los trece fonemas que se pronuncian. Un análisis más exhaustivo en cuanto a las características geométricas (conocidas de ahora en adelante como *geometricFeats*) se encuentra en el Anexo A, donde, con el objetivo de facilitar su visualización, se han agrupado las 18 componentes tal y como se especifica en la Tabla 5.1.

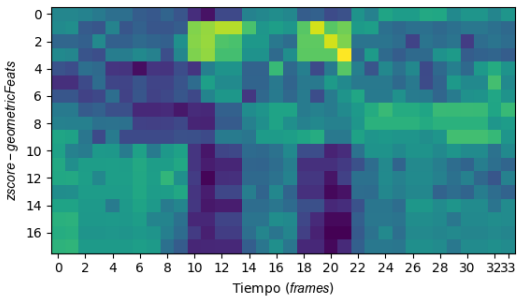

Figura 5.2: Secuencia de características geométricas

5.2. EigenLips

El siguiente tipo de características que exploraremos en nuestro proyecto va a estar relacionado con la apariencia, es decir, con la información visual contenida

en los píxeles que conforman la boca del interlocutor. Para ello, hemos optado por emplear la conocida técnica no supervisada PCA, de la que ya hemos hecho mención en el estado del arte. Como bien sabemos, a menudo los datos con los que trabajamos pueden contener características que son redundantes o que presentan una alta correlación. Si logramos reducir la dimensión de estos datos, prescindiendo de aquellos que no aporten información discriminativa, estamos contribuyendo favorablemente sobre el sistema final, ya que disminuiremos la carga computacional al tener que procesar un menor número de características y, por otro lado, al tratarse de datos decorrelados, puede beneficiarse su modelado mediante GMMs, como es nuestro caso. Estas son las principales razones por las que PCA ha ganado una amplia popularidad en el ámbito del *Machine Learning*, ya que se encargará de identificar la transformación lineal más apropiada con la que proyectar los datos sobre un nuevo espacio de representación. Un espacio donde se haya reducido la dimensión de estos datos y, además, se preserve gran parte de la varianza. En otras palabras, se trata de reducir el número de características pero con la intención de mantener la máxima varianza posible, es decir, encontrar las características con mayor capacidad discriminativa.

En nuestro caso, vamos a emplear PCA para obtener lo que en la literatura se ha bautizado como *eigenLips*. Este concepto se encuentra influenciado por los estudios realizados en materia de reconocimiento facial, donde se extraían las famosas *eigenFaces* [22]. Volviendo a nuestro proyecto, el primer paso es obtener las regiones bucales de todo nuestro *corpus*. Para ello, se ha hecho uso del *software* descrito en el Capítulo 4 con el que hemos extraído la región bucal ceñida, tal y como sugiere el recuadro verde de la Figura 5.1. Entonces, con el objetivo de poder aplicar PCA, redimensionamos cada una de estas regiones a un tamaño fijo de 32x16 píxeles, definiendo, entonces, cada imagen con 512 componentes. Se ha decidido este tamaño debido a las limitaciones respecto a la resolución de los vídeos, ya comentadas en el Capítulo 3, y en base al artículo publicado por Fung y Mak [29], donde se planteaba la lectura de labios a partir de imágenes con una dimensión reducida. Por otro lado, se ha considerado oportuno emplear estas imágenes en su versión de escala de grises, en lugar de su formato RGB original. Por último, con el objetivo de facilitar la tarea en futuras etapas del reconocimiento, aquellas regiones en las que la boca no estuviera alineada adecuadamente, es decir, se encontrase inclinada o ladeada debido a la naturaleza sin restricciones en las se obtuvieron los datos, han sufrido una transformación, concretamente una rotación, de forma que todas ellas pasarán a estar normalizadas. Este proceso se ha llevado a cabo calculando el ángulo presente entre los *landmarks* que delimitan las comisuras labiales. Una vez calculado dicho ángulo, bastaría con aplicar la rotación correspondiente. Dicho proceso puede observarse en la Figura 5.3.

Figura 5.3: Proceso esquematizado del alineamiento bucal

Con todos estos detalles aclarados ya pudimos aplicar PCA sobre nuestros datos de entrenamiento, indicando que se preservase el 80 % de la varianza presente en el *corpus*. Fue así como logramos obtener las componentes principales

con las que poder representar cada imagen únicamente con diez características. Estos componentes, también denominados *eigenLips*, pueden visualizarse en la Figura 5.4. Observamos como cada uno de ellos centra su atención sobre diferentes aspectos de la boca. El primer componente principal es aquel que mayor información relevante consigue capturar; es por ello que resalta esencialmente las comisuras labiales, ya que éstas son una de las partes que mayor deformación sufren durante la locución. En cuanto al resto de *eigenLips*, algunos destacan el contorno de los labios, mientras que otros enfatizan zonas donde hallaríamos los dientes o la lengua. Estos son aspectos que no conseguimos capturar cuando empleamos características puramente geométricas y que son de vital importancia para interpretar el habla visualmente.

Figura 5.4: *EigenLips* obtenidos tras aplicar PCA

En un principio, se mantuvo la idea de aplicar PCA sobre una región bucal amplia que incluso abarcase parte del mentón, tal y como muestra el recuadro azul de la Figura 5.1. Con todo esto, se pretendía aportar una mayor cantidad de información por *frame* que pudiera mejorar la distinción de los diferentes fonemas articulados por el hablante. Sin embargo, al inspeccionar los *eigenLips* obtenidos, observamos que se otorgaba mayor relevancia a zonas no relacionadas con los labios, ya que al seleccionar un área tan amplia en ocasiones se capturaba parte del cuello o incluso del fondo del plató de televisión. Estas zonas contenían gran parte de la varianza presente en el *corpus* seleccionado, por lo que al final eran destacadas por la técnica PCA como características discriminativas. Por el contrario, si estuvieramos trabajando con un sistema *end-to-end*, donde el módulo encargado de extraer las características fuera estimado en función de los errores obtenidos en la transcripción final, igual el sistema podría verse beneficiado de una región más amplia. No obstante, este estudio ha sido planteado en nuestro trabajo futuro, tal y como se indica en la Sección 8.2.

5.3. Características mediante Autoencoders

Como ya introdujimos en el estado del arte (Sección 2.4.2), es posible delegar la responsabilidad de extraer las características visuales sobre redes neuronales [29, 61, 64]. Normalmente, cuando los datos con los que trabajamos son imágenes, se emplean CNNs, las cuales han demostrado prestaciones de calidad en numerosas tareas enfocadas al reconocimiento y clasificación de imágenes [44, 79]. En esencia, las redes convolucionales van aplicando una serie de filtros con los que van reduciendo la dimensión de la imagen que recibió como entrada, a la vez que va obteniendo características visuales de gran relevancia. Esta reducción, permite al modelo extraer características en distintos niveles de profundidad que, sumados al aumento en número de filtros que se produce en

cada etapa, va generando cada vez mapas convoluciones más amplios y representativos. No obstante, nuestra tarea no consiste en una clasificación, puesto que desconocemos el fonema que se está pronunciando en cada uno de los fotogramas que componen el conjunto de datos. Debemos encontrar, entonces, una forma de entrenar nuestra red convolucional con el objetivo de sintetizar la información contenida en cada *frame*. Por ello, y en base a las publicaciones consultadas en la literatura, se ha optado por construir un Autoencoder Convolucional. A grandes rasgos, esta arquitectura, mostrada en la Figura 5.5, se divide en dos componentes:

- Encoder: este módulo se construye, tal y como se ha sugerido previamente, mediante una red convolucional. Su objetivo consiste en obtener una representación abstracta y de reducida dimensión a partir de la imagen que haya recibido como entrada. En nuestro caso, esta representación codificada se enfrenta a un MLP constituido por un número fijo de neuronas completamente conectadas. El número de neuronas definidas en esta etapa determinará la dimensión con la que representaremos la imagen.

- Decoder: al igual que el Encoder, consistiría en una red convolucional pero esta vez con la intención de reconstruir la imagen original a partir de esa representación compacta. Podría enterderse, tal y como refleja la Figura 5.5, como un proceso inverso a la convolución, en el que iríamos ampliando la dimensión del mapa convolucional a la vez que reducimos la profundidad de éste hasta alcanzar el nivel deseado.

De este modo, ya somos capaces de intuir el objetivo de un Autoencoder Convolucional: reconstruir la imagen original, partiendo de una representación abstracta y comprimida. Por lo tanto, una vez hayamos entrenado nuestro modelo, prescindiremos del *decoder*, ya que la parte que realmente nos interesa es aquella que nos permita representar cada *frame* que compone el *corpus* con el propósito de extraer las secuencias de características para cada locución. Estamos hablando del *encoder*, con el que obtendremos dicha representación.

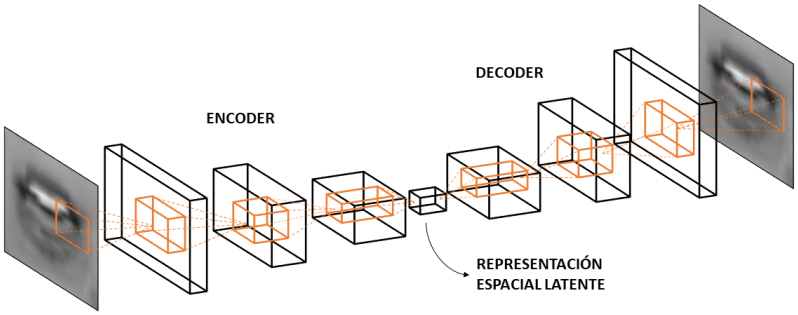

Figura 5.5: Arquitectura general de un Autoenconder Convolucional

No obstante, nos falta un detalle por definir y que es de vital importancia. Nos referimos al elemento que permite realizar la comparación entre ambas imágenes, es decir, la original y la reconstruida. Y será, a partir de este indicador, como la red podrá estimar adecuadamente sus parámetros. Estamos hablando

de la métrica conocida como Índice de Similitud Estructural (SSIM, por sus siglas en inglés). A diferencia del Error Medio Cuadrado (MSE, por sus siglas en inglés), si las intensidades de los píxeles difieren considerablemente, no significa que el contenido de éstas sea distinto, sino que simplemente puede ser una versión más oscura[1]. Debido a este tipo de inconvenientes, nuestra decisión se ha inclinado por establecer SSIM como función de pérdida o *loss*, ya que se centraría en aspectos relacionados con la información estructural a la hora de comparar dos imágenes. Por lo tanto, haciendo uso de la fórmula descrita en el artículo publicado por Wang y otros autores [85], podemos definir la función de pérdida que empleamos en nuestro proyecto (Ecuación 5.2). Esta métrica SSIM, haría su comparación ya no sólo teniendo en cuenta la media (μ) y varianza (σ) de la intensidad presente en los píxeles, sino que se implementaría con la intención de ir procesando las imágenes mediante una ventana de cierto tamaño. De este modo, logramos definir una pérdida robusta frente a los cambios estructurales que se percibirían entre ambas imágenes. Por otro lado, se introducen las denominadas constantes de regularización, representadas mediante los símbolos c_1 y c_2. Estas constantes tienen como objetivo evitar la inestabilidad de ciertas regiones donde la media o la varianza se encuentren cercanas a cero.

$$
\begin{aligned}
loss(x, y) &= 1 - SSIM(x, y) \\
SSIM(x, y) &= \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}
\end{aligned}
\tag{5.2}
$$

En último lugar, describiremos la arquitectura que hemos definido para el *encoder*, cuyas especificaciones quedan expuestas, de forma detallada, sobre la Figura 5.6. No obstante, antes de avanzar, es necesario saber que las imágenes que emplearemos, tanto como para entrenar como validar la red neuronal, son las mismas que se describieron cuando se obtuvieron los *eigenLips*. En otras palabras, serán imágenes de 32x16 píxeles, en escala de grises y normalizadas de forma que todas las bocas se encuentren adecuamente alineadas, tal y como sugiere la Figura 5.3. Por otro lado, no se ha considerado del todo necesario implementar un *data augmentation* que proporcionase imágenes con ciertos desplazamientos o diferentes escalas, puesto que estos matices ya se encuentran presentes en nuestro conjunto de datos, tal y como se mencionó en el Capítulo 3, al no ser planos grabados en un entorno controlado.

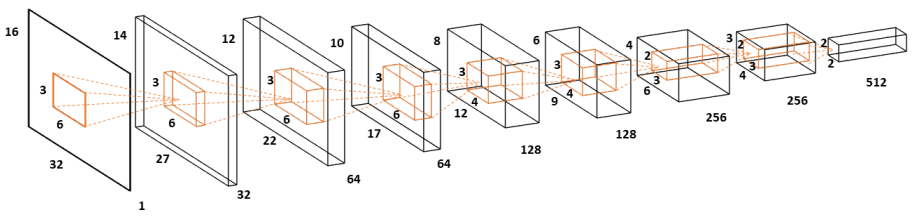

Figura 5.6: Arquitectura en detalle del Encoder diseñado

La principal influencia a la hora construir la topología que define el *encoder*

[1]https://www.pyimagesearch.com/2014/09/15/python-compare-two-images/

ha sido el artículo publicado por Fung y Mak [29], ya que expone dos características en común con nuestro proyecto, aunque aborde la clasificación a nivel de palabra. La primera de ellas es que se propone la tarea de la lectura de labios dentro de un escenario donde las imágenes disponibles presentan una baja resolución. De hecho, con el objetivo de replicar su modelo, se optó por adoptar las dimensionas expuestas en el artículo: 32x16 píxeles. El otro rasgo en común es que los autores decidieran emplear un Autoencoder como extractor de características. Sin embargo, ¿por qué decidimos centrar nuestra atención sobre este artículo y no en otro de los que hemos citado al inicio de esta sección? La razón que ha impulsado esta decisión ha sido observar cómo han adaptado el procesamiento de la imagen a la fisionomía bucal. Tal y como se puede observar en la Figura 5.6, el operador de convolución o *kernel* adopta una forma rectangular buscando una similitud con la boca. De esta forma, se pretende capturar características visuales con cierta coherencia y que de verdad puedan representar los detalles que interesen de cara a la reconstrucción. Además, observamos cómo estas dimensiones van variando a lo largo de las convoluciones y a medida que se reduce el tamaño de los mapas convolucionales. La única diferencia con los autores anteriormente citados sería que no empleamos más de un *frame* a la hora de obtener la representación latente, ya que esto podría disminuir la longitud de las secuencias, que ya es reducida, si se compara con la tarea del reconocimiento del habla acústica. Esta diferencia se debe también a que ellos tratan el problema de clasificación de palabras mediante una topología *end-to-end*. En cuanto a las especificaciones del *decoder*, son fácilmente deducibles a partir de la Figura 5.6, ya que sería, tal y como se sugirió previamente, un proceso inverso. Con todo esto, en la Figura 5.7 se observan varias reconstrucciones obtenidas por el Autoencoder definido, donde observamos resultados de gran calidad a pesar de enfrentarse a difentes escalas o fisionomías. Estas características latentes, de ahora en adelante conocidas como *deepFeats*, constan, únicamente, de 16 componentes.

