UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

"Sistema de Reconocimiento Multimodal de Emociones para Interacción Humano - Robot"

T E S I S

Para obtener el Título de:

Maestro en Robótica

Presenta:

Luis Alberto Pérez Gaspar

Director:

Dr. Santiago Omar Caballero Morales

Co-Director:

Dr. Felipe Trujillo Romero

Huajuapan de León, Oaxaca, Agosto de 2015

A la persona que toda mi vida me ha apoyado y guiado, mi mamá, que siempre ha vencido cualquier obstáculo y me ha demostrado que siempre es posible salir adelante. A mi tía y mi abuelita que son mis segundas mamás y que vieron siempre por mi. A mi abuelito por ser esa figura paterna que siempre ha estado conmigo y me ha dado sus consejos a lo largo de mi vida.

Agradecimientos

Quiero agradecer en primer lugar a mi director de tésis, el Dr. Santiago Omar Caballero Morales, por haberme dado las herramientas necesarias para el desarrollo de este trabajo. Asimismo, reconocer su gran capacidad para realizar investigación, ya que fue una pieza fundamental en mi formación y que me brindó la oportunidad de participar en conferencias y publicaciones. Gracias profesor, porque ante cualquier duda nunca se cansó de explicarme, también gracias por escucharme ante cualquier problema que tuviera.

A mi co-director el Dr. Felipe Trujillo Romero por haber aportado sus ideas y dado sus valiosos consejos para la tésis y para cada proyecto, siempre con el objetivo de mejorar el trabajo. Gracias profesor por siempre haber tenido tiempo para escucharme y orientarme tanto en lo académico como en lo profesional. Reconozco su excelente nivel como investigador y conocimientos en diversas áreas.

A mis sinodales la Dr. Lluvia Carolina Morales Reynaga, el Dr. Agustín Santiago Alvarado, el Dr. Enrique Guzmán Ramírez y el M.C. Luis Anselmo Zarza López, por haber invertido parte de su tiempo en revisar este trabajo de investigación y por haber aportado sus observaciones para que fuera un mejor trabajo.

A mis profesores el Dr. Anibal Arias, Dr. Hugo Ramírez Leyva, Dr. Carlos García, Dr. Manuel Arias, M.C. Mario Lomeli y la Dra. Irma Salinas quienes contribuyeron con sus conocimientos a mi desarrollo y que siempre mantuvieron ese entusiasmo al impartir sus clases. Les agradezco infinitamente esa calidad y atención para conmigo.

Resumen

En los humanos la habilidad para interpretar las emociones es muy importante para lograr una comunicación efectiva. En el aspecto tecnológico el reconocimiento automático de emociones representa uno de los retos más importantes para lograr una comunicación humano-robot intuitiva, comprensiva y natural. En este campo de investigación la presente tesis contribuye con el desarrollo de un sistema de reconocimiento multimodal de emociones para su integración en un robot humanoide. Esto con el propósito de crear el medio tecnológico para que un robot sea capaz de interactuar con usuarios permitiendo un diálogo más natural. En comparación con investigaciones similares el presente trabajo aborda de manera integral los siguientes puntos:

- Reconocimiento de emociones en voz y rostro para usuarios mexicanos;
- Desarrollo de base de datos multimodal para el análisis y desarrollo del sistema de reconocimiento (base de datos creada con participantes mexicanos, Base de Datos MX);
- Identificación de factores que afectan el desempeño del reconocimiento de emociones en voz y rostro;
- Integración de optimización evolutiva dentro del desarrollo del sistema para determinar parámetros importantes para mejorar su desempeño;
- Desarrollo de sistema de diálogo contextual para interacción por voz con el sistema robótico acorde con el estado emocional reconocido;
- Conexión del robot humanoide Bioloid con el sistema de reconocimiento e importación de librerías de movimiento realizadas con el software RoboPlus.

Para el desarrollo del sistema de reconocimiento se consideraron las emociones de Enojo, Felicidad, Neutro y Tristeza. Las técnicas de Redes Neuronales Artificiales (Artificial Neural Networks, ANNs) y Modelos Ocultos de Markov (Hidden Markov Models, HMMs) fueron utilizadas para el desarrollo de los sub-sistemas de reconocimiento visual y vocal respectivamente.

Experimentos realizados con la base de datos MX y una base de datos estándar (Base JAFFE) proporcionaron información acerca de la dependencia que hay entre la tasa de

reconocimiento y la base de datos utilizada para el desarrollo del sistema. La integración de optimización evolutiva mediante Algoritmos Genéticos (Genetic Algorithms, GAs) fue considerada para dar solución a esta situación. Las mejoras en desempeño con la integración de GAs fue estadísticamente significativa.

De igual manera los desempeños de los sub-sistemas de reconocimiento visual y vocal proporcionaron información acerca de las emociones que son reconocidas de mejor manera por cada uno de ellos. Este análisis de desempeños fue considerado para una ponderación de sub-sistemas para su integración en el sistema de reconocimiento multimodal.

Para añadir un esquema de aplicación al reconocimiento de emociones la integración de un sistema de diálogo contextual fue considerado. Este sistema de diálogo fue desarrollado mediante Máquinas de Estado Finito (Finite State Machines, FSMs) el cual funciona mediante reconocimiento de voz en modo de habla continua.

Finalmente en pruebas en-vivo con usuarios mexicanos diferentes a los participantes de la Base MX el sistema multimodal tuvo una precisión en el reconocimiento de emociones mayor al 95 %.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En años recientes el desarrollo de robots para asistir a los humanos en diversas actividades cotidianas ha tenido un auge significativo. Como ejemplo se puede mencionar a los robots de asistencia para personas con discapacidad motora o de edad avanzada [59] o robots de servicio [61]. Sin embargo una parte de la que se ha hablado últimamente y que está sujeta a investigación es la relación afectiva entre un robot y un humano.

Adentrándonos en el tema de interacción humano-robot, uno de los diversos pioneros de los llamados "robots sociales" fue la compañía Sony, la cual introdujo en el mercado a AIBO, un robot mascota diseñado para entretenimiento y diversión [27]. Al surgir este nuevo juguete se cambió la perspectiva de tener una mascota en casa. Este hecho fue un caso favorable para la aceptación de los robots en el hogar, ya que al realizarse un estudio en personas de diferentes edades se demostró que tuvo cierta empatía con los usuarios. Una investigación realizada con niños de preescolar [39] demostró que cuando se comparó a AIBO con un perro normal se obtuvo una mayor aceptación al sistema robótico, cambiando de igual forma el comportamiento hacia una actitud positiva y de alegría. Foros de discusión en internet sobre esta plataforma notificaron que se trataba como a un animal real, el cual podía proveer de compañía y satisfacción emocional. Kahn *et al.* [38] muestra los resultados obtenidos de pruebas con diferentes personas acerca de AIBO, argumentando que no sólo se trataba de un juguete con baterías, micrófono y cámara sino de un verdadero ser con inteligencia artificial capaz de interactuar con el humano de forma recíproca de acuerdo a sus emociones.

Por otra parte, el desarrollo de una interfaz inteligente multimodal para monitorear

la salud y estado de ánimo de pacientes a distancia fue presentado en [47]. Esta interfaz denominada MOUE (a Model Of User's Emotions) realizaba el monitoreo de gestos faciales y señales fisiológicas (ritmo cardíaco, temperatura) por medio de una computadora portable conectada a la muñeca del paciente y una cámara. Estos datos se trasladaban a una computadora central en donde se realizaba el procesamiento de éstos. La computadora central contaba con una pantalla que mostraba una animación de una persona "avatar" que reflejaba los estados emocionales del paciente dependiendo de lo que observaba la cámara (rasgos faciales) y las señales fisiológicas recibidas. El avatar podía, ya sea interactuar con el humano dependiendo de su estado de ánimo con una serie de preguntas, o simplemente reflejaba la emoción del paciente.

En el caso del robot KISMET [10] se estudia el rol que juega la emoción y el comportamiento expresivo durante una interacción social humano-robot. Este robot es capaz de reconocer intenciones afectivas a través del tono de la voz y de integrarse en situaciones sociales de una manera muy natural, además de expresar sus propias emociones.

Un aspecto importante de estos robots es la ingeniería necesaria para lograr una interacción con los humanos de una manera más natural, inteligente y eficiente. El desarrollo de robots amigables, para un usuario humano, que reconozcan emociones se ha vuelto un tema muy importante de estudio y un factor que contribuye a mejorar la interacción humano-robot [82, 42]. Esto es principalmente importante para robots humanoides [102].

En la Universidad Tecnológica de la Mixteca se han realizado proyectos previos dentro del campo del reconocimiento de expresiones y de la integración del factor emocional para lograr una interacción humano-computadora más eficiente:

- En [12] fue presentado el desarrollo de un sistema de detección de contornos para rostros y sus características faciales (ojos, nariz, boca) con futura aplicación para el reconocimiento de expresiones emocionales. Dentro de los retos encontrados para la tarea de localización de un rostro en una imagen se identificaron la pose, la oclusión parcial, y las condiciones de iluminación. La detección del rostro fue abordada usando la técnica *Snake* de contornos activos, suavizado Gaussiano, y el operador *Sobel* de detección de bordes. Las características faciales fueron detectadas mediante histogramas y un modelo geométrico. Dentro de las restricciones de este sistema se tuvieron las siguientes: detección de rostros de personas con piel

clara, sin oclusiones, rotaciones o inclinaciones en el rostro, la ropa del usuario debía ser contrastante con el color de su piel, y el fondo debía ser de color claro y uniforme. Pruebas realizadas con la base de datos de rostros GTAV mostraron un porcentaje de detección de características faciales de 98.32 % - 99.04 %.

- En [8] se presentó el desarrollo de avatares emocionales foto-realistas para una interacción humano-computadora más natural. Estos avatares se integraron dentro de una interfaz llamada "Memo" (Mensajero Emocional). El diseño de los avatares se realizó mediante Ingeniería Kansei y el propósito de la interfaz fue el intercambio de mensajes entre contactos de forma instantánea. La interfaz se probó con 10 usuarios mexicanos (5 hombres y 5 mujeres) y se encontró que el uso de los avatares mejoraron la experiencia afectiva de interacción entre los usuarios.

En comparación con el trabajo presentado en [12] en esta tesis se plantea lo siguiente:

- detección del rostro completo sin hacer segmentación de características faciales (es decir, no se extraen los ojos, la nariz o la boca para su análisis independiente);
- reconocimiento de la expresión emocional del rostro;
- reconocimiento adicional de la emoción a partir de los patrones de voz del usuario;
- integración de computación evolutiva para el mejor funcionamiento de ambos procesos de reconocimiento de emoción;
- desarrollo de un sistema que integre ambos procesos de reconocimiento de emoción (sistema multimodal);
- desarrollo de un sistema de diálogo por voz que soporte una conversación sencilla entre el usuario humano y un ente artificial (robot humanoide) de acuerdo a la emoción reconocida por el sistema multimodal.

Respecto al trabajo presentado en [8] en esta tesis se plantea el reconocimiento de patrones visuales y acústicos (expresiones faciales y voz) para determinar el estado emocional de un usuario y establecer un diálogo con un sistema robótico. No se considera un estudio acerca de las características que debe tener un ser artificial (avatar) para mejorar

la interacción con el usuario. De igual manera, tampoco se considera el cambio en el estado de ánimo del usuario como resultado de la interacción con el sistema robótico. Sin embargo, las técnicas presentadas en esta tesis pueden contribuir a extender los objetivos presentados en [8] para una interacción afectiva más eficiente entre personas y robots de servicio.

Por lo tanto, en general, la contribución de esta tesis es la integración de técnicas de reconocimiento de patrones, así como su mejora mediante optimización evolutiva, para el reconocimiento multimodal de emociones. En particular el sistema considera a usuarios mexicanos y las características del lenguaje español para su desarrollo y análisis. Esto con el propósito de crear el medio tecnológico para futuros desarrollos en México, como el presentado en [8], que estén enfocados a lograr una interacción humano-computadora, o humano-robot, más natural y eficiente.

1.1. Planteamiento del Problema y Limitaciones

En el aspecto tecnológico el reconocimiento automático de emociones representa uno de los retos más importantes para lograr una comunicación humano-robot intuitiva, comprensiva y natural [42, 76].

En la actualidad existen diversos tipos de robots domésticos que realizan tareas específicas. Ahora bien, la necesidad de tener robots sociales que puedan interactuar con la gente es cada vez mayor ya que se pretende que el robot pueda establecer una relación más afectiva con el humano dependiendo de su estado de ánimo.

La ejecución de órdenes por parte de los robots es un tema ampliamente abarcado, pero ahora se requiere también de una parte afectiva con el humano. Es por esto que el problema va más allá de sólo realizar funciones que faciliten la vida a las personas, orientándose a que el robot tenga la capacidad de comunicarse con los usuarios de acuerdo a la emoción que perciba. Es decir, que el robot tenga una respuesta diferente para una persona feliz, enojada o triste, con el objetivo fundamental de mejorar la comunicación y/o mantener una conducta positiva. De acuerdo a lo mencionado anteriormente, el desarrollo de un sistema de reconocimiento multimodal de emociones es vital para tener una interacción humano-robot de una forma más natural como la tendría un humano con otro.

El desarrollo de dicho sistema representa muchos retos entre los cuales se pueden mencionar los siguientes:

- Poca disponibilidad de recursos (bases de datos) visuales, acústicos, y audiovisuales para análisis y desarrollo, sobretodo de recursos gratuitos [54].

- Tasa de reconocimiento variable (30.00 % - 95.00 %) dependiendo del usuario, de la técnica utilizada para el reconocimiento de emociones, y del medio para detectar la emoción (gestos faciales, voz) [81, 102, 60, 69, 22].

El presente trabajo busca contribuir a la solución de los retos mencionados mediante la aportación de alternativas de mejora para el desarrollo de un sistema multimodal de emociones. Una de las limitantes de estas alternativas establece una acción del robot de acuerdo a la emoción detectada, no se abarca el verificar que haya habido un cambio de emoción en el usuario como resultado de esta acción ni la medición del mismo. Esta investigación está orientada sólo al reconocimiento de emociones multimodal y la integración de este sistema con un robot humanoide.

1.2. Justificación

Un sistema de reconocimiento de emociones implementado en sistemas robóticos es importante dado que dicho sistema puede mejorar la interacción entre el robot y el humano [10, 39, 38]. Esto a su vez abre el campo de aplicación para otros proyectos de investigación como aquellos enfocados a cambiar los niveles de sociabilidad de las personas bajo la presencia de entes artificiales, o aquellos enfocados en cambiar el estado de ánimo del usuario humano mediante dinámicas que realice con el robot.

Entre los puntos a resaltar de la importancia de este trabajo para otras investigaciones se pueden mencionar los siguientes:

- Al tener la capacidad de identificar emociones y actuar con base en ellas se puede modelar un comportamiento más humano y natural como lo tendría una persona normal, permitiendo la interpretación y reafirmación de acciones o intenciones durante la interacción [101].

- Capacidad del robot para interpretar las emociones del usuario e influir su estado si éste se encontrara triste [36].

- Desarrollo de robots de compañía dotados con inteligencia artificial mejorada [95] para aplicaciones de distracción y como medio para liberar tensiones a través del diálogo.

- Robots de asistencia para personas (por ejemplo, bebés [96] y su ajuste a diversos tipos de usuarios [34]).

- Se puede obtener un sistema portable a cualquier robot para que pueda implementar el reconocimiento y la interacción.

- Utilización del robot con el sistema de reconocimiento multimodal en centros para el cuidado de gente adulta u hospitales para niños.

Para el desarrollo del presente trabajo de tesis se utilizó el robot humanoide Bioloid Premium [71] ya que es una plataforma accesible en cuanto a disponibilidad y costo. De igual manera esta plataforma es ampliamente utilizada por investigadores de diversos campos en la robótica y es de las más presentadas en el Torneo Mexicano de Robótica y la RoboCup [74].

1.3. Hipótesis

La integración de optimización evolutiva puede reducir la variabilidad de la tasa de reconocimiento, y mejorar el desempeño general, de un sistema de reconocimiento multimodal de emociones al evaluarse con bases de datos y pruebas en-vivo. Esta optimización puede determinar mejores estructuras para los reconocedores de voz y visión del sistema multimodal para obtener una precisión conjunta y consistente mayor al 95.00 %.

1.4. Objetivos

1.4.1. Objetivo General

Desarrollar un sistema de reconocimiento multimodal de emociones para usuarios mexicanos con una precisión mayor del 95.00 % y su aplicación en actividades de interacción humano-robot. Para esto se utilizará el robot humanoide Bioloid y las técnicas de HMMs y ANNs para la construcción del sub-sistema de voz y visión respectivamente. Así mismo se aplicarán los GAs para mejorar el desempeño del sistema. Finalmente se realizará un sistema de diálogo a través del uso FSMs.

1.4.2. Objetivos Específicos

- Creación de una base de datos audiovisual con participantes mexicanos que provea de las características visuales y acústicas necesarias para el desarrollo del sistema multimodal. La base de datos debe presentar los siguientes estados emocionales: Enojo, Felicidad, Neutro y Tristeza.
- Adaptación de la técnica de Análisis de Componente Principal (del inglés Principal Component Analysis, PCA) para la extracción de características para el reconocimiento de emociones visual.
- Desarrollo de un sub-sistema estable de reconocimiento de emociones por voz. La técnica de HMMs es propuesta para dicho sistema.
- Desarrollo de un sub-sistema estable de reconocimiento de emociones en expresiones faciales. La técnica de ANNs unidireccionales es propuesta para dicho sistema.
- Integración de optimización evolutiva mediante GAs para la mejora estadísticamente significativa de ambos sub-sistemas.
- Integración de ambos sub-sistemas para su trabajo conjunto en la tarea de reconocimiento multimodal de emociones en usuarios mexicanos.
- Desarrollo de un sistema de diálogo acerca de un día escolar que permita la administración del sistema multimodal para llevar una conversación entre el robot y

el usuario, bajo el esquema de pregunta-respuesta (el robot pregunta y el usuario responde).

- Implementación del sistema multimodal para su uso con el robot humanoide Bioloid.

1.5. Sistema de Reconocimiento Multimodal de Emociones Propuesto

En la Figura 1.1 se presenta el diagrama a bloques del sistema de reconocimiento multimodal propuesto y las técnicas utilizadas para el desarrollo del mismo.

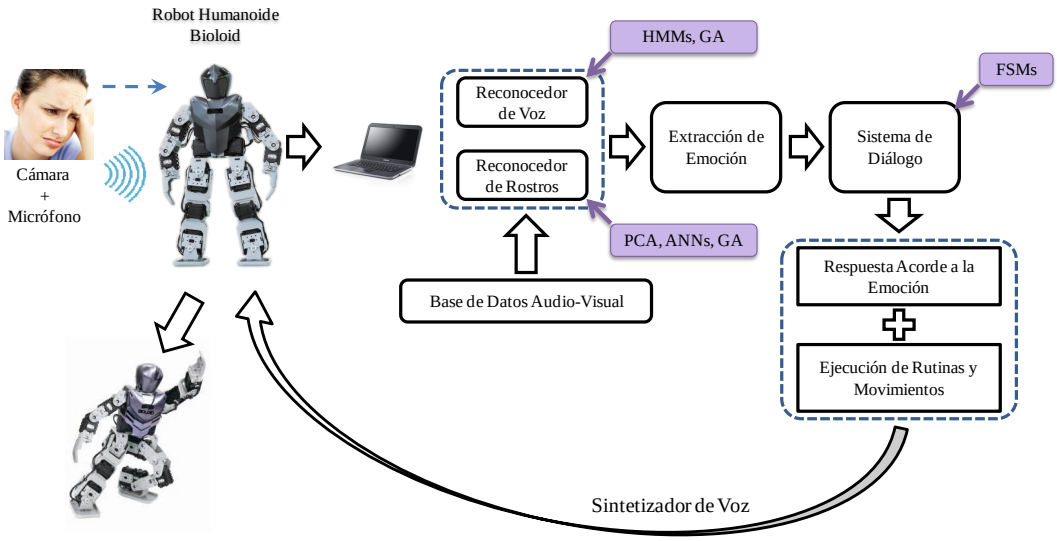


Figura 1.1: Diagrama de componentes y técnicas para el sistema de reconocimiento multimodal de emociones

El sistema multimodal de reconocimiento de emociones que se propone incluye el uso del robot humanoide Bioloid, el desarrollo de una base de datos y el desarrollo de dos sub-sistemas: el sub-sistema de voz y el sub-sistema de visión, ya que al considerar la integración de ambos se tiene el sistema multimodal completo. A continuación se describen de manera general los componentes del sistema:

- **Base de Datos Audio-Visual.** Para la elaboración de la base de datos se reclutaron personas de nacionalidad mexicana, de las cuales se capturaron imágenes con fon-

do blanco para la toma de rostros y se grabaron frases predefinidas expresando las emociones consideradas. Cada frase formulada estuvo orientada a los siguientes estados de ánimo: Neutro, Feliz, Enojo, Tristeza. Esto proporcionó información visual y acústica para el aprendizaje supervisado de los sub-sistemas de visión y voz respectivamente.

- **Reconocedor de Voz.** En lo que concierne al reconocedor de voz (sub-sistema de voz), se trabajó con HMMs mediante la herramienta computacional HTK de acceso libre, la cual permite construir y manipular estos modelos. HTK es usada en muchas aplicaciones relacionadas con el reconocimiento de voz. Así mismo se integró un GA para obtener las estructuras de los HMMs que mejor representen los fonemas de las vocales dependiendo de su emoción.

- **Reconocedor de Rostros.** Para el reconocedor de rostros (sub-sistema de visión) se emplearon las técnicas de PCA y ANNs. PCA fue utilizada como técnica de reconocimiento y de extracción de características faciales asociadas a cada emoción (expresión). Las caraterísticas extraídas se modelaron mediante ANNs para su posterior reconocimiento (clasificación). De igual manera se integró un GA para obtener la estructura de las ANNs más apropiada para modelar los patrones faciales de cada expresión emocional.

- **Extracción de la Emoción.** En esta etapa se determina la emoción considerando la respuesta de cada sub-sistema (voz y visión). Para esto se aplicaron sumas ponderadas, asignando mayor peso a la respuesta del sub-sistema que mejor desempeño tuvo al clasificar una emoción en particular.

- **Sistema de Diálogo.** Este sistema se encarga de administrar un diálogo con el usuario de acuerdo a la emoción reconocida. El robot humanoide Bioloid realiza acciones de acuerdo a la emoción reconocida por el sistema multimodal. Como caso de studio el vocabulario del sistema de diálogo se consideró dentro de un contexto de conversación acerca de las actividades cotidianas que realiza un estudiante en un día. El sistema de diálogo se desarrolló con FSMs con un enfoque retroalimentado para su constante construcción y adaptación.

- **Rutinas y Movimientos.** Se consideró el desarrollo de rutinas para que el humano

pueda interactuar con el robot. Estas rutinas consistieron de secuencias específicas de movimientos del robot propias de una actividad de distracción, pudiendo abarcar estiramientos, ejercicios anti-estrés, expresiones, bailes y chistes. Los movimientos del robot se modelaron con la herramienta RoboPlusMotion del sistema Bioloid y se consideraron tres rutinas para cada emoción. El sistema robótico se enlazó inalámbricamente con la computadora a través de una adaptación con los módulos de radiofrecuencia Zigbee.

1.6. Estructura de la Tesis

El trabajo realizado en la presente tesis se encuentra estructurado de la siguiente manera:

- **Capítulo 2: Marco Teórico**. En este capítulo se presentan antecedentes de trabajos realizados en reconocimiento automático de emociones. De igual manera se presentan en detalle las técnicas que dan sustento al desarrollo del sistema de reconocimiento multimodal propuesto (técnicas consideradas para extracción de características, clasificación de patrones y optimización).

- **Capítulo 3: Base de Datos Emocional**. En este capítulo se presentan los detalles de creación de la base de datos emocional con participantes mexicanos para el análisis y desarrollo de los sub-sistemas de voz y visión del sistema multimodal.

- **Capítulo 4: Sub-sistemas de Reconocimiento de Emociones en Voz y en Expresiones Faciales**. En este capítulo se presentan los detalles teóricos y técnicos del desarrollo de los sub-sistemas de reconocimiento de emociones basados en voz y en expresiones (gestos) faciales.

- **Capítulo 5: Optimización Evolutiva e Integración de Sistema Multimodal**. En este capítulo se presentan los detalles técnicos de los GAs diseñados para encontrar los parámetros que mejoren estadísticamente el desempeño de los sub-sistemas de reconocimiento de emociones. De igual manera se presenta la integración de los sub-sistemas optimizados en el sistema multimodal.

- **Capítulo 6: Sistema de Diálogo y Resultados de Interacción Multimodal con el Robot Humanoide.** En este capítulo se presentan los detalles de desarrollo del sistema de diálogo para el sistema de reconocimiento multimodal y las rutinas de movimiento para el robot humanoide. También se presentan los resultados de pruebas en-vivo con nuevos usuarios reclutados para el sistema multimodal y un experimento final que incluye la interacción del usuario con el robot humanoide a través del sistema de diálogo.

- **Capítulo 7: Conclusiones y Trabajo a Futuro**. En este capítulo se presentan las conclusiones generales del presente trabajo al igual que las futuras extensiones del mismo.

1.7. Publicaciones

El trabajo presentado en esta tesis se ha difundido en las siguientes publicaciones:

- Pérez-Gaspar, L., Caballero-Morales, S.O., Trujillo-Romero, F. "Integración de Optimización Evolutiva para el Reconocimiento de Emociones en Voz", Research in Computing Science, Vol. 93, pp. 9-21, 2015 (ISSN: 1870-4069).

- Pérez-Gaspar, L., Caballero-Morales, S.O., Trujillo-Romero, F. "Factores en el Reconocimiento Facial de Emociones y la Integración de Optimización Evolutiva", Research in Computing Science, Vol. 91, pp. 45-56, 2015 (ISSN: 1870-4069). **Proyecto Ganador del 3er Lugar en el Congreso Mexicano de Inteligencia Artificial (COMIA 2015) en Mayo de 2015** [32].

- Pérez-Gaspar, L., Trujillo-Romero, F., Caballero-Morales, S.O., Ramírez-Leyva, F.H. "Curve Fitting Using Polygonal Approximation for a Robotic Writing Task", In Proc. of the 2015 International Conference on electronics, Communications and Computers (CONIELECOMP 2015), Cholula, Puebla, Mexico, 25-27 February 2015, p. 184-189, 2015.

- Pérez-Gaspar, L., Caballero-Morales, S.O., Trujillo-Romero, F. "Error Modelling Approach based on Artificial Neural Networks for Face Emotion Recognition", Research in Computing Science, Vol. 78, p. 21-30, 2014 (ISSN: 1870-4069).

Capítulo 2

Marco Teórico

A través de los años los robots se han utilizado para la ejecución de tareas repetitivas y pesadas, de tareas peligrosas que involucren algún riesgo físico o biológico para el humano. Durante los años de 1950 en adelante, se desarrollaron novelas de ciencia ficción que involucraban a los robots pero ya no para la ejecución de tareas. Estas historias fueron lideradas principalmente por Isaac Asimov en donde se contemplaba la relación de los robots con los humanos de una forma más natural. Estas historias, con el paso de los años, poco a poco se han ido convirtiendo en realidad. Es por esto que en este trabajo de tesis se desarrolla una parte fundamental en el proceso de la evolución de los robots al desarrollar un sistema de reconocimiento de emociones.

En este capítulo se presenta el marco teórico relacionado con el desarrollo de sistemas de reconocimiento de emociones (enfoques y resultados más relevantes) al igual que casos de aplicación en sistemas robóticos. Esta información se complementa con la presentación de las bases teóricas de las herramientas utilizadas para el presente trabajo como son técnicas de extracción de características, reconocimiento de patrones (voz y expresiones faciales), modelado de secuencias para sistemas de diálogo, y optimización evolutiva.

2.1. Estado del Arte en Reconocimiento de Emociones

El reconocimiento de emociones es un tema estudiado últimamente en el campo de la robótica. Esto se debe a la constante evolución de la tecnología, la cual da pauta para

desarrollar agentes robóticos con mayor inteligencia. En la literatura existen diversos trabajos realizados con la finalidad de desarrollar sistemas capaces de reconocer emociones. A continuación se presenta una reseña de dichos trabajos.

2.1.1. Reconocimiento en Expresiones Faciales

En [40] se presentó la aplicación de GAs y ANNs para el reconocimiento de siete emociones (neutro, miedo, felicidad, tristeza, enojo, disgusto y sorpresa) en un solo usuario. Las partes del rostro que se consideraron para el reconocimiento fueron los labios y un ojo. Previo al procesamiento de los patrones faciales se aplicó una ecualización para mejorar el contraste y tener uniformidad, después se aplicaron filtros para suavizar la imagen y por último se aplicó el método Sobel para detección de bordes. Para la extracción de características se consideraron los siguientes métodos: perfil de proyección, perfil de contorno, y momentos. Un GA se utilizó para determinar formas elípticas para extracción de regiones representativas de los labios y el ojo. La función de elegibilidad del GA estaba basada en tres ecuaciones para determinar las distancias que mejor describieran la posición de las elipses dentro de las regiones a analizar. Para el proceso de clasificación (reconocimiento) se utilizaron dos modelos de ANN unidireccionales (conocidas como *feedforward* en inglés). En estos modelos de ANN las neuronas de cada capa reciben información de las neuronas de la capa precedente y transmiten la información a las neuronas de la capa siguiente. Estas ANNs usaron como entradas la parte alta y baja de los labios y el ojo. La arquitectura de la primera ANN consistía de tres neuronas en la capa de entrada, 20 neuronas en la capa oculta, y tres neuronas en la capa de salida para codificar de manera binaria las siete emociones a reconocer ($3\times20\times3$). Para la segunda ANN la estructura $3\times20\times7$ fue considerada en donde cada emoción a reconocer tenía asignada un bit en la salida. Las ANNs se entrenaron con el algoritmo de *backpropagation* con 1000 *epochs*, un rango de aprendizaje de 0.0001 y una función de activación $\frac{1}{1+e^{-x}}$. Con estos parámetros la primera ANN obtuvo un reconocimiento promedio para un solo usuario de 83.57 % en tanto que la segunda ANN obtuvo un promedio de 85.13 %.

Estas siete emociones también fueron reconocidas en [50] mediante la aplicación de PCA, Patrones Binarios Locales (Local Binary Pattern, LBP) y Máquinas de Soporte

Vectorial (Support Vector Machine, SVMs). Inicialmente las imágenes de un usuario pasan por un proceso de normalización de brillo, ecualización de histogramas y filtrado. Después PCA es aplicada para la extracción de características globales de la imagen completa en escala de grises. Dado que la captura de imágenes se implementó a partir de una cámara de vídeo, las regiones del rostro no son estáticas y cambian de acuerdo al tiempo. Esto puede afectar el desempeño de PCA y para abordar esta situación se integró la técnica de Patrones Binarios Locales (Local Binary Pattern, LBP) para la extracción de características. Finalmente para la clasificación de las características extraídas se utilizó la técnica de SVMs dando una tasa de reconocimiento final de 93.75 % para un usuario.

En [25] el reconocimiento de las mismas siete emociones se realizó con la base de datos FEEDTUM. En dicho trabajo las imágenes fueron pre-procesadas para determinar, con el algoritmo de Canny, los bordes de dos regiones en particular: ojos y boca. PCA fue aplicada para reducción de dimensionalidad de estas regiones y un conjunto de 14 ANNs (una ANN para cada región asociada a cada emoción) fue definido para la clasificación de cada una de ellas. La arquitectura de cada ANN consistió de una capa oculta con 20 neuronas y función de transferencia *tansig* mientras que la función de transferencia para la capa de salida fue *lineal*. El algoritmo de aprendizaje (entrenamiento) fue el de Levenberg-Marquart. Las tasas de reconocimiento obtenidas fueron de 46.00 % a 80.00 % con un promedio de 70.00 %.

La aplicación de ANNs también fue presentada en [65] en donde la base de datos JAFFE (Japanese Female Facial Expression) [52, 51] fue utilizada para reconocer las emociones de enojo, miedo, felicidad y neutro. Las imágenes de la base de datos fueron pre-procesadas con una ecualización por histogramas, un filtro adaptativo de media y el Algoritmo de Optimización de Alimentación Bacteriana (Bacterial Foraging Optimization Algorithm, BFOA) para reducción de ruido. Las regiones que se consideraron importantes fueron los ojos, los labios y la boca. Estas regiones fueron dadas a una ANN para su entrenamiento cuya configuración era la siguiente: estructura $625 \times 75 \times 4$ con tasa de aprendizaje de 0.5, 5000 *epochs*, un error de 0.001 y una función de transferencia de tipo *sigmoide*. La tasa de reconocimiento promedio para las emociones consideradas fue alrededor de 90.00 %.

También en el trabajo presentado en [41] se hizo uso de la base de datos JAFFE.

En este caso se utilizó para el reconocimiento de cinco emociones principales (enojo, felicidad, tristeza, disgusto y sorpresa). Las técnicas de PCA y SVD (Singular Value Decomposition) se aplicaron para el reconocimiento y extracción de características. Se consideraron dos regiones faciales principales: ojos-cejas y boca. Los resultados obtenidos para el reconocimiento de cada emoción fueron: (a) 95.00 % para Felicidad, (b) 70.00 % para Disgusto, (c) 85.00 % para Sorpresa, (d) 60.00 % para Enojo y (e) 90.00 % para Tristeza. En general la tasa promedio de reconocimiento fue del 80.00 %.

El uso de PCA y la base de datos JAFFE también se presentó en [89] y [31]. En [89] se obtuvieron las siguientes tasas de reconocimiento para las emociones de Enojo, Felicidad, Tristeza, Disgusto, Miedo y Sorpresa: 100.00 %, 100.00 %, 100.00 %, 85.00 %, 87.00 % y 75.00 % respectivamente con una tasa general de 91.16 %. En cambio en [31] con la aplicación adicional del método Sobel para detección de bordes se obtuvieron las siguientes tasas de precisión en el reconocimiento: 97.14 %, 90.00 %, 82.86 %, 90.00 %, 95.71 % y 91.43 % respectivamente con una tasa general del 91.19 %.

