APLICACIÓN DE LAS MÁQUINAS DE SOPORTE VECTORIAL AL RECONOCIMIENTO DE HABLANTES

UNIVERSIDAD AUTÓNOMA METROPOLITANA
MAESTRÍA EN CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN
Juan Gabriel Pedroza Bernal
pedrozafm@yahoo.com.mx

Asesores de Investigación:
Dr. Alfonso Prieto Guerrero
Dr. John Henry Goddard Close

22 de Junio 2007

Índice general

0.1. Introducción General

Actualmente es difícil concebir un sistema de cualquier índole, sin considerar aspectos relacionados con la seguridad del mismo. El término de seguridad, a pesar de tener diferentes acepciones, guarda la idea de preservar las condiciones de un estado considerado normal o aceptable. Así, la relación entre el reconocimiento de personas y la seguridad tiene un margen amplio, pues puede encontrarse tanto en aplicaciones forenses o legales como en controles de acceso a algún sistema; por ejemplo para determinar la identidad de una persona, su participación en un hecho, la validez de una petición de acceso, etc., en cualquier caso resulta evidente la necesidad de determinar la certeza o falsedad en la identidad de una persona.

El reconocimiento requiere de una base de registros conformada por algún patrón de referencia, que sea único para cada persona. Con esta base de registros se puede, conjuntamente con algún sistema automático y/o mediante alguna metodología, determinar si dada una persona cualquiera, ésta cuenta con registro o no y en su caso exhibir su identidad. Dos problemas fundamentales que enfrenta la realización de un sistema de reconocimiento son: establecer y definir la(s) característica(s) que serán consideradas como únicas para cada persona y determinar el(los) procedimiento(s) o metodología(s) mediante los cuales se realizará el reconocimiento. Al respecto, la biometría ha proporcionado el sustento científico para realizar sistemas de reconocimiento automático, basándose en la medición de propiedades fisiológicas del cuerpo humano como son las huellas digitales, el iris del ojo y también la voz.

En particular el reconocimiento mediante voz, se diferencia de los anteriores en que la voz no tiene, a priori, alguna característica que pueda ser considerada como única y en que la representación matemática de señales de voz es considerada como un proceso aleatorio, lo cual modifica drásticamente la naturaleza del problema de reconocimiento, pues las huellas digitales o el iris del ojo, son considerados en general como patrones gráficos fijos, al igual que otros como la forma de la cara o la de las manos. La necesidad de considerar un tratamiento de la señal de voz como un proceso aleatorio, estriba principalmente en los siguientes hechos:

- Físicamente, la voz es una onda mecánica de sonido que se propaga en el medio y es por ello continua, sin embargo su representación a través de cualquier sistema computacional es en forma discreta. La técnica que permite esta transformación es llamada muestreo y los datos obtenidos del mismo, dada una señal, no son predecibles a priori.

- La señal de voz guarda un alto grado de variabilidad en sus características para una misma persona, pues ésta es influenciada por factores ambientales y emocionales. Adicionalmente, los sistemas de grabado de voz provocan, de forma intrínseca, una distorsión en la señal recibida.

Debido a la trascendencia de este tipo de sistemas de reconocimiento y al interés científico que generan, en diferentes países existen esfuerzos de investigación donde se aplican técnicas y metodologías para proporcionar sistemas cada vez más seguros y confiables y con mejores parámetros de desempeño. Sin embargo la mayoría de estos trabajos han sido desarrollados en idiomas diferentes al español de México, por lo que este trabajo presenta aspectos específicos del desarrollo de sistemas de reconocimiento en nuestro idioma. Adicionalmente se plantea la aplicación de las Máquinas de Soporte Vectorial, como una técnica de clasificación recientemente utilizada y de la cual se tienen expectativas favorables. También se describe la conformación de una base de registros de voz en español bajo la cual se realiza la implementación de un sistema de verificación. Esta base de registros de voz sienta un precedente que puede ser aprovechado en futuras investigaciones en esta materia. Parte esencial del trabajo es la descripción del procesamiento al que fue sometida la voz, hasta obtener información representativa de la misma.

De esta forma el objetivo general de la aplicación que se presenta, consiste en la implementación y evaluación del desempeño de un sistema de verificación de hablantes, basado en Máquinas de Soporte Vectorial como sistema de clasificación. Esta implementación debe tratar, principalmente, con la variabilidad de la voz, con la influencia de factores intrínsecos (como la del módulo de grabación) y con el ruido presente en el ambiente. De igual forma debe considerar aspectos relacionados con la representación y representatividad de la voz, así como su procesamiento hasta la conformación de un modelo. Un objetivo adicional es la evaluación del sistema, para ello se requiere comparar su desempeño con el obtenido mediante otro sistema de clasificación. La comparación del desempeño requiere analizar el error obtenido en los sistemas, lo cual permite concluir sobre su comportamiento.

Con base en los objetivos planteados, este documento se estructura de la siguiente forma:

En el primer capítulo se presenta un esquema general sobre los sistemas de verificación por voz. También un estado del arte sobre los sistemas de reconocimiento por voz, el cual hace énfasis en diferentes desarrollos científicos y comerciales en esta área, así como sus características principales y destaca sus parámetros de desempeño. En el capítulo 2 se describe la conformación de una base de registros de voz que considera diferentes aspectos de la variabilidad de la voz y permite la operación de un sistema de reconocimiento así como la valoración del desempeño del mismo. Posteriormente se expone el procesamiento al que son sometidos los registros de voz, hasta la obtención de los llamados vectores de características, los cuales sintetizan el contenido frecuencial de la voz para cada hablante. Se finaliza este capítulo con una amplia descripción de la teoría en la que se basan las Máquinas de Soporte Vectorial como una técnica que permite la clasificación o separación de conjuntos de vectores. En el capítulo 3 se describe el procedimiento seguido para realizar el entrenamiento del sistema de clasificación así como las pruebas que lo evalúan.

Adicionalmente se presenta el comportamiento del sistema con voz en tiempo real. Con el objetivo de comparar los resultados obtenidos mediante el uso de Máquinas de Soporte Vectorial como sistema de clasificación, en el capítulo 4 se presentan dos metodologías de clasificación adicionales; Modelos de Mezclas Gaussianas y Redes Neuronales. Ambas son implementadas sobre los datos de los registros procesados. Por último, en el capítulo 5, se dan las conclusiones pertinentes y las actividades que pueden derivar de forma inmediata del presente trabajo.

Capítulo 1

Verificación de Hablantes

1.1. Introducción

El reconocimiento de hablantes es un área que depende en gran parte del procesamiento de señales y que se ocupa de la detección, identificación y verificación automática de hablantes, dichos conceptos son descritos a continuación:

Detección: Consiste en determinar a partir de la voz de un individuo determinado, si éste participó o no en una secuencia o trama de voz grabada. Un sistema de detección establece la participación de una persona en una grabación dada.

Identificación: Consiste en determinar la identidad de un individuo a partir de su voz; el individuo proporciona su voz al sistema el cual determina el nombre del hablante o si éste es desconocido.Un sistema de identificación determina la identidad de una persona a partir de su voz o en su caso indica que el hablante es desconocido.

Verificación: Consiste en determinar si un individuo tiene permiso de acceso o no a partir de su voz, de acuerdo a una base de patrones de voz establecida previamente. Un sistema de verificación determina si la voz proporcionada es de quien dice ser.

Existe trabajo considerable realizado en diferentes organizaciones e instituciones en reconocimiento de hablantes; universidades, industria y laboratorios de diferentes partes del mundo, sin embargo la mayor parte de la investigación se ha encaminado a la verificación de personas a través de línea telefónica [29]. Por la naturaleza de este proyecto resulta de interés describir los antecedentes que se tienen en materia de identificación y verificación, únicamente.

1.2. Esquema General de Verificación de Hablantes

En general los sistemas de verificación actuales pueden describirse mediante cuatro módulos: adquisición, procesamiento, verificación y decisión.

Adquisición: En esta fase se consideran aspectos sobre el grabado de la voz. Actualmente se utilizan convertidores analógicos digitales que muestrean la voz en un intervalo de 8000 Hz a 20,000 Hz y con una resolución de 12 a 32 bits [17]. La frecuencia de muestreo es asignada de acuerdo a la calidad de voz a manejar. Para calidad telefónica se emplean 8000 Hz. En la conformación de las bases de registros se consideran aspectos, principalmente, sobre la variabilidad de la voz y sobre otros, específicos del sistema de verificación.

Procesamiento: En esta fase se considera la extracción de características de las tramas de voz de los usuarios, que permitirán la operación del sistema de verificación. La extracción de vectores de características es realizada por medio del denominado análisis Cepstral. Una fase previa a dicho análisis puede consistir en la obtención de los coeficientes de predicción lineal (LPC, por sus siglas en inglés) [34], [17], ó bien la aplicación de un banco de filtros cubriendo el espectro en frecuencias de la señal [17]. Este análisis es también conocido como Mel-Warped Cepstrum, pues el banco de filtros es aplicado en una escala de frecuencias denominada Mel, la cual intenta emular el comportamiento fisiológico del oído humano [34].

Verificación: En esta fase se considera el proceso de generación del modelo correspondiente a cada hablante y la metodología para realizar la evaluación de los datos característicos a fin de determinar la mejor correspondencia con los modelos. Con este fin son empleados distintos procedimientos matemáticos. Algunos de los más referidos en los artículos de investigación son los Modelos de Mezclas Gaussianas (GMM, por sus siglas en inglés) [17] y los Modelos Ocultos de Markov (HMM, por sus siglas en inglés) [34]. También son reportados como procedimientos matemáticos las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) [45], [54] y las redes neuronales artificiales [4].

Decisión: En esta fase se considera la aplicación de un modelo matemático que, con base en los resultados obtenidos, minimice la posibilidad de que suceda alguno de los dos errores conocidos: la aceptación de un solicitante inválido y el rechazo de un solicitante válido. La tabla 1.1 define, de manera lógica, ambos tipos de error; bajo el supuesto de que **H={Muestra de voz de un usuario válido}**, $\neg H$ es la voz de un usuario no enrolado en el sistema y por tanto no válido. El sistema se limita a aceptar o rechazar la petición, entonces se tienen cuatro posibilidades de las cuales dos de ellas, E_1 y E_2, representan los dos tipos de error mencionados.

.	Acción del Sistema	
Propuesta	**Acepta**	**Rechaza**
H	**B**	E_1
$\neg H$	E_2	**B**

Tabla 1.1: Definición lógica de los errores posibles en un sistema de verificación.

El modelo para tratar con estos errores generalmente está basado en probabilidades y su objetivo es determinar un umbral de decisión. Esta fase está estrechamente ligada con la evaluación del desempeño del sistema. Actualmente se ha generalizado el uso de las curvas o gráficas DET (*Detection Error Tradeoff*), las cuales proporcionan un medio para comparar el desempeño de este tipo de sistemas.

La figura 1.1 muestra un esquema de un sistema de verificación general. En él pueden identificarse los cuatro módulos descritos previamente: adquisición, procesamiento, verificación y decisión.

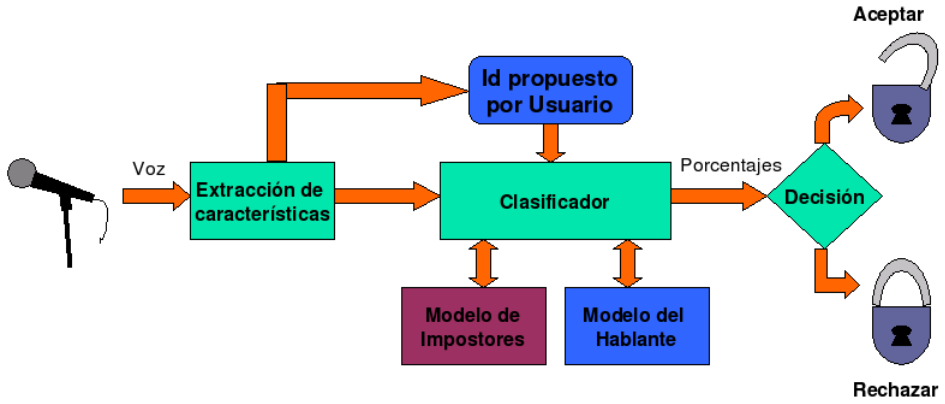

Figura 1.1: Esquema general de un sistema de verificación de hablantes.

1.3. Estado del Arte

En la tabla 1.2 se muestra el avance cronológico que se ha dado en la investigación e implementación de sistemas de reconocimiento automático de hablantes [34]. En este tabla se muestra que la verificación y la identificación pueden implementarse con o sin dependencia de frases específicas (*"text-dependent"* y *"text-independent"*, respectivamente), a su vez, dichas frases pueden tipificarse como provenientes de usuarios dispuestos o voluntarios o de

usuarios indispuestos o involuntarios. En el error reportado se identifica si el valor corresponde a pruebas de verificación (v), o a pruebas de identificación (i), así como la duración de las frases utilizadas para el reconocimiento.