Figura 5.7: Ejemplos de reconstrucción obtenidos mediante Autoencoder. Columna izquierda: imagen original. Columna derecha: imagen reconstruida.

6 | Desarrollo del Sistema Automático

Este capítulo trata de exponer los fundamentos teóricos en los que se basa el sistema dedicado a modelar la lectura de labios automática. Para su construcción haremos uso de la reconocida herramienta Kaldi [71], cuyo propósito consiste en proporcionar el soporte necesario para la construcción de sistemas enfocados a las Tecnologías del Lenguaje, permitiendo ya no sólo la definición de sistemas tradicionales, sino la edificación y modelado de sistemas *Deep Learning*. No obstante, tal y como se ha ido comentando a lo largo de la memoria y en base a la literatura respecto al reconocimiento convencional del habla acústica, el sistema desarrollado en este proyecto pertenecerá al paradigma tradicional, donde el pilar esencial va estar constituido por los modelos denominados GMM-HMM, aunque este tipo de sistemas, tal y como describimos en este capítulo, se compone de otros módulos imprescindibles de cara al reconocimiento del habla. Por otra parte, faltaría comentar numerosos aspectos relacionados con la parametrización de cada uno de estos módulos, así como detalles respecto a la forma con la que se evalúan este tipo de sistemas.

En cuanto a los fundamentos de implementación empleados a la hora de desarrollar la herramienta Kaldi, se ha dedicado el Anexo B para describir, de un modo intuitivo, la integración de todos los módulos que componen dicho sistema y cómo logran aunar todo su conocimiento para interpretar el habla. Si, por otro lado, se desea comprender con gran detalle este tipo de aspectos, se recomienda encarecidamente, bajo la sugerencia de los autores de Kaldi, leer el artículo *Speech Recognition with Weighted Finite-State Transducers* [59].

Todos los detalles expuestos en este capítulo toman como referencia principal la publicación elaborada por Gales y Young, conocida como *The Application of Hidden Markov Models in Speech Recognition* [30]. En ella encontramos una explicación minuciosa en torno al reconocimiento del habla mediante estos sistemas tradicionales, así como todos aquellos refinamientos que han hecho posible que éstos adquirieran prestaciones de gran calidad ante escenarios realistas.

6.1. Esquema general del sistema

La arquitectura escogida en este proyecto es conocida en la literatura, por norma general, como un sistema GMM-HMM, a pesar de que éste se encuentre compuesto por otros módulos, tal y como hemos avanzado en la introducción de este capítulo. Concretamente, dispondremos de un GMM-HMM por cada fonema que hayamos definido en nuestra tarea. La combinación de todos estos modelos, encargados de identificar las unidades básicas del lenguaje, constituye el Modelo Óptico. Cabe destacar que este módulo es conocido como Modelo Acústico en el ámbito del reconocimiento del habla a través del audio pero, dada la naturaleza de nuestros datos, se ha optado por bautizarlo de esta manera. No obstante, el sistema se compone de dos modelos adicionales, cuya función es de gran importancia, puesto que permiten simular el habla natural a partir de los fonemas reconocidos por el Modelo Óptico. Estamos hablando, en primer lugar, del Modelo Léxico, encargado, tal y como veremos más adelante, de combinar dichos fonemas con el objetivo de formar las palabras; por último, el Modelo de Lenguaje es el responsable de combinar las palabras proporcionadas por el Modelo Léxico con el objetivo de construir las oraciones finales. Toda la arquitectura esbozada hasta el momento queda reflejada sobre la Figura 6.1.

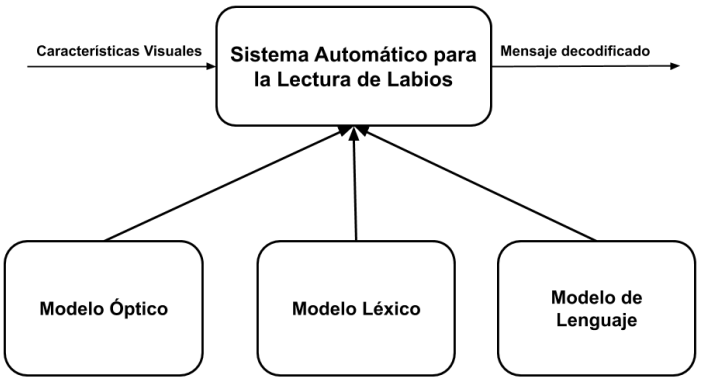

Figura 6.1: Esquema de un Sistema Automático para la Lectura de Labios

Como bien sabemos, y bien sugiere la Figura 6.1, nuestro sistema es alimentado con las características visuales extraídas a partir de la señal visual. Cada muestra del *corpus* está constituida por una secuencia de vectores de características, cuya longitud dependerá del número de *frames* presentes en el vídeo. Esta secuencia puede representarse del siguiente modo: $\boldsymbol{Y} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, donde T se correspondería con el número de *frames*. Por otro lado, el resultado proporcionado por este tipo de sistemas debe ser una oración o secuencia de palabras que denotaremos como $\hat{\boldsymbol{w}} = w_1, \ldots, w_L$, donde L indica el número total de palabras en el mensaje decodificado, el cual deseamos que sea lo más parecido posible a la transcripción de referencia. De esta forma, a través del formalismo matemático, podemos definir la siguiente expresión

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\{P(\boldsymbol{w}|\boldsymbol{Y})\} \tag{6.1}$$

donde se da a entender que el sistema tratará de encontrar, en función de las caracteristicas visuales $\boldsymbol{Y}$, la secuencia de palabras $\boldsymbol{w}$ que mayor probabilidad presente en base al conocimiento que haya adquirido. No obstante, debido a que el término $P(\boldsymbol{w}|\boldsymbol{Y})$ es díficil de modelar directamente, aplicaremos la Regla de Bayes con el objetivo de construir el sistema bajo una aproximación generativa. En otras palabras, el sistema iría generando, de forma progresiva, una serie de secuencias de palabras, conformando así el espacio de búsqueda. Por cada una de estas secuencias se va calculando su probabilidad acumulada y, dependiendo de este valor y ciertos criterios que después comentaremos en la Sección 6.5, algunas de estas alternativas o secuencias se descartan, aligerando el coste computacional del algoritmo. De este modo, cuando se observe la secuencia de palabras que con mayor probalidad genera la secuencia de características, cesará su búsqueda. Con todo esto, la Ecuación 6.1 quedaría tal que

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\{p(\boldsymbol{Y}|\boldsymbol{w})P(\boldsymbol{w})\} \tag{6.2}$$

donde la probabilidad $p(\boldsymbol{Y}|\boldsymbol{w})$ es determinada por el Modelo Óptico y el término $P(\boldsymbol{w})$ viene proporcionado por el Modelo de Lenguaje.

Por último, es necesario indicar que, para ofrecer sistemas factibles, deberán aplicarse distintas técnicas, como puede ser la Programación Dinámica, con el objetivo de acelerar el cómputo que acarrea este tipo de algoritmos. En nuestro

caso se trata, principalmente, del algoritmo Viterbi [72] o, por otro lado, de las influencias por parte de la filosofía basada en la Ramificación y Poda, tal y como habríamos sugerido previamente al describir el proceso de decodificación.

6.2. Modelo Óptico

La unidad básica de la que se compone el lenguaje es conocida como fonema, permitiéndonos distinguir los diferentes sonidos articulados durante la producción del habla. El reconocimiento de estos fonemas se encuentra vinculado a lo que en la literatura se conoce como Modelo Acústico. Sin embargo, al afrontar en nuestro caso la tarea de la lectura de labios automática, realmente estaríamos trabajando con los visemas como unidad mínima. Por tanto, este módulo pasaría a estar bautizado como Modelo Óptico, puesto que su reconocimiento sería a partir de características procedentes de la señal visual. No obstante, existe una cierta correspondencia o relación entre ambos, tal y como se sugirió al inicio de esta memoria. Debido a ello, se ha optado por mantener los fonemas para poder tratar la decodificación del habla natural. Por lo tanto, cada palabra w que compone la secuencia $\boldsymbol{w}$, mencionada previamente, se descompone, a su vez, en una secuencia de sonidos, expresado del siguiente modo $\boldsymbol{q}^{(w)} = q_1, \ldots, q_{K_w}$, donde K_w sería el número de fonemas que constituyen la palabra en cuestión.

Comenzamos, estableciendo la expresión respecto al primer término expuesto en la Ecuación 6.2, cuyo valor ya indicamos que estaría determinado por el Modelo Óptico. Entonces, en la Ecuación 6.3, observamos cómo la probabilidad de que una secuencia de características visuales $\boldsymbol{Y}$ sea generada por una secuencia de palabras $\boldsymbol{w}$ depende tanto de la probabilidad condicionada de dicha secuencia de características en base a una ristra de fonemas $\boldsymbol{Q}$ como de la probabilidad de que estos fonemas reproduzcan el mensaje $\boldsymbol{w}$. Además, con el objetivo de permitir multiples pronunciaciones, se introduce un sumatorio sobre toda secuencia de fonemas posible. A menudo este aspecto es aproximado mediante el operador *max*, pues normalmente existen pocas alternativas a la hora de pronunciar el mensaje, haciendo el sumatorio previamente indicado fácilmente tratable.

$$p(\boldsymbol{Y}|\boldsymbol{w}) = \sum_{\boldsymbol{Q}} p(\boldsymbol{Y}|\boldsymbol{Q}) P(\boldsymbol{Q}|\boldsymbol{w}) \tag{6.3}$$

El segundo término de la Ecuación 6.3, el cual relaciona el mensaje $\boldsymbol{w}$ con una secuencia de fonemas $\boldsymbol{Q}$, se encuentra determinado por cómo es de probable la reconstrucción, palabra a palabra, de este mensaje a partir de la ristra de fonemas planteada. Este detalle queda integrado en la Ecuación 6.4, donde L hace referencia al número de palabras que constituyen la oración.

$$p(\boldsymbol{Q}|\boldsymbol{w}) = \prod_{l=1}^{L} P(\boldsymbol{q}^{w_l}|w_l) \tag{6.4}$$

Por otra parte, antes de abordar el primer término, es necesario conocer, aunque sea a grandes rasgos, ciertos detalles respecto a los Modelos Ocultos de Markov (HMM, por sus siglas en inglés), ya que éstos constituyen, esencialmente, el Modelo Óptico. De hecho, tal y como introdujimos en la sección anterior, dispondremos de un HMM dedicado al reconocimiento de cada uno de los fonemas definidos en el proyecto, incluidos aquellos dedicados al silencio. Será,

mediante la combinación de éstos, como lograremos interpretar el habla. A fin de cuentas, un HMM se representa mediante un conjunto de estados y transiciones, definiendo lo que se conoce como su topología. Normalmente, la topología convencional en un sistema enfocado al reconocimiento del habla, consiste en un modelo de tres estados con transiciones consigo mismo y el siguiente estado, tal y como sugiere la Figura 6.2. Las transiciones se encargarían de modelar la dinámica temporal contenida en la propia naturaleza de las características visuales, permitiendo, gracias a los bucles, el ajuste o alineamiento con pronunciaciones de duración variable, ya que no siempre hablamos a la misma velocidad. Por otro lado, la razón por la que se ha optado por emplear tres estados gira en torno a la influencia que sufre la producción del fonema por parte de su contexto circundante. De esta forma, el cuerpo principal del fonema propiamente dicho se encontraría en el estado central, mientras los otros estados aproximarían las diferentes versiones con las que empezaría o finalizaría la pronunciación de éste, dependiendo de los fonemas entre los que se encuentre. No obstante, podemos apreciar un cuarto estado, representado con un tamaño más reducido, cuya única función consiste en permitir la combinación de los HMMs definidos en el sistema. De este modo es como obtendríamos la secuencia de fonemas resultado de la predicción.

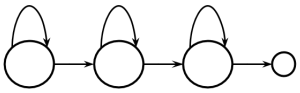

Figura 6.2: Topología clásica de un HMM dedicado al reconocimiento del habla

Sin embargo, ¿cómo es posible que los HMMs modelen este comportamiento? Pues bien, como sabemos, estos sistemas son alimentados por la secuencia de características visuales $\boldsymbol{Y} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, así como por la transcripción asociada a ésta. Eso sí, con el objetivo de mantener la coherencia a la hora de estimar el Modelo Óptico, estas transcripciones han sido convertidas a su correspondiente secuencia de fonemas. Teniendo en cuenta estos detalles, nos disponemos a entrenar nuestro modelo. Para entender este proceso, es necesario conocer que su principal propósito es aprender cómo debe alinear los HMMs para que éstos logren ajustarse, de la mejor forma posible, con cada una de las muestras que componen el *corpus*. Para ello, dado que al principio no dispone de ningún conocimiento a priori, divide a partes iguales la secuencia $\boldsymbol{Y}$ entre los HMMs que representan la transcripción fonética asociada. De esta forma puede suponer, en un principio, con qué *frames* o vectores de características debe alimentar a cada uno de los HMMs. Para que esto sea posible, cada vez que el modelo transita a un estado, ya sea el siguiente o él mismo, se consume un *frame* de la secuencia de características. Será entonces, a partir de varias pasadas sobre el *corpus* de entrenamiento, cómo este modelo irá estimando sus parámetros con el objetivo de ir afinando el alineamiento temporal. Para ello, los HMMs disponen de dos tipos de parámetros, tal y como se sugiere en la Figura 6.3. Por un lado, tenemos la probabilidad de transitar de un estado i a otro j, indicado en la figura bajo la nomenclatura a_{ij}, mientras que, por otra parte, tenemos la probabilidad de emitir un *frame* y_t desde un estado j, indicado, en este caso, como $b_j(y_t)$. A diferencia de las probabilidades de transición que se estiman mediante una aproximación basada en el conteo, cada estado debe emitir un conjunto de in-

formación que yace sobre un espacio de representación contínua. Por ello, este tipo de modelos se conocen como HMM Continuos. De hecho, para poder modelar este tipo de distribuciones de probabilidad emplearemos los Modelos de Mixturas de Gaussianas (GMM, por sus siglas en inglés), tal y como describiremos en la Sección 6.2.1, constituyendo de este modo el modelo GMM-HMM mencionado a lo largo del proyecto.

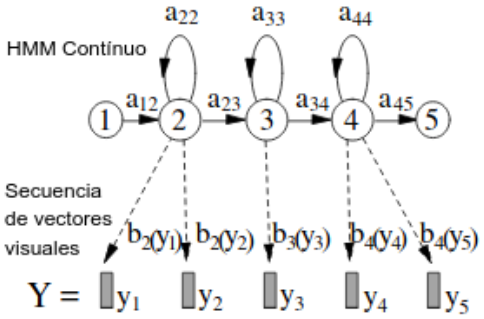

Figura 6.3: Estructura y modelado temporal de un GMM-HMM [30]

A raíz de este comportamiento, surgen dos aspectos que debemos tener en cuenta cuando construimos este tipo de sistemas. Se trata de las siguientes asunciones de independencia condicional:

- En primer lugar, la transición a un estado concreto es, dado el estado previo, condicionalmente independiente del resto de estados que hayan sido visitados durante el proceso.
- Del mismo modo, la emisión producida en un instante concreto es, dado el estado donde se genera, condicionalmente independiente del resto de emisiones.