En [69] el reconocimiento de las emociones de Enojo, Felicidad, Tristeza, Miedo y Neutro en vídeo fue presentado mediante la aplicación de ANNs Auto-asociativas (AANNs). Este tipo de ANNs son del tipo *feedforward* y para dicho trabajo tuvieron una estructura con cinco capas en donde las capas de entrada y salida se configuraron con funciones de transferencia *lineal* y las tres capas ocultas (intermedias) con funciones de transferencia no-lineal del tipo *tanh*. El algoritmo estándar de *backpropagation* (retro-propagación) fue utilizado para el aprendizaje de las AANNs. Las regiones consideradas como representativas para la extracción de características fueron los ojos y la boca y se construyó una AANN para cada emoción y región. Los resultados finales de reconocimiento con una base de datos creada para dicho trabajo presentaron una tasa promedio de 87.00 %.

La técnica de lógica difusa también se ha aplicado dentro de este campo. En [87] el reconocimiento de las emociones de felicidad, tristeza, enojo, miedo, disgusto, sorpresa y neutro fue abordado mediante la utilización de lógica difusa para la detección y segmentación de rostros. PCA fue utilizada para reducción de dimensionalidad y obtención de características. La clasificación fue realizada mediante el cálculo de distancias euclidianas. Experimentos realizados con la base de datos FACES Collection [24] dieron una tasa de reconocimiento promedio de 96.66 %.

En [18] el reconocimiento de felicidad, tristeza, disgusto, enojo, sorpresa y miedo fue realizado mediante la aplicación directa de lógica difusa. La base de datos Cohn-Kanade (CK+) [49] fue utilizada para dicho trabajo. Las imágenes inicialmente fueron pre-procesadas con un filtro Wiener de tamaño 5×5 para suavizamiento considerando la media y varianza de los pixeles adyacentes. Después las características se definieron como vectores que representaran las distintas distancias entre los elementos del rostro que definieran cada expresión (cejas, ojos, nariz y boca). Finalmente los resultados obtenidos con el clasificador basado en lógica difusa fueron (a) 100.00 % para felicidad, (b) 93.00 % para tristeza, (c) 85.00 % para disgusto, (d) 85.00 % para enojo, (e) 93.00 % para sorpresa y (f) 70.00 % para miedo. Estas emociones también fueron reconocidas con un sistema difuso del tipo Mamdani en el trabajo presentado en [62] con la base de datos FEEDTUM. El enfoque abordado para la extracción de características fue el de realizar un trazado de ocho líneas (horizontales y verticales) en la región del rostro a fin de determinar el cambio de patrones en ojos, cejas, frente, nariz, dientes, mejillas, labios y barba para cada emoción tomando como referencia una expresión neutra o normal. El resultado del sistema difuso con 14 reglas dió una tasa general de reconocimiento de 89.33 %.

Una integración de lógica difusa con GAs fue presentada en [35] para las emociones de felicidad, tristeza, disgusto, enojo, sorpresa y miedo. En dicho trabajo un GA fue diseñado para ajustar los parámetros de las funciones de membresía del sistema difuso. Para la extracción de características se consideraron varios factores como la abertura de los ojos, de la boca, la contracción de las cejas, ancho de la boca, existencia de dientes en la expresión y grosor de labios. Experimentos con la base de datos RaFD (Radboud Faces Database) [43] mostraron una tasa de reconocimiento general del 93.96 %.

Finalmente el sistema difuso presentado en [70] con la base de datos JAFFE obtuvo una tasa general del 92.33 % para las mismas emociones. La extracción de características para dicho trabajo consistió de la segmentación en cinco regiones principales del rostro para las cuales se calculó su energía (sumatoria de los pixeles al cuadrado), media y varianza.

2.1.2. Reconocimiento en Voz

En el campo del reconocimiento automático de emociones ha habido investigaciones enfocadas a la identificación de emociones en la señal de voz. En [19] el reconocimiento de las emociones de Tristeza, Enojo, Sorpresa, Miedo, Felicidad y Disgusto con la base de datos emocional BHUDES (Beihang University Database of Emotional Speech) en lenguaje Chino Mandarín fue realizado con los métodos de SVM, Fisher, PCA y ANN. La extracción de características espectrales de la señal de voz fue realizada mediante bandas de frecuencia en escala de Mel. Los métodos de PCA y Fisher fueron utilizados para la reducción de dimensionalidad bajo los siguientes enfoques: selección y extracción de características más discriminativas o representativas. Para el modelado y reconocimiento de las características extraídas se utilizaron SVMs y ANNs. Los resultados obtenidos para cada método de extracción y reconocimiento fueron los siguientes: (a) 50.16 % para Fisher+SVM, (b) 43.15 % para PCA+SVM, (c) 40.43 % para Fisher+ANN, y (d) 39.16 % para PCA+ANN. Las pruebas reportaron una confusión significativa entre Felicidad y Sorpresa.

En [16] una SVM multi-clase fue desarrollada para el reconocimiento de cinco emociones (Enojo, Miedo, Felicidad, Neutro y Tristeza). Los Coeficientes Cepstrales en las Frecuencias de Mel (Mel-Frecuency Cepstral Coefficients, MFCCs), histogramas de periodicidad y patrones de fluctuación fueron usados para la extracción de características. Experimentos con la base de datos emocional de voz danesa DES (Danish Emotion Speech) presentaron tasas de reconocimiento del 64.77 %, 78.41 %, 79.55 % y 78.41 % para funciones Kernel Lineal, Polinomial, RBF (Radial Basis Function) y Sigmoide respectivamente usando la SVM multi-clase. Una confusión significativa fue observada entre felicidad y enojo. También con el enfoque de SVMs el trabajo presentado en [98] reportó tasas de reconocimiento de 77.16 %, 65.64 %, 83.73 % y 70.59 % respectivamente para las emociones de Enojo, Felicidad, Neutro y Tristeza.

Una aplicación sobre un sistema robótico fue presentada en [4] para el Robot MEXI (Machine with Emotionally eXtended Intelligence). Este robot presentaba la peculiaridad de poder interactuar con los humanos a través del habla, ya que poseía un módulo de síntesis de voz integrado con una entonación emocional. El sistema era capaz de reconocer cinco emociones (enojo, miedo, tristeza, felicidad y neutro) y fue desarrollado con el sistema PROSBER a base de lógica difusa. Las tasas de reconocimiento repor-

tadas por este sistema en configuración dependiente e independiente del usuario fueron alrededor del 84.00 % y 60.00 % respectivamente.

Dentro de las técnicas de codificación de características espectrales para reconocimiento de emociones en voz las más utilizadas han sido MFCCs [16, 77, 26, 58, 92, 13], NUPLP (Non-Uniform Perceptual Linear Prediction) [103] y LPCCs (Linear Predictive Cepstral Coefficients) [79, 63, 53]. En cuanto a la implementación del sistema de reconocimiento las técnicas de HMMs [13, 79, 63, 100, 99], SVMs [80, 92, 19, 16, 94, 98] y ANNs [26, 19, 94] han sido las más utilizadas.

En general el desempeño de los reconocedores de emociones en voz es similar o un poco menor al de los reconocedores basados en expresiones faciales [2]. Para SVMs con seis emociones se han reportado tasas de 43.15 % a 50.16 % [19] en tanto que para cinco y cuatro emociones se han reportado tasas de 64.77 % a 79.55 % [16] y de 75.33 % a 78.16 % [94] respectivamente. Para el caso de HMMs se han reportado tasas de 94.32 % y 76.12 % para cuatro emociones [13, 44], 62.50 % [63] y 87.00 % [99] para cinco emociones, y 86.00 % para siete emociones [79]. El uso de ANNs ha reportado tasas de 39.16 % a 40.43 % para seis emociones [19] y de 68.50 % [26] a 71.87 % [94] para cuatro emociones.

2.1.3. Reconocimiento Multimodal en Interacción Humano-Robot

El desarrollo de un sistema multimodal involucra la integración de dos sistemas: (a) reconocimiento acústico o vocal, y (b) reconocimiento visual. Un ejemplo de este sistema se presenta en [85] para el reconocimiento de emociones a través de expresiones faciales y patrones de voz. En dicho trabajo se utilizó un HMM Triple (Triple Hidden Markov Model, THMM) para sincronizar las características de voz con las características de las expresiones faciales en el dominio del tiempo para el reconocimiento integrado. Para el reconocimiento de expresiones faciales se seleccionó la distancia entre los ojos, entre el ojo y la nariz, entre la boca y la nariz, y entre los extremos de la boca (ancho). Para la señal de voz se consideraron 48 características prosódicas y 16 formantes. Las características de energía y tono fueron importantes para el sistema de voz. De manera independiente para el reconocimiento de cinco emociones cada sistema tuvo el siguiente desempeño:

- Sistema visual: 90.24 % para Sorpresa, 87.18 % para Felicidad, 90.28 % para Enojo, 80.00 % para Miedo, 87.67 % para Tristeza, y 89.06 % para Neutral (promedio = 87.40 %).

- Sistema acústico: 86.58 % para Sorpresa, 71.79 % para Felicidad, 76.39 % para Enojo, 74.67 % para Miedo, 80.82 % para Tristeza, y 98.43 % para Neutral (promedio = 81.45 %).

Estos resultados mostraron un desempeño menor en el sistema de voz con respecto al sistema visual. Dicha situación se comentó en la Sección 2.1.2. A pesar de esto, la integración de ambos sistemas puede dar un mayor desempeño que el del mejor sistema independiente. Para el trabajo presentado en [85] el desempeño del sistema multimodal fue el siguiente: 93.90 % para sorpresa, 94.87 % para felicidad, 93.05 % para enojo, 88.00 % para miedo, 93.15 % para tristeza, y 96.87 % para neutral (promedio = 93.31 %). Como se observa un incremento de 5.91 % fue obtenido por el sistema multimodal (93.31 %) con respecto al sistema de visión que tuvo el mejor desempeño de manera independiente (87.40 %).

Este comportamiento también fue observado en los trabajos presentados en [11, 33]. En [11] se realizó el reconocimiento de las emociones de felicidad, neutro, enojo y tristeza utilizando las técnicas de PCA y SVM. Mientras que el sistema de voz tuvo un desempeño general de 71.00 %, y el sistema de visión un desempeño de 85.00 %, el sistema multimodal tuvo un desempeño mayor de 89.00 %. En cambio, en [33] para siete emociones (neutro, miedo, felicidad, tristeza, enojo, disgusto, sorpresa) con las técnicas de MFCCs para extracción de características (sistema de voz), PCA y LDA, se obtuvo un desempeño multimodal de 98.00 %. Sin embargo, estos trabajos usaron solo un usuario para el entrenamiento de las técnicas de reconocimiento. Trabajos que han considerado más usuarios para el entrenamiento de las técnicas de reconocimiento han reportado tasas generales de 64.00 % (tres emociones, 11 usuarios) [78] y 82.00 % (seis emociones, ocho usuarios) [93]. El desarrollo de sistemas multimodales es importante para lograr un mejor reconocimiento de emociones para una interacción humano-robot eficiente.

Dentro del campo de la interacción multimodal humano-robot se pueden encontrar desarrollos de robots sociales tanto para fines comerciales como académicos. En la Fi-

gura 2.1 se muestra el robot KISMET que es uno de los robots sociales más conocidos [10]. KISMET fue desarrollado en el Instituto Tecnológico de Massachusetts (MIT) por Cynthia Breazeal para investigaciones de interacción humano-robot afectiva. Este robot estaba dotado de cuatro cámaras y varios motores los cuales podían gesticular partes de su cara para representar varias emociones (enojo, miedo, felicidad, cansancio, disgusto, sorpresa, tristeza, interés y calma). KISMET podía reconocer la intención del usuario a través del tono de la voz y encajar en conversaciones de una forma natural.

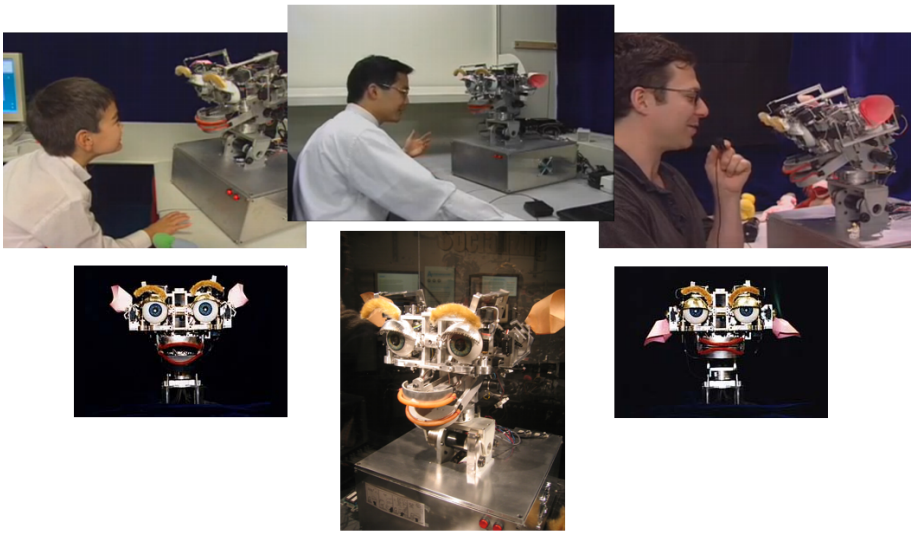


Figura 2.1: Robot Social KISMET.

Como sistema descendiente de KISMET se tiene al robot JIBO [17] que se presenta en la Figura 2.2. Este es un robot social desarrollado para estar en la casa de las personas y será introducido al mercado para su venta en el año 2015. Este robot también fue desarrollado por Cynthia Breazeal quien comenta que JIBO es diferente a los demás sistemas robóticos porque trata al usuario de una forma más natural o humana, involucrando sentimientos. Este robot está dotado de diferentes habilidades que lo hacen ser un robot de compañía muy atractivo al usuario humano. Entre éstas se pueden mencionar la de recordar eventos importantes, tomar fotografías, relatar historias, y establecer conexión Wi-Fi para comunicación multimodal con otras personas a través de su cámara y la de dispositivos móviles.

Otro robot introducido al mercado como robot de entretenimiento para todas las edades es el famoso AIBO de la compañía de Sony el cual se presenta en la Figura 2.3.

Figura 2.2: Robot Social JIBO.

AIBO es un robot mascota diseñado para entretenimiento y diversión el cual generó empatía entre las personas [27]. Una investigación realizada con niños de preescolar [39] demostró que cuando se comparó a AIBO con un perro normal se obtuvo una mayor aceptación al sistema robótico, cambiando de igual forma el comportamiento hacia una actitud positiva y de alegría. Foros de discusión en internet sobre esta plataforma notificaron que se trataba como a un animal real, el cual podía proveer de compañía y satisfacción emocional. En [38] se mostraron resultados obtenidos en pruebas con diferentes personas acerca de AIBO, argumentando que no sólo se trataba de un juguete con baterías, micrófono y cámara sino de un verdadero ser con inteligencia artificial capaz de interactuar con el humano de forma recíproca de acuerdo a sus emociones.

2.1.4. Comparación y Puntos Relevantes

En la Tabla 2.1 se presenta el compendio de los trabajos más relevantes en el ámbito del reconocimiento automático de emociones a partir de expresiones faciales y patrones de voz. De igual forma, en la Tabla 2.2, se muestran los trabajos multimodales que integran voz y visión.

Algunos puntos relevantes de estos trabajos son los siguientes:

- En general las tasas de reconocimiento son mayores en los sistemas de reco-

Figura 2.3: Perro Robótico AIBO de Sony

nocimiento basados en expresiones faciales en comparación con los sistemas basados en voz [2]. Sin embargo la integración de ambos tipos de sistemas puede presentar una tasa mayor que la de cualquier sistema independiente [85, 11].

- La mayoría de los trabajos realizados con expresiones faciales han sido desarrollados y evaluados con bases de datos especializadas como son FEEDTUM [10, 25], JAFFE [41, 89, 31, 70, 65], FACES [87], CK+ [18], y RaFD [35]. Estos trabajos también reportan las tasas más altas de reconocimiento de emociones.

- Mientras que la mayoría de los sistemas basados en expresiones faciales considera un número de seis emociones [10, 35, 70, 89, 31, 18] para los sistemas basados en voz el número considerado es de cuatro [94, 13, 98, 44, 26] y cinco [16, 4, 99].

- La mayoría de las bases de datos de voz usadas para reconocimiento de emociones se encuentran en idiomas extranjeros como idiomas Asiáticos [19, 98, 99, 26], Danés [16], Alemán [1, 79] e Inglés [44, 7, 79].

- Para los desarrollos y pruebas de sistemas de visión se han considerado datos de un usuario [40, 50], 10 usuarios [65, 41, 89, 31, 70], 20 usuarios [69] y más de 20 usuarios[18, 35]. Como se presentó en la Tabla 2.1 de manera general los trabajos reportados en la literatura han considerado a 10 usuarios.

- Para los desarrollos y pruebas de sistemas de voz hay trabajos que han utilizado a

Tabla 2.1: Trabajos Relevantes en Reconocimiento de Emociones: Ne=Neutro, Mi=Miedo, Fe=Felicidad, Tr=Tristeza, En=Enojo, Di=Disgusto, So=Sorpresa, Ab=Aburrimiento.

Trabajo	Emociones	Patrón	Base de Datos	Técnica de Reconocimiento	Desempeño
[40]	7 (Ne, Mi, Fe, Tr, En, Di, So)	Rostro	Propia (1 usuario)	ANN	83.57 % - 85.13 %
[50]	7 (Ne, Mi, Fe, Tr, En, Di, So)	Rostro	Propia (1 usuario)	SVM	93.75 %
[25]	7 (Ne, Mi, Fe, Tr, En, Di, So)	Rostro	FEEDTUM (18 usuarios)	ANN	70.00 %
[65]	4 (Ne, Mi, Fe, En)	Rostro	JAFFE (10 usuarios)	ANN	90.00 %
[41]	5 (Fe, Tr, En, Di, So)	Rostro	JAFFE (10 usuarios)	PCA (Dist. Eucl.)	80.00 %
[89]	6 (Mi, Fe, Tr, En, Di, So)	Rostro	JAFFE (10 usuarios)	PCA (Dist. Eucl.)	91.16 %
[31]	6 (Mi, Fe, Tr, En, Di, So)	Rostro	JAFFE (10 usuarios)	PCA (Dist. Eucl.)	91.19 %
[69]	5 (Ne, Mi, Fe, Tr, En)	Rostro	Propia (20 usuarios)	ANN	87.00 %
[87]	7 (Ne, Mi, Fe, Tr, En, Di, So)	Rostro	FACES Collection	PCA (Dist. Eucl.)	96.66 %
[18]	6 (Mi, Fe, Tr, En, Di, So)	Rostro	Cohn-Kanade CK+ (123 usuarios)	Lógica Difusa	87.67 %
[62]	6 (Mi, Fe, Tr, En, Di, So)	Rostro	FEEDTUM (18 usuarios)	Lógica Difusa	89.33 %
[35]	6 (Mi, Fe, Tr, En, Di, So)	Rostro	RaFD (67 usuarios)	Lógica Difusa	93.96 %
[70]	6 (Mi, Fe, Tr, En, Di, So)	Rostro	JAFFE (10 usuarios)	Lógica Difusa	92.33 %
[19]	6 (Mi, Fe, Tr, En, Di, So)	Voz	BHUDES (15 usuarios, Chino Mandarín, 20 frases/emoción)	SVM, ANN	39.16 % - 50.00 %
[16]	5 (Ne, Mi, Fe, Tr, En)	Voz	DES (4 usuarios, Danés, 88 frases)	SVM	64.77 % - 79.55 %
[98]	4 (Ne, Fe, Tr, En)	Voz	Material de TV (Chino Mandarín, 721 frases)	SVM	74.28 %
[4]	5 (Ne, Mi, Fe, Tr, En)	Voz	Propia (4 usuarios, Alemán, 260-280 frases)	Lógica Difusa	84.00 % (Dep.), 60.00 % (Indep.)
[13]	4 (Ne, Fe, Tr, En)	Voz	Propia (6 usuarios, Español, 10 frases/emoción)	HMM	94.32 %
[79]	7 (Ne, Mi, Fe, Tr, En, Di, So)	Voz	Propia (5 usuarios, Alemán e Inglés, 100 frases/emoción)	HMM	86.00 %
[63]	5 (Ne, Ab, Fe, Tr, En)	Voz	Propia (34 usuarios, Chino Mandarín, 3400 frases)	HMM	62.50 %
[100]	7 (Ne, Ab, Mi, Fe, Tr, En, Di)	Voz	BERLIN (10 usuarios, Alemán, 10 frases/emoción)	HMM	89.00 %
[99]	5 (Ne, Mi, Fe, Tr, En)	Voz	Propia (Chino Mandarín)	HMM	87.00 %
[44]	4 (Ne, Fe, Tr, En)	Voz	Propia (1 usuario, Inglés, 151-263 frases/emoción)	HMM	76.12 %
[80]	5 - 7	Voz	Varios Corpora (Cruzamiento de 6 Corpora)	SVM	81.00 %
[92]	7 (Ne, Ab, Mi, Fe, Tr, En, Di)	Voz	BERLIN (10 usuarios, Alemán, 10 frases/emoción), SUSAS (32 usuarios, Inglés)	SVM, HMM	90.00 % (BERLIN), 83.00 % (SUSAS)
[94]	4 (Ne, Fe, Tr, En)	Voz	Propia (2033 frases)	GMM, SVM, ANN	72.61 % (GMM), 75.33 %-78.16 % (SVM), 69.86 %-71.87 % (ANN)
[26]	4 (Ne, Fe, Tr, En)	Voz	Propia (Malabari, 700 frases)	ANN	55.00 %-68.50 %

Tabla 2.2: Sistemas de Reconocimiento Multimodales de Emoción: Ne=Neutro, Mi=Miedo, Fe=Felicidad, Tr=Tristeza, En=Enojo, Di=Disgusto, So=Sorpresa, Ab=Aburrimiento.

Trabajo	Emociones	Patrón	Base de Datos	Técnica de Reconocimiento	Desempeño
[11]	4 (Ne, Fe, En, Tr)	Rostro-Voz	Propia (1 actriz con 612 frases)	PCA, SVM	Voz: 71.00 %, Vision: 85.00 %, Multimodal: 89.00 %
[33]	7 (Ne, Mi, Fe, Tr, En, Di, So)	Rostro-Voz	Propia (1 hombre, 120 frases)	MFCC, PCA, LDA Gaussianas	Voz: 53.00 %, Visión: 98.00 %, Multimodal: 98.00 %
[93]	6 (Ne, Mi, Fe, Tr, En, So)	Rostro-Voz	Propia (500 videos, 8 usuarios)	MFCC, LDA Fisher	Multimodal: 82.00 %
[78]	3 (Fe, Tr, En)	Rostro-Voz	Propia (10 mujeres, 11 hombres, 10.5 hrs conversación humano-humano espontánea)	SVM	Multimodal: 64.00 %
[84]	7 (Ne, Mi, Fe, Tr, En, Di, So)	Rostro-Voz	Propia (1384 muestras)	THMM	Multimodal: 85.00 %

un usuario [44], cuatro a seis usuarios [16, 4, 13, 79], 10 a 15 usuarios [19, 100, 92] y 34 usuarios [63]. Se aprecia una mayor cantidad de trabajos que han considerado de cuatro a 10 usuarios.

- Para los trabajos que han desarrollado sistemas multimodales (ver Tabla 2.2) se han utilizado bases de datos con un usuario [11, 33], ocho usuarios [93] y 21 usuarios [78].

Estos puntos representan áreas de investigación que pueden extenderse con técnicas, recursos, analisis, condiciones y alternativas de desarrollo adicionales. El presente trabajo de tesis puede contribuir en este aspecto mediante las siguientes propuestas:

- Desarrollo de un sistema de reconocimiento multimodal de emociones con usuarios mexicanos.

- Establecer los parámetros de una base de datos emocional audio-visual para el desarrollo y análisis de desempeño del sistema de reconocimiento multimodal. La base de datos creada para el presente trabajo considera participantes mexicanos de la región Centro-Sur del país.

- Integración de GAs para encontrar las estructuras de HMMs y ANNs más apropiadas para el desarrollo del sistema de reconocimiento multimodal.

- Evaluación de desempeño con pruebas en vivo con usuarios mexicanos diferentes de aquellos considerados para la creación de la base de datos audio-visual. Se

establece una tasa de reconocimiento objetivo mayor de 95.00 %.

- Administración del sistema de reconocimiento multimodal de emociones mediante el desarrollo de un sistema de diálogo con un robot humanoide.

2.2. Técnicas de Desarrollo

2.2.1. Análisis de Componente Principal (Principal Component Analysis, PCA)

La técnica de Análisis de Componente Principal es un tipo de método analítico basado en subespacios el cual puede estimar datos en su forma original a partir de vectores característicos con pequeña dimensionalidad. PCA puede remover ruido e información redundante mediante la búsqueda de los elementos y estructuras más importantes de los datos originales. De igual manera PCA tiene las siguientes ventajas [50]:

- Es simple de implementar.
- No tiene límite de parámetros.
- La resolución de los datos originales en el espacio dimensional reducido es conveniente.

En el campo del reconocimiento de patrones, PCA se ha utilizado como técnica de reducción de dimensionalidad y extracción de características [50, 25, 87, 19]. Sin embargo también se ha utilizado como técnica de reconocimiento de patrones [41, 89, 31, 90]. Particularmente para el reconocimiento de rostros PCA es la base de la técnica de Eigenrostros [90, 83, 29].

En este contexto la técnica de PCA para reducción de dimensionalidad y reconocimiento de patrones faciales se presenta a continuación [90]:

1. Considere un conjunto de J imágenes de entrenamiento en donde cada imagen tiene dimensión n$\times$m = N pixeles.
2. Cada j-ésima imagen se convierte en un vector Γ_j de dimensión $1\times N$:

$$[\Gamma_1 \ \Gamma_2 \ , ..., \ \Gamma_J] \tag{2.1}$$

3. Una vez que se tienen todas las imágenes como vectores Γ_j se encuentra el rostro promedio Ψ el cual está definido por:

$$\Psi = \frac{1}{J}\sum_{j=1}^{J}\Gamma_j \tag{2.2}$$

4. A cada vector Γ_j se le resta el vector promedio Ψ:

$$\phi_j = \Gamma_j - \Psi \tag{2.3}$$

   Los vectores ϕ_j se almacenan en una matriz A la cual tendrá dimensiones $N \times J$:

$$A = [\phi_1 \ \ \phi_2 \ \ , ..., \ \ \phi_J] \tag{2.4}$$

5. De la matriz A, la matriz de covarianza L es obtenida como [83]:

$$L = A^T A \tag{2.5}$$

   Esta matriz es usada para obtener los eigenvectores $v = eig(L)$ los cuales son la base para los eigenrostros definidos por $u = Av$.

6. Finalmente al considerar R como el número de eigenrostros (aquellos con los eigenvalores más altos) un nuevo rostro Γ puede ser transformado en sus componentes de eigenrostro de la siguiente manera:

$$\Omega = u^T(\Gamma - \Psi) = [\omega_1 \ \ \omega_2 \ \ , ..., \ \ \omega_R]^T \tag{2.6}$$

7. La reducción de dimensionalidad es lograda al ser la dimensión de Ω igual al número de muestras de entrenamiento (J) [29]. También Ω representa las características de las muestras de entrenamiento las cuales pueden ser usadas para reconocimiento: los pesos ω_r describen las contribuciones de cada eigenrostro en la representación de la imagen del rostro de entrada. Este vector puede ser usado para reconocimiento de rostros/emociones encontrando la distancia euclidiana e más pequeña entre los vectores de pesos del rostro de entrada y los rostros de entrenamiento de la siguiente manera:

$$e = \|\Omega_{entrada} - \Omega\| \tag{2.7}$$

Si el índice de la imagen se asocia con una etiqueta que especifique la identidad del usuario entonces esto puede ser usado para el reconocimiento de rostros. Sin embargo, si la etiqueta de la imagen consiste de solamente el estado emocional expresado por el rostro (sin importar la identidad del usuario) entonces esto puede ser usado para el reconocimiento de emociones.

2.2.2. Redes Neuronales Artificiales (Artificial Neural Networks, ANNs)

Las Redes Neuronales Artificiales, o Artificial Neural Networks son una técnica de aprendizaje dentro del campo de la inteligencia artificial. Las ANNs consisten en modelos computacionales que están inspirados en las redes neuronales biológicas del Sistema Nervioso Central (SNC) de los seres vivos, en particular del cerebro. Las ANNs se presentan generalmente como sistemas de "neuronas" interconectadas que pueden estimar valores asociados a unas entradas, y que son capaces de aprender y reconocer patrones [6]. Por lo tanto las ANNs pueden verse como un procesador masivo paralelo, distribuido, hecho de procesadores simples que son naturalmente capaces de almacenar conocimiento obtenido de la experiencia y hacerlo útil.

En la Figura 2.4 se presentan las partes de una neurona humana los cuales tienen repesentación en el ámbito de las neuronas artificiales. Como se presenta en la Figura 2.5 una neurona artificial recibe una o más entradas (que representan una o más dendritas) las cuales se integran para producir una salida (que representa el axón de la neurona). Cada entrada (X_i) tiene asociada un peso (W_i) y la integración de las entradas se da mediante la suma ponderada de las mismas. Esta suma entonces se pasa a través de una función conocida como función de activación o de transferencia [6].

En la Figura 2.5 se presenta el tipo de ANN más sencillo que es el Perceptrón. Con este tipo de ANN se pueden clasificar elementos que pertenezcan a dos clases linealmente separables. De acuerdo con esta información se tienen los siguientes elementos básicos de una ANN:

- Neuronas y pesos (ponderaciones).
- Reglas de activación para cada neurona:

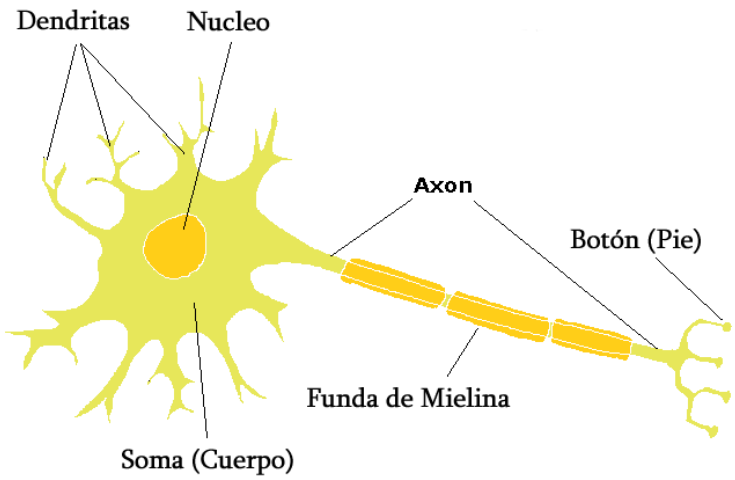


Figura 2.4: Partes de una neurona humana.

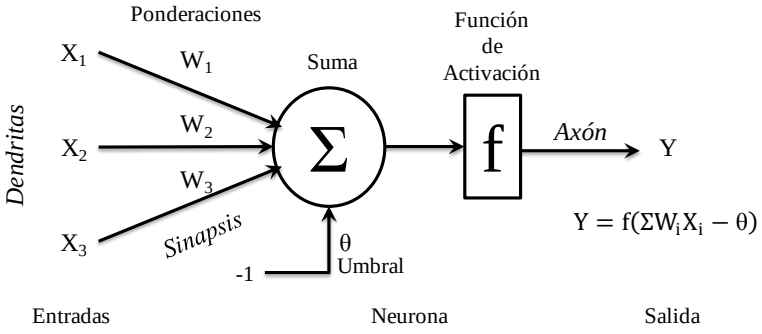


Figura 2.5: Perceptrón simple con una neurona y tres Entradas.

- Lineal:

$$f(x) = kx, \text{ en donde } k = \text{constante}. \tag{2.8}$$

- Escalón:

$$f(x) = \begin{cases} 1 & \text{si } x > t \\ 0 & \text{si } x \leq t \end{cases} \tag{2.9}$$

- Sigmoide:

$$f(x) = \frac{1}{1 - e^{-x}}. \tag{2.10}$$

- Tangente Hiperbólica:

$$f(x) = tanh(x). \tag{2.11}$$

- Topografía o Topología: Reglas de interacción entre neuronas. Entre estas se pueden mencionar las siguientes: (a) Una Capa, (b) Multicapa, (c) Recurrente, y (d) Mixta. Un ejemplo de una ANN multicapa (MultiLayer Perceptron, MLP) se presenta en la Figura 2.6. Este tipo de ANN tiene capas ocultas (hidden layers) entre las capas de neuronas para la(s) entrada(s) y la(s) salida(s). Una ANN MLP puede clasificar entradas que pertenezcan a dos o más clases que no sean linealmente separables.

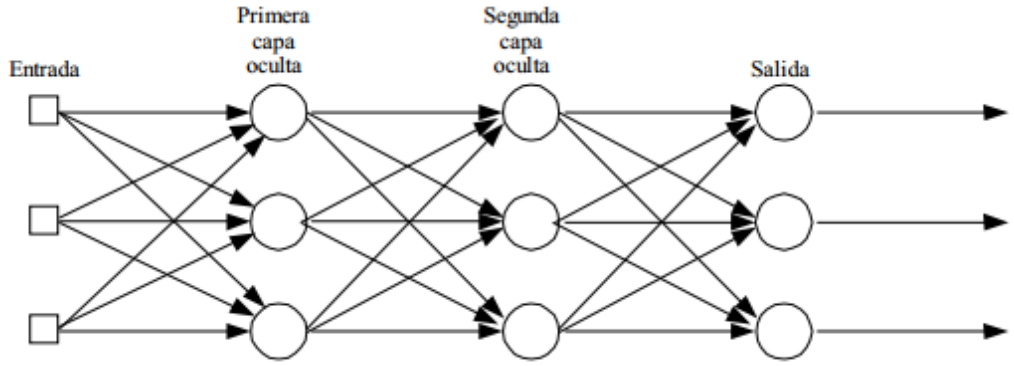


Figura 2.6: Perceptrón multicapa.

- Reglas de aprendizaje: Estas reglas o métodos están enfocados a ajustar las ponderaciones asociadas a cada neurona para que un grupo de datos de entrada produzca un grupo deseado de datos de salida. Existen dos tipos de aprendizaje o entrenamiento: (a) supervisado y (b) no supervisado.

En el aprendizaje supervisado los datos de entrada y la salida deseada para esa información es conocida. En el no supervisado no existe la salida deseada por lo que se tienen clases, que son grupos diferenciables entre sí, en donde a los valores de entrada corresponden ponderaciones asignadas por el sistema de acuerdo a su selección.

