UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MANIZALES

Desarrollo de un sistema de manipulación de un robot a través de movimientos de la boca y de comandos de voz

Tesis de Maestría 2009

Alexánder Ceballos Arias

aceballosa@unal.edu.co

Asesores:

Flavio Augusto Prieto Ortiz

Juan Bernardo Gómez Mendoza

DIEEC

Universidad Nacional de Colombia Sede Manizales

Grupo de Percepción y Control Inteligente

Desarrollo de un sistema de manipulación de un robot a través de movimientos de la boca y de comandos de voz

por:

Alexánder Ceballos Arias

TESIS DE MAESTRÍA

Presentada a:

Departamento de Ingeniería Eléctrica, Electrónica y Computación

Facultad de Ingeniería y Arquitectura

En cumplimiento de los Requerimientos

para el Grado de

MAGISTER EN INGENIERÍA - AUTOMATIZACIÓN INDUSTRIAL

Asesores:

Flavio Augusto Prieto Ortiz

Juan Bernardo Gómez Mendoza

Universidad Nacional de Colombia

Sede Manizales

A mi madre y a mis hermanos.

Abstract

In recent years audio-visual speech recognition has emerged as an active field of research thanks to advances in pattern recognition, signal processing and machine vision. Its ultimate goal is to allow human-computer communication using voice, taking into account the visual information contained in the audio-visual speech signal, whether to cope with the difficulties of a noisy environment, or when trying to recognize the emotion exhibited by the speaker.

This document presents a command's automatic recognition system using audio-visual information. Due to the fact that this work is part of the project "Automatic Segmentation and Classification of Lip Postures and Voice Commands in order to Control a Laparoscopic Robot", the system is expected to control a robot, in particular the laparoscopic robot da Vinci.

Hidden Markov Models have been used as technique for speech recognition using Hidden Markov Model Toolkit as computational tool. The audio signal is treated using the Mel Frequency Cepstral Coefficients parametrization method. Besides, features based on the points that define the mouth's outer contour according to the MPEG-4 standar are used in order to extract the visual speech information.

It becomes necessary to deal with lip tracking in video sequences. Lip tracking is still an open issue in terms of research due to the mouth's shape, texture and color complexity, the illumination changes and the background scenery. In this document an outer lip tracking algorithm based in shape and restrictions given in standard MPEG-4 is proposed. The video sequence does not have markers or any kind of makeover in order to highlight the lips. The algorithm is strong in presence of beard, skin tone and image's quality.

Resumen

En años recientes, el reconocimiento audio-visual del habla ha surgido como un campo activo de investigación, debido a los avances en reconocimiento de patrones, procesamiento de señales y visión por computador. Su objetivo final es permitir la comunicación hombre-máquina usando la voz, teniendo en cuenta la información visual contenida en la señal de habla audio-visual, para lidiar con las dificultades de un ambiente ruidoso, o para tratar de reconocer las emociones exhibidas por el locutor.

En este documento se presenta un sistema de reconocimiento automático de comandos usando información audio-visual. Debido a que este trabajo se enmarca en el proyecto "Segmentación y Clasificación Automática de Posturas Labiales y Comandos de Voz para el Control de un Robot Laparoscópico", el sistema pretende controlar un robot, en particular el robot laparoscópico da Vinci.

Se emplean los modelos ocultos de Markov como técnica de reconocimiento del habla, utilizando Hidden Markov Model Toolkit como herramienta computacional. La señal de audio se parametriza usando los coeficientes cepstrales en frecuencia de Mel, mientras que para extraer la información visual del habla, se usan características basadas en los puntos que definen el contorno externo de la boca según el estándar MPEG-4.

Se hace necesario hacer seguimiento preciso de la boca sobre secuencias de video. La complejidad de forma, textura y color de la boca, y los cambios de iluminación y fondos de los posibles escenarios, hacen que este sea aún un problema abierto. En este documento se propone un algoritmo para el seguimiento del contorno externo de la boca, sin utilizar marcadores o alguna clase de maquillaje para resaltar los labios, basado en apariencia y en restricciones morfológicas definidas en el estándar MPEG-4.

Y en un futuro no muy distante, las máquinas podrían incluso superar a los humanos en su capacidad de razonamiento.

— Justin Rattner, director tecnológico de Intel

Agradecimientos

Muchas gracias a todas las personas relacionadas con mi desarrollo como investigador, principalmente a los integrantes del grupo de trabajo académico PCI, y en especial a Flavio Prieto y a Juan Bernardo Gómez por brindarme la oportunidad de expandir mis horizontes.

A la Universidad Nacional de Colombia, que mediante el programa de becas para estudiantes sobresalientes de posgrado, hizo posible que realizara mis estudios de maestría. Al programa ECOS Franco-Colombiano (ECOS-Nord/COLCIENCIAS/ICFES/ICETEX), que como parte del proyecto Segmentación y Clasificación Automática de Posturas Labiales y Comandos de Voz para el Control de un Robot Laparoscópico, financió mi pasantía en el Laboratoire Ampère de l'Institut National des Sciences Apliques.

A mi familia y a mis amigos, por su apoyo incondicional y porque me han soportado incluso cuando parece imposible.

Gracias.

Índice

Lista de figuras **xii**

Lista de tablas **xiv**

Índice de algoritmos **xv**

1 Introducción **1**

2 Revisión de la literatura **3**

2.1 Características audio-visuales 4

2.1.1 Características acústicas de la voz 6

2.1.2 Características visuales usadas para reconocimiento del habla 8

2.2 Técnicas de reconocimiento automático del habla 13

2.2.1 Comparación de plantillas o patrones utilizando técnicas de programación dinámica 14

2.2.2 Redes neuronales (NN) 14

2.2.3 Modelos ocultos de Markov (HMM) 15

2.3 Métodos de integración 16

2.3.1 Modelos de lenguaje 17

3 Sistema de reconocimiento del habla usando sólo audio **19**

3.1 Reconocimiento del habla basado en fonemas usando audio 19

3.1.1 Experimento I: sistema ASR usando TIMIT 20

3.1.2 Experimento II: sistema ASR usando VidTIMIT 34

3.2 Sistema de reconocimiento de habla basado en palabras usando audio 37

3.2.1 Experimento III: sistema ASR usando datos adquiridos en el laboratorio 38

4 Sistema de reconocimiento audio-visual del habla **40**

4.1 Sistema de reconocimiento de habla basado en fonemas usando video 40

4.1.1 Seguimiento del contorno externo de los labios 41
4.1.2 Experimento IV: sistema de reconocimiento visual del habla usando VidTIMIT . 51
4.2 Sistema de reconocimiento de habla basado en palabras usando video 52
4.2.1 Experimento V: sistema de reconocimiento visual del habla usando datos adquiridos en el laboratorio 54
4.2.2 Experimento VI: sistema AVSR usando datos adquiridos en el laboratorio 55
4.2.3 Comparación 56

5 Comunicación con el robot **59**
5.1 Consideraciones 59
5.2 Implementación 63

6 Conclusiones y trabajo futuro **65**

A Bases de datos usadas en reconociemiento audio-visual del habla **68**
A.1 Base de datos propia 69

Apéndices **68**

B Freeduino **72**

Bibliografía **80**

Lista de figuras

2.1 Diagrama de bloques del sistema usado en [22] 5
2.2 Segmentación de la boca para interfaz hombre-máquina ([33, 34]) 6
2.3 Tracto vocal ([35]) 7
2.4 Características visuales usadas en sistemas AVSR ([26]) 9
2.5 Conjunto de características visuales usadas en [50] 10
2.6 Parámetros de Animación de la Cara ([30]) 12
2.7 Diagrama de bloques de un sistema de reconocimiento del habla 14
2.8 Estructura de una red neuronal 15
2.9 HMM usado para el reconocimiento del habla 16
2.10 Extensión del modelo HMM multi-cadena 17

3.1 Red de palabras representada con reglas de sintaxis 21
3.2 Programa usado para adquisición y etiquetado del audio 24
3.3 Palabra etiquetada a nivel de trifonemas 28
3.4 Frase alineada 29
3.5 Matrices de confusión para el caso de alineación de palabras 29
3.6 Matrices de confusión para el caso de busqueda de palabras claves en habla continua . . 30
3.7 Alineación de un archivo de audio adquirido en el laboratorio 31
3.8 Matriz de confusión de palabras claves en habla continua sobre datos adquiridos en el laboratorio 32
3.9 Matriz de confusión de palabras aisladas usando fonemas 33
3.10 Matriz de confusión de comandos aislados 33
3.11 Alineación de los datos de la base de datos VidTIMIT 35
3.12 Matriz de confusión reconociendo palabras claves sobre la base de datos VidTIMIT . . . 35
3.13 Archivo de audio adquirido en el laboratorio alineado usando el sistema entrenado con la base de datos VidTIMIT 36

3.14 Matriz de confusión de palabras claves sobre datos adquiridos en el laboratorio usando el sistema entrenado con la base de datos VidTIMIT 36

3.15 Matriz de confusión sobre datos adquiridos en el laboratorio usando el enfoque de palabras aisladas y empleando el sistema entrenado con la base de datos VidTIMIT 36

3.16 Consola de comando del sistema da Vinci 38

3.17 Número de estados de los modelos Vs la probabilidad logarítmica 39

3.18 Matriz de confusión sobre datos adquidos en el laboratorio usando el enfoque de palabras aisladas 39

4.1 Inicialización del algoritmo de seguimiento asistido 42

4.2 Ponderación de la similitud por una función de probabilidad normal 44

4.3 Seguimiento de los 10 puntos que conforman el contorno externo de la boca en una secuencia de video cada 10 cuadros 46

4.4 Seguimiento del contorno externo de los labios sobre tres secuencias de video de la base de datos VidTIMIT 47

4.5 Algunas características visuales probadas 49

4.6 Alineación de datos de VidTIMIT usando características visuales 51

4.7 Matriz de confusión de palabras claves sobre la base de datos VidTIMIT usando sólo información visual 52

4.8 Seguimiento del contorno externo de los labios sobre tres secuencias de video adquiridas en el laboratorio. Los resultados son mostrados cada 700 cuadros 53

4.9 Matriz de confusión sobre datos adquiridos en el laboratorio empleando únicamente características visuales y usando enfoque de palabras aisladas 55

4.10 Diagrama de bloques del sistema ASVR 56

4.11 Matriz de confusión sobre datos adquiridos en el laboratorio usando el enfoque de palabras aisladas para un sistema de reconocimiento de comando audio-visual 56

4.12 Respuesta del sistema ante el ruido acústico 57

5.1 Segmentación de la señal de audio en palabras aisladas 63

5.2 Matriz de confusión del sistema AVSR sobre 5 secuencias de 18 comandos 64

A.1 Datos adquiridos en el laboratorio 69

B.1 Plataforma abierta de desarrollo Freeduino 72

Lista de tablas

2.1 Visemas y fonemas relacionados. 13

4.1 Localización recomendada para los puntos característicos del contorno externo de la boca (el punto 7.1x corresponde al punto de rotación de la cabeza) 45

4.2 Unidades de los parámetros de animación de la cara. 48

4.3 Porcentaje de palabras correctamente reconocidas 57

5.1 Caraterísticas de duración de los comandos en inglés empleados 60

5.2 Caraterísticas de duración de los comandos en francés empleados 60

5.3 Caraterísticas de duración de los comandos en español empleados 60

5.4 Porcentaje de palabras reconocidas con el enfoque de palabras aisladas para seis comandos del habla inglesa, usando información audio-visual 62

5.5 Porcentaje de palabras reconocidas con el enfoque de palabras aisladas para siete comandos del habla francesa, usando información audio-visual 62

5.6 Porcentaje de palabras reconocidas con el enfoque de palabras aisladas para seis comandos del habla española, usando información audio-visual 62

Acrónimos

ASR	Automatic Speech Recognition
AVSR	Audio Visual Speech Recognition
MM	Markov Model
HMM	Hidden Markov Model
DTW	Dynamic Time Warping
LPC	Linear Predictive Coding
FT	Fourier Transform
DCT	Discrete Cosine Transform
FDPs	Facial Definition Parameters
FAPs	Facial Animation Paratemers
NN	Neural Network
PCA	Pricipal Component Analysis
LDA	Linear Discriminant Analysis
PCI	Percepción y Control Inteligente
MFCCs	Mel-frequency cepstral coefficients
HTK	Hidden Markov Model Toolkit
BEEP	British English Example Pronciation Dictionary
MLF	Master Label File

Índice de algoritmos

1 Seguimiento asistido de puntos del contorno externo 41
2 Restricciones 45
3 Caracteristicas 50
4 Segmentación audio 64

1 Introducción

Uno de los objetivos de la inteligencia artificial es conseguir que las máquinas interactúen con los humanos de manera natural. El reconocimiento automático del habla, al igual que la visión por computador, es una rama de la inteligencia artificial que pretende que un sistema informático imite algún comportamiento del hombre, en este caso, que la máquina pueda interpretar la información contenida en la señal de voz.

Esta tesis se enmarca en el proyecto "Segmentación y Clasificación Automática de Posturas Labiales y Comandos de Voz para el Control de un Robot Laparoscópico", cuyo objetivo es desarrollar una interfaz que de forma natural controle 3 grados de libertad de un robot.

El sistema da Vinci es un sistema de cirugía laparoscópica que consiste en una consola de control, una camilla, cuatro brazos robóticos y un sistema de visión de alto rendimiento. La consola de control puede ser ubicada a un lado de la mesa de cirugía o incluso en una habitación adyacente, lo que permite que el médico utilice el sistema sin portar mascarilla. Mientras que el cirujano observa imágenes tridimensionales a través de un sistema de visión estéreo, la cámara y los instrumentos son controlados mediante manipuladores de mano, haciendo el cambio entre ellos con pedales. Cuando el médico está manipulando la cámara, pierde control sobre los instrumentos, y algunas veces es necesario volver a posicionarlos. Para evitar dicha situación, se hace deseable controlar los movimientos de la cámara y de los intrumentos a la vez.

En una primera aproximación se usaron movimientos de la cara y gestos de la boca para controlar la cámara, pero se hizo evidente que no es posible mover la cabeza sin dejar de usar el sistema de visión esteréo. Una forma natural para comandar los movimientos de la cámara sin necesidad de mover la cabeza, es emplear un sistema de reconocimiento automático del habla con el fin de reconocer un conjunto pequeño de comandos para manejar los 3 grados del libertad de la cámara.

También se pretende robustecer al sistema ante el ruido acústico, al emplear características visuales del habla como información adicional, pues cuando una persona se encuentra en ambientes ruidosos,

usa el movimiento de los labios como parte de información. De hecho, se ha estimado que observar al hablante equivale a una ganancia de 15 dB en la relación señal a ruido [1, 2].

En este documento se exploran diferentes enfoques al implementar un sistema de reconocimiento del habla. Como características acústicas se usan los índices de Mel, y como información visual, se utilizan características basadas en el estándar MPEG-4.

Con el fin de extraer las características visuales del habla en las secuencias de video, se hace necesario hacer seguimiento de la boca. En este trabajo se propone un algoritmo de seguimiento del contorno externo de los labios y extracción de características basado en el estándar MPEG-4. El algoritmo es robusto a la resolución de la imagen, a la presencia de barba y al tono de la piel.

Se presentan diferentes enfoques al entrenar un sistema de reconocimiento de habla para controlar un robot, en especial el robot laparoscópico da Vinci. En todos los casos se emplean los modelos ocultos de Markov usando Hidden Markov Model Toolkit como herramienta computacional. Se muestra que la mejor forma de enfrentarse al desafío, es utilizando el enfoque de reconocimiento de palabras aisladas y usando palabras como unidades básicas. También se muestra que un sistema de reconocimiento audio-visual del habla tiene un desempeño superior cuando hay presencia de ruido acústico en la señal de entrada, aunque el sistema de reconocimiento que usa la información de audio tenga mejor desempeño en condiciones ideales.

En este trabajo se provee de sistemas de reconocimiento audio-visuales de comandos en inglés, en francés y en español capaces de controlar 3 grados de libertad de un robot. También se muestra cómo usar este sistema como parte de una interfaz hombre-máquina para transmitir la decisión tomada a través de puerto USB.

2 Revisión de la literatura

El campo del procesamiento de la señal de voz ha sido objeto de estudio intenso en las últimas tres décadas, debido principalmente a los avances en las técnicas de procesamiento digital de señales y reconocimiento de patrones, además de la capacidad de proceso de los sistemas de cómputo. Su objetivo final es desarrollar interfaces hombre-máquina, que permitan que el ser humano se comunique de manera natural con los distintos dispositivos como robots, sistemas telefónicos o el computador de escritorio [3, 4, 5, 6, 7].

La comunicación entre personas tiene en cuenta gran diversidad de conocimiento, lo que permite sortear dificultades como el ruido ambiental, el acento y la concatenación de palabras, además de asuntos gramaticales [8]. El reconocimiento automático del habla o ASR (Automatic Speech Recognition), es un campo difícil de tratar, debido principalmente a las variaciones de fonación (los locutores no hablan igual), las ambigüedades en la señal acústica (no toda la información presente está relacionada con el habla), la falta de cuidado del hablante, la variación en la frecuencia y duración de los fonemas, y la presencia de ruido o interferencias [9, 10, 1, 2]. Los sistemas actuales están restringidos a ambientes controlados, a ser utilizados con un grupo de hablantes reducido o requieren de posicionamiento especial del micrófono resultando en interfaces poco naturales.

El procesamiento de voz se divide en tres temas de interés: codificación, síntesis y reconocimiento del habla, los cuales aún son problemas abiertos de investigación. La codificación de voz fue ampliamente estudiada en la década de los ochenta y principios de los noventa, los esfuerzos se concentraron en el desarrollo de algoritmos de parametrización de la señal [11, 12], la extracción de la frecuencia fundamental [13], y el análisis y modelado de la envolvente espectral [14]. Mientras que en la síntesis el gran paso se produjo a mediados de los noventa, cuando se introdujeron los algoritmos basados en la concatenación de unidades pregrabadas [12, 15, 16, 17]. Con respecto a los sistemas de reconocimiento del habla, los dos bloques fundamentales consisten en un sistema de parametrización de la señal de voz y un sistema de reconocimiento de patrones [9, 3, 18]. Aunque el grado de desarrollo alcanzado en los sistemas de codificación es satisfactorio, no lo es para la síntesis ni para el reconocimiento del habla.

En cara al problema de reconocimiento automático del habla se han propuesto estrategias basadas en varias aproximaciones, pero la técnica de reconocimiento de patrones que mejores resultados ha

ofrecido para descifrar la señal de voz hasta el momento, es aquella basada en la teoría de decisión estadística. Esta técnica permite encontrar la secuencia de patrones que tiene la mayor probabilidad de estar asociada a la secuencia de observaciones acústicas de entrada.

Los modelos ocultos de Markov o Hidden Markov Models (HMMs por sus siglas en inglés) son modelos estadísticos cuya salida es una secuencia de símbolos o cantidades y poseen mejores tasas de identificación de habla distorsionada y normal que aquéllas basadas en plantillas o en otras aproximaciones [19]. Entre las razones de su popularidad sobresale el hecho de que la señal de voz puede ser vista como una señal estacionaria a trozos, es decir, se puede asumir que en un corto tiempo la señal puede ser modelada como un proceso estacionario [18, 20]. Además, los HMMs pueden ser entrenados automáticamente, son simples y computacionalmente viables [18].

Por otro lado, cuando una persona se encuentra en ambientes ruidosos o posee dificultades auditivas, trata de mejorar el reconocimiento del habla usando el movimiento de los labios como parte de información. De hecho, se ha estimado que observar al hablante equivale a una ganancia de 15 dB en la relación señal a ruido [1, 2]. Es por este motivo que el reconocimiento de habla audio-visual ha surgido como un campo activo de investigación [10]. Los esfuerzos se han concentrado en la representación visual del habla, debido a que se ha tratado de mejorar la confiabilidad al usar la información visual [21, 22, 23].

Al utilizar este esquema se han obtenido buenos resultados, por ejemplo, en [24] se utilizan características de color para extraer la información visual de la boca, mientras que en [25] extraen los contornos de la boca desde imágenes de intesidad de gris. En ambos se usan sistemas de integración basados en HMMs, capaces de preservar la naturaleza de sincronización entre el audio y el video, mejorando el reconocimiento cuando se agrega ruido sintético a la señal de habla. En [26] se muestran los enfoques usados para enfrentarse al problema del reconocimiento audio-visual del habla, así como algunos de los resultados más significativos.

2.1 Características audio-visuales

En los últimos años se han propuesto varios algoritmos para mejorar la robustez y la precisión del reconocimiento automático del habla en ambientes ruidosos. En [27], se usa un perceptrón multicapa para combinar características de audio y visuales y para compensar la pérdida de información causada por el ruido, además de un posproceso basado en información contextual. Se demostró que el uso de otras fuentes de información pueden mejorar la precisión de los sistemas, y se propuso como investigación futura usar otras fuentes, como el tema del que se habla, y otros métodos de integración de características más robustos. En [28] se incluyen variables auxiliares en los sistemas de reconocimiento

del habla como el pitch, y proponen usar la energía de la señal.

La extracción de características acústicas para el reconocimiento automático del habla ha sido ampliamente estudiada. Con generalidad se usan los coeficientes cepstrales en las frecuencias de Mel, debido a que están basados en la percepción auditiva humana. Se derivan de la Transformada de Fourier (FT) o de la Transformada Discreta del Coseno (DCT). En MFCC las bandas de frecuencia están situadas logarítmicamente (según la escala Mel), que modela la respuesta auditiva humana más apropiadamente que las bandas espaciadas linealmente de FT o DCT [29].

En la extracción de características acústicas, la señal de audio es muestreada en ventanas de tiempo, que son transformadas en características espectrales. Cada ventana de tiempo es representada como un vector de alrededor de 39 características de información espectral, de información de energía y de cambios espectrales.

Por otro lado, en los últimos años se han concentrado los esfuerzos en la representación visual del habla, debido a que se ha tratado de mejorar la confiabilidad al usar la información visual [22, 23, 21], tal como lo hacen los humanos en ambientes ruidosos o con personas con acento, justificándose en que ésta es invariante al ruido acústico. La selección de características acústicas ha sido ampliamente estudiada, y los esfuerzos actuales se concentran en la extracción de las características visuales y en la selección del modelo de integración audio-visual.

Para la extracción de características visuales, se usan algunas relaciones geométricas de los diámetros de la boca, así como los Parámetros de Animación de la Cara (FAPs) contenidos en el estándar MPEG4. Se usan los grupos 8 y 2 que describen los contornos interno y externo de los labios [30, 31].

En [22], se describe un sistema audio visual de reconocimiento automático del habla. Como características visuales usaron los puntos que describen los contornos interno y externo de la boca en los FAPs (Figura 2.1). Se usó PCA para disminuir la dimensión del vector de características. Como características de audio, se usaron los coeficientes Cepstrales en Frecuencia de Mel.

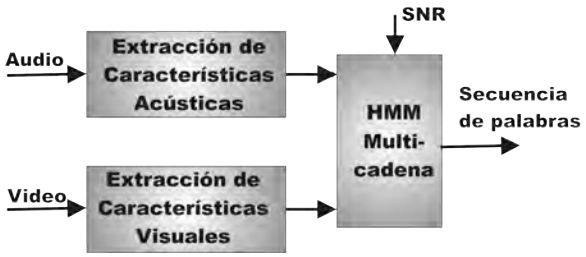


Figura 2.1: Diagrama de bloques del sistema usado en [22]

En el grupo de investigación Percepción y Control Inteligente (PCI) se han desarrollado algunos trabajos para el reconocimiento de fonemas usando información visual. En [32], se presentaron los diferentes algoritmos y procedimientos para la extracción automática de características faciales, basadas en

la región del contorno de la boca, para el proceso de reconocimiento de 5 fonemas vocales del lenguaje español. Además, en el 2006 se propuso un método para la segmentación y extracción de características faciales en secuencias de video en tiempo real, para ser usado en una interfaz hombre-máquina [33, 34]. El método usa diferentes algoritmos de segmentación, basados en píxeles y restricciones morfológicas para extraer el área de la boca. Se generó un pequeño pero suficiente conjunto de características gestuales, que permitió controlar tres grados de libertad de un robot manipulador (Figura 2.2).

Figura 2.2: Segmentación de la boca para interfaz hombre-máquina ([33, 34])

2.1.1 Características acústicas de la voz

El habla se genera cuando la forma de onda de la fuente glotal de una frecuencia fundamental pasa a través del tracto vocal, el cual, debido a su forma, tiene unas características de filtrado particulares. Pero muchas características de la fuente glotal, como su frecuencia, los detalles de pulso glotal, etc., no son necesarios para distinguir los diferentes fonemas. La información más útil para la detección del fonema es la del filtro, es decir, la posición exacta del tracto vocal (Figura 2.3). Si se sabe la forma del tracto vocal, se puede saber cual fonema fue producido. Esto sugiere que las características útiles para la detección de fonemas se encuentran al deconvolucionar (separar) la fuente y el filtro y mostrar sólo el filtro del tracto vocal.

