Integración de optimización evolutiva para el reconocimiento de emociones en voz

Luis-Alberto Pérez-Gaspar, Santiago-Omar Caballero-Morales, Felipe Trujillo-Romero

Universidad Tecnológica de la Mixteca, Huajuapan de Leon, Oaxaca, México

luis_335450@hotmail.com,scaballero@mixteco.utm.mx, ftrujillo@mixteco.utm.mx

Resumen. En este artículo se presenta el desarrollo de un sistema de reconocimiento de emociones basado en la voz. Se consideraron las siguientes emociones básicas: Enojo, Felicidad, Neutro y Tristeza. Para este propósito una base de datos de voz emocional fue creada con ocho usuarios Mexicanos con 640 frases (8 usuarios × 4 emociones × 20 frases por emoción). Los Modelos Ocultos de Markov (Hidden Markov Models, HMMs) fueron usados para construir el sistema de reconocimiento. Basado en el concepto de modelado acústico de vocales específicas emotivas un total de 20 fonemas de vocales (5 vocales × 4 emociones) y 22 fonemas de consonantes fueron considerados para el entrenamiento de los HMMs. Un Algoritmo Genético (Genetic Algorithm, GA) fue integrado dentro del proceso de reconocimiento para encontrar la arquitectura más adecuada para el HMM para cada vocal específica emotiva. Una tasa de reconocimiento total aproximada del 90.00 % fue conseguida con el reconocedor de voz construido con los HMMs optimizados.

Palabras clave: Reconocimiento de Emociones por Voz, Algoritmos Genéticos, Modelos Ocultos de Markov.

1. Introducción

Avances tecnológicos recientes en el campo de la Robótica se han centrado en el desarrollo de sistemas capaces de identificar un estado emocional de forma precisa. Dentro de este contexto la computación emocional está siendo integrada en los Robots con el propósito de establecer una interacción más natural y fluida con los humanos. Este objetivo es particularmente importante para el desarrollo de tecnología de asistencia, diagnóstico psiquiátrico y detección de mentiras [12].

Investigaciones sobre el reconocimiento de emociones en la voz han sido reportadas en la literatura. Por ejemplo en [11] una comparativa de desempeño de las técnicas de discriminante lineal de Fisher, Máquinas de Soporte Vectorial (Support Vector Machine, SVM), Análisis de Componente Principal (Principal Component Analysis, PCA) y Redes Neuronales Artificiales (Artificial Neural Networks, ANN) fue presentada para el reconocimiento de emociones.

El discriminante lineal de Fisher y PCA fueron usados para la extracción de características mientras que SVM y ANNs fueron usados para la clasificación. Seis emociones (Enojo, Felicidad, Tristeza, Sorpresa, Miedo y Disgusto) fueron consideradas. La tasa de reconocimiento para el sistema Fisher+SVM fue del 50.16 % mientras que para el sistema PCA+ANN fue del 39.16 %. El trabajo reportó una confusión significativa entre Felicidad y Sorpresa, al igual que la necesidad de mejorar la tasa de reconocimiento para Miedo y Disgusto.

Otro sistema de reconocimiento basado en la voz fue implementado para el Robot MEXI (Machine with Emotionally eXtended Intelligence) [2]. Este sistema permitía un diálogo natural con los usuarios humanos a través de un sistema de síntesis de voz con acento emocional. Cinco emociones (Enojo, Miedo, Tristeza, Felicidad y Neutro) fueron consideradas y la clasificación fue llevada a cabo con un sistema basado en lógica difusa llamado PROSBER. Las tasas de reconocimiento obtenidas fueron aproximadamente del 84.00 % para un sistema dependiente de usuario y del 60.00 % para un sistema independiente de usuario.

En [16] un reconocimiento multimodal de emociones para Enojo, Felicidad, Sopresa, Miedo, Tristeza y Neutro fue desarrollado usando FAPS (Facial Animation Parameters) y la técnica de Lipschitz para características acústicas. Modelos Ocultos de Markov Triples (Tripled Hidden Markov Models, THMMs) fueron implementados para realizar la sincronización del audio con las secuencias de patrones visuales y su clasificación. Para el sistema de voz una tasa de reconocimiento de 81.44 % fue obtenida mientras que para el sistema visual la tasa fue de 87.40 %. Sin embargo para el sistema multimodal (voz+visión) la tasa de reconocimiento fue alrededor de 93.30 %.