Entre los métodos de aprendizaje se pueden mencionar al aprendizaje de Boltzmann (Boltzmann Learning, BL), aprendizaje Hebbiano (Hebbian Learning, HL), y aprendizaje Competitivo (Competitive Learning, CL) [6]. Sin embargo el algoritmo más común para ANNs del tipo MLP es el de Retropropagación (Back-Propagation, BP) y el de Levenberg-Marquardt (LM). Más información acerca de los tipos de ANNs y los algoritmos de aprendizaje se pueden consultar en [6, 75]

- Información acerca del entorno (contexto) en donde las ANNs van a ser utilizadas (datos de entrenamiento). Entre los datos de entrada y de salida utilizados para el entrenamiento se pueden incluir números reales, enteros y binarios.

Finalmente entre las características importantes de las ANNs se pueden mencionar las siguientes:

- Aprendizaje: Una ANN puede modificar su comportamiento y reacción al entorno.
- Generalización: La ANN, una vez entrenada, puede ser tolerante a cambios en sus entradas.
- Abstracción: Una ANN es capaz de encontrar las características principales en un conjunto de datos.

2.2.3. Modelos Ocultos de Markov (Hidden Markov Models, HMMs)

Los Modelos Ocultos de Markov, o Hidden Markov Models son una técnica de reconocimiento de patrones utilizada dentro del campo del reconocimiento de voz. Un HMM es un modelo estocástico en el cual el sistema modelado se asume que es un proceso de Markov, en donde los estados no son directamente visibles, pero variables influenciadas por los estados (las observaciones, patrones a ser clasificados) si son visibles [37, 67, 97]. Un HMM está representado por transiciones (arcos, a_{ij}) y estados (nodos, q_i) como se presenta en la Figura 2.7.

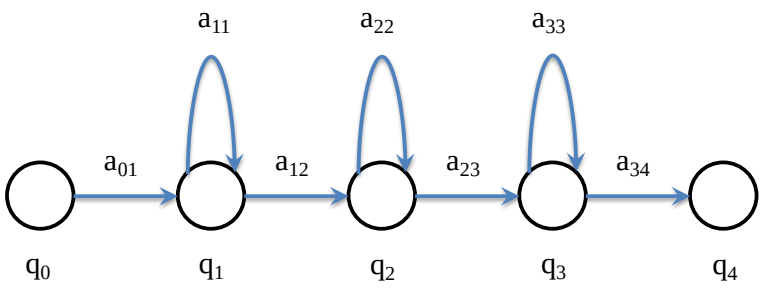


Figura 2.7: Modelo oculto de markov con estructura izquierda-a-derecha con tres estados emisores.

De manera formal los parámetros que definen a un HMM se denotan mediante la expresión $\lambda = (A, B, \pi)$ [37] en donde:

- $Q = \{q_0,\ q_1,\ ...,\ q_N\}$, define el conjunto de estados, en donde q_0 y q_N son estados no-emisores (no asociados con observaciones). Cada estado tiene asociado una función de probabilidad que modela la emisión/generación de ciertas observaciones (véase $B = \{b_i(\mathbf{o}_t)\}$).

- $A = \{a_{01},\ a_{02},\ ...,\ a_{NN}\}$, representa la matriz de probabilidades de transición A, en donde cada a_{ij} representa la probabilidad de moverse del estado i al estado j. $\sum_{j=1}^{N} a_{ij} = 1\ \forall i$.

- $O = \{o_0,\ o_1,\ ...,\ o_t\}$, define el conjunto de observaciones que le son dadas al Modelo de Markov para su entrenamiento.

- $B = \{b_i(\mathbf{o}_t)\}$, un conjunto de *Probabilidades de Observación*. Cada término representa la probabilidad de que un vector observado $\mathbf{o}_t$ sea generado o emitido por un estado i. El modelado de $b_j(\boldsymbol{o}_t)$ se hace por medio de una **Mezcla de Gaussianas** [37, 97]:

$$b_j(\mathbf{o}_t) = \sum_{k=1}^{K} C_{jk} N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \tag{2.12}$$

en donde K denota el número de componentes gaussianos, C_{jk} es el peso para la k-ésima distribución gaussiana que satisface $\sum_{k=1}^{K} C_{jk} = 1$, y $N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ denota a una sola función de densidad gaussiana con vector de media $\boldsymbol{\mu}_{jk}$ y matriz

de covarianza $\mathbf{\Sigma}_{jk}$ para el estado j. Esta gaussiana puede ser expresada como:

$$N(\mathbf{o}_t, \boldsymbol{\mu}_{jk}, \mathbf{\Sigma}_{jk}) = \frac{1}{\sqrt{(2\pi)^n \left|\mathbf{\Sigma}_{jk}\right|}} e^{-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{jk})' \mathbf{\Sigma}_{jk}^{-1} (\mathbf{o}_t - \boldsymbol{\mu}_{jk})} \tag{2.13}$$

en donde n es la dimensionalidad de $\mathbf{o}_t$, y (') denota la transpuesta del vector.

- $\pi = \{\pi_i\}$, una distribución inicial de estados, en donde $\pi_i = Pr(q_0 = i),\ 1 \leq i \leq N$, y $\sum_{i=1}^{N} \pi_i = 1$.

La estimación y ajuste de los parámetros de los HMMs para propósitos de reconocimiento de patrones se aborda mediante los siguientes problemas específicos:

- Problema de Decodificación o Búsqueda: Dada la secuencia observada $\boldsymbol{O}$ y el modelo $\lambda = (A, B, \pi)$, estimar la secuencia de estados Q que mejor describa las observaciones.

- Problema de Evaluación: Dada la secuencia observada $\boldsymbol{O}$ y el modelo (λ), estimar de manera eficiente la probabilidad de observar dicha secuencia dado el modelo ($Pr(\boldsymbol{O}|\lambda)$).

- Problema de Aprendizaje: Dada una secuencia de observaciones $\boldsymbol{O}$ de un conjunto de entrenamiento, estimar o ajustar A y B para describir con más precisión dicha información (maximizar $Pr(\boldsymbol{O}|\lambda)$).

Estos problemas se abordan mediante los algoritmos de Baum-Welch y Viterbi [67, 97] los cuales ajustan y evalúan los parámetros de los HMMs para determinar aquellos que mejor representen (o tengan mayor probabilidad de representar) observaciones presentes.

2.2.4. Máquinas de Estado Finito (Finite State Machines, FSMs)

Una Máquina de Estados de Estado Finito o FSMs por sus siglas en Inglés puede modelar procesos en donde una salida o resultado dependa del estado de una entrada o estímulo. Este tipo de FSM se conoce como Autómata de Estado Finito o Transductor de Estado Finito. De manera muy similar a una Cadena de Markov o HMM, un FSM tiene asociados estados y probabilidades de transición entre los mismos. De manera formal un FSM está definido por el siguiente conjunto de elementos [57, 56, 55]:

- Un FSM siempre parte de su estado inicial (s_0) y procesa una secuencia de símbolos de entrada Σ (alfabeto de entrada). El sistema contiene un conjunto finito de estados S.

- Los símbolos de Σ deben provenir de un conjunto finito. La interpretación de los símbolos depende de la tarea que se tenga que resolver.

- El FSM convierte los símbolos de Σ en términos de un conjunto de símbolos de salida Γ (alfabeto de salida). Los símbolos de Γ también deben provenir de un conjunto finito. Los estados con las salidas del FSM se denominan como estados finales (s_f).

- La conversión de símbolos $\Sigma \rightarrow \Gamma$ se determina mediante probabilidades o pesos de transición entre estados E.

- ϵ representa un símbolo vacío.

En la Figura 2.8(a) se muestra un ejemplo de un FSM cuyo proceso empieza en un estado inicial (0) y termina en cualquiera de los estados finales (1) o (2). Por convención el nodo que representa el estado inicial se representa como una circunferencia mientras que los estados finales son representados con dobles circunferencias [55]. El FSM acepta el símbolo de entrada $i \in \Sigma$ y lo traduce a uno de salida $\{o_1, o_2\} \in \Gamma$ dependiendo del costo dado por los pesos $\{w_1, w_2\} \in E$. Ésto se representa mediante $i : o_1/w_1, i : o_2/w_2$. En la Figura 2.8(b) se presenta una aplicación del FSM para la traducción de la palabra *PLANE* en inglés a español. Con una probabilidad o peso de 0.60 el FSM determina que esta palabra significa *AVIÓN* en español mientras que con el 0.40 determina que la palabra significa *PLANO*.

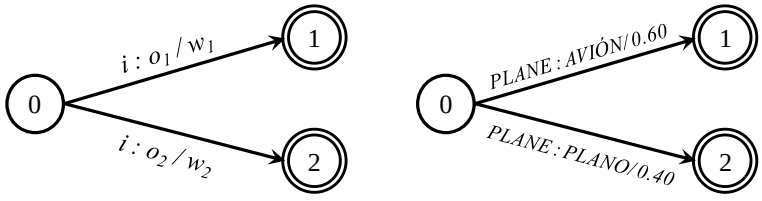


Figura 2.8: Ejemplos de máquina de estado finito para traducción.

Dadas estas características un FSM se puede utilizar para el desarrollo de sistemas de diálogo en donde una entrada (palabra o secuencia de palabras de estímulo) tiene asociada una acción (palabras de respuesta). Por ejemplo, la secuencia de entrada "¿Cómo estas?" tiene asignadas las siguientes respuestas: "Bien", "Muy bien", "Muy bien, gracias", "Ok", "Muy bien, ¿y tu?", "Mal", "Muy mal", "Estoy triste", etc.

Un sistema de diálogo tiene una amplia diversidad de respuestas las cuales dependen de los tipos y contextos de las entradas. Por lo tanto un sistema de diálogo se integra de una red de FSMs dado que para cada frase de entrada y salida se tiene un FSM. Esto sin embargo puede repercutir en tener una red de FSMs demasiado grande lo cual puede afectar la eficiencia del proceso de búsqueda de salidas correspondientes a un conjunto de entradas.

Para hacer más eficiente la red de FSMs se pueden aplicar las siguientes operaciones [55]:

- Determinización: El principal propósito de la determinización es eliminar transiciones redundantes en un FSM. Un FSM es determinístico si y sólo si cada uno de sus estados tiene a lo mucho una transición con cualquier tipo de etiqueta (o símbolo) de entrada. Esta etiqueta debe ser diferente de ϵ (vacío). De esta manera se reducen las transiciones con las mismas etiquetas.

- Minimización: Todo FSM determinístico se puede minimizar. Con esta operación se pueden eliminar estados redundantes. En un sistema de diálogo esta operación es útil cuando se tienen varias combinaciones de frases que pueden tener el mismo tipo de respuesta o etiqueta.

2.2.5. Algoritmos Genéticos (Genetic Algorithms, GAs)

Los Algoritmos Genéticos son técnicas de optimización heurística que se basan en el proceso natural de supervivencia y adaptación de los individuos más aptos en una población [30].

Los individuos con mejor aptitud para sobrevivir/adaptarse a un entorno se ganan un derecho (o son más probables) a reproducirse con otros individuos de igual o mejor aptitud. Estos individuos se convierten en "padres" de nuevas generaciones de "hijos" que heredarán las características que los hicieron más aptos para sobrevivir y adaptarse

al entorno. Estas características se van mejorando en cada ciclo de reproducción de manera generacional.

Dentro del contexto de los GA los individuos ("padres" e "hijos") representan soluciones viables para un problema en particular. En la Figura 2.9 se presentan los elementos principales de un GA los cuales se describen en las siguientes secciones.

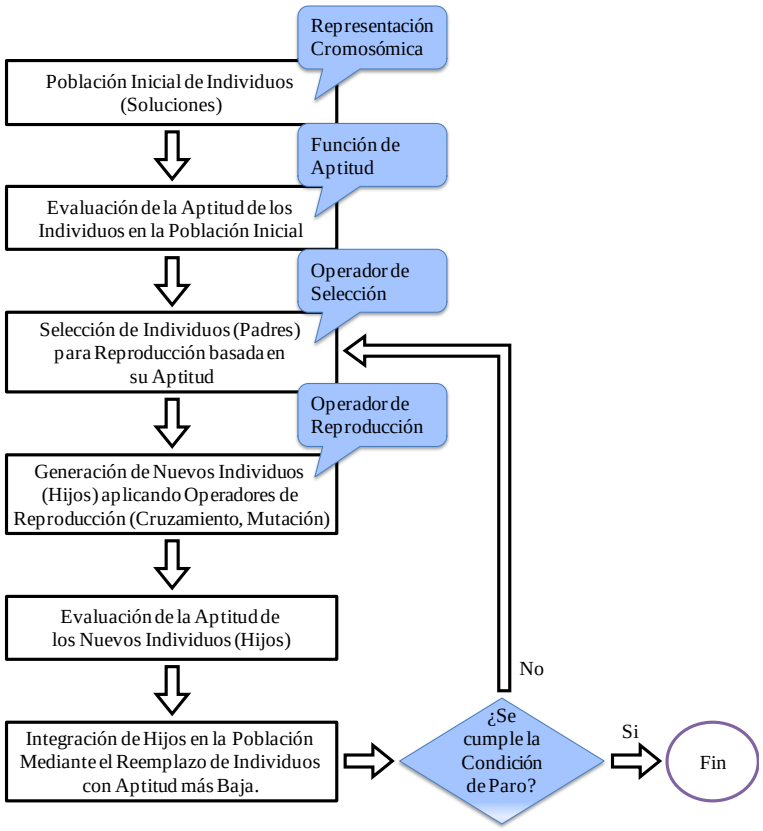


Figura 2.9: Diagrama de flujo de un algoritmo genético.

Representación Cromosómica

Inicialmente en un GA se tiene una Población Inicial de individuos los cuales representan soluciones progenitoras (x, padres). Estas soluciones sirven de base para generar soluciones descendientes (x', hijos) a través de operadores de reproducción.

Para realizar la reproducción, las soluciones deben representarse como un vector de variables de manera similar a un cromosoma con secuencias de genes. Esta codificación usualmente se hace de manera binaria. En la Figura 2.10 se presenta un ejemplo de la

representación cromosómica para dos individuos (números enteros).

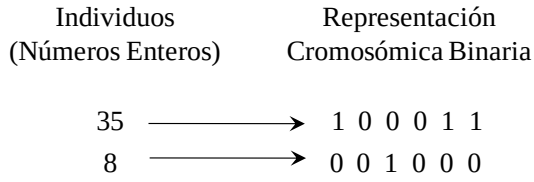


Figura 2.10: Ejemplo de representación cromosómica binaria.

La forma de representar los individuos en su forma cromosómica depende del tipo de problema a solucionar. Por ejemplo, para un problema de ruta más corta el cromosoma puede consistir de la secuencia de puntos a visitar, y cada ruta (individuo) tendría una secuencia de puntos diferente. De igual manera si un problema tiene soluciones dadas por las variables x, y, z el cromosoma se puede representar mediante el vector **v** = [x, y, z].

Función de Aptitud

A cada individuo se le asigna un valor o puntuación relacionado con su capacidad para resolver el problema considerado. En la naturaleza esto equivale al grado de adaptabilidad o aptitud de un organismo para competir por recursos dentro su entorno o sobrevivir a las adversidades del mismo ("supervivencia de el más fuerte"). Cuanto mayor sea la adaptación de un individuo a un entorno mayores serán sus oportunidades (probabilidades) para reproducirse.

Este grado o valor de aptitud depende del tipo de problema a resolver. Usualmente una "función objetivo" es considerada para evaluar la efectividad de una solución para resolver un problema. Para el caso de problemas de minimización y/o maximización la función objetivo determina el valor $f(x)$ asociado a un determinado individuo x. Si x aumenta el valor de $f(x)$ y el problema es de minimización entonces se trata de un individuo con poca aptitud. Sin embargo, si el problema es de maximización entonces x tiene buena aptitud.

Operador de Selección

El operador de selección tiene como objetivo determinar qué individuos se podrán reproducir basándose en sus valores de aptitud. Existe una diversidad de métodos para

realizar esta operación. Dentro de éstos se pueden mencionar los siguientes:

- Ruleta: En este método cada individuo tiene una parte de la ruleta, mayor o menor, en función de su valor de aptitud. Al hacer girar la ruleta (número al azar r) el individuo con mejor aptitud tendrá mayor probabilidad de ser seleccionado. El método se describe a continuación:

  1. Para cada individuo en la población x_i calcule el valor de aptitud f_i.
  2. Calcule la probabilidad de escoger cada individuo x_i como $p_i = \frac{f_i}{\sum_{k=1}^{N} f_k}$, en donde N es el tamaño de la población.
  3. Calcule la probabilidad acumulada q_i para cada individuo x_i como $q_i = \sum_{j=1}^{i} p_j$.
  4. Generar un número aleatorio uniforme $r \in \{0, 1\}$.
  5. Si $r < q_1$ entonces seleccione el primer individuo (x_1), en caso contrario seleccione el individuo x_i tal que $q_{i-1} < r \leq q_i$.
  6. Repetir los Pasos 4-5 N veces.

- Torneo: Este método implica la ejecución de varios "torneos" entre algunos individuos de la población elegidos al azar. El ganador de cada torneo (el que tiene el mejor valor de aptitud) se selecciona para reproducción. La presión de selección se ajusta cambiando el tamaño del torneo. El método se describe a continuación:

  1. Seleccione el tamaño del torneo (número de k individuos elegidos al azar de la población).
  2. Ordene de mejor a peor (mayor a menor valor de aptitud) los individuos en el torneo.
  3. La probabilidad de selección para cada individuo está dada por $q_i = p(1 - p)^m$ en donde $m = 0, ..., k - 1$.

Operadores de Reproducción

La reproducción se realiza mediante dos tipos de operadores: Cruzamiento (Cruce) y Mutación. Mediante estos operadores se realiza la recombinación y cambios de genes

(valores en el cromosoma) de los individuos seleccionados (padres) para la creación de nuevos individuos (hijos). La aplicación de estos operadores depende de la codificación del cromosoma y por ende del tipo de problema. Dentro de los métodos más utilizados para la reproducción se tienen los siguientes:

- Cruzamiento. Mediante este tipo de operador se da la recombinación de información genética de los individuos. Este operador fomenta la exploración del espacio de solución con el fin de evitar el encontrar soluciones óptimas locales. En la Figura 2.11 se presentan los métodos de cruzamiento más utilizados.

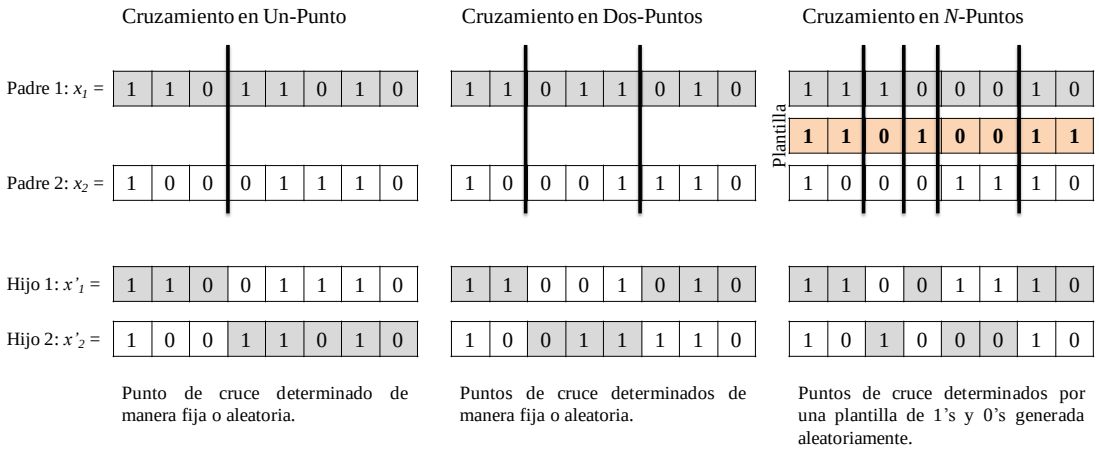


Figura 2.11: Métodos de cruzamiento binario.

- Mutación. Con este operador se pueden obtener nuevos individuos mediante pequeños cambios (mutaciones) en soluciones progenitoras. La mutación fomenta la explotación del espacio de solución adyacente a soluciones con altos valores de aptitud. En la Figura 2.12 se presentan los métodos de mutación más utilizados.

Condición de Paro

No hay una condición de paro específica para el GA. En caso de conocer el valor de aptitud óptimo el GA puede detener su ejecución al encontrar la solución cuya aptitud tenga este valor. De igual manera se puede detener si después de un número determinado de generaciones o iteraciones del GA ya no hay cambios significativos en el mejor individuo encontrado.

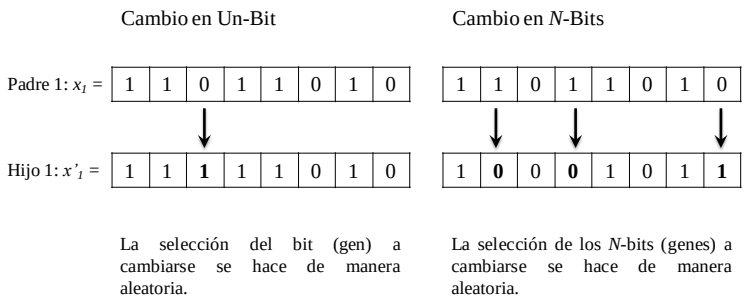


Figura 2.12: Métodos de mutación binaria.

2.2.6. Robot Humanoide Bioloid Premium

Uno de los sistemas robóticos más usados en investigación debido a su bajo costo en comparación con otras plataformas es el Bioloid Premium de la compañía Sur-Coreana Robotis. Este sistema se puede configurar en diferentes formas como se muestra en la Figura 2.13. En el presente trabajo la forma humanoide es de particular interés.

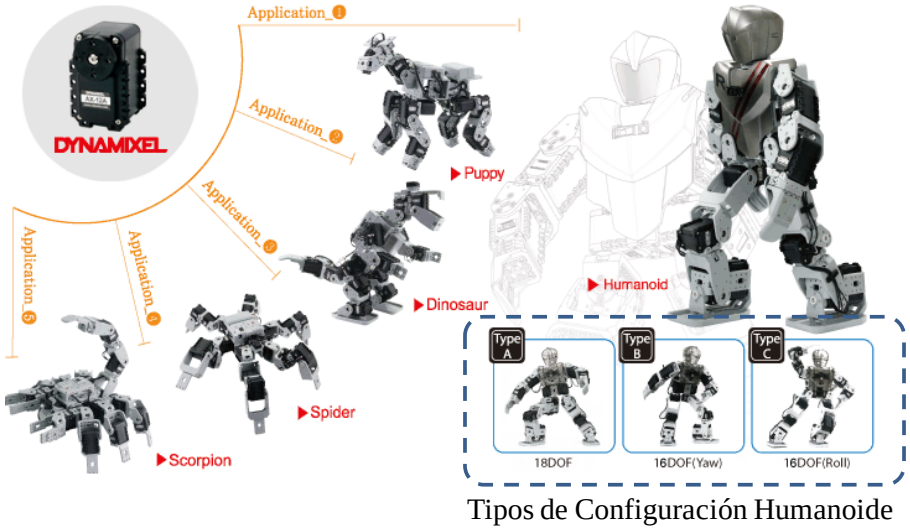


Figura 2.13: Configuraciones del sistema Bioloid.

El sistema Bioloid en configuración humanoide es uno de los más usados en las competencias de robots de pelea y la RoboCup [74]. Este sistema robótico cuenta con un CPU CM-530 basado en el controlador ARM Cortex (32 bits). Como se presenta en la Tabla 2.3 dependiendo del tipo de configuración humanoide (A, B o C) el sistema Bioloid puede constar de 16 o 18 servomotores dynamixel modelo AX-12A con par

de torsión de 15.3 Kg·cm (a 11.1 V). En este trabajo se seleccionó la configuración humanoide Tipo A. Esto dado que el robot puede tener una mayor movilidad en las piernas debido a sus 2 GDL (Grados de Libertad) extra sobre las otras configuraciones.

Tabla 2.3: Tipos de configuración humanoide del sistema Bioloid.

Configuración	Características
Tipo A	18 servomotores, Peso: 1.7 Kgs. Altura: 39.7 cm. Pasos estables y omnidireccionales
Tipo B	16 servomotores, Peso: 1.6 Kgs. Altura: 39.7 cm. Capacidad de cambiar de dirección mientras marcha y 2 servomotores disponibles para otras funciones
Tipo C	16 servomotores, Peso: 1.6 Kgs, Altura: 38.6 cm. Capacidad de ir de lado mientras marcha y 2 servomotores disponibles para otras funciones

Además de los servomotores dynamixel y su controlador ARM el sistema cuenta con los siguientes dispositivos:

- Transmisores Zigbee modelos zig-100 y zig-110A: Los transmisores sirven para operar al robot inalámbricamente. Estos transmisores operan entre 2.7 V y 3.6 V y pueden alcanzar velocidades de transmisión entre 9600 bps y 115200 bps. Permiten una comunicación UART usando una frecuencia de 2.4 GHz. Más información de estos dispositivos se puede encontrar en su manual [72].

- Control remoto RC-100: Este control permite operar al robot de manera inalámbrica mediante los transmisores de radio frecuencia zig-100 o el dispositivo bluetooth BT-100. El control cuenta con 11 botones en total, 10 para el control del robot y uno para encendido.

- Sensores: Cuenta con giroscopio de dos ejes el cual es usado para ajustar la postura del robot mientras camina. Cuenta también con un sensor DMS para medir distancias entre un rango de 10.0 cm a 80.0 cm, sensores infrarrojos para detectar objetos, y un receptor IR para recibir comandos del control remoto.

Capítulo 3

Base de Datos Emocional

Para poder desarrollar sistemas de reconocimiento de emociones se necesitan bases de datos apropiadas. Esto dada la etapa de aprendizaje de patrones inherente en el desarrollo del sistema de reconocimiento. Como se presentó en la Tabla 2.1 en cuanto al reconocimiento de emociones en expresiones faciales la mayoría de los trabajos realizados utilizan bases de datos estándar como JAFFE y FEEDTUM. Son pocos los trabajos de investigación que hacen uso de información propia para sus desarrollos y experimentos (o que evalúan con patrones en-vivo).

Respecto a los trabajos realizados en reconocimiento de emociones por voz, la mayoría de las bases de datos que existen actualmente están desarrolladas en otros idiomas (p.e., Inglés, Francés y Alemán) [20]. Estos recursos emocionales no pueden adaptarse fácilmente para el desarrollo de sistemas para personas mexicanas cuya lengua es el español dadas las diferencias de pronunciación (fonética) y acentos.

Una situación general de ambos tipos de recursos emocionales (expresiones faciales y voz) es que usualmente son creados bajo condiciones ideales de poco ruido, postura, iluminación y claridad. Estas condiciones no aplican en ambientes más cotidianos (por ejemplo, un salón de clases).

Debido a esta situación se creó una base de datos de patrones visuales (expresiones faciales) y vocales (voz emotiva) para el desarrollo del sistema multimodal propuesto. Condiciones ambientales de iluminación y ruido menos estrictas para la captura de los datos fueron consideradas con el fin de contar con datos más reales.

En las Tablas 2.1 y 2.2 (Capítulo 2) se presentaron comparativas de diversos trabajos

enfocados en el reconocimiento de emociones en voz, expresiones faciales, y ambos (multimodal). Estas comparativas, además de presentar el porcentaje de desempeño en el reconocimiento de emociones, presentaron información acerca de las bases de datos que se utilizaron para la obtención de sus resultados. Un dato importante es el número de usuarios presentes en dichas bases de datos. Tomando como referencia dichos trabajos previos, la mayoría de los sistemas realizados en visión consideraron entre uno y diez usuarios [40, 50, 65, 41, 89, 31, 70]. Los sistemas realizados en voz consideraron de cuatro a seis usuarios [16, 4, 13, 79], y 10 a 15 usuarios [19, 100, 92]. La mayoría de los sistemas multimodales han considerado entre uno y ocho usuarios [11, 33, 93].

Dado que la mayoría de los trabajos previos consideraron bases de datos con aproximadamente 10 usuarios, se determinó este número de usuarios como apropiado para la base de datos con usuarios mexicanos. En el presente trabajo de tesis se consideraron ocho usuarios para la base de datos de voz (MX-Voz) y nueve usuarios para la base de datos de expresiones faciales (MX-Expresiones). Sin embargo, para pruebas en-vivo se recolectaron datos de 10 usuarios diferentes de aquellos de las bases de datos. Al considerar estos nuevos datos (voz y expresiones faciales) con los de la base de datos inicial se tiene un total de 18 usuarios para los experimentos con voz, y 19 usuarios para los experimentos con expresiones faciales (visión). En las siguientes secciones se presentan los detalles de las bases de datos MX-Voz y MX-Expresiones.

3.1. Base de Datos de Expresiones Faciales Mexicana (MX-Expresiones)

El contar con una base de datos de expresiones faciales de personas Mexicanas fue importante para el desarrollo del presente trabajo. Esto porque las expresiones para la misma emoción pueden ser diferentes entre culturas (considerando las bases de datos disponibles como JAFFE y FEDDTUM) y personas. También porque las características faciales entre personas de diferentes nacionalidades podrían afectar el desempeño del sistema para un determinado usuario.

Para tener muestras representativas de rostros de personas para la base de datos de expresiones faciales Mexicana (MX-Expresiones) se reclutaron personas de las regiones

Este y Sur-Oeste de México. Para el desarrollo de la base MX-Expresiones se tomó como referencia la base de datos JAFFE (Japanese Female Facial Expression) de personas Japonesas la cual ha sido ampliamente usada para el reconocimiento de emociones (véase Tabla 2.1). La base JAFFE consiste de 213 imágenes de 10 mujeres Japonesas que presentan aproximadamente tres diferentes expresiones para cada una de las siguientes emociones: Enojo, Felicidad, Neutro, Tristeza, Disgusto, Miedo y Sorpresa. Las imágenes se encuentran en escala de grises con un tamaño de 256×256 pixeles en formato ".TIFF". En la Figura 3.1 se muestran algunos ejemplos de esta base de datos.

Figura 3.1: Imágenes de muestra de la base de datos de expresiones faciales JAFFE.

Para la base de datos MX-Expresiones nueve participantes mexicanos (tres hombres y seis mujeres) fueron reclutados. Estos participantes no fueron actores profesionales y sus datos generales se presentan en la Tabla 3.1.

Tabla 3.1: Perfiles de los participantes de la base de datos MX-Expresiones.

Usuario	Au	Me	Lu	Je	Ta	Ne	Mi	Fe	Ke
Género (Hombre, Mujer)	M	M	H	H	M	M	H	M	M
Edad (Años)	53	20	25	21	32	32	30	12	10
Estado de Origen	Veracruz	Veracruz	Veracruz	Veracruz	Oaxaca	Oaxaca	Oaxaca	Veracruz	Veracruz

La inducción de la emoción fue realizada mediante el uso de escenarios descritos por textos emocionales [20]. Estos textos fueron diseñados para la base de datos de voz emocional mexicana MX-Voz que se presenta en la Sección 3.2 . De esta manera los participantes expresaron las emociones de Enojo (EN), Felicidad (FE), Neutro (NE) y Tristeza (TR) [44, 98]. Para tener consistencia con la base de datos JAFFE se tomaron

tres muestras para cada emoción. Las muestras fueron capturadas con un fondo blanco en condiciones estándar de iluminación [48]. El equipo utilizado para la captura fue la cámara integrada en una computadora laptop con una resolución de 640×480 pixeles en formato ".TIFF" a color. De esta manera la base de datos MX-Expresiones se integró de 108 imágenes (9 participantes × 3 muestras × 4 emociones). Algunos ejemplos de esta base de datos se presentan en la Figura 3.2.

Figura 3.2: Imágenes de muestra de la base de datos de expresiones faciales mexicana (MX-Expresiones).

3.2. Base de Datos de Voz Emocional Mexicana (MX-Voz)

Para la creación del corpus emocional de voz mexicano las siguientes condiciones fueron consideradas [13, 46]:

- Estímulo textual de diferentes longitudes para cada emoción.
- Significancia semántica de los estímulos textuales.
- Debe haber suficientes ocurrencias de todas las vocales con todas las emociones consideradas en el texto de estímulo.

Los voluntarios para la base de datos emocional estuvieron dentro del grupo de edades de los 16 a los 53 años y no fueron actores profesionales. Para tener una pronunciación estándar mexicana estos voluntarios fueron reclutados de las regiones este y sur-oeste de México. Finalmente un total de cinco mujeres y tres hombres fueron considerados para el corpus emocional de voz. Sus datos generales se presentan en la Tabla 3.2.

Tabla 3.2: Perfiles de los participantes de la base de datos MX-Voz.

Usuario	Au	Me	Lu	Je	Li	Ta	Ne	Mi
Género (Hombre, Mujer)	M	M	H	H	M	M	M	H
Edad (Años)	53	20	25	21	16	32	32	30
Estado de Origen	Veracruz	Veracruz	Veracruz	Veracruz	Veracruz	Oaxaca	Oaxaca	Oaxaca

3.2.1. Estímulo Textual Emocional

Previo a la grabación de las muestras de voz para la base de datos el estímulo textual para cada emoción fue diseñado. Esto fue importante para tener muestras de voz con la entonación emocional apropiada. El texto de estímulo para Enojo, Felicidad y Tristeza consistió de frases que fueron concebidas en el contexto de situaciones de la vida cotidiana. Para Neutro las frases fueron consideradas de cultura general. En total se diseñaron 20 frases para cada emoción, tomando como referencia las bases BERLIN (10 usuarios, 10 frases/emocion) y BHUDES (15 usuarios, 20 frases/emocion) (ver Tabla 2.1), estas frases diseñadas se presentan en la Tabla 3.3.

Para asegurar el modelado acústico apropiado de las vocales mediante HMMs un mínimo de seis ocurrencias fue establecido. En la Tabla 3.4 el número de muestras por vocales para cada grupo de frases emocionales es presentado. Nótese que el mínimo es de 19 muestras (“u” con Tristeza) lo cual es mayor que el número mínimo considerado de seis ocurrencias.

Finalmente, en un salón a puerta cerrada, las frases fueron grabadas con la herramienta Wavesurfer [9] en formato “.WAV” con una frecuencia de muestreo de 48,000 Hz. En la Figura 3.3 se muestra la interfaz de este software.

La distancia entre el micrófono (micrófono interno de una computadora tipo laptop) y el usuario fue de alrededor de 60 cm. En la Figura 3.4 se muestra la posición adoptada por el participante al momento de la grabación.