El objetivo de la codificación de voz es parametrizar la señal de tal forma que se pueda extraer la información relevante. La mayoría de estos sistemas se basan en el análisis de la potencia espectral en tiempo corto, para lograrlo, la señal se divide en tramas cortas, estas tramas pueden considerarse entonces cuasiestacionarias y ser sometidas a análisis espectral y obtener así un vector de características [36].

El primer paso para la parametrización es elevar la cantidad de energía en las frecuencias altas. Si se observa, por ejemplo, el espectro de voz segmentada de las vocales, hay más energía en las frecuencias bajas que en las altas. Este desnivel de energía sobre las frecuencias es causado por la naturaleza del pulso glotal (posición de la lengua). Incrementar la energía de las frecuencias altas hace que la

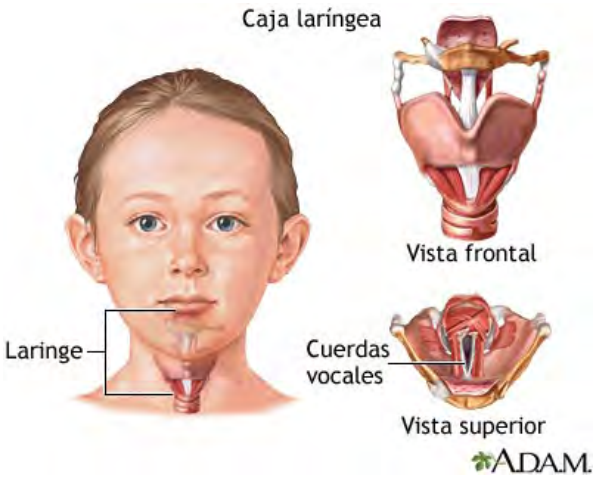


Figura 2.3: Tracto vocal ([35])

información de estas esté disponible para el modelo acústico y mejora la precisión de la detección del sonido. La parametrización puede hacerse tanto en el dominio del tiempo como en el de la frecuencia.

En el dominio del tiempo se analizan la energía local, la tasa de cruces por cero y la autocorrelación, el análisis de la señal es rápido, sencillo y de fácil interpretación. Mientras que el dominio de la frecuencia es utilizado por su mayor potencia para caracterizar la información de la señal de voz, al analizar la potencia espectral en escala logarítmica de las tramas de voz.

Codificación predictiva lineal (LPC)

Se concentra en modelar las resonancias de la garganta al generar la voz. Para esto se modela el tracto vocal como un tubo acústico sin pérdidas ni bifurcaciones. Los efectos del tracto vocal en la señal de excitación crean una serie de resonancias, modelando el tracto como un filtro todo polos $H(z)$. Mediante el modelo AR se pueden obtener los coeficientes del filtro sin calcular el espectro. LPC se utiliza para el sintetizado de voz, como método de compresión de voz, para digitalizar y encriptar la voz para su envío por un canal de capacidad limitada [9, 16, 17, 37, 38], y en música para combinar el sonido de instrumentos con la voz [39].

Índices de Mel

El cepstrum es otra manera útil para separar la fuente del filtro. Debido a que el espectro de potencias de la señal se obtiene aplicando la transformada de Fourier a las tramas de voz de ventanas que se solapan, aparecen armónicos de la frecuencia fundamental de las tramas. Este efecto se puede subsanar agrupando los conjuntos de componentes cercanos en unas 20 bandas de frecuencias antes de calcular el logaritmo de la potencia. Cada filtro hace un promedio pesado de las componentes espectrales

presentes en su banda, caracterizando el tracto vocal con la envolvente espectral suavizada. Es común usar la escala de resolución perceptual del oído humano, haciendo que las bandas que abarcan los filtros sean más anchas para frecuencias superiores a 1 kHz. Esta escala recibe el nombre de escala Mel. El logaritmo de la energía a la salida de los filtros en escala Mel da lugar a los coeficientes MFCC. Los coeficientes MFCC han demostrado ser los que mejores resultados dan como técnica de parametrización, teniendo en cuenta el compromiso entre costo computacional y resultados obtenidos [3, 36].

Modelos auditivos

Estas técnicas se basan en la percepción de la voz humana para parametrizar el habla, intentando reproducir el comportamiento de la membrana basilar del oído.

Perceptual Linear Prediction (PLP) Este modelo usa la resolución espectral en la banda crítica, las curvas de igual potencia y la ley de intensidad-potencia para calcular el espectro de la voz [3, 36, 40].

Generalized Synchrony Detector (GSD) Este modelo promedia la respuesta de un banco de filtros que refleja la respuesta de la membrana basilar a los estímulos acústicos [36, 41, 42].

Ensemble Interval Histogram (EIH) Este modelo se basa en el cómputo de los histogramas de las frecuencias de activación de los filtros con los que modela la membrana basilar, generando una representación de la voz con alta resolución espectral [36, 43, 44].

Los Modelos Acústicos no siguieron desarrollándose, aunque los resultados son buenos, debido a su costo computacional y de almacenamiento.

2.1.2 Características visuales usadas para reconocimiento del habla

Cuando una persona se encuentra en ambientes ruidosos o posee dificultades auditivas, trata de mejorar el reconocimiento del habla, al usar el movimiento de los labios como parte de información. La lectura de los labios es aun un problema abierto de visión artificial. Los esfuerzos se han concentrado en la representación visual del habla, debido a que se ha tratado de mejorar la confiabilidad al usar la información visual [22, 23, 21], tal como lo hacen los humanos en ambientes ruidosos o con personas con acento, justificándose en que ésta es invariante al ruido acústico.

El primer desafío para el desarrollo de un sistema audio-visual de reconocimiento del habla, Audio Visual Speech Recognition o AVSR por sus siglas en inglés, es la selección de las características visuales

y su extracción desde el video. Debido a que la información visual del habla se encuentra principalmente en la región de la boca, la mayoría de algoritmos localizan primero la cara y después la región de la boca. Una vez encontrada la región de interés, se parametriza ya sea la forma de la boca o su movimiento.

Las características visuales usadas para el reconocimiento del habla pueden dividirse en tres grupos [22]. i) Las características de nivel alto, o características de forma, en las que los parámetros de un modelo que define los contornos de la boca son usados como características. Estos métodos dependen de la precisión de detección de los contornos, la cual se dificulta bajo condiciones adversas como la rotación de la cámara o el ruido, pero poseen la ventaja de que pocos parámetros deben ser obtenidos por cada cuadro de video y tienen una dimensión baja, lo que se refleja en la viabilidad del sistema [10, 22]. ii) Las características de nivel bajo, o características de apariencia, que son obtenidas como resultado de transformaciones de los píxeles de la imagen en la región de la boca. Varias aproximaciones han sido estudiadas, como análisis de componentes principales (PCA) [27, 45], análisis de discriminantes lineales (LDA) [9, 23, 28, 46], transformada del coseno [21] y transformaciones lineales de máxima verosimilitud (MLLT) [47]. Las características visuales de bajo nivel no requieren algoritmos complejos y contienen información que no pueden capturar las de alto nivel. iii) Las características combinadas, que mezclan la forma y la apariencia de la boca simplemente concatenándolas o usando modelos estadísticos ([26]),(Figura 2.4). En general, el vector de características visuales captura información dinámica, al incluir la primera y segunda derivada temporal. Además, debido a que la frecuencia de muestreo del audio es mayor que la del video, las características visuales deben ser interpoladas [26].

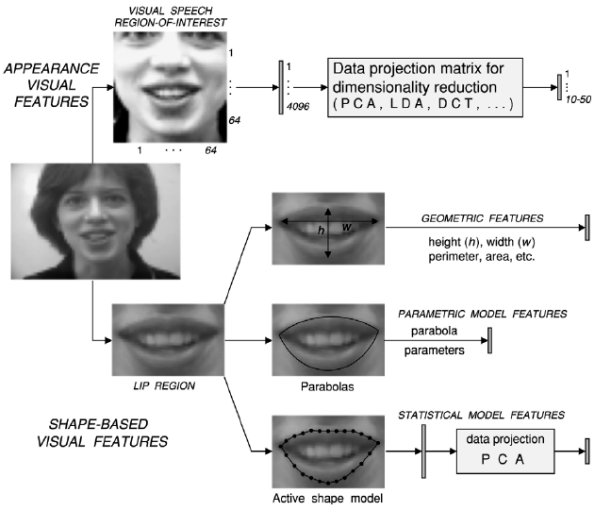
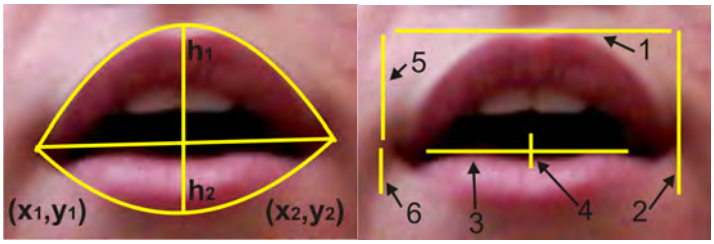


Figura 2.4: Características visuales usadas en sistemas AVSR ([26])

Características visuales de alto nivel

Una vez localizada la región de interés, existen varios métodos para extraer los contornos de la boca. Entre los más populares se destacan los snakes o contornos activos, que son curvas elásticas representadas por puntos de control, y el uso de curvas paramétricas como parábolas. En [48], se emplea un modelo de forma activa, el cual se basa en 5 curvas parabólicas, 3 describen el contorno externo de la boca y 2 el interno, y el vector de características consiste en 12 parámetros que representan la forma de la boca. Se mostró que el rendimiento de un sistema AVSR usando este modelo para hallar las características visuales, es superior que el de aquellos que usan modelos de forma activa basados en imágenes en escala de grises y a color que no tienen restricciones de forma. En [49], se realizó una comparación rigurosa entre posibles características geométricas usadas para el reconocimiento del habla, y se concluyó que la apertura vertical de los labios es la más relevante. También se mostró que el vector conformado por las aperturas vertical y horizontal de los labios, así como las primeras derivadas de los ángulos de las esquinas, producen el mejor resultado.

Se han empleado otras características de alto nivel, por ejemplo en [50], se usaron dos conjuntos de características visuales, los cuales se pueden apreciar en la Figura 2.5. Se propuso una técnica para separación robusta de señales mezcladas de habla, al identificar cuando cada hablante se encuentra en silencio, usando la envolvente de la señal obtenida al unir las características de audio y el movimiento de los labios.



(a) (b)

Figura 2.5: Conjunto de características visuales usadas en [50]

Un sistema AVSR que permite traducir inglés-japonés y japonés-inglés se presenta en [51]. Se trata de un traductor capaz de generar los movimientos de habla del locutor y sincronizarlos con la voz traducida al reemplazar la región de la boca, conservando la expresión facial. Aún es necesario modelar el movimiento de la lengua y mejorar el método de interpolación de la forma de los labios para obtener una apariencia más natural.

Algunos algoritmos y procedimientos para la extracción automática de características faciales, basadas en la región del contorno de la boca, para el proceso de reconocimiento de 5 fonemas vocales del lenguaje español fueron presentados en [32]. Un método para la segmentación y extracción de

características faciales en secuencias de video en tiempo real, para ser usado en una interfaz hombre-máquina fue propuesto en [33, 34]. El método usa diferentes algoritmos de segmentación, basados en píxeles y restricciones morfológicas para extraer el área de la boca. Se generó un pequeño conjunto de características gestuales, que permitió controlar tres grados de libertad de un robot manipulador.

Características visuales de bajo nivel

Para representar el movimiento del área de la región de la boca, se han propuesto varias aproximaciones basadas en operaciones matemáticas lineales y no lineales sobre todos los píxeles de la imagen [27, 25, 24]. El flujo óptico es definido como la distribución de las velocidades aparentes en el movimiento de patrones brillantes en una imagen. Las características visuales pueden ser calculadas de forma robusta sin extraer los contornos ni la localización de la boca, por lo tanto, es más razonable usar el movimiento de los labios que la forma para el reconocimiento del habla. En [52], se usaron los valores máximos y mínimos de la integral del flujo óptico como características visuales, estas características son especialmente útiles para detectar los períodos de pausa o silencio, logrando una reducción del 30% en el error relativo, con una relación señal a ruido de 10dB, al realizar el reconocimiento usando el audio sólo cuando se presume que la persona está hablando.

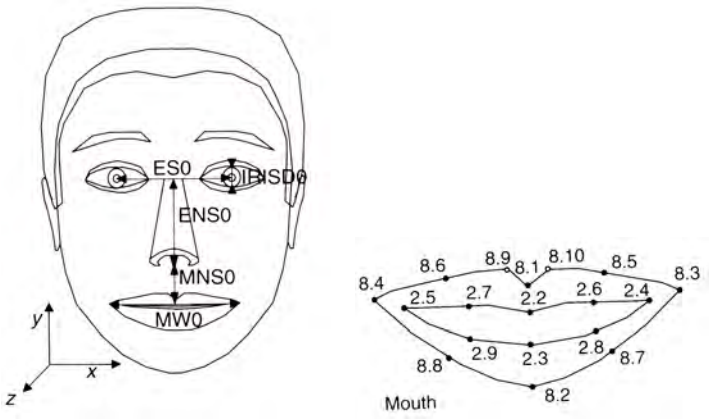
En [53], la región de interés es encontrada para luego hallar los puntos característicos (esquinas y bordes de la boca), con el fin de calcular vectores de movimiento mediante un algoritmo de comparación de bloques. En este algoritmo, la región de interés es dividida en un conjunto de rectángulos pequeños y sólo se consideran las traslaciones sobre este conjunto. Usando como entrada los vectores de movimiento para un sistema de reconocimiento basado en redes neuronales, se mostró que el sistema posee mejor rendimiento que aquel que sólo emplea la información de audio.

Se han explorado otras fuentes de información visual que sean insensibles a las condiciones de iluminación. En [46, 47], se empleó un casco con sensores infrarrojos para extraer características visuales de la región de la boca. Con el fin de calcular 30 características de movimiento, se utilizó LDA sobre los 100 primeros coeficientes de la trasformada discreta del coseno, obtenidos en cuadros seguidos de video. Estos coeficientes contienen información que no puede ser conservada con características de alto nivel, como la presencia de los dientes y de la lengua, pero se hace necesario realizar corrección tanto de escala como de orientación. Al realizar la comparación con técnicas que tienen en cuenta toda la cara, se observó que el uso de un casco con infrarrojos es una forma viable de desarrollar sistemas de reconocimiento AVSR.

Comparación entre características visuales de alto nivel y características visuales de bajo nivel

Para el reconocimiento visual del habla se utiliza tanto la forma de la boca, como su apariencia, pero ambas técnicas proveen casi la misma información. Desafortunadamente, existen pocos trabajos que desarrollen una comparación entre el desempeño de sistemas AVSR basados en los enfoques de alto y bajo nivel. En [54], se compararon dos grupos de características visuales, el primero de alto nivel, basado en los Parámetros de Animación de la Cara (FAPs), soportados en el estándar MPEG-4 y, el segundo de bajo nivel, basado en PCA realizado sobre la imagen de intensidad de la región de la boca. El sistema fue evaluado para varios niveles de ruido, y para diferentes valores de dimensión de los vectores de características, ofreciendo notables mejorías para ambos enfoques y para todos los niveles de ruido probados.

Estándares MPEG-4



(a) FAPUs medidos en la cara en estado neutro

(b) Gupos 2 y 8 de los FAPs

Figura 2.6: Parámetros de Animación de la Cara ([30])

MPEG-4 ha surgido debido a la necesidad de estandarizar los objetos virtuales de video real y sintético. En él se incluyen la codificación de video, la compresión de la geometría y la sincronización entre audio y video. En dicho estándar se define un conjunto complejo de parámetros FDP (Facial Definition Parameters), utilizados para la estandarización de la cara. También se especifican los parámetros de animación de la cara o FAP (Facial Animation Paratemers), que corresponden a una acción particular de deformación de un modelo de cara que está en estado neutro, lo que permite la animación de modelos de cara sintéticos. Los FAPs se miden en unidades específicas FAPU (Face Animation Parameter Units) [30]. En la Figura 2.6 se aprecian las medidas antropométricas normalizadas empleadas en el

Tabla 2.1: Visemas y fonemas relacionados.

Visema	Fonemas	Ejemplo
0	ninguno	
1	p, b, m	put, bed, mill
2	f, v	far, voice
3	T, D	think, that
4	t, d	Tip, doll
5	k, g	call, gas
6	Ts, dZ, S	chair, join, she
7	s, z	Sir, zeal
8	n, l	Lot, not
9	r	red
10	A:	car
11	e	bed
12	I	Tip
13	Q	top
14	U	book

estándar, los cinco FAPU miden la distancia entre los ojos (ES0), el diámetro del iris (IRISD0), la separación entre los ojos y la nariz (ENS0), la separación entre la boca y la nariz (MNS0) y el ancho de la boca (MW0).

El estándar MEPG-4 define 68 FAPs divididos en 10 grupos. En reconocimiento del habla, generalmente se usan el grupo 2 y 8, que describen el movimiento del contorno interno y externo de la boca respectivamente. Mientras que para la síntesis visual del habla, se usa el grupo 1 que define 14 visemas claramente distingibles (Tabla 2.1). Un visema es el patrón visual del referencia de un fonema, y un visema puede corresponder a varios fonemas.

2.2 Técnicas de reconocimiento automático del habla

En un sistema de reconocimiento automático del habla se intenta descifrar la información contenida en la señal de voz. El reconocimiento se realiza al comparar la señal con un conjunto de patrones, devolviendo la secuencia que con mayor probabilidad la representa ([55]) (Figura 2.7).

Se han utilizado diferentes técnicas al enfrentarse al problema del reconocimiento automático del habla [9], entre ellas citamos: las técnicas probabilísticas basados en la Teoría de la Decisión de Bayes y la Teoría de la Información [18], [21, 20], y las Técnicas de Comparación de Patrones y de Programación Dinámica [55, 56, 57].

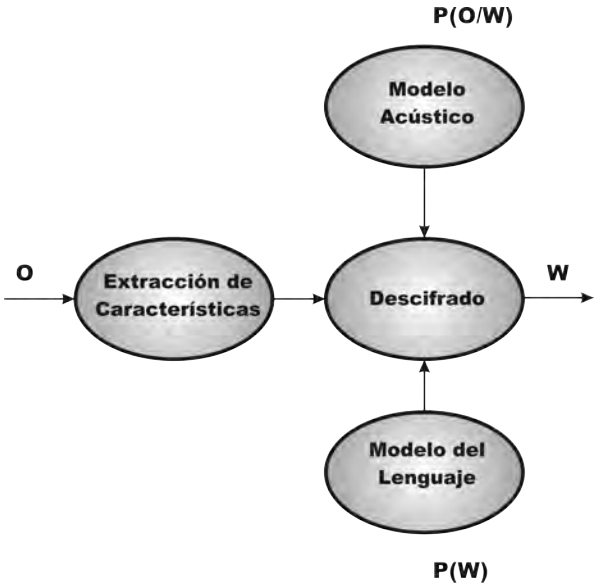


Figura 2.7: Diagrama de bloques de un sistema de reconocimiento del habla

2.2.1 Comparación de plantillas o patrones utilizando técnicas de programación dinámica

Las plantillas son simplemente secuencias de características ordenadas en el tiempo que representan las unidades a reconocer. Para realizar la comparación se debe implementar un alineamiento no lineal y contar con una medida de distancia apropiada. La técnica más utilizada para dicha comparación es conocida como DTW (Dynamic Time Warping) [56], la cual es usada para tratar el problema de reconocimiento de habla continua y aislada con una cierta independencia del locutor. DTW es un algoritmo de medida de similitud entre dos secuencias, que pueden variar en tiempo y velocidad. Actualmente se usa en sistemas de reconocimiento automático del habla basados en Modelos Ocultos de Markov [57].

2.2.2 Redes neuronales (NN)

Las redes neuronales son estructuras de procesamiento paralelo de información, formadas por numerosos nodos simples conectados entre sí mediante pesos y agrupados en diferentes capas, entre ellas la capa de entrada y la capa de salida. (Figura 2.8). La respuesta de la Red Neuronal puede ser representada como una proyección no lineal del espacio de entrada, y esta respuesta puede ser dinámica o estática. Se ha demostrado que una Red Neuronal puede aproximar cualquier función de entrada-salida, pues tiene la capacidad de aprender a partir de pares observación-objetivo.

Debido a las capacidades que poseen, se han convertido en una de las herramientas más atractivas

para la solución del problema del reconocimiento de habla. Se han conseguido resultados comparables a los obtenidos con otros métodos ya clásicos, consiguiendo modelar la variable tiempo, permitiendo no sólo clasificaciones muy acertadas, sino además la segmentación de la señal de entrada [58, 59]. En [60], se presenta un sistema de control de navegación de un robot móvil en tiempo real, basado en HMMs y una red neuronal difusa para representar el significado de las palabras dependiendo del contexto y del estado actual de la máquina.

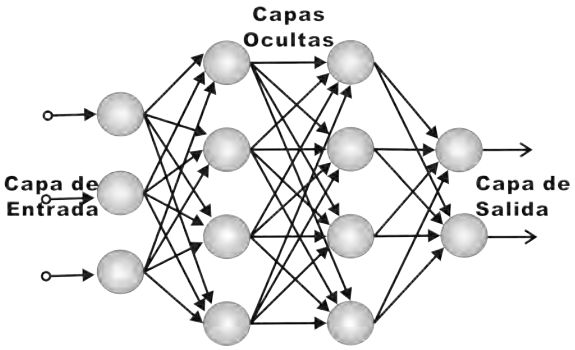


Figura 2.8: Estructura de una red neuronal

2.2.3 Modelos ocultos de Markov (HMM)

El modelado estocástico de la señal de habla ha proporcionado los mejores resultados hasta la fecha, tanto para el reconocimiento de habla aislada como continua y para independencia del locutor [19].

Entre las razones de su popularidad sobresale el hecho de que la señal de voz puede ser vista como una señal estacionaria a trozos, es decir, en un corto tiempo la señal es modelada como un proceso estacionario [18, 20]. Los HMMs albergan una descripción probabilística del fenómeno que modelan. Poseen un conjunto de estados finitos, un estado inicial y un estado final. La forma en que se realizan los cambios de un estado a otro es determinada por una función de transición de estados, denominada función de probabilidad de transición entre los estados a_{ij} (probabilidad de pasar al estado j dado que se está actualmente en el estado i). También existe una función de probabilidad B la cual es una distribución sobre todas las posibles salidas. Los símbolos de observación son denotados como V y el parámetro $b_{j(k)}$ describe la probabilidad de que el estado j observe al símbolo k del conjunto de salidas.

El proceso de modelado de unidades acústicas se divide en 3 partes. i) Evaluación. En general es suficiente con encontrar la mejor secuencia y su probabilidad, es decir, dada una observación Y y un modelo λ, determinar la probabilidad $p(Y|\lambda)$ de que el modelo genere esa observación. Con este fin, existen algoritmos que permiten ahorrar muchos cálculos, como el algoritmo de Viterbi [18, 20, 61]. La

máxima probabilidad acumulada se obtiene multiplicando la probabilidad de observación del estado por la máxima probabilidad acumulada entre todos los caminos que llegan a él. ii) Decodificación. A partir de las secuencias más probables encontradas con el algoritmo de Viterbi $X = x1, x2, ..., xn$, se determina la cantidad de veces que el k-ésimo símbolo ha sido asignado al j-ésimo estado del modelo, y hallar una buena estimación de la probabilidad de que el j-ésimo estado del modelo emita el k-ésimo símbolo observable, es decir $b_{j(k)}$. iii) Entrenamiento. Consiste en el cálculo de los parámetros que caracterizan el modelo. Pueden ser hallados dado un conjunto de secuencias observadas, aplicando de forma repetitiva la búsqueda de la mejor secuencia y posterior reestimación de las probabilidades [62, 63, 64]. El algoritmo de reestimación de Baum-Welch hace posible obtener una función de pertenencia con salida no binaria y utilizarla para pesar las evidencias de las secuencias de entrenamiento en la reestimación de las probabilidades del modelo [18]. Los HMM más utilizados en ARS poseen una estructura muy simple denominada de izquierda a derecha, en la cual la transición entre un estado a otro sólo es posible si $j > i$ (Figura 2.9).

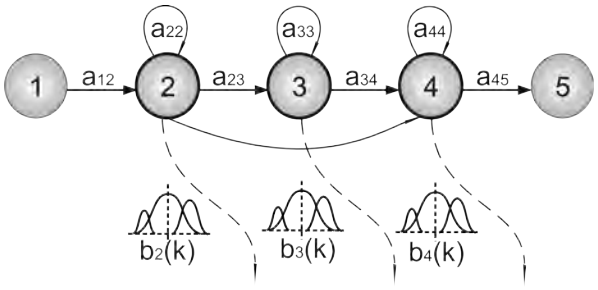


Figura 2.9: HMM usado para el reconocimiento del habla

2.3 Métodos de integración

La selección del conjunto de características visuales y del método de integración audio-visual son problemas activos de investigación. Las técnicas de integración audio-visual se pueden clasificar en técnicas de fusión de características y métodos de fusión de decisión [21]. En la fusión de características o integración temprana, los vectores de observación son obtenidos al concatenar las características de audio y visuales, y en ocasiones, se pueden emplear técnicas de reducción de dimensión [65]. En la fusión de decisión o integración tardía, la verosimilitud condicional de clase en los modelos ocultos de Markov, es combinada en diferentes etapas, y el resultado es usado para el reconocimiento.