A pesar de que la información presentada en la tabla 1.2 no es suficiente para identificar características particulares sobre cada una de las implementaciones, como por ejemplo la(s) metodología(s) de prueba realizadas tanto en la verificación como en la identificación, es posible establecer una tendencia, la cual consiste en la mejoría en la precisión de los sistemas, a coste de un incremento en el tiempo de las tramas de voz usadas, ésto conjuntamente con un incremento en las dimensiones de las bases de voz empleadas, aspecto que puede considerarse natural, dado el desarrollo que a la par se ha suscitado en las tecnologías de la información. Aún con esta mejoría en el desempeño, la existencia de un porcentaje de error, obliga al uso de sistemas auxiliares adjuntos en aplicaciones donde se requiere o desea una mayor confiabilidad. Dichos sistemas auxiliares pueden consistir, por ejemplo, en tarjetas con información estadística particular que defina a cada usuario del sistema de reconocimiento [34]. En la práctica, se sabe que existe siempre la posibilidad de error en cualquier tarea de reconocimiento, sin embargo se desea minimizarla.

Aunque el primer registro reportado en la tabla 1.2 corresponde a 1974, se sabe que los primeros esfuerzos por realizar sistemas automáticos de reconocimiento fueron hechos a principios de los años 1950s. A continuación se da una breve historia sobre el reconocimiento por voz desde las primeras décadas.

En 1952 en los Laboratorios Bell, los científicos Davis, Biddulph y Balashek [37] construyeron una máquina para separar dígitos de secuencias de voz proporcionadas por un solo hablante. Este sistema consistía en medir las frecuencias de resonancia en la región vocal de cada dígito pronunciado para así discriminarlos en las frases pronunciadas. Posterior a ello y de forma independiente, los científicos Olson y Belar [26] en los Laboratorios RCA intentaron reconocer 10 sílabas distintas de un único hablante, en 10 palabras monosilábicas pronunciadas por el mismo. Este sistema, desarrollado en 1956, también se basó en la obtención de las características espectrales de la voz en regiones vocales y no vocales. En 1959, en el University College del Reino Unido, los científicos Fry y Denes [15] trataron de desarrollar un sistema de reconocimiento de fonemas para cuatro sonidos vocales y nueve consonantes. Para ello usaron un analizador de espectros y un verificador de patrones como sistema de decisión. Lo novedoso de este sistema fue el uso de información estadística de cada uno de los fonemas involucrados a fin de reconocerlos en las palabras pronunciadas por los hablantes, las cuales consistían de más de un fonema.

En este mismo periodo, en los Laboratorios Lincoln del MIT, fue desarrollado un sistema de reconocimiento de vocales. Este sistema fue hecho por Forgie y Forgie en 1959

AUTOR(ES)	ORG.	PROCESAMIENTO	METODOLOGÍA	CALIDAD	TEXTO	PERSONAS	ERROR
Atal, 1974. [6]	AT&T	Cepstrum	Reconocimiento de Patrones	Laboratorio	Dependiente	10	i:2 %, 0.5s v:2 %, 1.0s
Markel, Davis, 1974. [32]	STI	LP	Estadísticas de largo plazo	Laboratorio	Independiente	17	i:2 %, 39.0s
Furui, 1981. [44]	AT&T	Cepstrum Normalizado	Reconocimiento de Patrones	Teléfono	Dependiente	10	v:0.2 %, 3.0s
Schwartz, et. al., 1982. [43]	BBN	LAR	PDF no paramétrico	Teléfono	Independiente	21	i:2.5 %, 2.0s
Li and Wrench, 1983. [38]	ITT	LP y Cepstrum	Reconocimiento de Patrones	Laboratorio	Independiente	11	i:21 %. 3.0s i:4 %, 10.0s
Doddington, 1985. [21]	TI	Filter Bank	Discrete Time Warpping	Laboratorio	Dependiente	200	i:0.8 %, 6.0s
Soong, et. al., 1985. [20]	AT&T	LP	VQ, Likelihood Ratio Distortion	Laboratorio	10 Dígitos Individuales	100	i:5.0 %, 1.5s i:1.5 %, 3.5s
Higgins, Wohlford, 1986. [3]	ITT	Cepstrum	DTW Likelihood Scoring	Laboratorio	Independiente	11	v:10.0 %, 2.0s v:4.5 %, 10s
Attilli, et. al., 1988. [27]	RPI	Cepstrum, LP, Autocorr.	Estadísticas a Largo Plazo	Laboratorio	Dependiente	90	v:1.0 %, 3.0s
Higgins, et. al., 1991. [1]	ITT	LAR, LP, Cepstrum	DTW Likelihood Scoring	Oficina	Dependiente	186	v:1.7 %, 10.0s
Tishby, 1991. [41]	AT&T	LP	HMM (AR mix)	Teléfono	10 Dígitos individuales	100	v:2.8 %, 1.5s v:0.8 %, 3.5s
Reynolds, 1995. [12],Reynolds, Carlson, 1995. [13]	MIT-LL	Mel-Cepstrum	HMM-GMM	Oficina	Dependiente	138	i:0.8 %, 10.0s v:0.12 %, 10s
Che, Lin, 1995. [7]	Rutgers	Cepstrum	HMM	Oficina	Dependiente	138	i:0.56 %, 2.5s i:0.14 %, 10s v:0.62 %, 2.5s
Colombi, et. al., 1996. [28]	AFIT	Cep, Eng dCep, ddCep	HMM monófono	Oficina	Dependiente	138	i:0.22 %, 10.0s v:0.28 %, 10s
Reynolds, 1996. [11]	MIT-LL	Mel-Cepstrum , MeldCepstrum	HMM, GMM	Teléfono	Independiente	416	v:16.0 %, 3s v:8.0 %, 16s v:5 %, 30s
J.P. Campbell, 1997. [34]	JHU	LPCC, LAR, LSPF	DTW	Teléfono	Dependiente	YOHO, 138	v:0.5 % EER
X. Dong, et. al., 2001 [56]	DFEEC-UEN	LPC	GMM, SVM	Teléfono	Independiente	YOHO, 138	v:0.5 % EER
H. Flengei, et. al., 2002 [23]	DIZ-UZ	MFCC, Speaker Clustering	SVM	Teléfono	Independiente	OGI, 60	v:3.1 % EER
W.M. Campbell, et. al., 2004 [55]	MIT-LL	Phone and word sequences	SVM	Teléfono	Independiente	NIST-2003,739	v:2.5 % EER
V. Wan, et. al., 2005 [51]	DCS-US	Sequence Discrimination, PLPCC	SVM, Score Space Kernels	Teléfono	Independiente	PolyVar, 38	v:4.0 % EER
S. Raghavan, et. al. 2006 [45]	MSU	MFCCs	SVM-RBF	———	Independiente	NIST-2003	Min DCF 0.1406
W.M. Campbell, et. al., 2006 [54]	MIT-LL	MFCC	GMM supervectors, SVM	Teléfono	Independiente	NIST-2005	v:3.77 % EER
K. Daoudi, et. al., 2007. [36]	IRIT-CNRS	LFCC	SVM-PoS	———	Independiente	NIST-SRE 2003, 2004	min DCF 0.0469

Tabla 1.2: Cronología del desarrollo de aplicaciones en reconocimiento de personas por voz.

[35] y consistió en el reconocimiento de 10 vocales incrustadas en frases con el formato \b\-vocal-\t\, de forma independiente al hablante considerado. Nuevamente fue utilizado un analizador basado en un banco espectral de filtros y en una estimación, variable en el tiempo, para reconocer la vocal pronunciada. Posteriormente, en los años sesentas, fueron introducidas diferentes ideas inicialmente por laboratorios de origen japones; una de las primeras fue presentada por Suzuky y Nakata del Radio Research Lab en Tokyo [31], la cual consistió en un sistema de reconocimiento de vocales implementado en hardware; éste contaba con un banco de filtros como analizador de espectro y la salida de cada filtro era conectada (y multiplicada por un peso asociado) a un circuito lógico complejo que funcionaba como sistema de decisión de vocales. Un segundo ejemplo es el sistema desarrollado en la Universidad de Kyoto por Sakai y Doshita en 1962 [47]. La implementación fue en hardware y consistió de un segmentador de voz junto con un analizador de cruces por cero (*zero-crossing*), el cual detectaba las secuencias de la trama correspondientes a voz.

En esta misma década (1960s) se iniciaron tres proyectos que tuvieron efectos importantes en el desarrollo de sistemas de reconocimiento [10]:

- El primer proyecto fue iniciado por Martin conjuntamente con los Laboratorios RCA[48], para proporcionar soluciones a problemas relacionados con la no uniformidad de las escalas temporales, en secuencias de voz. En este proyecto se desarrollaron diferentes métodos de normalización en la escala tiempo, basados en la detección de inicio y fin de segmentos de voz, de forma que fue posible reducir la variabilidad en los porcentajes de reconocimiento para segmentos de voz fonéticamente equivalentes. Martin, por último, desarrolló un método adecuado y fundó una de las primeras compañías (Threshold Technology) que desarrollaron productos comercialmente aptos basados en reconocimiento por voz.

- Casi a la par del proyecto anterior, en la Unión Soviética, Vintsyuk [49] propuso una metodología basada en programación dinámica para alinear en el tiempo dos pronunciaciones del mismo hablante. Este trabajo manejó conceptos sobre la metodología que hoy se conoce como *Dynamic Time Warping* (DTW), así como algoritmos rudimentarios para reconocimiento de palabras. Su trabajo permaneció desconocido hasta principios de los noventas, cuando fue estudiado y empleado por otros investigadores.

- El último proyecto considerado en esta década fue iniciado por Reddy en 1966 [16], mismo que fue pionero en el campo que hoy se conoce como *Reconocimiento Continuo de Voz (Continuous Speech Recognition)*, el cual se basa en el reconocimiento dinámico de fonemas. Este proyecto dio inicio a uno posterior en la Universidad Carnegie Mellon. Las primeras demostraciones de este último se dieron en dicha universidad en 1973; el sistema llamado *Hersay I*, era capaz de recibir información hablada y hacer una interpretación semántica para determinar la instrucción correspondiente, seleccionándola

de los movimientos posibles de las piezas en el juego de ajedrez. La conclusión fundamental fue que es posible emplear información sintáctica, semántica y contextual de la voz, para reducir el número de opciones posibles a ser consideradas en un sistema que entienda, aunque en forma limitada, el lenguaje humano.

En los años 1970s se cuentan múltiples ejemplos de desarrollos, algunos de ellos mostrados en la tabla 1.2, los cuales permitieron la conformación de áreas de estudio. De esta manera es posible caracterizar esta década a partir del área de investigación o metodología(s) empleada(s) en los proyectos de investigación [10]:

- En el área de *Reconocimiento de Palabras (Isolated-Word Recognition)* o *Reconocimiento Discreto de Voz (Discrete-Utterance Recognition)*, se tienen referencias como la proporcionada por Velichko y Zagoruyco en la Unión Soviética [52]. Sus estudios proporcionaron ideas acerca del uso del Reconocimiento de Patrones en Reconocimiento por Voz. Las investigaciones de Sakoe y Chiba [25] en Japón proporcionaron información sobre como usar de forma exitosa, los métodos de programación dinámica para reconocimiento de hablantes. Por su parte Itakura en los Estados Unidos de Norteamérica, mostró como las ideas de codificación de voz mediante LPC, pueden ampliarse al reconocimiento por voz mediante el uso de una métrica basada en estos parámetros espectrales [18].

- En el área de *Dictado Automático de Vocabulario Extenso (Large Vocabulary Automatic Speech Dictation)*, se tienen referencias del trabajo desarrollado en la IBM, en donde los investigadores se centraron en tres tareas principales: el llamado Lenguaje Raleigh para consultas simples a bases de datos [8], el Lenguaje Laser de Texto para Patentes usado para transcribir patentes mediante Laser [19] y el Sistema de Tareas de Oficina usado para dictado de cartas simples [19].

- Como una división del área anterior, se generaron investigaciones en lo que fueron los Laboratorios AT&T Bell. Los investigadores realizaron una serie de experimentos con el objetivo de hacer sistemas de reconocimiento de voz completamente independientes del hablante para aplicaciones en telecomunicaciones [39], particularmente servicios automáticos a distancia mediante la conversación entre una máquina y un hablante. Con este fin se realizaron diferentes algoritmos para tratar con la gran variabilidad de las palabras y expresiones de una extensa población. Los primeros resultados de esta investigación se obtuvieron una década después y proporcionaron diferentes modelos para la voz.

En la década de 1980s se identifican las siguientes características principales [10]:

- Prevaleció un enfoque hacia el reconocimiento de palabras conjuntas o frases (*connected word recognition)*, a diferencia de la década anterior en donde el enfoque principal fue

el reconocimiento de palabras y fonemas de forma individual. Así el objetivo de esta nueva generación de investigaciones fue el reconocimiento de palabras concatenadas o cadenas de palabras pronunciadas de forma continua o fluida, con base en patrones de palabras concatenadas. Una aplicación de estas investigaciones fue la desarrollada por la compañía Nippon Electric Corporation, donde Sakoe [24] programó un sistema dinámico de dos niveles para lograr el reconocimiento de frases pronunciadas de forma fluida.

- La metodología en las investigaciones se caracterizó por un cambio en la tecnología al pasar de un enfoque basado en plantillas o patrones de referencia a métodos de modelado basado en estadísticas. Entre los principales métodos puede identificarse los Modelos Ocultos de Markov (*Hidden Markov Models, HMM*).

Aún cuando la diversidad de aplicaciones e investigaciones se incrementó en la década pasada (1990s) es posible mencionar los siguientes aspectos:

- Los laboratorios de la DARPA (*Defense Advanced Research Projects Agency*), continuaron con los proyectos iniciados en la década pasada relacionados con reconocimiento de vocabulario extenso y con sistemas de reconocimiento de secuencias continuas de voz [10]. También dando énfasis a la recuperación de información hablada de sistemas en vuelo y a la transcripción de noticias o mensajes transmitidos.

