Ingeniare. Revista Chilena de Ingeniería
ISSN: 0718-3291
facing@uta.cl
Universidad de Tarapacá
Chile

Ceballos, Alexander; Serna-Morales, Andrés F.; Prieto, Flavio; Gómez, Juan B.; Redarce, Tanneguy
Sistema audiovisual para reconocimiento de comandos
Ingeniare. Revista Chilena de Ingeniería, vol. 19, núm. 2, agosto, 2011, pp. 278-291
Universidad de Tarapacá
Arica, Chile

Ingeniare. Revista chilena de ingeniería, vol. 19 Nº 2, 2011, pp. 278-291

Sistema audiovisual para reconocimiento de comandos

Audiovisual system for recognition of commands

Alexander Ceballos[1] Andrés F. Serna-Morales[1] Flavio Prieto[2]
Juan B. Gómez[1, 3] Tanneguy Redarce[3]

Recibido 17 de mayo de 2010, aceptado 8 de abril de 2011
Received: May 17, 2010 Accepted: April 8, 2011

RESUMEN

Se presenta el desarrollo de un sistema automático de reconocimiento audiovisual del habla enfocado en el reconocimiento de comandos. La representación del audio se realizó mediante los coeficientes cepstrales de Mel y las primeras dos derivadas temporales. Para la caracterización del vídeo se hizo seguimiento automático de características visuales de alto nivel a través de toda la secuencia. Para la inicialización automática del algoritmo se emplearon transformaciones de color y contornos activos con información de flujo del vector gradiente ("GVF snakes") sobre la región labial, mientras que para el seguimiento se usaron medidas de similitud entre vecindarios y restricciones morfológicas definidas en el estándar MPEG-4. Inicialmente, se presenta el diseño del sistema de reconocimiento automático del habla, empleando únicamente información de audio (ASR), mediante Modelos Ocultos de Markov (HMMs) y un enfoque de palabra aislada; posteriormente, se muestra el diseño de los sistemas empleando únicamente características de vídeo (VSR), y empleando características de audio y vídeo combinadas (AVSR). Al final se comparan los resultados de los tres sistemas para una base de datos propia en español y francés, y se muestra la influencia del ruido acústico, mostrando que el sistema de AVSR es más robusto que ASR y VSR.

Palabras clave: Reconocimiento audiovisual del habla, modelo oculto de Markov (HMM), coeficientes de Mel, contorno activo, pseudotono, estándar MPEG-4, puntos FAPs, seguimiento de características.

ABSTRACT

We present the development of an automatic audiovisual speech recognition system focused on the recognition of commands. Signal audio representation was done using Mel cepstral coefficients and their first and second order time derivatives. In order to characterize the video signal, a set of high-level visual features was tracked throughout the sequences. Automatic initialization of the algorithm was performed using color transformations and active contour models based on Gradient Vector Flow (GVF Snakes) on the lip region, whereas visual tracking used similarity measures across neighborhoods and morphological restrictions defined on MPEG-4 standard. First of all, we show the design process for an isolated word audio speech recognition system (ASR) using Hidden Markov Models. Next, we show the design process for a speech recognition system using only video features (VSR,) and both audio and video features combined (AVSR). Finally, we compare the results of the three systems on our database in Spanish and French language, showing that AVSR outperforms AVR and VSR under increased acoustic noise conditions in the sequences.

Keywords: Audiovisual speech recognition, hidden Markov models (HMM), Mel's coefficients, active contours, pseudo tone, MPEG-4 standard, FAP points, tracking features.

[1] Departamento de Ingeniería Eléctrica, Electrónica y de Computación, Universidad Nacional de Colombia Sede Manizales. Manizales, Colombia. Email: {aceballosa, asernam, jbgomezm}@unal.edu.co
[2] GAUNAL, Universidad Nacional de Colombia Sede Bogotá. Bogotá, Colombia. Email: faprietoo@unal.edu.co
[3] Laboratoire Ampère, Institut National des Sciences Appliquées (INSA–Lyon), Lyon, France. Email: {juan-bernado.gomez-mendoza, tanneguy.redarce} @insa-lyon.fr

INTRODUCCIÓN

El problema de reconocimiento automático del habla en señales de audio se ha tratado regularmente a través del modelado de las señales, utilizando técnicas como Redes Neuronales [14] o Modelos Ocultos de Markov [19], las cuales reportan buenos resultados en la literatura. Sin embargo, cuando las condiciones acústicas son adversas, su desempeño se ve afectado. Recientemente, el reconocimiento audiovisual del habla se ha convertido en un campo activo de investigación gracias a los avances en áreas como el procesamiento digital de señales, la visión de máquina y el reconocimiento de patrones [16, 22]. Su objetivo final es permitir la comunicación hombre-máquina usando información audiovisual del habla para combatir las dificultades de un ambiente ruidoso o para tratar de reconocer las emociones exhibidas por el locutor.

Se sabe de los sistemas de comunicación que el análisis visual de la región de la boca del hablante suministra información importante. En particular, los humanos visualizamos el contorno de los labios para mejorar la comprensión del habla [7]. En los trabajos de Campbell [3] se muestra que cuando el oyente tiene información visual de la región de la boca del hablante, la relación señal a ruido (SNR) puede incrementarse hasta en 15 dB. En [29] y [30] se aborda el problema del reconocimiento audiovisual del habla desde una perspectiva neurofisiológica. Se estudia el efecto de la atención en la integración humana de la información audiovisual para el reconocimiento del habla y se realizan experimentos en los que las señales de audio y vídeo son contaminadas con ruido para generar el efecto McGurk [31] y ponderar su importancia en tareas de reconocimiento visual, reconocimiento auditivo y reconocimiento audiovisual del habla.

Una posible aplicación es en algunos sistemas robóticos de cirugía, donde el cirujano tiene las manos ocupadas manipulando diferentes herramientas mediante un joystick, y debe intercambiar los controles, usando pedales, para reubicar algunos dispositivos, por ejemplo la cámara endoscópica. El uso de un sistema audiovisual de reconocimiento del habla en un sistema de este tipo permitiría que la reubicación de dispositivos se realice mediante el uso de comandos hablados.

En este documento se propone un sistema totalmente automático para el reconocimiento de órdenes usando información audiovisual. Se emplean modelos ocultos de Markov como técnica de reconocimiento del habla; la señal de audio se caracteriza usando los coeficientes cepstrales en frecuencia de Mel, mientras que para la señal visual se emplean características basadas en los puntos que definen el contorno externo de la boca. Para el seguimiento del contorno externo de la boca sobre secuencias de vídeo, se propone un algoritmo basado en apariencia que utiliza las restricciones morfológicas del estándar MPEG-4. Este algoritmo no requiere el uso de marcadores o alguna clase de maquillaje para resaltar los labios. A diferencia del sistema presentado en [4-5], donde la selección inicial de los puntos del contorno externo de la boca se realiza manualmente, lo que lo convierte en un sistema semiautomático, el sistema presentado aquí es totalmente automático. Para esta inicialización automática de los puntos se utilizó un contorno activo basado en flujo del vector gradiente [21], el cual a su vez es inicializado por la envolvente convexa de la imagen de pseudotono normalizada del labio [6].