En [66], se presenta un HMM multi-cadena que asume la sincronización del audio y el video, y permite que el cálculo de la verosimilitud en cada estado se haga independientemente y sea ponderada la contribución relativa de cada cadena dependiendo del nivel de ruido acústico. Presenta el inconveniente

de no poder describir los estados asincrónicos del audio y el video. Una extensión de este modelo es el HMM de producto multi-cadena [66, 67, 68, 69]. Aquí se representa cada estado del HMM como un par de estados (uno para el audio y otro para el video) (Figura 2.10), permitiendo asincronía entre las cadenas a nivel de estado. Finalmente, se puede tratar cada secuencia con un HMM diferente, lo que hace que la asincronía se maneje al nivel de modelo y no de estado, pero no mantiene la dependencia obvia entre las características de audio y visuales [70].

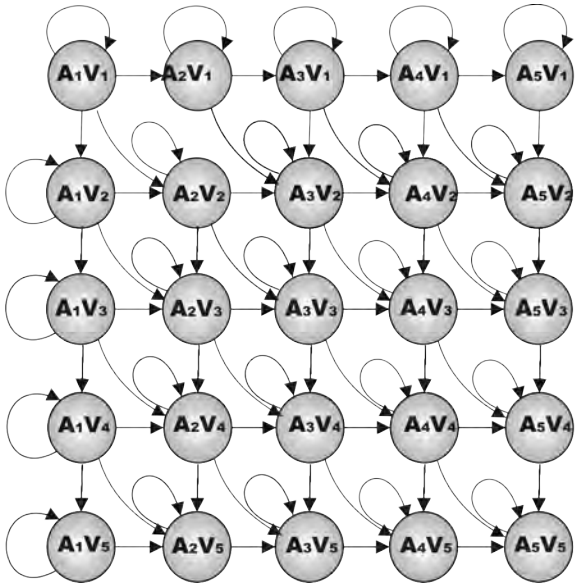


Figura 2.10: Extensión del modelo HMM multi-cadena

En [21], se proponen dos modelos estadísticos para la integración audio-visual en sistemas de reconocimiento de palabra aislada, que permiten asincronía en los estados de las secuencias de observación de audio y visuales, preservando su correlación en el tiempo. En [22], para la integración audio-visual, primero se interpoló a fin de sincronizar los datos, y la verosimilitud logarítmica de las características audio-visuales fue combinada usando pesos que representan la fiabilidad. En [27] utilizan un perceptrón multicapa para combinar características de audio y visuales y para compensar la pérdida de información causada por el ruido.

2.3.1 Modelos de lenguaje

Según el objetivo de reconocimiento, los sistemas pueden ser clasificados en los siguientes tipos: reconocimiento de palabra aislada, en el cual las palabras están separadas por pausas; reconocimiento de palabras clave, en el cual se reconocen ciertas palabras dentro de habla continua; y finalmente, reconocimiento de habla conectada o continua, en donde se trata de descifrar la secuencia de palabras

contenida en la señal de entrada, sabiendo que dicha señal no contiene pausas separadoras entre las palabras [55].

Por otro lado, los sistemas de reconocimiento del habla basados en HMMs pueden usar como unidades básicas los fonemas o las palabras. Aunque no existe manera directa de definir el número de estados para cada cadena, se ha asumido que para el manejo de fonemas basta con emplear 3 estados activos [71]. Cuando las cadenas representan palabras, la arquitectura del modelo debe ser supuesta de antemano. Debido a que el rendimiento del sistema depende fuertemente del número de estados y de la función de probabilidad de cada estado, varias configuraciones deben ser probadas para cada palabra.

3 Sistema de reconocimiento del habla usando sólo audio

Los modelos ocultos de Markov han demostrado su aplicabilidad en el problema de reconocimiento del habla durante las últimas tres décadas. De hecho, han mostrado su potencial para el reconocimiento de patrones dinámicos, como el clima o el mercado.

Independientemente de la técnica de reconocimiento de patrones empleada, se debe escoger la unidad con la cual las características de entrada serán comparadas, es decir, el reconocimiento de patrones de voz puede hacerse a nivel de fonemas, de palabras e incluso de frases, dejando el problema de interpretación y sintaxis en un estado superior.

En este capítulo se presentan las aproximaciones con las cuales se abordó el problema de reconocimiento de comandos para controlar un robot. Se usaron fonemas y palabras como unidades básicas y como características los índices de Mel. Además, se exploró el enfoque de reconocimiento de palabras claves y el de reconocimiento de palabras aisladas, en el cual las palabras están separadas entre sí por pausas y no por silencios. Para ambos se emplearon las herramientas provistas en el HTK (Hidden Markov Model Toolkit), que es un paquete desarrollado en C por el Departamento de Ingeniería de la Universidad Cambridge, con el apoyo de la Corporación Microsoft, y distribuido en Internet bajo licencia Open Source [61]. El pricipal objetivo de HTK es la manipulación y desarrollo de Modelos Ocultos de Markov para la investigación en reconocimiento del habla, ofreciendo herramientas capaces de manipular diferentes formatos de archivos de audio, e incluso algoritmos para la extracción de características acústicas como los índices de Mel. Sin embargo, y debido a que es modular, puede ser empleado en la investigación de otras aplicaciones que usen HMMs [72].

3.1 Reconocimiento del habla basado en fonemas usando audio

La aproximación más natural en sistemas ASR es usar fonemas como las unidades básicas que conforman el habla. Es de esta forma como los humanos reconocemos la información contenida en la señal

de voz, al concatenar sonidos que forman las palabras, siendo capaces, además, de ignorar aquellos sonidos que no conllevan a una respuesta lógica y de separar palabras que unimos cuando hablamos de forma continua, pues tenemos en cuenta información apriori obtenida con años de experiencia.

Otra de las ventajas de usar fonemas en sistemas ASR es que las palabras a reconocer no tienen que estar dentro del diccionario de entrenamiento, pues cualquier palabra puede ser construida como la combinación de los modelos entrenados. Por otro lado, se hace necesario el uso de extensas bases de datos para que los modelos de los fonemas sean robustos a la pronunciación y al acento.

Desde hace ya varios años se han recolectado bases de datos de audio con el fin de construir sistemas ASR. En general, estas bases de datos poseen suficiente información para tratar de hacer el sistema independiente del género, la pronunciación o el acento del locutor, pero debido a que los esfuerzos de investigación se concentran en el inglés, la mayoría de bases de datos se encuentran en este idioma (Apéndice A), aunque también existen algunas en japonés.

De hecho, debido a que el proyecto *Segmentación y clasificación automática de posturas labiales y comandos de voz para el control de un robot laparoscópico*, se realiza en un marco de cooperación con la república francesa, se hizo contacto con *European Languaje Resources Association - ELRA*, entidad encargada del estudio y preservación de los lenguajes europeos, y que distribuye a través del *ELDA (European Languaje Distrubution Agency)* las bases de datos usadas en ingeniería del lenguaje, para adquirir una base de datos audio-visual en francés. Sin embargo, no fue posible, pues ellos no disponen de alguna base audio-visual para desarrollo de sistemas de reconocimiento del habla en francés.

Cuando el interés ha sido el estudio de la viabilidad de los distintos enfoques para el reconocimiento del habla, las bases de datos constan de pocas palabras, como por ejemplo los números. Pero cuando el objetivo es el desarrollo de un sistema de reconocimiento de habla de vocabulario extenso, se han usado bases de datos más grandes que incluyen la pronunciación de algunas frases celebres.

Además, estas bases de datos creadas en el marco de proyectos de investigación, han sido puestas a disposición del público con algún costo, pues son uno de los resultados palpables de dichos proyectos. Este producto incluye, aparte de los archivos de audio, otra información obtenida durante el preprocesamiento, como los archivos de etiquetado, los cuales son una representación (en distintos formatos) del habla en el tiempo, a nivel de fonemas y palabras.

3.1.1 Experimento I: sistema ASR usando TIMIT

Para crear el sistema de reconocimiento del habla basado en voz, se necesita una base de datos, preferiblemente etiquetada a nivel de fonemas. La base de datos TIMIT es probablemente la base de datos más empleada a nivel mundial en el desarrollo de sistemas ASR de habla continua. Es el resultado de los esfuerzos conjuntos de Massachusetts Institute of Technology (MIT), Stanford Research Institute

(SRI) y Texas Instruments (TI) bajo el patrocinio de Defense Advanced Research Projects Agency - Information Science and Technology Office (DARPA-ISTO). TIMIT contiene 10 frases diferentes pronunciadas por 630 hablantes de ambos sexos y de 8 regiones dialécticas de Estados Unidos, para un total de 6300 muestras [73]. A continuación se mencionará paso a paso el procedimiento para desarrollar un sistema ASR usando HTK.

1. Grámatica Se hace necesario definir algunas reglas del lenguaje, por ejemplo, qué palabras pueden ser reconocidas y en qué orden deberían aparecer. Si estas reglas de sintaxis son simples, HTK ofrece una herramienta sencilla para su elaboración. Consiste en especificar un grupo de variables y expresiones regulares que describen las palabras a reconocer. Por ejemplo, se podría crear un archivo de texto con las siguientes líneas:

```
$comcam = START | STOP;
$direction = UP | DOWN | LEFT | RIGHT;
$comzoom = IN | OUT;
(SILENCIO (( $comcam CAMERA ) | ( MOVE $direction) | (ZOOM
$comzoom)) SILENCIO )
```

   Donde $ significa que la palabra que sigue es una variable, | expresa alternativas, los paréntesis () se usan para subexpresiones, las llaves {} para indicar que puede haber ninguna o una repetición, <> una o más repeticiones y los corchetes [] representan términos opcionales.

   De hecho, este archivo sitáctico es entendible para los humanos, pero la máquina necesita verlo como una red de palabras (Figura 3.1), para hacer esta conversión HTK ofrece la herramienta HParse que traduce las reglas gramaticales en un archivo de entramado a nivel de palabras de la forma Backus-Naur extendida, la cual es una representación de una red de estados finitos donde los nodos representan palabras.

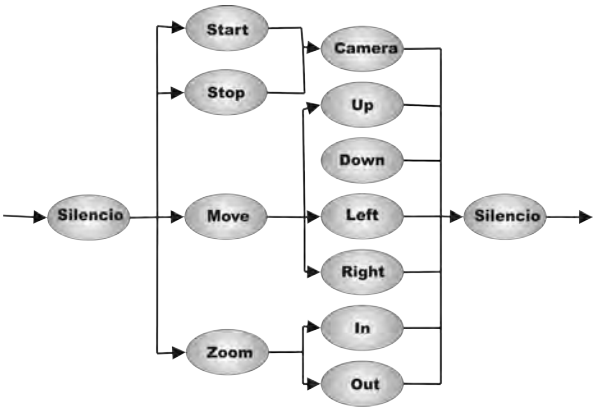


Figura 3.1: Red de palabras representada con reglas de sintaxis

Aplicando HParse al archivo escrito anteriormente se genera:

```
N=17 L=23 I=0 W=!NULL          J=3 S=3 E=4
I=1 W=!NULL                    J=4 S=5 E=4
I=2 W=SILENCIO                 J=5 S=2 E=5
I=3 W=START                    J=6 S=4 E=6
I=4 W=!NULL                    J=7 S=6 E=7
I=5 W=STOP                     J=8 S=9 E=7
I=6 W=CAMERA                   J=9 S=10 E=7
I=7 W=!NULL                    J=10 S=11 E=7
I=8 W=MOVE                     J=11 S=12 E=7
I=9 W=UP                       J=12 S=14 E=7
I=10 W=DOWN                    J=13 S=15 E=7
I=11 W=LEFT                    J=14 S=2 E=8
I=12 W=RIGHT                   J=15 S=8 E=9
I=13 W=ZOOM                    J=16 S=8 E=10
I=14 W=IN                      J=17 S=8 E=11
I=15 W=OUT                     J=18 S=8 E=12
I=16 W=SILENCIO                J=19 S=2 E=13
J=0 S=16 E=1                   J=20 S=13 E=14
J=1 S=0 E=2                    J=21 S=13 E=15
J=2 S=2 E=3                    J=22 S=7 E=16
```

Puede ser entendido de la siguiente forma: es una red de palabras que posee 17 nodos (numerados con I) y 23 enlaces (numerados con J), los nodos !NULL son nodos a los que llegan o salen enlaces pero no tienen asociadas palabras, como el nodo de inicio de la red y el de fin, estos nodos deberían reducir el número de enlaces en el caso de que haya palabras en paralelo o ciclos en la estructura de la red. Si se desea, se puede ponderar cada enlace por una probabilidad de transición logarítmica que se coloca en el cuarto campo del enlace. Los otros dos campos, S y E, son el nodo de partida y el de llegada respectivamente.

2. Definición del diccionario

   Después de definir la gramática, debe hacerse lo mismo con el diccionario. El diccionario debe poseer todas las palabras ordenadas alfabéticamente, que aparezcan tanto en el entrenamiento como en las pruebas, con su respectiva representación con base en las cadenas de Markov (fonemas) y con todas las posibles pronunciaciones.

   HTK ofrece la herramienta HDMan que prepara el diccionario de pronunciación desde una o más

fuentes. HDMan puede leer una lista de comandos escrita en un archivo de texto para editar y unir copias de uno o más diccionarios. Por ejemplo puede usarse comandos para añadir silencios al final de cada palabra, reemplazar símbolos de fonemas acentuados por fonemas sin acento o unir dos fonemas consecutivos en uno.

Como fuentes se usaron el diccionario definido para el habla inglesa "British English Example Pronciation dictionary-BEEP" propiedad de Oxford University [74], el cual sólo puede ser usado para propósitos de invetigación, y un diccionario propio de las palabras que están dentro de la base de datos, pero no se encuentran en BEEP, además de los comandos definidos en la gramática.

Se hizo necesario hacer varias modificaciones tanto en el diccionario BEEP, como en archivo de la lista de frases contenida en TIMIT para usarlas con las herramientas ofrecidas por HTK. Por ejemplo, se quitaron los símbolos que no se usan del diccionario, como la exclamación, el puto final, la interrogación o las comillas, y se colocó el identificador de la frase precediéndola, pues en TIMIT está primero la frase y depués el identificador. Se usó el lenguaje de programación Matlab para hacer dichas modificaciones.

Antes de continuar con la creación del diccionario, se debe poseer una lista de las palabras usadas en la base de datos. Esta lista debe tener las palabras ordenadas alfabéticamente pero sin repetición. Finalmente se puede usar HDMan teniendo como entrada las dos fuentes y la lista de palabras, se obtiene un diccionario con la siguiente forma. HDMan también puede crear una lista de los fonemas, la cual debe ser usada en el entrenemiento, pues obviamente representa los HMMs.

```
A ah

A ax

A ey

ABDOMEN ae b d ax m ax n

ABILITY ax b ih l ih t iy

ABOUT ax b aw t

ABRUPTLY ax b r ah p t l iy

ABSENCES ae b s ax n s ih z

ABYSS ax b ih s

ACCEPTANCE ax k s eh p t ax n s

ACCESS ae k s eh s
```

3. Archivos de etiquetado

   Se deben tener los achivos de audio con sus respectivas transcripciones a nivel de fonemas. TIMIT posee los archivos de audio en formato WAV NIST (National Institute of Standards and

Technology), pero para manipular estos archivos fácilmente con HTK deben estar en el formato propio, lo cual puede hacerse usando HCopy, que es una herramienta para copiar uno o varios archivos en un archivo de salida y también puede usarse para hacer conversión de formatos o parametrización. En cuanto a las transcripciones, en TIMIT la escala de tiempo está dada en muestras, mientras que HTK emplea unidades de 100 ns. Debido a que la frecuencia de muestreo de TIMIT es de 16000 Hz, para convertir las muestras en tiempo en unidades HTK deben multiplicarse por 625 ((1/(16000X100E − 9)), como se puede observar a continuación.

```
0 9640 h#                    0 6025000 sil
9640 11240 sh                6025000 7025000 sh
11240 12783 iy               7025000 7989375 iy
12783 14078 hv               7989375 8798750 hv
14078 16157 ae               8798750 10098125 ae
16157 16880 dcl              10098125 10550000 dcl
16880 17103 d                10550000 10689375 d
```

Si alguno de estos archivos no se tiene, HSLab puede ser usado tanto para la adquisición de audio como para la manipulación de archivos de etiquetado. Por ejemplo se podría cargar un archivo de audio, determinar los límites de las unidades del habla y guardar los cambios (Figura 3.2). HSLab es la única herramienta de HTK con entorno gráfico.

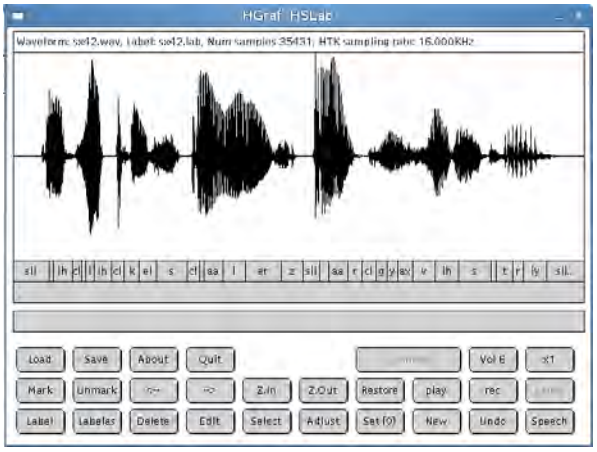


Figura 3.2: Programa usado para adquisición y etiquetado del audio

Siendo una base de datos grande, también se puede crear un archivo MLF (Master Label File) para los fonemas y otro para las palabras, lo cual hace más rápido el entrenamiento. Un MLF es un documento que contiene el conjunto completo de todas las transcripciones. HTK ofrece la herramienta HLEd, que es un simple editor para manipular archivos de etiquetado usando los comandos almacenados en un archivo de texto.

4. Extracción de características

   Antes de comenzar el entrenamiento del sistema ASR, se debe decidir qué características se usarán como entrada. En este caso son los primeros 12 índices de Mel, la energía de la señal y las primeras dos derivadas temporales, dando un total de 39 características por muestra. El análisis se hace en ventanas de 20 milisegundos con traslape del 50 %. Esto equivale a una observación cada 10 milisegundos, siendo la observación el vector de características acústicas de cada muestra. Además, debido a que estas características son dependientes del volumen de la señal de audio de entrada, se normalizaron todos los archivos antes del cálculo de los índices de Mel.

   De nuevo, mediante la herramienta HCopy puede hacerse esta parametrización de la señal. En la línea de comandos de linux puede teclearse

   ```
   HCopy -C config -S lista.scp
   ```

   Donde *config* es un archivo de configuración de parámetros y lista.scp es la lista de los archivos a codificar. En *config* debe especificarse el formato de la fuente, el tamaño de la ventana y del traslape, el uso de las derivadas temporales y de la energía, el número de filtros (NUMCHANS) y la escala de los coeficientes (CEPLIFTER) según la fórmula 3.1, debido a que en general los coeficientes de orden superior son mucho menores, y aunque HTK no tiene problemas con esto, es lo que se hace normalmente y evita problemas numéricos.

$$c'_n = \left(1 + \frac{L}{2} \sin \frac{\pi n}{L}\right) c_n \tag{3.1}$$

```
 # Coding parameters
TARGETKIND = MFCC_0_D_A
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 200000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
ENORMALISE = F
```

5. Creación y entrenamiento de los modelos

   Ahora se pueden entrenar los modelos. Primero se crea el prototipo de cada cadena de Markov. No existe forma directa de definir el número de estados, la matriz de transición o la función de densidad de probabilidad para cada estado, pero en la literatura se ha asumido con generalidad, que para fonemas, el modelo debe tener tres estados activos y ser de propagación hacia adelante (izquierda derecha), también es necesario que posea otros dos estados, uno al comienzo y otro al final, para permitir la creación de redes de cadenas (fonemas), que forman las palabras. La suposición más simple, es que cada estado posee una distribución de probabilidad normal, pero puede ser refinado, al tratar de modelarla con una mezcla de gaussianas (2.9).

   El prototipo es inicializado con la media y la covarianza global al usar HCompV, quedando de la forma

```
~ o <STREAMINFO> 1 39 <VECSIZE> 39 <NULLD><MFCC_D_A_0><DIAGC>
~ h "proto"
<BEGINHMM>
   <NUMSTATES> 5
      <STATE> 2
         <MEAN> 39
         -9.78 -6.97 -5.13 -9.38 ...
         <VARIANCE> 39
         1.03 5.99 7.95 8.26 ...
         <GCONST> 1.389399e+02
      <STATE> 3
         <MEAN> 39
         ...
         <VARIANCE> 39
         ...
         <GCONST> ...
      <STATE> 4
         <MEAN> 39
         ...
         <VARIANCE> 39
         ...
         <GCONST> ...
      <TRANSP> 5
```

0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.6	0.4	0.0
0.0	0.0	0.6	0.4	0.0
0.0	0.0	0.0	0.7	0.3
0.0	0.0	0.0	0.0	0.0

<ENDHMM>

Una vez definido el prototipo, se repite para cada uno de los HMMs, y se estiman los parámetros para cada modelo de Markov con el algoritmo de Baum-Welch basado en el principio de máxima verosimilitud, al maximizar los parámetros para que los tres estados dados por los fonemas en los archivos etiquetados, produzcan la observación dada en el archivo de características. HERest es empleado para realizar la estimación de los parámetros, y para poder usarlo se necesita, además de la definición de los modelos obtenida al clonar el prototipo, la transcripción de los archivos de entrenamiento a nivel de fonemas y los archivos de parametrización (índices de Mel). Este paso se repite hasta que el promedio de la probalilidad logarítmica por observación se estabilice.

Después se realinean todos los datos de entrada, es decir, se busca en el diccionario la pronunciación de cada una de las palabras presentes, se selecciona la que más se asemeja dados el vector de características y las cadenas de Markov, y se redefinen los límites en el tiempo. Esta tarea de reconocimiento se hace con HVite, que es un reconocedor de palabras que compara el archivo de audio (o el vector de características) contra la red de HMMs, usando el algoritmo de Viterbi, el cual busca la probabilidad máxima de que algún modelo haya producido la observación. Estando realineados los datos, se reentrena hasta que de nuevo se estabilice la probabilidad logarítmica.

También se agrega un modelo de silencio impulsivo, que permite que los modelos absorban silencios que se encuentran dentro de los registros sin necesidad de pasar a la palabra siguiente, y se refina al crear otras cadenas de Markov que modelan mejor el habla. Estas cadenas se hacen para tener en cuenta el contexto de los fonemas, formando así los trifonemas y bifonemas. Por ejemplo, en la Figura 3.3 se muestra la transcripción de la palabra *bliblical* contenida en la frase *biblical scholars argue history* (Figura 3.2), a nivel de fonemas, de trifonemas y de palabras.

Los modelos para los trifonemas y bifonemas son hechos simplemente clonando los monofonemas, al tener en cuenta la evidencia dada por los datos de entrada, y así crear un modelo por cada tres fonemas (en el interior de las palabras, o cuando se unen con otras), por cada dos fonemas (cuando se encuentran al principio o al final de la palabra). Con el fin de crear los archivos de etiquetado a nivel de trifonemas se usa de nuevo HLEd, y para clonar los HMMs, se emplea un editor que usa los comandos almacenados en un archivo de texto para transformar, clonar o

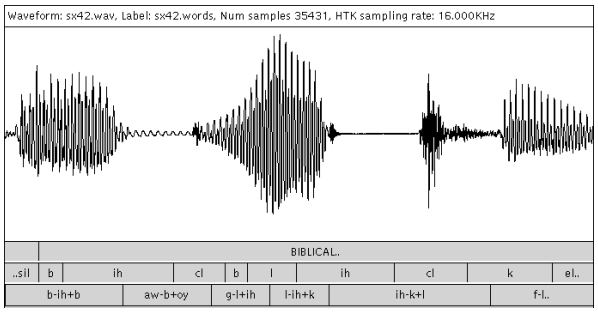


Figura 3.3: Palabra etiquetada a nivel de trifonemas

editar la estructura de HHMs llamado HHEd.

Obviamente, algunos modelos deben compartir un conjunto de parámetros. Ésto se consigue al crear un árbol de decisión, para definir cuáles estados deben ser compartidos usando HHEd.

Una vez creados los nuevos HMMs se entrena de nuevo. Aquí termina el diseño y la implementación del sistema ASR para habla continua. Los estados de los HMMs poseen una función de densidad de probabilidad normal, pero, usando HHEd y reentrenando se podría hacer que la función de probabilidad de cada estado sea modelada con una mezcla de gaussianas. El sistema fue refinado al usar tres gaussianas por cada estado.