Cada participante pronunció las 20 frases de cada emoción produciendo un total de 80 muestras de voz. De esta manera la base de datos MX-Voz consistió de 80 frases $\times$ 8 participantes = 640 muestras de voz emocionales.

Tabla 3.3: Frases de estímulo para los diferentes estados emocionales.

Frases para Enojo

1. Que yo no tuve la culpa.
2. Te digo que te vayas.
3. Mira… haz lo que quieras desde este momento.
4. Yo no te voy a estar soportando.
5. Nunca me incluyes en tus planes.
6. Ya me tienes harto, ya deja de hablar.
7. Ya no te quiero volver a ver.
8. El maestro nos reprobó a todos por tu culpa.
9. Jamás te vuelvas a acercar a mí.
10. Y ahora ¡¿qué es lo que quieres?!
11. Se me olvido mi dinero y tengo que regresarme otra vez.
12. ¡Firulais deja de estarme mordiendo!
13. Te explique muchas veces y no me entiendes.
14. Ya no me pidas dinero.
15. Me cancelaron mi servicio de cable y ya lo había pagado.
16. Esta es una situación que no está a discusión.
17. ¡Corre que la ayuda es urgente inútil !
18. Tú solo me usas a tu conveniencia.
19. Hubieras luchado un poco más, ¡pero no quisiste!
20. No hagas eso que estás de luto.

Frases para Neutro

1. Paris es la capital de Francia.
2. Estudio en la Universidad Tecnológica de la Mixteca.
3. Tengo la edad de 25 años.
4. El año tiene 4 estaciones.
5. La letra "j" es la única que no aparece en la tabla periódica.
6. La Mona Lisa es una obra de Leonardo da Vinci.
7. El océano Atlántico es más salado que el Pacífico.
8. El sistema solar tiene 8 planetas.
9. La araña Sidney es la más venenosa y puede matar un humano en 15 minutos.
10. El graznido de un pato no hace eco.
11. Los mosquitos tienen 47 dientes.
12. El monte Everest es el más alto del mundo.
13. No se puede estornudar con los ojos abiertos.
14. La velocidad del estornudo oscila entre 70 y 150 km/h.
15. El edificio Burj en Dubai es el más alto del mundo con 828 metros.
16. Uruguay es un país de América del Sur.
17. Un avestruz es un ave que no vuela.
18. Una úlcera gástrica ocurre en el estómago.
19. La urticaria es una enfermedad de la piel.
20. El universo es muy grande.

Frases para Felicidad

1. ¡Me gane un viaje todo pagado a Florida!
2. Saqué 10 de promedio en mi escuela.
3. ¡Me compre un billete de lotería y gané!
4. Estoy feliz porque me ascendieron de puesto.
5. ¡Tuve unas súper vacaciones!.
6. Que deliciosa estuvo mi comida.
7. Me regalaron pases dobles para ir al cine.
8. Ganamos el primer lugar en el torneo de futbol.
9. Me regalaron un auto en mi cumpleaños.
10. Gané una beca para estudiar en el extranjero.
11. ¡Me encontré 500 pesos tirados en la calle!
12. Ya casi es navidad y veré a toda mi familia junta otra vez.
13. Esta noche iré a cenar con mis amigos y después al cine.
14. Me regalaron el celular que tanto quería.
15. ¡Me regalaron una pantalla 3D extraordinaria!.
16. ¡Es un día fabuloso!
17. ¡Sucedió el milagro que esperaba!
18. Hemos logrado una victoria más.
19. Jugaremos todo el día, es genial.
20. Nadé con una tortuga galápagos.

Frases para Tristeza

1. Quisiera volver el tiempo para ver a mi familia.
2. Por más que se hizo no se pudo salvar mi perrito.
3. No entiendo porque ya no quieres verme.
4. Nadie me comprende ni me escucha.
5. Me siento solo, no tengo a nadie con quien platicar.
6. Mi vida es un fracaso… ya no puedo seguir.
7. Por favor dame otra oportunidad.
8. Quisiera que entendieras lo que yo siento.
9. Reprobé todas las materias y me sacarán de la escuela.
10. Mi mejor amigo acaba de fallecer ayer.
11. Acaban de atropellar mi gatito recién nacido.
12. Maté a mi perrito porque no tenía como alimentarlo.
13. Quisiera dormir y no despertar nunca.
14. Me haces mucha falta te extraño.
15. Tu recuerdo sigue aquí todavía.
16. Tuve una sensación como si estuvieras conmigo.
17. La vida es dura y el universo conspira contra mí.
18. Trato de unir los pedazos de mi corazón.
19. Siento que me hundo en un mar de lágrimas.
20. Tu crueldad está acabando con mi mundo.

Tabla 3.4: Registro de número de vocales por grupo de frases de estímulo.

Vocal	Enojo	Felicidad	Neutro	Tristeza
a	65	86	92	83
e	83	94	115	86
i	38	46	60	58
o	54	54	74	65
u	23	28	35	19

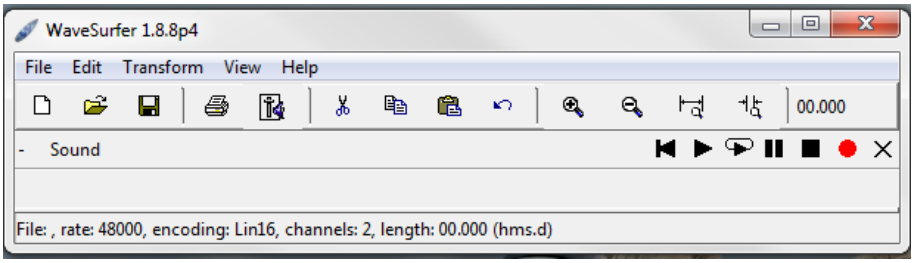


Figura 3.3: Herramienta Wavesurfer para grabación de muestras de voz.

Figura 3.4: Modelo de captura de frases entre usuario y computadora.

3.2.2. Etiquetado Ortográfico y Desarrollo de Transcriptor Fonético

Un paso importante en el desarrollo de un corpus de voz para entrenamiento de un sistema de reconocimiento es el etiquetado. Mediante este paso, también conocido como "transcripción" se relaciona en el tiempo el estímulo textual con las muestras de voz [97]. De esta manera se puede identificar el segmento de la señal de voz que representa una palabra escrita y así facilitar su aprendizaje por parte del sistema de reconocimiento.

El etiquetado ortográfico consiste en identificar los segmentos de la señal de voz que corresponden a palabras. En cambio, el etiquetado fonético consiste en identificar los sub-segmentos que forman una palabra, los cuales se conocen como fonemas. Un fonema representa una abstracción del sonido de un elemento del alfabeto (vocal o consonante). Por ejemplo, el sonido de la consonante "d" se describe con el fonema "/d/". En cambio "h" no tiene sonido en el español mexicano (por ende no hay fonema que la represente). Los sistemas de reconocimiento de voz se modelan a nivel fonema, de tal manera que con un conjunto finito de sonidos se pueden formar una amplia variedad de palabras.

El etiquetado ortográfico y fonético del corpus de voz emocional se realizó mediante la herramienta Wavesurfer [9]. El estilo de etiquetado (formato de tiempo para las etiquetas) fue el de "HTK Transcription". HTK (Hidden Markov Model Toolkit) [97] es la herramienta de software utilizada para la construcción del sistema de reconocimiento de voz. Esto se presenta en la Sección 4.1.

Si bien el etiquetado ortográfico es directo de realizar, el etiquetado fonético requiere de mayor cuidado. Esto dadas las variaciones de los sonidos de algunas consonantes las cuales dependen de la coarticulación de la voz. El idioma español mexicano está conformado por un total de 27 fonemas (22 consonantes + 5 vocales) los cuales pueden definir cualquier palabra en el idioma [64]. En la Tabla 3.5 se muestran estos fonemas.

Tabla 3.5: Fonemas del idioma español mexicano (nn=ñ).

Fonemas								
Z	_D	_G	_N	_R	a	b	d	e
f	g	i	k	ks	l	m	n	nn
o	p	r	r(	s	t	tS	u	x

En la literatura se ha encontrado que las propiedades espectrales de los sonidos de las vocales son un indicador confiable de las emociones en la voz [44, 45]. Debido a esto las vocales pueden ser usadas para el reconocimiento de emociones si se les considera fonéticamente independientes en la creación de un sistema de reconocimiento de voz estándar [13]. De esta forma cuando uno vocaliza una emoción, los fonemas principalmente afectados corresponden a las vocales, por lo que una vocal "a" expresada con Enojo puede ser diferente de una "a" expresada con Tristeza o Felicidad. Esto permite el modelado acústico de vocales específicas emotivas [13]. Bajo este concepto se tienen un total de 4 emociones $\times$ 5 vocales = 20 fonemas de vocales para el etiquetado fonético del corpus de voz emocional, dando un total de 22 consonantes + 20 vocales = 42 fonemas.

En la Figura 3.5 se muestra el etiquetado ortográfico de la frase "EL UNIVERSO ES MUY GRANDE". Note que cada palabra tiene adjunta el marcador "_N". Esto fue realizado para identificar la emoción con la cual fue pronunciada esta palabra. De esta manera los identificadores para las palabras de cada frase emocional fueron "_E" para Enojo, "_F" para Felicidad, "_N" para Neutro y "_T" para Tristeza. Por otro lado la palabra "SIL" representa los segmentos de silencio antes y después de la frase.

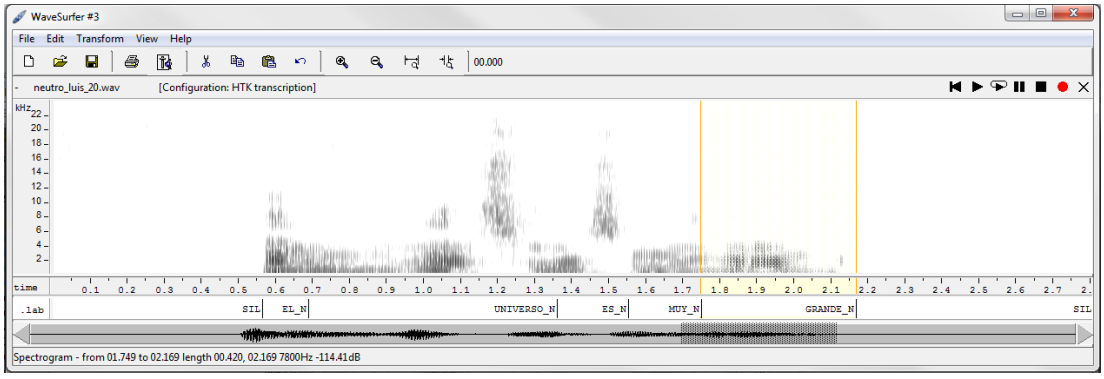


Figura 3.5: Etiquetado ortográfico con Wavesurfer de una frase con emoción neutra.

En la Figura 3.6 se muestra el etiquetado fonético para la misma frase. Los fonemas que representan las vocales ahora tienen el marcador "_n" para identificar la emoción. De manera análoga a las etiquetas ortográficas los identificadores para las vocales de las palabras emocionales fueron "_e" para Enojo, "_f" para Felicidad, "_n" para Neutro y "_t" para Tristeza. A nivel fonema "sil" representa el silencio.

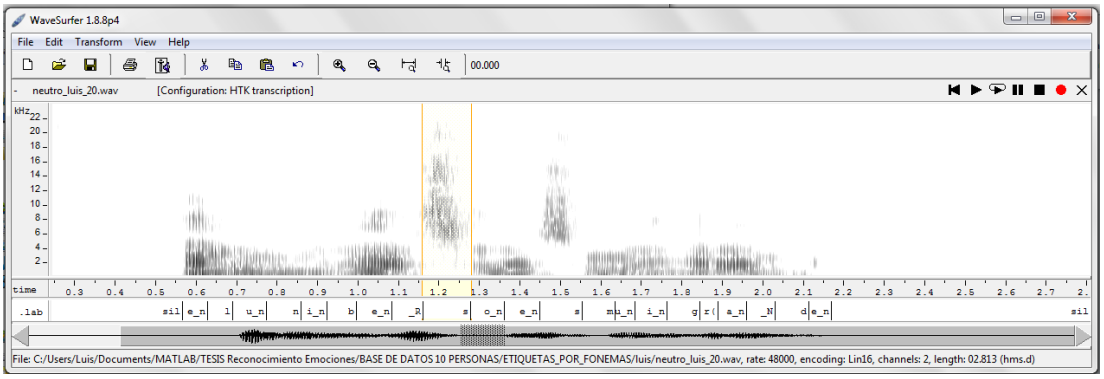


Figura 3.6: Etiquetado fonético con Wavesurfer de una frase con emoción neutra.

Para obtener la secuencia de fonemas que integran cada palabra considerando las variaciones de pronunciaciones de consonantes se desarrolló un transcriptor fonético basado en TranscribEmex [64]. Este recurso fue muy importante para crear el diccionario fonético del sistema de reconocimiento de voz (el cual se presenta en la Sección 4.1) y obtener transcripciones automáticas de nuevas palabras.

El transcriptor fonético consideró 50 reglas gramaticales y acústicas para las diferentes combinaciones de vocales y consonantes dentro de una palabra. Estas reglas se presentan en extenso en el Apéndice A. Algunas de las reglas se presentan a continuación:

- Si la consonante “q” (fonema /k/) aparece antes de una vocal “u” y la vocal “e” o “i” le sigue a ésta entonces la vocal “u” no tiene sonido y el fonema asociado (por ejemplo: /u_e/) no se incluye en la transcripción. Por ejemplo, si las siguientes palabras fueron pronunciadas con la emoción Enojo:

  QUE = /k/ /e_e/

  QUIEN = /k/ /i_e/ /e_e/ /_N/

- Si la consonante “n” aparece al principio de la palabra el fonema asociado en la transcripción es /n/. Sin embargo si la consonante aparece al final, o le sigue una consonante entonces el fonema que representa su sonido es /_N/. Por ejemplo si las siguientes palabras fueron pronunciadas con la emoción Felicidad:

  NIEVE = /n/ /i_f/ /e_f/ /b/ /e_f/

CANCION = /k/ /a_f/ /_N/ /s/ /i_f/ /o_f/ /_N/

- Si la consonante “d” aparece al principio de una palabra o si una vocal o la consonante “r” le sigue, entonces el fonema que representa su sonido es /d/. Sin embargo, si “d” aparece al final de la palabra o después de una vocal, el sonido asociado es mejor descrito con el fonema /_D/. Por ejemplo si las siguientes palabras fueron pronunciadas con la emoción Neutro:

  DRAGON = /d/ /r(/ /a_n/ /g/ /o_n/ /_N/
  DIGNIDAD = /d/ /i_n/ /_G/ /n/ /i_n/ /d/ /a_n/ /_D/

- Si la consonante “g” aparece al final de una palabra su sonido es representado con /_G/ al igual que cuando le sigue una consonante (a excepción de la “r” y “l”). Sin embargo, si las consonantes “r” o “l” , o las vocales “a”, “o”, o “u” aparecen después de la consonante “g”, entonces el sonido es mejor descrito con el fonema /g/. Cuando la vocal “e” o “i” aparece después de la “g” entonces el fonema correcto es /x/. Por ejemplo si las siguientes palabras fueron pronunciadas con la emoción Tristeza:

  GLOBO = /g/ /l/ /o_t/ /b/ /o_t/
  GRITAR = /g/ /r(/ /i_t/ /t/ /a_t/ /_R/
  GENIO = /x/ /e_t/ /n/ /i_t/ /o_t/
  GITANA = /x/ /i_t/ /t/ /a_t/ /n/ /a_t/

- Cuando la consonante “t” se encuentra al principio o al final de alguna palabra, su fonema correcto es /t/; pero cuando una consonante “b” le sigue a la consonante “t” entonces el fonema que representa su sonido es /_D/. Por ejemplo para la siguiente palabra pronunciada con emoción Neutra se tiene:

  FUTBOL = /f/ /u _n/ /_D/ /b/ /o_n/ /l/

- Para la consonante “y” se tienen diversos casos. Si “y” se encuentra al final de la palabra necesariamente suena como la vocal /i/. Si se encuentra al principio el fonema correcto debe ser /Z/. Ahora bien, cuando se tenga alguna combinación

intermedia dentro de una palabra tendremos dos casos: (1) cuando “y” se encuentra entre dos consonantes, entonces su sonido está dado por el fonema /i/, y (2) cuando la “y” se encuentra entre vocales su fonema será /Z/. Algunos ejemplos para palabras pronunciadas con la emoción Neutra se presentan a continuación:

SOY = /s/ /o_n/ /i_n/

YO = /Z/ /o_n/

GLADYS = /g/ /l/ /a_n/ /d/ /i_n/ /s/

YOYO = /Z/ /o_n/ /Z/ /o_n/

Para el caso de las vocales, con excepción de la “u´´, todas pasan directo a su forma fonética (/a/ /e/ /i/ /o/) en cualquier palabra, y para el caso de la “h” no existe un fonema porque es muda (por ejemplo “HOYO” $\rightarrow$ /o_n/ /Z/ /o_N/ si fue pronunciada con la emoción Neutro). Finalmente el transcriptor añade al final de la transcripción fonética de cada palabra el fonema “sp” (short pause, pausa corta):

GENIO = /x/ /e_n/ /n/ /i_n/ /o_n/ /sp/

El fonema “sp” sirve para indicar los pequeños silencios entre palabras de una frase y son necesarios para la construcción del sistema de reconocimiento de voz.

Capítulo 4

Sub-sistemas de Reconocimiento de Emociones en Voz y en Expresiones Faciales

Como se presentó en la Figura 1.1 el sistema multimodal propuesto integra dos sistemas: (1) Reconocedor de Voz - Voz Emocional, y (2) Reconocedor de Rostros - Expresiones Faciales. Al considerar el sistema multimodal como el sistema principal, los reconocedores de voz emocional y de expresiones faciales se consideran sub-sistemas.

En el presente capítulo se describen los pasos de desarrollo de estos sub-sistemas usando la base de datos multimodal creada con personas mexicanas. Estos sub-sistemas también representan la base de referencia para las técnicas de mejora que se presentan en el siguiente capítulo.

4.1. Sub-sistema de Reconocimiento de Emociones en Voz

En la Figura 4.1 se muestran los componentes de un sistema de reconocimiento de voz general. La implementación de estos componentes se realizó mediante la herramienta HTK [97]. Como se puede observar el corpus de entrenamiento (base de datos de voz) es el componente principal para el desarrollo del sistema. Los detalles de este componente se presentaron en el Capítulo 3.

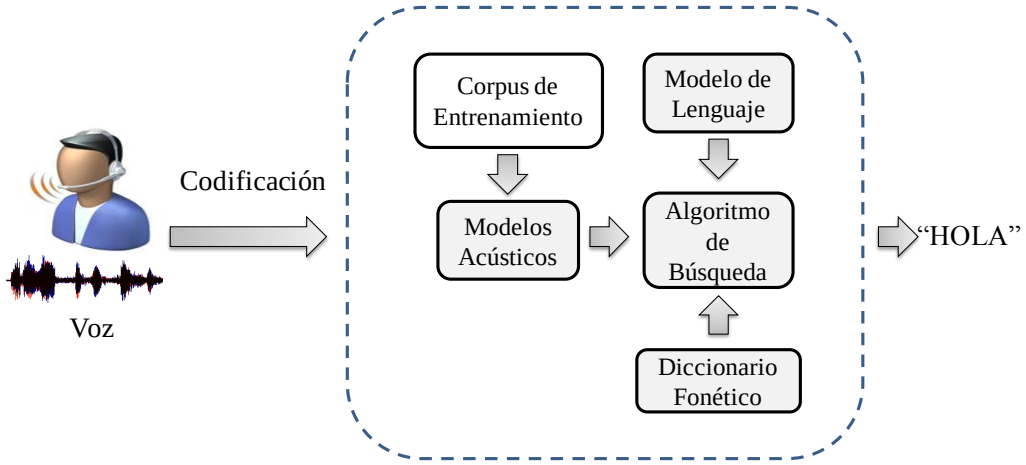


Figura 4.1: Diagrama de componentes de un sistema de reconocimiento de voz

Las muestras del corpus de voz, al igual que la voz que se recibe en-vivo (en-línea) necesitan pasar por un proceso de codificación para extracción de características espectrales [97]. En la Sección 2.1.2 se presentaron varios trabajos que utilizaron MFCCs como técnica de codificación de voz.

Aunque existen técnicas alternativas para codificación, los MFCCs tienen propiedades que los hacen más eficientes para aplicaciones de reconocimiento de voz. Los MFCCs pueden representar con más fidelidad los componentes espectrales de la voz en frecuencias mayores a 3KHz. Esto dado el proceso de filtrado en la escala no-lineal de Mel que asemeja la capacidad de percepción del oído humano a diferentes frecuencias [21, 37, 97]. Por lo tanto los MFCCs pueden modelar de mejor manera la percepción humana en comparación con técnicas como DCT (Transformada de Coseno Discreta) que consideran el proceso de filtrado en escalas lineales. Los MFCCs permiten una compresión de audio eficiente y un procesamiento más rápido de la señal de voz en el reconocedor [21, 97].

Para este trabajo la codificación del corpus de voz consistió de 12 MFCCs con consideración de coeficientes de energía, delta y de aceleración [97]. A continuación se presentan los detalles del desarrollo de los demás componentes del reconocedor de voz.

4.1.1. Modelos Acústicos

Todos los fonemas identificados en la base de datos de voz emocional deben ser modelados para ser reconocidos. Como se presentó en la Sección 2.1.2 una de las técnicas más usadas y eficientes para modelado fonético es la de HMMs. En la Figura 4.2 se presenta la arquitectura de HMM más utilizada para el modelado acústico de fonemas [97].

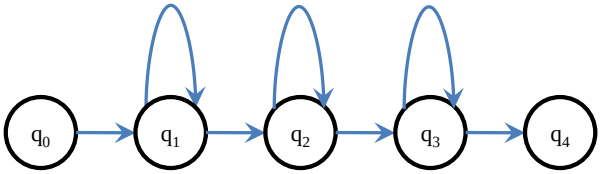


Figura 4.2: Estructura HMM Bakis de izquierda-a-derecha con tres estados emisores.

Cada fonema fue modelado con un HMM lo cual dió como resultado un conjunto de 44 HMMs (42 fonemas + sil + sp). Mediante HTK el entrenamiento supervisado de los HMMs con el algoritmo de Baum-Welch [97, 37] fue realizado.

4.1.2. Diccionario Fonético

El diccionario fonético es el componente que determina la secuencia de fonemas que forma una palabra. Es un elemento primordial para reconocedores de palabras que funcionan a base de fonemas.

El diccionario consta de todas las palabras existentes en el sistema y las secuencias de fonemas que forman cada palabra. Para crear el diccionario de manera dinámica (para cualquier palabra en el español mexicano) se desarrolló el transcriptor fonético presentado en la Sección 3.2.2.

Dado que cada palabra puede ser hablada con cualquier emoción cada palabra en el diccionario fonético aparece múltiples veces. Por ejemplo, para la palabra “CASA” el diccionario tiene las siguientes entradas:

CASA_E = /k/ /a_e/ /s/ /a_e/

CASA_F = /k/ /a_f/ /s/ /a_f/

CASA_N = /k/ /a_n/ /s/ /a_n/

CASA_T = /k/ /a_t/ /s/ /a_t/

4.1.3. Modelo de Lenguaje (ML)

El Modelo de Lenguaje (ML) contiene información estadística de las secuencias válidas de palabras en un lenguaje. El idioma español contiene varias estructuras acerca de las secuencias de palabras que forman oraciones coherentes. Por lo tanto el ML contiene información estadística del vocabulario, las estructuras de las frases existentes, y la relación que las palabras pueden tener entre sí.

Dado esto el ML guía el proceso de reconocimiento mediante la restricción de secuencias reconocidas a secuencias que son estadísticamente más probables que otras. Por ejemplo, es común (y correcto) decir "EL PÁJARO ES AZUL", pero no es correcto decir "EL AZUL PÁJARO ES". En este caso un ML le asignaría una mayor probabilidad a la primera secuencia que a la segunda.

Información de estas estructuras se puede extraer de texto representativo. Para este trabajo el estímulo textual de la base de datos de voz (las etiquetas ortográficas del corpus MX-Voz) se usó como texto representativo. El esquema de ML utilizado fue el de bigramas (secuencias válidas de pares de palabras) y se construyó con HTK. Los detalles técnicos pueden consultarse en [97].

Es importante mencionar que cualquier oración o secuencia de palabras puede ser pronunciada con cualquier emoción. De esta forma las estructuras gramaticales en un lenguaje aplican a cualquier frase sin importar la emoción con la cual fue pronunciada. El modelado de las vocales específicas emotivas implica independencia entre las mismas (es decir, una "a" con Enojo es diferente de una "a" con Tristeza). Sin embargo no se implica una independencia entre palabras del vocabulario (es decir, gramáticamente la palabra "HOLA" pronunciada con Enojo no es diferente de la misma palabra pronunciada con Felicidad).

A pesar de que las palabras habladas con una determinada emoción tienen un identificador (_E, _F, _N o _T) estas palabras existen para todas las emociones. Por esta razón el ML fue integrado por el conjunto completo de 80 frases del corpus MX-Voz considerando que cada una de ellas puede ser expresada con todas las emociones. Esto dió como resultado un total de 80 frases $\times 4$ emociones = 320 frases para la estimación del ML para el sistema de reconocimiento de voz. Con esta acción también se evita que haya un sesgo o influencia del ML en el reconocimiento del estado emocional.

4.1.4. Algoritmo de Búsqueda

El algoritmo de búsqueda se encarga de integrar la información estadística de los modelos acústicos (HMMs), el diccionario fonético y el ML para poder determinar las frases existentes en el ML (vocabulario del sistema) que mejor describen una señal de voz recibida.

La búsqueda o reconocimiento de voz se realiza en la siguiente secuencia de pasos:

- Se recibe la señal de voz **O** a ser reconocida.
- Se realiza la codificación (extracción de características) de la señal de voz **O** (esto es, se convierte a MFCCs).
- Los modelos acústicos del sistema (HMMs) proveen de patrones de observaciones de acuerdo a las probabilidades de sus elementos (probabilidades de transición y probabilidades de emisión, véase Sección 2.2.3). Estos patrones se comparan con los de la señal codificada **O** y aquellos HMMs con la máxima probabilidad de semejanza se consideran como modeladores de la señal **O**. Por lo tanto la primera salida del algoritmo de búsqueda consiste de secuencias de HMMs (fonemas) que mejor describen a **O**.
- El Diccionario Fonético restringe las secuencias de fonemas obtenidas en el paso anterior para producir palabras.
- Finalmente el algoritmo de búsqueda aplica el ML para restringir la secuencia de palabras propuestas por el Diccionario Fonético para formar secuencias válidas de palabras (frases).

El algoritmo de Viterbi [37] es ampliamente utilizado para la tarea de encontrar la secuencia más probable de estados de HMMs que pudiera haber generado una secuencia de observaciones **O**. Este algoritmo, al igual que los algoritmos de entrenamiento supervisado para los HMMs y estimación del ML, se implementó con la herramienta HTK [97].

Inicialmente el sistema de reconocimiento da como resultado la mejor secuencia de palabras que describen la frase hablada las cuales, de acuerdo a la emoción, llevan un

identificador específico. La emoción de la frase es estimada contando el número de vocales específicas emotivas dentro de las palabras reconocidas. El identificador (_e,_f,_n,_t) con el mayor número de vocales define la emoción dominante.

4.1.5. Adaptación de Usuario

El entrenamiento supervisado de los HMMs se puede realizar de manera eficiente cuando hay disponibilidad de corpora (plural) de voz. De aquí se definen dos tipos de sistemas de reconocimiento:

- Dependiente de usuario (DU): Los HMMs de este sistema se entrenan con las muestras de voz del usuario que va a usar el sistema. Por lo tanto dará buenos resultados sólo cuando sea usado por el usuario que proporcionó las muestras de voz para su entrenamiento.

- Independiente de usuario (IU): Los HMMs de este sistema se entrenan con muestras de voz de usuarios que pueden ser diferentes del usuario final del sistema. Estos sistemas requieren de técnicas de adaptación de usuario para poder usarse con usuarios diferentes de aquellos que proporcionaron las muestras de voz para entrenar el sistema.

En la práctica los sistemas de reconocimiento de voz son IU ya que el crear sistemas DU para cada usuario no es factible dado el requerimiento de tiempo para hacer el corpus de voz correspondiente y el entrenamiento del sistema. Una técnica eficiente para adaptar el sistema IU a las características acústicas de un usuario en particular es la conocida como Regresión Lineal de Máxima Probabilidad (Maximum Likelihood Linear Regression, MLLR) [97].

Las técnicas de adaptación de usuario normalmente requieren de algunas muestras de voz del usuario (datos de adaptación) para estimar "transformaciones" que ajusten los parámetros de los HMMs a su voz. La adaptación es supervisada cuando hay conocimiento de las palabras pronunciadas por el usuario, y es no supervisada cuando no se tiene dicha información. En el presente proyecto, MLLR se utilizó como técnica de adaptación, la cual se basa en el supuesto de que un conjunto de transformaciones lineales se puede usar para reducir la diferencia entre los modelos de un sistema de reconocimiento de voz y las muestras de voz para adaptación. Estas transformaciones son

aplicadas sobre la media y varianza de las mixturas de gaussianas de los HMMs del sistema base (véase Sección 2.2.3), teniendo el efecto de ajustar dichos parámetros de tal manera que aumente la probabilidad de que los HMMs del sistema generen los datos de adaptación.

La implementación de la adaptación de usuario con MLLR fue realizada con HTK. El estímulo textual para obtener muestras de voz para adaptación se obtuvo del estímulo textual de la base de datos MX-Voz.

4.2. Sub-sistema de Reconocimiento de Emociones en Expresiones Faciales

Para el desarrollo del reconocedor de expresiones faciales se consideraron las siguientes alternativas:

- **Sistema ANN Preliminar**: Extracción de regiones faciales y ANNs para el reconocimiento de patrones y modelación de errores de reconocimiento. La base de datos JAFFE se utilizó para evaluar este sistema y explorar el desempeño de las ANNs para el reconocimiento de patrones faciales.

- **Sistema PCA**: Pre-procesamiento de imágenes con extracción de rostros y PCA para reducción de dimensionalidad y reconocimiento de patrones. La base de datos JAFFE y MX-Expresiones fueron utilizadas para evaluar esta técnica cuando se considera una base de datos no estándar como la MX-Expresiones junto con una base estándar con condiciones controladas de iluminación como lo es la JAFFE.

- **Sistema PCA+ANN**: Pre-procesamiento de imágenes con extracción de rostros, PCA para reducción de dimensionalidad, y ANNs para el reconocimiento de patrones. La base de datos JAFFE y MX-Expresiones fueron utilizadas para la evaluación de PCA integrado con ANNs para el reconocimiento de expresiones faciales.

Considerando que hay tres imágenes de cada emoción para cada usuario en las bases de datos JAFFE y MX-Expresiones se escoge una imagen X para el entrenamiento del

sistema y otra imagen Y para la prueba del mismo. La imagen restante, Z es considerada para propósitos de mejora (véase Capítulo 5). Note que $X \neq Y \neq Z$. A continuación se presenta una descripción de los sistemas diseñados.

4.2.1. Sistema ANN Preliminar

Al trabajar con sistemas de visión para el reconocimiento de emociones se necesitan aplicar técnicas de pre-procesamiento a las imágenes que se quieren evaluar. Esto se hace con el fin de extraer características que mejor definan la emoción que se presenta. Sin embargo, como trabajo preliminar se propuso mejorar el desempeño del clasificador aplicando una etapa de post-procesamiento, evitando así realizar algún pre-procesamiento a las imágenes de entrada.

El sistema preliminar que se desarrolló hizo uso de las Redes Neuronales Artificiales (ANNs) para la tarea de clasificación utilizando la base de datos JAFFE y las emociones básicas de Enojo, Felicidad, Neutro y Tristeza. De cada imagen del conjunto de entrenamiento X se identificaron las regiones del rostro que mejor representan los gestos propios de una emoción. Las regiones identificadas como las más representativas fueron las cejas, los ojos y la boca. Para reducir el procesamiento computacional las imágenes de entrenamiento fueron reducidas por un factor de 0.75 antes de extraer las regiones representativas [60]. De esta manera las imágenes pasaron de tener un tamaño de 256×256 pixeles a un tamaño de 64×64 pixeles. También para reducir la dispersión y variabilidad entre los valores de los pixeles las imágenes fueron normalizadas. En la Figura 4.3 se muestra la extracción de las regiones representativas y su reordenamiento en un solo vector p_{ij}, en donde i representa la i-ésima emoción (1=Enojo, 2=Felicidad, 3=Neutro, y 4=Tristeza) y j representa el j-ésimo usuario (j = 1,...,10 usuarios de la base de datos JAFFE).

Los vectores de características de las muestras de entrenamiento de todos los usuarios se agruparon como se muestra en la Figura 4.4 para su modelamiento con una ANN. La estructura de entrenamiento de la ANN también se muestra en la Figura 4.4. Cada vector columna p_{ij} que corresponde a la i-ésima emoción de la j-ésima persona se ordenan para formar una sola matriz con 40 columnas en total (4 emociones $\times$ 10 personas). Esta matriz es la entrada de la ANN y para el entrenamiento de la misma se definió co-

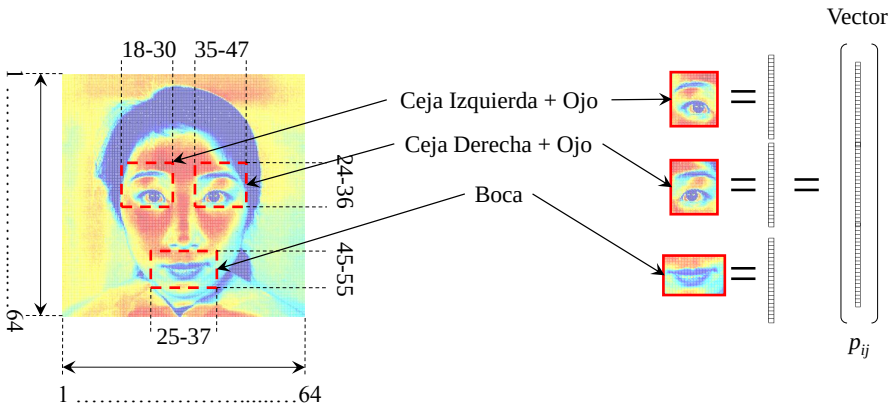


Figura 4.3: Regiones representativas para la extracción de características en el Sistema ANN Preliminar.

mo salida requerida un vector fila de 40 elementos (4 emociones $\times$ 10 personas) el cual contiene secuencia de números enteros que identifican la emoción de cada imagen en la matriz de entrada ("1" para Enojo, "2" para Felicidad, "3" para Neutro, y "4" para Tristeza).