- El uso de la tecnología de reconocimiento por voz se incrementó dentro de redes de telecomunicación tradicionales (teléfono) para automatizar y mejorar la operación de servicios [5].

- Se intensificó el uso de Modelos Ocultos de Markov como metodología en los sistemas de reconocimiento [7], [28], [41]. Adicionalmente se inicia el uso de Modelos de Mezclas Gaussianas (*Gaussian Mixture Models, GMM*) como una metodología basada en estadísticas [11], [12], [13].

A pesar de que la identificación es el caso más general y enriquecedor en el ámbito científico, comparado con la verificación, la tendencia en las aplicaciones comerciales en la última década (2000s) es la siguiente [17]:

- La mayoría de los sistemas comerciales consiste en verificación por voz.

- Estos sistemas están basados en esquemas donde se tienen usuarios dispuestos o voluntarios, es decir que desean ser identificados por el sistema.

- Son dependientes de texto; se pronuncian frases pequeñas proporcionadas por el sistema. Generalmente las claves de acceso son dígitos individuales.

- Siguen usando las principales metodologías de la década pasada. Adicionalmente se proponen nuevas metodologías como las Máquinas de Soporte Vectorial [23], [45], [51], [53], [55], [56] (*Support Vector Machines, SVM*) y Redes Neuronales Artificiales [42] (*Artificial Neural Networks, ANN*).

Existen diferentes compañías que han desarrollado aplicaciones comerciales de sistemas de reconocimiento [34]: ITT, Lernout & Hauspie, T-NETIX, Veritel y Voice Control Systems. El objetivo principal de sus aplicaciones se orienta al resguardo de seguridad en controles de acceso, banco por teléfono y manejo de tarjetas de crédito vía telefónica, con el objetivo de reducir transacciones fraudulentas.

La tendencia en el uso de la voz como característica biométrica en sistemas de reconocimiento y el que éstos se realicen con calidad telefónica reside, principalmente, en los siguientes hechos [17]:

- En muchas regiones el teléfono alámbrico es el único o principal medio de comunicación, por lo que el reconocimiento por voz resulta natural como sistema de seguridad.
- Los sistemas telefónicos constituyen un sistema que provee una red que, además de ser popular, tiende a ser ubícuita y desde la cual se pueden obtener las muestras de voz.
- Debido a los puntos anteriores, no existe necesidad de un procesamiento adicional en la captura de las muestras, a diferencia de sistemas de reconocimiento basados en otras características biométricas, como las huellas digitales o la cara.
- En general, el costo y tipo de la infraestructura requerida para la implementación es reducido y accesible.
- Se cuenta actualmente con una experiencia en esta área en investigación, desarrollo y evaluaciones, de más de 50 años, la cual incluye aplicaciones comerciales.

Las restricciones aplicadas a estos sistemas, generalmente con la finalidad de incrementar la precisión, en algunos casos pueden volverlos inflexibles a modificaciones o alteraciones, por ejemplo en el ambiente de operación o para usuarios no dispuestos o involuntarios. Para estos casos la verificación por voz independiente de texto es más adecuada.

A pesar de la amplia experiencia acumulada en el reconocimiento de hablantes, no es inmediato determinar alguna dirección clara en el futuro de las investigaciones. Sin embargo existen retos que pueden orientarlas, como son la pérdida de consistencia de los sistemas frente a variaciones en el canal y otras condiciones de error, como por ejemplo en el micrófono o la existencia de un SNR (*Signal to Noise Ratio*) bajo.

Muchos de los sistemas basan su operación en características acústicas de la voz (por ejemplo su espectro en frecuencia), por lo que las variaciones en el canal influyen considerablemente en el desempeño global. Un desacoplamiento o rompimiento de dicha dependencia, en alguna forma aún no considerada, haría posible una consistencia más robusta en las aplicaciones [17]. Adicionalmente existen otros niveles en el sistema de audición humano, que no han sido explorados o que se están empezando a explorar, por ejemplo la habilidad del oído humano para reconocer voces que le son familiares. Esta perspectiva motiva la posibilidad de emular y aprovechar hábitos del habla de largo plazo en los sistemas automáticos.Las líneas de investigación que apuntan hacia estos tópicos han emergido recientemente y quizá requieran de algunas de las técnicas ya desarrolladas o incluso de su fusión o combinación.

Capítulo 2

Sistema Automático de Verificación

En este capítulo se presenta la aplicación de una metodología para la creación de un sistema automático de verificación de personas a partir de la voz, basado en la teoría de Máquinas de Soporte Vectorial como sistema de clasificación.

El proceso de verificación requiere, como parte esencial, la creación de módelos correspondientes a los usuarios considerados como válidos e inválidos. Esta tarea se basa en la hipótesis fundamental de que es posible asociar de forma biunívoca registros de voz de cada hablante con el mismo. Cuando se logra satisfacer esta hipótesis en alguna implementación, se cuenta entonces con un procedimiento de asociación biunívoca. De esta manera, dada una muestra de voz y a través del mismo procedimiento, es posible determinar si ésta corresponde o no al usuario que solicita ser verificado. Es entonces necesario conformar un modelo para cada voz y establecer así un procedimiento de asociación. En las siguientes secciones se exponen los planteamientos al respecto.

2.1. Metodología

En la figura 2.1 se muestra un esquema que plantea la metodología utilizada en este trabajo. Los etapas son implementadas de izquierda a derecha y de arriba hacia abajo. Dicho esquema incluye la comprobación del sistema con voz en tiempo real.

La metodología propuesta consiste en conformar la base de registros de voz para una número fijo de personas. A partir de ésta es posible obtener los vectores de características para cada hablante. Dicho conjunto de vectores permitirá definir los conjuntos de entrenamiento y de prueba. El conjunto de entrenamiento será usado por el sistema de soporte vectorial para conformar el modelo correspondiente a cada hablante. La formación de modelos es determinante en la creación del sistema automático. Una vez construidos los modelos

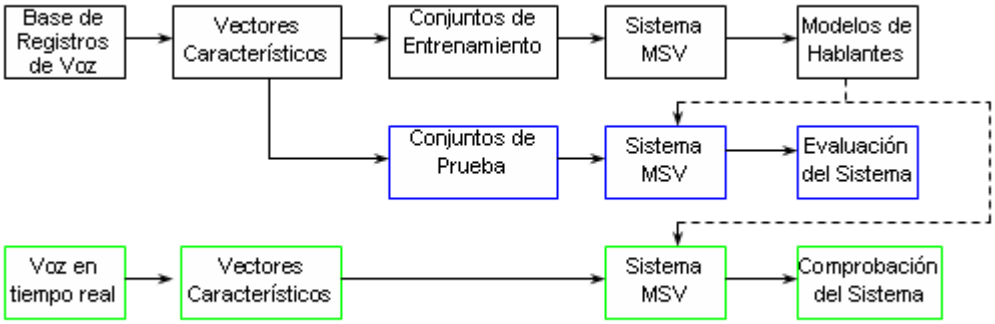

Figura 2.1: Metodología para implementar un sistema automático de verificación mediante Máquinas de Soporte Vectorial.Las flechas continuas muestran el orden de aplicación de cada etapa. Las flechas punteadas muestran la dependencia de ciertas etapas a los modelos creados.

de los hablantes, se aplicarán los conjuntos de prueba para obtener una evaluación del desempeño del sistema implementado.

A fin de comprobar la operación se realizarán pruebas con voz en tiempo real. Para ello la voz será grabada desde un micrófono, sometida a la extracción de vectores de características y después ingresada al sistema con soporte vectorial, para determinar la operación ante solicitudes válidas e inválidas.

2.2. Base de Registros de Voz

La creación de una base de registros de voz tiene como principal objetivo contar con muestras representativas de voz de cada una de las personas que participarán en el sistema, sea como usuarios válidos (enrolados) o como usuarios no válidos. En este último caso, las muestras de voz proporcionadas por los participantes permitirán formar, aunque no totalmente, el modelo de los usuarios no válidos, conocido comunmente como complemento o *background*. Se debe considerar que para un usuario, denotado por U_i, el resto de los usuarios enrolados también servirá para conformar el *background*. La tabla 2.1 establece la notación que será válida a partir de aquí.

Notación	Significado
U_i	Un hablante que participa con su voz en el sistema de verificación.
$\{U_j\}_{j=1}^{N}$	Conjunto que denota a todos los usuarios o hablantes (válidos e inválidos) involucrados en el sistema, referido como universo de usuarios. En este caso el total de usuarios es denotado por N.
$\{U_i\}^c$	Conjunto que denota al complemento de usuarios a U_i respecto del universo de usuarios.
S_i	Muestra de voz del usuario i.
$\{x_{ik}\}_{k=1}^{l}$	Vectores de características del usuario U_i
M_i	Modelo del usuario i.

Tabla 2.1: Notación utilizada para denotar los elementos del sistema de verificación.

Para la conformación de la base de registros de voz pueden identificarse dos puntos principales a considerar:

Representatividad. Los registros deben ser representativos por su contenido de información. Para ello se consideran algunos aspectos de la variabilidad de la voz en las escalas de tiempo y amplitud, así como el contenido de fonemas y sílabas de las frases pronunciadas.

Universalidad. La conformación de conjuntos complemento a un usario ($\{U_i\}^c$) es únicamente respecto al conjunto total o universo de usuarios, es decir no incluye de forma real a todos los usuarios no válidos. La conformación de complementos para cada usuario será universal en la medida en la que se incluya a potenciales usuarios no válidos del sistema.

A pesar de que formalmente el lenguaje no es una característica intrínseca del ser humano, pues es adquirido de manera empírica a través del tiempo y con la adecuada socialización, si puede considerarse como una característica distintiva de la especie. La voz es una sucesión lógica de sonidos básicos o fundamentales llamados fonemas. Físicamente, dichos fonemas son ondas mecánicas esféricas con una representación matemática particular, la cual puede ser periódica. Cada lenguaje, entendido éste como un conjunto de símbolos que permiten la comunicación verbal (v.g. español, ingés, francés), posee un conjunto finito de fonemas y ciertas reglas de ordenamiento que determinan la lógica y semántica en las secuencias o sucesiones pronunciadas, las cuales pueden incluir intervalos de silencio. La creación de un sistema automático que determine si una persona es quien dice ser (verificación), se basa, como se ha mencionado, en la hipótesis de que las características biológicas y fisiológicas del sistema del habla son tales, que la voz generada posee propiedades cualitativas y cuantitativas que son únicas para cada hablante. Esta hipótesis fundamental es sin embargo empírica, pues

no se tiene en el campo de la biometría referencias que demuestren la unicidad en alguna(s) propiedad(es) de la voz.

Aún si se supone una validación correcta de la hipótesis, la gran variabilidad en las propiedades cuantitativas y cualitativas de la voz, impide la comparación directa de fonemas, como técnica de reconocimiento de hablantes. En consecuencia es necesario considerar el tratamiento, análisis y comprobación de la(s) hipótesis propuesta(s) y la implementación de metodologías que traten con la variabilidad de la voz y con los factores que inciden en dicha variabilidad, a fin de reducir el porcentaje de error e incrementar el desempeño del proceso de verificación.

Idealmente la creación de un sistema de verificación universal por voz es deseable, sin embargo actualmente no es viable por razones de implementación. De esta manera la base de registros de voz deberá considerar tanto la variabilidad de la voz como la orientación o fin al que se destinará, con el objetivo de acotar sus alcances y con ello también sus características. Algunos de los aspectos específicos que pueden ser considerados para la conformación de una base de registros se enlistan a continuación [34], [30], [46]:

1. **Sesiones de grabado.** A fin de considerar la variabilidad de la voz en diferentes días, es posible realizar para cada persona, grabaciones de voz en dos sesiones con al menos cinco días de separación.

2. **Velocidad de lectura.** Consiste en el grabado de voz a la velocidad usual del hablante y también a mayor y menor velocidad de la usual. La variabilidad de la voz en la escala temporal, es una de las dificultades que se presenta en el reconocimiento de fonemas y palabras. Considerar este aspecto en el grabado de los registros, a pesar de ser un tanto subjetivo, amplía la utilidad de los mismos y su representatividad.

3. **Frases de prueba.** El objetivo de estas frases es contar con datos de prueba para verificar el comportamiento del sistema ante intrusos; en este caso se tienen frases particulares para cada hablante así como frases que pronuncian por igual todos los hablantes. Estas frases también permiten el uso de los registros de voz en sistemas que son dependientes e independientes del texto pronunciado.

4. **Niveles de ruido.** La presencia de ruido en el ambiente de grabación no favorece el proceso de verificación pues dificulta el procesamiento de la señal real. Bajo esta consideración, la grabación debiera realizarse en un ambiente controlado sin ninguna otra influencia audible diferente a la voz. Sin embargo, si el sistema será usado en un ambiente donde existe influencias de sonido externas (por ejemplo en una oficina), lo adecuado sería considerar dichos niveles de ruido en los registros de voz, pues las muestras que se ingresen al sistema lo incluirán.