Entonces, tras conocer el comportamiento que gobierna este tipo de modelos morfológicos, la secuencia de fonemas $\boldsymbol{Q}$, anteriormente mencionada, puede interpretarse como una concatenación o combinación de HMMs, ya que cada uno de éstos está asociado con un fonema en concreto. De esta forma, logramos conformar un modelo compuesto gracias a los estados sin emisión comentados previamente. Con todo esto, ya podemos definir la probabilidad óptica como:

$$p(\boldsymbol{Y}|\boldsymbol{Q}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta}, \boldsymbol{Y}|\boldsymbol{Q}) \tag{6.5}$$

donde $\boldsymbol{\theta} = \theta_0, \theta_1 \ldots, \theta_{T+1}$ representa la secuencia de estados visitada a través del modelo compuesto $\boldsymbol{Q}$ que logra generar la secuencia de características $\boldsymbol{Y}$, la cual no tiene por qué ser única, tal y como sugiere el sumatorio introducido. Observamos que ha sido necesario introducir dos entradas adicionales: θ_0 y θ_{T+1}, las cuales se corresponden con los estados carentes de emisión que constituyen el inicio y final de la secuencia. Al final, si ahondamos sobre esta ecuación, podemos intuir cómo el término $p(\boldsymbol{\theta}, \boldsymbol{Y}|\boldsymbol{Q})$ estaría determinado por las probabilidades contenidas en el proceso que ha hecho posible construir la secuencia de HMMs,

es decir, las transiciones y emisiones de éstos. De este modo, se ha obtenido la expresión reflejada en la Ecuación 6.6, donde recorremos la secuencia hasta alcanzar el estado final θ_{T+1}.

$$p(\boldsymbol{\theta}, \boldsymbol{Y}|\boldsymbol{Q}) = a_{\theta_0\theta_1} \prod_{t=1}^{T} b_{\theta_t}(\boldsymbol{y_t}) a_{\theta_t\theta_{t+1}} \tag{6.6}$$

En último lugar, faltaría por exponer el proceso por el cual es posible la estimación de los parámetros ópticos $\boldsymbol{\lambda} = [\{a_{ij}\}, \{b_j()\}]$. No obstante, dado que no lo consideramos un factor tan esencial para los propósitos del proyecto, una explicación detallada se encuentra en el artículo que ha constituido la principal referencia de este capítulo [30], donde aprenderíamos cómo el algoritmo *forward-backward* [5], cuyo origen reposa sobre la popular técnica *expectation-maximisation* (EM) [23], logra estimar eficientemente estos parámetros a partir de un *corpus* apropiadamente recopilado. No obstante, un factor que sí debemos tener en cuenta es que hemos descrito lo que se conoce como un *modelo monofónico*, ya que trataríamos de interpretar el habla a través de fonemas independientes del contexto. Este hecho supone problemas a la hora de capturar las variaciones dependientes del contexto producidas durante la locución de un discurso. Una forma de mitigar este problema sería considerar un HMM por cada posible contexto que puede envolver a un fonema, pasando de N modelos, como sería el caso descrito a lo largo de la sección, a N^3 HMMs, como por ejemplo podría ser el modelo */p/+/**a**/+/r/*. Este tipo de sistemas son conocidos como *modelos trifónicos* y no han podido evaluarse en este proyecto, ya que no disponíamos de una cantidad suficiente de datos que pudiera estimar adecuadamente tal cantidad de modelos (aunque, en numerosas ocasiones, algunos de éstos compartirían sus parámetros tras un proceso de *clustering*). Y es que, a pesar de tener cerca de 4 horas de material audiovisual, no podemos entrenar adecuadamente todos estos modelos con un ratio de frecuencia de 30 *frames* por segundo, puesto que no alcanzaríamos ni el medio millón de *frames*. Esto no hubiera ocurrido si se hubiera tratado de datos puramente acústicos, cuya frecuencia de extracción genera 100 *frames* de información por segundo. Debido a todo esto, este proyecto únicamente considera los modelos *monofónicos*.

6.2.1. Modelos de Mixturas de Gaussianas

En nuestro caso, tal y como adelantamos en la sección anterior, los estados que componen los HMMs descritos deben emitir, en cada instante de tiempo, un conjunto de información continua. Concretamente, tratarán de replicar los vectores de características visuales que hayamos definido, siendo el objetivo final transitar entre los diferentes HMMs a fin de generar la secuencia de palabras más probable en base a las características visuales que haya recibido. Esto es posible mediante el empleo de Modelos de Mixturas de Gaussianas (GMM, por sus siglas en inglés), ya que modelar esta distribución de probabilidad con una única Gaussiana sería una tarea compleja. Estamos hablando de intentar aproximar o capturar la naturaleza con la que evolucionan las características visuales y esto requiere de modelos más complejos y potentes como pueden ser los GMMs, dotados de una gran flexibilidad que le permite modelar datos caracterizados por ser asimétricos y multi-modales. De hecho, estas características estarían presentes en las tareas de las Tecnologías del Habla, donde encontramos múl-

tiples *speakers* con diferentes acentos, fisionomías o costumbres a la hora de comunicarse.

Con todo esto, introducimos la Ecuación 6.7 para definir la distribución de probabilidad $b_j(\boldsymbol{y})$ respecto a la emisión de un vector de características visuales $\boldsymbol{y}$ desde un estado j. En ella, observamos cómo todas las Gaussianas $\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}^{(jm)}, \Sigma^{(jm)})$ estarían involucradas en el cómputo de dicha distribución, eso sí, ponderadas por un peso c_{jm} estimado adecuadamente para un estado j y Gaussiana m concretos. Además, este último componente debe satisfacer la condición de ser consistente, tal y como sugiere la Ecuación 6.8. De hecho, este factor de ponderación se estima de una forma similar a como son estimadas las probabilidades de transición, es decir, mediante una filosofía de conteo. Por otro lado, en cuanto al aprendizaje general de este tipo de modelos GMM, es necesario saber que sus bases se encuentran respaldadas por el algoritmo EM, mencionado con anterioridad durante este capítulo.

$$b_j(\boldsymbol{y}) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}^{(jm)}, \Sigma^{(jm)}) \tag{6.7}$$

$$\sum_{m=1}^{M} c_{jm} = 1, \quad c_{jm} \geq 0 \tag{6.8}$$

Esta incorporación fue uno de los aspectos que propulsó la construcción de sistemas de gran calidad enfocados al reconocimiento del habla, alcanzando el estado del arte. No obstante, también influenciaron otras técnicas que lograron refinar este tipo de sistemas hasta lo que hoy en día conocemos dentro del paradigma tradicional. Todas ellas son mencionadas con cuidadoso detalle en el artículo que ha constituído la principal referencia del capítulo [30]. Estaríamos hablando de técnicas relacionadas con aspectos respecto a la normalización, la modelización de las covarianzas y la duración del discurso, así como una distinción en función del género del *speaker* y proyecciones sobre las características que componen el conjunto de datos empleado.

6.2.2. Beneficios de los coeficientes Delta-Delta

Como bien sabemos, la extracción de características pretende proporcionar una representación compacta de la señal en cuestión que estemos tratando. Esta representación debería proveer de información capaz de dotar al sistema del conocimiento para discernir entre las palabras que componen el mensaje contenido en la señal. Además, deberían tener en cuenta cómo van a ser modeladas por el sistema. Por ejemplo, si se emplean distribuciones Gaussianas diagonales entonces las características extraídas en cada *frame* no deberían presentar correlación alguna. No obstante, existe otro factor que puede llegar a influir notablemente sobre las prestaciones de nuestro sistema: las asunciones de independencia condicional que implican los HMMs y que ya comentamos en la Sección 6.2. Entonces, con el objetivo de compensar estas limitaciones, se ha optado por estudiar la metodología empleada a la hora de extraer las características en el reconocimiento del habla acústico. Nos referimos al popular esquema de codificación *Mel-Frequency Cepstral Coefficients* (MFCC) [21].

Resulta que estos coeficientes MFCC no lograban paliar las limitaciones que suponían las asunciones anteriormente citadas. Para ello, fue necesario introdu-

cir sobre cada uno de los vectores $\boldsymbol{y_t}$ de la secuencia de características completa $\boldsymbol{Y}$ los coeficientes conocidos como *delta-delta* ($\Delta\Delta$). Esto implica aumentar la dimensión de estos vectores de características, alcanzando 54 componentes en el caso de que empleásemos las características geométricas constituidas por 18 métricas. Al final, estos coeficientes son las derivadas de primer y segundo orden. De hecho, lo podemos comprobar en la Ecuación 6.9, definida a partir de los detalles expuestos en el *HTK Book* [88] (herramienta predecesora al desarrollo de Kaldi), donde observamos cómo el coeficiente *delta* en un *frame* concreto d_t depende, principalmente, de las diferencias presentes entre los valores contenidos por los vectores de características en distintos instantes de tiempo. Este contexto temporal, tanto en vistas al futuro como al pasado, queda establecido por la variable Θ.

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(y_{t+\theta} - y_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{6.9}$$

De esta forma, siguiendo la fórmula expuesta en la Ecuación 6.9, podemos obtener la derivada de primer orden, también conocida como *velocidad*, a partir de las características originales. A su vez, a partir de la *velocidad* podremos obtener la derivada de segundo orden o *aceleración*. Una vez hayamos realizado todos los cálculos necesarios ya podemos conformar las nuevas características, ahora mucho más completas ya que incorporan información relacionada con la evolución que sufren a lo largo de un contexto temporal, paliando, de este modo y en cierta medida, las asunciones que implica el empleo de HMMs. Es por ello que este tipo de aproximación temporal será estudiada también dentro de nuestro proceso experimental.

6.3. Modelo Léxico

Sin embargo, no podemos abordar la construcción de oraciones o frases partiendo directamente desde los fonemas reconocidos por el Modelo Óptico. En otras palabras, se requiere de un módulo que haga de intermediario y permita llevar a cabo nuestros objetivos finales. Estamos hablando del Modelo Léxico, responsable, tal y como se mencionó previamente, de construir las palabras a partir de los fonemas obtenidos en la etapa anterior. Al final, se trata de un modelo de estados finitos, encargado de relacionar ciertas secuencias de fonemas con su correspondiente palabra. Por lo tanto, el tamaño de éste dependerá del vocabulario que estemos empleando. Por otro lado, para su correcto funcionamiento debemos hacer frente a un incoveniente: aquellas palabras que presenten una misma representación fonética, como por ejemplo “vaca” y “baca”, cuya secuencia de fonemas sería "b a k a". Por lo tanto, con el objetivo de solventar este problema, se añaden los llamados símbolos de desambiguación, dotando al sistema de una robustez frente a este tipo de situaciones.

6.4. Modelo de Lenguaje

El Modelo de Lenguaje, tal y como se ha avanzado en la Sección 6.1, es uno de los pilares fundamentales de este tipo de sistemas, ya que es el responsable de combinar las palabras proporcionadas por el Modelo Léxico y generar, de este

modo, el mensaje que el sistema estaría interpretando a partir de la señal visual. En nuestro caso, implementamos un Modelo de Lenguaje basado en los denominados *n-gramas*, gracias a las herramientas proporcionadas por el *toolkit* SRILM [77], conocido ampliamente en el ámbito de la Lingüística Computacional.

Si entendemos $\boldsymbol{w} = w_1, \ldots, w_L$ como una oración o secuencia de L palabras, la probabilidad a priori, requerida en la Ecuación 6.1, quedaría definida como

$$P(\boldsymbol{w}) \approx \prod_{l=1}^{L} P(w_l | w_{l-1}, w_{l-2}, \ldots, w_{l-N+1}) \tag{6.10}$$

donde observamos cómo la probabilidad de una secuencia de palabras estaría determinada por la probabilidad que posee cada una de las componentes o palabras. A su vez, este último término nos muestra cómo la probabilidad de una palabra concreta w_l depende de las $N-1$ palabras anteriores o, mejor dicho, depende de su historia. De esta forma, recorriendo el conjunto de textos dedicados al entrenamiento de estos sistemas, se estiman adecuadamente las cuentas vinculadas a cada ocurrencia de *n-gramas* que se vaya encontrando. Por ejemplo, si definimos un Modelo de Lenguaje basado en *3-gramas*, la probabilidad de una palabra concreta w_l sería, aproximadamente, la relación entre el número de veces que se ha observado dicha palabra junto a una historia determinada $w_{l-2}w_{l-1}$ y las veces que se ha contemplado, a lo largo del texto completo, la historia correspondiente, tal y como se muestra en la Ecuación 6.11.

$$P(w_l | w_{l-1}, w_{l-2}) \approx \frac{C(w_{l-2}w_{l-1}w_l)}{C(w_{l-2}w_{l-1})} \tag{6.11}$$

No obstante, como bien podemos suponer, nuestro *corpus* de entrenamiento no puede contener toda la casuística de un lenguaje. Por lo tanto, tras el aprendizaje suelen producirse un gran número de sucesos no contemplados que adquirirán una probabilidad nula, pudiendo provocar, por ejemplo, que ciertas frases bien construidas sean consideradas como erróneas. Este problema es resuelto mediante técnicas de suavizado que se encargan de descontar una cierta cantidad de probabilidad de los sucesos observados para distribuirla posteriormente, en base a unos criterios establecidos, entre los sucesos no contemplados. Uno de los suavizados más conocidos en el ámbito es la técnica *Backoff*, así como el uso de descuentos *Witten-Bell* o *Kneser-Ney*.

En último lugar, este tipo de modelos son evaluados mediante una métrica denominada *perplejidad*, cuya fórmula se expone en la Ecuación 6.12 y donde L, recordamos, sería el número de palabras presentes en el texto completo. Para comprender mejor la interpretación de este indicador es necesario contemplar el Modelo de Lenguaje como un generador de texto automático. Entonces, cuando nos encontremos en un momento determinado, tendremos en cuenta tanto la última palabra escrita como la historia que le acompaña para decidir cuál es la palabra más probable con la que poder continuar el texto. De este modo, la *perplejidad* es, a menudo, interpretada como el número de alternativas entre las que podemos elegir a la hora de predecir la siguiente palabra del mensaje.

$$\begin{aligned} H &= -\lim_{L\to\infty} \frac{1}{L} \log_2(P(\boldsymbol{w})) \\ &\approx -\frac{1}{L} \sum_{l=1}^{L} \log_2(P(w_l | w_{l-1}, w_{l-2}, \ldots, w_{l-N+1})) \end{aligned} \tag{6.12}$$

6.5. Conceptos respecto a la parametrización

A medida que hemos avanzado por el capítulo se ha ido describiendo cómo funcionan este tipo de sistemas enfocados al reconocimiento del habla. No obstante, con el objetivo de poder comprender de una forma más apropiada la experimentación expuesta en el Capítulo 7, es necesario conocer los parámetros respecto al entrenamiento del Modelo Óptico y la decodificación del mensaje contenido en la señal visual.