Lo que resta del documento tiene la estructura que se describe a continuación. Se presenta el diseño del sistema audiovisual de reconocimiento de órdenes, descompuesto en tres partes: i) metodología para el diseño de sistemas de reconocimiento de habla empleando sólo información de audio (ASR) y empleando Modelos Ocultos de Markov (HMMs); ii) diseño del sistema de reconocimiento visual del habla (VSR), enfatizando en la extracción inicial de los puntos a seguir en la secuencia de video; iii) finalmente, se explica cómo se pasa de estos dos sistemas al de reconocimiento audiovisual (AVSR). En la sección de resultados se evalúa el sistema para órdenes en español y francés, se muestra el desempeño frente al ruido y se compara con los resultados obtenidos en [4-5]. Por último, se presentan algunas conclusiones del trabajo.

DISEÑO DEL SISTEMA AUDIOVISUAL DE RECONOCIMIENTO DE ÓRDENES

Para el diseño del sistema de reconocimiento audiovisual de órdenes se implementa primero un sistema de reconocimiento del habla a partir de señales de audio (ASR) usando Modelos Ocultos de Markov (HMMs); posteriormente, se propone

un sistema para el reconocimiento visual del habla (VSR); finalmente, se combinan los dos. A continuación se hace una breve descripción de los HMMs y se presentan los dos sistemas antes mencionados.

Modelos ocultos de Markov (HMMs)

Los HMMs son una de las metodologías más populares en la solución de problemas relacionados con el reconocimiento del habla. Los HMMs son modelos estadísticos en los que la señal de entrada se procesa como una señal estacionaria a trozos, y cuya salida es una secuencia de símbolos [24]. En la literatura se reporta que para este tipo de problemas los HMMs ofrecen mejor desempeño que los sistemas basados en plantillas o en redes neuronales [25]. Los HMMs hacen una descripción probabilística de un fenómeno y representan temporalmente una secuencia de variables. Los modelos tienen un número finito de estados N, denotados como $S = \{S_1, S_2, \ldots, S_N\}$, y el estado en el instante t como q_t. Un HMM describe un fonema o palabra como una transición finita de estados en un solo sentido (Figura 1). Se define la probabilidad de que la señal de voz en el instante t pertenezca a cada estado como b_j, mientras que la forma en que el modelo cambia de un estado a otro es dada por la probabilidad de transición de estados a_{ij} [26-27].

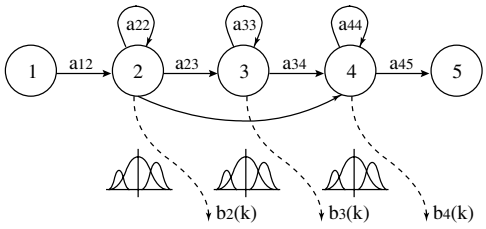


Figura 1. Estructura de un modelo oculto de Markov empleado en ASR

Para definir la configuración más adecuada del HMMs para la tarea de reconocimiento del habla variamos la cantidad de estados del sistema y el número de funciones gaussianas por estado, durante la etapa de entrenamiento. En [28] se muestra que para más de 20 estados hay un crecimiento logarítmico de la probabilidad, debido a un sobreajuste de los datos, y que una configuración usando 10 estados y una función gaussiana por estado es la que presenta mejor desempeño. Como se verá más adelante, esta es la configuración que se empleó para la implementación del sistema.

Sistema de reconocimiento del habla usando sólo audio (ASR)

El problema de reconocimiento automático del habla (ASR–"Audio Speech Recognition") mediante el análisis de señales de audio ha sido ampliamente abordado en la literatura [15, 19]. Por lo tanto, en este trabajo se utilizaron técnicas y herramientas conocidas para el reconocimiento de comandos de voz, como las provistas en el HTK (Hidden Markov Model Toolkit), que es un paquete distribuido bajo licencia Open Source [23]. El principal 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 [8]. Adicionalmente, y debido a que es modular, puede ser empleado en la investigación de otras aplicaciones que usen HMMs, como el reconocimiento audiovisual del habla.

Entre los aspectos a considerar en el desarrollo de un sistema ASR usando HTK, tenemos: i) Definición de las reglas de sintaxis del lenguaje, por ejemplo, qué palabras pueden ser reconocidas y en qué orden deberían aparecer. ii) Definición del diccionario, el cual 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. iii) Etiquetado de archivos, es decir, obtener los archivos de audio con sus respectivas transcripciones a nivel de fonemas. iv) Definición y extracción de características, correspondientes a los primeros 12 índices de Mel, la energía de la señal y las primeras dos derivadas temporales de los índices; el análisis se hace en ventanas de 20 milisegundos con traslape del 50%, siendo la observación el vector de características acústicas de cada muestra. Además, debido a que estas características son dependientes de la amplitud de la señal de audio de entrada, se normalizaron todos los archivos con respecto al máximo valor de amplitud antes del cálculo de los índices de Mel. Finalmente, v) Creación y entrenamiento de los modelos: primero se crea el prototipo de cada cadena de Markov; aunque no existe forma directa de definir el número de estados, la matriz de transición ni la función de densidad de probabilidad para

cada estado, en la literatura se ha probado que tres estados son suficientes y, generalmente, se emplean modelos de propagación hacia adelante (de izquierda a derecha) para simplificar los cálculos y reducir el número de parámetros. 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, aunque puede usarse una mezcla de gaussianas.

Reconocimiento del habla basado en fonemas

La aproximación más natural en sistemas ASR es usar fonemas como unidades básicas que conforman el habla, ya que de esta forma los humanos reconocemos la información contenida en la señal de voz, concatenando sonidos que forman las palabras; siendo capaces además de ignorar aquellos sonidos que no conllevan a una respuesta lógica y separar palabras que unimos cuando hablamos de forma continua. 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. Sin embargo, para que los modelos de los fonemas sean robustos a la pronunciación y al acento, se hace necesario el uso de extensas bases de datos durante el entrenamiento.

La base de datos audiovisual VidTIMIT, desarrollada por C. Sanderson [17], posee 430 frases en inglés de 43 sujetos (hombres y mujeres) en su mayoría de origen australiano, aunque también hay presencia de extranjeros, sobre todo de origen oriental, que no hablan inglés como lengua materna. En esta base de datos, 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. Antes del diseño del sistema ASR, la base de datos debe ser acondicionada. Para el caso del audio, consiste en segmentar la señal en unidades básicas, y teniendo en cuenta que se desea diseñar un sistema que reconozca palabras (comandos) que no están dentro de la base de datos, estas unidades deben ser fonemas. Por lo tanto, se etiquetaron manualmente los fonemas de 80 frases de la base de datos VidTIMIT. Posteriormente, se entrenó un sistema ASR para hacer reconocimiento de palabras claves en habla continua; cuando se buscaron 21 palabras de las 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%. Este resultado se debe a que, como se mencionó, el audio en la base de datos VidTIMIT presenta un alto nivel de ruido.