Para cada uno de los locutores se asignaron varias frases distintas, pero todos ellos debieron pronunciar dos frases comunes: “she had your dark suit in greasy wash water all year” y “don’t ask me to carry an oily rag like that”. Por otro lado, en la base de datos TIMIT se sugiere usar una parte para el entrenamiento y otra para realizar las pruebas, en los archivos de audio sugeridos para realizar las pruebas hay 336 muestras de estas dos frases.

Usando como grámatica las siguientes líneas se puede usar el sistema entrenado para realizar la alineación sobre los archivos de prueba:

```
(SILENCIO ((SHE HAD YOUR DARK SUIT IN GREASY WASH WATER ALL YEAR
)|(DONT ASK ME TO CARRY AN OILY RAG LIKE THAT)) SILENCIO)
```

Siendo así, se puede observar el resultado de alinear los datos sobre una muestra en la Figura 3.4. Las primeras dos líneas muestran la alineación lograda por el sistema con una gaussiana por estado y en la segunda cuando se emplea una mezcla de tres gaussianas para modelar cada estado. Es evidente que el sistema trabaja bien con los dos modelos para el caso de definir los límites de las palabras cuando se conoce el orden de éstas en la frase.

En la Figura 3.5 se puede corroborar que los resultados del sistema para alineación de palabras son buenos, la figura muestra las matrices de confusión como una imagen en escala de grises, donde el

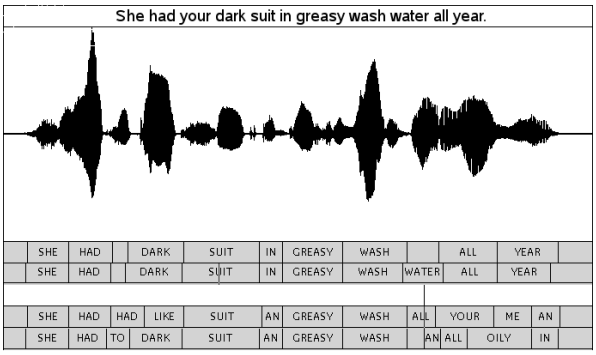
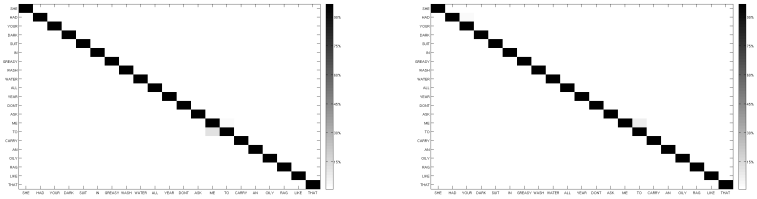


Figura 3.4: Frase alineada

blanco representa 0% de palabras reales (filas) siendo reconocidas por el sistema como las palabras de las columnas. En el resultado ideal, sólo la diagonal principal debe ser negra y el resto de la matriz debe ser blanca.



(a) Usando una gaussiana (b) Usando tres gaussianas

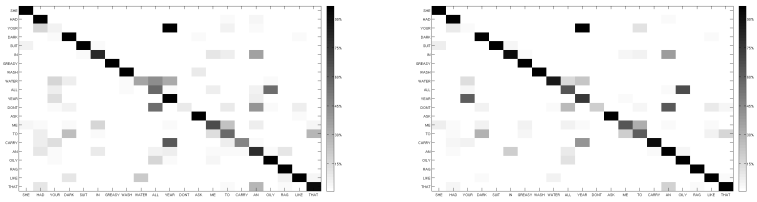
Figura 3.5: Matrices de confusión para el caso de alineación de palabras

Una mejor prueba del sistema es buscar palabras claves en habla continua (Sección 2.3.1). Esto es, buscar las palabras en la frase, sabiendo que puede haber silencio entre ellas o estar unidas, sin presumir el orden en que se encuentran, si están en la muestra, o si se repiten. Esto puede conseguirse al usar la siguiente gramática:

```
$com = SHE | HAD | YOUR | DARK | SUIT | IN | GREASY | WASH |
WATER | ALL | YEAR | DONT | ASK | ME | TO | CARRY | AN | OILY |
RAG | LIKE | THAT | SILENCIO ;
({$com})
```

Los resultados de buscar las palabras que se encuentran en las dos frases repetidas por todos los hablantes como palabras claves en habla continua, se presentan en las dos últimas líneas de la Figura 3.4. Se puede apreciar que el sistema reconoce algunas palabras, los resultados pueden ser analizados de mejor forma al observar la Figura 3.6, en la cual queda claro que tanto al modelar los estados con una o tres gaussianas, el sistema tiene problemas con la palabra *your*, que en la mayoría de los casos

fue reconocida como *year*. También se hace claro que aunque en la matriz de confusión los datos se concentran en la diagonal principal, el sistema comete errores representados como los datos que no se encuentran sobre la diagonal, siendo más recurrentes aquellos más oscuros.



(a) Usando una gaussiana (b) Usando tres gaussianas

Figura 3.6: Matrices de confusión para el caso de busqueda de palabras claves en habla continua

Para el análisis de los resultados HTK ofrece la herramienta HResults, la cual lee un conjunto de archivos de etiquetado y los compara con un archivo de transcripción de referencia. Los resultados de analizar con esta herramienta el desempeño del sistema usado en el reconocimiento de palabras claves, con una sola gaussiana por estado, en la base de datos TIMIT y para las 21 palabras de dos frases comunes para todos los hablantes es:

```
SENT: Correct=0.60 [H=2, S=334, N=336]
WORD: Corr=60.45, Acc=56.44 [H=2381, D=569, S=989, I=158,
N=3939]
```

En la primera línea (SENT) se da la precisión a nivel de frase dada por los archivos de etiquetado que son idénticos a los de referencia, mientras que en la segunda (WORD) expresa la precisión a nivel de palabra. H es el número de etiquetas correctas, D el número de palabras borradas, S el número de sustituciones, I el de inserciones y N el total de etiquetas definidas en los archivos de transcripción.

El porcentaje de etiquetas correctamente reconocidas está dado por (Ecuación 3.2):

$$\%\text{Correct} = \frac{H}{N} \times 100\% \tag{3.2}$$

Mientras que la precisión es calculada como (Ecuación 3.3):

$$\text{Acc} = \frac{H - I}{N} \times 100\% \tag{3.3}$$

De estos datos estadísticos el más representativo para este caso es el porcentaje de palabras correctamente reconocidas.

Al realizar el mismo análisis pero con tres gaussianas por estado se obtiene lo siguiente:

```
SENT: Correct=0.30 [H=1, S=335, N=336]
WORD: Corr=66.65, Acc=63.31 [H=2532, D=398, S=869, I=127,
N=3799]
```

Se puede inferir que se hace una mejora del 6% en el porcentaje de palabras correctamente reconocidas, al utilizar tres gaussianas en vez de una, para el sistema de reconocimiento de habla basado en fonemas, reconociendo palabras claves en habla continua sobre la base de datos TIMIT.

El sistema también debe ser probado con datos que no se encuentren dentro de la base de datos TIMIT. Con este propósito, se adquirieron las mismas dos frases de 5 hablantes distintos. Cada persona repitió cada frase tres veces, dando un total de 30 muestras. Es de aclarar que ninguna de estas personas habla inglés como lenguaje materno (3 son colombianos y 2 árabes), lo que influye obviamente en el resultado, pues el sistema fue entrenado con personas nacidas en Estados Unidos.

En el caso de alineación de palabras sobre los archivos de audio adquiridos en el laboratorio, el desempeño del sistema fue bueno tanto para una, como para tres gaussianas por estado, como se puede ver en las primeras dos líneas de la Figura 3.7. Tanto para el cálculo de la matriz de confusión, como para el análisis usando HResults se utilizó como referencia el archivo de etiquetado arrojado al realizar la alineación con tres gaussianas. Las últimas dos líneas de la Figura 3.7 muestran el resultado de realizar el reconocimiento de palabras claves en habla continua, usando un ASR basado en fonemas con una y tres gaussianas por estado, sobre un archivo de audio adquirido en el laboratorio. Se observa un resultado similar al obtenido sobre la base de datos TIMIT, la matriz de confusión es presentada en la Figura 3.8. El procentaje de palabras reconocidas correctamente fue del 43.91% mientras que la precisión fue el 42.76%

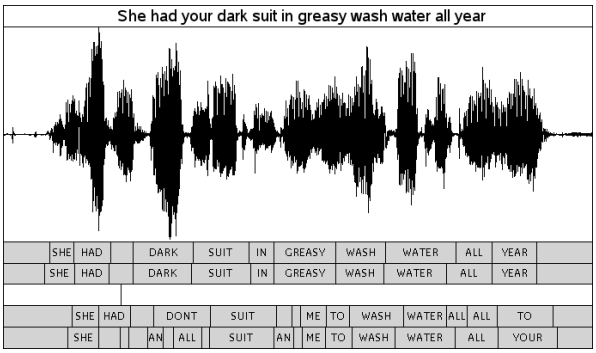


Figura 3.7: Alineación de un archivo de audio adquirido en el laboratorio

El sistema ha sido entrenado con el propósito de utilizarlo para reconocer comandos para manipular un robot, por eso también se probó usando palabras que puedan ser usadas en los siguientes comandos:

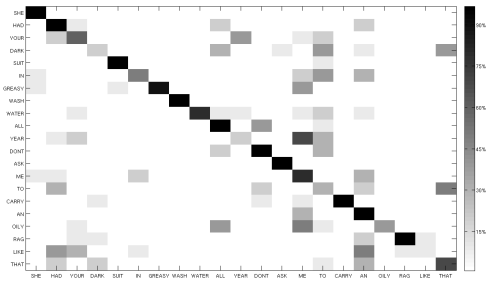


Figura 3.8: Matriz de confusión de palabras claves en habla continua sobre datos adquiridos en el laboratorio

- Star camera
- Stop camera
- Move up
- Move down
- Move left
- Move right
- Zoom in
- Zoom out

Primero se usó el enfoque de reconocimiento de palabras aisladas, es decir, palabras que se encuentran separadas por pausas, que a diferencia del reconocimiento de palabras claves, no se encuentran dentro de la misma muestra, sino que cada palabra es un archivo de audio diferente (o es manejado en el buffer como tal). Se tomaron 364 muestras de 8 personas (2 árabes, 2 franceses, 1 mexicano, 2 colombianos, 1 rumano), y se usó la siguiente gramática para indicar que sólo había una palabra por muestra.

```
$com = START | STOP | UP | DOWN | LEFT | RIGHT | IN | OUT |
CAMERA | MOVE | ZOOM ;
([SILENCIO] $com [SILENCIO])
```

La matriz de confusión del sistema basado en fonemas, usado en el reconocimiento de 11 palabras aisladas, sobre archivos de audio adquiridos en el laboratorio, se ve en la Figura 3.9, se observa que el sistema hace reconocimiento, pero existe una alta tasa de confusión. El procentaje de palabras reconocidas correctamente es de 43,96% al igual que la precisión, pues se trata de palabras aisladas y por lo tanto no hay incerciones de palabras.

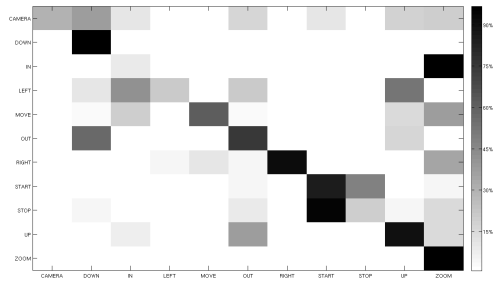


Figura 3.9: Matriz de confusión de palabras aisladas usando fonemas

El sistema se probó entonces usando los comandos aislados, en lugar de palabras aisladas, lo que mejoró el desempeño al establecer reglas, por ejemplo, cuáles palabras pueden preceder a otras, definir varias pronunciaciones o penalizar el silencio. Se usó la siguiente gramática:

```
$comcam = START | STOP;
$direction = UP | DOWN | LEFT | RIGHT;
$comzoom = IN | OUT;
(SILENCIO (( $comcam CAMERA ) | ( MOVE $direction) | (ZOOM
$comzoom)) SILENCIO )
```

El reconocimiento se realizó sobre 319 muestras (de las mismas 8 personas), modelando cada estado con tres gaussianas, obteniendose un procentaje de palabras reconocidas correctamente de 68.65% y la matriz de confusión de la Figura 3.10. El sistema pudo reconocer los comandos aunque no estaban dentro de la base de entrenamiento, gracias a que el modelo se basa en fonemas, pero presenta una alta tasa de confusión entre el comando “Stop camera” y “Star camera”. El sistema aún depende de la pronunciación (acento) y de las características del ambiente.

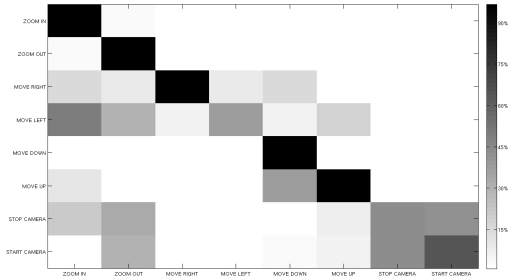


Figura 3.10: Matriz de confusión de comandos aislados

3.1.2 Experimento II: sistema ASR usando VidTIMIT

La base de datos TIMIT no cuenta con información visual, por eso se hace necesario usar otra para el entrenamiento del sistema. La base de datos VidTIMIT fue puesta a disposición de la comunidad científica para propósitos de investigación en el año 2008 [75]. Posee 430 frases en inglés de 43 sujetos (hombres y mujeres), los locutores en su mayoría son de origen australiano, pero también hay presencia de extranjeros, sobre todo de origen oriental, que no hablan inglés como lengua materna. Esta base de datos fue desarrollada por Conrad Sanderson mientras realizaba sus estudios de doctorado en la Universidad Griffith de Australia. Los datos de audio presentan un alto nivel de ruido, las imágenes se encuentran en formato jpg y las frases no están etiquetadas, ni por fonemas ni por palabras (ver el Apéndice A).

Antes de comenzar con el diseño del sistema ASR (descrito en la Sección 3.1.1), se debe acondicionar la base de datos. Para el caso del audio, consiste en segmentar la señal en unidades básicas, y debido a que se desea diseñar un sistema que reconozca palabras (comandos) que no están dentro de la base de datos, estas unidades deben ser fonemas.

Se etiquetaron manualmente 80 frases de la base de datos VidTIMIT, basados en los fonemas para el habla inglesa dados en BEEP y usando HSLab. Debido a que los datos adquiridos en el laboratorio, al igual que la base de datos TIMIT tienen una frecuencia de muestreo de 16 KHz, los datos de la base de datos VidTIMIT fueron submuetreados de 32KHz a 16KHz, y de igual forma que con TIMIT, la señal de audio fue normalizada. Se intentó reducir el ruido, pero debido a las características que éste presenta, es difícil de filtrar sin modificar las características de la señal de habla.

Se entrenó un sistema ASR con la información obtenida de las 80 frases, el cual fue usado para alinear toda la base de datos y así obtener los archivos de etiquetado de las 430 frases. El sistema ASR incluyó otro HMM que modela el ruido presente, además del modelo del silencio que se usó con TIMIT. El sistema se probó sobre todos los hablantes de VidTIMIT y con las mismas dos frases comunes de TIMIT (86 frases en total). El sistema alineó bien las frases de la base de datos que no estaban dentro del entrenamiento, tanto para hombres como para mujeres (Figura 3.11).

Una vez etiquetada la base de datos VidTIMIT a nivel de fonemas, se entrenó otro sistema ASR con el fin de hacer reconocimiento de palabras claves en habla continua. Cuando se buscaron las 21 palabras de las dos frases como palabras claves en habla continua, el sistema tuvo un porcentaje de palabras reconocidas correctamente de 53.90% y una precisión de 52.29%. El desempeño se puede observar en la matriz de confusión dada en la Figura 3.12, y aunque los datos se concentran en la diagonal principal no se puede decir que el sistema funciona para los datos de VidTIMIT.

Al realizar las pruebas con los datos adquiridos en el laboratorio, se pueden obtener buenos resultados con este sistema para el caso de alineación de las palabras como se ve en la Figura 3.13, si no se

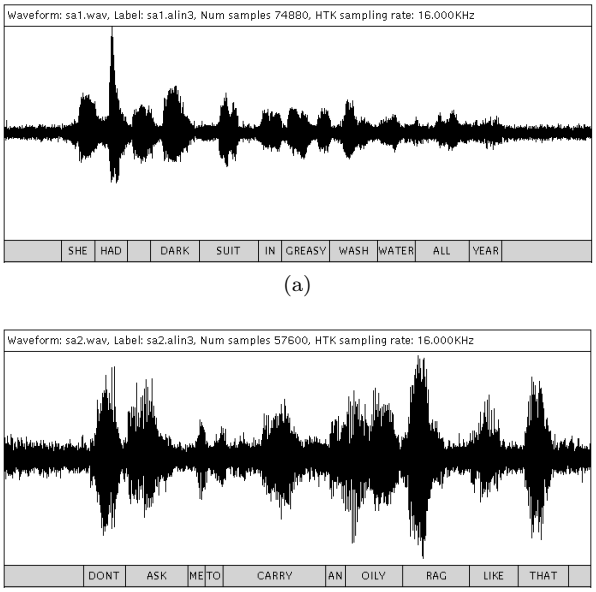


Figura 3.11: Alineación de los datos de la base de datos VidTIMIT

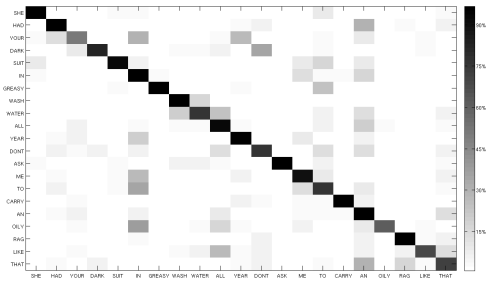


Figura 3.12: Matriz de confusión reconociendo palabras claves sobre la base de datos VidTIMIT

tiene en cuenta los modelos del silencio ni del ruido.

Sin embargo, cuando se trata de reconocimiento de las 21 palabras de las dos frases, como palabras claves en habla continua, los resultados no son aceptables, pues el porcentaje de palabras reconocidas correctamente es de 27.73% mientras la precisión es de -16.81%. El desempeño puede ser observado en la matriz de confusión de la Figura 3.14

El sistema ha sido entrenado con el propósito de utilizarlo para reconocer comandos para manipular un robot, por eso también se realizó el análisis sobre las palabras usadas en los comandos del experimento anterior. El procedimiento se hizo con el enfoque de reconocimieto de palabras aisladas, obteniéndose las matrices de confusión de la Figura 3.15, donde se puede observar que el sistema no reconoce eficientemente las palabras, ni los comandos, con un porcentaje de palabras reconocidas

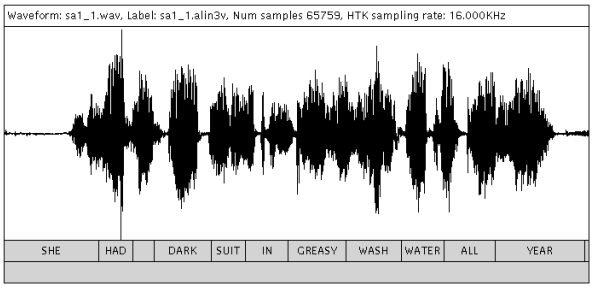


Figura 3.13: Archivo de audio adquirido en el laboratorio alineado usando el sistema entrenado con la base de datos VidTIMIT

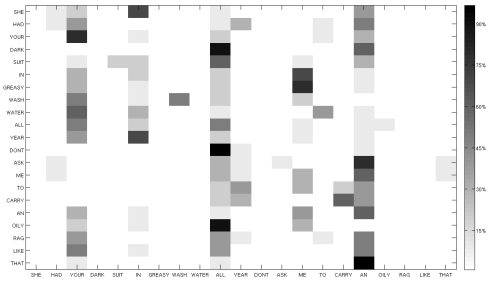
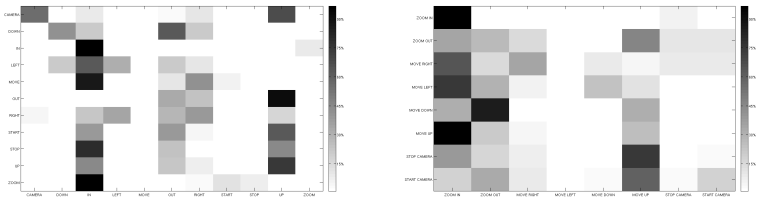


Figura 3.14: Matriz de confusión de palabras claves sobre datos adquiridos en el laboratorio usando el sistema entrenado con la base de datos VidTIMIT

correctamente del 19.23 % y del 25.86% respectivamente.



(a) Palabras aisladas (b) Comandos aislados

Figura 3.15: Matriz de confusión sobre datos adquiridos en el laboratorio usando el enfoque de palabras aisladas y empleando el sistema entrenado con la base de datos VidTIMIT

3.2 Sistema de reconocimiento de habla basado en palabras usando audio

El entrenamiento de sistemas ASR basados en fonemas necesita grandes cantidades de datos, como los presentes en la base de datos TIMIT. Sin embargo, usando un sistema ASR entrenado con esta base para el reconocimiento de palabras claves en habla continua y para reconocimiento de palabras aisladas no fue bueno.

Debido a que el interés particular es el desarrollo de un sistema de reconocimiento de comandos usando información audio-visual, y a que la base de datos TIMIT no cuenta con video, se entrenó un sistema ASR usando las señales de audio brindadas por la base de datos VidTIMIT. Los resultados tampoco fueron aceptables comparados con los obtenidos con la base de datos TIMIT, debido principalmente al ruido presente y a la necesidad de etiquetar a nivel de fonemas la misma base de datos, lo que indudablemente introduce errores. Además, de los resultados anteriores se puede intuir que la respuesta del sistema depende fuertemente de las condiciones de adquisición, de la relación señal a ruido, así como del acento de los locutores.

Por otra parte, se intentó conseguir una base de datos para realizar los mismos experimentos en francés y en español, lo cual no fue posible. Por tal motivo se diseñó y adquirió una base de datos que posee expresiones en inglés, en francés y en español. También cabe recalcar, que aparte de la dificultad de adquirir una base de datos extensa y de etiquetar estos datos a nivel de fonemas del habla francesa, española o inglesa, el problema a tratar en este documento no se centra en hacer reconocimiento en habla continua, como lo sería para un sistema de dictado, sino en hacer reconocimiento de comandos, lo cual cae en el dominio de reconocimiento de palabras aisladas.

En la consola de comando del sistema da Vinci, el cirujano se ubica de forma tal que la única manera de registrar la boca para la implementación del sistema, es haciendo las tomas desde un ángulo bajo (Figura 3.16). Además, teniendo en cuenta que el sistema ASR usará como unidades básicas palabras en lugar de fonemas, los comandos a reconocer deben encontrarse dentro de la base de entrenamiento. Siendo así, la base de datos adquirida en el laboratorio consta de 6 comandos en inglés, 7 en francés y 6 en español:

Comandos en inglés:

“Left”, “Right”, “Up”, “Down”, “Go back”, “Go forward”.

Comandos en francés:

“À gauche”, “À droite”, “Monter”, “Reculer”, “Arriere”, “Avant”, “Descendre”.

Comandos en español:

“Izquierda”, “Derecha”, “Arriba”, “Abajo”, “Atrás”, “Adelante”.

(a) (b) (c)

Figura 3.16: Consola de comando del sistema da Vinci

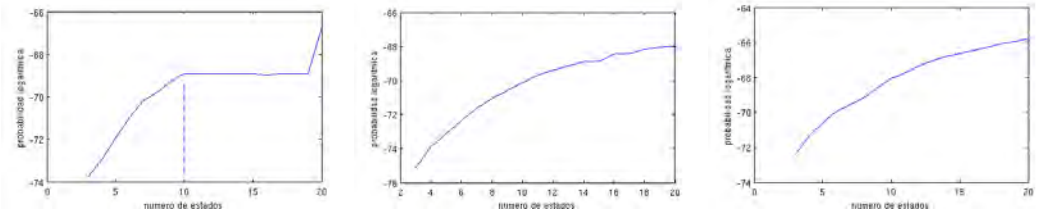
Se grabaron secuencias de video de 18 locutores diciendo 4 frases y 13 comandos dos veces (Apéndice A). Entre los sujetos había 4 mujeres y 14 hombres de diferentes zonas geográficas (Francia, países árabes, Vietnam, Nigeria, México y Colombia). En cuanto a los comandos en español, se grabaron secuencias de video de 18 personas, entre los locutores se encontraban 5 mujeres y 13 hombres todos de nacionalidad colombiana.