Finalmente en [6] 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-Frequency Cepstral Coefficients, MFCCs), Histogramas de Periodicidad y Patrones de Fluctuación fueron usados para la extracción de características. Experimentos realizados con la base de datos de voz emocional danesa DES (Danish Emotion Speech) presentaron las siguientes tasas de reconocimiento: 64.77 % con función Kernel Lineal, 78.41 % con función Polinomial, 79.55 % con función RBF y 78.41 % con función Sigmoide. Una confusión significativa fue observada entre Felicidad y Enojo.

En este artículo se aborda el reconocimiento de emociones considerando el Español Mexicano. Para esto se desarrolló una base de datos de voz emocional con usuarios Mexicanos. Para la tarea de reconocimiento se utilizó la técnica de modelado acústico de vocales específicas emotivas con HMMs [5]. Mientras que en otros trabajos una estructura HMM estándar es considerada para el reconocimiento de emociones en la voz [5,8,17] en este trabajo un Algoritmo Genético (GA) fue diseñado para encontrar la estructura más adecuada para los HMMs de cada vocal específica emotiva. Los resultados obtenidos mostraron que las características acústicas asociadas a las vocales de cada emoción requieren estructuras específicas de HMMs lo cual puede mejorar su reconocimiento.

La estructura del presente trabajo es la siguiente: en la Sección 2 los detalles del sistema de reconocimiento basado en voz son presentados. Estos detalles

incluyen la creación del corpus de voz emocional y el transcriptor fonético asociado para el entrenamiento supervisado de los HMMs. Después en la Sección 3 se presenta el diseño del GA para encontrar la estructura más adecuada de los HMMs para el modelado acústico. Los resultados del sistema HMM con la optimización del GA son presentados y discutidos en la Sección 4. Finalmente en la Sección 5 se presentan las conclusiones y el trabajo a futuro.

2. Sistema de reconocimiento emocional por voz

Para el desarrollo de un sistema de reconocimiento de emociones es importante contar previamente con una base de datos apropiada para el entrenamiento (modelado) del mismo. Para el presente trabajo una base de datos de voz (corpus) emocional fue requerida. Aunque existen corpora de voz de este tipo para propósitos de investigación la mayoría de los mismos se encuentran en lenguajes extranjeros (por ejemplo, Inglés [3,9,15] y Alemán [1,15]). Estos recursos no pueden ser fácilmente adaptados para otros lenguajes porque hay diferencias fonéticas entre ellos.

Dada esta situación fue necesaria la creación de un corpus de voz emocional Mexicano. Las siguientes condiciones fueron consideradas para el desarrollo de este recurso [5,13]:

- estímulo textual de diferentes longitudes para cada emoción;
- significancia semántica de los estímulos textuales;
- deben haber suficientes ocurrencias de las vocales específicas emotivas y consonantes 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. Un total de cinco mujeres y tres hombres fueron considerados para el corpus de voz emocional.

2.1. Base de datos de voz

Previo a la grabación de las muestras de voz se diseñó el estímulo textual para cada emoción. Esto fue importante para tener muestras de voz con la entonación emocional apropiada. Debido a que se ha encontrado en la literatura que las propiedades espectrales de los sonidos de las vocales son un indicador confiable de las emociones en la voz [9,10] éstas 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 [5]. De esta forma es considerado que una vocal “a” expresada con Enojo es diferente de una “a” expresada con Tristeza o Felicidad. Esto permite el modelado acústico de vocales específicas emotivas [5].

Para este trabajo las siguientes emociones fueron consideradas: Enojo, Felicidad, Neutro y Tristeza [5,18,20]. 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. Se diseñaron 20 frases para cada emoción y algunos ejemplos son presentados en la Tabla 1.

Tabla 1. Muestra de Frases de Estímulo Diseñadas para cada Emoción.