Sistema de reconocimiento de habla basado en palabra aislada

Al igual que en otras bases de datos como VidTIMIT [17], BANCA [36] o CUAVE [37], el interés principal es desarrollar un sistema audiovisual de reconocimiento automático del habla, por lo cual se trata de abarcar un espectro amplio de acentos y de incluir tanto hombres como mujeres en las muestras. Debido a la falta de disponibilidad de corpus para el reconocimiento audiovisual del habla tanto en francés como en español [34], se ha construido una base de datos propia, que posee expresiones en francés y en español. Esta base de datos está compuesta por 2 frases comunes en francés, 7 comandos en francés y 6 en español. La información de vídeo se encuentra en formato NTSC, con una frecuencia de muestreo de 29,97 fotogramas por segundo (fps) y con cuadros de 720×480 píxeles, mientras que el audio está muestreado a 32 kHz. Para los comandos en francés se grabaron 18 sujetos, 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 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. Las dos frases en francés son: "Je ne sais pas quelle sera l'arme utilisée pour la troisième guerre mondiale, mais la quatrième se fera à coups de bâtons et de gourdins" y "Les ordinateurs ne servent à rien, ils ne peuvent donner que des réponses". Además, los siguientes comandos en francés: "À gauche", "À droite", "Monter", "Reculer", "Arrière", "Avant" y "Descendre"; y en español: "Izquierda", "Derecha", "Arriba", "Abajo", "Atrás" y "Adelante".

En la construcción de esta base de datos se hizo un esfuerzo por registrar diferentes matices idiomáticos, por lo que la selección se realizó de manera aleatoria tratando de incluir la mayor variabilidad disponible en las personas de la muestra. Sin embargo, en el caso general, estas variabilidades son infinitas debido a que están asociadas a cada hablante, a

Tabla 1. Matriz de confusión usando el enfoque de palabra aislada para idioma español.

	Derecha	Izquierda	Adelante	Atrás	Arriba	Abajo
Derecha	51	0	0	0	0	0
Izquierda	0	41	0	0	0	1
Adelante	0	0	50	0	0	0
Atrás	1	0	0	49	0	0
Arriba	0	0	1	0	45	0
Abajo	0	0	0	0	0	51

Tabla 2. Matriz de confusión usando el enfoque de palabra aislada para idioma francés.

	À droite	À gauche	Arrière	Avant	Descendre	Monter	Reculer
À droite	35	0	0	0	0	0	0
À gauche	0	35	0	0	0	0	0
Arrière	0	0	34	0	0	0	0
Avant	0	0	0	34	1	0	0
Descendre	0	0	0	0	35	0	0
Monter	0	0	0	0	0	35	0
Reculer	0	0	0	0	0	0	32

sus raíces, su nacionalidad y su cultura, por lo que una base de datos que incluya la totalidad de estas variaciones sería demasiado grande y se saldría del alcance de este trabajo. La principal limitante de esto corresponde a la parte visual, ya que el costo computacional de almacenamiento del vídeo no hace posible que se tengan tantos sujetos como aquellas bases de datos de sólo audio, ni que se tengan suficientes palabras para considerarlas fonéticamente balanceadas [35].

Para probar el algoritmo de reconocimiento se entrenó un sistema ASR totalmente automático para el reconocimiento de comandos en español y francés. Los sistemas usan como unidades básicas palabras en lugar de fonemas, y el objetivo es el reconocimiento de palabras aisladas y no de palabras claves en habla continua. Se usó el 70% de la base de datos para el entrenamiento y el 30% para la evaluación. En la literatura se ha asumido 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 variando desde 3 hasta 20 el número de estados, y usando una, dos o tres funciones gaussianas por estado. Se encontró el mejor resultado con el uso de 20 estados con una función gaussiana por estado. La respuesta del sistema ASR usando dicha configuración, puede ser observada en la matriz de confusión de las Tablas 1 y 2.

Sistema de reconocimiento visual del habla (VSR)

El primer paso en la construcción de un sistema visual de reconocimiento (VSR–"Visual Speech Recognition") es la extracción de características visuales. Estas características pueden ser divididas en características de alto nivel, características de bajo nivel y características combinadas. Los modelos paramétricos que definen el contorno de los labios son usados como características de alto nivel o características de forma [1, 10]: las características de bajo nivel, o características de apariencia, son obtenidas de transformaciones a nivel de píxel de la región de la boca [13, 11]; finalmente, las características combinadas mezclan características de forma y apariencia de la región de la boca, concatenando las características o usando modelos estadísticos [2]. Generalmente, el vector de características visuales captura información dinámica como la primera y la segunda derivada de los índices de Mel. Las características visuales deben ser interpoladas debido a que la frecuencia de muestreo del audio es mayor que la frecuencia de muestreo del vídeo.

El estándar MPEG-4 fue creado debido a la necesidad de estandarizar objetos virtuales de vídeo real y sintético. Este estándar incluye codificación de vídeo, compresión geométrica y sincronización audio-vídeo. Adicionalmente, contiene un conjunto de parámetros de definición del rostro (FDPs–"Face Definition Parameters"), usados para la estandarización del rostro, y otro conjunto de parámetros para animación del rostro (FAPs–"Face Animation Parameters"). Los FAPs pueden ser usados para describir movimientos del rostro (modelos deformados) con respecto al estado neutro del modelo del rostro. En el estándar se definen 68 FAPs divididos en 10 grupos. Los grupos 2 y 8 (Figura 2) son utilizados en reconocimiento del habla

y describen los movimientos del contorno interno y externo de los labios, respectivamente.

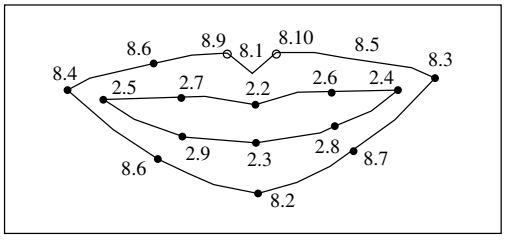


Figura 2. FAPs de los grupos 2 y 8 del estándar MPEG4.