3.2.1 Experimento III: sistema ASR usando datos adquiridos en el laboratorio

Se decidió entrenar tres sistemas ASR, uno para el reconocimiento de comandos de habla inglesa, otro de habla francesa y otro de habla española, que usen como unidades básicas palabras en lugar de fonemas y cuyo objetivo sea el reconocimiento de palabras aisladas y no de palabras claves en habla continua. Con este fin se hizo necesaria la implementación de una base de datos en el laboratorio, la cual tuvo que ser segmentada y etiquetada a nivel de palabras. Los sistemas fueron implementados 50 veces empleando aleatoriamente el 70 % de los datos para el entrenamiento y el 30 % para realizar pruebas.

Se ha asumido en la literatura que cuando las cadenas de Markov representan fonemas, el número de estados activos es 3, pero cuando los modelos representan palabras, la configuración de las cadenas de Markov debe ser seleccionada. Por esta razón se hizo el entrenamiento como se ilustra en la Sección 3.1.1 variando desde 3 hasta 20 el número de estados y usando 1, 2 y 3 funciones gaussianas por estado. La Figura 3.17 muestra que para los comandos de habla inglesa, en 10 estados el entrenamiento ha convergido, y aunque existe un mayor valor de probabilidad logarítmica alrededor de 20 estados, ésta es debida al sobreajuste. En inglés no se encontraron mejores resultados que el obtenido con 10 estados y usando una función gaussiana por estado, por lo tanto se mantuvo esta configuración en los experimentos posteriores. En francés y en español, aún en 20 estados la probabilidad logarítmica es

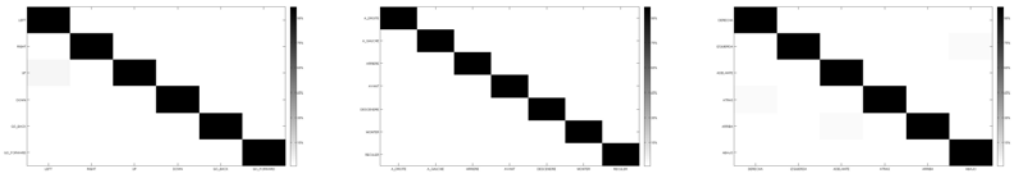
creciente, y en ambos idiomas no se obtuvo mejor resultado que al usar 20 estados con una función gaussiana por estado, y por lo tanto se empleó esta configuración en los experimentos posteriores.



(a) inglés (b) francés (c) español

Figura 3.17: Número de estados de los modelos Vs la probabilidad logarítmica

La respuesta de los sistemas usando dicha configuración puede ser observada en la matriz de confusión de la Figura 3.18. Para el cálculo de estas matrices se usaron todos los datos de la base de datos. Se puede ver que la respuesta de los sistemas fue muy buena, en especial si se compara con las obtenidas al usar como unidades básicas fonemas, debido principalmente a la simplificación del problema, dejando el problema de segmentación y de interpretación de las palabras a un estado de abstracción superior.



(a) Comandos en inglés (b) Comandos en francés (c) Comandos en español

Figura 3.18: Matriz de confusión sobre datos adquidos en el laboratorio usando el enfoque de palabras aisladas

Al usar la herramienta HResults sobre el 30% de los datos, destinados para pruebas para cada una de las 50 repeticiones, se obtuvo un promedio de porcentaje de palabras correctamente reconocidas del 98.46% para los comandos en inglés, del 98.94% en francés, y de 97.54% para los comandos en español, con desviación estándar de 1.64%, 1.34% y 1.42% respectivamente.

4 Sistema de reconocimiento audio-visual del habla

Los primeros tres experimentos se realizaron usando únicamente la información dada por las características de audio, especifícamente los índices de Mel. El sistema ASR será usado para identificar comandos que servirán para controlar un robot, con el fin de hacer el sistema menos sensible al ruido acústico, se realizó el mismo procedimiento hecho con el audio para las características de video que brindan información visual del habla, y se exploró la viabilidad de usar tanto características de audio como de video en el sistema de reconocimiento de comandos.

De nuevo, el sistema de reconocimiento se implementó usando modelos ocultos de Markov (HMM) y se emplearon tanto fonemas como palabras como unidades básicas. En cuanto al conjunto de características visuales, se emplearon características de alto nivel, las cuales se basan en la forma de los labios. Y como herramienta computacional se utilizó HTK, haciendo que el conjunto de características cumpla con el formato para archivos propios de HTK (Capítulo 3).

En este capítulo se describen los algoritmos implementados tanto para hacer seguimiento de los labios sobre la secuencia de video, como para calcular el conjunto de características. Después se muestran los resultados de diferentes sistemas de reconocimiento del habla usando únicamente características visuales y usando información audio-visual.

4.1 Sistema de reconocimiento de habla basado en fonemas usando video

Se pretende mejorar la robustez del sistema al incluir la información de video. Las características fueron extraídas de la base de datos VidTIMIT usando los puntos que definen el contorno externo de la boca, según MPEG-4 (Sección 2). Los estándares de animación del cuerpo y de la cara definidos en MPEG-4 están basados en la estructura ósea y muscular del ser humano, y aunque no permiten que se generen todos los movimientos, pues algunos son propios de cada persona, es el esfuerzo más cercano hasta ahora y es el estándar que se usa en este momento en la industria cinematográfica.

Debido a que la base de datos VidTIMIT cuenta con 103543 imágenes, se hace poco práctico ubicar manualmente los 10 puntos del contorno externo de la boca sobre cada una. Por lo tanto, se diseñó un algoritmo asistido de seguimiento de los puntos del contorno externo de la boca (Algoritmo 1), el cual se implementó en Matlab. La calidad de las imágenes de la base de datos VidTIMIT no es muy buena, pues se encuentran en formato jpeg, tienen una resolución de 512x384 píxeles, son de toda la cara y la región de interés es de 100x50 píxeles aproximadamente.

El seguimiento de los labios es aún un problema abierto de visión artificial debido a la complejidad de las formas, colores, texturas y los cambios de iluminación [76]. Este problema ha sido exitosamente tratado para vistas laterales y con el fondo controlado [77], pero para vistas frontales y sin marcadores de labios ha mostrado ser más complicado. Para el caso de imágenes en escala de grises, los métodos fallan en localizar los límites de la boca en áreas de contraste pobre como el labio inferior, y además son muy sensibles a cambios de iluminación o sombras.

4.1.1 Seguimiento del contorno externo de los labios

Algoritmo 1 Seguimiento asistido de puntos del contorno externo

Entradas: Video en forma de secuencia de imágenes $c_1, c_2, \ldots, c_n \in C$.
Salida: La secuencia de puntos del contorno externo de la boca para todos los cuadros de video S_{nx10}.
Los siguientes pasos se realizan para todos los cuadros.
[***Paso 1:***] Localización de la región de interés.
[***Paso 2:***] Ubicación de los 10 puntos del contorno externo de la boca $p_1, p_2, \ldots, p_{10} \in P$ para el primer cuadro de video.
[***Paso 3:***]
para todos los puntos del contorno de cuadro anterior $S_{n-1,i}$ **hacer**
  Calular la similitud con los píxeles de la vecindad en el cuadro presente C_n.
  Escoger el candidato a punto actual $S_{n,i}$ como el píxel donde la similitud es mayor (más cercana a la unidad)
fin para
[***Paso 4:***] Una vez obtenidos los 10 puntos candidatos, aplicar algunas restricciones de forma:
$\mathrm{restricciones}(S_n)$ (Algoritmo 2)

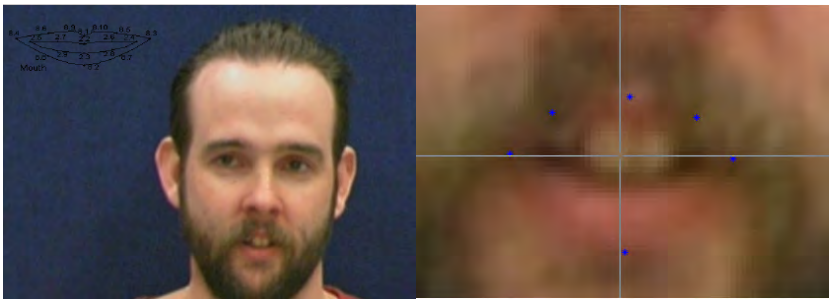
El algoritmo que se propone para el seguimiento del contorno externo de la boca, está basado en apariencia y en restricciones morfológicas definidas en el estándar MPEG-4 (Algortimo 1). El algoritmo usa el grupo 8 que describe el contorno externo de los labios [31] pues algunos estudios psicológicos han sugerido que es el que más influencia tiene en la lectura de los labios. Además, en [22] se muestra que el uso del grupo 2, que describe el contorno interno de la boca, no aumenta significativamente el rendimiento de un sistema de reconocimiento automático de habla, y los algoritmos usados son

significativamente más costosos que los del contorno externo.

En general, los pasos en un sistema de seguimiento de la boca son: detección de la región de la boca, localización de los labios (inicialización), seguimiento de los labios y la extracción de características [78]. El seguimiento es explicado con detalle en el Algortimo 1.

Paso 1, detección de la región de la boca La región de interés es localizada de forma asistida únicamente en el primer cuadro de video de la secuencia.

Paso 2, localización de los labios Para iniciar el algoritmo de seguimiento de los labios, se hace necesaria la ubicación exacta de los puntos que describen el contorno externo de la boca. Debido a que la segmentación robusta de la boca ante la presencia de barba, tono de piel, cambios de iluminación, presencia de lengua y calidad de la imagen aún es un problema abierto y sólo se han obtenido buenos resultados para la extracción del contorno sobre imágenes de alta definición, la inicialización en este caso se hace de forma manual. Se ubican manualmente sobre el primer cuadro de video los 10 puntos (4.1).



(a) Interfaz del programa

(b) Ubicación manual de los 10 puntos

Figura 4.1: Inicialización del algoritmo de seguimiento asistido

Paso 3, seguimiento de los labios Con el fin de realizar el seguimiento de los labios se usa una medida de similitud entre cuadros de video, además de algunas restricciones morfológicas dadas en el estándar MPEG-4.

La medida de similitud se hace sobre los píxeles pertenecientes a la vecindad de cada uno de los 10 puntos que definen el contorno externo. Primero, se calcula la distancia en el espacio de color (R,G,B) de la ventana centrada en el punto hallado en el cuadro de video anterior (V) con las ventanas en el cuadro presente, centradas en cada uno de los píxeles de la vecindad de interés (V_{ij}). El cuadro presente es además comparado con el primer cuadro de la secuencia se video, el cual posee información altamente confiable, debido a que los puntos del contorno de la boca de este cuadro no fueron calculados (Ecuación 4.1).

$$d_{ij} = \| V - V_{ij} \| + \| V_1 - V_{ij} \| \tag{4.1}$$

La distancia será mínima si el color (R,G,B) de cada uno de los píxeles de la ventana del cuadro de video anterior concuerda exactamente con los de alguna de las ventanas del cuadro de video actual y su valor máximo será definido por el tamaño de las ventanas. Con el fin de normalizar la distancia y usarla como medida de similitud, la distancia es usada como el argumento de la función exponencial negativa. Siendo así, el rango se encuentra entre 1 y 0, 1 para una total concordancia y 0 para cuando las ventanas son totalmente diferentes (Ecuación 4.2).

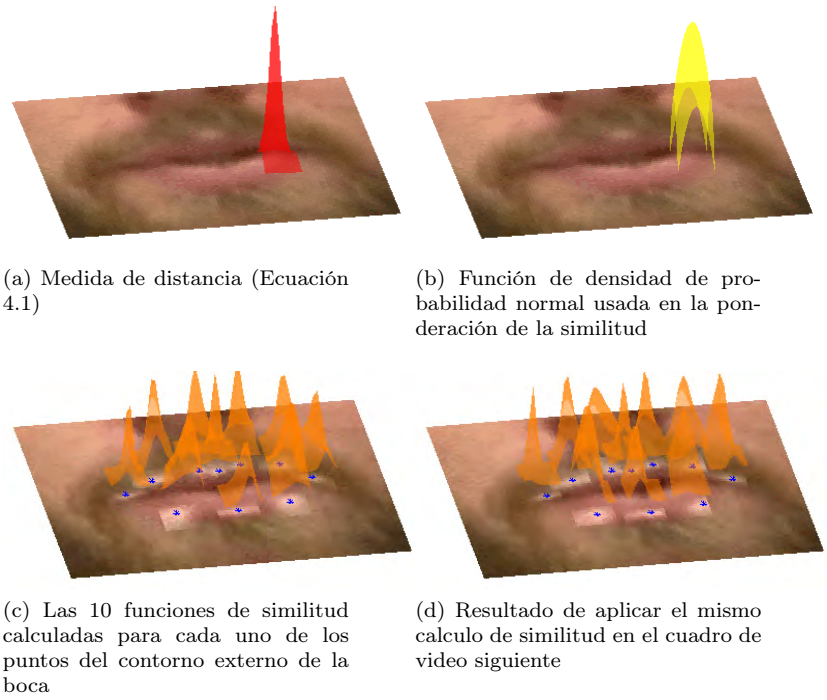
$$c_{ij} = e^{-d_{ij}} \tag{4.2}$$

La similitud es entonces ponderada usando una función de densidad de probabilidad normal con media en el punto del cuadro de video anterior y con desviación estándar igual al tamaño del vecindario (Figura 4.2b). Así se consigue dar más peso a aquellos píxeles cercanos al punto hallado en el cuadro de video anterior, debido a que es más probable que correspondan al punto en el cuadro de video actual. Se escoge como candidatos a cada uno de los 10 puntos que conforman el contorno externo de la boca, aquellos cuya similitud sea más cercana a la unidad. En las Figuras 4.2c y 4.2d se observa el resultado de este procedimiento en dos cuadros seguidos, los píxeles iluminados representan la probabilidad de los píxeles de convertirse en cada punto.

Con el objetivo de hacer el seguimiento de los labios más robusto, se hace que los puntos candidatos hallados con la similitud máxima, cumplan las restricciones morfológicas de la boca, así como las restricciones sugeridas en el estándar MPEG-4 (Tabla 4.1) (Algoritmo 2). La forma de la boca está caracterizada por ser simétrica (simetría reflexiva sobre el eje vertical). Para satisfacer las restricciones de simetría reflexiva se usan dos polinomios de segundo grado (parábolas) con eje de simetría vertical (Ecuación 4.3), debiéndose rotar primero todos los puntos de modo que el punto 4 y el punto 3 queden a 0 grados. Entonces, se ajusta el polinomio superior con los candidatos a puntos 4, 6, 9, 10, 5 y 3, y el polinomio inferior con los candidatos 4, 8, 2, 7 y 3, teniendo en cuenta las ubicaciones sugeridas por el estándar MPEG-4 en la Tabla 4.1.

$$y = ax^2 + bx + c \tag{4.3}$$

En cuanto a las restricciones morfológicas sugeridas en la Tabla 4.1, cabe recalcar que el punto 7.1, el cual corresponde al punto de rotación de la cabeza, para este caso es desconocido. Por tal motivo, las abscisas tanto del punto 1 como del punto 2, correspondientes al punto medio entre



(a) Medida de distancia (Ecuación 4.1)

(b) Función de densidad de probabilidad normal usada en la ponderación de la similitud

(c) Las 10 funciones de similitud calculadas para cada uno de los puntos del contorno externo de la boca

(d) Resultado de aplicar el mismo calculo de similitud en el cuadro de video siguiente

Figura 4.2: Ponderación de la similitud por una función de probabilidad normal

los vértices de la boca, son hechas iguales a $(8.3x + 8.4x)/2$, teniendo en cuenta que aunque no se pueda satisfacer la restricción dada en la Tabla 4.1, la boca es simétrica. En el mismo orden de ideas, para los puntos 9 y 10 pertenecientes al arco de cupido (para los cuales no se definen restricciones en la Tabla 4.1), se igualan las abscisas.

Con los puntos 9 y 10 se usan aún más restricciones, pues tampoco se permite un movimiento entre un cuadro de video y otro superior al 20 % de la distancia media al punto 1, ni que alguno de los dos haga un cruce por el eje vertical.

También se ajustan las abscisas de los puntos 5, 6, 7 y 8 según la Tabla 4.1, y se hallan las ordenadas de los puntos 6, 9, 10 y 5 al evaluar el polinomio que modela la parte superior de los labios, y de los puntos 8 y 7 al evaluar el polinomio que modela la parte inferior. Finalmente, se debe invertir la rotación hecha.

Con el fin de observar el funcionamiento del Algoritmo 1, en la Figura 4.3 se presenta el resultado sobre una secuencia de video cada 10 cuadros. En este caso se utilizó un vecindario de búsqueda de $11x11$ píxeles y una ventana para el cálculo de la similitud de $11x11$ píxeles. Siendo así, la función de ponderación es una función de distribución normal centrada en cada uno de los 10 puntos del contorno externo del cuadro de video anterior, y con desviación 11 (el tamaño del vecindario). Las zonas iluminadas alrededor de cada punto representan la probabilidad dada

Tabla 4.1: Localización recomendada para los puntos característicos del contorno externo de la boca (el punto 7.1x corresponde al punto de rotación de la cabeza)

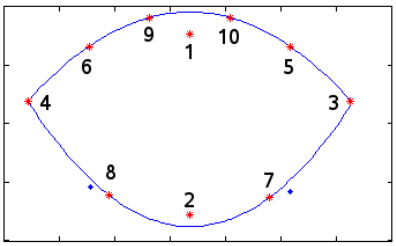
FP	Descripción	Localización recomendada
8.1	Punto medio del contorno externo del labio superior	7.1x
8.2	Punto medio del contorno externo del labio inferior	7.1x
8.3	Esquina izquierda del contorno externo de los labios	
8.4	Esquina derecha del contorno externo de los labios	
8.5	Punto medio entre FP 8.3 y 8.1 en el contorno externo del labio superior	$(8.3x + 8.1x)/2$
8.6	Punto medio entre FP 8.4 y 8.1 en el contorno externo del labio superior	$(8.4x + 8.1x)/2$
8.7	Punto medio entre FP 8.3 y 8.2 en el contorno externo del labio inferior	$(8.3x + 8.2x)/2$
8.8	Punto medio entre FP 8.4 y 8.2 en el contorno externo del labio inferior	$(8.4x + 8.2x)/2$
8.9	Punto superior derecho del arco de cupido	
8.10	Punto superior izquierdo del arco de cupido	

Algoritmo 2 Restricciones

Entradas: Los 10 puntos del contorno externo de la boca $p_1, p_2, \ldots, p_{10} \in P$.

Salida: Los 10 puntos del contorno externo de la boca $p_1, p_2, \ldots, p_{10} \in P$.

[***Paso 1:***] Rotar los puntos de forma tal que el punto p_4 quede a 0^o con el punto p_3.



[***Paso 2:***] Encontrar la parábola más cercana a los puntos p_4, p_6, $_9$, p_{10}, P_5 y p_3 (parábola superior).

[***Paso 3:***] Encontrar la parábola más cercana a los puntos p_4, p_8, p_2, p_7 y p_3 (parábola inferior).

[***Paso 4:***] Hacer la abcisa del punto p_1 como el promedio entre la abcisa del punto p_3 y del p_4.

$p_{1x} = \frac{p_{3x}+p_{4x}}{2}$

[***Paso 5:***] Hacer $p_{2x} = \frac{p_{3x}+p_{4x}}{2}$

[***Paso 6:***] Hacer $p_{6x} = \frac{p_{4x}+p_{9x}}{2}$

[***Paso 7:***] Hacer $p_{5x} = \frac{p_{3x}+p_{10x}}{2}$

[***Paso 8:***] Hacer $p_{3x} = \frac{p_{7x}+p_{2x}}{2}$

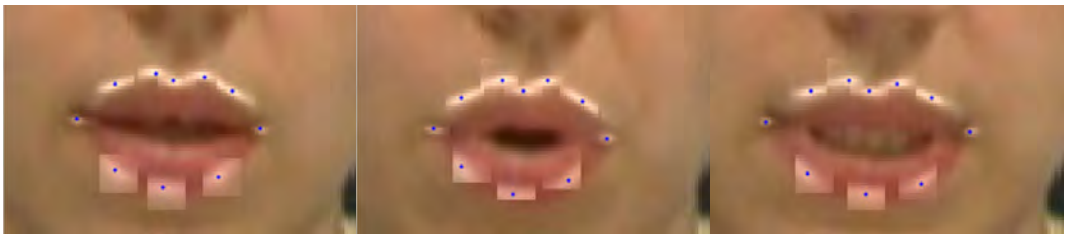
[***Paso 9:***] Hacer $p_{8x} = \frac{p_{4x}+p_{2x}}{2}$

[***Paso 10:***] Proyectar los puntos p_6, p_9, p_{10} y p_5 a la parábola superior.

[***Paso 11:***] Proyectar los puntos p_7, y p_8 a la parábola inferior.

[***Paso 12:***] Deshacer la rotación hecha en el Paso 1.

por la medida de similitud, de que los píxeles sean los nuevos puntos que describen el contorno externo de la boca.



(a) Cuadro 1 (b) Cuadro 11 (c) Cuadro 21

Figura 4.3: Seguimiento de los 10 puntos que conforman el contorno externo de la boca en una secuencia de video cada 10 cuadros

En la Figura 4.4 se aprecia el seguimiento del contorno externo de la boca sobre tres secuencias pertenecientes a la base de datos VidTIMIT. Las secuencias poseen alrededor de 100 cuadros de video, siendo así los resultados son mostrados cada 30 cuadros. Dado que el ancho de la boca sobre la base de datos VidTIMIT tiene una media de 55 píxeles con una desviación estándar de 6,76 píxeles, se usó una ventana para el cálculo de la similitud de 21x21 y un vecindario de búsqueda de 5x5 píxeles.

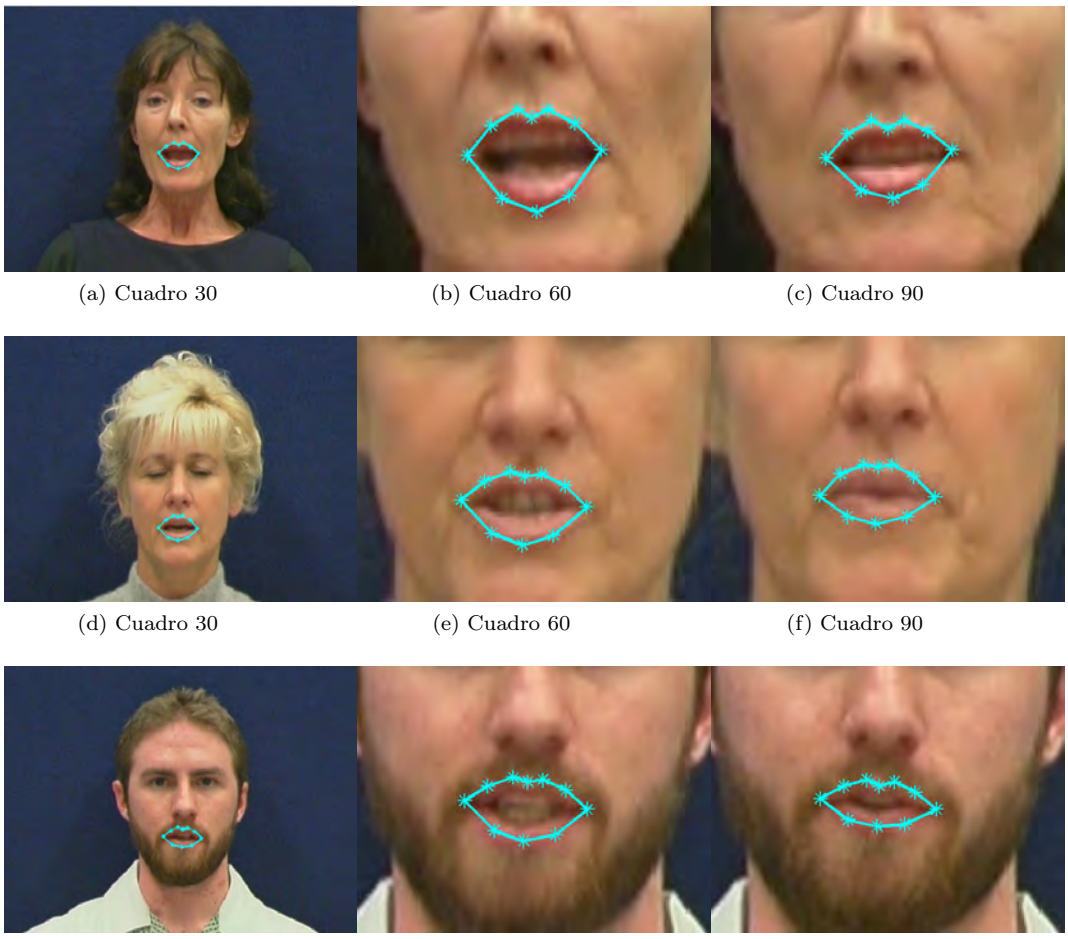
Paso 4, extracción de características Una vez encontrados los puntos en la secuencia de video, las características de la forma de la boca deben ser calculadas. Teniendo los 10 puntos sobre toda la secuencia de video, se pueden encontrar, entre muchas otras, el área de la región dentro de los labios, la redondez, el factor de forma, la relación entre el eje horizontal y el vertical, el perímetro y diferentes relaciones geométricas entre los puntos.