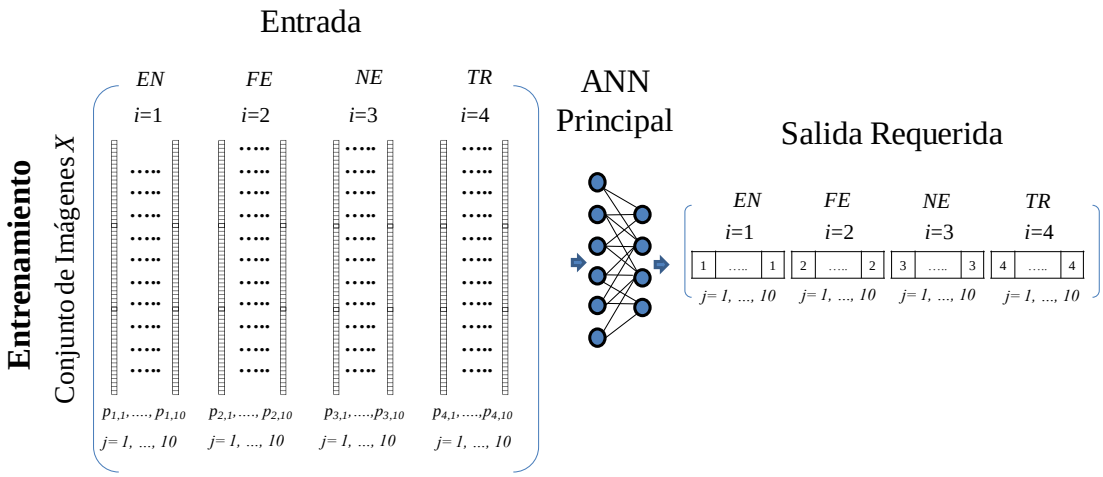


Figura 4.4: Estructura del Sistema ANN Preliminar.

En este sistema preliminar en lugar de considerar un pre-procesamiento complejo se optó por un post-procesamiento que consistió en una ANN correctiva para los errores de clasificación [14]. Para obtener resultados de clasificación para estimación de error se usó el conjunto de imágenes Z. Para modelación y corrección del error se usaron cuatro ANNs, una para cada emoción como se presenta en la Figura 4.5.

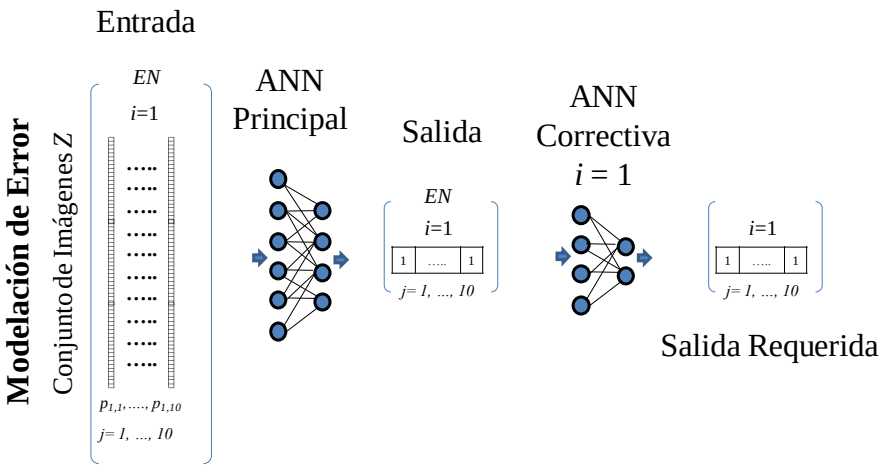


Figura 4.5: Estructura de ANN Correctiva i=1 para el Sistema ANN Preliminar.

4.2.2. Sistema PCA

Un sistema basado en PCA para la reducción de dimensionalidad y el reconocimiento de emociones fue implementado para tener una referencia de comparación con los trabajos presentados en [41, 89, 31]. Para la realización de los experimentos la base de datos MX-Expresiones y JAFFE tuvieron el siguiente pre-procesamiento:

- En su forma original las imágenes de la base de datos MX-Expresiones son en color y contienen información no solo del rostro del usuario sino también de otros elementos que no son relevantes para el proceso de reconocimiento (cabello, hombros, cuello, fondo). Para eliminar estos elementos y hacer una extracción eficiente del rostro del usuario se implementó un sistema detector de rostros basado en el algoritmo de Viola y Jones [91]. El mismo sistema de detección de rostros fue aplicado a la base de datos JAFFE.

- Las regiónes faciales extraídas de ambas bases de datos se ajustaron a un tamaño de 256 $\times$ 256 pixeles. Para la base de datos MX-Expresiones una conversión de RGB a escala de grises con una corrección de contraste por un factor de 0.7 fue realizado. Algunos ejemplos de las regiones faciales extraídas son mostrados en la Figura 4.6.

La implementación de PCA para el reconocimiento de emociones se describe a continuación:

Figura 4.6: Imágenes pre-procesadas de las bases de datos MX-Expresiones y JAFFE para los Sistemas PCA y PCA+ANN

- Para la extracción de características cada imagen de entrenamiento (conjunto X) fue transformada en un vector columna p_{ij} en donde i=índice de la emoción y j=índice del usuario (i=1,...,4 y j=1,...,J en donde J=9 para la base de datos MX-Expresiones y J=10 para la base de datos JAFFE).

- Los vectores p_{ij} que representan las imágenes de una base de datos fueron arreglados como se presenta en la Figura 4.7 para crear la matriz S, la cual es la base para la reducción de dimensionalidad. El número total de columnas en S es obtenido como $J \times I$ mientras que el número de filas es $H = 256 \times 256$ (no hubo reducción de escala en este caso).

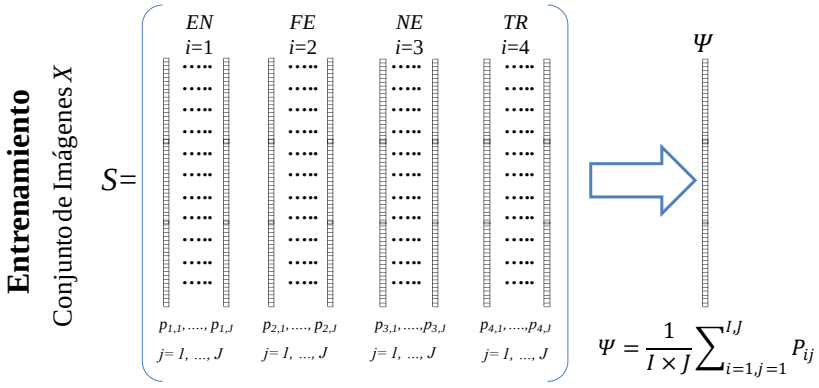


Figura 4.7: Arreglo de vectores para entrenamiento del Sistema PCA.

- Después de que S es creada, el vector columna que corresponde a la media, Ψ, es obtenido y después es restado de cada vector columna p_{ij} en S. Esto produce la

matriz A de vectores columna $\Phi_{ij} = p_{ij} - \Psi$ en donde:

$$A = [\Phi_{11}...\Phi_{1J} \quad \Phi_{21}...\Phi_{2J} \quad \Phi_{31}...\Phi_{3J} \quad \Phi_{41}...\Phi_{4J}] \tag{4.1}$$

- De la matriz A la matriz de covarianza L es obtenida como $L = A^T A$ [83]. Esta matriz es usada para obtener los eigenvectores $v = eig(L)$ los cuales son la base para los eigenrostros definidos por $u = Av$. Finalmente al considerar R como el número de eigenrostros (aquellos con los eigenvalores más altos) un nuevo rostro Γ puede ser transformado en sus componentes de eigenrostro de la siguiente manera:

$$\Omega = u^T(\Gamma - \Psi) = [\omega_1, \omega_2, ..., \omega_R]^T \tag{4.2}$$

- La reducción de dimensionalidad es lograda al ser la dimensión de Ω igual al número de muestras de entrenamiento ($I \times J$) [29]. También Ω representa las características de las muestras de entrenamiento las cuales pueden ser usadas para reconocimiento: los pesos ω_r describen las contribuciones de cada eigenrostro en la representación de la imagen del rostro de entrada. Este vector puede ser usado para reconocimiento de rostros/emociones encontrando la distancia euclidiana e más pequeña entre los vectores de pesos del rostro de entrada y los rostros de entrenamiento de la siguiente manera:

$$e = \|\Omega_{entrada} - \Omega\| \tag{4.3}$$

- El enfoque de eigenrostros determina mediante la mínima distancia euclidiana el rostro de la base de datos (implícita en la matriz S) que mejor se asocia a una imagen de entrada. Al ubicar el rostro con la mínima distancia se puede determinar la identificación del usuario si la imagen en S tiene una etiqueta con dicha información. Para el reconocimiento de la emoción cada imagen en la matriz S tiene una etiqueta que especifica la emoción que expresa.

4.2.3. Sistema PCA+ANN

El Sistema PCA+ANN integra los enfoques de los sistemas previos. Con PCA se realiza la reducción de dimensionalidad del conjunto de imágenes de entrenamiento X

la cual es representada por la matriz Ω. Esta matriz es normalizada y dada como entrada a una ANN para el entrenamiento. Para la ANN la salida requerida fue considerada como un arreglo de vectores binarios representando la emoción asociada a cada vector p_{ij} (o ω_r). En la Figura 4.8 se presenta la estructura de la entrada y los datos de entrenamiento para la ANN.

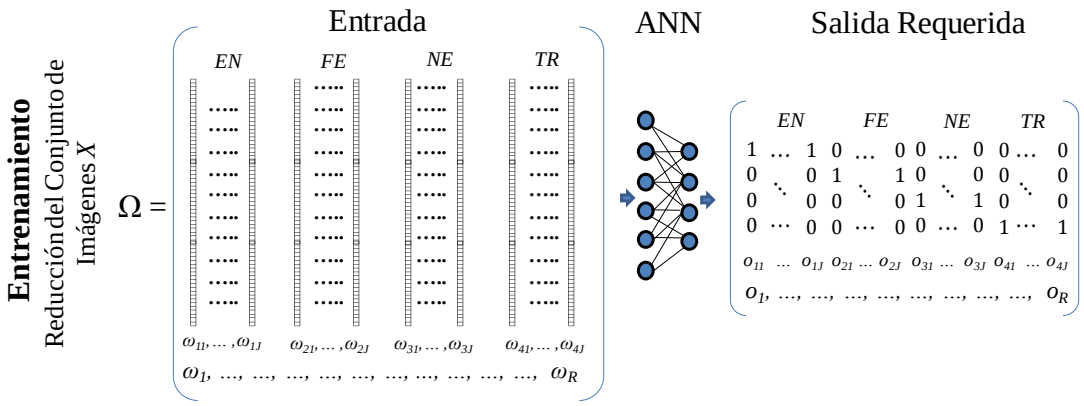


Figura 4.8: Estructura del Sistema PCA+ANN.

Capítulo 5

Optimización Evolutiva e Integración de Sistema Multimodal

Dentro de los objetivos del presente trabajo se consideró el de mejorar el desempeño de los sistemas de reconocimiento de emociones. En el capítulo anterior se presentaron los sub-sistemas de reconocimiento base para el reconocimiento de expresiones y emociones en voz. En ambos sub-sistemas se realiza un pre-procesamiento de los patrones visuales y acústicos a ser reconocidos. Sin embargo, algo que no ha sido presentado en la literatura para este problema es la mejora de la arquitectura del elemento fundamental del reconocedor, esto es, la técnica de modelado.

Para el sub-sistema de voz la técnica de modelado consiste de HMMs mientras que para el de expresiones la técnica consiste principalmente de ANNs. En este capítulo se presenta la mejora de estas técnicas mediante GAs. De igual manera se hace una validación estadística de los resultados de desempeño de esta técnica de mejora. Esto fue importante para seleccionar el tipo de sub-sistema de reconocimiento de expresiones (ANN Preliminar, PCA, PCA+ANN) para la integración del sistema multimodal.

El tipo de GA aplicado para optimización en este trabajo se conoce como micro-GA dado el tamaño considerado de población inicial [5]. Este tipo de GA puede converger de manera rápida después de algunas generaciones, generando soluciones de igual calidad que un GA convencional que puede tener poblaciones de hasta 1000 individuos [5].

5.1. Optimización del Sub-sistema de Reconocimiento de Emociones en Voz

Como se presentó en la Sección 4.1 la estructura HMM más usada para el modelado acústico de fonemas es la conocida como Bakis (izquierda-a-derecha con tres estados emisores). Para el caso específico de los fonemas que representan vocales específicas emotivas se considera la hipótesis de que otras estructuras de HMMs pueden ser más adecuadas para su modelado acústico. En la Figura 5.1(a) se muestra la estructura Bakis estándar antes mencionada (ahora identificada como "Bakis Tipo A") en tanto que las Figuras 5.1(b) y 5.1(c) muestran dos alternativas adicionales: "Bakis Tipo B" y "Ergódica" respectivamente.

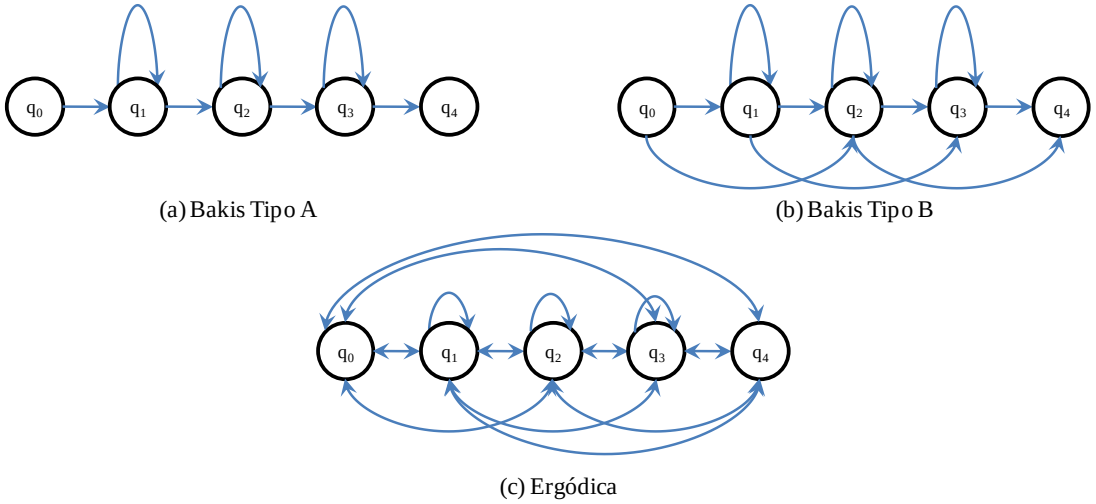


Figura 5.1: Estructuras de HMMs para modelado acústico de los fonemas de las vocales específicas emotivas.

El problema de identificar la estructura de HMMs más apropiadas para cada vocal específica emotiva es abordado con un GA (en la Sección 2.2.5 se pueden consultar los detalles técnicos generales de los GAs). En la Figura 5.2 se presenta el cromosoma del micro-GA diseñado para la optimización de la estructura de los HMMs.

El cromosoma se definió con una codificación binaria de 20 genes (2 bits cada uno) el cual representa el tipo de la estructura de HMM para la vocal/emoción asociada (4 emociones $\times$ 5 vocales). Solamente las estructuras de las vocales específicas emotivas fueron consideradas para optimización. Los modelos HMM para las consonantes man-

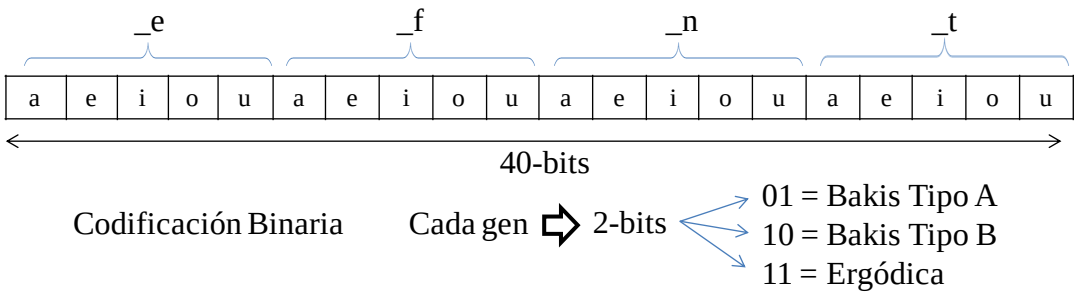


Figura 5.2: Cromosoma utilizado para la optimización del reconocedor de vocales específicas emotivas.

tuvieron su estructura estándar "Bakis Tipo A". El valor de la aptitud de los individuos (función objetivo) fue medido como la tasa de clasificación obtenida con el conjunto completo de HMMs.

Una vez que el conjunto de estructuras para los HMMs y la representación del cromosoma fueron definidos las características de operación del GA fueron establecidas. Estas características son mostradas en la Tabla 5.1.

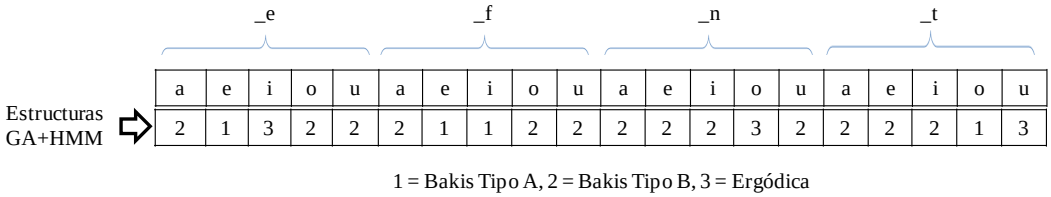
Tabla 5.1: Parámetros de configuración del algoritmo genético para el sistema de voz.

Parámetro	Descripción
Población Inicial	10 Individuos
Operadores de Reproducción	Cruzamiento: Uniforme aleatorio en N-puntos
	Mutación: Simple en 1-bit
Método de Selección	Ruleta
Función Objetivo (Aptitud)	Tasa de Clasificación Máxima
Generaciones	20

Para encontrar las estructuras de HMMs más adecuadas cada conjunto de frases de la base de datos MX-Voz (véase Sección 3.2) fue dividido en: (a) frases de entrenamiento y (b) frases para optimización (evaluación de aptitud). El conjunto de entrenamiento consistió de las últimas ocho frases de cada conjunto emocional (frases 13 a 20) y las frases de optimización consistieron de las seis frases intermedias (frases 7 a 12). Las primeras seis frases (frases 1 a 6) fueron consideradas para la evaluación preliminar del GA.

En la Figura 5.3 se presenta el vector fila resultante con las estructuras de HMMs para cada vocal específica emotiva. También se presenta el desempeño preliminar del reconocimiento de emociones con estas estructuras sobre las frases de evaluación para

todos los usuarios. Este desempeño es comparado con el del reconocedor base en donde todos los HMMs tienen la misma estructura estándar (Bakis Tipo A). Como se presenta, el conjunto de HMMs encontrados por el GA obtuvieron una ganancia significativa del 5.20 % (75.00 % - 80.20 %) sobre las frases de evaluación. En este conjunto se observa una combinación de todas las estructuras consideradas (Bakis Tipo A, Bakis Tipo B, Ergódica) en donde la estructura Bakis Tipo B tiene más presencia.



Desempeño Preliminar de Reconocimiento de Emociones

Conjunto HMM	Conjunto de Optimización (6 frases)	Conjunto de Evaluación (6 frases)
Estándar (Bakis Tipo A)	77.08%	75.00%
GA+HMMs (Bakis Tipo A, B, o Ergódica)	85.41%	80.20%

Figura 5.3: Sistema GA+HMMs: configuración de HMMs para las vocales específicas emotivas y tasa de reconocimiento preliminar del sistema de voz.

5.1.1. Análisis de Resultados

Para la evaluación final del enfoque evolutivo con el micro-GA para el reconocimiento de emociones basado en voz dos esquemas fueron considerados:

- Esquema de Prueba A (dependiente de usuario): bajo este esquema 40 frases (10 primeras frases $\times$ 4 emociones) de cada usuario fueron consideradas para entrenamiento de los HMM adicionalmente a las 560 frases (20 frases $\times$ 4 emociones $\times$ 7 usuarios restantes) de los otros usuarios. Finalmente el desempeño del reconocimiento es evaluado con el resto de las 40 frases del hablante en cuestión (10 últimas frases $\times$ 4 emociones).

- Esquema de Prueba B (independiente de usuario): bajo este esquema 40 frases (10 primeras frases $\times$ 4 emociones) de cada usuario fueron consideradas para adaptación de usuario (con la técnica MLLR). Los HMMs fueron entrenados solamente

con las 560 frases (20 frases $\times$ 4 emociones $\times$ 7 usuarios restantes) de los otros usuarios. Finalmente el desempeño del reconocimiento es evaluado con el resto de las 40 frases del hablante en cuestión (10 últimas frases $\times$ 4 emociones).

Los desempeños de las estructuras estándar (Bakis Tipo A solamente) y las estructuras GA+HMM (véase Figura 5.3) fueron evaluados bajo ambos esquemas de prueba y los resultados son presentados en la Tabla 5.2 y Tabla 5.3.

Tabla 5.2: Tasa de reconocimiento promedio del sistema de voz: HMMs Estándar (Base con Bakis Tipo A).

Esquema de Prueba A

Usuario	Género	Enojo	Felicidad	Neutro	Tristeza
Lu	M	100.00	50.00	100.00	80.00
Ta	F	100.00	80.00	100.00	90.00
Au	F	80.00	85.00	80.00	100.00
Mi	M	70.00	70.00	100.00	85.00
Me	F	75.00	70.00	90.00	90.00
Je	M	100.00	30.00	75.00	50.00
Li	F	70.00	40.00	20.00	75.00
Ne	F	80.00	100.00	90.00	90.00
	Promedio	84.38	65.63	81.88	82.50
				Promedio Total	78.59

Esquema de Prueba B

Usuario	Género	Enojo	Felicidad	Neutro	Tristeza
Lu	M	100.00	50.00	100.00	100.00
Ta	F	100.00	70.00	100.00	100.00
Au	F	100.00	100.00	80.00	100.00
Mi	M	70.00	80.00	100.00	90.00
Me	F	95.00	90.00	100.00	100.00
Je	M	80.00	100.00	70.00	90.00
Li	F	75.00	80.00	75.00	70.00
Ne	F	90.00	100.00	100.00	80.00
	Promedio	88.75	83.75	90.63	91.25
				Promedio Total	88.59

Tabla 5.3: Tasa de reconocimiento promedio del sistema de voz: GA+HMMs.

Esquema de Prueba A

Usuario	Género	Enojo	Felicidad	Neutro	Tristeza
Lu	M	100.00	60.00	100.00	90.00
Ta	F	100.00	90.00	100.00	90.00
Au	F	80.00	70.00	80.00	100.00
Mi	M	100.00	65.00	100.00	90.00
Me	F	65.00	90.00	100.00	90.00
Je	M	100.00	20.00	85.00	25.00
Li	F	60.00	45.00	80.00	90.00
Ne	F	80.00	100.00	100.00	90.00
	Promedio	85.63	67.50	93.13	83.13
				Promedio Total	82.34

Esquema de Prueba B

Usuario	Género	Enojo	Felicidad	Neutro	Tristeza
Lu	M	100.00	60.00	100.00	100.00
Ta	F	100.00	90.00	100.00	90.00
Au	F	100.00	100.00	80.00	100.00
Mi	M	70.00	60.00	90.00	90.00
Me	F	95.00	100.00	90.00	100.00
Je	M	90.00	100.00	90.00	80.00
Li	F	90.00	60.00	90.00	70.00
Ne	F	100.00	100.00	100.00	80.00
	Promedio	93.13	83.75	92.50	88.75
				Promedio Total	89.53

Para ambos sistemas (HMMs Estándar y GA+HMMs) el esquema de prueba independiente de usuario presentó un desempeño mayor que el del esquema dependiente de usuario. Para la validación estadística de la mejora obtenida con el GA se hizo uso de la prueba no paramétrica de Wilcoxon de una muestra. Esto dado que los resultados no tienen una distribución normal. La prueba de Wilcoxon puede determinar si la media de un conjunto de datos difiere de un valor en específico (referencia).

Para el Esquema de Prueba A (dependiente de usuario) se consideró como valor de referencia el promedio total obtenido con los HMMs Estándar (78.59 %). Al analizar el conjunto de datos del Esquema de Prueba A correspondiente al reconocimiento con los GA+HMMs se obtuvo que hay una diferencia significativamente estadística con p = 0.065 (considerando $p < 0.10$).

Sin embargo bajo el Esquema de Prueba B (independiente de usuario) la mejora obtenida con los GA+HMMs no fue estadísticamente significativa. Considerando como valor de referencia el promedio total obtenido con los HMMs Estándar (88.59 %) la prueba de Wilcoxon determinó que el conjunto de datos correspondiente al reconocimiento con los GA+HMMs no era estadísticamente diferente dado $p = 0.147$ ($p > 0.10$). A pesar de que se obtuvieron mejoras para Enojo y Neutro con los GA+HMMs bajo el esquema independiente de usuario no hubo una mejora para Tristeza.

Para ambos sistemas y esquemas de prueba Felicidad fue la emoción con la tasa más baja de reconocimiento. Considerando el uso de la estructura estándar solamente para las vocales específicas emotivas de Tristeza (Estándar+GA+HMMs) bajo el Esquema de Prueba B el desempeño total del sistema se presenta en la Tabla 5.4. Este desempeño (90.16 %) es marginalmente significativo comparado con el valor de referencia de los HMMs Estándar (88.59 %) al tener $p = 0.091$.

Tabla 5.4: Tasa de reconocimiento promedio del sistema de voz: Estándar+GA+HMMs.

Esquema de Prueba B					
Sistema	Enojo	Felicidad	Neutro	Tristeza	Promedio
HMMs Estándar	88.75	83.75	90.63	91.25	88.59
GA+HMMs	93.13	83.75	92.50	88.75	89.53
Estándar+GA+HMMs	93.13	83.75	92.50	91.25	90.16

De esta manera la configuración Estándar+GA+HMMs fue seleccionada para el subsistema de reconocimiento de emociones en voz para el sistema multimodal. Es de notar que el desempeño obtenido es mayor que el de otros sistemas con número similar de usuarios y mismo número de emociones (véase Tabla 2.1).

5.2. Optimización del Sub-sistema de Reconocimiento de Emociones en Expresiones Faciales

Cuando una ANN es definida se deben considerar algunas características que son necesarias para un desempeño eficiente. Estas características residen en el número de capas ocultas (o intermedias), el número de neuronas por capa, y la función de activación o transferencia para cada capa. Algunas veces estas características se definen de acuerdo a la experiencia. Otras veces estas características se definen en base a trabajos previos en donde se obtenía un buen desempeño con un determinado conjunto de características.

En general no hay un consenso acerca de las características de una ANN para propósitos de reconocimiento de expresiones. Por lo tanto se estableció la aplicación de un GA para determinar una estructura de ANN apropiada para este propósito.

5.2.1. Optimización del Sistema ANN Preliminar y Análisis de Resultados

Para la optimización de la "ANN Principal" del sistema preliminar (véase Figura 4.4) se consideraron las alternativas de GAs presentadas en la Figura 5.4.

En la Sección 4.2 se mencionó que hay tres imágenes de cada emoción para cada usuario en las bases de datos JAFFE y MX-Expresiones. Debido a esto se escogió una imagen X para el entrenamiento del sistema y otra imagen Y para la prueba del mismo. La imagen restante, Z, es utilizada para la evaluación de aptitud (función objetivo) de las configuraciones estimadas por el GA.

Para el entrenamiento de la "ANN Principal" con el conjunto X las siguientes características y parámetros se establecieron: (1) ANN del tipo *Feedforward* entrenada con el algoritmo de propagación hacia atrás RPROP (Resilient Backpropagation), (2) 1000 *epochs*, y (3) un error de 0.0001. La evaluación de la aptitud con el conjunto Z hizo que cada GA convergiera a una configuración específica para la estructura de la "ANN Principal". Estas configuraciones se muestran en la Tabla 5.5.

El reconocimiento de emociones realizado sobre el conjunto Z para la evaluación de aptitud de configuraciones de estructuras se utilizó para el entrenamiento de las "ANNs Correctivas". Debido a que los vectores de error que se modelan con las "ANNs Co-

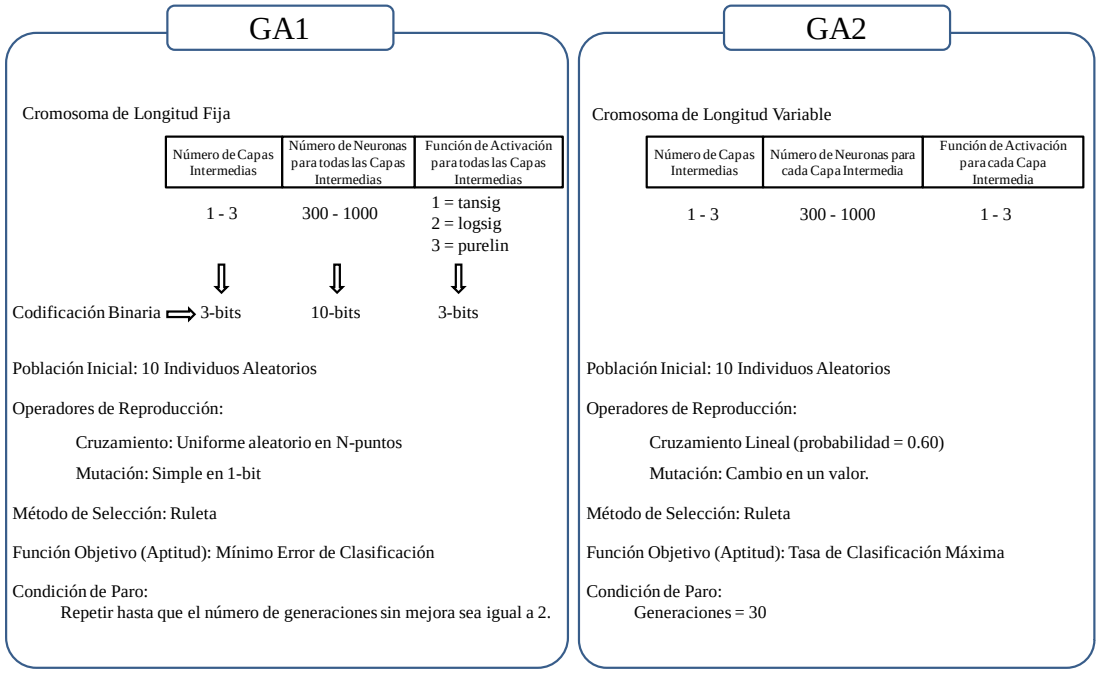


Figura 5.4: Parámetros de configuración del algoritmo genético para el Sistema ANN Preliminar.

Tabla 5.5: Sistema GA1/2+ANN Preliminar: Configuración de ANNs para el Sistema ANN Preliminar de reconocimiento de expresiones.

Algoritmo Genético	**Estructura**
GA1	Capas Intermedias = 2 Neuronas = 961 Función de Activación para las Capas Intermedias = *logsig* Función de Activación para la Capa de Salida = *purelin* (fija)
GA2	Capas Intermedias = 3 Neuronas = [Capa1=859, Capa2=456, Capa3= 144] Función de Activación para las Capas Intermedias = [Capa1= *purelin* Capa2= *logsig* Capa3= *tansig*] Función de Activación para la Capa de Salida = *purelin* (fija)

rrectivas" son más pequeños que los usados para la "ANN Principal" sus estructuras fueron más simples. Las "ANNs Correctivas" no fueron sujetas a optimización y tuvieron las siguientes características y parámetros: (1) ANN del tipo *Feedforward* entrenada con el algoritmo de propagación hacia atrás SCG (Scaled Conjugate Gradient), (2) 1000 *epochs*, (3) un error de 0.0001, (4) dos capas intermedias (Capa1=400 neuronas, Capa2= 200 neuronas), y (5) función de activación *purelin*.

Análisis de Resultados

En la Tabla 5.6 se presentan los resultados de reconocimiento del Sistema ANN Preliminar (sin ANN Correctiva) optimizado con cada GA. Estos resultados se obtuvieron con el conjunto de imágenes Y (prueba).

Tabla 5.6: Tasa de reconocimiento de expresiones promedio del Sistema GA1/2+ANN Preliminar.

Sistema ANN Preliminar	Enojo	Felicidad	Neutral	Tristeza	Tasa Promedio
Con GA1	90.00%	90.00%	90.00%	30.00%	75.00%
Con GA2	100.00%	80.00%	80.00%	50.00%	77.50%

Como se presenta en la Tabla 5.6 el desempeño del Sistema ANN Preliminar optimizado con GA2 fue mayor que el desempeño obtenido con el sistema optimizado con GA1 ($77.50\,\% > 75.00\,\%$). Sin embargo ambos desempeños se encuentran debajo de las tasas de reconocimiento presentadas en la literatura para la base de datos JAFFE (véase Tabla 2.1).

Esto hasta cierto punto era esperado dado que no hubo implementación en este sistema de algún método complejo para extracción de características o pre-procesamiento. En la Tabla 5.7 se presenta la tasa de reconocimiento del Sistema ANN Preliminar optimizado con GA2 (el que demostró mayor eficiencia) y las ANNs Correctivas.

Tabla 5.7: Tasa de reconocimiento de expresiones promedio del Sistema GA2+ANN Preliminar con ANNs Correctivas.

Sistema ANN Preliminar	Enojo	Felicidad	Neutral	Tristeza	Tasa Promedio
Con GA2 + ANNs Correctivas	80.00%	80.00%	100.00%	80.00%	85.00%

A pesar de haber obtenido un incremento adicional de 5.00 % en la tasa de reconocimiento el desempeño del sistema siguió siendo menor que el reportado en la literatura. No obstante este trabajo preliminar proporcionó información importante con respecto a los siguientes puntos:

- El GA (GA1 o GA2) puede proporcionar configuraciones de estructuras más eficientes para el reconocimiento de expresiones faciales.
- La ANN Correctiva puede mejorar el desempeño del sistema de reconocimiento pero involucra un proceso adicional de reconocimiento que puede extender el tiempo de respuesta en aplicaciones en-vivo. De igual manera la mejora en el desempeño pueder ser mínimo si la tasa de reconocimiento ya es alta [14].
- La segmentación de ojos y boca no es exacta o eficiente por lo que hay un ruido significativo dentro de los vectores de características.