Para extraer las características visuales de alto nivel del habla es necesario realizar un seguimiento de la boca en las secuencias de vídeo. Este seguimiento es un problema difícil debido a la complejidad de la forma, color, textura de los labios y cambios inesperados de iluminación [20]. Para este trabajo se modificó el algoritmo para el seguimiento de los labios propuesto en [4], el cual está basado en restricciones de apariencia y forma, como las definidas en el estándar MPEG-4. El algoritmo asume que el vídeo contiene imágenes frontales del hablante y no se requiere el uso de marcadores labiales. La modificación consistió en hacer automático el proceso de seguimiento de los labios en lugar de la inicialización manual que este requería. En [1], los autores mostraron que el uso de los FAPs del grupo 2, asociados al labio interno, no incrementa significativamente el desempeño de los sistemas de reconocimiento automático del habla con respecto a los sistemas que sólo realizan el seguimiento del labio externo. Por lo anterior, el algoritmo se diseñó solamente para el seguimiento de los FAPs del grupo 8, asociados al contorno externo de los labios (Figura 2).

Algoritmo de seguimiento automático del labio

El algoritmo propuesto en [4] está compuesto por los siguientes pasos: *i*) Localización de la región de interés, *ii*) Ubicación de 10 puntos del contorno externo de la boca, *iii*) Seguimiento del contorno de los labios, y *iv*) Cálculo de las características de la región de la boca. En dicho algoritmo, los pasos *i*) y *ii*) se realizan de forma manual. Para el sistema aquí propuesto, todo el seguimiento de los labios se realiza de forma automática, por lo que dichos pasos han sido sustituidos por los siguientes: *i*) Realce de la región de la boca, y *ii*) Inicialización automática del contorno de los labios. A continuación describiremos esta metodología.

Paso 1: Realce de la región de la boca

En el espacio RGB, los píxeles de la piel y los labios tienen componentes diferentes. En ambos casos, la componente roja es predominante; sin embargo, la diferencia entre el rojo y el verde es mayor para los labios que para la piel [6]. En [12], Hulbert & Poggio propusieron un pseudotono, definido en la ecuación (1), con el objetivo de maximizar esta diferencia.

$$h(x,y) = \frac{R(x,y)}{G(x,y) + R(x,y)} \tag{1}$$

Buscando reducir el ruido, sobre la imagen se aplica un filtro pasa-bajas. Posteriormente, la señal RGB es corregida para reducir la dependencia de la luminancia; la corrección que se utilizada proviene de la ley de Michaelis-Menten [18], la cual modela la adaptación de la visión humana ante los cambios de iluminación. Las nuevas componentes, notadas por *(Rc, Gc, Bc)*, son calculadas como se muestra en la ecuación (2), donde *L(x,y)* es la luminancia del píxel *(x,y)* y *a, b* son parámetros de corrección con valores ajustados heurísticamente en *a=0,4 y b=0,8*,

$$\begin{aligned} R_c(x,y) &= \frac{R(x,y)}{R(x,y) + (b-a)L(x,y) + a} \\ G_c(x,y) &= \frac{G(x,y)}{G(x,y) + (b-a)L(x,y) + a} \\ B_c(x,y) &= \frac{B(x,y)}{B(x,y) + (b-a)L(x,y) + a} \end{aligned} \tag{2}$$

Finalmente, el pseudotono es normalizado usando la ecuación (3), donde *max(h)* y *min(h)* son, respectivamente, el valor máximo y mínimo del pseudotono en toda la imagen. La Figura 3a ilustra el realce de la boca utilizando este procedimiento.

$$h_{norm}(x,y) = \frac{h(x,y) - \min(h)}{\max(h) - \min(h)} \tag{3}$$

Paso 2: Inicialización automática del contorno labial

Buscando que la selección inicial de los 10 FAPs del grupo 8 del estándar MPEG-4 sea un proceso automático y robusto, debemos suministrar al sistema únicamente la imagen de la región de la boca en la que estamos interesados. En la primera imagen de la secuencia de vídeo, el contorno de los labios es extraído utilizando una GVF snake o contorno activo de flujo del vector gradiente [21]. Este contorno es usado como inicialización para los siguientes pasos del algoritmo y los 10 FAPs del grupo 8 (puntos de interés) son calculados usando geometría euclidiana. El procedimiento se describe en el Algoritmo 1. En la Figura 3 se observa el resultado del proceso de inicialización y localización de los 10 FAPs de interés.

Para tener una medida de qué tan adecuado es el algoritmo de inicialización automática (Algoritmo 1), se compararon los puntos seleccionados automáticamente contra los puntos seleccionados manualmente con el fin de computar el error de inicialización. En cada una de las secuencias de vídeo se calculó la distancia euclídea entre cada punto manual y su correspondiente punto automático. Debido a que la distancia entre la cámara y el parlante no es la misma, las distancias medidas en píxeles no son comparables entre un vídeo y otro, por lo que para calcular el error de inicialización se normalizaron las distancias utilizando el ancho de la boca. Una vez normalizadas, estas distancias ya no están medidas en píxeles, sino que son cantidades adimensionales. El error calculado para 40 inicializaciones fue $e=0,0520$ y su desviación estándar fue $\sigma=0,0448$. La normalización significa que el error es igual a 1 cuando la distancia entre el punto automático y el punto manual es igual al ancho de la boca; teniendo en cuenta esto, los errores de inicialización obtenidos son pequeños, y como se verá en la sección de resultados, no afectan de ninguna forma el desempeño del sistema de reconocimiento, por lo que esta metodología de inicialización se considera adecuada para este problema. En la Figura 4 se muestra una comparación entre los puntos obtenidos manual y automáticamente.

Algoritmo 1: Inicialización Automática del Contorno Labial.

[Entrada:]	Imagen de pseudotono corregido y normalizado $h_{norm}(x, y)$
[Salida:]	10 puntos del contorno externo de la boca definidos en el grupo 8 del estándar MPEG-4.
[Paso 1:]	Aplicar un umbral automático a $h_{norm}(x, y)$ para seleccionar las regiones con mayor intensidad. Ver Figura 3b.
[Paso 2:]	Debido al ruido y la iluminación no uniforme, deben llenarse algunos huecos de la imagen umbralizada aplicando una dilatación morfológica. Adicionalmente, esta dilatación garantiza que la inicialización de la GVF Snake contenga en su interior a la región labial.
[Paso 3:]	Usando código de cadena, encontrar y eliminar las regiones pequeñas, conservando únicamente la región de mayor tamaño. Ver Figura 3c.
[Paso 4:]	Calcular la envolvente convexa de la región de mayor tamaño y utilizarla como inicialización del algoritmo GVF Snakes. Ver Figura 3d.
[Paso 5:]	Calcular la fuerza externa GVF usando la imagen de pseudotono normalizada $h_{norm}(x,y)$ como parámetro de entrada [21]. Ver Figura 3e.
[Paso 6:]	Deformar iterativamente la GVF Snake (contorno activo) hasta minimizar su funcional de energía (ecuación (4)). El mínimo de esta función se obtiene cuando la snake alcanza la posición y la forma del contorno deseado. Ver Figura 3f. $$E_{snake} = \int_s E_{\text{int}}\left[V(s)\right] + E_{ext}\left[V(s)\right] ds$$ $$E_{\text{int}}\left[V(s)\right] = \underbrace{\frac{\alpha}{2}\left\|V'(s)\right\|^2}_{E.\ Contorno} + \underbrace{\frac{\beta}{2}\left\|V''(s)\right\|^2}_{E.\ Contorno} \quad (4)$$
[Paso 7:]	Encontrar los 10 puntos del grupo 8 del estándar MPEG-4 basados en las posiciones descritas en la Tabla 3. Ver Figura 3f.