5. **Edades y sexo de los hablantes.** Se sabe que el intervalo de frecuencias audibles que conforman una señal de voz, está determinada en parte, por el sexo del hablante, por lo que es de importancia adquirir registros de ambos sexos a fin de considerar la variabilidad genérica de la voz. No existe entonces restricción para la participación de hombres y mujeres a menos que el sistema, por la naturaleza de su uso, lo requiera. Por otro lado se sabe que la voz es sensible a la edad en un mismo hablante y que la mayor variabilidad de ésta sucede en la adolescencia, por ello debe identificarse y considerarse de forma adecuada la población que estará involucrada en el sistema de reconocimiento, para garantizar la estabilidad temporal en el desempeño del mismo.

6. **Número de hablantes.** Se considera que cada hablante genera un registro de voz. El número de hablantes determinará directamente la información almacenada. En teoría, dada la hipótesis de unicicidad de la voz, el incrementar el número de registros no incrementa la complejidad de la operación de un sistema de reconocimiento. En realidad el número de registros es un factor importante para conformar un sistema confiable, ya que incrementa la probabilidad de encontrar correlación entre los registros de voz.

7. **Sistema de grabado.** Es posible considerar la influencia del sistema de grabado en los datos obtenidos, para ello pueden utilizarse micrófonos de diferente calidad para realizar grabaciones simultáneas o en sesiones diferidas. Este aspecto agrega un factor que permite adecuar los registros de voz a diferentes sistemas de grabado. También permite analizar la influencia de estos medios en los sistemas de reconocimiento.

8. **Espontaneidad en la voz.** Este aspecto considera diferencias en la pronunciación entre voz por lectura de texto y por pronunciación espontánea. Si se considera que los sistemas de reconocimiento pueden operar solicitando la pronunciación de un texto leído o la repetición de una frase asignada, deben considerarse ambos tipos de grabación.

9. **Variación del canal.** Es posible variar la distancia entre el micrófono y el hablante a fin de considerar la influencia de variaciones en el canal de transmisión, lo cual puede representar una condición real de operación del sistema de reconocimiento.

10. **Distribución de fonemas y sílabas.** Una muestra de voz es representativa en la medida en la que incluye los fonemas y sílabas del idioma en la que se obtiene. Las frases pronunciadas deben tener una distribución de fonemas y sílabas que corresponde con la distribución usual del idioma.

2.2.1. Protocolo de Grabación

Con base en los aspectos planteados previamente es posible establecer un protocolo de grabación, el cual permite uniformizar la adquisición de muestras de voz de cada uno de los participantes. Las tareas que integran este protocolo se describen a continuación:

a) *Pronunciación del nombre del hablante.* Esta tarea permite la identificación de los registros de voz y proporciona una frase particular para cada usuario la cual le es familiar y permite la grabación de una muestra de voz con espontaneidad.

b) *Pronunciación de los dígitos del 0 al 9.* La pronunciación solicitada al hablante inicia con el número cero y hasta el nueve. De esta manera se genera una primera frase equivalente para todos los hablantes así como el registro de pronunciación individual de los 10 dígitos de forma espontánea.

c) *Pronunciación individual de dos cadenas de dígitos.* Se generaron de forma aleatoria las cadenas de dígitos *4, 2, 3, 1, 7* y *8, 6, 0, 1, 5*, las cuales son pronunciadas con una pausa intermedia entre ambas por todos los participantes. Esta tarea porporciona la pronunciación de dígitos en desorden y de forma individual lo cual permite contar con información sobre posibles diferencias con respecto a la pronunciación en forma ordenada, también proporciona una frase general obtenida por lectura.

d) *Repetición de una cadena de dígitos.* Se genera de forma aleatoria y para cada hablante una cadena de 5 dígitos (ver Tabla 2.2). La cadena es repetida dos veces con una pausa intemedia entre cada una. Esta tarea proporciona información sobre la repetición de cadenas de dígitos leídos individualmente y sobre posibles similitudes o diferencias en la pronunciación.

e) *Pronunciación de frases de texto particulares.* Se seleccionaron para cada hablante, dos frases de texto de diferentes fuentes, considerando aquellas que tuvieran alrededor de 15 palabras y sin considerar a priori su contenido fonético o silábico (ver Tabla 2.2). Esta tarea proporciona frases leídas de texto y particulares para cada hablante, las cuales pueden ser consideradas como aleatorias.

f) *Pronunciación de una misma frase de texto.* Todos los participantes leyeron la siguiente frase:
Fue Platón quien dio su estructura dialéctica, mucho más amplia y sutil, al nuevo género literario, aportándole un espíritu mucho más audaz y una claridad artística que anteriormente no tenía. Es por ello que está considerado el más grande de los prosistas helénicos; sus diálogos escritos alcanzaron la madurez y perfección como paradigma clásico de las discusiones filosóficas a cielo abierto.
Esta frase proporciona información representiva de la pronunciación de cada hablante, de los diferentes fonemas y sílabas que conforman el español. También trata con la pronunciación leída de texto, equivalente para todos los participantes.

g) *Pronunciación de texto con diferente rapidez.* Los participantes leyeron el texto de la tarea f) con mayor y posteriormente con menor rapidez de lo que usualmente lo hacen. Con ello se cuenta con información representativa sobre posibles modificaciones de la

voz ante circunstancias como el nerviosismo o la calma que favorecen la presencia de errores.

Las siete tareas descritas previamente, conforman el protocolo de grabación que fue usado para obtener cada uno de los registros de voz S_j los cuales generan la base de registros denominada *Voces-MCyTI*. Algunas de las características consideradas para la obtención de dichas muestras son las siguientes:

1. Las sesiones de grabación fueron realizadas en una oficina de la UAM Iztapalapa, a puerta cerrada, pero sin mayor cuidado respecto a la presencia de fuentes de ruido externas.

2. Después de cada una de las tareas del protocolo, se realizó una pausa y una reanudación de la grabación. En las tareas de lectura se solicitó al participante se familiarizara con la frase a pronunciar antes de realizar la grabación.

3. Los días y horarios de grabación fueron dispersos, dependiendo de la disposición de los participantes.

4. El micrófono para las grabaciones fue de calidad media. Las muestras S_j fueron obtenidas a una frecuencia de muestreo de 8000Hz considerando el uso del sistema a través de línea telefónica. No hubo control estricto respecto de la distancia entre el hablante y el micrófono.

5. La distribución de edades de los 17 participantes se ilustra en la figura 2.2.

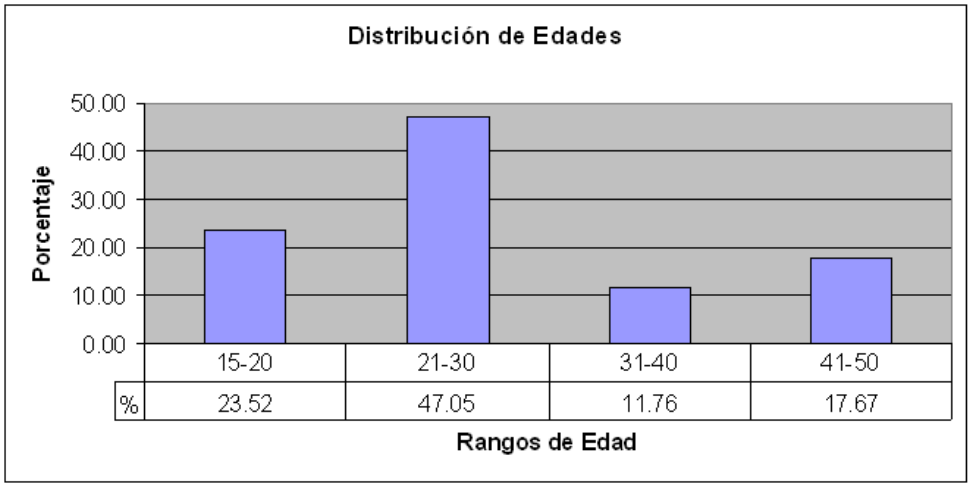

Figura 2.2: Distribución de edades de los participantes en la base *Voces-MCyTI*.

En las tablas 2.2 y 2.3 se muestran algunas de las frases particulares pronunciadas por los participantes así como la referencia al archivo de sonido generado.

Archivo	**Tarea d)**	**Tarea e)**	**Fecha de grabación**
mcyti_1.wav	4,6,2,3,9	1) Alemania abre sus puertas al mundo con la duda de qué tan lejos llegará su selección. 2) Venderán cerveza en los estadios. Para los teutones tomar esta bebida es casi una religión.	23/05/2006
mcyti_2.wav	3,2,9,7,1	1) La única que te da el doble de puntos del programa de recompensas vida bancomer. 2) Cada veintemil puntos acumulados puedes canjearlos por un boleto de avión, viaje redondo, a cualquier parte del país.	15/06/2006
mcyti_3.wav	7,1,3,6,2	1) La danzonera dimas fue fundada en 1921 por Amador Pérez Torres mejor conocido como Dimas. 2) Chávez, lector de poesía, precisó que con esta reedición rinde homenaje a López Velarde y a Fermín Revueltas.	19/10/2006
mcyti_4.wav	5,9,4,8,0	1) El caifán Oscar Chávez cantará danzonetes, sus clásicas *por tí* y *la niña de Guatemala*, en su concierto. 2) Chávez agregó: Todo lo que ha generado mi disco *Chiapas* es para apoyar a las comunidades indígenas.	18/10/2006
mcyti_5.wav	8,5,0,2,1	1) Erasmo Capilla nació en Jalapa, estudió música con su padre y después en la Universidad Veracruzana. 2) A los 14 años de edad ganó el Concurso Nacional de Violín para la Juventud Mexicana.	18/10/2006
mcyti_6.wav	4,6,9,7,3	1) Para desmantelar el amarillismo, la producción de *Vuelo 93* se ha cuidado en documentar la historia. 2) Asimismo, los actores fueron escogidos en virtud a su semejanza física con las personas reales.	18/10/2006
mcyti_7.wav	6,7,9,5,4	1) La hipertensión arterial y el hábito de fumar favorecen la muerte de millones de personas. 2) Reducir el consumo de grasas, incrementar la ingesta de frutas y verduras y modificar hábitos de vida.	18/10/2006
mcyti_8.wav	2,0,8,3,1	1) A veces la ciencia olvida que los hombres son el otro componente vital del proceso reproductivo. 2) De los trecemil ochocientos sesenta embarazos reportados mil quinientos seis se interrumpieron.	19/10/2006
mcyti_9.wav	1,3,7,6,0	1) En su más reciente película *Munich*, el primer palestino asesinado por los servicios secretos. 2) En la inexorable realidad de Gaza no hay Aladino ni lámpara maravillosa que valgan.	19/10/2006

Tabla 2.2: Frases particulares pronunciadas por los participantes en la base *Voces-MCyTI*.

Archivo	**Tarea d)**	**Tarea e)**	**Fecha de grabación**
mcyti_10.wav	2,9,5,4,8	1) La historia era increíble, en efecto, pero se impuso a todos, porque sustancialmente era cierta. 2) Verdadero era también el ultraje que había padecido; sólo eran falsas las circunstancias.	23/10/2006
mcyti_11.wav	1,4,9,3,6	1) Nació en el sudeste de Asia, no es muy alta y de color verde pálido, más conocida como *noni*. 2) Las últimas investigaciones indican que la planta está repleta de químicos que revierten el cancer.	20/10/2006
mcyti_12.wav	8,7,1,0,4	1) Un indígena asiste en Sucre, Bolivia, a la presentación del programa *Microsoft Windows* en Quechua. 2) Evo Morales anunciará la nacionalización de las concesiones para la exportación de las riquezas naturales.	25/10/2006
mcyti_13.wav	3,6,5,9,2	1) En la opuesta margen resplandecía, bajo el último sol o bajo el primero, la evidente Ciudad de los Inmortales. 2) Esta ciudad es tan horrible que su mera existencia y perduración contamina el presente y el porvenir.	26/10/2006
mcyti_14.wav	5,3,1,0,8	1) La muerte hace preciosos y patéticos a los hombres; todo es irrecuperable y azaroso. 2) Entre los inmortales, en cambio, cada acto y cada pensamiento es el eco de otros que en el pasado lo antecedieron.	23/10/2006
mcyti_15.wav	4,9,2,7,6	1) La mujer quiere resistir, pero dos hombres la han tomado del brazo y la hechan sobre Otálora. 2) Que le han permitido el amor, el mando y el triunfo, porque ya lo daban por muerto.	26/10/2006
mcyti_16.wav	0,2,7,1,9	1) Le encontraron la misma sustancia pocos días después de haberse proclamado campeon del Tour de Francia. 2) El entrenador Trevor denunció el uso de esteroides THG, entre deportistas de Estados Unidos.	07/11/2006
mcyti_17.wav	3,6,0,4,9	1) Básteme recordar que el desertor malhirió o mató a varios de los hombres de cruz. 2) No iba a consentir el delito de que se matara a un valiente y se puso a pelear contra los soldados.	16/11/2006

Tabla 2.3: Frases particulares pronunciadas por los participantes en la base *Voces-MCyTI*.

2.2.2. Análisis Fonológico y Silábico

A continuación se presenta el análisis fonológico y silábico de la frase pronunciada en las tareas f) y g) del protocolo de grabación utilizado. El análisis consisitió en separar por fonemas y sílabas dicha frase y obtener la distribución de frecuencias correspondiente. Esta distribución es la que corresponde con la usual del español [30], criterio que determinó la selección de la misma.