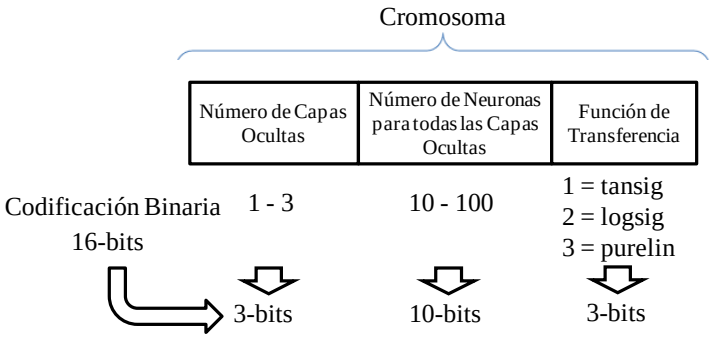
Debido a estos puntos se procedió a considerar el rostro completo del usuario para realizar el reconocimiento de la expresión. De igual manera se procedió a realizar la reducción de dimensionalidad y extracción de características con PCA y la consideración de ambas bases de datos (MX-Expresiones y JAFFE) para los experimentos. Los resultados se presentan y discuten en la siguiente sección.

5.2.2. Optimización del Sistema PCA+ANN y Análisis de Resultados

Como se presentó en la Sección 4.2 se desarrollaron dos sistemas con extracción de características y reducción de dimensionalidad: PCA y PCA+ANN. Estos dos sistemas se entrenaron con las bases de datos MX-Expresiones y JAFFE y las imágenes tuvieron un pre-procesamiento en donde se realizó la extracción del rostro de cada usuario mediante el método descrito en [91].

En contraste con el Sistema ANN Preliminar, en los Sistemas PCA y PCA+ANN se realizaron pruebas con diferentes imágenes para los conjuntos X, Y y Z. Esto dió como resultado un análisis más completo del efecto que tiene la imagen seleccionada para entrenamiento, prueba, y optimización, en el desempeño del sistema de reconocimiento final.

Para el Sistema PCA no hubo aplicación de GA ya que se utilizó este sistema como punto de referencia (sistema base). En cambio si hubo una aplicación de GA para optimizar la estructura de la ANN del Sistema PCA+ANN (como en el caso del Sistema ANN Preliminar). Los parámetros de configuración del GA para la optimización de la ANN del Sistema PCA+ANN se presenta en la Figura 5.5.



Población Inicial	10 individuos
Operadores de Reproducción	Cruzamiento aleatorio Uniforme en N-puntos/ Mutación Simple en 1 bit
Operador de Selección	Ruleta
Función de Aptitud	Tasa de Clasificación Máxima
Condición de Paro	10 Iteraciones

Figura 5.5: Parámetros de configuración del algoritmo genético para el Sistema PCA+ANN.

Para el entrenamiento de la ANN se utilizó el conjunto de imágenes X y el algoritmo de propagación hacia atrás RPROP (Resilient Backpropagation) con 1000 *epochs* y un error de 0.0001. La optimización se hizo con el conjunto de imágenes Z (evaluación de aptitud) y la prueba final se hizo sobre el conjunto Y.

Análisis de Resultados

Para los experimentos de reconocimiento de emociones una validación cruzada fue llevada a cabo. De esta manera diferentes imágenes se fueron usando para los conjuntos X, Y y Z. Para esto las imágenes disponibles por emoción fueron enumeradas de la 1 a la 3 y los esquemas presentados en la Tabla 5.8 fueron considerados para el entrenamiento (X), optimización del GA (Z) y evaluación del sistema (Y, prueba). Esto fue

realizado para determinar la influencia de la muestra considerada para entrenamiento sobre el desempeño del reconocimiento con una muestra diferente (muestra de prueba).

Tabla 5.8: Esquemas de imágenes considerados para los conjuntos X, Z, y Y de Entrenamiento, Optimización y Evaluación de los Sistemas de Reconocimiento PCA, PCA+ANN, y PCA+GA+ANN.

Conjunto	Esquema	S_1	S_2	S_3	S_4	S_5	S_6
X	Entrenamiento	1	1	2	2	3	3
Z	Optimización GA	3	2	3	1	2	1
Y	Prueba	2	3	1	3	1	2

Para el análisis del reconocimiento de expresiones se consideraron los siguientes factores:

- Sistema de Reconocimiento: (a) sistema basado en PCA (Sistema PCA) como se presentó en la Sección 4.2.2, (b) sistema basado en PCA para reducción de dimensionalidad y ANN como técnica de reconocimiento (Sistema PCA+ANN) como se presentó en la Sección 4.2.3, y (c) Sistema PCA+ANN en donde la ANN es optimizada con el GA presentado en la Figura 5.5 (Sistema PCA+GA+ANN).
- Emociones: (a) Enojo, (b) Felicidad, (c) Neutro, (d) Tristeza.
- Base de Datos de Expresiones Faciales: (a) base de datos propia de usuarios mexicanos (MX-Expresiones), (b) base de datos estándar (JAFFE), y (c) la unión de las bases de datos MX-Expresiones+JAFFE.
- Esquema de Entrenamiento-Optimización-Prueba: S_1, S_2, S_3, S_4, S_5 y S_6 como se presenta en la Tabla 5.8.

Un análisis factorial con Minitab (v17.2.1) fue realizado para determinar el impacto de cada factor en la tarea de reconocimiento de emociones. En la Figura 5.6 las gráficas de interacción para la tasa media de reconocimiento de emociones a través de todos los factores son presentadas.

Considerando el "Sistema de Reconocimiento" (método de clasificación) como el factor principal las Figuras 5.6(a),5.6(b) y 5.6(c) presentan la siguiente información:

- Las tasas de reconocimiento más altas para Enojo, Felicidad y Neutro son obtenidas con el sistema integrado PCA+GA+ANN en comparación con los sistemas

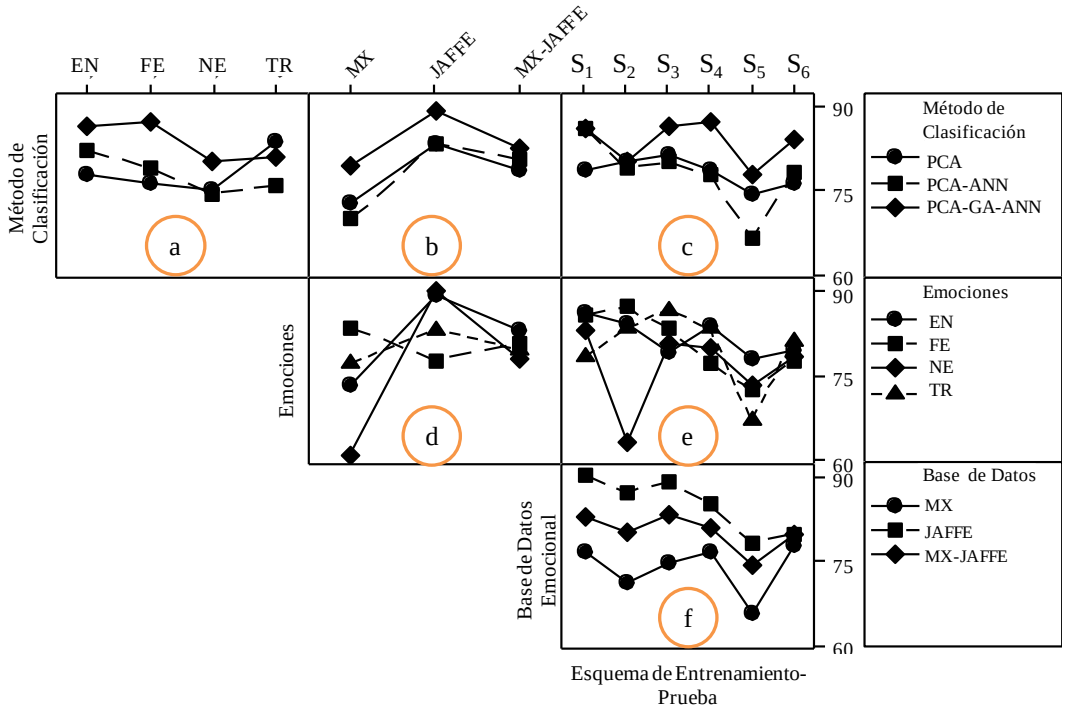


Figura 5.6: Gráficas de interacción de la tasa de reconocimiento de expresiones promedio del Sistema PCA, PCA+ANN y PCA+GA+ANN.

PCA y PCA+ANN. Para Tristeza la tasa de reconocimiento más alta es obtenida con el Sistema PCA.

- Las tasas de reconocimiento más altas son obtenidas para la base de datos JAFFE en comparación con la base de datos MX-Expresiones. Para ambas bases de datos una tasa mayor de reconocimiento es conseguida con el Sistema PCA+GA+ANN.

- Independientemente del esquema de Entrenamiento-Optimización-Prueba el Sistema PCA+GA+ANN presenta una mayor tasa de reconocimiento que los otros sistemas. Sin embargo hay una variabilidad significativa entre las tasas de reconocimiento de estos esquemas.

Cuando se consideran las “Emociones” como el factor principal, en la Figura 5.6(d) se observa que Felicidad es la emoción que mejor se reconoce en la base de datos MX-Expresiones. En tanto esta emoción tiene la tasa de reconocimiento más baja en la base de datos JAFFE. La emoción Neutro tiene la tasa de reconocimiento más baja en la base MX-Expresiones y se reconoce mejor en la base de datos JAFFE. La Figura 5.6(e) da

evidencia del efecto de los esquemas de Entrenamiento-Prueba sobre el desempeño del reconocimiento. La tasa de reconocimiento para Neutro es severamente afectada bajo el esquema S_2 (entrenando con la imagen 1, optimizando con la imagen 2, y probando con la imagen 3). Una disminución general en la tasa de reconocimiento es observada bajo el esquema S_5 (entrenando con la imagen 3, optimizando con la imagen 2, y probando con la imagen 1).

Finalmente cuando se considera la "Base de Datos de Emociones Faciales" como el factor principal la Figura 5.6(f) presenta la base de datos JAFFE con la mayor tasa de reconocimiento a través de todos los esquemas de Entrenamiento-Optimización-Prueba. La base de datos MX-Expresiones tiene la tasa más baja a través de todos los esquemas. También se observa una disminución general en la tasa de reconocimiento bajo el esquema S_5. En general el Sistema PCA+GA+ANN presenta la mayor tasa promedio de reconocimiento (83.7462 %) en comparación con el Sistema PCA+ANN (77.9083 %) y el Sistema PCA (78.2501 %). Minitab v17.2.1 fue utilizado para realizar una prueba ANOVA para evaluar la significancia estadística de estos resultados. Esta prueba concluyó que el desempeño del Sistema PCA+GA+ANN es estadísticamente diferente del desempeño de los Sistemas PCA y PCA+ANN con $p < 0.05$ (p = 0.003).

Un tema importante acerca de estos experimentos es la estructura de la ANN obtenida para cada esquema de Entrenamiento-Prueba en el Sistema PCA+GA+ANN. La Tabla 5.9 presenta un resumen general de estas estructuras para cada esquema y base de datos.

En general el GA determinó más capas ocultas (3) para la base de datos JAFFE que para la base MX-Expresiones (1). Cuando ambas bases se unen el GA determinó un valor intermedio. El número promedio de neuronas estuvo dentro del rango de 43 a 57 neuronas y para todas las bases de datos y esquemas de prueba la tercera función de transferencia (*purelin*) fue definida. Tomando los valores promedio de las ANNs, una estructura general de ANN fue estimada para cada base de datos y todos los esquemas de Entrenamiento-Prueba. Como se presenta en la Tabla 5.9, con la ANN general para ambas bases de datos, la tasa de reconocimiento total incrementó de 83.7462 % a 87.7979 %.

Sin importar el sistema de reconocimiento una tasa mayor fue obtenida con la base de datos JAFFE en comparación con la base de datos propia MX-Expresiones. Como se

Tabla 5.9: Tasa de reconocimiento de expresiones promedio del PCA+GA+ANN con Estructura Específica y Promedio para la ANN.

Base de Datos MX

Estructura Específica para la ANN para cada Esquema de Entrenamiento-Optimización-Prueba									**Estructura Promedio para la ANN**				
Ent	Opt	Pru	Estructura	EN	FE	NE	TR	Promedio	EN	FE	NE	TR	Promedio
1	2	3	[1 71 3]	80.37	88.88	45.55	61.85	69.16	88.89	88.89	44.44	66.67	72.22
1	3	2	[2 78 3]	84.07	88.51	72.96	77.40	80.74	88.89	88.89	77.78	77.78	83.34
2	1	3	[1 56 3]	88.88	100.00	77.77	79.25	86.48	88.89	100.00	77.78	88.89	88.89
2	3	1	[2 64 3]	72.96	97.03	77.77	72.96	80.18	77.78	100.00	77.78	88.89	86.11
3	1	2	[1 30 3]	82.59	88.88	77.77	100.00	87.31	88.89	88.89	77.78	100.00	88.89
3	2	1	[1 42 3]	59.25	90.00	60.37	79.62	72.31	55.56	88.89	77.78	77.78	75.00
Estructura Promedio = [1.33 56.83 3] ≈ [1 57 3]				Tasa de Reconocimiento Global =				**79.36**	Tasa de Reconocimiento Global =				**82.41**

Base de Datos JAFFE

Estructura Específica para la ANN para cada Esquema de Entrenamiento-Optimización-Prueba									**Estructura Promedio para la ANN**				
Ent	Opt	Pru	Estructura	EN	FE	NE	TR	Promedio	EN	FE	NE	TR	Promedio
1	2	3	[3 62 3]	94.00	95.66	89.66	92.66	93.00	100.00	100.00	90.00	100.00	97.50
1	3	2	[3 73 3]	93.33	90.66	100.00	82.66	91.66	100.00	100.00	100.00	80.00	95.00
2	1	3	[3 50 2]	92.66	79.33	94.00	91.66	89.41	100.00	100.00	100.00	100.00	100.00
2	3	1	[3 28 3]	100.00	85.66	100.00	90.00	93.92	100.00	90.00	100.00	90.00	95.00
3	1	2	[2 59 2]	84.33	80.66	87.00	79.00	82.75	90.00	90.00	100.00	80.00	90.00
3	2	1	[1 18 2]	98.00	71.33	96.33	75.00	85.17	100.00	80.00	100.00	80.00	90.00
Estructura Promedio = [2.5 48.33 2.5] ≈ [3 48 3]				Tasa de Reconocimiento Global =				**89.32**	Tasa de Reconocimiento Global =				**94.58**

Base de Datos MX + JAFFE

Estructura Específica para la ANN para cada Esquema de Entrenamiento-Optimización-Prueba									**Estructura Promedio para la ANN**				
Ent	Opt	Pru	Estructura	EN	FE	NE	TR	Promedio	EN	FE	NE	TR	Promedio
1	2	3	[3 52 3]	89.12	86.14	60.52	78.77	78.64	89.47	89.47	68.42	84.21	82.89
1	3	2	[2 79 3]	91.40	91.40	84.03	74.38	85.30	94.74	94.74	84.21	73.68	86.84
2	1	3	[1 22 2]	88.94	85.61	82.28	88.42	86.31	94.74	89.47	84.21	94.74	90.79
2	3	1	[2 56 3]	90.87	83.85	80.70	87.71	85.78	94.74	84.21	78.95	89.47	86.84
3	1	2	[2 35 3]	83.85	83.15	83.50	80.52	82.76	89.47	89.47	84.21	84.21	86.84
3	2	1	[1 11 2]	81.57	83.33	75.43	65.96	76.57	94.74	89.47	73.68	63.16	84.21
Estructura Promedio = [1.8 42.5 2.66] ≈ [2 43 3]				Tasa de Reconocimiento Global =				**82.56**	Tasa de Reconocimiento Global =				**86.40**
				Tasa de Reconocimiento Total =				**83.75**	**Tasa de Reconocimiento Total =**				**87.80**

presenta en la Tabla 5.9 con el sistema propuesto PCA+GA+ANN una tasa de reconocimiento de 94.58 % fue alcanzada con la base de datos JAFFE y 82.41 % con la base de datos MX-Expresiones a través de diferentes esquemas de Entrenamiento-Prueba. En particular la tasa de reconocimiento de 94.58 % para la base de datos JAFFE es consistente con lo presentado en la literatura (véase Tabla 2.1). La integración del GA para optimización de la estructura de la ANN condujo a mejoras estadísticamente significativas sobre las tasas de reconocimiento para ambas bases.

Una situación interesante respecto a la base de datos JAFFE es que fue creada con condiciones adecuadas de postura e iluminación. Estas condiciones no fueron muy estrictas para la creación de la base MX-Expresiones ya que fue considerado el representar a los usuarios en ambientes menos restringidos. Esto puede explicar la diferencia entre las tasas de reconocimiento presentadas en la Figura 5.6 y la Tabla 5.9.

5.3. Integración del Sistema Multimodal con los Sub-sistemas Optimizados

A la forma de adquirir la información del mundo real mediante diferentes canales se le denomina multimodalidad. Para el reconocimiento de emociones la multimodalidad se da con la integración de los sub-sistemas de voz y visión (expresiones). Esto con el objetivo de contemplar las fortalezas de cada sub-sistema independiente para tener una respuesta unificada que sea más efectiva.

Antes de hacer la integración es necesario determinar las mejores configuraciones de los sub-sistemas. De los análisis de desempeño presentados en las Secciones 5.1.1 y 5.2.2 se determinaron como los mejores sub-sistemas de reconocimiento los siguientes:

- Reconocimiento de Voz: Sistema Estándar+GA+HMMs Independiente de Usuario (véase Tabla 5.4).
- Reconocimiento de Expresiones: Sistema PCA+GA+ANN con la Estructura Promedio de 1 Capa Oculta con 57 Neuronas y Función de Activación *purelin* (véase Tabla 5.9). Este es el sistema optimizado para usuarios mexicanos (base de datos MX-Expresiones).

La respuesta de cada sub-sistema se puede representar con un vector fila de cuatro columnas (una columna para cada emoción). Como ejemplo para una frase/expresión de entrada el vector [60.00,20.00, 10.00,10.00] representaría la salida del sub-sistema respectivo, mostrando que el patrón de entrada puede pertenecer a Enojo con 60.00 %, a Felicidad con un 20.00 %, o a Neutro/Tristeza con 10.00 %. Tomando el porcentaje mayor se determinaría que la emoción correcta es Enojo. Sin embargo este porcentaje mayor puede ser engañoso dado que cada sub-sistema podría ser más propenso a confundir una emoción en particular con otra.

En la Figura 5.7 se presenta un resumen de los resultados obtenidos con los sistemas Estándar+GA+HMMs Independiente de Usuario (véase Tabla 5.4) y PCA+GA+ANN con Estructura Promedio para MX-Expresiones (véase Tabla 5.9). En general el sub-sistema de voz reconoce con mayor precisión las emociones de Enojo, Neutro y Tristeza. En cambio el sub-sistema de visión reconoce con mayor precisión la emoción de Felicidad.

Esquema de Prueba B

	Sistema	Enojo	Felicidad	Neutro	Tristeza	Promedio
Sub-sistema de Voz	HMMs Estándar	88.75	83.75	90.63	91.25	88.59
	GA+HMMs	93.13	83.75	92.50	88.75	89.53
	Estándar+GA+HMMs	93.13	83.75	92.50	91.25	90.16

Estructura Promedio [1 57 3] para la ANN

		EN	FE	NE	TR	Promedio
Sub-sistema de Visión	Base de Datos MX	88.89	88.89	44.44	66.67	72.22
		88.89	88.89	77.78	77.78	83.34
		88.89	100.00	77.78	88.89	88.89
		77.78	100.00	77.78	88.89	86.11
		88.89	88.89	77.78	100.00	88.89
		55.56	88.89	77.78	77.78	75.00
		81.48	92.59	72.22	83.34	**82.41**

Figura 5.7: Comparación de tasas de reconocimiento de emociones promedio de los Sub-sistemas Estándar+GA+HMMs (Voz) PCA+GA+ANN (Expresiones).

Una vez obtenido este conocimiento se procedió a diseñar un sistema de ponderación para la unificación de las respuestas de los sub-sistemas (integración del sistema multimodal). En la Figura 5.8 se presenta el esquema de ponderación (asignación de pesos) para las respuestas de cada sub-sistema para cada emoción. Estos pesos se definen como ω_{ij} en donde i es el índice para el sub-sistema y j es el índice para la emoción. Note que para toda j:

$$\sum_{i=1}^{2} \omega_{ij} = 1. \tag{5.1}$$

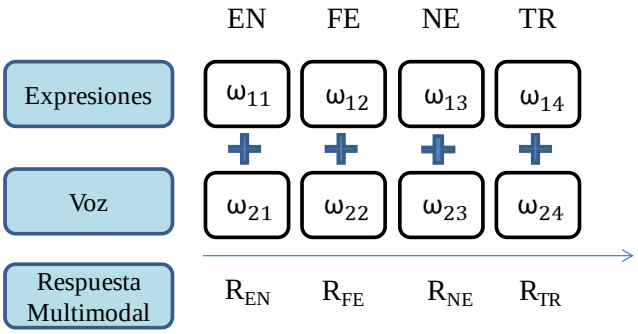


Figura 5.8: Estructura implementada para la integración de los sistemas de voz y expresiones.

De esta manera la respuesta final ($R_{Emocion}$) estaría dada por el porcentaje mayor obtenido al sumar los resultados de los sub-sistemas previamente multiplicados por su

respectivos pesos. En base a la comparativa de desempeños presentada en la Figura 5.7 se definieron los pesos presentados en la Tabla 5.10.

Tabla 5.10: Asignación de pesos para los sub-sistemas para cada emoción.

	Enojo	Felicidad	Neutro	Tristeza
Sistema	EN	FE	NE	TR
Visión	0.20	0.68	0.35	0.15
Voz	0.80	0.32	0.65	0.85

5.4. Interfaz Gráfica de Usuario

Para el manejo en-vivo del sistema multimodal se desarrolló una interfaz en Matlab utilizando la herramienta GUI (Graphical User Interface) [88] la cual permite la edición de ventanas, mensajes de texto, botones, etc. Esta herramienta tiene la capacidad de crear menús, submenús y asemejar su interfaz a cualquier ventana de interacción de un software. En la Figura 5.9 se muestra la interfaz para el reconocimiento multimodal de emociones.

La interfaz cuenta con diversas secciones las cuales se describen a continuación:

1. **Reconocimiento de Usuario:** Este botón tiene la tarea de la identificación del usuario cuya vista previa se muestra en la pantalla.

2. **Rostro de Usuario:** Este campo muestra el rostro encontrado en la pantalla principal. Se aplica el pre-procesamiento del Sistema PCA+GA+ANN.

3. **Corrección de Imagen:** En esta sección se puede activar una ecualización a la imagen o en su defecto aplicar contraste y brillo.

4. **Discurso Reconocido:** En este cuadro se muestran las palabras que fueron reconocidas por el sistema de voz.

5. **Global (Emoción Detectada):** En este recuadro se muestra la emoción detectada en conjunto con ambos sistemas (respuesta del sistema de ponderación).

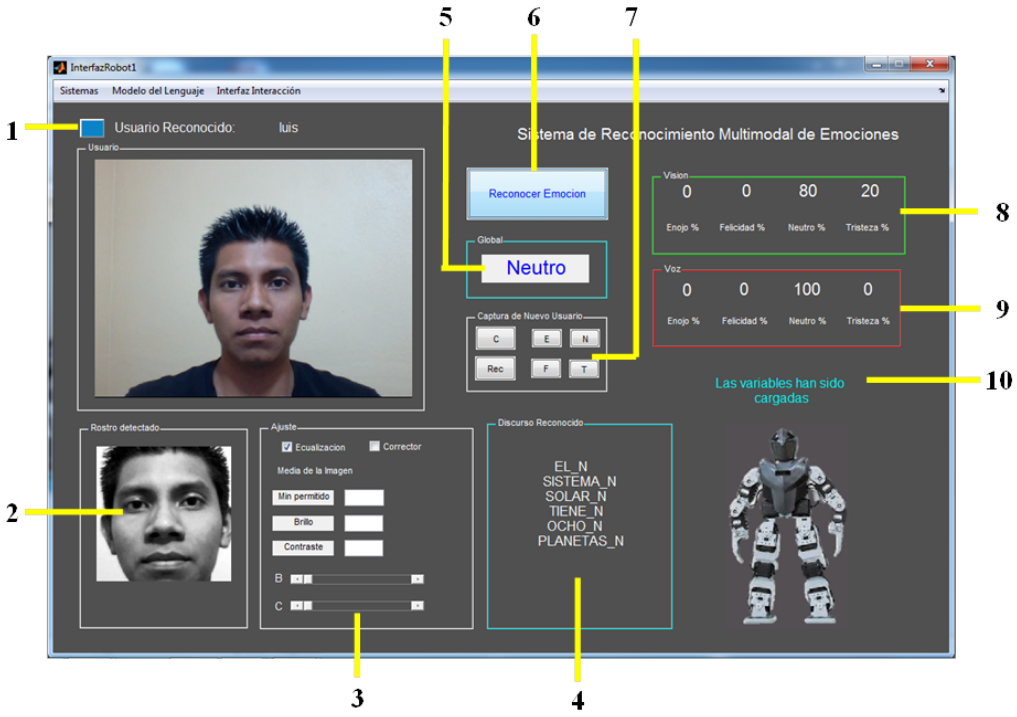


Figura 5.9: Interfaz del Sistema de Reconocimiento Multimodal de Emociones.

6. **Reconocer Emoción:** Con este botón se activa la toma de audio y video y se comienza el procesamiento de la información capturada para el reconocimiento de la emoción.

7. **Captura (Ingreso) de Nuevo Usuario:** En este módulo se realiza el ingreso de un nuevo usuario. En la Figura 5.10 se muestran los botones de este módulo. En lo que respecta al sistema de visión se tiene el botón "C" con el cual se introduce el nombre del nuevo usuario. Los botones con las letras "E", "F", "N" y "T" sirven para capturar muestras visuales del nuevo usuario para cada emoción (esto es, muestras de expresiones de Enojo, Felicidad, Neutro y Tristeza respectivamente). Esto representa una adaptación de usuario para el sub-sistema de visión.

   En lo que concierne a la adaptación del sub-sistema de voz, el botón "Rec" (Grabar) abre una ventana nueva, la cual permite la grabación de muestras de voz para adaptación de usuario. En la Figura 5.11 se muestra la ventana de adaptación que sale al oprimir el botón "Rec". En el campo "Nombre de Usuario" se puede introducir el nombre del usuario que va a grabar las muestras de voz. Los botones

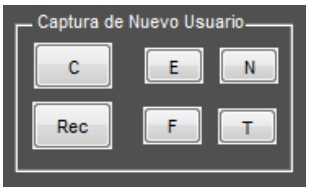


Figura 5.10: Módulo de ingreso de datos y muestras de imagen (expresiones) y voz para nuevos usuarios.

siguientes ("Frases para Enojo, "Frases para Felicidad", "Frases para Neutro" y "Frases para Tristeza") se utilizan para desplegar a su vez ventanas de grabación para las frases de adaptación. En la Figura 5.12 se muestra la ventana de grabación para las frases de estímulo de "Neutro" (las frases de estímulo consistieron de las últimas 10 frases de cada emoción del corpus MX-Voz). La grabación empieza cuando se presiona el boton del texto de estímulo y se detiene al volver a presionar el botón.

Figura 5.11: Ventana de adaptación de voz para nuevo usuario.

Regresando a la Figura 5.11 se tiene también el botón "Empezar la Adaptación". Este botón comienza con el proceso de adaptación MLLR con las frases grabadas para ajustar los HMMs del sistema de voz a las características acústicas del nuevo usuario. Asimismo se tiene el botón para eliminar todas las grabaciones hechas.

8. **Respuesta del Sistema de Visión:** En este cuadro se muestran los resultados del sub-sistema de reconocimiento de expresiones para cada emoción.

Figura 5.12: Ventana de grabación para las frases de adaptación de la emoción “Neutro”.

9. **Respuesta del sistema de Voz:** En este cuadro se muestran los resultados del sub-sistema de voz para cada emoción.

10. **Mensajes (Estado del Proceso):** En este apartado se muestran los mensajes de estado de la interfaz (por ejemplo, cuando las variabes son cargadas, cuando el sistema de visión o de voz ha sido entrenado, etc.).

Después de haber mencionado los componentes principales de la interfaz se procede a mencionar las opciones que se tienen en la barra de “Menú”. En la Figura 5.13 se muestran los sub-menús encontrados en la pestaña de “Sistemas”. En el sub-menú se tiene el entrenamiento de los sub-sistemas de visión y de voz en donde cada sub-sistema tiene sus propios parámetros de configuración. Esto con el propósito de construir los sub-sistemas con otras alternativas de configuración.

Para el sub-sistema de visión se pueden añadir nuevos usuarios para el entrenamiento y calcular las características de cada emoción. Para el entrenamiento de la ANN se pueden definir el número de neuronas, la función de activación/transferencia (FT), el número de capas ocultas, el error mínimo y el método de entrenamiento. En la Figura 5.14 se muestra la ventana de configuración de la ANN. La interfaz puede guardar la

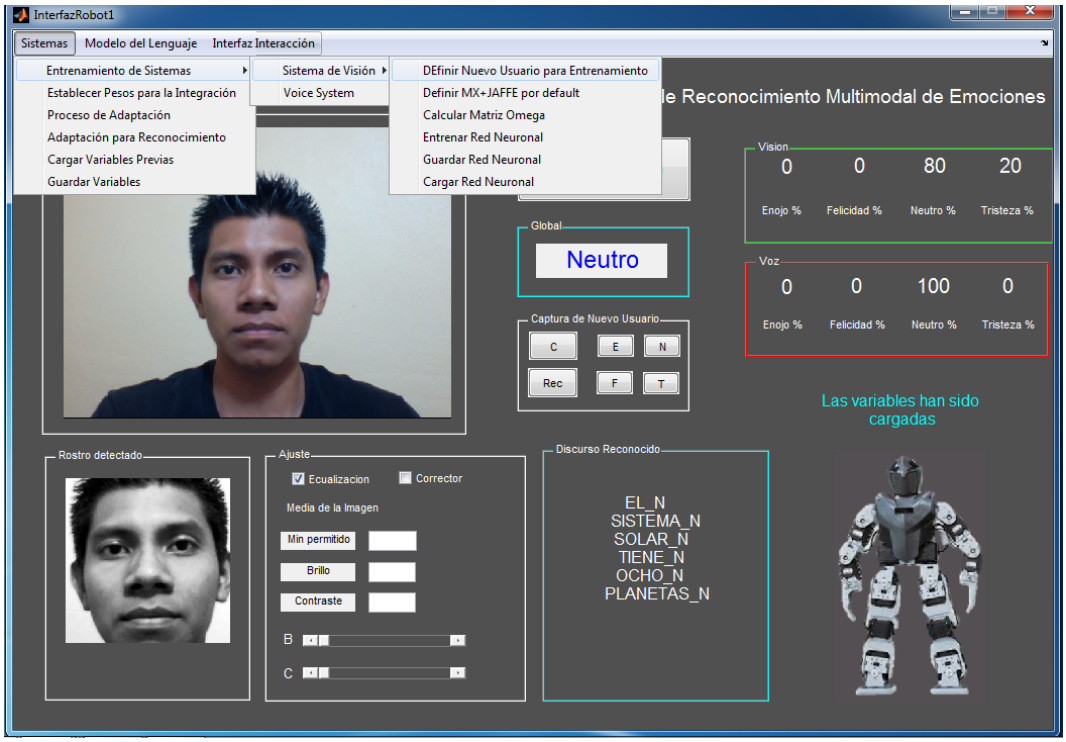


Figura 5.13: Menú de opciones para los sub-sistemas.

nueva ANN entrenada o puede cargar una ya existente (que haya sido guardada con anterioridad).

Dentro de las opciones se tiene el proceso de adaptación, el cual es usado para los usuarios originales del corpus MX-Voz. En este apartado se puede añadir el nombre de usuario para comenzar las transformaciones lineales, la ventana es mostrada en la Figura 5.15. Otra opción que se tiene es el ingreso del nombre del usuario para el reconocimiento.

También se tiene la opción de establecer los pesos para la ponderación de los subsistemas. Esto para realizar un ajuste de las ponderaciones si hay un cambio en las técnicas de reconocimiento en el futuro. La ventana para establecer los pesos es mostrada en la Figura 5.16.

En la Figura 5.17 se muestran las opciones del Modelo de Lenguaje para el subsistema de voz. Entre las opciones está la asignación del valor “Sigma” el cual controla la influencia del Modelo de Lenguaje en el proceso de reconocimiento de voz [97]. La opción de añadir nueva frase está disponible así como la opción para borrar frases.

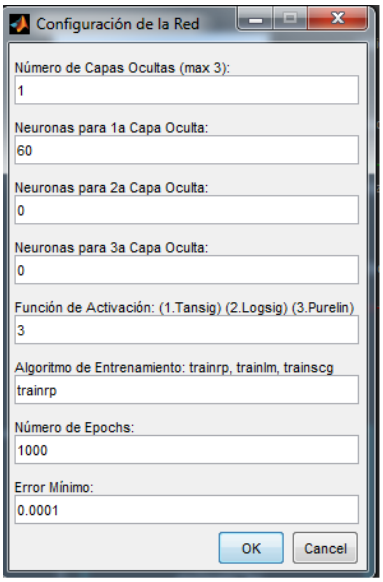


Figura 5.14: Configuración de parámetros para el entrenamiento de la ANN del subsistema de visión.

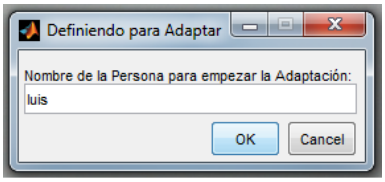


Figura 5.15: Ventana para adaptación de usuario del corpus MX-Voz.

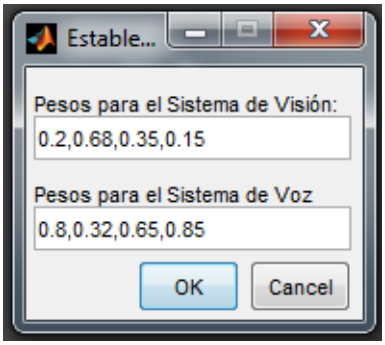


Figura 5.16: Ventana para edición de pesos.