Tabla 3. Localización de los FAPs para el contorno externo de los labios, recomendada por el estándar MPEG-4.

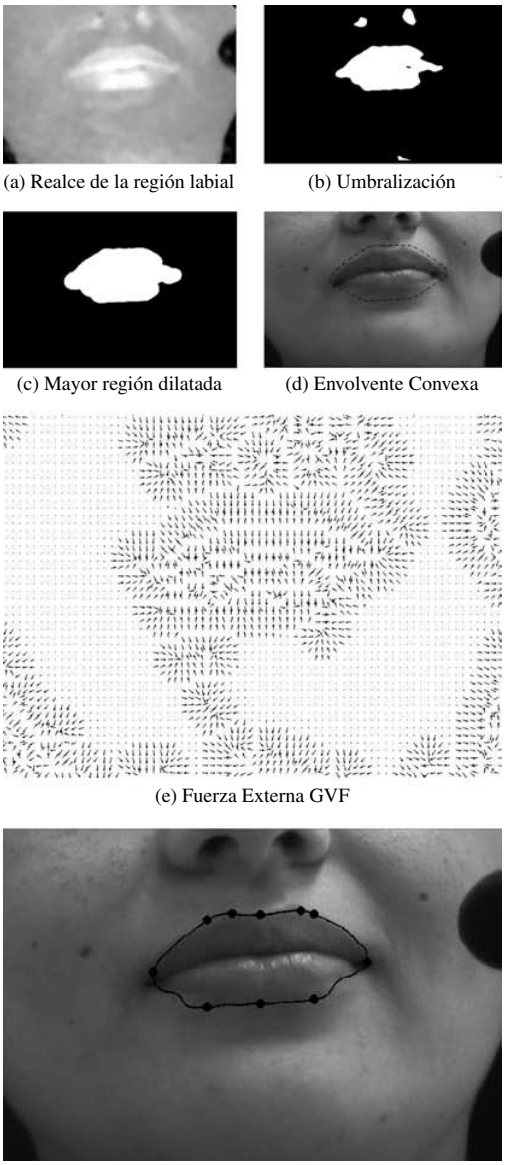
FAP	Localización recomendada
8.1	Punto medio del contorno externo del labio superior
8.2	Punto medio del contorno externo del labio inferior
8.3	Esquina izquierda del contorno externo de los labios
8.4	Esquina derecha del contorno externo de los labios
8.5	Punto medio entre FAPs 8.3 y 8.1 en el contorno externo del labio superior
8.6	Punto medio entre FAPs 8.4 y 8.1 en el contorno externo del labio superior
8.7	Punto medio entre FAPs 8.3 y 8.2 en el contorno externo del labio inferior
8.8	Punto medio entre FAPs 8.4 y 8.2 en el contorno externo del labio inferior
8.9	Punto superior derecho del arco de Cupido
8.10	Punto superior izquierdo del arco de Cupido

Paso 3: Seguimiento de los labios

Para realizar el seguimiento de los labios se usa una medida de similitud entre cuadros de vídeo y algunas restricciones morfológicas definidas 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 RGB, de las ventanas centradas en el punto hallado en el cuadro de vídeo anterior con las ventanas del 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 de vídeo (V_1). La ecuación (4) ilustra este procedimiento.

$$d_{ij} = \left\| V - V_{ij} \right\| + \left\| V_1 - V_{ij} \right\| \tag{4}$$

El mínimo de esta función ocurre cuando los dos vecindarios son idénticos. Para normalizar esta distancia y transformarla a una medida de similitud, se aplica una función exponencial, sopesada con una función gaussiana para dar más importancia a los píxeles centrales de la ventana



(a) Realce de la región labial (b) Umbralización

(c) Mayor región dilatada (d) Envolvente Convexa

(e) Fuerza Externa GVF

(f) Snake y FAPs resultantes

Figura 3. Algoritmo de inicialización automática del contorno labial.

(ecuación (5)). Finalmente, los FAPs con mayor valor de similitud son seleccionados como puntos en el nuevo cuadro.

$$s_{ij} = G_{\mu,\sigma}\left(x, y\right)\ e^{-d_{ij}} \tag{5}$$

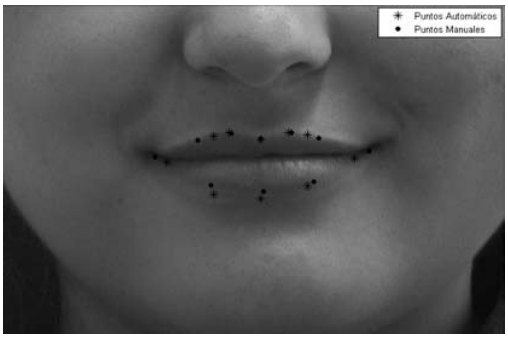


Figura 4. Comparación entre puntos seleccionados automática y manualmente.

Paso 4: Extracción de características

Tal y como se hizo en [5], una vez encontrados los puntos en la secuencia de vídeo, diversas características de la forma de la boca deben ser calculadas. Entre otras se calcularon: el área de la región dentro de los labios *(A)*, la redondez *(R)*, el factor de forma *(FF)*, la relación entre el eje horizontal y el vertical *(RHV)*, el perímetro *(Per)* y diferentes relaciones geométricas entre los puntos.

Evaluación de sistema VSR

Con el objeto de evaluar el Sistema de Reconocimiento Visual de Comandos, una vez extraídas las características mencionadas, se procedió a entrenar el sistema de reconocimiento visual de comandos usando el enfoque de palabra aislada para los dos idiomas: francés y español. El diccionario define la pronunciación de las palabras en función de los Modelos Ocultos de Markov (HMMs). En este caso, los HMMs representan las palabras mismas, por lo que el diccionario empleado para idioma español fue: "Izquierda", "Derecha", "Arriba", "Abajo", "Atrás" y "Adelante"; mientras que para francés el diccionario fue: "À gauche", "À droite", "Monter", "Reculer", "Arrière", "Avant" y "Descendre". Según el enfoque de palabra aislada, éstas están separadas por pausas en lugar de silencios, por lo que las características visuales son almacenadas en diferentes archivos, uno para cada muestra de los comandos presentes en los datos.