Comenzaremos describiendo los parámetros dedicados a la decodificación. Este proceso presenta una alta complejidad, puesto que supone la combinación de todos los módulos mencionados a lo largo del capítulo con el propósito de aunar el conocimiento necesario para interpretar el habla. Para ello, tal y como sugiere el Anexo B, se compone una estructura denominada *lattice*, donde cada modelo, definido a través de los conocidos transductores de estados finitos, se encontraría combinado con el resto, conformando así el sistema en su totalidad. No obstante, sobre éste identificamos numerosos parámetros que permiten ajustar el sistema a nuestros propósitos o a la naturaleza de nuestros datos. Entre ellos destacamos los siguientes:

- Ancho de búsqueda (*beam*): como bien sabemos, la decodificación estaría gobernada por el algoritmo de Vitebi [72], el cual se encarga de ir expandiendo progresivamente todo camino o secuencia de estados capaz de generar la señal visual recibida. Aquel que consiga generar la señal con la mayor de las probabilidades proporcionará, en base a los HMMs visitados, el mensaje asociado. No obstante, este proceso puede resultar prácticamente intratable debido al gran coste computacional que implica explorar un espacio de búsqueda tan amplio. Por ello aparece el parámetro conocido como *beam*. Dado que conocemos en todo momento la probabilidad parcial de los caminos a medida que se van expandiendo, podemos aplicar, tal y como se ha avanzado durante este capítulo, una filosofía basada en la Ramificación y Poda. Por lo tanto, el *beam* no es más que un valor que nos permite descartar aquellos caminos cuya probabilidad parcial no es prometedora, es decir, aquellos caminos, cuya probabilidad sea *beam* veces menor que la mayor probabilidad en el momento actual. De este modo, a mayor valor el sistema proprocionará predicciones más precisas pero será más lento, mientras que a menor valor aceleraremos el proceso de decodificación a costa de deteriorar la calidad de las predicciones.

- Malla o *Lattice*: este parámetro también estaría relacionado con el espacio de búsqueda y cómo limitar su coste computacional. Concretamente, se le asigna un valor entero con el que indicamos al sistema cuántos caminos alternativos, como máximo, debe mantener durante la búsqueda. En cuanto a sus consecuencias, observamos una analogía respecto al parámetro *beam*.

- Influecia del Modelo de lenguaje (*grammar-scale-factor*): viene a ser la influencia que presenta el Modelo de Lenguaje sobre la decodificación del mensaje. De esta forma, conseguimos que sistemas donde el Modelo Óptico esté pobremente estimado sea posible mejorar la calidad de sus predicciones.

- Influencia del Modelo Óptico (*optical-scale*): de un modo similar, la herramienta Kaldi también proporciona un factor relacionado con la influencia del modelo morfológico o Modelo Óptico, aunque este factor se encuentra en un rango distinto al valor que suele adquirir el parámetro anterior.

- Penalización de inserción (*word-insertion-penalty*): cuando hablamos no existen pausas significativamente notables que puedan indicar al sistema los limites entre las palabras del mensaje a decodificar. Por lo tanto, nuestro sistema puede proveer predicciones segmentando la señal en un número incorrecto de palabras. Por ello, introducimos el *word-insertion-penalty*, un factor con el que se promueve la decodificación de palabras largas frente a varias palabras cortas, aspecto que puede mejorar la decodificación final.

Por otra parte, el entrenamiento del Modelo Óptico viene a estar determinado por una serie de parámetros que guiarán el alineamiento temporal. Existen multitud de parámetros al respecto, incluso relacionados con la topología del HMM, como puede ser definir qué influencia presentarían cada tipo de transición durante la etapa de aprendizaje. No obstante, en este proyecto centraremos nuestra atención sobre los siguientes parámetros:

- Ancho de búsqueda (*beam*): se trata de un parámetro que presenta gran similitud con el homónimo anteriormente descrito. La única diferencia es que la limitación espacial durante la búsqueda no se realiza buscando el mensaje final, sino con la intención de encontrar la mejor forma de encontrar el alineamiento temporal, teniendo constancia de la señal y su transcripción asociada. En Kaldi podemos encontrar, principalmente, dos tipos de *beam*: *regular* y *retry*. El primero de ellos sería el *beam* aplicado, por norma general, durante la etapa de entrenamiento, mientras que el segundo se emplea cuando no se ha podido alcanzar un estado final con el *regular beam*. En ese caso, se volvería a repetir el proceso con un *beam* de mayor valor.

- Número total de Gaussianas (*totgauss*): A priori no podemos conocer cuál es el número más apropiado de componentes gaussianas a combinar en nuestro modelo GMM. Este detalle depende, esencialmente, de la naturaleza que presenten los datos con los que estemos trabajando. Por ello, Kaldi ha implementado una función por la que, en base a ciertos criterios y heurísticas, se va determinando, a medida que se desarrolla el aprendizaje, cuántas Gassianas serían las más adecuadas para modelar la emisión de los estados. Para ello, únicamente debemos especificar el número máximo de componentes que deseamos.

Si se desea conocer con todo lujo de detalles los aspectos de implementación, así como los parámetros que constituyen la herramienta Kaldi [71], se recomienda inspeccionar la documentación oficial[1] y las recetas proporcionadas junto a la instalación del *toolkit*.

[1] https://kaldi-asr.org/doc/

6.6. Evaluación

Una vez tengamos nuestro sistema entrenado es hora de evaluar sus prestaciones, es decir, cómo de bien logra interpretar el habla contínua. Para ello, haremos uso de la métrica conocida como *Word Error Rate* (WER), ampliamente empleada en el ámbito del reconocimiento del habla. Está métrica se define del siguiente modo:

$$WER = \frac{\sum_{i=0}^{I} S + I + D}{\sum_{i=0}^{I} N_i} \cdot 100 \tag{6.13}$$

donde recorremos cada muestra i que constituye el *corpus* de validación, computando, en cada caso, el número de sustituciones (S), inserciones (I) y borrados (D) que han sido necesarios aplicar sobre la predicción del sistema para que ésta fuese idéntica a la referencia. En otras palabras, el cálculo descrito hasta el momento se corresponde con la Distancia de Levenshtein [58]. Posteriormente, dividiremos este valor por el número de palabras que conformaron la secuencia de palabras predicha (N_i). De esta forma, logramos obtener un indicador intuitivo respecto a la precisión con la que nuestro sistema reconoce el habla. Aun así, es necesario saber que no se trata de una probabilidad, pudiendo exceder el 100 % en ciertas ocasiones donde sea necesario un gran número de inserciones.

Por otro lado, a menudo resulta complicado elaborar un método apropiado con el que realizar un análisis coherente y significativo a la hora de comparar las prestaciones de diferentes sistemas. Por ello, con el objetivo de dotar a nuestros experimentos de una evaluación adecuada, respaldaremos nuestra metodología al respecto sobre el artículo *Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation*, publicado por M. Bisani y H. Ney [7], a partir del cual hemos comprendido las razones, a modo de pregunta, por las que debemos aplicar una evaluación que nos proporcione unos intervalos de confianza de gran calidad. Son las siguientes:

- ¿Cómo puede variar la tasa de error si cambia el conjunto de validación?
- ¿Cuánto podemos confiar en la mejora que supone un sistema sobre otro?

La idea central del artículo es extraer muestras del *corpus* de validación aleatoriamente hasta completar B subconjuntos diferentes, en los cuales podemos encontrar muestras repetidas o replicadas. Este último detalle es lo que se conoce como extracción con reemplazamiento. Además, el número de subconjuntos extraídos suele ser del orden de 10^3 o 10^4. Entonces, por cada uno de estos subconjuntos b, computaremos, por cada una de sus muestras, el número de palabras que las compone n_s^{*b}, así como el número de errores encontrados respecto a la referencia e_s^{*b}. De este modo, obtenemos subconjuntos con el siguiente esquema:

$$X^{*b} = (n_1^{*b}, e_1^{*b}), \ldots, (n_s^{*b}, e_s^{*b}) \tag{6.14}$$

a partir de los cuales podremos obtener el denominado *Bootstrap Word Error Rate*, bajo el símbolo WER_{boot}. Este término será simplemente la media sobre

las tasas de error WER^{*b} alcanzadas por cada subconjunto b extraído anteriormente, tal y como sugiere la Ecuación 6.15.

$$WER_{boot} \approx \frac{1}{B} \sum_{b=1}^{B} WER^{*b} \tag{6.15}$$

Podríamos pensar que esta última tasa de error sería igual al WER descrito al inicio de esta sección. No obstante, este no es el caso, sino que existirán normalmente pequeñas variaciones entre ellas.

En cuanto a la incertidumbre o, más bien, el rango en el que puede variar la precisión de nuestro sistema, se calcula a través del error estándar se_{boot}, definido tal y como se expone en la Ecuación 6.16. De esta forma, podemos obtener un intervalo de confianza de calidad sobre la precisión que obtuvimos en un principio, simplemente computando la expresión $WER \pm 1{,}64 se_{boot}$.

$$se_{boot} \approx \sqrt{\frac{\sum_{b=1}^{B} (WER^{*b} - WER_{boot})^2}{B-1}} \tag{6.16}$$

7 | Experimentación

Este capítulo pretende condensar todo el proceso experimental llevado a cabo durante el desarrollo del proyecto. Como bien sabemos, estas pruebas van a ser realizadas haciendo uso del *toolkit* Kaldi [71], el cual ha permitido la construcción de sistemas enfocados a las tecnologías del lenguaje cuyas prestaciones han alcanzado el estado del arte en numerosas ocasiones [62, 83, 67]. Esta herramienta proporciona el soporte para componer sistemas de gran potencial, desde aproximaciones tradicionales hasta modelos híbridos o sistemas basados puramente en *Deep Learning*. En nuestro caso, tal y como se ha podido comprobar en el capítulo anterior, los experimentos van a elaborarse sobre un sistema GMM-HMM monofónico, debido a la insuficiencia de datos disponibles, lo que ha imposibilitado la estimación de modelos dependientes del contexto. Respecto a la evaluación de los diferentes experimentos realizados, haremos uso de la conocida métrica WER y la metodología *bootstrap* [7], ya explicadas en la Sección 6.6.

En un principio, se realizaron pruebas explorando, de forma sistemática, diferentes configuraciones en cuanto a los parámetros que principalmente modelan un sistema clásico de reconocimiento del habla. Estamos hablando de parámetros que influyen tanto en la etapa de entrenamiento como en la de decodificación y cuyos detalles expusimos anteriormente en la Sección 6.5. Sin embargo, a pesar de realizar numerosas pruebas, los resultados obtenidos no lograban ser mínimamente aceptables en ningún caso. Además, no reflejaban ningún indicio considerable que nos permitiese extraer algún tipo de conclusión al respecto.

Es por ello que, al final, se optó por definir un conjunto de pruebas en el que redujesemos la complejidad de la tarea. Para ello, se decidió emplear, en esta ocasión, un modelo de lenguaje cerrado, es decir, un Modelo de Lenguaje entrenado únicamente a partir del texto contenido en el conjunto de validación. De esta forma, centramos nuestros experimentos en torno al Modelo Óptico, el módulo encargado de predecir los fonemas; y por lo tanto, también fijamos nuestra atención sobre el estudio o análisis de las características visuales. Una vez hayamos extraído nuestras conclusiones retomaremos el sistema en condiciones reales para volver a realizar una serie de pruebas con ciertos aspectos ya establecidos. Ha sido este proceso experimental el que ha guiado el modo en el que se ha estructurado el capítulo, aunque antes de abordarlo comenzaremos describiendo algunos aspectos que permitan conocer los detalles respecto a las condiciones en las que se han realizado dichas pruebas, como puede ser la partición del *corpus* o el entrenamiento del Modelo de Lenguaje.

7.1. Partición del conjunto de datos

El primero de los pasos es dividir el conjunto de datos recopilado para la ocasión. Se trata, pues, de obtener un conjunto dedicado al entrenamiento del sistema y otro, con un menor número de segundos acumulados, para la validación o evaluación de éste. En nuestro caso, debido a nuestros objetivos, plantearemos la tarea bajo el prisma de la filosofía *speaker independent*. En otras palabras, ningún hablante o *speaker* presente en el conjunto de entrenamiento podrá estar en la partición dedicada a la evaluación del sistema. De esta forma, buscamos un sistema capaz de interpretar el habla producida por una persona que no ha sido tratada durante su aprendizaje, lo que indicaría una generalización de buena calidad ante escenarios realistas.

No obstante, como ya describimos en el Capítulo 3, nuestro *corpus* no presenta una participación relativamente uniforme de todos los hablantes, es decir, existe un desbalanceo. Sobre todo lo observamos con el primero de los *speakers*, ya que es el presentador principal del telediario y, por lo tanto, cubre gran parte del rodaje. Y no sólo eso, sino que otros hablantes apenas logran acumular una porción considerable. En base a estos factores, se han determinado los siguientes detalles respecto a la partición:

- Prescindiremos totalmente del primer *speaker*, puesto que, tras algunas pruebas, se ha constatado que introducir un hablante que presenta tal cantidad de segundos puede desvirtuar el aprendizaje del modelo óptico.
- En cuanto al conjunto de validación, se han escogido *speakers* cuya participación no logra superar los 55 segundos. Al final este conjunto dispone de 14 hablantes, constituyendo alrededor de un total de 455 segundos y 1625 palabras.
- El resto de *speakers*, a excepción del primer hablante, tal y como hemos avanzado en el primer punto, conforman el conjunto de entrenamiento para que el módulo óptico pueda estimar sus parámetros. Este conjunto presenta 43 hablantes, proporcionando cerca de 3 horas de datos.

Tras estos detalles, dado que prescindimos de uno de los *speakers* troncales del *corpus*, han variado ligeramente las estadísticas del conjunto de datos respecto a las expuestas en la Tabla 3.1. En estos momentos destacar, simplemente, que estaríamos tratando con poco más de 3 horas de locuciones, las cuales, a su vez, constituirían un vocacularios 2885 palabras diferentes.

7.2. Entrenamiento del Modelo de Lenguaje

El Modelo de Lenguaje, tal y como se indica en la Sección 6.4 y se demuestra en este capítulo, constituye una de las bases fundamentales en este tipo de sistemas. Del mismo modo, hemos conocido sus parámetros más relevantes: el tamaño del *n-grama*, así como el descuento y suavizado aplicado a la hora de estimar adecuadamente sus parámetros. En nuestro caso, tras realizar una serie de pruebas, hemos fijado el tamaño del *n-grama* en 4 palabras junto a un descuento *Witten-Bell* con suavizado *Backoff*, obteniendo así una perplejidad en torno a un valor de 70 alternativas sobre el conjunto de validación. Si empleábamos otros tamaños, o bien otros tipos de descuento, la perplejidad aumentaba. Este suceso empeoraba la eficiencia a la hora de decodificar el mensaje, ya que al contemplar un mayor número de alternativas el algoritmo requería más tiempo para llevar a cabo el cómputo correspondiente sin lograr mejorar la calidad de las predicciones.

Por otro lado, para poder estimar o entrenar este modelo, necesitamos de un *corpus* en el que encontremos una cantidad suficiente de texto que pueda representar la naturaleza del lenguaje deseado, en nuestro caso, el español. Para ello, se ha construido un *corpus* a partir de las transcripciones de numerosos telediarios emitidos por Radio Televisión Española (RTVE), siempre y cuando no coincidieran con el telediario empleado para construir el *dataset* dedicado al Modelo Óptico y estuvieran comprendidos entre las fechas del mismo, tal y como

se describe en el Capítulo 3. No obstante, era necesario pulir ciertos aspectos. Se trata de un preproceso que facilite la integración del sistema, puesto que todo los carácteres serán convertidos a su versión en minúsculas y los números o siglas serán transcritos adecuadamente. De este modo, tras el preproceso y al tratarse de un dominio cercano, entrenamos este modelo con un texto similar al presente en las transcripciones del *corpus* audiovisual.