Figura 5.17: Opciones de menú para el modelo de lenguaje.

Capítulo 6

Sistema de Diálogo y Resultados de Interacción Multimodal con el Robot Humanoide

Una vez integrado el sistema de reconocimiento multimodal de emociones se procedió a desarrollar el medio para poderse usar dentro de una interacción con un robot. Este medio consiste en un sistema de diálogo para la administración de la información multimodal para generar una respuesta en el robot.

En el presente capítulo se describe el sistema de diálogo y las librerías de movimientos de respuesta para el robot humanoide Bioloid los cuales se integran dentro de la interfaz del sistema multimodal. De igual manera se presentan los resultados de pruebas en-vivo del sistema multimodal final con usuarios mexicanos diferentes de los usuarios de las bases de datos MX-Expresiones y MX-Voz. En estas pruebas se obtuvo una tasa de reconocimiento de emociones promedio de 97.00 %.

6.1. Sistema de Diálogo

El sistema de diálogo fue desarrollado a base de FSMs. Para la implementación de este sistema se utilizó la herramienta FSM Library [3].

Como se presentó en la Sección 2.2.4 una FSM se puede utilizar para el desarrollo de sistemas de diálogo en donde una entrada (palabra o secuencia de palabras de estímulo)

tiene asociada una acción (palabras de respuesta). Para este trabajo se tomaron como símbolos de entrada las palabras generadas por el reconocedor de voz y como salida un conjunto de etiquetas asociadas a campos semánticos dentro del contexto de una conversación informal acerca del dia del estudiante.

Por ejemplo, para iniciar una conversación usualmente se tienen las siguientes palabras / frases: “Hola”, “Qué onda”, “Hola, ¿qué tal?”, etc. Para este tipo de frases de entrada se asocia una etiqueta denominada como “SALUDO” la cual denota su significado semántico. Esta etiqueta representa el alfabeto de salida la cual tiene asociada una secuencia de acciones a realizar por el robot (esto es, dar una contestación acorde a lo que dijo el usuario, en este caso devolver el saludo). En la Figura 6.1 se muestra la estructura de los FSMs para las respuestas del sistema de reconocimiento de voz.

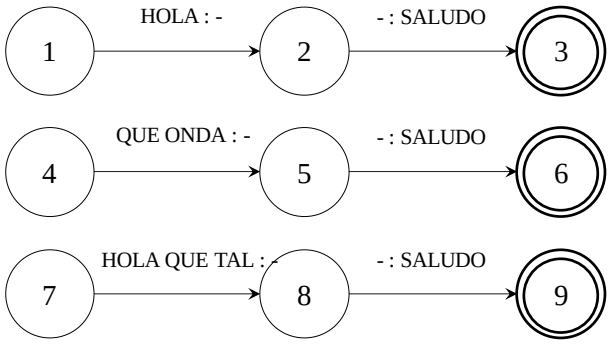


Figura 6.1: Estructura de FSMs para frases de entrada y etiquetas de salida.

Es importante mencionar que dadas las diversas frases de entrada y posibles respuestas la unión de todos los FSMs puede dar como resultado una red de FSMs muy extensa. Por lo tanto es necesario hacer una reducción y simplificación de la red final (esto es, eliminación de nodos y transiciones redundantes). Para esto se aplicaron las operaciones de minimización y determinización descritas en la Sección 2.2.4. Estas operaciones se realizaron con el FSM Library. En la Figura 6.2 se muestra la red simplificada de la unión de los FSMs mostrados en la Figura 6.1.

6.1.1. Frases de Diálogo: Día en la Escuela

Para tener un compendio de las diferentes frases que se pudieran tener en una conversación dentro del contexto de “Un Día en la Escuela” se hizo una entrevista a seis

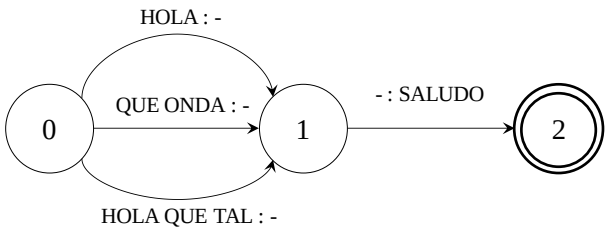


Figura 6.2: Estructura de FSM simplificado con operaciones de determinización y minimización.

usuarios y se hizo una revisión en blogs de internet acerca del tema. En la Figura 6.3 se presenta un ejemplo de las frases que se pueden tener en un relato dentro del contexto y las posibles respuestas que puede tener una persona (o en este caso el robot) con base a la emoción detectada. Estas conversaciones forman la base de los FSMs para el sistema de diálogo.

Es importante mencionar que las frases existentes en los diálogos se incluyen en el Modelo de Lenguaje del sistema de reconocimiento de voz ya que representan vocabulario adicional. El transcriptor fonético descrito en la Sección 3.2.2 fue importante para crear el Diccionario Fonético para este vocabulario.

Como se puede observar en la Figura 6.3, se tienen diversos caminos viables durante el desarrollo de la conversación. Las respuestas del robot consisten de frases de respuesta y actividades (por ejemplo, bailar) las cuales están asociadas a las etiquetas identificadas por la red de FSMs. Las frases de respuesta se manejan mediante un sintetizador de voz mientras que las actividades se realizan físicamente por el robot. Esto se aborda en la Sección 6.2.3. El sintetizador de voz utilizado para el robot fue “Acapela Text to Speech (Demo)” con el lenguaje “Spanish (Us)” y la voz “Valeria”.

Para dotar al robot de un sentido de seguimiento de la conversación se realizó un programa para que llevara la secuencia de la plática (etiquetas de salida de la red de FSMs) y evitar repetir preguntas. Es decir, se tomaron en cuenta los sucesos previos (respuestas del usuario) para poder dar una pregunta de seguimiento diferente cada vez. Finalmente en la Figura 6.4 se muestra la visualización de un fragmento de la red de FSMs del sistema de diálogo. En los arcos únicos finales (que van al nodo “36”) se encuentran las etiquetas de significado de la frase de entrada (que inician en el nodo “0”) y las cuales sirven para determinar la acción a realizar.

Extracto de Frases de Diálogo Humano-Robot

U: Usuario ***R: Robot***

R: Hola ¿Cómo estás?

U:

(1) Estoy Bien ¿y tú? / Bien ¿y tú? / Excelente ¿y tú? / Mal ¿y tú? /
(2) Bien gracias / bien
(3) Estoy…/ me siento…/ Cansado/ Feliz/ Enojado/Triste
(4) Mal, reprobé/ Mal no pase una materia
(5) Pues más o menos

R:

(1) Estoy bien y cuéntame ¿cómo te fue en la escuela?
(2) Qué bueno, me da gusto, y cuéntame ¿cómo te fue en la escuela?
(3) *Cansado* ***: Descansa un poco, relájate y duerme un rato para que repongas pilas.***
Feliz ***: Que bueno yo también… ¿quieres ver cómo me pongo de cabeza y me equilibro?***
Enojado ***: Trata de calmarte, siempre hay solución… checa estas rutinas para liberar tensión.***
Triste ***: No estés así, ¿que tal si bailo un rato?***
(4) Que malo, pero levanta el ánimo que puedes mejorar, ¿quieres verme hacer lagartijas?
(5) ¿Quieres que te cuente un chiste para animarte?

U:

(1) Me fue muy bien / Me fue mal / Bien / Mal /Pues bien / Pues mal
(2) Me fue muy bien / Me fue mal / Bien / Mal /Pues bien / Pues mal
(3) Cansado: Si tienes razón / Si descansaré un rato/ No, tengo otras cosas que hacer/
Feliz: OK / OK está bien / Si / Si quiero/ Sale vale/ A ver/ Va
Enojado: OK / No / No quiero /Deja de molestarme/ Ya cállate / Deja de hablar
Triste: OK está bien. / No gracias /Quiero estar solo
(4) OK / No quiero / Pero era mi examen final /
(5) OK/ OK está bien / A ver / Si/ Si quiero/ Órale / ¡Va! / Sale vale / No gracias / No me hables

Figura 6.3: Ejemplo de un diálogo entre el usuario humano y el robot considerando las emociones detectadas.

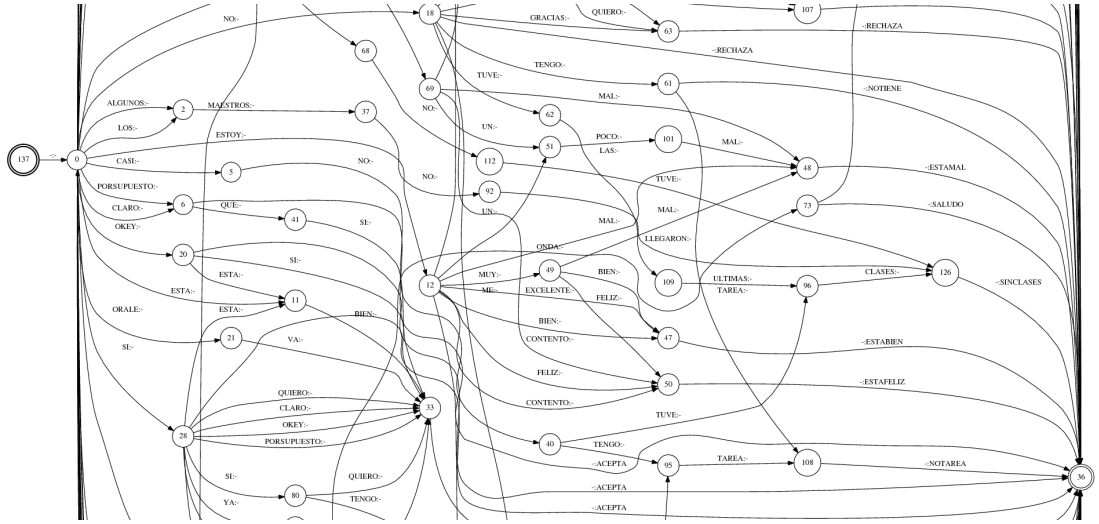


Figura 6.4: Fragmento de FSMs del sistema de diálogo.

6.2. Enlace de la Interfaz Multimodal con el Robot Humanoide Bioloid

6.2.1. Conexión Robot-Computadora

El sistema de comunicación estuvo basado en el módulo Zigbee para transmitir información entre la computadora y el robot. Para establecer una comunicación entre los dos sistemas es necesario tener los siguientes dispositivos: (a) USB2Dynamixel (convertidor USB a serial), (b) Zig2Serial (tarjeta acondicionadora en donde se monta el controlador Zig-100), (c) Zig-100 (Zigbee para la laptop), (d) Zig-110A (Zigbee para el robot) y (e) CM-530 (CPU del robot) [71].

Sin embargo en este caso en lugar de utilizar el USB2Dynamixel y Zig2Serial se optó por utilizar el convertidor USB-a-TTL serial. Esto dado que el convertidor ya cuenta con las salidas de 3.3V y 5V que requiere el dispositivo Zig100 lo cual lo hace más práctico y económico [66]. En la Figura 6.5 se muestra la arquitectura de la conexión del robot humanoide Bioloid con la interfaz multimodal en la computadora usando los dispositivos Zigbee y USB-a-TTL. El cable verde no es usado dado que es la salida de +5V, en cambio el cable amarillo si es usado dado que es de +3V que es el que necesita el Zig100. El cable naranja es Tierra (GND), el cable café conecta la terminal de transmisión (TX) del USB a la terminal de recepción (RX) del Zig100, y el cable rojo conecta la terminal de transmisión (TX) del Zig100 a la terminal de recepción (RX) del USB. En la Figura 6.6 se presenta la conexión final.

Finalmente el SDK de Robotis [73] fue utilizado para realizar la comunicación de datos entre la computadora y el robot con la conexión USB-Zigbee.

6.2.2. Creación de Movimientos con RoboPlus

Para desarrollar los movimientos del robot humanoide se utilizó el software RoboPlus que provee la empresa Robotis. La interfaz RoboPlus, mostrada en la Figura 6.7 contiene diferentes apartados que ayudan a realizar tareas específicas como la programación del robot, creación de movimientos o configuración de sus componentes como servomotores y sensores.

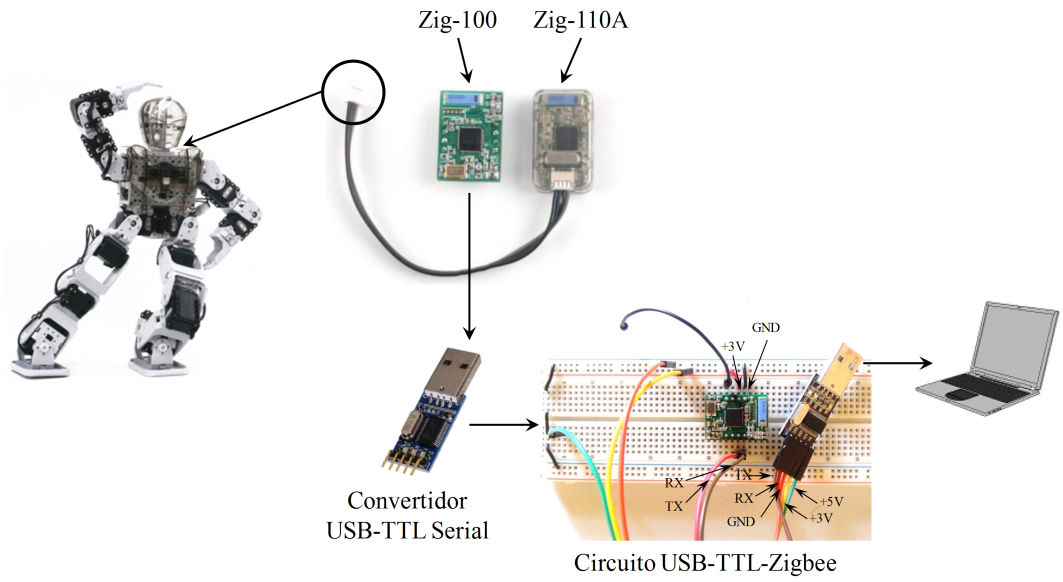


Figura 6.5: Conexión USB-Zigbee del robot humanoide Bioloid.

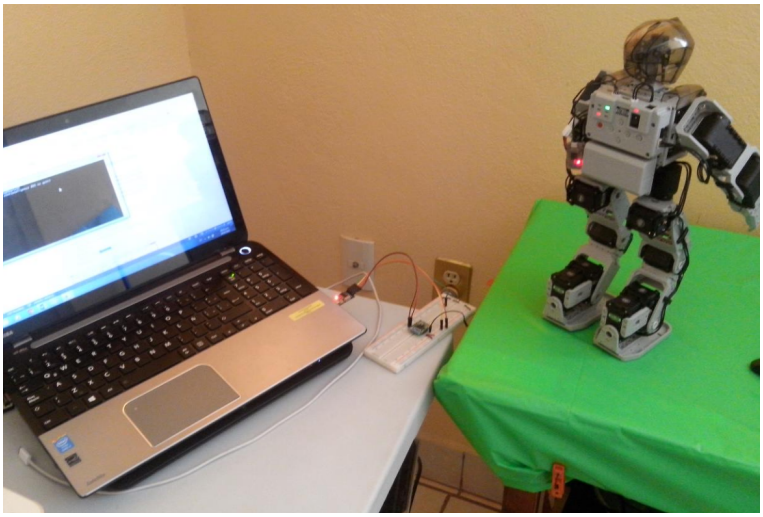


Figura 6.6: Conexión Final USB-Zigbee del robot humanoide Bioloid.

Figura 6.7: Interfaz Roboplus.

El robot Bioloid en su modalidad humanoide necesita de dos archivos para su correcto funcionamiento:

- Task (.TSK) File (Archivo de Tarea): Este archivo define las acciones que debe ejecutar el robot como son la lectura de datos recibidos, activación de sensores, realización de movimientos definidos por archivos .MNT, etc. En el programa .TSK se pueden añadir estructuras de programación como saltos y llamadas a funciones de una manera muy semejante a la programación en el lenguaje C. Para escribir un programa (archivo .TSK) se debe acceder a "RoboPlus Task" (véase Figura 6.7). En la Figura 6.8 se muestra un ejemplo de un programa .TSK en donde de acuerdo a un valor recibido (513, 514, 516) se ejecuta un movimiento identificado mediante un índice (226, 227, 228). Estos movimientos se encuentran almacenados en un archivo .MNT el cual se describe a continuación.

- Motion (.MNT) File (Archivo de Movimiento): Este archivo se encarga de definir los movimientos del robot los cuales se modelan a base de secuencias de pasos. El archivo .MTN se construye accediendo a "RoboPlus Motion" (véase Figura 6.7).

```
// Mis propios movimientos_____________________
ELSE IF ( ReceiveData  ==  513 )
{
    // Porque_____________________
    Motion Index Number  = 226
    CALL WaitMotion
}
ELSE IF ( ReceiveData  ==  514 )
{
    // Asustado_____________________
    Motion Index Number  = 227
    CALL WaitMotion
}
ELSE IF ( ReceiveData  ==  516 )
{
    // Saludo_____________________
    Motion Index Number  = 228
    CALL WaitMotion
```

Figura 6.8: Archivo .TSK: Recepción de datos y llamadas a índices de movimientos.

Un archivo .MNT está constituído por 255 índices o "páginas" y cada una de ellas es capáz de almacenar hasta 7 pasos que pueden definir un movimiento completo. Si un movimiento necesita de más pasos para ser completado es posible conectar más "páginas".

Cada página posee parámetros adicionales que pueden cambiarse, como el número de veces que se puede repetir una página (para repetición de movimientos), la velocidad con la que se reproduce la página (para movimientos rápidos/lentos), y la fuerza inercial (para movimientos fuertes/suaves). En cambio para cada paso se pueden establecer dos tipos de parámetros: (a) "pausa", el cual se refiere al tiempo entre el final del paso actual y el inicio del siguiente paso (su valor oscila entre 0 y 2.04 segundos), y (b) "tiempo", el cual es el tiempo que debe de durar el paso actual desde su inicio hasta su fin (su valor puede oscilar entre 0.072 y 2.04 segundos). En la Figura 6.9 se presenta el archivo .MNT asociado al archivo .TSK mostrado en la Figura 6.8.

En el archivo .TSK se observan las llamadas a los números de páginas de movimientos con el comando "Motion Index Number". En la Figura 6.9 se muestra la ventana de edición del archivo de movimientos en donde se muestran las páginas 226, 227 y 228 que representan las expresiones de "¿Porque?", "Asustado" y "Saludo" respectivamente. También se presenta un ejemplo de conexión entre 7 páginas (229 a la 235) para movimientos de "Flexiones". Esta conexión se realiza poniendo el número de página que se quiere unir con la actual en la casilla

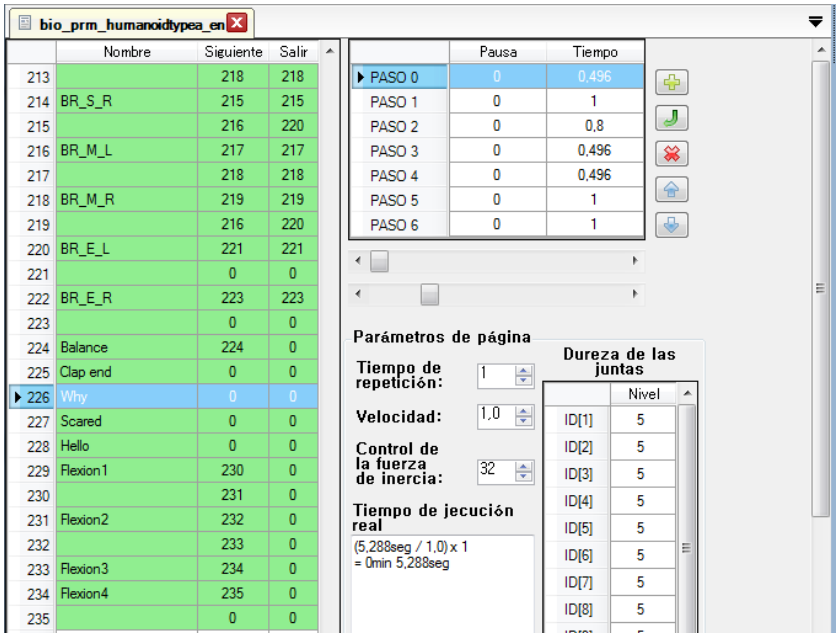


Figura 6.9: Archivo .MNT: Páginas de movimientos y pasos correspondientes.

"Siguiente".

6.2.3. Conjunto de Movimientos para Sistema de Diálogo

La interacción entre el robot y el usuario está dada por el diálogo entre ambos el cual depende de la emoción detectada. Dentro de las respuestas del robot se consideran movimientos que enfatizan las expresiones que conteste el robot. Por lo tanto el sistema de diálogo primero identifica la frase y el estado de ánimo del usuario, y después en base a la emoción detectada propone una rutina a realizar por el robot que sea más acorde al contexto de la situación.

De esta manera se realizaron los siguientes conjuntos de movimientos de respuesta emocional:

- **Rutina de Presentación:** Cuando se detecta un usuario el robot se presenta, saluda y describe el propósito que tiene, el de platicar para liberar un poco las tensiones que pudieran tenerse en un día escolar de un estudiante.

- **Expresiones:**

a. Interrogativa: Abre sus brazos a los costados expresando la pregunta "¿por qué?".

b. Asustado: Cubre con sus manos su rostro.

c. Aplausos: Alza los brazos y aplaude encima de su cabeza. Esta rutina es usada cuando la emoción de "Felicidad" es detectada.

d. Afirmación: Alza y baja varias veces el brazo derecho para reforzar alguna frase.

- **Rutinas de Ejercicio:** Se proponen ejercicios de estiramientos por parte del robot para inducir al usuario a hacerlos ya que ayudan a mantenerse flexible, reducen el dolor muscular, ayudan a evitar lesiones por movimientos bruscos y alivian la tensión originada por el día a día [15, 86]. Además las rutinas de estiramiento reducen el estrés ocasionado por estar sentados todo el día frente a una computadora [23]. Estas rutinas fueron propuestas en base a información encontrada en sitios de salud [28, 68] y se consideraron apropiadas para la emoción de "Enojo".

  a. Estiramientos:

    - Laterales con una mano alzada.
    - Flexión de mano rodilla de cada lado.
    - Estiramientos de los brazos hacia atrás.
    - Elevaciones de brazos con las manos tomadas por la parte posterior.

  b. Lagartijas: Posición boca abajo realizando tres flexiones con los codos para bajar y elevar el cuerpo.

  c. Baile: Realización de varios pasos para ejecutar una mini coreografía de una canción. Esta rutina es usada cuando la emoción de "Tristeza" es detectada.

- **Rutina de Distracción con Voz:** Rutina acompañada de voz y considerada especialmente para "Neutro".

  a. Cuenta Chistes Diversos: Elevación de brazos y movimientos cortos de piernas.

6.3. Resultados de Interacción en Pruebas En-Vivo

6.3.1. Tasa de Reconocimiento Emocional Multimodal

Para la validación final de los sub-sistemas de reconocimiento de emociones y del sistema multimodal se realizaron pruebas en-vivo con nuevos participantes mexicanos. Previo a esta etapa se llevó a cabo una recabación de muestras para adaptación de usuario dado el enfoque IU (Independiente de Usuario) de los sub-sistemas. En la Tabla 6.1 se presentan los datos generales de estos nuevos participantes (5 Hombres y 5 Mujeres).

Tabla 6.1: Perfiles de los participantes para la prueba final del Sistema Multimodal (Pruebas En-Vivo).

Usuario	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
Género (Hombre, Mujer)	H	H	M	H	H	H	M	M	M	M
Edad (Años)	27	40	22	26	28	27	44	40	45	28
Estado de Origen	Oaxaca	Oaxaca	Oaxaca	Oaxaca	Oaxaca	Oaxaca	Veracruz	Veracruz	Veracruz	Veracruz

Mediante las interfaces presentadas en la Figura 5.10 y 5.11 se grabaron las muestras de expresión y voz para los sub-sistemas respectivos. Para voz el material de adaptación consistió de las últimas 10 frases (11-20) de cada emoción del corpus MX-Voz y la evaluación se realizó con las primeras 10 frases (1-10) de cada emoción (véase Esquema de Prueba B, Sección 5.1.1). En la Tabla 6.2 se muestran los resultados de la evaluación del sistema de reconocimiento multimodal de emociones (Interfaz Multimodal, Figura 5.9) para los diferentes usuarios y las 10 frases de prueba.

De la Tabla 6.2 se puede observar una tasa de reconocimiento total de 97.0 % considerando a todos los usuarios y todas las emociones. Analizando las identificaciones incorrectas del estado emocional se puede ver que las emociones de "Felicidad'" y "Tristeza" son las que presentan más errores en su reconocimiento. Sin embargo para ambas la tasa de reconocimiento promedio es mayor o igual al 94.0 %. Note que la evaluación final fue realizada con 400 frases (10 usuarios $\times$ 10 frases $\times$ 4 emociones).

Tabla 6.2: Tasa de reconocimiento promedio del Sistema Multimodal de Emociones (Pruebas En-Vivo).

U	Frases de Enojo										Frases de Felicidad										Frases de Neutro										Frases de Tristeza										#	Tasa de Reconocimiento
	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10		
U1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	40	100.0
U2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	38	95.0
U3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	38	95.0
U4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	40	100.0
U5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	40	100.0
U6	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	39	97.5
U7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	39	97.5
U8	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	37	92.5
U9	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	40	100.0
U10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	37	92.5
	100.0%										**96.0%**										**98.0%**										**94.0%**											**97.0%**

U : Usuario

: Número de Identificaciones Correctas de la Emoción

✓ Identificación Correcta de la Emoción

✗ Identificación Incorrecta de la Emoción

Tasa de Reconocimiento Total

6.3.2. Prueba de Interacción Multimodal

A continuación se muestran diferentes capturas de un ejemplo de interacción realizada con el robot humanoide. También se muestra la evolución del diálogo entre el usuario y el robot con sus respectivas rutinas dependiendo de la emoción detectada.

- **Pregunta 1**
  - **Robot:** "Hola, ¿Cómo estás?"
  - **Usuario:** "Estoy mal (Enojado)"
  - **Robot:** "No estés molesto, mira, fíjate en esta rutina para liberar estrés. Empezamos levantando el brazo, y nos estiramos hacia este lado dos veces, ahora levantamos el otro brazo y repetimos lo mismo que para el primero y así vamos liberando tensiones a través del ejercicio, y finalmente estiramos los dos brazos hacia atrás, de verdad se siente el relajamiento, bueno espero y te hayan gustado estos estiramientos" + (RUTINA FLEXIONES). La ejecución de la rutina se presenta en la Figura 6.10.
- **Pregunta 2**
  - **Robot:** "Y cuéntame, ¿Cómo te fue en la escuela?"
  - **Usuario:** "Me fue mal (Enojado)"
  - **Robot:** "Quizá y unas cuantas lagartijas ayuden a contrarrestar el mal humor. Chécate cómo las hago. Con fuerza y firmeza manteniendo una postura con

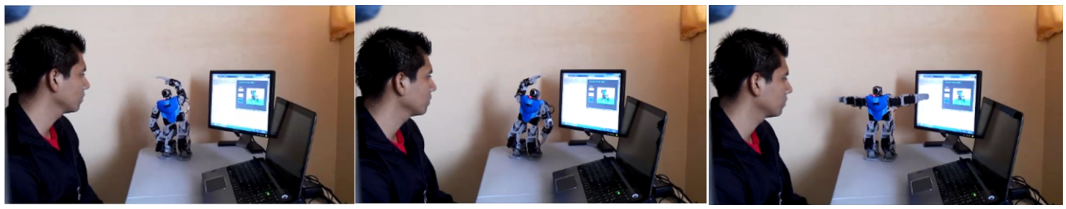


Figura 6.10: Pregunta 1: Diálogo y ejecución de la rutina “Flexiones”.

la espalda recta, que se sienta el esfuerzo, esto deberás hacer para quitar ese enojo” + (RUTINA LAGARTIJAS). La ejecución de la rutina se presenta en la Figura 6.11.

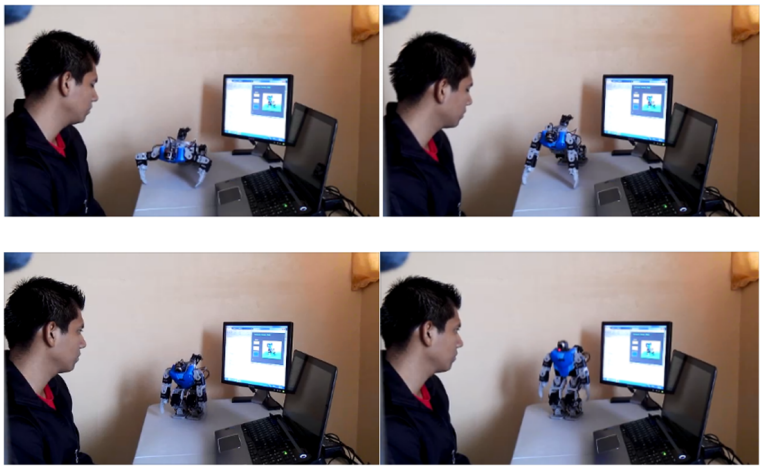


Figura 6.11: Pregunta 2: Diálogo y ejecución de la rutina “Lagartijas”.

- **Pregunta 3**
  - **Robot:** “Oye quiero saber algo, ¿Tienes exámenes?”
  - **Usuario:** “Si, si tengo (Triste)”
  - **Robot:** “Yo diría que te pusieras a estudiar en un rato más para que saques una muy buena calificación. Oye quiero que escuches este chiste” + (CHISTE1) + “Espero y te haya gustado, porque a mi si me gustó”. La ejecución de la rutina se presenta en la Figura 6.12.
- **Pregunta 4**
  - **Robot:** “Y dime, ¿Te encargaron mucha tarea? O casi no”

Figura 6.12: Pregunta 3: Diálogo y ejecución de la rutina “Chiste1”.

- **Usuario:** “Si, me encargaron mucha tarea (Triste)”
- **Robot:** “Pues ponte a hacer tu tarea y verás que terminarás pronto, Ahora checa qué canción bailaré, es muy conocida, espero que te rías al verme bailar” + (BAILE1) + “Bueno espero y te haya gustado mi baile y te hayas divertido”. La ejecución de la rutina se presenta en la Figura 6.13.

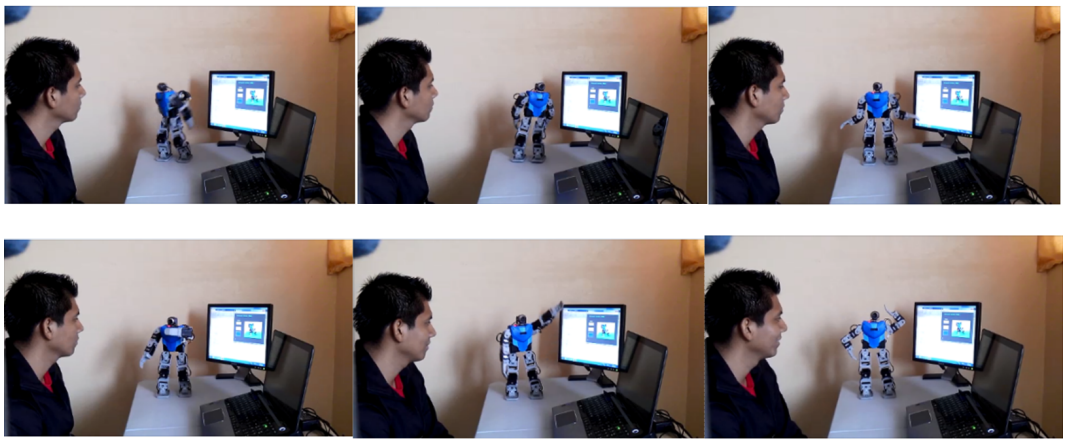


Figura 6.13: Pregunta 4: Diálogo y ejecución de la rutina “Baile1”.

- **Pregunta 5**
  - **Robot:** “¿Y tu tuviste todas tus clases?”
  - **Usuario:** “Si tuve todas (Enojado)”
  - **Robot:** “Que bien que tuviste todas tus clases, eso habla muy bien de tu escuela, aunque un descanso no cae nada mal (risas del robot). No te enojes, chécate este estiramiento” + (RUTINA ESTIRAMIENTO) + “Nos inclinamos sobre este lado flexionando la rodilla y después hacia el otro realizando

el mismo procedimiento, hay que tener en cuenta que debemos de ir respirando profundamente eso es de gran ayuda, ya para lo último llevamos los brazos hacia atrás para estirar parte de los hombros y pectorales. Así se debe de hacer, no es tan difícil y puede ayudar a la relajación de los músculos". La ejecución de la rutina se presenta en la Figura 6.14.

Figura 6.14: Pregunta 5: Diálogo y ejecución de la rutina "Estiramiento".

- **Pregunta 6**
  - **Robot:** Oye, "¿Y saliste temprano?"
  - **Usuario:** "Si salí temprano (Neutro)"
  - **Robot:** "Que bien que saliste temprano, así tuviste tiempo ya sea de llegar a tu casita antes o salir con tus amigos a dar la vuelta un rato. Tengo guardado otro chiste para que nos sigamos divirtiendo" + (CHISTE2). La ejecución de la rutina se presenta en la Figura 6.15.

Figura 6.15: Pregunta 6: Diálogo y ejecución de la rutina "Chiste2".

- **Despedida**
  - **Robot:** "Bueno pues espero y te hayas relajado y pasado bien durante nuestra conversación, me despido y siempre recuerda ser feliz". La ejecución de la rutina se presenta en la Figura 6.16.

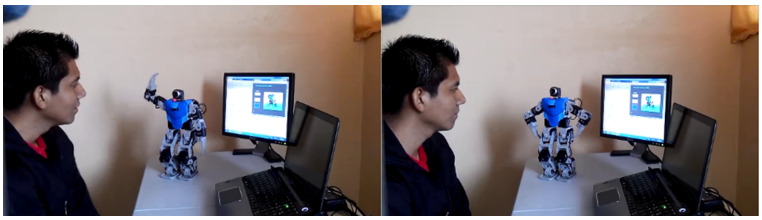


Figura 6.16: Diálogo y ejecución de la rutina “Despedida”.