Los resultados obtenidos se aprecian en las Tablas 4 y 5, donde se observa que la información brindada por el vídeo, sin considerar el audio, no es suficiente para distinguir entre comandos, aunque sí existe una tendencia a reconocer correctamente, por lo que la mayor concentración de datos está en la diagonal principal. Esto implica que las características visuales brindan información importante en la tarea de reconocimiento. En particular para idioma español, la mayor confusión se presenta para la palabra "Atrás", que es reconocida erróneamente como "Adelante", el 20% de las veces. Mientras que para idioma francés, la mayor confusión se presenta para la palabra "Arrière", la cual fue clasificada erróneamente el 70,58% de las veces.

Tabla 4. Matriz de confusión empleando únicamente características visuales y enfoque de palabra aislada en idioma español.

	Derecha	**Izquierda**	**Adelante**	**Atrás**	**Arriba**	**Abajo**
Derecha	34	4	9	3	1	0
Izquierda	5	31	3	0	3	0
Adelante	7	0	36	3	4	0
Atrás	4	4	10	27	2	3
Arriba	6	2	5	4	29	0
Abajo	0	1	1	6	3	40

Tabla 5. Matriz de confusión empleando únicamente características visuales y enfoque de palabra aislada en idioma francés.

	À droite	**À gauche**	**Arrière**	**Avant**	**Descendre**	**Monter**	**Reculer**
À droite	13	5	5	2	2	3	5
À gauche	3	14	4	5	2	2	5
Arrière	4	5	10	7	3	2	3
Avant	3	8	4	13	2	2	3
Descendre	2	3	5	3	18	2	2
Monter	3	2	7	4	3	11	5
Reculer	2	2	5	1	4	3	15

De forma similar al error de inicialización, el error de seguimiento es calculado como la distancia euclídea normalizada entre cada punto de las secuencias inicializadas automáticamente, y cada punto correspondiente en las secuencias inicializadas manualmente. Para el cálculo se tuvieron en cuenta aproximadamente 40,000 FAPs, para los cuales la media del error y su desviación estándar se muestran

en la Tabla 6. El error máximo de seguimiento se presenta para el FAP 8.2, el mismo para el cual se presenta la máxima desviación estándar. Cabe aclarar que todos los errores son menores a 0,0498, lo que es pequeño comparado con la unidad. Adicionalmente, los errores para los 10 FAPs son muy similares, lo que indica que ni el algoritmo de seguimiento ni el de inicialización están sesgados.

Tabla 6. Error de seguimiento y desviación estándar de los FAPs.

FAP	Error medio	Desviación
8.1	0,0289	0,0106
8.2	0,0498	0,0221
8.3	0,0226	0,0206
8.4	0,0226	0,0206
8.5	0,0269	0,0112
8.6	0,0252	0,0109
8.7	0,0440	0,0190
8.8	0,0432	0,0188
8.9	0,0295	0,0134
8.10	0,0303	0,0139

Sistema de reconocimiento audiovisual del habla (AVSR)

Para el Sistema de Reconocimiento Audiovisual del Habla (AVSR–"Audio Visual Speech Recognition") se empleó como entrada el conjunto conformado por la combinación de las características obtenidas del audio y las extraídas al hacer seguimiento de los labios. En la sección siguiente se presentan los resultados obtenidos con este sistema.

RESULTADOS

Sistema audiovisual de reconocimiento de comandos

Como se mencionó, este sistema fue entrenado siguiendo el procedimiento de las secciones anteriores. Sin embargo, se deben realizar algunos ajustes. El conjunto de entrenamiento debe ceñirse al formato HTK, 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 de características de audio se compone de los primeros 12 coeficientes de Mel y un término de energía; mientras que como características visuales se usan las 3 componentes principales (PCA) de los FAPs y la redondez del contorno externo de la boca.

De las secciones precedentes es claro que los mejores resultados de reconocimiento automático de comandos se hallaron al emplear un enfoque de reconocimiento audiovisual de palabras aisladas. En las Tablas 7 y 8 se presentan los resultados del reconocimiento para seis comandos del habla española y siete del habla francesa empleando este enfoque. Las palabras pronunciadas se encuentran en cada fila y son reconocidas por el sistema según cada columna. Se aprecia que el sistema reconoce correctamente casi todos los comandos, y *"adelante"* y *"arrière"* son las palabras con mayor confusión, pues son reconocidas erróneamente el 4% y 6,25% de las veces, respectivamente. A pesar de esto, la tasa de reconocimiento es alta, mayor al 93% en todos los casos; adicionalmente, no depende de la duración de las palabras.

En la Tabla 9 se muestran los porcentajes de comandos correctamente reconocidos para los sistemas ASR, VSR y AVSR. Los resultados se presentan para cadenas de Markov de 10 estados entrenadas 50 veces. Aunque el rendimiento del sistema AVSR fue menor que con el ASR, este sistema es menos sensible al ruido acústico como se evidencia más adelante.

Tabla 7. Porcentaje de palabras reconocidas por el sistema AVSR para seis comandos en español.

	Derecha	**Izquierda**	**Adelante**	**Atrás**	**Arriba**	**Abajo**
Derecha	51	0	0	0	0	0
Izquierda	0	41	0	1	0	0
Adelante	1	0	48	0	1	0
Atrás	0	0	0	49	1	0
Arriba	0	0	0	0	46	0
Abajo	0	0	0	0	0	51

Tabla 8. Porcentaje de palabras reconocidas por el sistema AVSR para siete comandos en francés.

	À droite	**À gauche**	**Arrière**	**Avant**	**Descendre**	**Monter**	**Reculer**
À droite	33	1	0	1	0	0	0
À gauche	0	35	0	0	0	0	0
Arrière	2	0	32	0	0	0	0
Avant	0	0	0	35	0	0	0
Descendre	1	0	0	0	34	0	0
Monter	0	0	0	0	0	34	1
Reculer	0	0	0	0	0	0	32

Tabla 9. Porcentaje de palabras correctamente reconocidas.

Idioma	**Características**	**Promedio**	**Desviación**
Español	Audio	97,54%	1,42%
	Vídeo	38,36%	5,06%
	Audio + Vídeo	94,01%	2,34%
Francés	Audio	98,94%	1,34%
	Vídeo	12,60%	5,06%
	Audio + Vídeo	94,01%	2,34%

Evaluación del sistema AVSR ante ruido acústico

De la Tabla 9, es claro que los mejores resultados se logran cuando sólo se usa información de audio. Sin embargo, como los resultados de reconocimiento de voz dependen fuertemente de las condiciones de adquisición y de la relación señal a ruido (SNR), se evaluó el sistema agregando diferentes niveles de ruido acústico. Con este fin, se contaminó la información de audio con ruido blanco gaussiano. Se hicieron pruebas con niveles de señal a ruido desde 100 dB hasta 1 dB.