En último lugar, cabe destacar que se realizaron experimentos combinando este *corpus* con datos de mayor tamaño, como pueden ser las transcripciones proporcionadas por el Parlamento Europeo [43]. Sin embargo, el dominio del texto adquiría ya un carácter más amplio, aumentando así la perplejidad del modelo. Por el contrario, cuando experimentamos con un Modelo de Lenguaje cerrado, como es el caso de la siguiente sección, la perplejidad de éste descendía a mínimos, alcanzando un valor alrededor de 2 alternativas.

7.3. Pruebas con Modelo de Lenguaje Cerrado

Como se ha avanzado anteriormente, debido a las dificultades para extraer conclusiones mediante la implementación de un sistema enfocado a un escenario de carácter realista, se optó por reducir o relajar la complejidad que presenta la tarea. Para ello, una forma que llamó nuestro interés fue emplear un modelo de lenguaje cerrado que, tal y como se ha descrito en la introducción de este capítulo, no es más que un modelo de lenguaje entrenado (siguiendo las pautas indicadas en la Sección 6.4) únicamente con el texto incluido en el conjunto de validación. De este modo, no sólo podemos elaborar un informe respecto a qué características visuales, ya sea de forma independiente o combinada, logran representar con mayor o menor éxito el movimiento labial, sino que podremos determinar también cuál sería la topología o arquitectura más adecuada de cara a la construcción del HMM. A fin de cuentas, hemos liberado al sistema de una gran variedad de alternativas a la hora de predecir el mensaje, centrando nuestros experimentos en el modelo óptico. Estas pruebas no dejan de ser relevantes pues, tal y como sugieren Fernández-López, Oriol Martínez y M. Sukno en una de sus publicaciones [26], un reconocimiento aceptable a nivel de fonemas no implica, necesariamente, prestaciones de buena calidad cuando se trata de decodificar el mensaje a nivel de palabras.

Por otro lado, en este período de pruebas no se exploran distintas formas de configurar los parámetros típicos en un sistema de reconocimiento del habla. Más bien, se ha decidido conveniente mantener los valores por defecto establecidos en las plantillas proporcionadas por la herramienta Kaldi, salvo por una excepción. Debido a la complejidad que supone modelar el silencio a través de la señal visual, se ha prescindido de los fonemas enfocados a su reconocimiento, detalle del que ya hicimos mención en la Sección 3.2 pero que también fue planteado por Thangthai en su tesis [80].

7.3.1. Experimentación respecto a la topología del HMM

Comenzaremos estudiando la topología que definirá nuestros HMMs, antes de avanzar al estudio sobre las características visuales. La forma en la que definamos esta topología va a dirigir el aprendizaje de nuestro sistema, ya que el alienamiento, tal y como describimos en la Sección 6.2, depende principalmente

de este factor. Entonces, teniendo en cuenta que la señal visual es capturada con un ratio de frecuencia menor al que observamos cuando se trata de datos acústicos, solamente se estudiarán topologías con tres estados o menos. Además, por la misma razón, se examina cómo afecta sobre la calidad del sistema añadir transiciones directas al estado final, permitiendo así alineamientos con mayor flexibilidad. Concretamente, se han estudiado las topologías reflejadas en la Figura 7.1, donde cada una de ellas ha sido etiquetada mediante una letra. Si observamos los resultados expuestos en la Tabla 7.1 y su representación gráfica en la Figura 7.2, antes de nada es necesario saber que dichos resultados han sido obtenidos empleando las características geométricas sin ningún tipo de normalización que no sea la que ya comentamos en la Sección 5.1 para reducir el impacto de que un *speaker* estuviera más o menos alejado de la cámara. Con todo esto, a partir de los resultados mencionados, extraemos las siguientes conclusiones:

- En primer lugar, constatamos que añadir saltos al estado final con el objetivo de paliar una frecuencia de *frames* reducida produce mejores resultados sobre el conjunto de *test*. Puede comprobarse si se observa la diferencia, en términos de WER %, existente entre las topologías A y B, así como entre C y D.
- Vemos que con el simple hecho de prescindir de un estado, como ocurre en el caso de las topologías C y D, logramos disminuir la tasa de error. Esta diferencia es sustancial entre A y C, aunque de forma mínima si comparamos B y C. El mejor resultado lo obtenemos si combinamos ambos factores: transiciones al estado final y un estado menos (Topología D), logrando alrededor de un 20 % de mejora respecto a la topología clásica (A). No obstante, si constituimos el HMM con único estado la precisión empeora pero, tal y como refleja la gráfica, no de forma tan significativa como en otros casos.

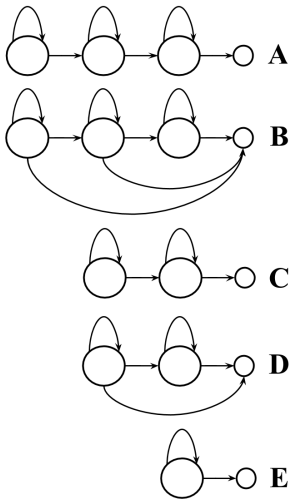

Figura 7.1: Topologías del HMM estudiadas

- Conforme a la conclusión anterior, podemos intuir, a nuestro entender y en base a la interpretación comentada en la Sección 6.2, que cuando

trabajamos con los visemas, al no existir unas diferencias tan notables como en el ámbito de los fonemas, la información contextual se encapsula en mayor grado o, en otras palabras, presenta una mayor influencia sobre la deformación actual que sufren los labios. De ahí que las prestaciones del sistema se beneficien al conformar los HMMs con dos estados en lugar de tres como sería el caso del modelado acústico convencional. O, en lugar de eso, simplemente se trata de una consecuencia debida al presentar un ratio de frecuencia inferior.

Tabla 7.1: Estudio sobre la topología del HMM expresado en WER %

Topología	**A**	**B**	**C**	**D**	**E**
WER %	93.3± 1.6	85.8±7.1	85.6±5.8	**71.8±7.0**	79.6±7.6

* Resultados obtenidos con las *geometricFeats* solamente normalizadas respecto la distancia a la que fue grabado el *speaker*

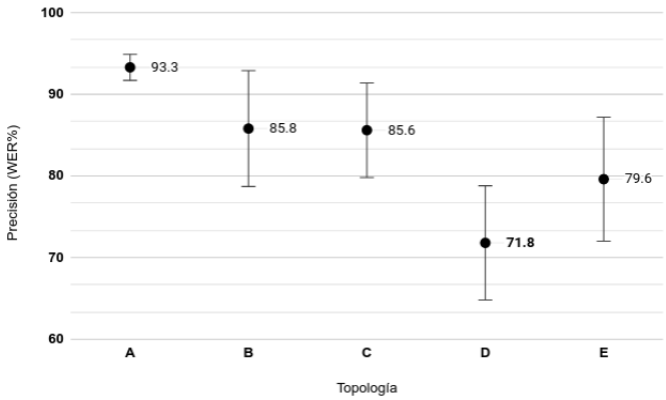

Figura 7.2: Gráfica sobre el estudio de la topología del HMM

7.3.2. Experimentación respecto a las características visuales

De acuerdo con la experimentación detallada en el anterior apartado, todo indica que la topología D proporciona los resultados de mayor calidad y, por lo tanto, podríamos considerarla como la arquitectura que mejor alinea la señal visual. De este modo, teniendo en cuenta dicho factor, centraremos nuestra atención en elaborar un análisis sobre las características visuales que describimos en el Capítulo 5, ya sea, como se ha mencionado en otra ocasión, de forma independiente o combinada, tal y como se muestra en las Tablas 7.2 y 7.3. En ella, además, evaluamos cómo influye sobre los resultados la incorporación de coeficientes temporales, previamente explicados en la Sección 6.2.2. Sin embargo, antes de comentar las conclusiones extraídas al respecto, es necesario exponer ciertos detalles en cuanto a la normalización *z-score*, ampliamente conocida en el ámbito del *Machine Learning*. Son los siguientes:

- *z-score per Speaker*: de esta manera fue como realizaron la normalización Lan y otros autores en una de sus publicaciones [48]. Se trata, pues, de

una normalización que se aplica con el objetivo de disminuir, en la medida de lo posible, las diferencias presentes entre un *speaker* y otro. En nuestro caso, consistiría en menguar las diferencias respecto a la fisionomía bucal.

- *z-score per Utterance*: no obstante, ¿qué ocurre si en cada locución o *utterance*, aunque pertenezcan al mismo *speaker*, se presentan condiciones dispares que pueden alterar el correcto aprendizaje y generalización del sistema? Estamos hablando de cambios en el plano cuando se graba la escena o, incluso, cambios intra-personales, como puede suceder con la barba o gafas. Por ello, con el propósito de reducir los inconvenientes que estos aspectos podrían acarrear, se plantea, además, una normalización por cada una de las locuciones.

Tabla 7.2: Estudio sobre las características visuales normalizadas per *Speaker*. Resultados expreseados mediante la métrica WER %

Normalización *z-score* per *Speaker*				
Características	***raw***	$\Delta\Delta_1$	$\Delta\Delta_2$	$\Delta\Delta_3$
geometricFeats	51.3±8.5	49.2±8.2	**37.7±8.2**	50.5±7.8
eigenLips	71.6±8.6	60.6±8.5	**57.1±8.3**	65.8±7.5
deepFeats	46.6±8.5	**31.7±7.3**	35.9±7.7	39.1±7.8
geometricFeats+eigenLips	29.6±7.5	30.3±6.5	34.2±7.2	**26.6±6.0**
geometricFeats+deepFeats	32.9±7.9	**29.8±7.1**	34.1±7.0	41.3±6.7
eigenLips+deepFeats	45.2±8.2	33.6±7.9	**23.7±6.4**	35.4±7.1
geometricFeats+eigenLips+deepFeats	30.4±6.8	**27.3±6.5**	33.4±6.5	41.5±6.7

raw: se refiere a las características en crudo, sin añadir ningún tipo de coeficiente temporal
$\Delta\Delta_x$: aplica sobre las características los coeficientes *delta-delta* con un contexto de x frames

Tabla 7.3: Estudio sobre las características visuales normalizadas per *Utterance*. Resultados expresados mediante la métrica WER %

Normalización *z-score* per *Utterance*				
Características	***raw***	$\Delta\Delta_1$	$\Delta\Delta_2$	$\Delta\Delta_3$
geometricFeats	46.1±8.8	49.8±8.6	**36.8±7.1**	48.7±8.0
eigenLips	66.6±8.8	**49.9±8.1**	55.6±8.5	53.0±8.4
deepFeats	72.4±7.9	58.9±8.2	54.8±8.9	**54.7±8.2**
geometricFeats+eigenLips	**26.1±6.2**	34.6±7.1	31.2±6.6	34.9±6.4
geometricFeats+deepFeats	**29.3±7.7**	36.5±7.0	33.0±6.9	41.3±7.6
eigenLips+deepFeats	38.0±7.5	29.6±6.8	**26.8±7.2**	31.0±7.2
geometricFeats+eigenLips+deepFeats	34.6±7.7	36.4±6.7	**34.6±6.8**	43.0±6.8

raw: se refiere a las características en crudo, sin añadir ningún tipo de coeficiente temporal
$\Delta\Delta_x$: aplica sobre las características los coeficientes *delta-delta* con un contexto de x frames

Una vez comprendidas las razones que han llevado a explorar estos dos tipos de normalización, ya podemos analizar los resultados obtenidos en la Tablas 7.2 y 7.3, donde estudiamos las características visuales de forma sistemática. Es, a partir de estos datos, de donde extraemos las siguientes conclusiones:

- Lo primero que debemos comentar es que en todos los casos de estudio planteados en la tabla, la normalización *z-score*, ya sea *per Speaker* como

per Utterance, mejora la calidad de las predicciones proporcionadas por el sistema si las comparamos con las obtenidas en la Tabla 7.1, donde empleamos las características planas. De hecho, vemos cómo se ha reducido la tasa de error del 71.8 WER % hasta un 46.1 WER % simplemente normalizando *per Utterance* y sin la necesidad de añadir ningún coeficiente temporal.

- Por otro lado, si fijamos nuestra atención únicamente sobre los experimentos que estudian las características de forma individual, podemos concluir los siguientes aspectos interesantes:

  - En primer lugar, observamos que cuando se trata de una normalización *per Speaker*, las características que mejor resultados proporcionan son las *deepFeats*, alcanzando un 31.7 WER % como mínimo. Estas características estarían seguidas por las *geometricFeats*, las cuales aportan una tasa de error del 37.7 WER %, dejando en último lugar a las obtenidas mediante PCA. De hecho, podemos advertir una dinámica similar en todos los casos, independientemente de si presentan coeficientes temporales o no.
  - Por contra, si realizamos una normalización *per Utterance*, las *deepFeats* son relegadas al último puesto a raíz de un drástico empeoramiento, mientras que tanto *geometricFeats* como *eigenLips* logran mejorar sus resultados anteriores, aunque sea de forma sutil. En esta ocasión, las *geometricFeats* serían las características que mejor representan la señal visual con un 36.8 WER %. No obstante, no alcanzan la tasa de error que destacábamos en el punto anterior.
  - Al final, concluimos que las *deepFeats* son las características que, sin combinarse, proporcionan las predicciones de mayor calidad.
  - No obstante, llama la atención que las características extraídas mediante técnicas *Deep Learning* presenten un comportamiento diferente respecto al resto de aproximaciones. Como hemos expuesto previamente, las *deepFeats*, mejoran sus resultados cuando son normalizadas *per Speaker*. Esto puede significar que, a diferencia del resto de aproximaciones, estas características ya de por sí logran capturar de una forma más apropiada información robusta frente a cambios en el plano o ciertas diferencias intra-personales. De este modo, una normalización *per Speaker* resulta más provechosa.
  - Por último, en cualquier tipo de características queda demostrado que la incorporación de coeficientes temporales afecta favorablemente al modelado visual. Dependiendo de la normalización o la naturaleza de las características será conveniente emplear un contexto de mayor o menor alcance, tal y como reflejan las tablas.

- Siguiendo la misma línea, si estudiamos los resultados obtenidos mediante la combinación de características, identificamos las siguientes conclusiones:

  - A primer golpe de vista, constatamos que la combinación de características, por norma general, produce una mejora a la hora de interpretar el habla visual. Esto nos hace pensar que las características

estudiadas se complementan y logran proporcionar una representación más robusta. No obstante, en ciertas ocasiones estos resultados se solapan, si observamos los intervalos de confianza, con las mejores tasas de error obtenidas cuando no combinábamos las características.