La distribución de fonemas se realizó considerando un total de 28, con base en la notación SAMPA (*Speech Assesment Methods Phonetics Alphabet*) para el español (ver Apéndice A.1). Así, la frase consta de un total de 318 fonemas y su distribución se encuentra resumida en la tabla 2.4 y esquematizada en la figura 2.3.

a	12.89 %	**e**	10.37 %	**o**	9.43 %	**s**	8.49 %
i	8.49 %	**n**	5.03 %	**t**	5.34 %	**d/D**	4.72 %
l	5.66 %	**k**	5.34 %	**u**	4.71 %	**r**	4.71 %
m	3.14 %	**p**	2.51 %	**T**	1.88 %	**b/B**	0.63 %
rr	2.20 %	**G**	0.63 %	**x**	0.63 %	**f**	1.26 %
j	0.94 %	**tS**	0.63 %	**L**	0.31 %	**jj/J/w**	0 %

Tabla 2.4: Distribución de fonemas en la frase de las tareas f) y g) del protocolo de grabación.

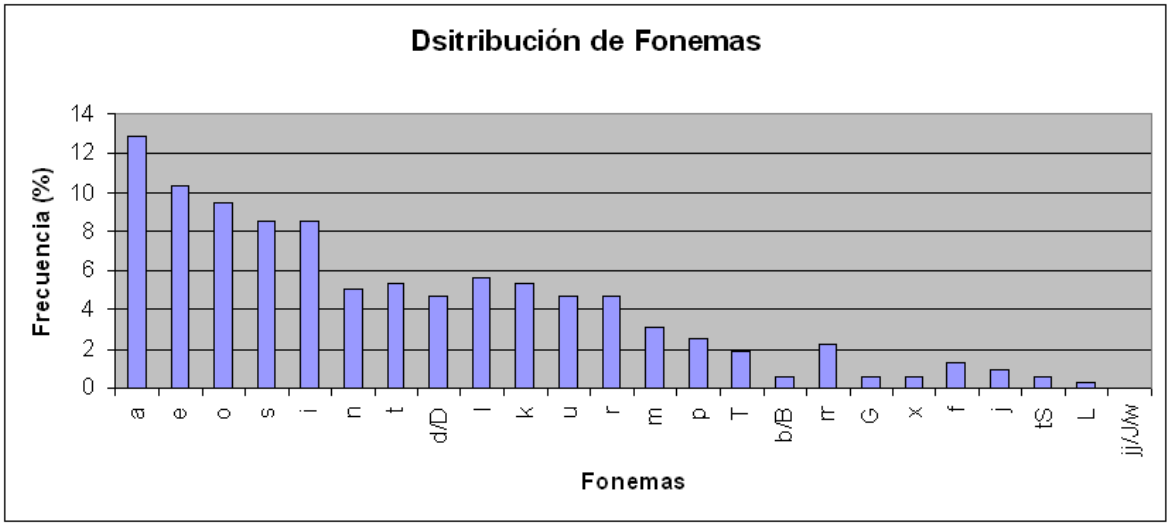

Figura 2.3: Gráfica de la distribución de fonemas en la frase de las tareas f) y g) del protocolo de grabación.

La distribución de sílabas se obtuvo considerando 8 tipos. Con base en ésto la frase considerada consta de un total de 136 sílabas y la distribución de éstas se encuentra resumida en la tabla 2.5 y esquematizada en la figura 2.4.

CLAVE	DESCRIPCIÓN	PORCENTAJE
CV	Consonante-Vocal	46.32 %
CVC	Consonante-Vocal-Consonante	22.79 %
V	Vocal	6.61 %
VC	Vocal-Consonante	8.08 %
CVV	Consonante-Vocal-Vocal	6.61 %
CCV	Consonante-Consonante-Vocal	3.67 %
O	Otras sílabas	5.88 %

Tabla 2.5: Distribución de sílabas de la frase de las tareas f) y g) del protocolo de grabación.

La finalidad de este análisis es contar con una frase balanceada que contenga la información correspondiente a los fonémas y sílabas que pueden ser pronunciados en el español y que sea así representativa del lenguaje.

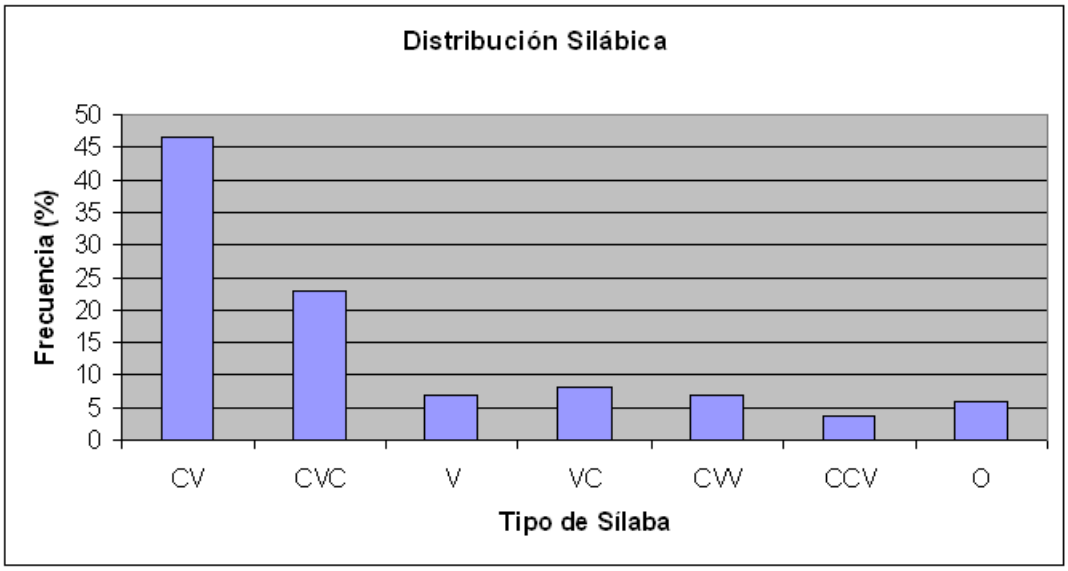

Figura 2.4: Gráfica de la distribución de sílabas de la frase de las tareas f) y g) del protocolo de grabación.

2.3. Procesamiento de la Voz

En la sección 1.3 (Estado del Arte) se hizo mención de las diferentes técnicas y procedimientos que pueden ser implementados para que, dada una señal de voz se obtengan valores numéricos que la caractericen o representen. En general, los objetivos principales de dicha representación son:

- Sintetizar la información contenida en la señal de voz, de manera que se reduzca el número de datos o valores que la representan.
- Evitar el manejo de información poco significativa, es decir los datos que dificultan las tareas de reconocimiento, como el ruido o los intervalos entre palabras (silencios).
- Minimizar el manejo de información redundante contenida en la señal de voz, por ejemplo los fonemas pronunciados.
- Facilitar la definición de un modelo que corresponda de manera única a la voz considerada.
- Proporcionar los datos de entrada, simplificar la operación y en su caso mejorar la efectividad de algún sistema de clasificación o verificación automática.

A continuación se describe brevemente el procedimiento que, en particular, se propone para la realización de este trabajo, mismo que se basa en el análisis de la voz por su contenido frecuencial. Por su importancia en el procesamiento de la voz, ciertos módulos serán descritos detalladamente en secciones posteriores. La figura 2.5 muestra la secuencia de módulos a los que se somete la señal de voz, $S(n)$, hasta obtener los vectores de características, $x(m)$.

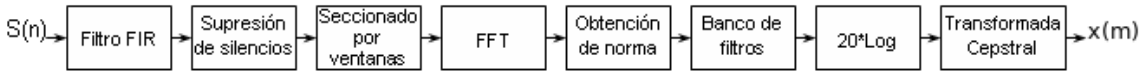

Figura 2.5: Esquema del procesamiento de voz para extracción de características

El procedimiento consiste en la obtención de los llamados coeficientes MFCC (*Mel Frecuency Cepstral Coeficients*), los cuales constituyen los vectores de características de la voz, con los que se pretende cumplir los objetivos establecidos previamente. Su uso es propuesto y descrito en diferentes artículos de investigación que han sido publicados recientemente y en los cuales se presenta como una técnica exitosa en el reconocimiento de hablantes [17], [36], [46].

1. Filtro FIR: La señal de voz es sometida a un filtro con respuesta al impulso finita. Este es un filtro pasa banda entre los 300Hz y 3300Hz, de orden 51, simétrico. Para

mejorar la resolución espectral (debido al truncamiento temporal de la señal), se aplica una ventana de Hamming antes del filtrado. La finalidad del filtro es eliminar las frecuencias correspondientes al ruido y obtener una mejora en el SNR.

2. **Supresión de silencios:** Se implementa un procedimiento para suprimir de la señal los intervalos entre palabras los cuales corresponden al ruido en el ambiente. Este procedimiento consiste en analizar el comportamiento de la potencia promedio de la señal en ventanas de longitud M.

3. **Seccionado por ventanas:** El seccionado o separación de la señal por ventanas es la preparación para aplicar la transformada de Fourier a la señal. El seccionado o segmentado de la señal se realiza en ventanas de tiempo donde el proceso es considerado estacionario. Para la voz un intervalo típico es de 20 a 30 ms. También se considera un traslape entre ventanas consecutivas de 10 ms.

4. **FFT:** Se aplica la transformada rápida de Fourier (*Fast Fourier Transform*) a cada una de las ventanas consideradas. Esta transformación puede ser realizada mediante diferentes algoritmos y usando una cantidad de puntos N variable. Generalmente N es una potencia de 2 y es más grande que el número de muestras M de las ventanas. Usualmente se usan 512 puntos.

5. **Obtención de norma:** De manera análoga a la transformada de Fourier continua, su similar discreta está en el plano complejo, por lo que es necesario obtener la norma o módulo de cada uno de los puntos obtenidos. El resultado de esta operación conforma una representación gráfica simétrica, por lo que sólo es necesario considerar la mitad de los puntos obtenidos por cada ventana analizada.

6. **Banco de filtros:** Con el fin de obtener una representación de la envolvente de cada uno de los espectros resultantes, se aplica a cada uno un banco de filtros. Este es una serie de filtros pasabanda que son multiplicados uno a uno con el espectro dado y promediados los resultados correspondientes. De esta manera se reduce nuevamente la longitud de los vectores pudiendo, dado un filtro, enfatizar el comportamiento del espectro en una frecuencia particular. El banco es definido por la forma de los filtros y la escala de frecuencias a la que son ubicados; frecuencia inicial, frecuencia media y frecuencia final. Los filtros pueden ser de forma triangular y estar ubicados de acuerdo a la escala de frecuencias de Mel, obtenida a partir de la siguiente ecuación [17]:

$$f_{MEL} = 1000log(1 + \frac{f_{lin}}{1000})/log(2) \tag{2.1}$$

7. **Transformación 20*Log:** En esta parte se obtiene el logaritmo de cada uno de los coeficientes obtenidos y se multiplica por 20 para obtener valores en decibeles (dB). Los datos obtenidos son llamados vectores espectrales (denotados por V).

8. Transformada Cepstral [46]: La última transformada aplicada a la señal, sobre cada uno de los vectores espectrales, consiste en obtener a partir de éstos, los coeficientes cepstrales, los cuales están definidos por la siguiente ecuación:

$$x(n) = \sum_{k=1}^{K} V_k cos\left(\frac{n\pi}{K}\left(k - \frac{1}{2}\right)\right),\ n = 1, 2, \ldots, L. \tag{2.2}$$

donde V_k son los elementos de un vector espectral y K es la longitud de éste. L es el número de coeficientes cepstrales que se desean calcular ($L \leq K$).

Con respecto a los parámetros involucrados en la extracción de vectores de características, se tienen valores típicos empleados y con los que se han obtenido resultados favorables [40]. Los parámetros iniciales utilizados en el procesamiento de los registros de voz S_j son resumidos en la tabla 2.6.

Parámetro	**Valor**
Banda del filtro FIR	300Hz-3300Hz
Orden del filtro FIR	51, Hamming
Tiempo de ventana (M)	30ms (240 muestras)
Tiempo de traslape	10ms (80 muestras)
Orden de FFT (N)	512 puntos
Orden del Filtro Mel (K)	31
Orden del Vector Cepstral (L)	13

Tabla 2.6: Parámetros iniciales usados para el procesamiento de los registros de voz S_j y la obtención de los vectores de características $x_j(n)$.

2.3.1. Supresión de Silencios

Una señal de ruido puede definirse como cualquier influencia externa o diferente de nuestra señal objetivo, en este caso la voz. Cuando se realiza la grabación de la voz, tanto la influencia del ambiente como la del propio sistema de grabación, alteran la señal, modificando los niveles reales con los que originalmente es emitida, es decir se produce una distorsión. Los efectos totales correspondientes a un sistema de grabación en un ambiente dado, pueden ser obtenidos grabando dicho ambiente. La señal obtenida considera el ruido del sistema y puede así ser analizada para determinar sus características. Cuando la grabación incluye la voz de un hablante, los niveles de ruido permanecen y se superponen a la señal de voz, de manera que la señal resultante es una suma directa de ambas.

Bajo estas condiciones se propone una metodología para determinar en dicha señal los intervalos correspondientes al habla. Dos aspectos importantes que justifican tener un procedimiento como éste son:

- La posibilidad de reducir el tamaño de los registros de voz, previo a cualquier procesamiento de los mismos, lo cual disminuiría el tiempo de ejecución del sistema.
- Omitir de los registros de voz, información que no corresponde a los hablantes y que, de permanecer, agregaría correlación entre los conjuntos de vectores de características, provocando mayor dificultad en el proceso de clasificación.