En las Figuras 5 y 6 se hace evidente que el comportamiento del sistema usando únicamente audio es muy inferior cuando la relación señal a ruido es inferior a 20 dB comparado con la señal sin ruido (Tabla 9). También se observa que para ambos sistemas (comandos en francés y español), el desempeño es igual o superior para todos los niveles de ruido cuando se usa el sistema audiovisual

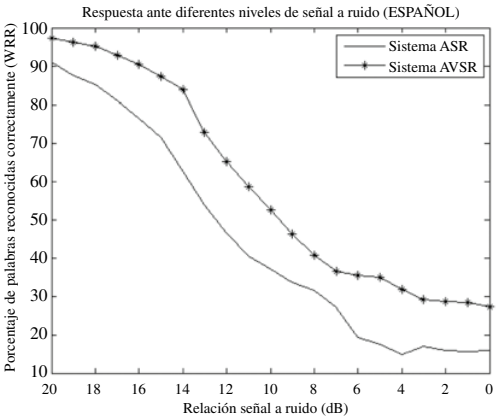


Figura 5. Respuesta del sistema ante ruido acústico en idioma español.

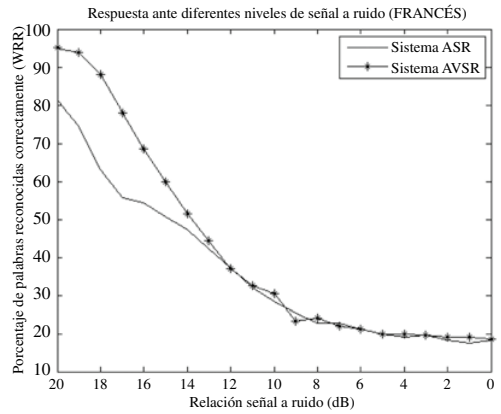


Figura 6. Respuesta del sistema ante ruido acústico en idioma francés.

comparado con el sistema que emplea sólo audio. Para el caso del sistema en francés, y debido a que todos los locutores hablan francés con fluidez, la respuesta del sistema audiovisual es mucho mejor para niveles de ruido altos. Es más evidente aún en español, donde la respuesta del sistema audiovisual es mucho mejor en todos los casos, debido a que los locutores hablan español como lengua materna.

Aunque el desempeño del sistema de reconocimiento utilizando características audiovisuales parece ser inferior al sistema que sólo emplea información acústica, hay que destacar que el comportamiento depende fuertemente de las condiciones de adquisición, de la relación señal a ruido, así como del acento de los locutores. Aunque al realizar el análisis sin ruido no fue evidente, los resultados muestran

que el enfoque de reconocimiento audiovisual del habla tiene un desempeño superior cuando hay presencia de ruido acústico en la señal de entrada. De hecho, el porcentaje de palabras correctamente reconocidas por el sistema ASR cae alrededor de 15% cuando se añade ruido gaussiano a la señal de audio, haciendo que la relación señal a ruido sea de 20 dB. El sistema audiovisual de reconocimiento presenta mejores resultados para todos los niveles de ruido, en especial para los comandos en español.

Adicional al ruido, existen otros inconvenientes producidos por la exposición, registro y organización de las señales (ruido exterior, superposición de sonidos, silencios, retrasos en el vídeo, entre otros), que no fueron considerados en este trabajo. Sin embargo, deben plantearse y tratarse en una aplicación futura con el fin de dar mayor robustez a los sistemas desarrollados.

CONCLUSIONES

Se presentó el desarrollo de un sistema automático de reconocimiento audiovisual de órdenes. Los mejores resultados se lograron al emplear palabras como unidades básicas, el enfoque de reconocimiento de palabras aisladas y Modelos Ocultos de Markov. Con el fin de hacer el sistema robusto ante la presencia de ruido, se utilizó información audiovisual en el diseño del sistema. De hecho, cuando el ruido es apenas una centésima parte de la señal, el porcentaje de palabras correctamente reconocidas por el sistema ASR cae drásticamente, y el sistema audiovisual de reconocimiento presenta mejores resultados para todos los niveles de ruido.

La selección del modelo y las características es una tarea muy importante durante el diseño de un sistema de reconocimiento de habla, y depende de la aplicación específica. En este trabajo se emplearon los coeficientes de Mel como características de audio y características de alto nivel basadas en la forma de la boca como características visuales. Se exploraron características basadas en los parámetros de animación faciales (FAPs), definidos en el estándar MPEG-4, específicamente aquellas 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]. En cuanto a los tres modelos presentados, los resultados mostraron que el enfoque de reconocimiento audiovisual del habla tiene un mejor desempeño superior en presencia de ruido acústico en la señal de entrada, que los enfoques visual o de audio independientemente.

El algoritmo de inicialización automático del contorno labial es adecuado para esta aplicación, puesto que los errores de inicialización y seguimiento fueron pequeños en todos los casos, y los resultados de reconocimiento audiovisual no se vieron afectados y fueron comparables con los obtenidos con el sistema semiautomático presentado en [4] y [5].

Se evidenció que los resultados dependen fuertemente del acento de las personas cuyas secuencias de audio y vídeo hacen parte de la base de datos de entrenamiento. Como trabajo futuro, se construirán bases de datos en otros idiomas e incluyendo mayor variabilidad idiomática, para probar los algoritmos desarrollados.

En los trabajos de Rodríguez-Bravo [32, 33] y Sanz [34] se realizan estudios de la expresividad del habla en los que se describe el proceso fisiológico de producción de la voz, en donde se evidencia que las variaciones en intensidad, tono y timbre son utilizadas para dar información sobre el estado, el contexto y la voluntad del hablante. Adicionalmente, se analizan aspectos importantes en la construcción de corpus orales que incluyan diferentes estilos expresivos. En el contexto de este trabajo, inicialmente no se trabajó en reconocer la expresividad de la persona que emite el comando. Sin embargo, es un trabajo útil para la investigación futura, ya que un sistema de reconocimiento automático debe ser capaz de reconocer eficientemente las instrucciones sin importar las variaciones expresivas del hablante.

AGRADECIMIENTOS

Los autores agradecen el apoyo brindado por el programa Franco-Colombiano ECOS-NORD (ECOSNord/COLCIENCIAS/ICFES/ICETEX).

REFERENCIAS

[1] P.S. Aleksic and A.K. Katsaggelos. “Comparison of MPEG-4 facial animation parameter groups with respect to audiovisual speech recognition performance”. Proceedings of the IEEE

International Conference on Image Processing (ICIP), pp. 501-504. Genoa, Italy. Marzo.

[2] N.S. Alothmany. "Classification of visemes using visual cues". Tesis para optar al grado de Doctor. Swanson School of Engineering, University of Pittsburgh. Pittsburgh, USA. 2009.

[3] R. Campbell. "The processing of audiovisual speech: empirical and neural bases". Philosophical Transactions of the Royal Society B. Vol. 363, Issue 1493, pp. 1001-1010. September, 2007.

[4] A. Ceballos, J.B. Gómez y F. Prieto. "Seguimiento del contorno externo de la boca en imágenes de vídeo". Revista Ingenierías. Universidad de Medellín. Vol. 8 Nº 14, pp. 129-144. Enero-Junio 2009.