- En cuanto a las características geométricas, determinamos que, a pesar de no ser por una excesiva diferencia, la combinación de éstas con los *eigenLips* proporciona mejores resultados que en el caso de combinarlas con las *deepFeats*. Estos resultados parecidos podría indicarnos que las características profundas y los *eigenLips* fuesen de una naturaleza similar. Sin embargo, este aspecto no sería coherente con futuras deducciones. Por lo tanto, al final, tras la observación anterior de que ambos resultados no distan excesivamente y tras apreciar que, en el caso de combinarlas con las *deepFeats*, la precisión mejora si normalizamos *per Utterance*; concluimos que en ambas combinaciones las características dominantes serían las geométricas.
- Respaldando parte del comentario anterior, comprobamos cómo la combinación de *eigenLips* junto *deepFeats*, obtiene las predicciones con menor tasa de error, alcanzando un 23.7 WER %. Podemos intuir que ambas características se ayudan mutuamente, ya que cada una de ellas, por separado, conseguía sus mejores resultados con una normalización distinta.
- No obstante, en el supuesto de que no empleásemos coeficientes temporales, la combinación de características geométricas junto a los *eigenLips* superaría con creces las prestaciones alcanzadas por la combinación *eigenLips+deepFeats*.
- En último lugar, destacar que la combinación conjunta de todas las características conforma una representación de gran dimensión, hecho que puede provocar dificultades a la hora de modelar los datos estadísticamente. De hecho, a diferencia de lo que ocurría con las características de forma individual, aquí no siempre mejoran los resultados la incorporación de coeficientes temporales. Al final, estamos hablando de 44 componentes por cada *frame* y, aunque los resultados no sean despreciables en cuanto a calidad se refiere, no alcanzan las prestaciones vistas en otros experimentos.

A modo de conclusión, determinamos que la combinación de *eigenLips* junto *deepFeats*, siempre y cuando hagamos uso de los coeficientes Delta-Delta y una normalización *per Speaker*, proporciona las predicciones de mayor calidad. Por lo visto, gracias a las características de apariencia contenidas en los *eigenLips* y el gran potencial que demostraron las *deepFeats* de cara a la reconstrucción (Figura 5.7), lo que nos sugiere que deben poseer ciertos detalles relacionados con la geometría y fisionomía bucal, se ha conseguido una representación de gran calidad.

7.4. Pruebas con Modelo de Lenguaje Abierto

Una vez decidida cuál de las combinaciones de características visuales estudiadas en la sección anterior ha proporcionado la mejor tasa de reconocimiento

y, por lo tanto, entendemos que representa con mayor calidad los movimientos y articulaciones producidas en los labios del emisor, ya podemos retomar el sistema en condiciones realistas, es decir, con un Modelo de Lenguaje abierto, teniendo en cuenta que este modelo ha sido entrenado siguiendo las pautas ya descritas en la Sección 7.2. Por otro lado, es preciso recordar que el silencio ha sido deshabilitado durante el aprendizaje, debido a la complejidad que resulta su modelado, ya comentado a lo largo de la memoria.

Desafortunadamente, no se obtuvieron resultados mínimamente aceptables, alcanzando, en el mejor de los casos, una tasa de error del 91.8 WER %. No obstante, la experimentación no se limitó a una única prueba, sino que se llegaron a estudiar diferentes configuraciones respecto a los parámetros descritos en la Sección 6.5. Se optó por examinar, en primer lugar, los parámetros relacionados con el entrenamiento del Modelo Óptico para después acabar estudiando cómo configurar el proceso de decodificación. Concretamente, se exploraron, por cada uno de estos parámetros, valores cercanos a los establecidos por defecto en una de las recetas de la herramienta Kaldi. Del mismo modo, también se exploraron valores relativamente extremos con la finalidad de apreciar alguna dinámica que pudiera guiar otra serie de experimentos. Sin embargo, tal y como se ha avanzado previamente, no se obtuvieron resultados de una calidad considerable.

8 | Conclusiones

Este capítulo pretende englobar, desde una perpectiva general, las conclusiones extraídas tras el desarrollo del proyecto. Por otro lado, tal y como sugiere la estructura definida, trataremos de describir cuáles han sido, bajo nuestra consideración, los principales factores que han causado los resultados obtenidos dentro de un entorno realista. Más adelante, comentamos cómo los estudios cursados durante la carrera académica han permitido llevar a cabo el trabajo fin de máster. Y por último, se dedica una sección con el objetivo de impulsar y guiar el progreso del proyecto hacia futuras líneas de investigación.

Comenzamos reconociendo la importancia e influencia que ha supuesto el estudio de la literatura respecto a la lectura de labios automática, permitiéndonos conocer en detalle la multitud de aproximaciones y experimentos llevados a cabo en el ámbito. No obstante, centrando nuestra atención sobre el proyecto realizado, la principal conclusión que hemos podido extraer gira en torno a la extracción de características, cuerpo central de éste, tal y como se ha ido mencionando a lo largo de la memoria. Concretamente, se ha realizado un amplio estudio respecto a las características visuales, estudiando aproximaciones basadas tanto en tecnologías tradicionales como en técnicas *Deep Learning*. Estas aproximaciones, enfocadas a aspectos relacionados con la geometría o la apariencia bucal, han permitido explorar diferentes representaciones de la señal visual, ya que incluso se ha analizado su comportamiento cuando éstas eran combinadas entre ellas. De este modo, tras realizar una serie de experimentos con un Modelo de Lenguaje cerrado, pudimos concretar que la combinación de *eigenLips* junto a *deepFeats* proporciona la representación de mayor calidad. Fue entonces, una vez determinado cómo debíamos representar el movimiento labial y cuál debía ser la topología empleada por el HMM, cuando nos adentramos en los experimentos bajo un escenario realista, es decir, estableciendo un Modelo de Lenguaje abierto. En este caso, a pesar de explorar diferentes configuraciones en cuanto a la parametrización del sistema, no se alcanzaron resultados mínimamente aceptables. Las razones que consideramos culpables de este suceso son expuestas y planteadas en la siguiente sección.

No obstante, para poder llevar a cabo toda esta serie de experimentos ha sido necesario el desarrollo de dos aspectos fundamentales. El primero de ellos tiene que ver con la recolección de datos, elemento de vital importancia y que, además, nos ha permitido conocer la complejidad inherente a la tarea. Por ello, cabe destacar que se ha recopilado un *corpus* dedicado esencialmente a la lectura de labios automática. Eso sí, enfocando la recopilación de estos datos al reconocimiento del habla continua en español, constituyendo al final alrededor de cuatro horas de material audiovisual, distribuidas entre numerosos *speakers*. Por otra parte, se ha construido un sistema preliminar dedicado a la lectura de labios automática bajo el paradigma tradicional pero que, desafortunadamente, no ha proporcionado resultados de una calidad aceptable. Si concretamos, se trata de un sistema GMM-HMM monofónico, debido a la insuficiencia de datos a la hora de entrenar un Modelo Óptico dependiente del contexto. Sin embargo, es necesario destacar que en ningún momento, incluso durante el aprendizaje del sistema, se ha empleado información procedente del canal auditivo.

Por último, este proyecto ha permitido, sin lugar a dudas, comprender de una forma completa todo el proceso que conlleva la construcción de un sistema en el ámbito del *Machine Learning*, desde la recogida de datos hasta la obtención y análisis de los resultados obtenidos, además de entender los fundamentos teóricos que soportan este tipo de sistemas.

8.1. Problemas identificados a raíz de los resultados

En el transcurso del proyecto, concretamente en la etapa de experimentación, pudimos observar numerosos problemas que surgían durante el período de aprendizaje del sistema. Problemas que podrían ser los causantes de los desafortunados resultados que se obtuvieron cuando pretendimos tratar la tarea bajo un escenario realista. Además, a éstos se sumarían aspectos relacionados, incluso, con la complejidad de la tarea propuesta o las limitaciones de un paradigma tradicional. De esta forma, se ha optado por destacar, con el fin de guiar futuros desarrollos, los siguientes problemas identificados:

- El primero de ellos se encuentra fuertemente vinculado al ratio de frecuencia con el que se extrae la información procedente del canal visual, ya que esta frecuencia está limitada por las condiciones en las que fue grabado el vídeo. Estamos hablando, en nuestro caso, de 30 *frames* por segundo. Si lo comparamos con la cantidad de información obtenida a partir de la señal acústica, vemos cómo esta última aporta, prácticamente, el triple de información. ¿Qué supone este detalle? Pues supone problemas a la hora de alinear los datos, puesto que, como bien sabemos y se describe en la Sección 6.2, este alineamiento depende, en gran medida, del número de *frames* con los que alimentar los HMMs. De hecho, han sido numerosas las ocasiones en las que hemos observado como algunas muestras no podían ser alineadas adecuadamente, ya que en ella se pronunciaban demasiados fonemas respecto al número de *frames* que representaban la señal y que debían dividirse entre los HMMs. Por ello, se pensó que añadir transiciones a la topología de éstos podría solventar el problema al aportar una mayor flexibilidad, aunque en ciertos casos podría generar alineamientos forzados a causa de una cantidad insuficiente de información y que, en verdad, no se corresponderían con la naturaleza de la señal.

- Otro problema, ya conocido, sería la dificultad que entraña la tarea de leer los labios. Muchos de los detalles al respecto pueden encontrarse en la Sección 3.2 pero debemos destacar la ambigüedad visual, aspecto que agrava sus consecuencias al existir movimientos de lengua o garganta que no pueden ser observados. Por otro lado, nuestro *corpus* presenta un amplio abanico de *speakers* y variabilidad que, igualmente, no es posible capturar de una forma estadísticamente apropiada con la cantidad de datos recopilados hasta el momento.

- Partiendo del anterior punto principalmente, en base al problema de la ambigüedad visual, es necesario indicar que en nuestro proyecto se han realizado las pruebas con un sistema monofónico, limitando considerablemente el modelado relacionado con el contexto y su influencia. De hecho, compartimos la opinión que sugerían Fernández-López y M.Sukno en la revisión que publicaron en 2018 [27]. En ella, respaldaban que las futuras vías de investigación debían versar, principalmente, en el modelado de la información contextual. Es por ello que una de nuestras propuestas es recopilar una mayor cantidad de datos con la finalidad de comprobar cómo el contexto afecta a la hora de decodificar el habla visualmente.

- En último lugar, sabemos que no existe todavía un consenso respecto a cuáles son las características más apropiadas a la hora de representar el habla visual, al contrario de lo que ocurre en el ámbito del reconocimiento del habla acústico. Normalmente, la etapa de extracción de características se realiza independientemente al entrenamiento del sistema, es decir, el sistema trata de adaptarse a los datos que se le han sido suministrados, como es nuestro caso. No obstante, el sistema podría beneficiarse si la extracción de características no fuera estática, sino que fuera estimándose a medida que se van identificando los errores en la transcripción decodificada. Estamos hablando de emplear una arquitectura *end-to-end*, donde todos los parámetros que componen al sistema estuvieran modelados por y para reducir los fallos en la transcripción. Además, podría suponer mejoras respecto al alineamiento temporal.

A pesar de haber obtenido resultados coherentes y de gran calidad cuando estudiábamos las características visuales, los aspectos comentados en esta sección han podido deteriorar las prestaciones del sistema cuando se enfrentaba a un escenario realista, donde el espacio de búsqueda no estuviera tan restringido por el Modelo de Lenguaje. Factor que nos induce a pensar, como principal incoveniente, la existencia de numerosas ambigüedades a la hora de decodificar el habla a través, únicamente, de la señal visual. No obstante, la identificación y consideración de estos y otros aspectos puede guiar y enfocar nuestros futuros desarrollos sobre el campo de investigación.

8.2. Futuras líneas de investigación

La lectura de labios automática, tal y como se ha podido constatar a lo largo de la memoria, es una tarea que continúa abierta a la investigación. A pesar de los resultados obtenidos, en este proyecto hemos centrado nuestros esfuerzos en construir un sistema preliminar, destacando la recopilacion del *corpus* y el estudio realizado respecto a las características visuales. No obstante, tal y como se ha observado durante la descripción del estado arte, quedan muchos aspectos a explorar. Algunos de ellos ya han sido sugeridos en la Sección 8.1 a medida que comentábamos los problemas identificados durante la experimentación. Debido a todas estas razones, se proponen las siguientes líneas de investigación futuras:

- En primer lugar, tampoco debemos dar por sentada la experimentación respecto a las características visuales, ya que existen altenativas que todavía no hemos estudiado. Sería el caso, por ejemplo, de las características basadas en el movimiento, las cuales pueden obtenerse mediante el empleo de conocidas técnicas como *Optical Flow* [78].

- Por otro lado, hacemos referencia al conjunto de datos recopilado. Se trata de continuar aumentando la cantidad de segundos que lo componen, ya no sólo para proporcionar un soporte consolidado a la gran variabilidad presente en este tipo de *corpus* recogidos sin ningún tipo de restricción, sino también con la finalidad de poseer un *corpus* que permita la estimación y el aprendizaje de sistemas basados en técnicas *Deep Learning*, ya que este tipo de sistemas presenta una gran cantidad de parámetros.

- A raíz del plantemiento anterior, podemos deducir que otra vía sobre la que se apoyan nuestros trabajos futuros es la construcción de un sistema *end-to-end*, aprovechando los beneficios que aporta un aprendizaje directo, además de la gran capacidad de abstracción que han demostrado las arquitecturas *Deep Learning*. De hecho, antes de probar este tipo de sistemas, también tenemos la intención de explorar los denominados DNN-HMM, un paradigma que combina el modelo clásico de los HMMs con las *Deep Neural Netwoks*, encargadas de sustituir los modelos GMM a la hora de modelar la generación de la señal visual, ya que, tal y como sugieren Hinton y otros autores [38], este formalismo supone mejoras sobre la mixtura de Gaussianas convencional.

Bibliografía

[1] Nasir Ahmed, T_ Natarajan y Kamisetty R Rao. «Discrete cosine transform». En: *IEEE transactions on Computers* 100.1 (1974), págs. 90-93.

[2] Iryna Anina, Ziheng Zhou, Guoying Zhao y Matti Pietikäinen. «Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis». En: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. IEEE. 2015, págs. 1-5.

[3] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson y Nando de Freitas. «LipNet: Sentence-level Lipreading». En: *ArXiv* abs/1611.01599 (2016).

[4] Enrique Bailly-Bailliére, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad Popovici, Fabienne Porée y col. «The BANCA database and evaluation protocol». En: *International conference on Audio-and video-based biometric person authentication*. Springer. 2003, págs. 625-638.

[5] Leonard E Baum, Ted Petrie, George Soules y Norman Weiss. «A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains». En: *The annals of mathematical statistics* 41.1 (1970), págs. 164-171.

[6] Yoshua Bengio. «Learning deep architectures for AI». En: *Foundations and trends® in Machine Learning* 2.1 (2009), págs. 1-127.

[7] Maximilian Bisani y Hermann Ney. «Bootstrap estimates for confidence intervals in ASR performance evaluation». En: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE. 2004, págs. I-409.

[8] CM Bishop. *Pattern Recognition and Machine Learning, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds*. 2006.

[9] Gary Bradski. «The opencv library». En: *Dr Dobb's J. Software Tools* 25 (2000), págs. 120-125.

[10] William Chan, Navdeep Jaitly, Quoc V Le y Oriol Vinyals. «Listen, attend and spell». En: *arXiv preprint arXiv:1508.01211* (2015).