La supresión de intervalos que no corresponden a voz, parte de la hipótesis de que es posible discriminar a partir de la intensidad de un punto de la señal grabada, si éste corresponde a voz o a ruido. Si suponemos que $S_j(n)$ representa una grabación, puede considerarse que es medida como una diferencia de potencial E (Volts). Entonces $S_j^2(n)$ es proporcional a la potencia de la señal:

$$P = E.I = E\frac{E}{R} = \frac{1}{R}E^2. \tag{2.3}$$

Al graficar $S_j(n)$ y $S_j^2(n)$ para algunas muestras, se observa que el comportamiento en la magnitud de esta última contrasta en mejor forma los periodos correspondientes a voz de los de ruido. De esta forma se usa $S_j^2(n)$ en lugar de $S_j(n)$ y se considera como la potencia de la señal. El procedimiento que se propone para realizar la discriminación de puntos correspondientes a ruido de los de voz se describe a continuación.

1. Se asumen dos hipótesis necesarias: a) Se considera que existen condiciones de grabación invariantes en cuanto a los niveles de ruido presentes en el sistema. Lo anterior significa que no existen fuentes ni sumideros de ruido en el tiempo de grabación de S_j. b) Cada uno de los registros S_j que sean sometidos al proceso de supresión de silencios, deben tener un intervalo de silencio o pausa del hablante de r_j segundos ($r_j > 0$) al inicio de la trama. La figura 2.6 muestra un segmento inicial del registro S_1.

2. Se obtiene la potencia promedio por segmentos de la señal con base en la siguiente ecuación:

$$P_j(n)_{prom} = \frac{1}{M}\sum_{k=n}^{n+M} S_j^2(k), \tag{2.4}$$

   es decir se obtiene el promedio de la potencia de la señal en un segmento a partir de la muestra n y hasta la muestra $n + M$. La figura 2.7 muestra un segmento inicial de la gráfica de $S_1^2(n)$ mientras que la figura 2.8 permite observar el efecto de la ecuación 2.4 para la misma muestra en el mismo segmento.

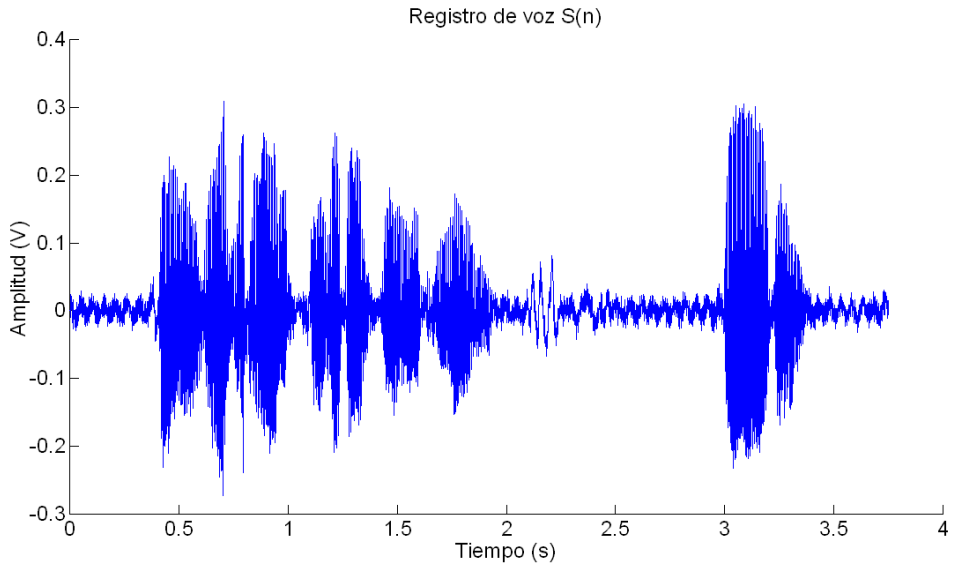

Figura 2.6: Segmento inicial del registro S_1.

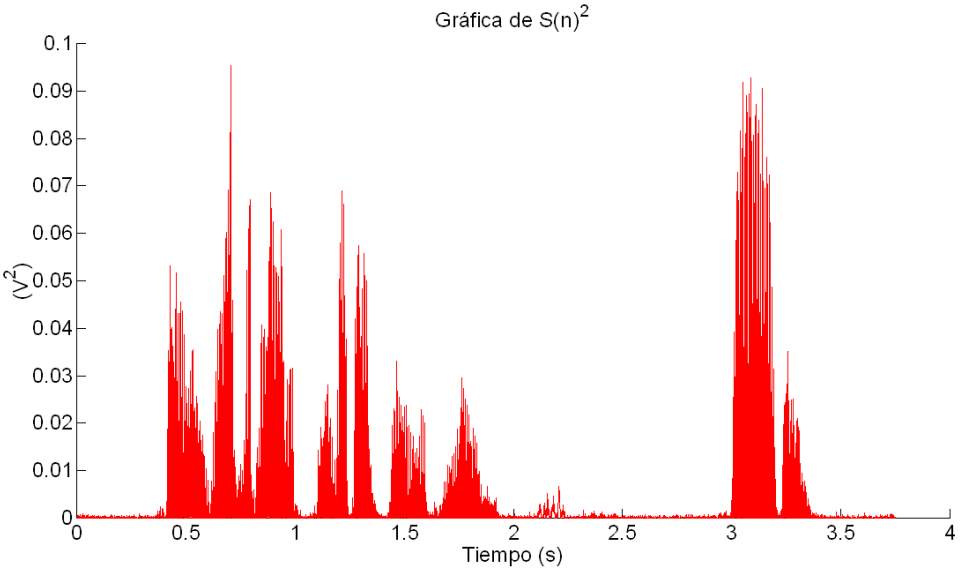

Figura 2.7: Esquema de $S_1^2(n)$.

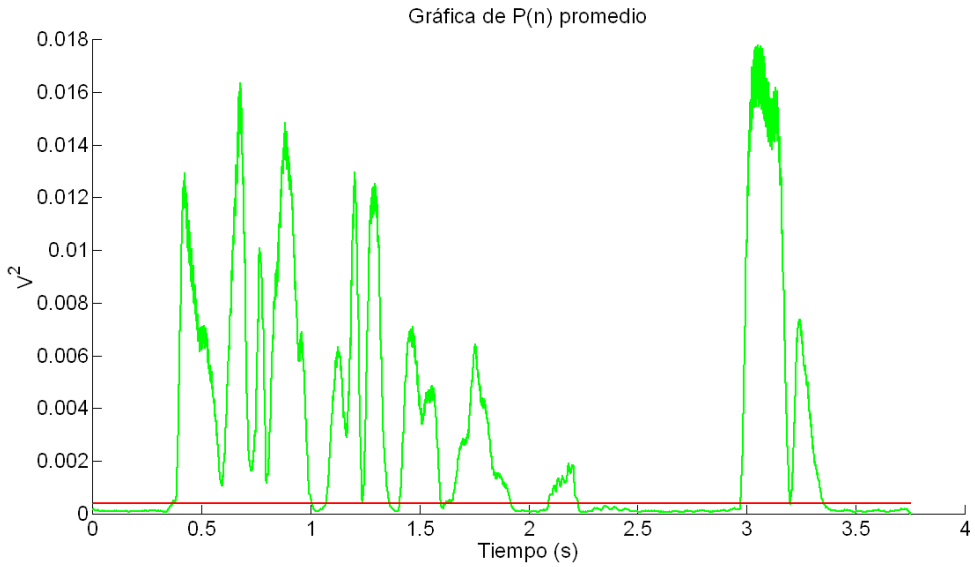

Figura 2.8: Gráfica de $P_1(n)_{prom}$ y nivel de umbral U_1.

3. Con base en $P_j(n)_{prom}$ y dado un valor de umbral u_j, se filtra $S_j(n)$ mediante la siguiente función:

$$S_j(n)_{voz} = \begin{cases} S_j(n) & \text{si } P_j(n)_{prom} \geq u_j \\ 0 & \text{si } P_j(n)_{prom} < u_j \end{cases} \tag{2.5}$$

La señal $S_j(n)_{voz}$ permite discriminar los intervalos de voz de las pausas hechas por el hablante y que corresponden al ruido del sistema. Para ello es suficiente con eliminar de $S_j(n)_{voz}$ los puntos de la señal para los que $P_j(n)_{prom} < u_j$.

4. Con base en las hipótesis del punto 1, se puede establecer que los r_j segundos iniciales de cada señal S_j corresponden a una muestra del ruido total del sistema y que esta muestra es estadísticamente equivalente a cualquiera otra de la señal, de longitud r_j y que corresponda también a ruido. Por lo tanto parámetros como el promedio, la varianza y/o la desviación estandar de la muestra, deben ser equivalentes o aproximados para cualquier muestra de ruido, considerada en las mismas condiciones. Por lo anterior el nivel de umbral u_j que proponemos se establece de la siguiente forma:

$$u_j = \frac{1}{r_j f_s} \sum_{k=1}^{r_j f_s} S_j^2(k) + c_j \sqrt{\sigma_{S_j}^2}, \tag{2.6}$$

donde f_s es la frecuencia de muestreo, $\sigma_{S_j}^2$ está dada por

$$\sigma^2_{S_j} = E\left[\left(S_j^2(n) - \frac{1}{r_j f_s}\sum_{k=1}^{r_j f_s} S_j^2(k)\right)^2\right], \tag{2.7}$$

y $c_j > 1$ es una constante que incrementa el valor de umbral ajustándolo para cada registro S_j.

En la figura 2.9 se esquematiza el comportamiento del procedimiento de supresión de silencios para un segmento inicial del registro de voz S_1.

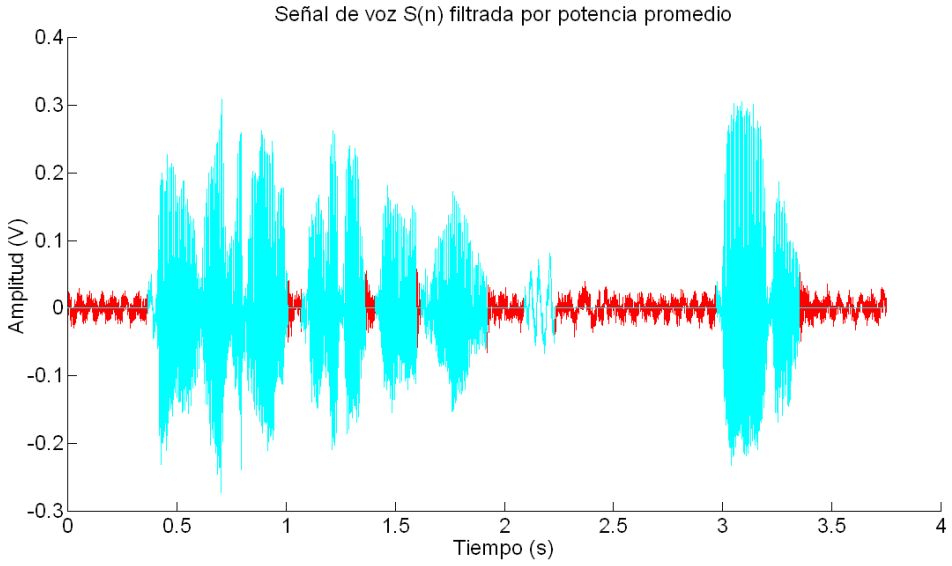

Figura 2.9: Señal con supresión de silencios $S_1(n)_{voz}$ superpuesta a la señal original.

Este procedimiento está basado en la medición de las propiedades de una muestra representativa de ruido, la cual se toma del inicio del registro S_j. La efectividad de la función 2.5, para establecer las pausas o intervalos de silencio de un hablante, está determinada por el valor de umbral u_j, ya que éste discrimina los puntos que deben omitirse de la señal, con base en el comportamiento promedio de la potencia en un segmento. Lo anterior implica que al incrementarse, en promedio, la potencia del nivel de ruido con respecto a la potencia promedio de la voz, la supresión de silencios puede eliminar también segmentos de voz. Un parámetro que permite medir esta relación es la llamada razón señal a ruido (*Signal to Noise Ratio, SNR*) definida en la siguiente ecuación [33]:

$$SNR = 10log\left(\frac{P_{\text{señal}}}{P_{\text{ruido}}}\right), \text{ en } dB. \tag{2.8}$$

En consecuencia es deseable incrementar el valor del SNR en los registros de voz para mejorar la efectividad de la supresión de silencios. Debido a ésto el filtro FIR tiene prioridad respecto de la etapa de supresión en el proceso de extracción de los vectores de características.

2.3.2. Banco de Filtros Mel

La aplicación de la transformada de Fourier discreta a los registros de voz, permite obtener información sobre el contenido frecuencial de la misma por cada ventana o segmento analizado. Sin embargo la cantidad de información obtenida es excesiva y puede complicar la clasificación o separación de vectores. Se da entonces la necesidad de sintetizar la información contenida en el espectro de frecuencias de una forma discreta. La aplicación de un banco de filtros al espectro frecuencial obedece al planteamiento dado previamente.

Como se ha mencionado la elección de un banco de filtros en la escala Mel tiene fundamento en que ésta emula en mejor forma, la resolución con la que el oído humano funciona. Esto significa que enfatiza el contenido en ciertos intervalos de frecuencias, de forma análoga al comportamiento fisiológico del oído humano.