Alternativamente durante una segunda interacción se empezó el diálogo con un estado emocional diferente al mostrado inicialmente. En este caso se respondió con “Felicidad” a las Preguntas 2 y 6:

- **Pregunta 2**
  - **Robot:** “Y cuéntame, ¿Cómo te fue en la escuela?”
  - **Usuario:** “Me fue muy bien (Feliz)”
  - **Robot:** “Mira como aplaudo, todos debemos estar así, siempre manteniendo una sonrisa en nuestra cara, además de que es bueno para el cuerpo” + (RUTINA APLAUSO). La ejecución de la rutina se presenta en la Figura 6.17.

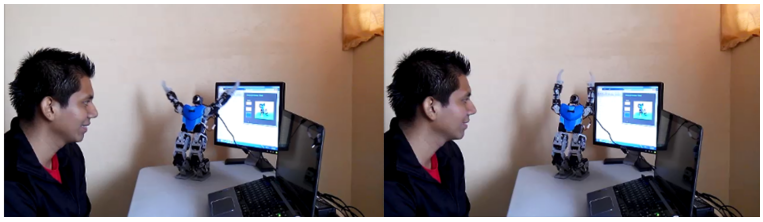


Figura 6.17: Pregunta 2: Diálogo y ejecución de la rutina “Aplauso”.

- **Pregunta 6**
  - **Robot:** “Oye, ¿Y saliste temprano?”
  - **Usuario:** “Si salí temprano (Feliz)”

- **Robot:** “Que bien que saliste temprano, así tuviste tiempo ya sea de llegar a tu casita antes o salir con tus amigos a dar la vuelta un rato. Hago como King Kong porque estoy igual de feliz que tu (risas robot)” + (RUTINA KING KONG). La ejecución de la rutina se presenta en la Figura 6.18.

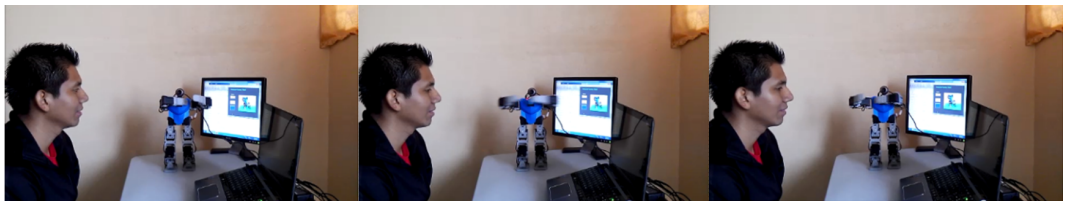


Figura 6.18: Pregunta 6: Diálogo y ejecución de la rutina “King Kong”.

Capítulo 7

Conclusiones y Trabajo a Futuro

En la Tabla 7.1 se muestra una revisión de todos los sistemas desarrollados en esta tesis. Se presentan las tasas de clasificación global para las cuatro emociones consideradas (Enojo, Felicidad, Neutro, Tristeza).

El Sistema de Reconocimiento Multimodal de Emociones estuvo conformado por el sub-sistema de voz "Estándar+GA+HMMs" y el sub-sistema de visión "PCA+GA+ANN" dado a que estos sistemas presentaron la máxima tasa de clasificación con las bases de datos creadas con usuarios mexicanos (MX-Voz y MX-Expresiones). Los resultados obtenidos para el desarrollo del sub-sistema de visión mostraron que el reconocimiento de una emoción puede ser dependiente de las características de la base de datos y de la técnica de reconocimiento misma.

En cuanto al sub-sistema de voz se determinó que el enfoque IU puede ofrecer tasas de reconocimiento mayores que el enfoque DU. Para el enfoque IU es necesario tener una etapa de adaptación la cual requiere menor tiempo que el entrenamiento de un sistema DU. Una característica importante a comentar es que el reconocimiento de emociones tanto para hombres como para mujeres estuvo equilibrado (es decir, tuvieron tasas de reconocimiento similares). Esto a pesar de que la base de datos MX-Voz estuvo conformada en su mayoría por Mujeres. Esto conlleva a recalcar el buen modelado de las emociones sin importar el género mediante el enfoque de vocales específicas emotivas y la técnica de adaptación de usuario MLLR (Regresión Lineal de Máxima Probabilidad).

El desempeño de estos sub-sistemas fue mejorado mediante el uso del método de computación evolutiva de los GA. La integración de los GAs para mejorar las técnicas

Tabla 7.1: Revisión de sistemas desarrollados para el reconocimiento de emociones.

Fuente	**Sub-sistemas de Visión (Expresiones Faciales)**	**Tasa de Reconocimiento (%)**		**Global**
Tabla 5.6	ANN Preliminar con GA1 (JAFFE)	75.00		
Tabla 5.6	ANN Preliminar con GA2 (JAFFE)	77.50		
Tabla 5.7	ANN Preliminar con GA2+ANN Correctiva (JAFFE)	85.00		
Figura 5.6 (b)	PCA (MX-Expresiones)	72.22		
Figura 5.6 (b)	PCA (JAFFE)	83.33		78.25
Figura 5.6 (b)	PCA (MX-Expresiones+JAFFE)	78.25		
Figura 5.6 (b)	PCA+ANN (MX-Expresiones)	69.91	Estructura ANN	
Figura 5.6 (b)	PCA+ANN (JAFFE)	83.33	Fija	77.90
Figura 5.6 (b)	PCA+ANN (MX-Expresiones+JAFFE)	80.48	[2 40 2]	
Figura 5.6 (b)	PCA+GA+ANN (MX-Expresiones)	79.36	Estructura ANN	
Figura 5.6 (b)	PCA+GA+ANN (JAFFE)	89.32	Específica con	83.75
Figura 5.6 (b)	PCA+GA+ANN (MX-Expresiones+JAFFE)	82.56	GA	
Tabla 5.9	PCA+GA+ANN (MX-Expresiones)	82.41	Estructura ANN	
Tabla 5.9	PCA+GA+ANN (JAFFE)	94.58	Promedio con GA	87.80
Tabla 5.9	PCA+GA+ANN (MX-Expresiones+JAFFE)	86.40		

Fuente	**Sub-sistemas de Voz**	**Tasa de Reconocimiento (%)**	
		Esquema A (Dependiente de Usuario)	**Esquema B (Independiente de Usuario)**
Tabla 5.2	HMMs Estándar (MX-Voz)	78.59	88.59
Tabla 5.3	GA+HMMs (MX-Voz)	82.34	89.53
Tabla 5.4	Estándar+GA+HMMs (MX-Voz)	82.19	90.16

Fuente	**Sistema Multimodal**	**Tasa de Reconocimiento (%)**
Tabla 6.2	Con los sub-sistemas: Estándar+GA+HMMs (Voz) PCA+GA+ANN (Expresiones Faciales)	97.00

de clasificación dio mejores tasas de reconocimiento en comparación con los sistemas cuyos parámetros son estimados de manera empírica o basados en trabajos previos.

La integración del GA para optimización de la estructura de la ANN para el sub-sistema de visión condujo a mejoras estadísticamente significativas sobre las tasas de reconocimiento para las bases de datos MX-Expresiones y JAFFE (base estándar). El Sub-sistema "PCA+GA+ANN" obtuvo las mayores tasas de reconocimiento de emociones en expresiones faciales a través de diferentes esquemas de Entrenamiento-Prueba. Sin embargo de manera particular se encontró que algunos sub-sistemas como PCA pueden ser más eficientes para alguna emoción en particular (por ejemplo, PCA para "Tristeza" y ANN para "Enojo", "Felicidad" y "Neutro").

Para el sub-sistema de voz las estructuras de HMMs estimadas con el GA estadísticamente mejoraron el desempeño del reconocimiento de emociones bajo los esquemas IU y DU. Las estructuras para HMMs encontradas por el GA para el modelado de vocales específicas emotivas consistieron de una combinación de las estructuras Bakis Tipo A, Bakis Tipo B y Ergódica, en donde la estructura Bakis Tipo B tuvo más presencia. Sin embargo la estructura estándar, Bakis Tipo A, se encontró más eficiente para el modelado de las vocales de la emoción "Tristeza".

Para el objetivo de la presente tesis el Sistema Multimodal presentó una tasa de reconocimiento de emociones de 97.00 % en usuarios mexicanos diferentes de aquellos que participaron en la creación de las bases de datos MX-Voz y MX-Expresiones. Los resultados globales presentados en la Tabla 7.1 dan soporte a la hipótesis planteada en la Sección 1.3:

"La integración de optimización evolutiva puede reducir la variabilidad de la tasa de reconocimiento, y mejorar el desempeño general, de un sistema de reconocimiento multimodal de emociones al evaluarse con bases de datos y pruebas en-vivo. Esta optimización puede determinar mejores estructuras para los reconocedores de voz y visión del sistema multimodal para obtener una precisión conjunta y consistente mayor al 95.00 %."

Con la integración de la optimización evolutiva se obtuvo un incremento en el reconocimiento de estas emociones. Para el sub-sistema de voz se alcanzó un promedio del 90.00 % mientras que para el sub-sistema de visión el desempeño se incrementó en un 10.00 % sobre los sistemas previos sin optimización. En pruebas en-vivo, integrando

ambos sub-sistemas con GAs, se obtuvo una tasa del 97.00 % sobre 400 frases evaluadas. Por lo tanto, las técnicas presentadas en este trabajo de tesis representan una contribución para el desarrollo de sistemas más eficientes de Interacción Humano-Robot.

Este trabajo se puede extender en los siguientes puntos para el mejoramiento de los sistemas de reconocimiento de emociones:

- Mejorar el reconocimiento de emociones en voz bajo el esquema IU. De igual manera considerar más alternativas para el tipo de las estructuras de HMMs para la optimización mediante GAs.

- Probar los diferentes sub-sistemas de visión con otras bases de datos como FEED-TUM y FACES collection. De igual manera realizar una comparativa entre PCA y Fisher para la extracción de características que discriminen clases de mejor manera en lugar de extraer características que mejor describan a una clase.

- Emplear técnicas de optimización alternativas a los GAs como Búsqueda Tabú.

- Considerar otras técnicas de codificación para la extracción de características espectrales para hacer más eficiente la detección de emociones en voz.

- Hacer uso de cámaras de profundidad para modelar el rostro y los rasgos faciales de cada emoción a través de puntos tridimensionales. El modelado de la dispersión de puntos podría implementarse con Gaussianas.

- Implementar un sistema multimodal que involucre la psicología de las emociones y tratar de verificar y corregir el estado de ánimo del usuario.

- Implementar el sistema robótico en centros para el cuidado de niños o gente adulta, además de poderse utilizar como instructor en terapias.

Bibliografía

[1] Alter, K., Rank, E., and Kotz, S.A. Accentuation and emotions - two different systems ? In *Proc. ISCA Workshop Speech and Emotion*, volume 1, pages 138–142, 2000.

[2] Anagnostopoulos, C.N., Iliou, T., and Giannoukos, I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artif Intell Rev*, 43:155–177, 2015.

[3] AT&T Technology Solutions. AT&T FSM Library. In *http://www.research.att.com/export/sites/att_labs/library/documents/licensing_data_sheets/fsmlibrary_factsheet_20090925.pdf*, Consultado el 19/05/2015.

[4] Austermann, A., Esau, N., Kleinjohann, L., and Kleinjohann, B. Fuzzy emotion recognition in natural speech dialogue. In *Proc. of the 2005 IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2005)*, pages 317–322, 2005.

[5] Bakare, G.A., Venayagagamoorthy, G.K., and Aliyu, U.O. Reactive Power and Voltage Control of the Nigerian Grid System Using Micro-Genetic Algorithm. In *Proc. of the Power Engineering Society General Meeting*, pages 1916–1922, 2005.

[6] Basheer, I.A. and Hajmeer, M. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, 43:3–31, 2000.

[7] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Archy, S., Russell, M., and Wong, M. "you stupid tin box" - children interacting with the AIBO robot: A cross-

linguistic emotional speech corpus. In *Proc. Language Resources and Evaluation (LREC '04)*, pages 171–174, 2004.

[8] Benítez-Saucedo, A. *Avatares Emocionales en la Mensajería Instantánea*. Tesis de Maestría en Medios Interactivos. Universidad Tecnológica de la Mixteca, 2014.

[9] Beskow, J. and Sjolander, K. *Wavesurfer*. KTH: The department of Speech, Music and Hearing, 2013.

[10] Breazeal, C. Emotion and sociable humanoid robots. *Int. J. Human-Computer Studies*, 59:119–155, 2003.

[11] Busso, C., Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Deng, Z., Lee, S., Neumann, U, and Narayanan, S. Analysis of Emotion Recogition Using Facial Expressions, Speech and Multimodal Information. In *Proc. Int. Conf. Multimodal Interfaces (ICMI 2004)*, pages 205–211, 2004.

[12] Caballero-Barbosa, A. *Extracción Automática del Contorno de Rostros y sus Características Faciales*. 2007.

[13] Caballero, S. Recognition of emotions in mexican spanish speech: An approach based on acoustic modelling of emotion-specific vowels. *The Scientific World Journal*, pages 1–13, 2013.

[14] Caballero, S.O. and Cox, S.J. Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers. *EURASIP J. Adv. Signal Processing*, 2009:1–14, 2009.

[15] Cardio Smart. Exercise: The Key to Good Health. In *https://www.cardiosmart.org/̃media/Documents/Fact %20Sheets/es-US/zu1880.ashx*, Consultado el 11/05/2015.

[16] Chaavan, V.M. and Gohokar, V.V. Speech emotion recognition by using SVM-classifier. *International Journal of Engineering and Advanced Technology*, 1(5):11–15, 2012.

[17] Chambers, S., Breazel, C., Atkins, A., Revis, M., Asher, J., Craft, A., Westelman, R., Kotelly, B., and Smith, L. Meet Jibo, The World's First Family Robot. In *http://www.jibo.com/*, Consultado el 19/02/2015.

[18] Chaturvedi, A. and Tripathi, A. Emotion Recognition using Fuzzy Rule-base System. *International Journal of Computer Applications*, 93(11):25–28, 2014.

[19] Chen, L., Mao, X., Xue, Y., and Cheng, L. Speech emotion recognition: Features and classification models. *Digital Signal Processing*, 22:1154–1160, 2012.

[20] Cowie, R., Douglas-Cowie, E., and Cox, C. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18(4):371–388, 2005.

[21] Davis, S.B. and Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, 28(4):357–366, 1980.

[22] Devillers, L., Tahon, M., Sehili, M., and Delaborde, A. Inference of Human Beings' Emotional States from Speech in Human - Robot Interactions. *International Journal of Social Robotics*, pages 1–13, 2015.

[23] Deya Cano. Estiramiento vital para sentirse bien. In *http://enforma.salud180.com/nutricion-y-ejercicio/estiramiento-vital-para-sentirte-bien*, Consultado el 11/05/2015.

[24] Ebner, N.C., Riediger, M., and Lindenberger, U. FACES-A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42:351–362, 2010.

[25] Filko, D. and Martinovic, G. Emotion Recognition System by a Neural Network Bases Facial Expression Analysis. *AUTOMATIKA*, 54(2):263–272, 2013.

[26] Firoz-Shah, A., Vimal-Krishnan, V.R., Raji-Sukumar, A., Jayakumar, A., and Babu-Anto, P. Speaker independent automatic emotion recognition from speech:

a comparison of MFCCs and discrete wavelet transforms. In *Proc. of the International Conference on Advances in Recent Technologies in Communication and Computing*, pages 528–531, 2009.

[27] Fujita, M. On activating human communications with pet-type robot AIBO. *Proceedings of the IEEE*, 92(11):1804–1813, 2004.

[28] GeoSalud. Ejercicio: ¿cómo empezar? In *http://www.geosalud.com/Enfermedades %20Cardiovasculares/ejercicio.htm*, Consultado el 11/05/2015.

[29] Gil, A., Benavides, M., Guilarte, Y., and Marquez, M. Sistema para el reconocimiento e identificación de rostros a través de fotografías. *Revista Ciencia e Ingeniería*, 29(2):131–136, 2008.

[30] Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Co., 1989.

[31] Gosavi, A.P. and Khot, S.R. Facial Expression Recognition Using Principal Component Analysis. *International Journal of Soft Computing and Engineering*, 3(4):258–262, 2013.

[32] Guerrero, A. Exponen en aguascalientes avances en inteligencia artificial. In *http://www.conacytprensa.mx/index.php/centros-conacyt/1824-exponen-avances-de-ia-en-aguascalientes*. CONACYT - Agencia Informativa, Consultado el 10/06/2015.

[33] Haq, S., Jackson, P.J.B., and Edge, J. Audio-visual Feauture Selection and Reduction for Emotion Classification. In *Proc. Auditory-Visual Speech Processing (AVSP 2008)*, pages 185–190, 2008.

[34] Hong, J.-W., Han, M.-J., Song, K.-T., and Chang, F.-Y. A Fast Learning Algorithm for Robotic Emotion Recognition. In *Proc. of the International Symposium on Computational Intelligence in Robotics and Automation (CIRA 2007)*, pages 25–30, 2007.

[35] Ilbeygi, M. and Hosseini, H. A novel fuzzy facial expression recognition system based on facial feature extraction from color face images. *Engineering Applications of Artificial Intelligence*, 25:130–146, 2012.

[36] Jang, K.-D. and Kwon, O.-W. Speech Emotion Recognition for Affective Human-Robot Interaction. In *Proc. of SPECOM 2006*, pages 419–422, 2006.

[37] Jurafsky, D. and Martin, J.H. *Speech and Language Processing*. Pearson: Prentice Hall, 2009.

[38] Kahn, P.H., Freier, N.G., Friedman, B., Severson, R.L., and Feldman, E.N. Social and Moral Relationships with Robotic Others? In *Proc. of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, pages 20–22, 2004.

[39] Kahn, P.H., Friedman, B., Pérez-Granados, D.R., and Freier, N.G. Robotic pets in the lives of preschool children. *Interaction Studies*, 7(3):405–436, 2006.

[40] karthigayan, M., Rizon, M., Nagarajan, R., and Sazali, Y. Genetic algorithm and neural network for face emotion recognition. In *Affective Computing*, pages 57–68. InTech, 2008.

[41] Kaur, M., Vashisht, R., and Neeru, N. Recognition of Facial Expressions with Principal Component Analysis and Singular Value Decomposition. *International Journal of Computer Applications*, 9(12):36–40, 2010.

[42] Kulic, D. and Croft, E.A. Affective state estimation for human-robot interaction. *IEEE Transactions on Robotics*, 23(5):991–1000, 2007.

[43] Langner, O., Dotsch, R.G., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., and van Knippen- berg, A. Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, 24(8):1377–1388, 2010.

[44] Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. Emotion recognition based on phoneme classes. In *Proc. Int. Conf. Spoken Language Processing (ICSLP '04)*, volume 1, pages 889–892, 2004.

[45] Li, A., Fang, Q., Hu, F., Zheng, L., Wang, H., and Dang, J. Acoustic and articulatory analysis on Mandarin Chinese vowels in emotional speech. In *Proc. of the Int. Symp. Chinese Spoken Language Processing ISCSLP*, pages 38–43, 2010.

[46] Lin, Y.L. and Wei, G. Speech emotion recognition based on HMM and SVM. In *Proc. Int. Conf. Machine Learning and Cybernetics*, volume 8, pages 4898–4901, 2005.

[47] Lisetti, C., Nasoz, F., LeRouge, C., Ozyer, O., and Alvarez, K. Developing multimodal intelligent affective interfaces for tele-home health care. *Int. J. Human-Computer Studies*, 59:245–255, 2003.

[48] López, J., Cearreta, I., Garay, N., López, K., and Beristain, A. Creación de una base de datos emocional bilingüe y multimodal. In Redondo, M.A., Bravo, C., and Ortega, M., editors, *Proc. of the 7th Spanish Human Computer Interaction Conference, Interaccion'06*, pages 55–66, 2006.

[49] Lucey, P., Cohn, J., Kanade, T., Saegih, J., Ambadar, Z., and Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proc. of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101, 2010.

[50] Luo, Y., Wu, C., and Zhang, Y. Facial Expression Recognition Based on Fusion Feature of PCA and LBP with SVM. *Optik - International Journal for Light and Electron Optics*, 124(17):2767–2770, 2013.

[51] Lyons, M. The Japanese Female Facial Expression (JAFFE) Database. In *http://www.kasrl.org/jaffe.html*, Consultado el 19/06/2014.

[52] Lyons, M., Budynek , J., and Akamatsu, S. Automatic Classification of Single Facial Images. *IEEE Trans. Patt. Anal. Mach. Intell.*, 21:1357–1362, 1999.

[53] Mao, X., Chen, L., and Fu, L. Multi-level speech emotion recognition based on HMM and ANN. In *Proc. of the World Congress on Computer Science and Information Engineering*, pages 225–229, 2009.

[54] Martín de Diego, I., Serrano, A., Conde, C., and Cabello, E. Técnicas de reconocimiento automático de emociones. *Revista Electrónica Teoría de la Educación: Educación y Cultura en la Sociedad de la Información*, 7(2):107–127, 2006.

[55] Mohri, M. Minimization algorithms for sequential transducers. *Theoretical Computer Science*, 234:177–201, 2000.

[56] Mohri, M., Pereira, F., and Riley, L. Weighted automata in text and speech processing. In *Proc. of the 12th European Conference on Artificial Intelligence*, pages 257–286, 1996.

[57] Mohri, M., Pereira, F., and Riley, L. Weighted finite state transducers in speech recognition. *Computer Speech and Language*, 16:69–88, 2002.

[58] Neiberg, D., Elenius, K., and Laskowski, K. Emotion recognition in spontaneous speech using GMMs. In *Proc. of INTERSPEECH*, pages 809–812, 2006.

[59] Odashima, T., Onishi, M., Riken, N., Hirano, S., Mukai, T., and Luo, Z. Development of the Tactile Sensor System of a Human-Interactive Robot "RI-MAN". *IEEE Transactions on Robotics*, 24(2):505–512, 2008.

[60] Owusu, E., Zhan, Y., and Mao, Q. R. A neural-AdaBoost based facial expression recognition system. *Expert Systems with Applications*, 41:3383–3390, 2014.

[61] PAL Robotics. Reem - humanoid robot. In *http://reemc.pal-robotics.com/en/*, Consultado el 19/06/2014.

[62] Pal, S.S. and Hasan, M. Facial Expression Recognition Using Fuzzy Logic. *International Journal of Science and Research*, 3(6):851–854, 2014.

[63] Pao, T.L., Liao, W.Y., Chen, Y.T., Yeh, J.H., Cheng, Y.M,, and Chien, C.S. Comparison of several classifiers for emotion recognition from noisy Mandarin speech. In *Proc. of the 3rd International Conference on International Information Hiding and Multimedia Signal Processing (IIHMSP 2007)*, pages 23–26, 2007.

[64] Pineda, L., Villaseñor, L., Cuétara, J., Castellanos, H., Galescu, L., Juárez, J., Llisterri, J., and Pérez, P. The corpus dimex100: Transcription and evaluation. *Language Resources and Evaluation*, 44:347–370, 2010.

[65] Pooja, R.N. and Kaur, S. Hybrid Technique for Human Face Emotion Detection. *International Journal of Advanced Computer Science and Applications*, 1(6):91–101, 2010.

[66] Prolific Technology. USB-TTL serial converter datasheet. In *http://www.prolific.com.tw/UserFiles/files/ds_pl2303HXD_v1_4_4.pdf*, Consultado el 19/05/2015.

[67] Rabiner, L. A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proc. IEEE*, volume 37, pages 257–286, 1989.

[68] Ramírez, S, Jurado, N., Sanchez, L., and Marco, M.R. Terapias alternativas en el manejo del dolor. In *http://www.enfermeriadeciudadreal.com/terapias-alternativas-en-el-manejo-del-dolor-73.htm*, Consultado el 11/05/2015.

[69] Rao, K.S., Saroj, V.K., Maity, S., and Koolagudi, S.G. Recognition of emotions from video using neural networks models. *Expert Systems with Applications*, 38(10):13181–13185, 2011.

[70] Rasoulzadeh, M. Facial Expression Recognition using Fuzzy Inference System. *International Journal of Engineering and Innovative Technology*, 1(4):1–5, 2012.

[71] Robotis. *Bioloid Premium, Quick Start: Assembly and Program Download Manual*. Robotis Co., Ltd., 2012.

[72] Robotis. *ZIG-100/110A e-Manual*. Robotis Co., Ltd., 2012.

[73] Robotis Co., Ltd. Zigbee SDK. In *http://support.robotis.com/en/software/zigbee_sdk.htm*, Consultado el 11/05/2015.

[74] Röfer, T., Laue, T., Burchardt, A., Damrose, E., Fritsche, M., Müller, J., and Rieskamp, A. B-Human: Team Description for RoboCup 2008. In L. Iocchi, H. Matsubara, A. Weitzenfeld, and C. Zhou, editors, *RoboCup 2008: Robot Soccer World Cup XII*, pages 1–6, 2008.

[75] Rojas, R. *Neural Networks - A Systematic Introduction*. Springer-Verlag, Berlin, 1996.

[76] Samani, H.A. and Saadatian, E. A Multidisciplinary Artificial Intelligence Model of an Affective Robot. *International Journal of Advanced Robotic Systems*, 9:1–11, 2012.

[77] Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *Proc. of INTERSPEECH*, pages 2253–2256, 2007.

[78] Schuller, B., Muller, R., Hornler, B., Konosu, H., and Rigoll, G. Audiovisual Recognition of Spontaneous Interest within Conversations. In *Proc. Int. Conf. Multimodal Interfaces (ICMI 2007)*, pages 30–37, 2007.

[79] Schuller, B., Rigoll, G., and Lang, M. Hidden Markov model-based speech emotion recognition. In *Proc. of the International Conference on Multimedia and Expo*, pages 401–404, 2003.

[80] Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Trans Affect Comput*, 1:119–131, 2010.

[81] Shih, F.Y., Chuang, C.-F., and Wang, P. S. P. Performance comparisons of facial expression recognition in JAFFE database. *International Journal of Pattern Recognition and Artificial Intelligence*, 22(3):445–459, 2008.

[82] Shin, Y.-G., Park, S.S., Kim, J.-N., Kim, J.-N., and Jang, D.-S. Development of a humanoid robot for emotion recognition. In *Proc. of the 5th WSEAS Int. Conf. on Computational Intelligence, Man-Machine Systems and Cybernetics*, pages 308–314, 2006.

[83] Slakovic, M. and Jevtic, D. Face recognition using eigenface approach. *Serbian Journal of Electrical Engineering*, 9(1):121–130, 2012.

[84] Song, M., Bu, J., Chen, C., and Li, N. Audio-Visual-Based Emotion Recognition: A New Approach. In *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR 2004)*, pages 1020–1025, 2004.

[85] Song, M., You, M., Li, N., and Chen, C. A Robust Multimodal Approach for Emotion Recognition. *Neurocomputing*, 71:1913–1920, 2008.

[86] Sport Life. Top 20 estiramientos mejora movilidad gana salud. In *http://www.sportlife.es/deportes/articulo/top-20-estiramientos-mejora-movilidad-gana-salud*, Consultado el 11/05/2015.

[87] Tayal, S. and Vijay, S. Human Emotion Recognition and Classification from Digital Colour Images Using Fuzzy and PCA Approach. *Advances in Computer Science*, pages 1033–1040, 2012.

[88] The Mathworks, Inc. Creating Graphical User Interfaces. In *http://www.mathworks.com/help/pdf_doc/matlab/buildgui.pdf*, Consultado el 11/05/2015.

[89] Thuseethan, S. and Kuhanesan, S. Eigenface based recognition of emotion variant faces. *Computer Engineering and Intelligent Systems*, 5(7):31–37, 2014.

[90] Turk, M. and Pentland, A. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[91] Viola, P. and Jones, M. J. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[92] Vlasenko, B., Schuller, B., Wendemut, A., and Rigoll, G. Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. In *Proc. of the 2nd International Conference on Affective Computing and Intelligent Interaction*, pages 139–147, 2007.

[93] Wan, Z. and Guan, L. Recognizing Human Emotion from Audiovisual Information. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2005)*, pages 1125–1128, 2005.

[94] Wu, C.H. and Liang, W.B. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Trans Affect Comput*, 2:10–21, 2011.

[95] Xu, L., Zhuge, Z., and Yang, J. Artificial emotion and its recognition, modeling and applications: an overview. In *Proc. of the Fifth World Congress on Intelligent Control and Automation (WCICA 2004)*, volume 3, pages 2380–2385, 2004.

[96] Yamamoto, S., Yoshitomi, Y., Tabuse, M., Kushida, K., and Asada, T. Recognition of a Baby's Emotional Cry Towards Robotics Baby Caregiver. *International Journal of Advanced Robotic Systems*, 10:1–7, 2013.

[97] Young, S. and Woodland, P. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

[98] Yu, F., Chang, E., Xu, Y. Q., and Shum, H.Y. Emotion detection from speech to enrich multimedia content. In *Proc. IEEE Pacific-Rim Conf. Multimedia 2001*, volume 1, pages 550–557, 2001.

[99] Yu, W. Research and implementation of emotional feature classification and recognition in speech signal. In *Proc. of the 2008 International Symposium on Intelligent Information Technology Application*, pages 471–474, 2008.

[100] Yun, S. and Yoo, C.D. Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model. In *Proc. of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pages 4169–4172, 2009.

[101] Zhang, L., Jiang, M., Farid, D., and Hossain, M.A. Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications*, 40:5160–5168, 2013.

[102] Zhang, S., Zhao, X., and Lei, B. Speech Emotion Recognition Using an Enhanced Kernel Isomap for Human-Robot Interaction. *International Journal of Advanced Robotic Systems*, 10:1–7, 2013.

[103] Zhou, Y., Zhang, J., Wang, L., and Yan, Y. Emotion recognition and conversion for Mandarin speech. In *Proc. of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 179–183, 2009.

Apéndice A

Reglas Fonéticas para el Transcriptor Automático

El transcriptor se desarrolló como una función en lenguaje de programación MATLAB. Esta función requiere dos argumentos:

- La palabra a transcribir (argumento textual) con su identificador emocional (por ejemplo, CASA_F, si la palabra se pronuncia con felicidad).
- El número “0” o “1” (argumento numérico) para especificar si se desea añadir el fonema de pausa-corta (/sp/) al final de la secuencia fonética de la palabra (CASA_F = /k/ /a_f/ /s/ /a_f/ /sp/).

Para obtener la transcripción de una palabra a partir de las letras del alfabeto que la forman se definió un conjunto de reglas gramaticales y acústicas. Esto fue importante porque la representación del sonido (fonema) de algunas letras depende del contexto en el cual se encuentra (esto es, las letras que se encuentran antes y después de la misma). En general, se definieron 50 reglas para la transcripción fonética de palabras las cuales se presentan en la Tabla A.1. La nomenclatura de los fonemas se tomó de la representación presentada en [64].

Tabla A.1: Reglas del Transcriptor Fonético.

Regla	Letra	Contexto (α) + Letra + (β)	Representación Fonética $z = \{e,f,n,t\}$	Tipo
1	A	(*) + A + (*)	/a_z /	vocal
2	E	(*) + E + (*)	/e_z /	vocal
3	I	(*) + I + (*)	/i_z /	vocal
4	O	(*) + O + (*)	/o_z /	vocal
5	U	(*)+U+(*)	/u_z /	vocal
		Casos Particulares:		
6		(G)+U+(I)	no se transcribe	
7		(Q)+U+(I)	no se transcribe	
8		(G)+U+(E)	no se transcribe	
9		(Q)+U+(E)	no se transcribe	
10	H	(*) + H + (*)	no se transcribe	
11	B	(*) + B + (*)	/b/	consonante
12	V	(*) + V + (*)	/b/	consonante
13	S	(*) + S + (*)	/s/	consonante
14	Z	(*) + Z + (*)	/s/	consonante
15	Q	(*) + Q + (*)	/k/	consonante
16	K	(*) + K + (*)	/k/	consonante
17	F	(*) + F + (*)	/f/	consonante
18	P	(*) + P + (*)	/p/	consonante
19	J	(*) + J + (*)	/x/	consonante
20	Ñ	(*) + Ñ + (*)	/nn/	consonante
21	T	(*) + T + (*)	/t/	consonante
		Casos Particulares:		
22		(*) + T + (B)	/_D/	consonante
	Y	Casos Particulares:		
23		Final de una Palabra	/i_z/	vocal
24		Inicio de una Palabra	/Z/	consonante
25		(consonante)+Y+(consonante)	/i_z/	vocal
26		(vocal)+Y+(vocal)	/Z/	consonante
27	L	(*) + L + (*)	/l/	consonante
		Casos Particulares:		
28		(*)+L+(L), (L)+L+(*)	/Z/	consonante
29	D	(*) + D + (*)	/_D/	consonante
		Casos Particulares:		
30		(*)+D+(vocal), (*)+D+(R)	/d/	consonante
31	G	(*)+G+(*)	/_G/	consonante
		Casos Particulares:		
32		(*)+G+(A), (*)+G+(U), (*)+G+(O), (*)+G+(R), (*)+G+(L)	/g/	consonante
33		(*)+G+(E), (*)+G+(I)	/x/	consonante
34	N	(*)+N+(*)	/n/	consonante
		Casos Particulares:		
35		Final de una Palabra	/_N/	consonante
	R	Casos Particulares:		
36		Final de una Palabra	/_R/	consonante
37		Inicio de una Palabra	/r/	consonante
38		(*)+R+(R), (R)+R+(*)	/r/	consonante
39		(*)+R+(vocal)	/r(/	consonante
	C	Casos Particulares:		
40		(*)+C+(A), (*)+C+(O), (*)+C+(U), (*)+C+(R), (*)+C+(L)	/k/	consonante
41		(*)+C+(E), (*)+C+(I)	/s/	consonante
42		(*)+C+(T)	/_G/	consonante
43		(*)+C+(H)	/tS/	consonante
44	M	(*)+M+(*)	/m/	consonante
		Casos Particulares:		
45		(*)+M+(P)	/_N/	consonante
	X	Casos Particulares:		
46		Inicio de una Palabra	/x/	consonante
47		(*)+X+(T), (*)+X+(P), (*)+X+(C)	/kS/	consonante
48		(vocal)+X+(vocal), Palabra Inicia con Vocal	/x/	consonante
49		(vocal)+X+(vocal), Palabra Inicia con Consonante	/kS/	consonante
50	W	(*)+W+(*)	/g/ + /u_z /	consonante + vocal

α = letra ubicada antes de la letra a transcribir, β = letra ubicada después de la letra a transcribir, * = cualquier letra (vocal o consonante).