[11] Alin G Chitu y Leon JM Rothkrantz. «Visual speech recognition automatic system for lip reading of Dutch». En: *Journal on Information Technologies and Control, vol. year vii* 3 (2009), págs. 2-9.

[12] Joon Son Chung, Andrew Senior, Oriol Vinyals y Andrew Zisserman. «Lip reading sentences in the wild». En: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, págs. 3444-3453.

[13] Joon Son Chung y Andrew Zisserman. «Lip reading in profile». En: *British Machine Vision Conference*. 2017.

[14] Joon Son Chung y Andrew Zisserman. «Lip reading in the wild». En: *Asian Conference on Computer Vision*. Springer. 2016, págs. 87-103.

[15] Martin Cooke, Jon Barker, Stuart Cunningham y Xu Shao. «An audio-visual corpus for speech perception and automatic speech recognition». En: *The Journal of the Acoustical Society of America* 120.5 (2006), págs. 2421-2424.

[16] Timothy F Cootes, Gareth J Edwards y Christopher J Taylor. «Active appearance models». En: *European conference on computer vision*. Springer. 1998, págs. 484-498.

[17] Timothy F Cootes, Christopher J Taylor, David H Cooper y Jim Graham. «Active shape models-their training and application». En: *Computer vision and image understanding* 61.1 (1995), págs. 38-59.

[18] Stephen J Cox, Richard W Harvey, Yuxuan Lan, Jacob L Newman y Barry-John Theobald. «The challenge of multispeaker lip-reading.» En: *Proc. International Conference on Auditory-Visual Speech Processing*. Citeseer. 2008, págs. 179-184.

[19] Andrzej Czyzewski, Bozena Kostek, Piotr Bratoszewski, Jozef Kotus y Marcin Szykulski. «An audio-visual corpus for multimodal automatic speech recognition». En: *Journal of Intelligent Information Systems* 49.2 (2017), págs. 167-192.

[20] Navneet Dalal y Bill Triggs. «Histograms of oriented gradients for human detection». En: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, págs. 886-893.

[21] Steven Davis y Paul Mermelstein. «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences». En: *IEEE transactions on acoustics, speech, and signal processing* 28.4 (1980), págs. 357-366.

[22] Kresimir Delac, Mislav Grgic y Panos Liatsis. «Appearance-based statistical methods for face recognition». En: *Proceedings of the 47th International Symposium ELMAR-2005 focused on Multimedia Systems and Applications, Zadar, Croatia*. 2005, págs. 151-158.

[23] Arthur P Dempster, Nan M Laird y Donald B Rubin. «Maximum likelihood from incomplete data via the EM algorithm». En: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), págs. 1-22.

[24] Stéphane Dupont y Juergen Luettin. «Audio-visual speech modeling for continuous speech recognition». En: *IEEE transactions on multimedia* 2.3 (2000), págs. 141-151.

[25] Virginia Estellers y Jean-Philippe Thiran. «Multipose audio-visual speech recognition». En: *2011 19th European Signal Processing Conference*. IEEE. 2011, págs. 1065-1069.

[26] Adriana Fernandez-Lopez, Oriol Martinez y Federico M Sukno. «Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database». En: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE. 2017, págs. 208-215.

[27] Adriana Fernandez-Lopez y Federico M Sukno. «Survey on automatic lip-reading in the era of deep learning». En: *Image and Vision Computing* 78 (2018), págs. 53-72.

[28] Cletus G Fisher. «Confusions among visually perceived consonants». En: *Journal of speech and hearing research* 11.4 (1968), págs. 796-804.

[29] Ivan Fung y Brian Mak. «End-to-end low-resource lip-reading with maxout CNN and LSTM». En: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, págs. 2511-2515.

[30] Mark Gales y Steve Young. *The application of hidden Markov models in speech recognition*. Now Publishers Inc, 2008.

[31] F. A. Gers, J. Schmidhuber y F. Cummins. «Learning to forget: continual prediction with LSTM». En: *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*. Vol. 2. 1999, 850-855 vol.2.

[32] Malik Ghallab, Dana Nau y Paolo Traverso. *Automated Planning: theory and practice*. Elsevier, 2004.

[33] Roland Goecke y J Bruce Millar. «The audio-video Australian English speech data corpus AVOZES». En: *Proceedings of the 8th International Conference on Spoken Language Processing INTERSPEECH*. 2004, págs. 2525-2528.

[34] Alex Graves, Santiago Fernández, Faustino Gomez y Jürgen Schmidhuber. «Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks». En: *Proceedings of the 23rd international conference on Machine learning*. 2006, págs. 369-376.

[35] Mihai Gurban y Jean-Philippe Thiran. «Information theoretic feature extraction for audio-visual speech recognition». En: *IEEE Transactions on signal processing* 57.12 (2009), págs. 4765-4776.

[36] Naomi Harte y Eoin Gillen. «TCD-TIMIT: An audio-visual corpus of continuous speech». En: *IEEE Transactions on Multimedia* 17.5 (2015), págs. 603-615.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren y Jian Sun. «Deep residual learning for image recognition». En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, págs. 770-778.

[38] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath y col. «Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups». En: *IEEE Signal processing magazine* 29.6 (2012), págs. 82-97.

[39] Dominic Howell. «Confusion modelling for lip-reading». Tesis doct. University of East Anglia, 2015.

[40] Dominic Howell, Stephen Cox y Barry Theobald. «Visual units and confusion modelling for automatic lip-reading». En: *Image and Vision Computing* 51 (2016), págs. 1-12.

[41] Vahid Kazemi y Josephine Sullivan. «One millisecond face alignment with an ensemble of regression trees». En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, págs. 1867-1874.

[42] Davis E King. «Dlib-ml: A machine learning toolkit». En: *The Journal of Machine Learning Research* 10 (2009), págs. 1755-1758.

[43] Philipp Koehn. «Europarl: A parallel corpus for statistical machine translation». En: *MT summit*. Vol. 5. Citeseer. 2005, págs. 79-86.

[44] Alex Krizhevsky, Ilya Sutskever y Geoffrey E Hinton. «Imagenet classification with deep convolutional neural networks». En: *Advances in neural information processing systems*. 2012, págs. 1097-1105.

[45] Kshitiz Kumar, Tsuhan Chen y Richard M Stern. «Profile view lip reading». En: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 4. IEEE. 2007, págs. IV-429.

[46] Yuxuan Lan, Richard Harvey, B Theobald, Eng-Jon Ong y Richard Bowden. «Comparing visual features for lipreading». En: *International Conference on Auditory-Visual Speech Processing 2009*. 2009, págs. 102-106.

[47] Yuxuan Lan, Barry-John Theobald y Richard Harvey. «View independent computer lip-reading». En: *2012 IEEE International Conference on Multimedia and Expo*. IEEE. 2012, págs. 432-437.

[48] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong y Richard Bowden. «Improving visual features for lip-reading». En: *Auditory-Visual Speech Processing 2010*. 2010, págs. 142-147.

[49] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu y Thomas Huang. «AVICAR: Audio-visual speech corpus in a car environment». En: *Eighth International Conference on Spoken Language Processing*. 2004, págs. 2489-2492.

[50] Eduardo Lleida, Alfonso Ortega, Antonio Miguel, Virginia Bazán, Carmen Pérez, M Zotano y Alberto de Prada. «RTVE2018 database description». En: *Vivolab and Corporación Radiotelevisión Española, Zaragoza, Spain* (2018). [Online] Available: `http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf`.

[51] Luca Lombardi y col. «A survey of automatic lip reading approaches». En: *Eighth International Conference on Digital Information Management (ICDIM 2013)*. IEEE. 2013, págs. 299-302.

[52] David G Lowe. «Distinctive image features from scale-invariant keypoints». En: *International Journal of Computer Vision* 60.2 (2004), págs. 91-110.

[53] Patrick J Lucey, Sridha Sridharan y David B Dean. «Continuous pose-invariant lipreading». En: *Interspeech September 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia* (2008), págs. 2679-2682.

[54] Iain Matthews, Timothy F Cootes, J Andrew Bangham, Stephen Cox y Richard Harvey. «Extraction of visual features for lipreading». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.2 (2002), págs. 198-213.

[55] Christopher McCool, Sebastien Marcel, Abdenour Hadid, Matti Pietikäinen, Pavel Matejka, Jan Cernockỳ, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy y col. «Bi-modal person recognition on a mobile phone: using mobile phone data». En: *2012 IEEE International Conference on Multimedia and Expo Workshops*. IEEE. 2012, págs. 635-640.

[56] Harry McGurk y John MacDonald. «Hearing lips and seeing voices». En: *Nature* 264.5588 (1976), págs. 746-748.

[57] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin y Gilbert Maitre. «XM2VTSDB: The extended M2VTS database». En: *Second international conference on audio and video-based biometric person authentication*. Vol. 964. 1999, págs. 965-966.

[58] Frederic P Miller, Agnes F Vandome y John McBrewster. «Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? Levenshtein distance, spell checker, hamming distance». En: (2009).

[59] Mehryar Mohri, Fernando Pereira y Michael Riley. «Speech recognition with weighted finite-state transducers». En: *Springer Handbook of Speech Processing*. Springer, 2008, págs. 559-584.

[60] Alfonso Ortega, Federico Sukno, Eduardo Lleida, Alejandro F Frangi, Antonio Miguel, Luis Buera y Ernesto Zacur. «AV@ CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition.» En: *Proc. International Conference on Language Resources and Evaluation*. 2004, págs. 763-767.

[61] Karel Paleček. «Extraction of features for lip-reading using autoencoders». En: *International Conference on Speech and Computer*. Springer. 2014, págs. 209-216.

[62] Vassil Panayotov, Guoguo Chen, Daniel Povey y Sanjeev Khudanpur. «Librispeech: an asr corpus based on public domain audio books». En: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, págs. 5206-5210.

[63] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis y Petros Maragos. «Adaptive multimodal fusion by uncertainty compensation with application to audio-visual speech recognition». En: *Multimodal Processing and Interaction*. Springer, 2008, págs. 1-15.

[64] Dharin Parekh, Ankitesh Gupta, Shharrnam Chhatpar, Anmol Yash y Manasi Kulkarni. «Lip reading using convolutional auto encoders as feature extractor». En: *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. IEEE. 2019, págs. 1-6.

[65] Adrian Pass, Jianguo Zhang y Darryl Stewart. «An investigation into features for multi-view lipreading». En: *2010 IEEE International Conference on Image Processing*. IEEE. 2010, págs. 2417-2420.

[66] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci y John N Gowdy. «CUAVE: A new audio-visual database for multimodal human-computer interface research». En: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. IEEE. 2002, págs. II-2017.

[67] Vijayaditya Peddinti, Daniel Povey y Sanjeev Khudanpur. «A time delay neural network architecture for efficient modeling of long temporal contexts». En: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015, págs. 2440-2444.

[68] Stavros Petridis, Jie Shen, Doruk Cetin y Maja Pantic. «Visual-only recognition of normal, whispered and silent speech». En: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, págs. 6219-6223.

[69] Thomas Plötz y Gernot A Fink. «Markov models for offline handwriting recognition: a survey». En: *International Journal on Document Analysis and Recognition (IJDAR)* 12.4 (2009), pág. 269.

[70] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg y Andrew W Senior. «Recent advances in the automatic recognition of audiovisual speech». En: *Proceedings of the IEEE* 91.9 (2003), págs. 1306-1326.

[71] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz y col. «The Kaldi speech recognition toolkit». En: *IEEE 2011 workshop on automatic speech recognition and understanding*. EPFL-CONF-192584. IEEE Signal Processing Society. 2011.

[72] Lawrence Rabiner y Biing-Hwang Juang. «Fundamentals of speech recognition. Prentice-Hall, Inc.» En: *Upper Saddle River, NJ, USA* (1993).

[73] Mohammad Hasan Rahmani y Farshad Almasganj. «Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features». En: *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*. IEEE. 2017, págs. 195-199.

[74] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou y Maja Pantic. «300 faces in-the-wild challenge: Database and results». En: *Image and vision computing* 47 (2016), págs. 3-18.

[75] Conrad Sanderson. *The vidtimit database*. Inf. téc. IDIAP, 2002.

[76] Themos Stafylakis y Georgios Tzimiropoulos. «Combining residual networks with LSTMs for lipreading». En: *Interspeech August 2017, Stockholm, Sweden* (2017), págs. 20-24.

[77] Andreas Stolcke. «SRILM – An extensible language modeling toolkit». En: *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*. 2002, págs. 901-904.

[78] Deqing Sun, Stefan Roth, JP Lewis y Michael J Black. «Learning optical flow». En: *European Conference on Computer Vision*. Springer. 2008, págs. 83-97.

[79] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke y Andrew Rabinovich. «Going deeper with convolutions». En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, págs. 1-9.

[80] Kwanchiva Thangthai. «Computer lipreading via hybrid deep neural network hidden Markov models». Tesis doct. University of East Anglia, 2018.

[81] Kwanchiva Thangthai, Richard W Harvey, Stephen J Cox y Barry-John Theobald. «Improving lip-reading performance for robust audiovisual speech recognition using DNNs.» En: *AVSP*. 2015, págs. 127-131.

[82] W Freeman Twaddell. «On defining the phoneme». En: *Language* 11.1 (1935), págs. 5-62.

[83] Karel Veselỳ, Arnab Ghoshal, Lukás Burget y Daniel Povey. «Sequence-discriminative training of deep neural networks.» En: *Interspeech*. Vol. 2013. 2013, págs. 2345-2349.

[84] Paul Viola y Michael J Jones. «Robust real-time face detection». En: *International journal of computer vision* 57.2 (2004), págs. 137-154.

[85] Zhou Wang, Alan C Bovik, Hamid R Sheikh y Eero P Simoncelli. «Image quality assessment: from error visibility to structural similarity». En: *IEEE transactions on image processing* 13.4 (2004), págs. 600-612.

[86] Svante Wold, Kim Esbensen y Paul Geladi. «Principal component analysis». En: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), págs. 37-52.

[87] Michael Wooldridge. *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[88] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey y col. «The HTK book». En: *Cambridge university engineering department* 3.175 (2002), pág. 12.

[89] Guoying Zhao, Mark Barnard y Matti Pietikainen. «Lipreading with local spatiotemporal descriptors». En: *IEEE Transactions on Multimedia* 11.7 (2009), págs. 1254-1265.

A | Naturaleza de las características geométricas

En este anexo, se pretende mostrar, con mayor lujo de detalles, un análisis sobre las características geométricas introducidas en la Sección 5.1 [1]. Por ello, se proporciona la Figura A.1. En ella podemos observar la naturaleza con la que evoluciona este tipo de representación a lo largo de un discurso, donde se pronuncia la palabra "completamente". No obstante, como ya sabemos, en nuestro caso disponemos de 18 características por cada instante de tiempo o *frame* y esto puede dificultar la tarea de comprender los datos, por lo que se ha optado por agrupar estas características siguiendo el esquema ofrecido por la Tabla 5.1, donde, por cada *frame*, se combinarán aquellas métricas que compartan un aspecto en común mediante la media aritmética. De esta forma, logramos distinguir seis métricas, tal y como sugiere la leyenda de la gráfica anteriormente citada. Gracias a esta representación logramos identificar los siguientes aspectos:

- En primer lugar, comprobamos que estas características visuales presentan una alta coherencia, puesto que observamos, por ejemplo, cómo la opertura de la cavidad bucal disminuye hasta alcanzar mínimos cuando nos encontramos alrededor del *frame* número 21. Esto nos demuestra que, aunque pueda degradarse esta calidad en muestras donde el locutor se encuentre más lejano a la cámara y, por lo tanto, se complique la tarea de localizar los *landmarks*, por norma general logramos capturar adecuadamente la geometría bucal.