Frases para Enojo
1 ¡Yo no te voy a estar soportando!
2 ¡Ya me tienes harto, ya deja de hablar!
3 …

Frases para Felicidad
1 ¡Me gané un viajo todo pagado a Florida!
2 ¡Me compré un billete de lotería y gané!
3 …

Frases para Neutro
1 El graznido de un pato no hace eco
2 La araña Sidney es la más venenosa y puede matar a un humano en 15 minutos
3 …

Frases para Tristeza
1 Mi mejor amigo acaba de fallecer ayer
2 Me haces mucha falta te extraño
3 …

Para asegurar el modelado acústico apropiado de las vocales un mínimo de seis ocurrencias fue considerado. En la Tabla 2 se presenta el número de muestras por vocales para cada grupo de frases emocionales. Nótese que el mínimo es de 19 muestras ("u" con Tristeza) lo cual es mayor que el número mínimo considerado de seis ocurrencias.

Tabla 2. 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

Las frases emocionales fueron grabadas en un salón a puerta cerrada con la herramienta Wavesurfer [4] en formato .WAV con una frecuencia de muestreo de 48000 Hz. La distancia entre el micrófono (micrófono interno de una compu-

tadora tipo laptop) y el usuario fue de alrededor de 60 cm. A cada voluntario se le pidió pronunciar cada una de las 20 frases por emoción llegando a un total de 80 muestras de voz por voluntario (80 frases $\times$ 8 usuarios = 640 frases).

2.2. Etiquetado fonético para el modelado acústico

Después de que las muestras de voz fueron grabadas, los archivos de audio fueron etiquetados a nivel palabra con Wavesurfer como se presenta en la Figura 1. Para identificar las palabras y (subsecuentemente) los fonemas de las vocales las cuales fueron pronunciadas con una emoción en particular un identificador fue añadido a la palabras y a las etiquetas fonéticas. Para cada emoción el identificador para las palabras fue *_E* para Enojo, *_F* para Felicidad, *_N* para Neutro y *_T* para Tristeza. Para las vocales (a nivel fonético) los identificadores fueron *_e*, *_f*, *_n* y *_t* respectivamente [5].

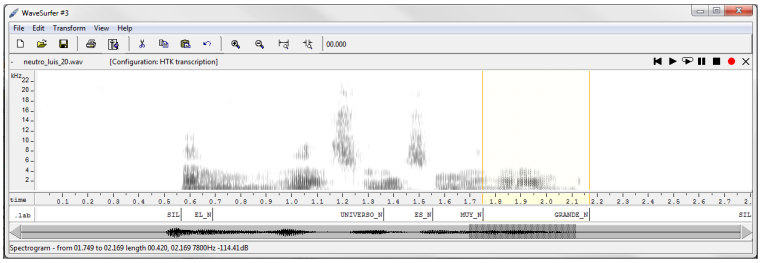


Fig. 1. Etiquetado a Nivel Palabra con Wavesurfer.

Una vez que el proceso de etiquetado a nivel palabra fue terminado se llevó a cabo el etiquetado a nivel fonético. En el Español Mexicano existen 27 fonemas (22 consonantes + 5 vocales) [7]. Debido a que un grupo de vocales fue considerado para cada emoción un total de 20 vocales (5 vocales $\times$ 4 emociones) fueron integradas dentro del repertorio fonético para el Español Mexicano lo cual llevó a un total de 42 fonemas (22 consonantes + 20 vocales).

Para obtener la secuencia de los fonemas para cada palabra un transcriptor fonético basado en TranscribEmex [14] fue desarrollado. El transcriptor fonético consideró aproximadamente 60 reglas gramaticales y acústicas para las diferentes combinaciones de vocales y consonantes dentro de una palabra. Algunas de las reglas se presentan a continuación:

- Si la consonante "q" (fonema /k/) aparece antes de la vocal "u" y la vocal "e‘" o "i" 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: "QUE" $\rightarrow$ /k/ /e_*n*/, "QUIEN" $\rightarrow$ /k/ /i_*n*/ /e_*n*/ /n/ si las palabras fueron pronunciadas con la emoción Neutro).