El área es calculada en la forma polar según la Ecuación 4.4, donde r corresponde a la distancia de cada uno de los 10 puntos hasta el centro de la boca, y Θ al ángulo en radianes de separación entre un punto y otro (Algoritmo 3).

$$A = \sum_{i=1}^{10} r_i^2 \Delta\Theta \tag{4.4}$$

Por su parte la redondez es hallada usando la Ecuación 4.5, en la cual A corresponde al área dentro del contorno y d al diámetro mayor equivalente al ancho de la boca, es decir a la distancia entre los puntos 3 y 4 que definen el contorno externo de la boca según el estándar MPEG-4 (Algoritmo 3).

$$R = \frac{4A}{\pi d^2} \tag{4.5}$$



(a) Cuadro 30 (b) Cuadro 60 (c) Cuadro 90

(d) Cuadro 30 (e) Cuadro 60 (f) Cuadro 90

(g) Cuadro 30 (h) Cuadro 60 (i) Cuadro 90

Figura 4.4: Seguimiento del contorno externo de los labios sobre tres secuencias de video de la base de datos VidTIMIT

Tabla 4.2: Unidades de los parámetros de animación de la cara.

IRISD0	diámetro del iris	IRISD=IRISD0/1024
ES0	separación de los ojos	ES=ES0/1024
ENS0	separación ojo-nariz	ENS=ENS0/1024
MNS0	separación boca-nariz	MNS=MNS0/1024
MW0	ancho de la boca	MW=MW0/1024
AU	unidad de angulo	10E-5 rad

El perímetro a su vez, se calcula al sumar las distancias entre los puntos como se indica en la Ecuación 4.6 (Algoritmo 3). Donde p_i corresponde a las coordenadas (x, y) del punto i.

$$\begin{aligned} Per = \|S_{i,5} - S_{i,3}\| + \|S_{i,10} - S_{i,5}\| + \|S_{i,1} - S_{i,10}\| + \|S_{i,9} - S_{i,1}\| + \|S_{i,6} - S_{i,9}\| + \\ \|S_{i,4} - S_{i,6}\| + \|S_{i,8} - S_{i,4}\| + \|S_{i,2} - S_{i,8}\| + \|S_{i,7}x - S_{i,2}\| + \|S_{i,3} - S_{i,7}\| \end{aligned} \tag{4.6}$$

Mientras que el factor de forma es encontrado al utilizar la Ecuación 4.7. En esta ecuación Per es el perímetro y A el área comprendida dentro del contorno.

$$FF = \frac{Per}{4\pi A} \tag{4.7}$$

Finalmente, la relación entre el eje vertical y horizontal de la boca es hallada al usar la Ecuación 4.8.

$$RHV = \frac{\|p_3 - p_4\|}{p_1 - p_2} \tag{4.8}$$

En la Figura 4.5 se muestra la dinámica de algunas características que pueden ser extraídas teniendo el contorno externo de la boca en una secuencia de video. Se puede observar que la respuesta del factor de forma, la redondez y la relación de los diámetros (ejes vertical y horizontal) en el tiempo es similar, mientras que el área y el perímetro se comportan de manera análoga.

Se decidió hacer uso de los FAPs que definen la deformación de los puntos característicos del contorno externo de la boca como características. Los FAPs miden la deformación de los puntos que definen la cara desde un modelo de la cara en estado neutro, y normalizados según algunas medidas antropométricas. Para definir los FAPs para un modelo de cara arbitrario, MPEG-4 define los FAPUs que sirven para escalar cualquier modelo de cara. Los FAPUs están definidos como fracciones de distancias entre puntos claves de la cara (Tabla 4.2) (Figura 2.6).

Para calcular los FAPs del contorno externo de la boca, se mide el desplazamiento de cada uno

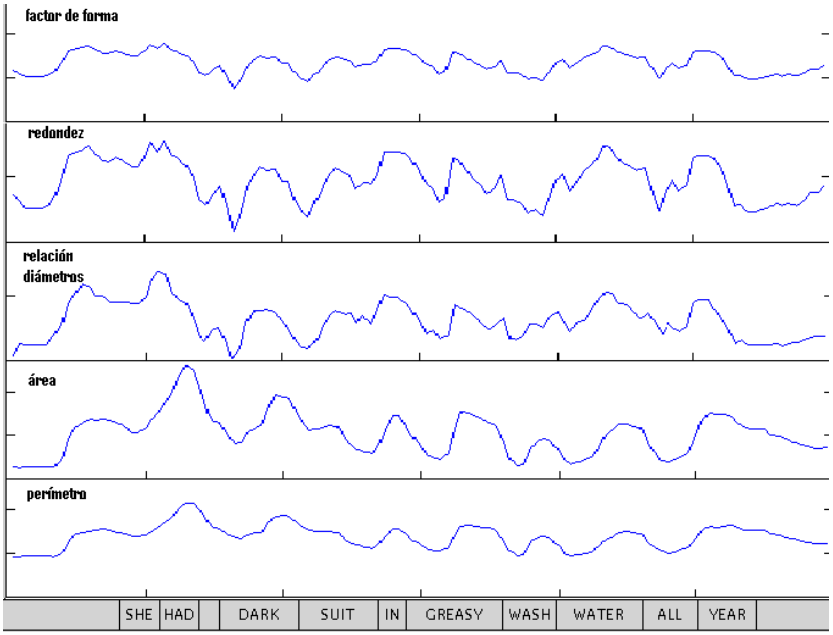


Figura 4.5: Algunas características visuales probadas

de los puntos con respecto a una boca en estado neutro, la cual es seleccionada de los cuadros dentro de la secuencia de video (Ecuación 4.9) (Algoritmo 3). Cabe recalcar que debe haber desplazamientos tanto positivos como negativos para definir las deformaciones de la boca desde un estado neutro. Estos desplazamientos son entonces normalizados respecto al ancho de la boca, el cual es el FAPU (MW0) para los grupos 2 y 8 que describen los contornos interno y externo de la boca.

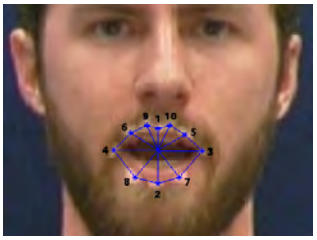
$$\text{FAPs} = \frac{S_{neutro} - S}{MW0} \tag{4.9}$$

También debe tenerse en cuenta que en el caso del audio contenido en la base de datos VidTIMIT, se hizo el análisis en ventanas de 20 ms con traslape del 50%, es decir que la frecuencia de observación es de 100 Hz, el video de la base de datos tiene 25 cuadros por segundo (25Hz), y por lo tanto debe hacerse interpolación del vector de características visuales para obtener una observación cada 10 ms (100Hz) (Algoritmo 3).

Algoritmo 3 Caracteristicas

Entradas: La secuencia de puntos del contorno externo de la boca para todos los cuadros de video S_{nx10}.

Salida: 4 caracteríticas de forma de los labios y las dos primeras derivadas temporales C_{mx12}



[***Paso 1:***] Cargar los 10 puntos característicos de la imagen con la boca en estado neutro S_{neutro}

[***Paso 2:***] Hallar el FAPU que se usa para la boca:
$MW0 = \|S_{neutro,4} - S_{neutro,4}\|$

[***Paso 3:***] Hallar los FAPs como la diferencia entre los puntos del cuadro actual con los de la boca en estado neutro y normalizarlos con $MW0$
$FAPs = \frac{Si - S_n}{MW0}$

[***Paso 4:***] Hacer PCA de los 10 FAPs de la boca y seleccionar los 3 mayores componentes principales $FAPs'$

[***Paso 5:***] Rotar y trasladar los puntos del cuadro actual S_i, de forma tal que los puntos $S_{i,3}$ y $S_{i,4}$ formen el eje x, y los puntos $S_{i,1}$ y $S_{i,2}$ el eje y.

[***Paso 6:***] Calcular la redondez usando la Ecuación 4.5

[***Paso 7:***] Hacer el vector de características como:
$C_i = [R\ FAPs']$

[***Paso 8:***] Calcular las primeras dos derivadas temporales usando la aproximación de diferencia central:
$C_i = \frac{C_{i+1} - C_{i-1}}{2}$

[***Paso 9:***] Realizar interpolación lineal para hallar C'_1 y C'_n
$C''_i = \frac{C'_{i+1} - C'_{i-1}}{2}$

[***Paso 10:***] Realizar interpolación lineal para hallar C''_1 y C''_n

[***Paso 11:***] Concatenar el vector de características con las derivadas $C_i = [C_i\ C'_i\ C''_i]$.

[***Paso 12:***] Hacer interpolación spline cúbica (de 25 cuadros por segundo a 1 muestra cada 10 mS)

4.1.2 Experimento IV: sistema de reconocimiento visual del habla usando VidTIMIT

Una vez calculado el conjunto de características se puede proceder de la misma forma que en la Sección 3. Primero se probará el sistema usando sólo información visual, y al igual que en el caso cuando se usó audio, se hará primero empleando como unidad básica fonemas, es decir, cada cadena de Markov modelará un fonema diferente del habla inglesa.

El algoritmo de seguimiento y extracción de características de audio presentado en esta sección fue usado sobre la base de datos VidTIMIT. Los modelos fueron entrenados usando la misma metodología que para el audio, pero usando el vector de características visuales que incluye las dos primeras derivadas temporales. El sistema que se empleó para el audio para la base de datos VidTIMIT posee un modelo para el silencio y otro para el ruido, pero las características de video no brindan información para diferenciar el ruido acústico, y por lo tanto sólo se modeló el silencio. Con el fin de usar las herramientas de HTK se debe guardar las características en formato HTK y especificar que la información que contiene está definida por el usuario, lo cual se consigue usando el siguiente archivo de configuración de parámetros, en cuanto a la grámatica y el diccionario se emplearon los mismos que para el caso de audio.

```
# Coding parameters

TARGETKIND = ANON
```

El modelo fue probado al igual que los anteriores, al tratar de alinear los datos de una de las dos frases comunes, pero como se puede ver en la Figura 4.6, las características visuales no ofrecen suficiente información para realizar la alineación de la señal de habla.

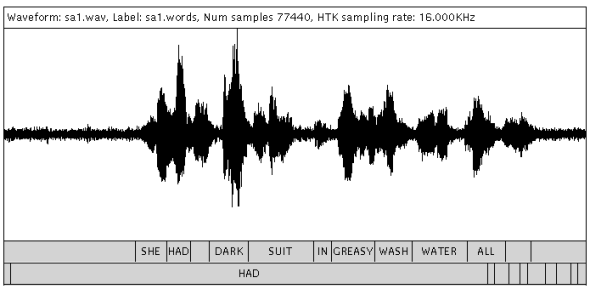


Figura 4.6: Alineación de datos de VidTIMIT usando características visuales

De nuevo se probó el sistema al tratar de reconocer las palabras que se encuentran en las dos frases comunes, pero en este caso, al no tener buenos resultados alineando las frases, se usó el enfoque de palabras aisladas, es decir, se guardó en archivos de audio independientes cada una de las palabras y se usó la siguiente grámatica:

```
$com = SHE | HAD | YOUR | DARK | SUIT | IN | GREASY | WASH |
WATER | ALL | YEAR | DONT | ASK | ME | TO | CARRY | AN | OILY |
RAG | LIKE | THAT;
($com)
```

Aún así, el sistema entrenado con las características de video basado en fonemas, en la mayoría de los casos no clasificó la entrada debido a que se quedaba anclado en un estado, y cuando la clasificó no lo hizo correctamente (Figure 4.7). Se obtuvo un porcetaje de palabras reconocidas correctamente de 14.23%, sin embargo este dato no es real, pues no se tiene en cuenta las entradas que no reconoció.

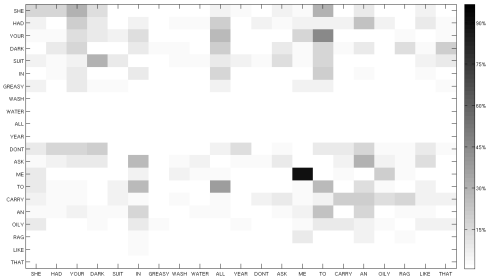


Figura 4.7: Matriz de confusión de palabras claves sobre la base de datos VidTIMIT usando sólo información visual

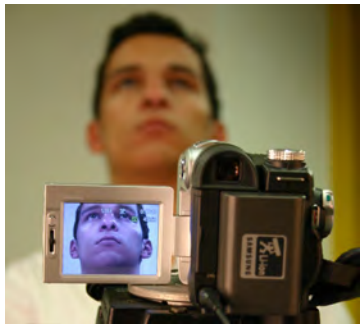
Los resulados no fueron aceptables para el modelo basado en fonemas usando video, debido principalmente, a la insuficiencia de datos, pues la base de datos VidTIMIT sólo posee 430 frases, comparadas con las 6300 de la TIMIT, y a que la información dada por las caracterśticas visuales no es tan rica como las de audio.

4.2 Sistema de reconocimiento de habla basado en palabras usando video

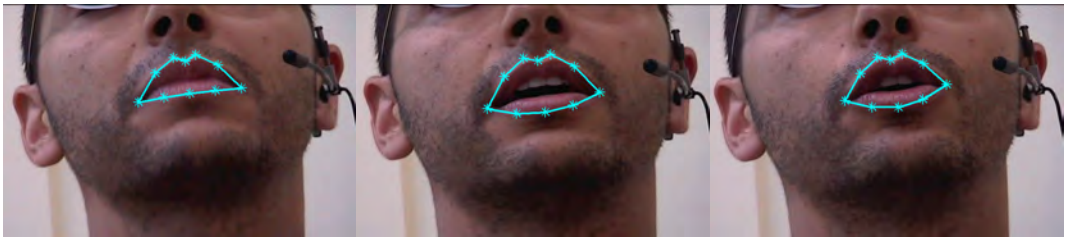
Debido a que los resultados usando como unidades básicas los fonemas del habla inglesa y con la base de datos VidTIMIT no fueron aceptables ni usando audio ni la información visual, además de la necesidad de implementar el sistema de reconocimiento de comandos en francés y en español, se adquieron datos audio-visuales en el laboratorio (Apéndice A).

El primer paso en la construcción del sistema de reconocimiento, es la extracción de características. Se empleó el Algoritmo 1 para realizar el seguimiento del contorno externo de la boca sobre las secuencias de video adquiridas. Para el caso de los datos del laboratorio, se empleó una ventana para el cálculo de la similitud de 11x11 píxeles y un vecindario de búsqueda de 11x11 píxeles. El ancho de

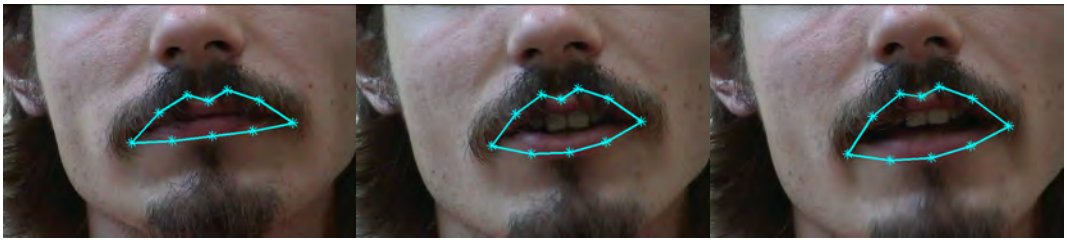
la boca tiene una media de 210.38 píxeles con desviación de 72,46. Los resultados de usar el algoritmo sobre estos datos se muestran en la Figura 4.8 cada 700 cuadros, pues las secuencias de video poseen 2500 cuadros aproximadamente.



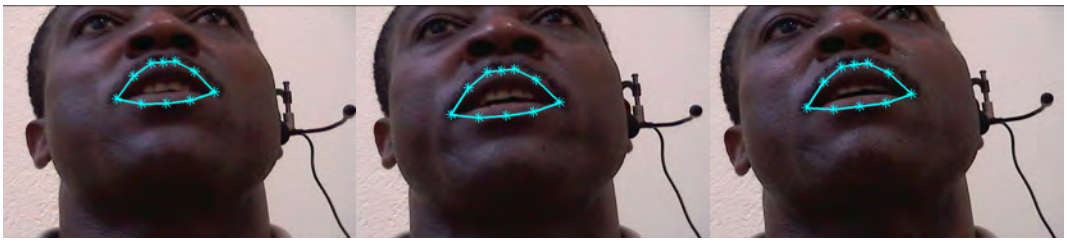
(a) Adquisición



(b) Cuadro 700 (c) Cuadro 1400 (d) Cuadro 2100



(e) Cuadro 700 (f) Cuadro 1400 (g) Cuadro 2100



(h) Cuadro 700 (i) Cuadro 1400 (j) Cuadro 2100

Figura 4.8: Seguimiento del contorno externo de los labios sobre tres secuencias de video adquiridas en el laboratorio. Los resultados son mostrados cada 700 cuadros

4.2.1 Experimento V: sistema de reconocimiento visual del habla usando datos adquiridos en el laboratorio

Una vez extraídas las características de audio sobre las secuencias de video adquiridas en el laboratorio, se procedió a entrenar 50 veces sistemas de reconocimiento de comandos usando el enfoque de palabra aislada, uno para los comandos en habla inglesa, otro para los comandos en habla francesa y otro para los comandos en habla española.

El diccionario define la pronunciación de las palabras en función de los HMMs. En este caso, los HMMs representan las mismas palabras. Por lo tanto, en el caso de habla inglesa se usó el siguiente diccionario:

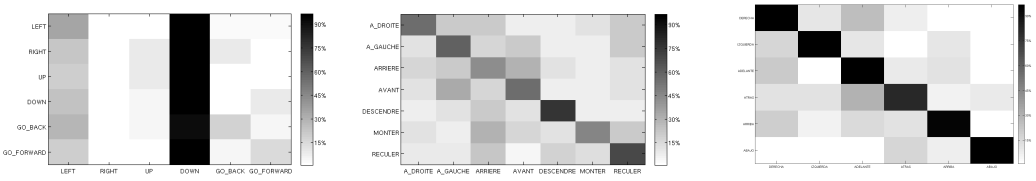
LEFT	LEFT
RIGHT	RIGHT
UP	UP
DOWN	DOWN
GO_BACK	GO_BACK
GO_FORWARD	GO_FORWARD

La gramática fue definida según el enfoque de palabra aislada (para el habla francesa):

```
$com = A_GAUCHE | A_DROITE | MONTER | RECULER | ARRIERE | AVANT |
DESCENDRE;
($com)
```

Cabe recordar que según este enfoque, las palabras están separadas por pausas en lugar de silencios, por lo tanto, las características visuales fueron almacenadas en diferentes archivos, uno para cada muestra de los comandos presentes en los datos.

Los sistemas de reconocimiento de comandos en inglés, en francés y en español fueron entrenados usando la metodología expuesta en la Sección 3. Los resultados se aprecian el la Figura 4.9, donde se ve que la información brindada por el video sin tener en cuenta el audio no es suficiente para distinguir entre comandos, en especial en el caso de los comandos en inglés, en el cual casi todas las muestras fueron reconocidas como DOWN, aunque también tiene importancia el hecho de que los datos no fueron tomados de personas que hablen inglés como lengua materna. Por otro lado, los resultados en francés y en especial en español muestran que aunque las características visuales aún no son suficientemente discriminantes para distinguir entre comandos, brindan información importante pues en las matrices de confusión los datos se concentran en la diagonal principal. Los resultados obtenidos con HResults para los sistemas con 10 estados y 20 estados se aprecian en la Tabla 4.3.



(a) inglés (b) francés (c) español

Figura 4.9: Matriz de confusión sobre datos adquiridos en el laboratorio empleando únicamente características visuales y usando enfoque de palabras aisladas

4.2.2 Experimento VI: sistema AVSR usando datos adquiridos en el laboratorio

Finalmente, se empleó el conjunto conformado por la combinación de las características obtenidas del audio y las extraídas al hacer seguimiento de los labios como entrada del sistema. A este enfoque se le conoce como integración temprana (Sección 2). Para implementar un sistema de reconocimiento audio-visual, es necesario realizar algunos ajustes. El conjunto debe ceñirse al formato HTK (Sección 3), es decir, en un archivo binario debe almacenarse cada característica en dos bytes, concatenado las primeras y las segundas derivadas temporales. Para el sistema desarrollado, el conjunto consiste en los primeros 12 coeficientes de Mel y un término de energía como características de audio, mientras que como caraterísticas visuales se usan los 3 componentes principales de los FAPs y la redondez del contorno externo de la boca.

También cabe aclarar que el video adquirido en el laboratorio se encuentra en formato NTSC, por lo tanto la frecuencia de muestreo para el video es de 29.97 cuadros por segundo (aproximadamente 30 Hz), mientras que las características de audio están muestradas a 100 Hz, lo que corresponde a ventanas de 20 ms con traslape de 10 ms. Por lo tanto las características de video fueron interpoladas de 30 Hz a 300 Hz para luego ser submuestreadas a 100 Hz (Algoritmo 3).

En la Figura 4.10 se presenta el diagrama de bloques del sistema de reconocimiento audio-visual del habla utilizado en los experimentos. Se aprecia claramente el modelo de integración temprana y el tratamiento que se le dió tanto al audio como la información visual.

El sistema audio-visual de reconocimiento de comandos fue entrenado mediante el procedimiento del Capítulo 3 usando el diccionario y la grámatica dadas en la Sección 4.2.1 y con el enfoque de reconocimiento de palabras aisladas. En la Figura 4.11 se aprecia la matriz de confusión obtenida al realizar reconocimiento sobre todos los comandos aislados de la base de datos adquirida en el laboratorio. Aunque el rendimiento fue menor que al usar únicamente información de audio (Sección 3.2.1), este sistema es menos sensible al ruido acústico como se verá en la Sección 4.2.3. Aún así, se

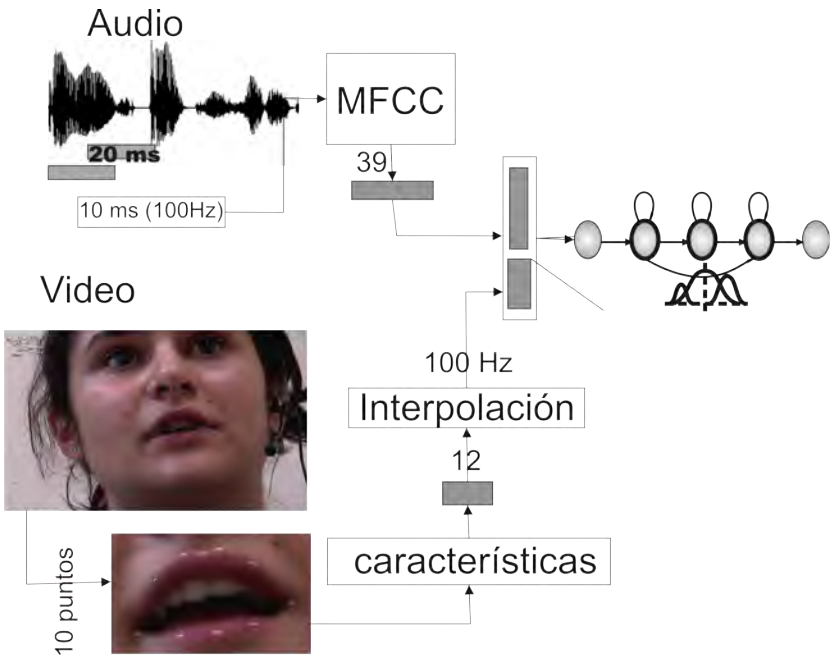
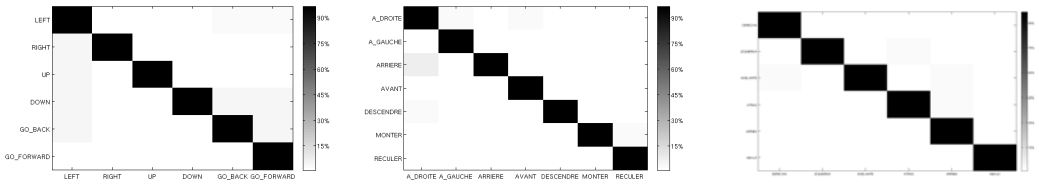


Figura 4.10: Diagrama de bloques del sistema ASVR

puede ver que la respuesta del sistema en inglés, en francés y en español es muy buena comparada con la obtenida para cuando se usan fonemas como unidades básicas.