- Por otro lado, prácticamente todas las métricas muestran una varianza razonable. Al igual que la opertura descrita en el ejemplo previo, el resto de medidas proporcionan una fuente de información sustancial. No obstante, cuando hablamos de la distancia entre la barbilla y el contorno inferior de los labios, encontramos una evolución bastante estable, sin identificar cimas o valles pronunciados. Aún así, se ha optado por mantener dicha métrica.

- Otro aspecto interesante es obsevar la relación existente entre la altura y la opertura a lo largo del discurso. Vemos como, aunque pertenezcan a magnitudes diferentes, ambas métricas siguen una naturaleza similar si nos fijamos en los valles y repuntes que se producen durante la locución. Una situación similar ocurre con el área bucal que se encontraría influenciada tanto por la altura como por la anchura que va adoptando la boca.

- Por último, destacar la métrica que se encarga de medir la distancia desde el contorno labial superior y la nariz, puesto que presenta un matiz importante. Observamos cómo esta distancia muestra una relación inversa, principalmente, con la opertura y altura de la boca. Una naturaleza comprensible, ya que es razonable que el labio superior se aproxime a la nariz cuando la altura y cavidad bucales aumenten.

[1] En este análisis tratamos con las características geométricas sin la normalización *z-score*

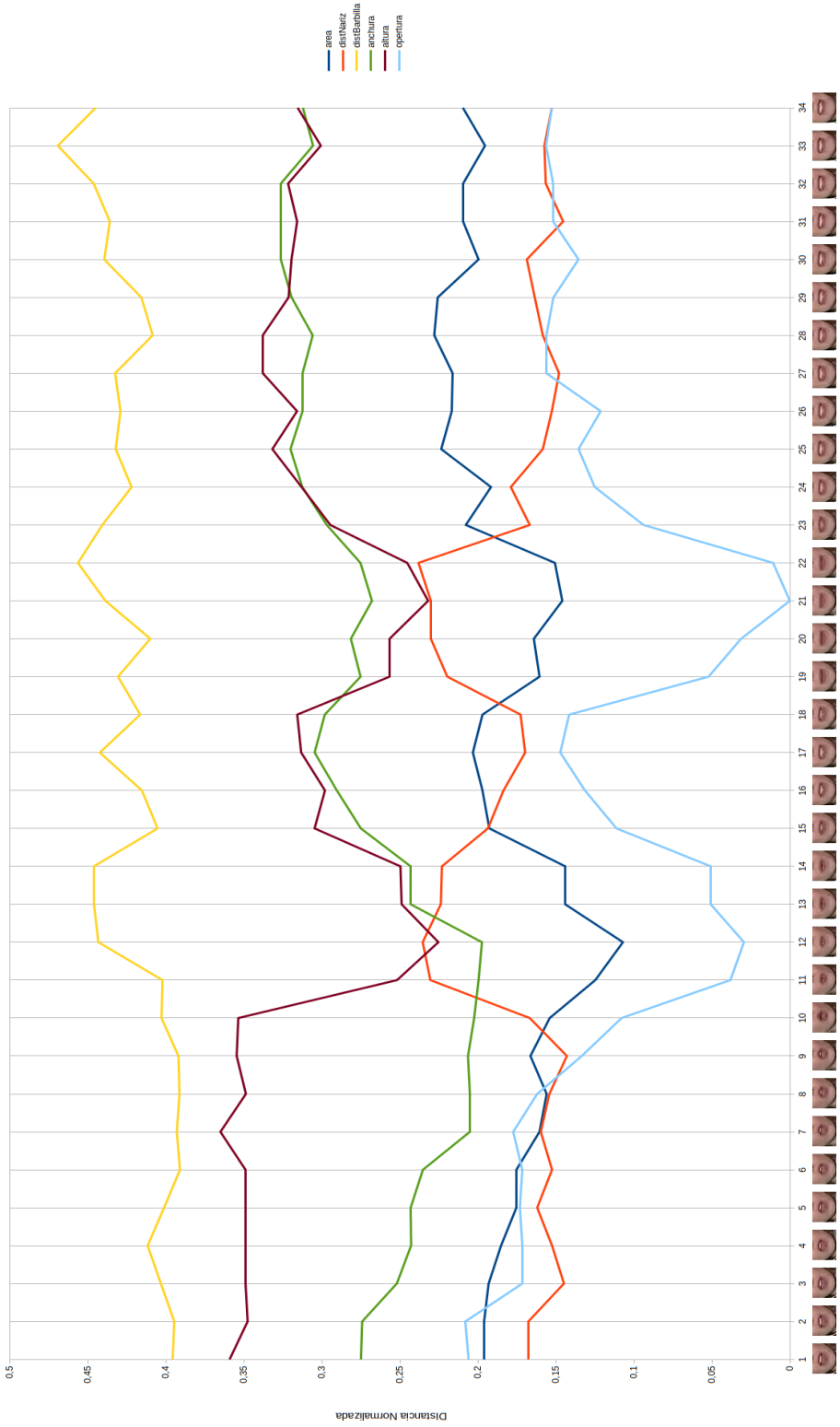

Figura A.1: Naturaleza de las características geométricas

B | Fundamentos de implementación en Kaldi

La construcción de un sistema enfocado al reconocimiento del habla es un proceso complejo, puesto que involucra la participación de múltiples módulos, cada uno encargado de un aspecto diferente pero que en conjunto proporcionan el soporte para interpretar el lenguaje. Como ya introdujimos en el Capítulo 6, nuestro sistema se compone de tres módulos: Modelo Óptico, Modelo Léxico y Modelo de Lenguaje. Estos tres modelos deberán aunar su conocimiento para poder reconocer el habla. Para ello, haremos uso de los conocidos Transductores de Estados Finitos en su versión ponderada, es decir, teniendo en cuenta cierto carácter estocástico. Una vez hayamos definido cada uno de los módulos, conformaremos una malla o *lattice* que combine a todos ellos de una forma eficiente con el objetivo de consolidar sistemas dedicados a escenarios realistas. Por lo tanto, en base a los fundamentos con los que se han implementado el *toolkit* Kaldi, este anexo considera como principal referencia el artículo *Speech Recognition with Weighted Finite-State Transducers* [59], cuya lectura es altamente recomendable si se desea conocer minuciosamente cualquier detalle al respecto.

Estos transductores proporcionan una representación natural para los componentes de este tipo de sistemas, incluidos los HMMs. De hecho, tal y como refleja la Figura B.1, la estructura que definen es considerablemente similar a la forma con la que representábamos la topología del HMM. A fin de cuentas, no es más que un Automáta de Estados Finitos cuyas transiciones se encuentran etiquetadas tanto por un símbolo de entrada como por un símbolo de salida, así como por la probabilidad asociada a esta transición. Concretamente, tal y como observamos en la figura previamente citada, un transductor tiene como propósito codificar una secuencia de símbolos de entrada en otra, siguiendo las reglas o pautas definidas en su arquitectura, siempre y cuando alcancemos un estado final, denotados por un doble círculo. A menudo podemos encontrarnos el símbolo ϵ o $< eps >$, indicando que o bien no se consume ningún símbolo de entrada o, por el contrario, éste es sustituido por vacío, permitiéndonos generar secuencias de una longitud distinta a la original. De esta forma, si introducimos la secuencia *accca* sobre el transductor de la Figura B.1, obtendríamos como salida la cadena *baaaa*. Además, podríamos saber con qué probabilidad fue generada esta secuencia, dependiendo del criterio establecido para su cómputo. Con todo esto, podemos deducir el gran potencial latente en este tipo de formalismos, ya que sobre cada transición podríamos incluir factores que determinasen la influencia de ciertos modelos, así como la penalización que implican ciertos parámetros a la hora de decodificar el mensaje final.

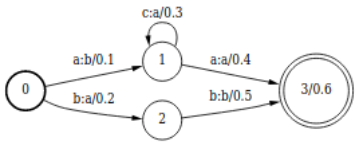

Figura B.1: Esquema de un Transductor de Estados Finitos Ponderado [59]

Por otro lado, en base a los fundamentos matemáticos relacionados con la Teoría de Automátas, seremos capaces de combinar distintos transductores mediante el operador de *composición*. A parte, podremos optimizar el resultado

de esta combinación gracias al operador de *determinación* con el que eliminaremos caminos o rutas redundantes. Y por último, con el mismo propósito que el anterior, aplicaremos el operador de *minimización* que nos permitirá definir un automáta equivalente pero con el menor número de estados y transiciones posibles. Los aspectos en cuanto a estos operadores, así como sus bases matemáticas, pueden inspeccionarse, con todo lujo de detalles, en el artículo que constituye la principal referencia del anexo.

Con todo esto, ya podemos empezar la construcción de nuestro sistema. En primer lugar, definimos el Modelo de Lenguaje o Gramática, denotado, de ahora en adelante, mediante el símbolo G. Tal y como sugiere la Figura B.2, se trata de un transductor encargado de relacionar las palabras con el objetivo de construir las oraciones finales. Las probabilidades que observamos serían inferidas a partir del modelo estimado mediante *n-gramas* que ya describimos en la Sección 6.4.

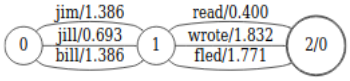

Figura B.2: Gramática definida mediante un transductor [59]

Por otra parte, tendríamos el denominado Modelo Léxico, relacionado, de ahora en adelante, con el símbolo L. Como ya sabemos y podemos deducir a partir del transductor expuesto en la Figura B.3, se trata de un modelo cuyo propósito consiste en vincular una secuencia de fonemas con su correspondiente palabra. En un principio observamos múltiples caminos alternativos o incluso ambigüedades que posteriormente serán solventadas, a menudo, mediante la introducción de los llamados símbolos de desambigüación, cuya nomenclatura comienza mediante el caracter #.

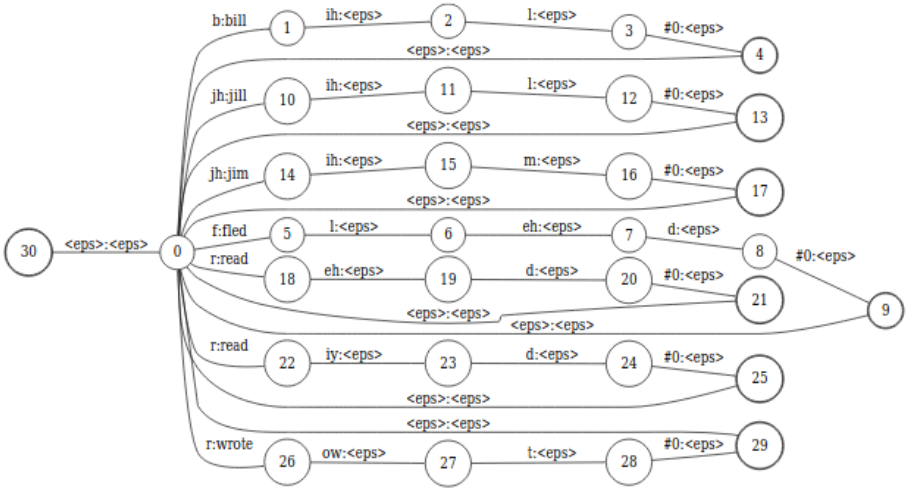

Figura B.3: Módelo Léxico definido mediante un transductor [59]

Suponiendo entonces que realizamos la unión o combinación de estos modelos mediante el operador de *composición*, tal y como sugiere la expresión:

$$L \circ G \tag{B.1}$$

obtendríamos, de esta forma, un transductor resultante que relacionaría las secuencias de fonemas con sus correpondientes secuencias de palabras, del mismo modo que sugiere la Figura B.4. No obstante, seguimos observando caminos ambiguos o redundantes que pueden entorpecer la eficiencia de nuestro sistema final. Por ello, aplicamos sobre este último transductor los operadores de *determinación* y *minimización*, modificando la anterior expresión tal que:

$$min(det(L \circ G)) \tag{B.2}$$

a partir de la cual obtendríamos como resultado el transductor contenido en la Figura B.5, esta vez constituido por el menor número de estados y transiciones posibles, facilitando la compresión de éste.

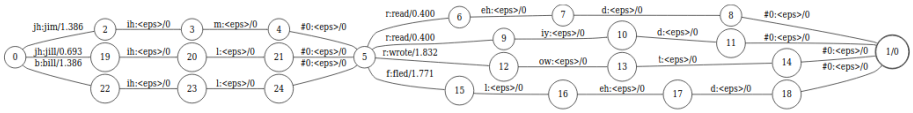

Figura B.4: Combinación de la Gramática y el Léxico [59]

Figura B.5: Optimización del transductor gramático-léxico [59]

No obstante, faltarían dos modelos por introducir antes de definir el sistema implementado en Kaldi. Se trata, en primer lugar, del modelo representado a través del símbolo C, encargado de relacionar los fonemas dependientes del contexto con sus respectivos fonemas independientes. En cierto modo, modela o representa la dependencia contextual. Por otro lado, tendríamos el modelo conocido mediante el símbolo H, ya que su objetivo consiste en modelar la relación existente entre las secuencias de HMMs con las secuencias de fonemas asociadas. Para ello, es necesario etiquetar adecuadamente nuestros modelos fónicos de forma que la cadena devuelta por el transductor nos indique qué HMMs han participado en la decodificación del mensaje final. De este modo, reformulando la Ecuación B.2, nuestro sistema quedaría definido mediante la expresión determinada en la Ecuación B.3, donde se realizarían refinamientos en distintos niveles de la composición. Además, se aplicarían otros detalles no mencionados, pero comentados en el artículo, que mejorarían nuestro sistema.

$$min(det(H \circ det(C \circ det(L \circ G)))) \tag{B.3}$$

A modo de conclusión, hemos sido testigos de la gran ventaja que presenta el empleo de los Transductores de Estados Finitos Ponderados, ya que permiten modelar, tal y como sugieren los autores del artículo de referencia, una relación

entre dos niveles de representación distintos, como puede ser la relación existente entre los HMMs y los fonemas o entre los fonemas y las palabras. Además, gracias a los conocidos avances en Teoría de Autómatas, se ha podido aplicar sobre estos modelos compuestos una serie de operadores con el objetivo de acelerar y optimizar los procesos que se deben computar sobre ellos. Además, tal y como sugieren los autores de la publicación, se tratarían de unos métodos de carácter general, pudiendo aplicar estas técnicas a otras tareas, como pueden ser la síntesis de voz, el reconocimiento de texto manuscrito, el análisis de secuencias biológicas o, en nuestro caso, la lectura de labios.