- 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 el fonema que representa su sonido es /_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: "DRAGON" → /d/ /r(/ /a_*e*/ /g/ /o_*e*/ /_N/, "DIGNIDAD" → /d/ /i_*t*/ /_G/ /n/ /i_*t*/ /d/ /a_*t*/ /_D/ si las palabras fueron expresadas con Enojo y Tristeza respectivamente).
- Si la consonante "g" aparece al final de una palabra su sonido es representado con /_G/. 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: GLOBO → /g/ /l/ /o_*e*/ /b/ /o_*e*/, GRITAR → /g/ /r(/ /i_*e*/ /t/ /a_*e*/ /_R/, GENIO → /x/ /e_*e*/ /n/ /i_*e*/ /o_*e*/, GITANA → /x/ /i_*e*/ /t/ /a_*e*/ /n/ /a_*e*/ si las palabras fueron pronunciadas con Enojo).

2.3. Modelo de lenguaje

El modelo de lenguaje es un elemento importante de cualquier sistema de reconocimiento de voz ya que contiene información estadística de las secuencias válidas de palabras en un lenguaje. El lenguaje Español contiene varias estructuras acerca de las secuencias de palabras que forman oraciones coherentes. Es importante mencionar que cualquier oración o frase puede ser pronunciada con cualquier emoción. De esta forma las estructuras gramaticales en un lenguaje aplican a cualquier emoción.

El modelado específico de emociones implica la independencia de vocales pero no de las palabras del vocabulario. A pesar de que las palabras habladas con una emoción dada tienen un identificador (_*E*, _*F*, _*N* o _*T*) estas palabras existen para todas las emociones. De esta forma el modelo de lenguaje fue integrado por el conjunto completo de 80 frases considerando que cada una de ellas puede ser expresada con todas las emociones. Esto llevó a un total de 80×4 emociones =320 frases para la estimación del modelo de lenguaje para el sistema de reconocimiento de voz. Esto también fue requerido para evitar un sesgo o influencia en el reconocimiento del estado emocional por parte del modelo de lenguaje.

Es importante mencionar que el reconocimiento de emociones es estimado contando el número de vocales dentro de las palabras reconocidas. El identificador (_*e*, _*f*, _*n*, _*t*) con el mayor número de vocales define la emoción dominante.

3. Optimización de HMMs con algoritmos genéticos

Todos los sonidos (fonemas) identificados en los archivos de audio con sus transcripciones fonéticas deben ser modelados para poder ser reconocidos. Entre

las técnicas usadas para modelado fonético los HMMs han sido ampliamente usados [19]. En la Figura 2(a) se presenta la estructura Bakis que es la más común para este propósito [19]. Sin embargo para el modelado acústico de vocales específicas emotivas otras estructuras pueden ser más adecuadas. La Figura 2(b) y la Figura 2(c) presentan estructuras HMM alternativas para el modelado acústico de fonemas. El problema de identificar la estructura HMM apropiada para cada vocal específica emotiva puede ser resuelto con un Algoritmo Genético (GA).

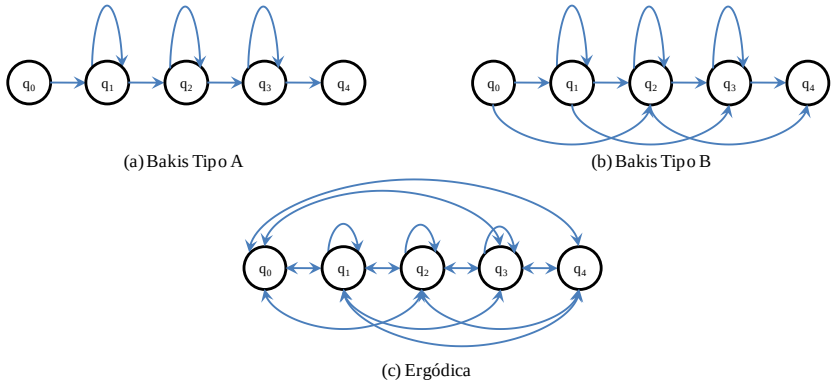


Fig. 2. Estructuras HMM para Modelado Acústico de los Fonemas de las Vocales.