A continuación se describe la conformación y aplicación de este banco de filtros en el procesamiento y extracción de características de la voz [17].

1. Se establece una frecuencia máxima de aplicación del banco de filtros. En este caso $f_{max} = 3300Hz$, debido a que el filtro FIR aplicado en etapas previas, limita el contenido frecuencial de los registros de voz S_j a dicho valor.

2. Con base en f_{max} y la dimensión del banco de filtros, dada por K, se obtiene un intervalo Δ de construcción del banco en una escala lineal de frecuencia:
$$\Delta = \frac{f_{max}}{K+1}. \tag{2.9}$$

3. Dado que la forma de cada filtro es triangular, la escala lineal de frecuencias establece los puntos de inicio, intermedio y final de cada filtro. Dicha escala queda dada por
$$F_{lineal}(i) = (i-1)\Delta, \quad i = 1, \ldots, K+2. \tag{2.10}$$

4. Se aplica la ecuación 2.1 a la escala de frecuencias lineal para obtener la correspondiente en escala Mel.
$$F_{MEL}(i) = 1000\frac{log(1+\frac{F_{lineal}(i)}{1000})}{log(2)}, \quad i = 1, \ldots, K+2. \tag{2.11}$$
Esta escala contiene también los puntos de inicio, intermedio y final de cada filtro triangular.

5. Cada filtro triangular tiene un traslape con el siguiente de manera que el punto intermedio de uno es el inicio del siguiente. El filtro es aplicado al espectro obtenido de cada segmento de voz del registro $S_j(n)_{voz}$; cada filtro se multiplica por los valores del espectro correspondientes y los resultados son promediados, de forma que se obtiene un vector espectral de longitud K como salida del procesamiento.

La figura 2.10 muestra una vista final del banco de filtros en la escala Mel. Hay que resaltar que la aplicación de este filtro se realiza sobre un espectro FFT de 256 puntos de forma que es necesario corresponder los valores de dicha escala con el espectro discreto.

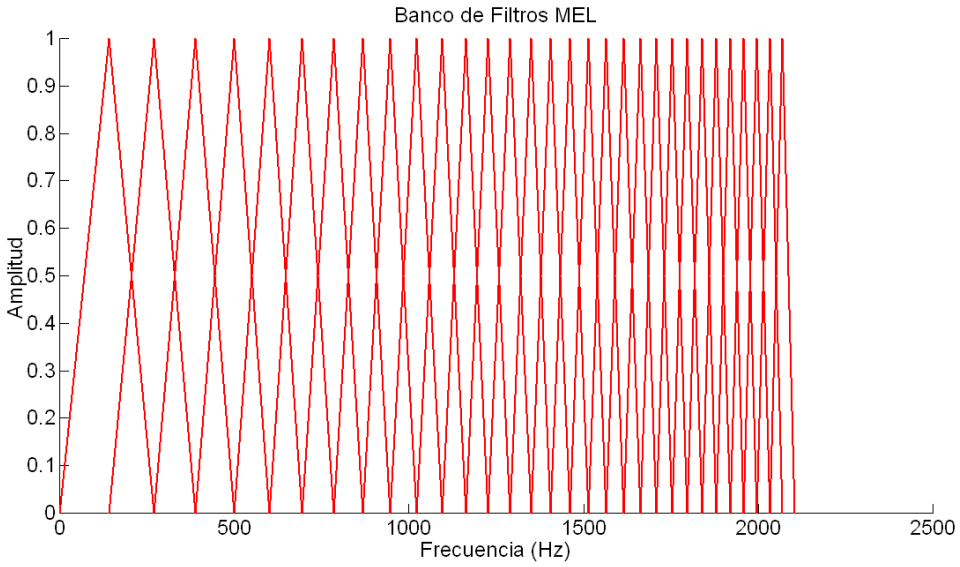

Figura 2.10: Gráfica correspondiente a un banco de 31 filtros en escala Mel.

2.4. Máquinas de Soporte Vectorial

Clasificar involucra necesariamente el concepto de conjunto; de manera general el proceso de clasificación consiste en realizar una separación de los elementos de un conjunto C en diferentes subconjuntos C_i, $i = 1, \ldots, P$, denominados clases, con base en la medición de las características que los elementos de C poseen. Una vez que se determinan las propiedades de los subconjuntos en los que se clasificará al conjunto original (modelos), los elementos de éste son comparados con cada uno de los modelos, para establecer a cual de ellos pertenecen. Matemáticamente este proceso puede entenderse como una función que mapea el conjunto C al conjunto de clases $\{C_i\}_{i=1}^{P}$. Se parte de la hipótesis de que, sin importar la naturaleza del conjunto C, sus elementos pueden ser representados de forma numérica.

Esta representación puede ser en $\mathbb{R}^n$, para algún $n \in \mathbb{N}$. Sin embargo bajo este planteamiento, la labor de clasificación no guarda dificultad alguna, ésta surge, por ejemplo, cuando se considera que los elementos del conjunto C son resultado de un conjunto finito de variables aleatorias en $\mathbb{R}^n$ denotado por

$$C = \{X_1, X_2, \ldots, X_k\},$$

donde X_j es una variable aleatoria discreta infinita o continua. Si consideramos que las clases C_i son una partición de C entonces, dadas las hipótesis, el proceso de clasificación puede no ser exhaustivo, por lo que, de manera práctica, C es formado con subconjuntos de valores representativos de cada una de las variables aleatorias X_j. Así, el objetivo de una Máquina de Soporte Vectorial (*Support Vector Machine*, SVM) consiste en modelar en cierta forma el comportamiento de cada una de las variables aleatorias X_j, de forma que se pueda determinar, dado un vector propuesto, a cual de ellas pertenece.

En particular, para la clasificación de voz puede considerarse, sin pérdida de generalidad, que el conjunto C está formado por dos variables aleatorias, es decir que

$$C = \{X_1, X_2\}.$$

Es posible representar a cada elemento del conjunto C, de la siguiente forma:

$$(x_j, y_j),\ j = 1, \ldots, l,$$

donde $x_j \in \mathbb{R}^n$, $y_j \in \{-1, 1\}$ y l es la cardinalidad de C. Si suponemos que tomamos una muestra representativa de cada una de las variables aleatorias, la representación dada previamente permite establecer al conjunto C de la siguiente forma:

$$\begin{aligned} C = C_1 \cup C_2, \quad C_1 \cap C_2 = \emptyset, \\ C_1 = \{x_1, \ldots, x_k\}, \\ C_2 = \{x_{k+1}, \ldots, x_l\}. \end{aligned}$$

Por las consideraciones hechas, la distribución de los puntos que conforman a C_1 y C_2 es desconocida a priori. Podemos entonces considerar, de forma general, dos casos; cuando C_1 y C_2 son linealmente separables y cuando no lo son [50]. La teoría dada a continuación puede ampliarse consultando [45] y [50].

2.4.1. Conjuntos Separables Linealmente

Se dice que C_1 y C_2 son linealmente separables cuando existe un hiperplano en $\mathbb{R}^n$ determinado por un vector w perpendicular al mismo de forma que

$$w \cdot x + b = 0, \quad b \in \mathbb{R}, \tag{2.12}$$

para cualquier punto x en el hiperplano y además

$$w \cdot x_i + b > 0, \quad \forall x_i \in C_1, \tag{2.13}$$
$$w \cdot x_j + b < 0, \quad \forall x_j \in C_2. \tag{2.14}$$

Puede obtenerse fácilmente que si C_1 y C_2 son linealmente separables entonces la existencia de un hiperplano tal, determinado por un vector w, no es única, de hecho existen una infinidad de tales vectores. Así, es necesario establecer un criterio que permita determinar cual de ellos se tomará para la clasificación. También es inmediato que si ambos conjuntos son separables entonces existe una distancia mínima entre ambos conjuntos. La figura 2.11 presenta un esquema del caso que se expone.

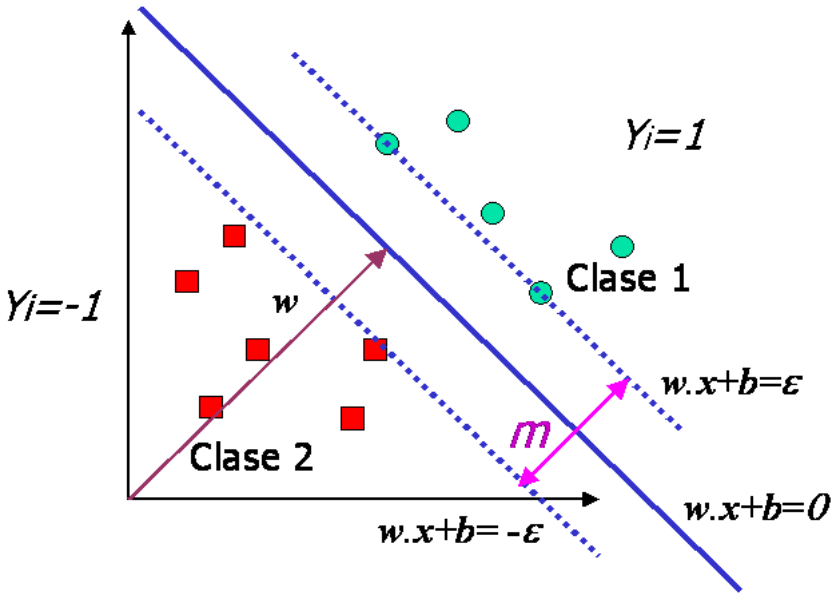

Figura 2.11: Esquema de dos conjuntos de vectores linealmente separados mediante un hiperplano que maximiza el margen m.

De la misma, podemos observar la construcción de dos hiperplanos paralelos al original, determinado éste por w, los cuales delimitan un margen entre los conjuntos y cuya magnitud m podemos relacionar con dichos hiperplanos de la siguiente forma: Supongamos que w está contenido en el hiperplano inferior entonces sucede que

$$w \cdot w + b = -\varepsilon, \tag{2.15}$$

para alguna constante $\varepsilon > 0$. Entonces el vector dado por

$$w + \frac{w}{\| w \|} m, \tag{2.16}$$

está incluido en el hiperplano superior, lo cual implica que

$$w \cdot \left(w + \frac{w}{\| w \|} m \right) + b = \varepsilon. \tag{2.17}$$

Restando las ecuaciones 2.15 y 2.17 y simplificando obtenemos

$$\| w \| m = 2\varepsilon \Rightarrow m = \frac{2\varepsilon}{\| w \|}. \tag{2.18}$$

Dada la igualdad 2.18 se infiere que si deseamos maximizar la magnitud del margen m es necesario minimizar la magnitud de w. Retomando nuevamente la notación (x_j, y_j) con $y_j \in \{1, -1\}$, para cada uno de los vectores x_j de C, entonces el problema de encontrar un hiperplano que separe a C_1 y a C_2, maximizando el margen de separación entre dichos conjuntos, queda planteado como:

$$min\{f(w) = \frac{1}{2\varepsilon} \| w \|^2\} \tag{2.19}$$

$$\begin{aligned} y_i\big(w \cdot x_i + b\big) &\geq \varepsilon, \\ i &= 1, \ldots, l, \\ \varepsilon > 0, b &\in \mathbb{R}. \end{aligned}$$

Este es un problema de optimización de tipo cuadrático sujeto a l restricciones en $\mathbb{R}^n$ y cuya solución puede darse a partir del uso de la teoría de multiplicadores de Lagrange en n variables.

Denotemos las restricciones que aparecen en el problema 2.19 de la siguiente forma:

$$g_i(w) = y_i(w \cdot x_i + b), \quad i = 1, \ldots, l, \tag{2.20}$$

entonces si denotamos por w_k a la k-ésima entrada de w sabemos que se cumple

$$\frac{\partial f}{\partial w_k}(w) = \sum_{i=1}^{l} \alpha_i \frac{\partial g_i}{\partial w_k}(w), \quad k = 1, \ldots, l \tag{2.21}$$

$$\alpha_i g_i(w) = \varepsilon, \quad i = 1, \ldots, l, \tag{2.22}$$

para ciertas constantes $\alpha_i \in \mathbb{R}$ a determinar. Sustituyendo en la ecuación 2.21 a la función f y a las funciones g_i se tiene el siguiente desarrollo:

$$\frac{\partial f}{\partial w_k} = \frac{1}{\varepsilon} w_k = \sum_{i=1}^{l} \alpha_i \frac{\partial g_i}{\partial w_k} = \sum_{i=1}^{l} \alpha_i y_i x_i(k) \Rightarrow$$

$$w = \varepsilon \sum_{i=1}^{l} \alpha_i y_i x_i.$$

Por otro lado de la igualdad 2.22 se tiene que

$$\frac{\partial \sum_{i=1}^{l} \alpha_i g_i(w)}{\partial b} = 0 \Rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0.$$

Por lo tanto la solución al sistema 2.19 está dada por

$$w = \varepsilon \sum_{i=1}^{l} \alpha_i y_i x_i \tag{2.23}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.24}$$

Para obtener el valor de cada una de las constantes α_i sustituimos la ecuación 2.23 en las restricciones iniciales 2.22 con lo que se obtiene la siguiente ecuación:

$$\alpha_j y_j \big(\varepsilon x_j \cdot \sum_{i=1}^{l} \alpha_i y_i x_i + b\big) = \varepsilon, \quad j = 1, \ldots, l. \tag{2.25}$$

Derivando parcialmente cada una de las ecuaciones dadas en 2.25 con respecto a α_j, se obtiene que

$$\alpha_j \parallel x_j \parallel^2 + y_j \sum_{i=1}^{l} \alpha_i y_i x_i \cdot x_j = -\frac{y_j b}{\varepsilon}, \quad j = 1, \ldots, l. \tag{2.26}$$

Las igualdades descritas en 2.26 conforman un sistema lineal de l ecuaciones con incógnitas $\alpha_i, i = 1, \ldots, l$, por lo que se tiene un sistema de la forma $A\alpha = B$, con $A \in \mathbb{R}^{lxl}$ y $B \in \mathbb{R}^l$, siendo α el vector de incógnitas. Este sistema tiene solución única sólo si A tiene inversa. Es importante recordar que el valor l es la cardinalidad del conjunto C, por lo que la complejidad en la obtención de una solución mediante alguna implementación, dependerá del orden de dicho conjunto. Los valores de las constantes α_i quedarán en términos de b y ε, de los cuales uno de ellos puede ser propuesto y el otro determinado, por ejemplo, con la ecuación 2.15.