(a) inglés (b) francés (c) español

Figura 4.11: Matriz de confusión sobre datos adquiridos en el laboratorio usando el enfoque de palabras aisladas para un sistema de reconocimiento de comando audio-visual

4.2.3 Comparación

En la Tabla 4.3 se muestran los porcentajes de palabras correctamente reconocidas obtenidos sobre el 30% de los datos adquiridos en el laboratorio usando la herramienta HResults y haciendo 50 repeticiones para cada idioma. Los sistemas se entrenaron usando sólo las características de audio, sólo las carecterísticas de video y el conjunto conformado por la unión de ambas. El enfoque empleado fue el de reconocimiento de palabras aisladas y usando como unidades básicas las palabras en lugar de

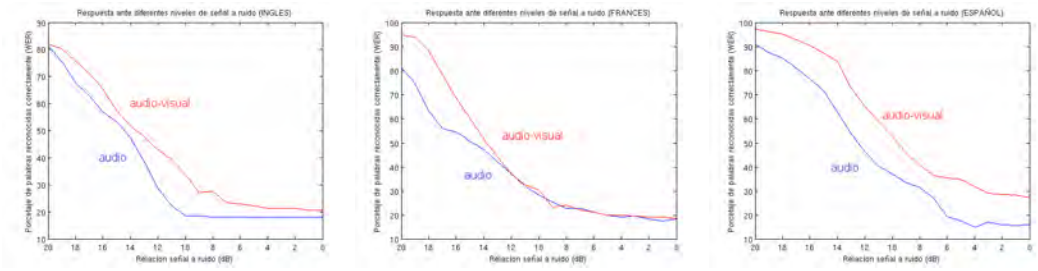
Tabla 4.3: Porcentaje de palabras correctamente reconocidas

		promedio	desviación estándar
inglés	audio	98.46%	1.64%
	visual	14.84%	6.49%
	audio-visual	90.28%	2.69%
francés	audio	98.94%	1.34%
	visual	12.60%	5.06%
	audio-visual	94.01%	2.34%
español	audio	97.54%	1.42%
	visual	38.26%	4.93%
	audio-visual	95.42%	4.93%

fonemas.

Es claro que los mejores resultados se lograron cuando sólo se usó información de audio, pero también se hace necesario probar el sistema para diferentes niveles de ruido acústico, pues como se vió en el Capítulo 3, los resultados dependen fuertemente de las condiciones de adquisición, de la relación señal a ruido, así como del acento de los locutores.

Con el fin de comparar el comportamiento ante la presencia de ruido, se contaminó la información de audio adquirida en el laboratorio con ruido blanco gaussiano. Se hicieron pruebas con niveles de señal a ruido desde 100 dB hasta 1 dB, donde 100 dB corresponde a un ruido 100 veces menor que la señal, 10 dB a un ruido 10 veces menor que la señal, 3 dB a un ruido cuya amplitud corresponde a la mitad de la amplitud de la señal y 1 dB a un ruido de igual amplitud que la señal.



(a) Sistema de reconocimiento de comandos en inglés (b) Sistema de reconocimiento de comandos en francés (c) Sistema de reconocimiento de comandos en español

Figura 4.12: Respuesta del sistema ante el ruido acústico

En la Figura 4.12 se hace evidente que el comportamiento del sistema usando únicamente audio es muy inferior cuando la relación señal a ruido es de 100 dB que cuando no se había agregado ruido. También se observa que tanto para el sistema de reconocimiento de comandos en francés, como en inglés y como en español, el desempeño es superior para todos los niveles de ruido cuando se usa el sistema audio-visual que cuando sólo se emplea el audio. Para el caso del sistema en francés, y debido a que todos los locutores hablan francés con fluidez, la respuesta del sistema audio-visual es mucho

mejor para niveles de ruido bajos. Es más evidente aún en español, donde la respuesta del sistema audio-visual es mucho mejor en todos los casos, debido a que los locutores hablan español como lengua materna.

5 Comunicación con el robot

5.1 Consideraciones

En los capítulos anteriores se probaron distintas configuraciones de sistemas ASR. Se evaluó un sistema ASR usando audio, basado en fonemas y únicamente para el habla inglesa. Se entrenaron los HMMs usando la base de datos TIMIT, la cual cuenta sólo con información acústica. El desempeño de este sistema fue bueno para el caso de alineación de palabras, en donde se conoce con anticipación cuáles son las palabras que aparecen en la secuencia y en qué orden, reconociendo claramente dos frases distintas, hallando los límites en el tiempo de cada una de las palabras (Sección 3.1.1, Figuras 3.4 y 3.5). Sin embargo, cuando el objetivo fue reconocer dentro de un conjunto de palabras definido, cuáles estaban presentes en la muestra, es decir, usando el enfoque de reconocimiento de palabras claves, el porcentaje de palabras correctamente reconocidas descendió al 60 % sobre los datos de la base de entrenamiento (Figura 3.8), y al 43 % sobre datos adquiridos en el laboratorio (Figura 3.8). Con el fin de hacer reconocimiento de algunos comandos que podrían ser usados para controlar un robot, se empleó el enfoque de reconocimiento de palabras aisladas, en el cual las palabras están separadas por pausas y no por silencios, es decir, que cada palabra es un archivo de audio distinto o es tratada en el buffer como tal. Usando el enfoque de reconocimiento de palabras aisladas empleando fonemas como unidades básicas, el porcentaje de palabras reconocidas correctamente fue de 44 % (Figura 3.9), mientras que cuando se usaron comandos aislados, es decir, parejas de palabras al especificar reglas gramaticales, el reconocimiento aumento al 69 % (Figura 3.10).

Con el objetivo de agregar información visual del habla, se empleó la base de datos VidTIMIT, la cual es una base de datos audio-visual disponible en internet desde el año 2008, pero que no posee información de etiquetado y tiene alta presencia de ruido acústico. Uno de los mayores desafíos al emplear ésta o cualquier otra base de datos para entrenar un sistema ASR basado en fonemas, es el etiquetado de las señales. Tomó 160 horas de trabajo hombre para etiquetar 80 frases, con la cuales se entrenó un sistema ASR usando sólo el audio, que luego fue utilizado para realizar alineación de las demás frases y así generar la información de etiquetado sobre las 430 secuencias de video de la base de datos. El desempeño de este sistema al emplearlo para alinear los datos de la base de datos, así como los

Tabla 5.1: Caraterísticas de duración de los comandos en inglés empleados

	Left	Right	Up	Down	Go back	Go forward
media (s)	0.57	0.55	0.44	0.55	0.66	0.78
desviación estándar (s)	0.14	0.11	0.14	0.10	0.14	0.11

Tabla 5.2: Caraterísticas de duración de los comandos en francés empleados

	À droite	À gauche	Arrière	Avant	Descendre	Monter	Reculer
media (s)	0.63	0.62	0.58	0.47	0.62	0.56	0.59
desviación estándar (s)	0.15	0.15	0.14	0.10	0.19	0.12	0.14

datos adquiridos en el laboratorio fue bueno (Figuras 3.11 y 3.13), mientras que cuando se utilizó para hacer reconocimiento de palabras claves en habla continua, el porcentaje de palabras correctamente reconocidas fue de 54% (Figura 3.12) para los datos de la base de datos VidTIMIT y del 28% (Figura 3.14) para las secuencias adquiridas en el laboratorio. Usando el enfoque de palabras aisladas para reconocer algunos comandos, empleando como unidades básicas fonemas, el desempeño fue del 19 %, y empleando reglas gramaticales para reconocer comandos aislados, se obtuvo un porcentaje de palabras reconocidas correctamente del 26 % (Figura 3.15).

Con el fin de utilizar información visual del habla, se emplearon los tres primeros componentes principales de los FAPs (Facial Definition Parameters) del contorno externo de la boca, definidos en el estándar MPEG-4, y la redondez del área comprendida por el contorno externo. Para incluir información dinámica, se extrajeron las primeras dos derivadas temporales de las características visuales.

Se intentó reproducir el mismo procedimiento usando sólo las características visuales con la base de datos VidTIMIT, pero los resultados no fueron aceptables (Figura 4.6 y Figura 4.7).

Debido a que el interés principal de este trabajo es implementar un sistema audio-visual de reconocimiento del habla, y a que los resultados usando fonemas como unidades básicas no fueron aceptables, se adquirió una base de datos en el laboratorio con comandos en inglés, francés y español. La duración de los comandos en inglés adquiridos puede verse en la Tabla 5.1, la de los comandos en francés en la Tabla 5.2, mientras que la de los comandos en español en la Tabla 5.3. Puede apreciarse que la pronunciación de los comandos por los locutores es altamente variable, y que además, los comandos en cada idioma tienen duraciones similares.

Tabla 5.3: Caraterísticas de duración de los comandos en español empleados

	Derecha	Izquierda	Adelante	Atrás	Arriba	Abajo
media (s)	0.95	1.05	1.11	0.96	0.90	0.96
desviación estándar (s)	0.19	0.19	0.24	0.18	0.25	0.27

Se realizaron los mismos experimentos, pero esta vez usando palabras como unidades básicas, es decir que cada cadena oculta de Markov representa una palabra. Siendo así, se probaron distintas configuraciones para cada uno de los sistemas, encontrando los mejores resultados para 10 estados y una gaussiana por estado para los comandos de habla inglesa, y 20 estados y una gaussiana por estado para los comandos tanto de habla francesa como española (Figura 3.17). Los resultados usando únicamente los MFCCs y empleando el enfoque de reconocimiento de palabras aisladas, se acercaron al 100 % para los tres idiomas (Figura 3.18).

Emplear únicamente las características que describen la forma de la boca para el reconocimiento no es suficiente (Figura 4.9). Más aún, el desempeño del los tres sistemas cae bruscamente cuando se usan tanto las características de audio como las de video, si se compara con el comportamiento cuando se emplea sólo información acústica (Tabla 4.3), pero cuando hay presencia de ruido acústico se hace evidente la ventaja de usar información visual. De hecho, cuando el ruido es apenas una centésima parte de la señal, el porcentaje de palabras correctamente reconocidas por el sistema ASR usando sólo información acústica cae en alrededor 15 %, y el sistema audio-visual de reconocimiento presenta mejores resultados para todos los niveles de ruido, en especial en la lengua española, teniéndose en cuenta que todas las muestras en este idioma fueron adquiridas de locutores que hablan español como lengua materna (Figura 4.11).

Por lo tanto, los mejores resultados al enfrentarse al desafío de reconocimiento automático de comandos, se hallaron al emplear palabras como unidades básicas y el enfoque de reconocimiento de palabras aisladas. Además, con el fin de hacer el sistema robusto ante la presencia de ruido, se debe utilizar información audio-visual del habla.

En la Tabla 5.4 se presentan los resultados del reconocimiento empleando este enfoque para algunos comandos del habla inglesa, las palabras pronunciadas se encuentran en cada fila y son reconocidas por el sistema según cada columna. Se aprecia que el sistema confunde casi todas las palabras algunas vez con “left”, y que el reconocimiento no depende de la duración de la palabras. Por ejemplo, confundió “left” con “go forward”. Aun así la tasa es muy alta, y para todos los comandos es superior al 90 %. Los resultados son mejores para el sistema de reconocimiento de comandos del habla francesa (Tabla 5.5) y española (Tabla 5.6).

En este capítulo se describe el procedimiento realizado con el fin de emplear los sistemas ASR entrenados para controlar un robot. El procedimiento fue llevado a cabo bajo el lenguaje de programación C++ usando QT versión 4 en sistema operativo linux (Open Suse 10), como interfaz física con el robot se utilizó el hardware de fuente abierta Freeduino (Apéndice B).

Tabla 5.4: Porcentaje de palabras reconocidas con el enfoque de palabras aisladas para seis comandos del habla inglesa, usando información audio-visual

	Left	Right	Up	Down	Go back	Go forward
Left	94%	0%	0%	0%	3%	3%
Right	3%	97 %	0%	0%	0%	0%
Up	3%	0%	97%	0%	0%	0%
Down	3%	0%	0%	91%	3%	3%
Go back	3%	0%	0	0%	94%	3%
Go forward	0%	0%	0%	0%	0%	100%

Tabla 5.5: Porcentaje de palabras reconocidas con el enfoque de palabras aisladas para siete comandos del habla francesa, usando información audio-visual

	À droite	À gauche	Arrière	Avant	Descendre	Monter	Reculer
À droite	94%	0%	0%	6%	0%	0%	0%
À gauche	0%	100%	0%	0%	0%	0%	0%
Arrière	6%	0%	94%	0%	0%	0%	0%
Avant	0%	0%	0%	100%	0%	0%	0%
Descendre	3%	0%	0%	0%	97%	0%	0%
Monter	0%	0%	0%	3%	0%	94%	3%
Reculer	0%	0%	0%	0%	0%	0%	100%

Tabla 5.6: Porcentaje de palabras reconocidas con el enfoque de palabras aisladas para seis comandos del habla española, usando información audio-visual

	Derecha	Izquierda	Adelante	Atrás	Arriba	Abajo
Derecha	100%	0%	0%	0%	0%	0%
Izquierda	0%	98 %	0%	2%	0%	0%
Adelante	2%	0%	96%	0%	0%	2%
Atrás	0%	0%	0%	98%	2%	0%
Arriba	0%	0%	0	0% %	100%	0%
Abajo	0%	0%	0%	0%	0%	100%

5.2 Implementación

Al emplear el enfoque de reconocimiento de palabras aisladas, se hace necesario realizar la segmentación de cada palabra en la secuencia de audio. Con este fin se normalizó la señal de audio de entrada y se utilizó la relación señal a ruido como criterio de decisión. Para el análisis de la señal, se empleó una ventana de 5000 muestras, que a razón de 32000 muestras por segundo, equivale a una ventana de 16 ms aproximádamente. La segmentación a nivel de palabras es mostrada con más detalle en el Algoritmo 4.

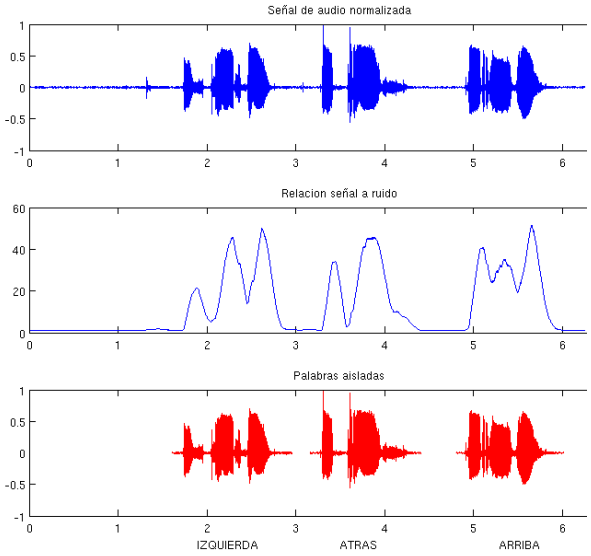


Figura 5.1: Segmentación de la señal de audio en palabras aisladas

En el Algoritmo 4 se asume que el locutor no está hablando en los primeros 16 ms de la secuencia de video, y se hace el análisis para ventanas de 16ms. Además, las características visuales deben ser extraídas con anticipación, pues en este trabajo no se resolvió el problema de seguimiento del contorno externo de la boca ni la extracción de características visuales en tiempo real. El resultado de usar el Algoritmo 4 sobre las primeras tres palabras de una señal de audio (200000 muestras) adquirida en el laboratorio es mostrado en la Figura 5.1, donde se aprecia la efectividad del método para segmentar palabras aisladas.

Una vez segmentada la señal de audio, las características acústicas y visuales combinadas son usadas como entrada al sistema de reconocimiento audio-visual del habla mostrado en la Sección 4.2.2. Este procedimiento se implementó para el sistema de reconocimiento de comandos en inglés, en francés y en español. El sistema en español fue probado con 5 sujetos, tres mujeres y dos hombres, pronunciando una secuencia de 18 comandos separados entre sí por más de un segundo y medio. Dicha secuencia es:

Izquierda, atrás, arriba, abajo, derecha, adelante, derecha, izquierda, abajo, atrás, adelante, arriba,

Algoritmo 4 Segmentación audio

Entradas: Señal de audio A.
Características visuales de la secuencia de video C.
Salida: Características audio-visuales de las palabras aisladas P
[***Paso 1:***] Calcular la energía del ruido como:
$E_R = \sum_{i=1}^{5000} |A_i|$
[***Paso 2:***]
para todas las muestras de la señal de audio A_i **hacer**
Calcular la energía de la señal como:
$E_S = \sum_{j=i}^{i+5000} |A_j|$
Calcular la relación señal a ruido como:
$SR_i = \frac{E_R}{E_S}$
fin para
[***Paso 3:***] La secuencia de muestras B es un candidato a palabra aislada cuando la relación señal a ruido sea mayor que 6.
B sigue siendo candidato mientras al menos 5000 muestras (16 ms) seguidas no tengan una relación señal a ruido menor a 6.
B es una palabra aislada si posee más de 5000 muestras cuya relación señal a ruido es mayor a 6.
[***Paso 4:***] B es normalizada y se extraen las características de audio para combinarlas en P con las características visuales dadas por C en el mismo intevalo de tiempo.

derecha, adelante, abajo, izquierda, atrás, arriba.

El desempeño fue del 100 %, el sistema segmentó y reconoció satisfactoriamente los 18 comandos en las 5 secuencias como se aprecia en la Figura 5.2, donde se muestra la matriz de confusión del sistema sobre los 90 comandos.

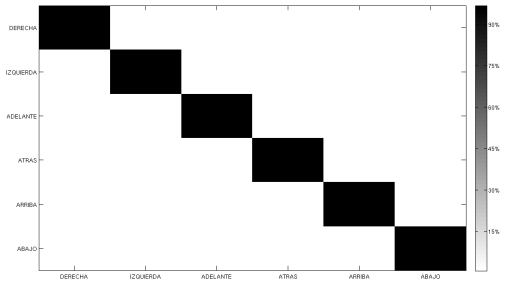


Figura 5.2: Matriz de confusión del sistema AVSR sobre 5 secuencias de 18 comandos

La decisión tomada por el sistema AVSR es enviada a través de puerto USB al sistema de desarrollo freeduino. La plataforma de desarrollo es usada como interfaz con el robot, codificando la palabra reconocida en uno de los puertos digitales de salida, en el cual se coloca sólo un bit en alto a la vez para cada uno de los 6 comandos.

6 Conclusiones y trabajo futuro

En este trabajo se presentaron diferentes enfoques para realizar reconocimiento del habla con el fin de controlar un robot, en especial el robot laparoscópico da Vinci. Debido a que los modelos ocultos de Markov son la técnica que mejores resultados ha presentado para el reconocimiento automático del habla, se utilizó Hidden Markov Model Toolkit como herramienta computacional.

Las cadenas ocultas de Markov modelan las unidades básicas del habla. Se entrenaron sistemas ASR usando tanto palabras como fonemas como unidades. Uno de los principales desafíos para entrenar un sistema ASR basado en fonemas, es generar la información de etiquetado sobre la base de datos, debido a la tediosidad de la tarea. Durante el desarrollo de este trabajo, se etiquetó a nivel de fonemas del habla inglesa las 430 secuencias de video de la base de datos VidTIMIT, la cual se encuentra disponible en internet.

Se exploró el reconocimiento del habla usando fonemas y palabras como unidades básicas, el enfoque de reconocimiento de palabras claves en habla continua y el de palabras aisladas, y empleando una gaussiana y una mezcla de tres gaussianas para modelar la probabilidad de observación de cada estado. Usando diferentes configuraciones, se mostró que la mejor forma de enfrentarse al desafío de reconocimiento de comandos para controlar un robot, es emplear el enfoque de reconocimiento de palabras aisladas usando palabras como unidades básicas, utilizando una sola función gaussiana por estado y utilizando 10 estados para los comandos en habla inglesa, y 20 estados tanto para los comandos de habla francesa como española (Figura 3.17). Con esta configuración, y usando sólo información de audio, el porcentaje de palabras correctamente reconocidas para los comandos en inglés, en francés y en español fue cercana el 100 % (Sección 3.2.1) (Figura 3.18). Usando información audio-visual y la misma configuración, el desempeño descendió alrededor del 93 % (Tabla 4.3, Figura 4.11).

Aunque al parecer el desempeño del sistema de reconocimiento utilizando características audio-visuales es inferior al sistema que sólo emplea información acústica, hay que destacar que el comportamiento depende fuertemente de la condiciones de adquisición, de la relación señal a ruido, así como del acento de los locutores (Sección 3.1.2). Aunque al realizar el análisis sin ruido no fue evidente, se mostró que el enfoque de reconocimiento audio-visual del habla tiene un desempeño superior cuando hay presencia de ruido acústico en la señal de entrada (Figura 4.12). De hecho, el porcentaje de palabras correctamente reconocidas por el sistema ASR usando sólo los MFCCs cae en alrededor 15 %,

cuando se añade ruido gaussiano a la señal de audio, haciendo que la relación señal a ruido sea de 20 dB (la amplitud de la señal es 100 veces la del ruido). El sistema audio-visual de reconocimiento presenta mejores resultados para todos los niveles de ruido, en especial para los comandos en español, los cuales fueron adquiridos de locutores que hablan este idioma como lengua materna (Figura 4.11).

Es muy importante ls selección del mejor modelo para el sistema de reconocimiento de habla según la aplicación, también lo es la selección del conjunto de características. En este trabajo se emplearon como características de audio los MFCC, que son los más empleados como método de parametrización de la señal acústica en las últimas décadas, mientras que para extraer la infomación visual, se emplearon características de alto nivel, basadas en la forma de la boca. Se exploraron características basadas en los parámetros de animación facial (FAPs), definidos en el estándar MPEG-4, específicamente aquéllas basadas en el contorno externo de la boca, pues se ha mostrado que el contorno interno no influye de manera significativa en el reconocimiento del habla [22]. También se exploraron otras características basadas en la geometría de la boca (Figura 4.5). Se seleccionaron 4 características: los tres primeros componentes principales de los 10 FAPs que definen en contorno externo de la boca, los cuales representan los tres ejes de movimiento de los labios, y la redondez.

El mayor desafío para extraer características de alto nivel en secuencias de video es el seguimiento de la boca. En este trabajo se presentó un modelo basado en restricciones morfológicas y en una medida de similitud a nivel píxeles para el seguimiento del contorno externo de la boca en imágenes a color. La propuesta mostró ser robusta ante la presencia de barba y el tono de piel, e incluso realizó el seguimiento tanto en imágenes con buena definición adquiridas en el laboratorio, como en imágenes con menor definición presentes en la base de datos VidTIMIT. También mostró ser fuerte ante cambios de iluminación y enfoque, pues no hubo control de iluminación y la cámara tenía autoenfoque (Sección 4.1.1).

El estudio de las características visuales se ha concentrado en la descripción de la forma de la boca y en realizar operaciones lineales y no lineales sobre la región de interés. Aunque no existen muchos trabajos que desarrollen una comparación rigurosa, es razonable pensar que el mejor enfoque es realizar operaciones a nivel de píxeles, ya que además de haber demostrado tener un desempeño similar a aquellas características basadas en la forma de los labios [54], son computacionalmente más viables en aplicaciones de reconocimiento del habla en tiempo real. Las características basadas en la forma son útiles para la síntesis visual del habla, el reconocimiento de emociones y la identificación de personas.

Con el fin de hacer automático el algoritmo de seguimiento, se debe resolver el problema de segmentación de los labios. En el grupo de trabajo académico PCI Se ha venido trabajando en segmentación robusta de la boca usando componentes de color y se ha pensado en usar características de textura con éste fin. De hecho, si desea implementarse el sistema en tiempo real, como características

visuales del habla se debe usar características de bajo nivel, basadas en la apariencia de la boca, debido a que no debe realizarse un procedimiento de seguimiento preciso de la boca.

Se ha mostrado que los resultados dependen fuertemente del acento de las personas cuyas secuencias de audio y video hacen parte de la base de datos de entrenamiento. Para los sistemas de reconocimiento de comandos en habla inglesa y francesa, se hace necesario la adquisición de una base de datos con locutores cuya lengua materna sea inglés y francés respectivamente.

Usando las herramientas provistas por HTK, se entrenó un sistema AVSR, el cual fue usado como núcleo de una interfaz hombre-máquina. La herramienta se desarrolló usando lenguaje de programación c++, y la comunicación con el computador se hizó a través de USB. La entrada al sistema es la secuencia de video para la cual se han extraído con anticipación las características visuales. La desición tomada por el sistema AVSR es codificada en una palabra de 6 bits, usando la plataforma de desarrollo Freeduino (Sección 5.2).

A Bases de datos usadas en reconocimiento audio-visual del habla

En los sistemas de reconocimiento de la voz usando audio, las bases de datos varían en el número de palabras (dígitos, algunas palabras, o vocabulario extenso) y en el número de sujetos, dependiendo del modelo de lenguaje que se desee usar.

Para el desarrollo y evaluación de sistemas de reconocimiento de palabras en habla continua, existen bases de datos que poseen la información fonético-acústica necesaria (gran número de palabras etiquetadas fonéticamente). Por ejemplo, la base de datos TIMIT cuenta con 6300 frases, dichas por 630 personas de 8 dialectos de Estados Unidos [73].