Un GA es una técnica de optimización heurística que se basa en el proceso natural de sobrevivencia y adaptación de los individuos más aptos en una población. Estos individuos al sobrevivir son más probables a reproducirse, convirtiéndose en "padres" de nuevas generaciones de "hijos" que heredarán las características de 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 posibles soluciones a un problema combinatorio.

El diagrama general de operación y módulos principales del GA para el presente trabajo se presentan en la Figura 3. El cromosoma para la optimización de la estructura de los HMMs consistió de 20 (2-bit) genes (5 vocales $\times$ 4 emociones) en donde cada gen contiene el tipo de estructura de HMM para la vocal/emoción asociada. Solamente las estructuras de las vocales específicas emotivas fueron consideradas para optimización. Los modelos HMM para las consonantes tuvieron una 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.

Para encontrar las estructuras de HMMs más adecuadas cada conjunto de frases fue dividido en: (a) frases de entrenamiento y (b) frases para optimización (evaluación de aptitud). El conjunto de entrenamiento consistió de las últimas 8

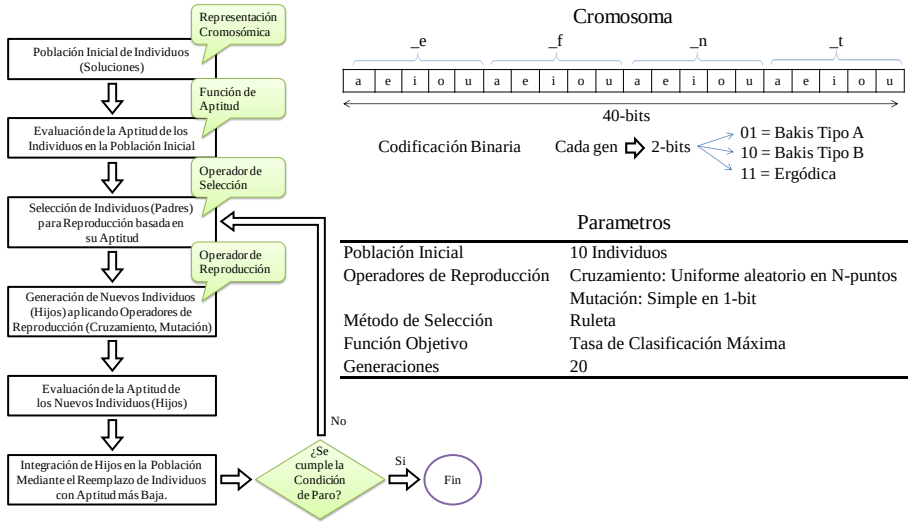


Fig. 3. Diagrama, Cromosoma y Parámetros de Configuración del Algoritmo Genético.

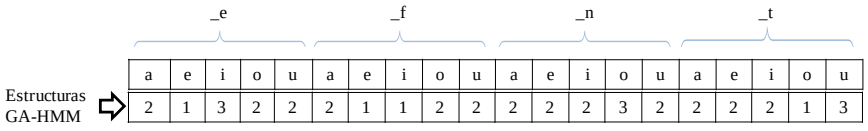
frases de cada conjunto emocional (frases 13 a 20) y las frases de optimización consistieron de las seis frases intermedias (frases 7 a 12). Las primeras seis frases (frases 1 a 6) fueron consideradas para la evaluación preliminar del GA.

En la Figura 4 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 de un reconocedor 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.

4. Resultados

Para la evaluación final del enfoque evolutivo con 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 × 4 emociones) de cada usuario fueron consideradas para entrenamiento de los HMM adicionalmente a las 560 frases (20 frases × 4 emociones × 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 × 4 emociones).



Desempeño Preliminar de Reconocimiento de Emociones

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

Fig. 4. GA-HMMs para las Vocales Específicas Emotivas y Desempeño Preliminar de Reconocimiento de 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. 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).

La herramienta HTK [19] fue utilizada para el desarrollo del sistema de reconocimiento de voz con HMMs y la implementación de la técnica de adaptación de usuario (Maximum Likelihood Linear Regression, MLLR) bajo el esquema independiente de usuario. Para la codificación de las muestras de voz se utilizó la técnica de vectores de características espectrales MFCC (Mel Frequency Cepstral Coefficients). Para esto se utilizaron 12 MFCCs más los coeficientes de energía, delta y de aceleración [19]. Los desempeños de las estructuras estándar (Bakis Tipo A solamente) y las estructuras GA-HMM (ver Figura 4) fueron evaluados bajo ambos esquemas de prueba y los resultados son presentados en la Tabla 3 y Tabla 4.