[5] A. Ceballos, J.B. Gómez, F. Prieto and T. Redarce. "Robot Command Interface using an Audiovisual Speech Recognition System". Lecture Notes in Computer Science. Vol. 5856, pp. 869-876. 2009.

[6] N. Eveno, A. Caplier and P.Y. Coulon. "New color transformation for lips segmentation". IEEE Fourth Workshop on Multimedia Signal Processing, Cannes, pp. 3-8. Francia. 2001.

[7] N. Eveno, A. Caplier and P.Y. Coulon. "Accurate and Quasi-Automatic Lip Tracking". IEEE Transactions on circuits and systems for video technology. Vol. 14, Issue 5, pp. 706-715. May, 2005.

[8] L. García. "Ecualización de histogramas en el procesado robusto de voz". Tesis para optar al grado de Doctor. Departamento de Teoría de la Señal, Telemática y Comunicaciones. Universidad de Granada. Granada, España. Diciembre 2007.

[9] 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.

[10] R. Goecke. "Current trends in joint audio-video signal processing: A review". Proceedings of the IEEE Eight International Symposium on Signal Processing and Its Applications ISSPA, pp. 70-73. Sydney, Australia. September, 2005.

[11] J. Huang, G. Potamianos, J. Connell and C. Neti. "Audiovisual speech recognition using an infrared headset". Speech Communication. Elsevier. Vol. 44, Issue 1, pp. 83-96. October, 2004.

[12] A.C. Hurlbert and T.A. Poggio. "Synthesizing a Color Algorithm from Examples". Science, New Series. Vol. 239, Issue 4839, pp. 482-485. January, 1988.

[13] M.W. Kim, J.W. Ryu and E.J. Kim. "Speech Recognition with Multi-modal Features Based on Neural Networks". Lecture Notes in Computer Science. Vol. 4233, pp. 489-498. 2006

[14] T.L. Kumar, T.K. Kumar and K.S. Rajan. "Speech Recognition Using Neural Networks". Proceedings of IEEE International Conference on Signal Processing Systems (ICSPS), pp. 248-252. Singapore. July, 2009.

[15] S. Nakagawa. "A Survey on Automatic Speech Recognition". IEICE Transactions on Information and Systems. Vol. 85, Issue 3, pp. 465-486. 2002.

[16] G. Potamianos, C. Neti, J. Luettin and I. Matthews. "Issues in Visual and Audiovisual Speech Processing". MIT Press, pp. 1-30. Boston, Massachusetts, USA. 2004.

[17] C. Sanderson and K.K. Paliwal. "Identity verification using speech and face information". Digital Signal Processing. Elsevier. Vol. 14, Issue 5, pp. 449-480. September, 2004.

[18] D.M. Schneeweis and J.L. Schnapf. "Photovoltage of rod and cones in the macaque retina". Science. Vol. 268, Issue 5213, pp. 1053-1056. May, 1995.

[19] E. Trentin and M. Gori. "A survey of hybrid ANN/HMM models for automatic speech recognition". Neurocomputing. Elsevier. Vol. 37, Issue 1-4, pp. 91-126. April, 2001.

[20] Z: Wu, P.S. Aleksic and A.K. Katsaggelos. "Lip tracking for MPEG-4 facial animation". Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI). Vol. 1, pp. 293-298. Pittsburgh, USA. January, 2003.

[21] C. Xu and J.L. Prince. "Generalized gradient vector flow external forces for active contours". Signal Processing, An International Journal. Elsevier. Vol. 71, pp. 131-139. 1998.

[22] T. Yoshida, K. Nakadai and H.G. Okuno. "Automatic Speech Recognition Improved by Two-Layered Audiovisual Integration for Robot Audition". Proceedings of the 9th IEEE-RAS International Conference

on Humanoid Robots, pp. 604-609. Paris, France. December, 2009.

[23] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland. "The HTK Book". Microsoft Corporation. 2000.

[24] R.J. Elliot, L. Aggoun and J.B. Moore. "Applications of mathematics". In I. Karatzas and M. Yor (Eds.). Hidden Markov Models. Estimation and Control. New York: Springer. 1995.

[25] S. Anderson and D. Kewley-Port. "Evaluation of speech recognizers for speech training applications". IEEE Transactions on Speech and Audio Processing. Vol. 3, Issue 4, pp. 229-241. July, 1995.

[26] L.R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition". Prodeedings of the IEEE, Vol. 77, Issue 2, pp. 267-296. February, 1989.

[27] D. Jurafsky and J.H. Martin. "Speech and Language Processing: An introduction to natural language processing". Prentice Hall, 1 Edition. February, 2000.

[28] J.B. Gómez, A. Ceballos, F. Prieto and T. Redarce. "Mouth gesture and voice command based robot command interface". IEEE International Conference on Robotics and Automation (ICRA), pp. 333-338. Kobe, Japan. July, 2009.

[29] A. Alsius, J. Navarra, R. Campbell and S. Soto-Faracao. "Audiovisual integration of speech falters under high attention demands". Current Biology. Elsevier. Vol. 15, Issue 9, pp. 839-843. May, 2005.

[30] D. Sanabria, C. Spence and S. Soto-Faracao. "Perceptual and decisional contributions to audiovisual interactions in the perception of apparent motion: A signal detection study". Cognition. Elsevier. Vol. 102, Issue 2, pp. 299-310. February, 2007.

[31] H. McGurk and J. MacDonald. "Hearing lips and seeing voices". Nature. Vol. 264, pp. 746-748. December, 1976.

[32] A. Rodríguez-Bravo. "Propuestas para una modelización del uso expresivo de la voz". Revista de Estudios de Comunicación. Nº 13, pp. 157-173. Universidad del País Vasco, Zarauz. 2002.

[33] A. Rodríguez-Bravo, P. Lázaro Pernias, N. Montoya, J.M. Blanco, D. Bernadas, J.M. Oliver y L. Longhi. "Modelización acústica de la expresión emocional en el español". Procesamiento del lenguaje natural. Nº 25, pp. 159-166. Universidad de Lérida. 1999.

[34] I.I. Sanz. "Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva". Tesis para optar al grado de Doctor. Universitat Ramón Llull. España, 2008.

[35] A. Ceballos. "Desarrollo de un sistema de manipulación de un robot a través de movimientos de la boca y de comandos de voz". Tesis para optar al grado de Magíster. Universidad Nacional de Colombia, Sede Manizales. Colombia. 2009.

[36] 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". Lecture Notes in Computer Science. Vol. 2688, pp. 625-638. 2003.

[37] E.K. Patterson, S. Gurbuz, Z. Tufekci and J.N. Gowdy, "CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research". International Conference on Acoustics Speech and Signal Processing. Vol. 2, Issue 68, pp. 2017-2020. 2002.