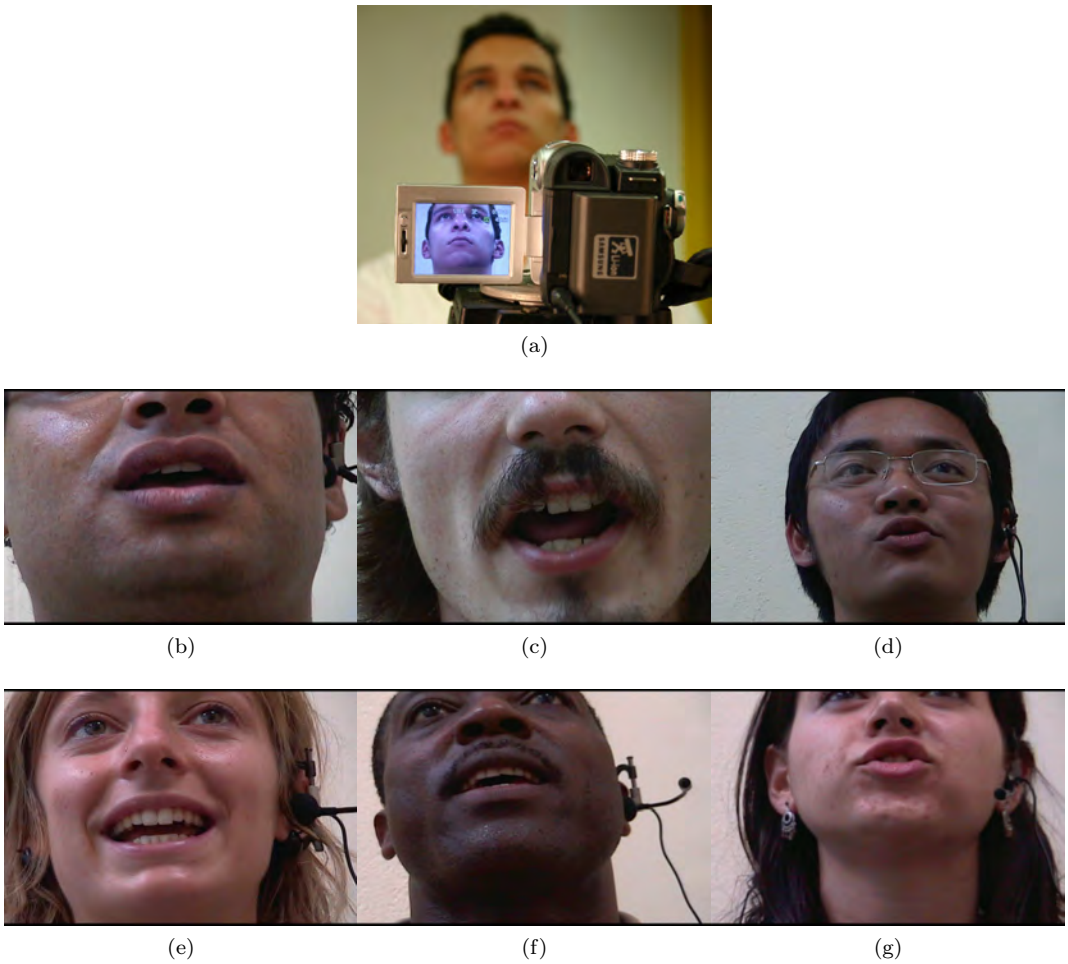
Por otra parte, y debido al interés de la comunidad científica sobre temas como el reconocimiento automático del habla usando características visuales y el reconocimiento de personas usando el rostro, se han desarrollado varias bases de datos que combinan información de video y de audio. Estas bases de datos se crearon, en general, dentro del marco de un proyecto de investigación con objetivos específicos. A causa, principalmente, del costo computacional de almacenamiento del video, estas bases de datos no poseen tantos sujetos como aquellas de sólo audio, o no poseen suficientes palabras para considerarlas fonéticamente balanceadas.

Una de las bases de datos con la que se implementó el sistema es la VidTIMIT, la cual posee información de audio y de video de 430 frases de 43 sujetos (hombres y mujeres) y está disponible en Internet [75]. Los datos de audio presentan un alto nivel de ruido, las imágenes se encuentran en formato jpg y las frases no están etiquetadas, ni por fonemas ni por palabras.

Algunas de las bases de datos que se han usado en sistemas similares son: AVOZES [79], BANCA [80], DAVID [81], M2VTS [82], XM2VTS [83] y CUAVE [84]. De estas, la única que puede ser interesante para el proyecto, ya que se encuentra etiquetada, posee pronunciaciones de números y de frases, el ambiente de grabación presenta condiciones similares, está hecha para propósitos de reconocimiento del habla y ha superado varios inconvenientes de sus predecesoras, es CUAVE. Ha habido inconvenientes en adquirir CUAVE, pues aunque se es libre para propósitos de investigación, la dirección de correo electrónico de contacto parece no estar en uso. Por otro lado, se ha buscado una

base de datos similar para desarrollar el sistema con palabras en francés. Se ha acudido al "European Languaje Resources Association - ELRA", entidad encargada del estudio y preservación de los lenguajes europeos, y que distribuye a través del ELDA (European Languaje Distribution Agency) las bases de datos usadas en ingeniería del lenguaje. No se ha conseguido el propósito, pues ellos no disponen de alguna base de datos audio visual para desarrollo de sistemas de reconocimiento del habla en francés.

A.1 Base de datos propia



(a) (b) (c) (d) (e) (f) (g)

Figura A.1: Datos adquiridos en el laboratorio

Desde sus orígenes, la investigación en procesamiento del lenguaje natural ha estado fuertemente guiada por una tendencia hacia el tratamiento de la lengua inglesa, arrastrada por las tendencias internacionales y por la falta de recursos en otras lenguas.

Para el análisis de los resultados del sistema de reconocimiento del habla usando audio, se hizo

necesario realizar una pequeña base de datos de palabras en inglés (ya que tanto TIMIT como VidTIMIT se encuentran en ese idioma), con palabras incluidas en los datos de entrenamiento y aquellas definidas para el control del robot. De nuevo, esta base de datos debió ser convertida en vectores de características, conservando la frecuencia de muestreo y normalizando la magnitud. También, se etiquetaron las frases en el tiempo a nivel de palabras. La base de datos consiste de 2 frases que son comunes para todos los sujetos tanto en la TIMIT como en la VidTIMIT y de palabras claves en habla continua que podrían controlar algunos grados de libertad de un robot. Cabe recalcar que el lenguaje materno de los sujetos no es inglés. En la base de datos adquirida se incluyen las dos frases: "She had your dark suit in greasy wash water all year" y "Don't ask me to carry an oily rag like that", y los comandos "move right", "move left", "move up", "move down", "zoom in", "zoom out", "start camera" y "stop camera".

También implementó otra base de datos audio-visual para ser empleada con el enfoque de reconocimiento de palabras aisladas. La base de datos consiste en 4 frases comunes (dos en inglés y dos en francés) y algunas palabras claves, tanto en inglés como en francés y en español. La información de video se encuentra en formato NTSC, con una frecuencia de muestreo de 29.97 fotogramas por segundo y con cuadros de 720X480 píxeles, mientras que el audio está muestreado a 32 KHz. Para los comandos en inglés y en francés se grabaron 18 sujetos que no poseen un dialecto en común, pues son de diferentes partes del mundo, el conjunto está compuesto de 4 mujeres y 12 hombres de diferentes zonas geográficas (Francia, paises árabes, Vietnam, Nigeria, México y Colombia). En cuanto a los comandos en español se grabaron 18 personas, 5 mujeres y 13 hombres de diferentes regiones de Colombia. La base de datos fue etiquetada a nivel de palabras, y las características visuales y de audio fueron extraídas. (Figura A.1).

Esta nueva base de datos incluye las mismas dos frases en inglés: "She had your dark suit in greasy wash water all year". "Don't ask me to carry an oily rag like that".

Dos frases en francés: "Je ne sais pas quelle será l'arme utilisée pour la troisiéme guerre mondiale, mais la quatriéme se ferá a cote du bâton et du gourdin". "Les ordinateurs ne servent à rien, ils ne peuvent doner que de reponses".

Comandos en inglés:

"Left".

"Right".

"Up".

"Down".

"Go back".

"Go forward".

Comandos en francés:

“À gauche”.

“À droite”.

“Monter”.

“Reculer”.

“Arriere”.

“Avant”.

“Descendre”.

Y comandos en español:

“Izquierda”.

“Derecha”.

“Arriba”.

“Abajo”.

“Atrás”.

“Adelante”.

Freeduino

B

Arduino es plataforma abierta para computación física basada en una tarjeta de entrada y salida muy sencilla. Puede ser utilizada para crear objetos interactivos independientes o conectados al computador.

La tarjeta Freeduino es un versión especial de la tarjeta Arduino diseñada para permitir un ensamblado fácil utilizando partes insertables (exceptuando el chip FT232L de comunicación USB), la tarjeta fue diseñada por Bill Westfield del equipo de Freeduino (Figura B.1).

Figura B.1: Plataforma abierta de desarrollo Freeduino

Freeduino es un proyecto de fuente abierta cuyo fin es replicar y publicar archivos de hardware compatibles con Arduino, permitiendo a los usuarios crear tarjetas funcional, eléctrica y físicamente compatibles con Arduino.

Miestras Arduino es una marca protegida por derechos de autor, Freeduino tiene licencia libre y no restringida, lo que significa que puede ser usada en cualquier aplicación y que se pueden modificar los archivos de hardware sin necesidad de crear una versión propia de la tarjeta Freeduino.

Bibliografía

[1] R. Campbell, "Audio-visual speech processing," *Elsevier*, pp. 562–569, 2006. 2, 3, 4

[2] ——, "The processing of audio-visual speech: empirical and neural bases," *Philosophical Transactions of The Royal Society B*, no. 363, p. 1001–1010, 2008. 2, 3, 4

[3] M. Grimm and K. Kroschel, *Robust Speech Recognition and Understanding*, 2007. 3, 8

[4] K. Homayounfar, "Rate adaptive speech coding for universal multimedia access," *IEEE Signal Processing Magazine*, p. 30–39, 2003. 3

[5] D. O. Shaughnessy, "Interacting with computers by voice: Automatic speech recognition and synthesis," *ICONIP 2006, Part II, LNCS 4233,*, pp. 489–498, 2006. 3

[6] B. Mellorf, C. Baber, and C. Tunley, "Evaluating automatic speech recognition as a component of a multi-input device human-computer interface," *in Proceedings of the International Conference on Spoken Language Processing*, pp. 1668–1671, 1996. 3

[7] T. Kubik and M. Sugisaka, "Use of a cellular phone in mobile robot voice control," in *Proceedings of the 40th SICE Annual Conference International Session Papers SICE 2001*, 2001, pp. 106–111. 3

[8] S. Shamma, "Relevance of auditory cortical representations to speech processing and recognition," p. 5, 2005. 3

[9] B. Juang and T. Chen, "The past, present, and future of speech processing," *IEEE Signal Processing Magazine*, vol. 15, pp. 24–48, 1998. 3, 7, 9, 13

[10] R. Goecke, "Current trends in joint audio-video signal processing: A review," p. 70–73, 2005. 3, 4, 9

[11] J. Hurtado, G. Castellanos, and J. Suarez, "Effective extraction of acoustic features after noise reduction for speech classification," in *In International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, (TCSET'02)*, 2002, pp. 245–248. 3

[12] J. Schroeter, J. Larar, and M. Sondhi, "Speech parameter estimation using a vocal tract/-cord model," in *In IEEE International Conference onAcoustics, Speech, and Signal Processing, (ICASSP '87)*, vol. 12. IEEE, 1987, pp. 308–311. 3

[13] N. Cheng, L. Mabiner, A. Rosenberg, and C. Mooonegal, "Some comparisons among several pitch detection algorithms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '76*, vol. 1. IEEE, 1976, pp. 332– 335. 3

[14] R. Hoory, A. Sagi, S. Shechtman, A. Sorin, Z. Shuang, and R. Bakis, "High quality sinusoidal modeling of wideband speech for the purposes of speech synthesis and modification," in *In IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP'06)*, vol. 1. IEEE, 2006, pp. I877–I880. 3

[15] R. G. Termens, J. O. Lafont, F. G. Portabella, and J. M. Roca, "Síntesis de voz utilizando difonemas: uniones entre vocales," *Procesamiento del lenguaje natural*, no. 21, pp. 69–74, 1997. 3

[16] J. Rothweiler, "Noise-robust 1200-bps voice coding," in *Proceedings of the Tactical Communications Conference: Technology in Transition*, vol. 1, 1992, pp. 65–69. 3, 7

[17] J. Macres, "Real-time implementations and applications of the US federal standard CELP voice coding algorithm," in *Proceedings of the Tactical Communications Conference: Technology in Transition*, vol. 1, 1992, pp. 41–45. 3, 7

[18] L. R. Rabiner, "A tutorial on hidden markov models and selected application in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. 3, 4, 13, 15, 16

[19] S. Anderson and D. Kewley-Port, "Evaluation of speech recognizers for speech training applications," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 229–241, 1995. 4, 15

[20] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, 1993. 4, 13, 15

[21] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, pp. 1–15, 2002. 4, 5, 8, 9, 13, 16, 17

[22] P. S. Aleksic and A. K. Katsaggelos, "Comparision of MPEG-4 facial animation parameter groups with respect to audio-visual speech recognition performance," *0-7803-9134-9/05*, vol. 7, no. 6, 2005 IEEE. xii, 4, 5, 8, 9, 17, 41, 66

[23] J. Kratt, F. Metze, R. Stiefelhagen, and A. Waibel, "Large vocabulary audio-visual speech recognition using the janus speech recognition toolkit," *DAGM 2004, LNCS 3175*, pp. 488–495, 2004. 4, 5, 8, 9

[24] R. Mersereau, X. Zhang, and M. Clements, "Audio-visual speech recognition by speechreading," *The 10th IEEE Digital Signal Processing (DSP) Workshop*, p. 1069–1072, 2002. 4, 11

[25] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, p. 141–151, 2000. 4, 11

[26] G. Potamianos, "Speech recognition, audio-visual," *Elsevier*, pp. 800–805, 2006. xii, 4, 9

[27] K. Myung, R. Joung, and K. Eun, *Speech Recognition with Multi-modal Features Based on Neural Networks*, ser. Memoria Técnica del Proyecto, 2004. 4, 9, 11, 17

[28] T. A. Stephenson, J. Escofet, M. Magimai-Doss, and H. Bourlard, "Dynamic bayesian network based speech recognition with pitch and energy as auxiliary variables," *IEEE 0-7803-7616-1/02*, pp. 637–646, 2002. 4, 9

[29] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An introduction to natural lanjjaguage processing, computational linguistics, and speech recognition.*, 2007. 5

[30] ISO/IEC, "Information technology-generic coding of audio-visual objects, part 2: Visual, ISO/IEC FDIS 14496-2 (final drafts international standard), ISO/IEC JTC1/SC29/WG11 N2502," 1998. xii, 5, 12

[31] I. S. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation: The Standard, Implementation and Applications.* Wiley, 2002. 5, 41

[32] A. Salazar, J. Hernández, and F. Prieto, "Automatic quantitative mouth shape analysis," *Computer Analysis of Images and Patterns*, pp. 416–423, 2007. 5, 10

[33] J. B. Gómez, F. Prieto, and T. Redarce, "Lips movement segmentation and features extraction in real time," *Innovative Algorithms and Techniques in Automation, Industrial Electronics and Telecommunications*, pp. 205–210, 2007. xii, 6, 11

[34] J. E. Hernández, F. Prieto, and T. Redarce, "Real-time robot manipulation using mouth gestures in facial video sequences," *Advances in Brain, Vision, and Artificial Intelligence*, pp. 224–233, 2007. xii, 6, 11

[35] U.S. National Library of Medicine and National Institutes of Health, "Medlineplus," *http://medlineplus.gov/spanish/*. xii, 7

[36] L. G. Martínez, "Ecualización de histogramas en el procesado robusto de voz," Ph.D. dissertation, Universidad de Granada, 2007. 6, 8

[37] C. Bao, "Harmonic excitation LPC (HE-LPC) speech coding at 2.3 kb/s," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03)*, vol. 1, 2003, pp. I784–I787. 7

[38] T. Tremain, "The government standard Linear Predictive Coding Algorithm: LPC-10," *Speech Technology*, pp. 40–49, 1982. 7

[39] Y. Kim., "A framework for parametric singing voice analysis/synthesis," pp. 123–126, 2003. 7

[40] K. Sim and M. Gales, "Discriminative semi-parametric trajectory model for speech recognition," *Computer Speech and Language*, vol. 21, pp. 669–687, 2007. 8

[41] S. Ahn and J. Werterkamp, "Cochlear modeling using a general purpose digital signal processor," in *Proceedings of the IEEE 1990 National Aerospace and Electronics Conference (NAECON 1990)*, vol. 1, 1990, pp. 57–63. 8

[42] M. Hunt and C. Lefebvre, "Speech recognition using a cochlear model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'86)*, vol. 11, 1986, pp. 1979–1982. 8

[43] T. Sreeniva, K. Singh, R. Niederjohn, and J. Heinen, "Spectral estimation properties of non-linear auditory models for noisy signals," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1989. Images of the Twenty-First Century*, vol. 2, 1989, pp. 679–680. 8

[44] K. Doh, J. Jae, K. Jae, and L. Soo, "Feature extraction based on zero-crossings with peak amplitudes forrobust speech recognition in noisy environments," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)*, vol. 1, 1996, pp. 61–64. 8

[45] B. Mak, J. T. Kwok, and S. Ho, "Using kernel PCA to improve eigenvoice speaker adaptation," *Proceedings of the Third Intemational Conference on Machine Learning and Cybemetics*, pp. 3062–3067, 2004. 9

[46] J. Luettin and S. Dupont, "Continuous audio-visual speech recognition," in *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume II*. Springer-Verlag, 1998, pp. 657–673. 9, 11

[47] J. Huang, G. Potamianos, J. Connell, and C. Neti, "Audio-visual speech recognition using an infrared headset," *Speech Communication*, vol. 44, pp. 83–96, 2004. 9, 11

[48] T. A. Faruquie, A. Majumdar, N. Rajput, and L. V. Subramaniam, "Large vocabulary audio-visual speech recognition using active shape models," *15th International Conference on Pattern Recognition (ICPR'00)*, vol. 3, pp. 106–109, 2000. 10

[49] M. N. Kaynak, Q. Zhi, A. D. Cheok, K. Sengupta, Z. Jian, and K. C. Chung, "Analysis of lip geometric features for audio-visual speech recognition," *IEEE Transactions on Systems, Man, and Cybernetics — part A: Systems and Humans*, vol. 34, no. 4, pp. 564–570, 2004. 10

[50] P. Aarabi and B. Mungamuru, "The fusion of visual lip movements and mixed speech signals for robust speech separation," *Information Fusion, Elsevier*, vol. 5, pp. 103–117, 2004. xii, 10

[51] S. Morishima, S. Ogata, K. Murai, and S. Nakamura, "Audio-visual speech translation with automatic lip synchronization and face tracking based on 3-d head model," *Proceedings in IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '02)*, vol. 2, pp. 2117–2120, 2002. 10

[52] S. Tamura, K. Iwano, and S. Purui, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*. Springer, 2005, ch. Chapter 3: A Robust Multimodal Speech Recognition Method using ICAL Flow Analysis, pp. 37–53. 11

[53] T.-L. Pao and W.-Y. Liao, "A motion feature approach for audio-visual recognition," in *48th Midwest Symposium on Circuits and Systems*, vol. 1, 2005, pp. 421–424. 11

[54] P. S. Aleksic and A. K. Katsaggelos, "Comparison of low- and high-level visual features for audio-visual continuous automatic speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing.(ICASSP '04)*, vol. 5. IEEE, 2004, pp. V917–920. 12, 66

[55] J. C. Pasamontes, *Estrategias de incorporación de conocimiento sintáctico y semántico en sistemas de comprensión de habla continua en español*, ser. Estudios de Lingüística Española, 2001. [Online]. Available: http://elies.rediris.es/elies12/ 13, 18

[56] M. K. Brown and L. R. Rabiner, "An adaptive, ordered, graph search technique for dynamic time warping for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-30, 4, pp. 656–667, 1982. 13, 14

[57] G. Fang, W. Gao, Member, and D. Zhao, "Online hierarchical transformation of hidden markov models for speech recognition," *IEEE Transactions on Systems, Man, and Cybernetics - part A: Systems and Humans*, vol. 37, no. 1, pp. 1–9, 2007. 13, 14

[58] O. A. A. Alim, N. Elboghdadly, and N. M. E. Shaar, "HMM / NN hybrids for continuous speech recognition," *Eighteenth National Radio Science Conference*, pp. 509–516, 2001. 15

[59] J. Stadermann and G. Rigoll, "Comparing nn paradigms in hybrid NN/HMM speech recognition using tied posteriors," *Automatic Speech Recognition and Understanding (ASRU)*, pp. 89–93, 2003. 15

[60] K. Pulasinghe, K. Watanabe, K. Izumi, and K. Kiguchi, "Modular fuzzy-neuro controller driven by spoken language commands," *IEEE Transactions on Systems, Man, and Cybernetics - part B: Cybernetics*, vol. 34, 1, pp. 293–302, 2004. 15

[61] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University and Microsoft Corporation, 2000. 15, 19

[62] X. Huang and K.-F. Lee, "On speaker-independent, speaker-dependent, and speaker- adaptive speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 150–157, 1993. 16

[63] K.-Y. Su and C.-H. Lee, "Speech recognition using weighted HMM and subspace projection approaches," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 69–79, 1994. 16

[64] J.-T. Chien, "Online hierarchical transformation of hidden markov models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 656–667, 1999. 16

[65] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," *IEEE International Conference of Acoustics, Speech, Signal Processing*, pp. 165–168, 2001. 16

[66] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," *IEEE International Conference of Acoustics, Speech, Signal Processing*, p. 169–172, 2001. 16, 17

[67] X. Liu, Y. Zhao, X. Pi, L. Liang, and A. V. Nefian, "Audio-visual continuous speech recognition using a coupled hidden markov model," *7th International Conference on Spoken Language Processing (ICSLP-2002)*, pp. 213–216, 2002. 17

[68] Y. Zhang, S. Levinson, and T. Huang, "Speaker independent audio-visual speech recognition," *IEEE International Conference on Multimedia and Expo*, vol. 2, p. 1073–1076, 2000. 17

[69] G. Lv, D. Jiang, R. Zhao, and Y. Hou, "Multi-stream asynchrony modeling for audio-visual speech recognition," *Ninth IEEE International Symposium on Multimedia 2007*, p. 37–44, 2007. 17

[70] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition (AVSR workshop 2000 final report)," Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, Md, USA, Tech. Rep., 2000. [Online]. Available: http://www.clsp.jhu.edu/ws2000/groups/av/speech/ 17

[71] R. C. Aguilar, "Diseño y manipulación de modelos ocultos de markov, utilizando herramientas HTK," *Ingeniare. Revista chilena de ingeniería*, vol. 15, no. 1, pp. 18–26, 2007. 18

[72] N. Moreau, *HTK Basic tutorial*. Technical University of Berlin, 2002. [Online]. Available: http://www.nue.tu-berlin.de/wer/moreau/HTK_basic_tutorial.pdf 19

[73] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, 1993. 21, 68

[74] T. Robinson, *British English Example Pronciation dictionary Version 1.0*. Oxford University Press (OUP) and the Medical research council (MRC), 1997. 23

[75] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing, Elsevier*, vol. 15, no. 4, pp. 449–480, 2004. 34, 68

[76] W. Zhilin, P. Aleksic, and A. Katsaggelos, "Lip tracking for MPEG-4 facial animation," in *In Fourth IEEE International Conference on Multimodal Interfaces Processing*, vol. 1, 2002, p. 293–298. 41

[77] M. Ramos, J. Matas, and J. Kittler, "Statistical chromaticity-based lip tracking with B-splines," in *In ICASSP '97: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 1997, pp. 29–73. 41

[78] J. Zhang, M. Kaynak, A. Cheok, and C. Ko, "Real-time lip tracking for virtual lip implementation in virtual environments and computer games," in *In The 10th IEEE International Conference on Fuzzy Systems Proceedings*, vol. 3, 2001, pp. 1359–1362. 42

[79] R. Göcke, *The Audio-Video Australian-English Speech Data Corpus AVOZES Documentation Version 1.2*. Canberra Laboratory, National ICT Australia, 2004. 68

[80] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Boree, B. Ruiz, , and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *AVBPA2003*, 2003, p. 625–638. 68

[81] C. Chibelushi, S. Gandon, J. Mason, F. Deravi, and D. Johnston, "Design Issues for a Digital Integrated Audio-Visual Database," in *IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication (Digest No: 1996/213)*, 1996, pp. 7/1 – 7/7. 68

[82] K. Messer, J. Matas, and J. Kittler, "Acquisition of a large database for biometric identity verification," in *BIOSIGNAL 98*, 1998, p. 70–72. 68

[83] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," in *In Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, 1999, p. 72–77. 68

[84] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research," in *ICASSP2002*, vol. 2, 2002, p. 2017–2020. 68