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 $p < 0.10$).

Tabla 3. Desempeño del Sistema de Reconocimiento de Emociones por Voz: HMMs Estándar.

Esquema de Prueba A						Esquema de Prueba B					
Usuario	Género	Enojo	Felicidad	Neutro	Tristeza	Usuario	Género	Enojo	Felicidad	Neutro	Tristeza
Lu	M	100.00	50.00	100.00	80.00	Lu	M	100.00	50.00	100.00	100.00
Ta	F	100.00	80.00	100.00	90.00	Ta	F	100.00	70.00	100.00	100.00
Au	F	80.00	85.00	80.00	100.00	Au	F	100.00	100.00	80.00	100.00
Mi	M	70.00	70.00	100.00	85.00	Mi	M	70.00	80.00	100.00	90.00
Me	F	75.00	70.00	90.00	90.00	Me	F	95.00	90.00	100.00	100.00
Je	M	100.00	30.00	75.00	50.00	Je	M	80.00	100.00	70.00	90.00
Li	F	70.00	40.00	20.00	75.00	Li	F	75.00	80.00	75.00	70.00
Ne	F	80.00	100.00	90.00	90.00	Ne	F	90.00	100.00	100.00	80.00
	Promedio	84.38	65.63	81.88	82.50		Promedio	88.75	83.75	90.63	91.25
				Promedio Total	78.59					Promedio Total	88.59

Tabla 4. Desempeño del Sistema de Reconocimiento de Emociones por Voz: GA-HMMs.

Esquema de Prueba A						Esquema de Prueba B					
Usuario	Género	Enojo	Felicidad	Neutro	Tristeza	Usuario	Género	Enojo	Felicidad	Neutro	Tristeza
Lu	M	100.00	60.00	100.00	90.00	Lu	M	100.00	60.00	100.00	100.00
Ta	F	100.00	90.00	100.00	90.00	Ta	F	100.00	90.00	100.00	90.00
Au	F	80.00	70.00	80.00	100.00	Au	F	100.00	100.00	80.00	100.00
Mi	M	100.00	65.00	100.00	90.00	Mi	M	70.00	60.00	90.00	90.00
Me	F	65.00	90.00	100.00	90.00	Me	F	95.00	100.00	90.00	100.00
Je	M	100.00	20.00	85.00	25.00	Je	M	90.00	100.00	90.00	80.00
Li	F	60.00	45.00	80.00	90.00	Li	F	90.00	60.00	90.00	70.00
Ne	F	80.00	100.00	100.00	90.00	Ne	F	100.00	100.00	100.00	80.00
	Promedio	85.63	67.50	93.13	83.13		Promedio	93.13	83.75	92.50	88.75
				Promedio Total	82.34					Promedio Total	89.53

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. 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. Desempeño del Sistema de Reconocimiento de Emociones por 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

5. Conclusiones

Las estructuras de HMMs estimadas con el GA estadísticamente mejoraron el desempeño del reconocimiento de emociones bajo el esquema de prueba dependiente de usuario (Esquema de Prueba A) de 78.59 % a 82.34 %. A pesar de que el desempeño de reconocimiento total fue mayor bajo el esquema de prueba independiente de usuario (Esquema de Prueba B) y se obtuvo un incremento adicional con el GA (88.59 % a 89.53 %) éste no fue estadísticamente significativo. Un incremento adicional marginalmente significativo (90.16 %) fue obtenido en el Esquema de Prueba B al considerar las estructuras HMM estándar para la emoción de Tristeza.

El trabajo a futuro se enfocará en mejorar el reconocimiento bajo el esquema de prueba independiente de usuario e incrementar el tamaño de la base de datos emocional. De igual manera mejorar el desempeño del GA para obtener incrementos más significativos (p.e., con $p < 0.05$) y contar con más alternativas para el tipo de las estructuras HMM para optimización. También es importante considerar la integración de otras técnicas de codificación para la extracción de características espectrales para hacer más eficiente la detección de la emoción. Finalmente el alcance del enfoque presentado en este trabajo debe evaluarse con

otras bases de datos de voz emocional y hacer una comparativa extensa con otros enfoques presentados en la literatura.

Referencias

1. Alter, K., Rank, E., Kotz, S.A.: Accentuation and emotions - two different systems ? In: Proc. ISCA Workshop Speech and Emotion. vol. 1, pp. 138–142 (2000)
2. Austermann, A., Esau, N., Kleinjohann, L., Kleinjohann, B.: Fuzzy emotion recognition in natural speech dialogue. In: Proc. of the 14th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2005) (2005)
3. Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Archy, S., Russell, M., 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) (2004)
4. Beskow, J., Sjolander, K.: WaveSurfer. KTH: The Department of Speech, Music and Hearing (2013)
5. Caballero, S.: Recognition of emotions in mexican spanish speech: An approach based on acoustic modelling of emotion-specific vowels. The Scientific World Journal pp. 1–13 (2013)
6. Chavan, V.M., Gohokar, V.V.: Speech emotion recognition by using SVM-classifier. International Journal of Engineering and Advanced Technology (IJEAT) 1(5), 11–15 (2012)
7. Cuétara, J.: Fonética de la Ciudad de México: Aportaciones desde las Tecnologías del Habla. Tesis de Maestría, Universidad Nacional Autónoma de México (UNAM), México. (2004)
8. Fernandez, R., Picard, R.: Modelling drivers' speech under stress. Speech Communication 40, 145–159 (2003)
9. Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S.: Emotion recognition based on phoneme classes. In: Proc. Int. Conf. Spoken Language Processing (ICSLP '04). vol. 1, pp. 889–892 (2004)
10. Li, A., Fang, Q., Hu, F., Zheng, L., Wang, H., Dang, J.: Acoustic and articulatory analysis on Mandarin Chinese Vowels in emotional speech. In: Proc. 7th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2010. pp. 38–43 (2010)
11. Lijiang, C., Mao, X., Xue, Y., Cheng, L.: Speech emotion recognition: Features and classification models. Digital Signal Processing 22, 1154–1160 (2012)
12. Lin, Y.-L., Wei, G.: Speech emotion recognition based on HMM and SVM. In: Proc. of the 2005 International Conference on Machine Learning and Cybernetics. vol. 8, pp. 4898–4901 (2005)
13. López, J.M., Cearreta, I., Garay, N., López de Ipiña, K., Beristain, A.: Creación de una base de datos emocional bilingüe y multimodal. In: Proc. of the 7th Spanish Human Computer Interaction Conference, Interaccion 2006. vol. 6, pp. 55–66 (2006)
14. Pineda, L., Villaseñor, L., Cuétara, J., Castellanos, H., Galescu, L., Juárez, J., Llisterri, J., Pérez, P.: The corpus DIMEX100: Transcription and evaluation. Language Resources and Evaluation 44, 347–370 (2010)
15. Schuller, B., Rigoll, G., Lang, M.: Hidden Markov model-based speech emotion recognition. In: Proc. of the International Conference on Multimedia and Expo. pp. 401–404 (2003)

16. Song, M., You, M., Li, N., Chen, C.: A robust multimodal approach for emotion recognition. Neurocomputing 71, 1913–1920 (2008)
17. Wagner, J., Vogt, T., André, E.: A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech. Affective Computing and Intelligent Interaction, Series "Lecture Notes in Computer Science" 4738, 114–125 (2007)
18. Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S.: An acoustic study of emotions expressed in speech. In: Proc. Int. Conf. Spoken Language Processing (ICSLP '04). vol. 1, pp. 2193–2196 (2004)
19. Young, S., Woodland, P.: The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department, UK. (2006)
20. Yu, F., Chang, E., Xu, Y. Q., Shum, H.Y.: Emotion detection from speech to enrich multimedia content. In: Proc. IEEE Pacific-Rim Conf. Multimedia 2001. vol. 1, pp. 550–557 (2001)