Así, el vector w dado en la ecuación 2.23 es conocido como el *vector de soporte* del hiperplano que separa a C_1 y a C_2, de donde deriva el nombre de Máquinas de Soporte Vectorial.

2.4.2. Conjuntos No Separables Linealmente

Puede considerarse que el sistema lineal de ecuaciones dado en 2.26, es una conclusión al tratar el caso de dos conjuntos C_1 y C_2, separables linealmente. Sin embargo dicho sistema

puede ser obtenido también para cualesquiera dos conjuntos en $\mathbb{R}^n$. Así, podemos decir que C_1 y C_2 no son separables linealmente cuando la matriz correspondiente al sistema 2.26 no tenga solución, en cuyo caso no es posible la construcción de un hiperplano que satisfaga las condiciones del problema planteado en 2.19.

El tratamiento para el caso de dos conjuntos no separables linealmente consiste en utilizar una función $\phi : \mathbb{R}^n \to \mathbb{R}^m$, con $n, m \in \mathbb{N}, m \geq n$ ó $m = \infty$. Dicha función mapea a los conjuntos $C_1 = \{x_1, \ldots, x_k\}$ y $C_2 = \{x_{k+1}, \ldots, x_l\}$, a un espacio de mayor dimensión, donde podemos denotarlos por $\Gamma_1 = \{\phi(x_1), \ldots, \phi(x_k)\}$ y $\Gamma_2 = \{\phi(x_{k+1}), \ldots, \phi(x_l)\}$ respectivamente. El objetivo de realizar este mapeo es que los conjuntos obtenidos Γ_1 y Γ_2 sean separables linealmente o en su caso se minimice el error mediante la separación con un hiperplano, es decir que el número de vectores clasificados incorrectamente sea mínimo. La figura 2.12 esquematiza la operación ideal de la función ϕ sobre un conjunto dado.

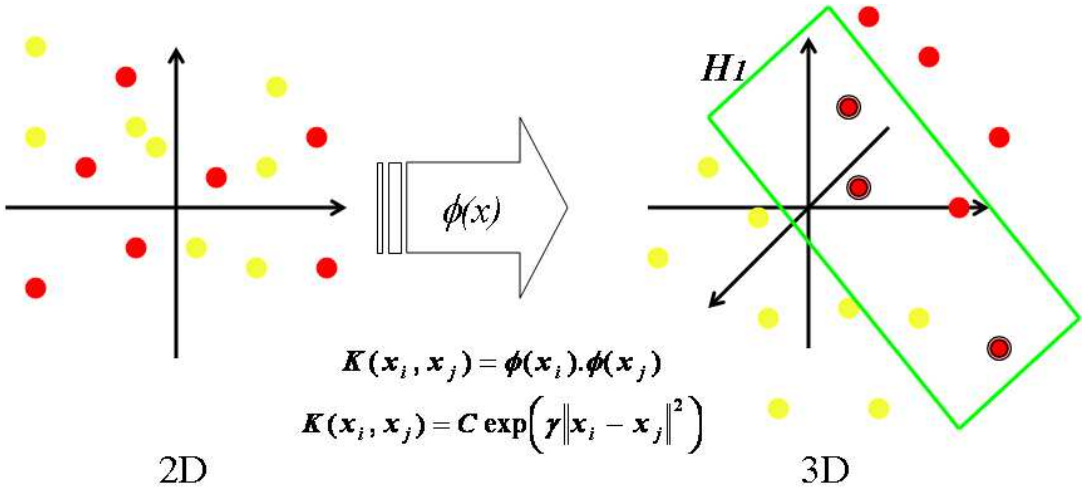

Figura 2.12: Esquema de la transformación de dos conjuntos no separables linealmente mediante la función ϕ.

Como se mencionó anteriormente, idealmente la función ϕ mapea los conjuntos de vectores no separables linealmente a un espacio de mayor dimensión. Éstos vectores son separables linealmente y la solución al problema original mediante el procedimiento explicado en la sección 2.4.1, es un hiperplano $H_1 \subset \mathbb{R}^m$. En este caso se considera que existen vectores de los conjuntos Γ_1 y Γ_2 que se encuentran contenidos en H_1. A dichos vectores se les conoce como vectores de soporte. Para obtener una solución en el espacio $\mathbb{R}^n$ original, se realiza el mapeo inverso de los vectores de soporte, los cuales determinarán las fronteras que separarán a los conjuntos C_1 y C_2. Se considera que los vectores que determinan estas fronteras conforman el modelo para C_1 (o equivalentemente para C_2).

No es posible determinar a priori, dados dos conjuntos de vectores $C_1, C_2 \subset \mathbb{R}^n$, una función ϕ que cumpla los objetivos descritos previamente, por lo que el procedimiento para determinarla no es constructivo. Por tanto el tratamiento para este caso está basado en la realización de ensayos con funciones ϕ conocidas. Así, la función ϕ es de especial importancia en la solución del problema de clasificación.

Del procedimiento dado en 2.4.1, puede observarse que las operaciones con vectores involucran el producto punto o producto interno canónico en $\mathbb{R}^n$. Éste proporciona una función que determina una norma y a su vez una métrica para el espacio:

$$\| x \| = \sqrt{x \cdot x}, \tag{2.27}$$

$$d(x, y) = \| x - y \| . \tag{2.28}$$

Tales norma y métrica, respectivamente, son empleadas también al separar los conjuntos Γ_1 y Γ_2 en $\mathbb{R}^m$. Así que el problema análogo al 2.19, planteado en este nuevo espacio es

$$min\{f(w) = \frac{1}{2\varepsilon} \| w \|^2, \quad w \in \mathbb{R}^m\} \tag{2.29}$$

$$\begin{aligned} y_i\big(w \cdot \phi(x_i) + b\big) &\geq \varepsilon, \\ i &= 1, \ldots, l, \\ \varepsilon > 0,\ b &\in \mathbb{R}. \end{aligned}$$

Por lo anterior la ecuación análoga a 2.26 en este nuevo espacio está dada por

$$\alpha_j \| \phi(x_j) \|^2 + y_j \sum_{i=1}^{l} \alpha_i y_i \phi(x_i) \cdot \phi(x_j) = -\frac{y_j b}{\varepsilon}, \quad j = 1, \ldots, l. \tag{2.30}$$

Se infiere de la ecuación 2.30, que la función ϕ puede ser vista como una que modifica la norma y métrica del espacio original (dadas en 2.27 y 2.28), por las siguientes:

$$\| x \|_\phi = \sqrt{\phi(x) \cdot \phi(x)}, \tag{2.31}$$

$$d_\phi(x, y) = \| x - y \|_\phi . \tag{2.32}$$

A la función definida por

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j), \tag{2.33}$$

se le conoce como función núcleo y su uso es más importante que el de la propia función ϕ, de la cual no se requiere su conocimiento en forma explícita, ya que es suficiente, como lo

muestra la ecuación 2.30, con establecer la función núcleo K para obtener una solución.

Algunos ejemplos de funciones núcleo que han sido sugeridas o empleadas en problemas de clasificación son las siguientes [22]:

Lineal:

$$K(x_i, x_j) = x_i \cdot x_j. \tag{2.34}$$

Polinomial:

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d, \quad \gamma > 0,\ r, d \in \mathbb{R}. \tag{2.35}$$

Función de Base Radial: (*Radial Basis Function, RBF*)

$$K(x_i, x_j) = c\ exp\big(-\gamma \parallel x_i - x_j \parallel^2 \big), \quad \gamma > 0,\ c \in \mathbb{R}. \tag{2.36}$$

Sigmoide:

$$K(x_i, x_j) = tanh(\gamma x_i \cdot x_j + r),\ \gamma, r \in \mathbb{R}. \tag{2.37}$$

Aún cuando existen diferentes funciones núcleo, es común el uso de la Función de Base Radial, por los resultados obtenidos durante la clasificación [9], [22]. Sin embargo puede optarse por el uso de otros núcleos dependiendo de los resultados obtenidos para un caso particular.

Es importante mencionar que dado el origen y la naturaleza de los vectores obtenidos en el proceso de extracción, se espera tratar con un caso de conjuntos no separables linealmente, por lo que en este trabajo se propone utilizar inicialmente la función de base radial (dada en 2.36), para realizar la clasificación.

Capítulo 3

Desempeño del Sistema

En el proceso de verificación de un hablante, es necesaria la conformación de un modelo que corresponda a la voz. Como se mostró en el capítulo 2, el sistema de verificación basado en Máquinas de Soporte Vectorial, requiere la extracción de vectores de características, mismos que son obtenidos mediante el análisis cepstral de los registros de voz. El conjunto total de vectores de cada usuario, puede ser dividido en subconjuntos para la creación de modelos de la voz de los hablantes y también para la realización de pruebas que evalúen al sistema. El subconjunto de vectores utilizado para conformar el modelo de un usuario se denomina conjunto de entrenamiento, éste puede considerarse como la clase o conjunto E_i. Para el proceso de clasificación, la otra clase se forma a partir de vectores del resto de los usuarios, válidos e inválidos. Este conjunto se conoce como *Background* del usuario, que será denotado por B_i. El sistema de clasificación determina entonces los vectores de soporte que separan a ambas clases, minimizando los errores de clasificación.

Una vez creado el modelo del usuario, pueden realizarse pruebas que lo validen, por ejemplo clasificando un conjunto $P_i \neq E_i$, de vectores de características del usuario. El objetivo es determinar el porcentaje de aciertos con un conjunto distinto al de entrenamiento. La cardinalidad del conjunto de prueba puede variarse para observar la efectividad en la clasificación. P_i puede consistir también de vectores de características de usuarios no válidos, lo cual prueba el sistema ante este tipo de solicitudes.

En este capítulo se describen las técnicas empleadas para la formación de conjuntos de entrenamiento y conjuntos de prueba así como los resultados obtenidos de las pruebas. Adicionalmente se presenta un procedimiento para evaluar el desempeño del sistema basado en curvas DET. Se presenta por último la validación de los modelos creados, mediante la realización de pruebas con muestras de voz de los usuarios en tiempo real.

3.1. Entrenamiento y Pruebas

Existen diferentes aspectos que deben considerarse para la formación de conjuntos de entrenamiento y prueba:

Cardinalidad de los conjuntos E_i y B_i. Inicialmente no se tiene un parámetro de referencia para asignar la cantidad de vectores adecuado para formar tales conjuntos o, incluso, si dado el sistema de clasificación usado, existen valores óptimos. Uno de los objetivos de este trabajo consiste en observar el comportamiento del sistema, en relación con los valores de cardinalidad propuestos para estos conjuntos.

Cardinalidad de P_i. El conjunto de prueba más simple consta de un sólo vector, el cual puede corresponder a un usuario válido o a uno inválido. Al incrementar la cardinalidad del conjunto de prueba se puede observar el comportamiento del sistema de verificación y así establecer un valor óptimo: el número mínimo de vectores para la operación aceptable del sistema.

Selección de los vectores de E_i. Como se ha planteado $E_i \subset \{x_{ik}\}_{k=1}^{l}$, sin embargo existen diferentes formas de realizar una selección de los vectores que forman al conjunto E_i. Por ejemplo se puede realizar una selección pseudoaleatoria. También pueden considerarse vectores que correspondan a una misma secuencia de voz del registro S_i. Otra técnica conocida es la de validación cruzada (*Cross Validation*), en la cual se realiza una partición del conjunto total de vectores, en K subconjuntos de igual cardinalidad. Posteriormente se toma cada uno de los K subconjuntos como conjunto de entrenamiento y se obtiene el porcentaje de acierto con los $K-1$ conjuntos restantes. El conjunto de entrenamiento que menor porcentaje de error proporcione, será considerado para conformar el modelo final del usuario.

Selección de los vectores de B_i. Dada la definición de B_i y su uso en la clasificación, la formación de este conjunto requiere que los vectores incluidos sean representativos del resto de los usuarios a U_i. Por ello la selección de éstos puede ser a partir de la misma técnica empleada para formar a cada uno de los E_i. Sin embargo debe realizarse un proceso de selección tal que minimice la cardinalidad de B_i, pues ésta incrementa también la complejidad del problema de clasificación, como se vió en la sección 2.4.2.

A continuación se da una descripción de las características bajo las que se realizó la formación de modelos y las pruebas al sistema de verificación basado en la base *Voces-MCyTI*.

Dado un usuario U_i, el conjunto total de sus vectores de características se denota por $\{x_{ik}\}_{k=1}^{l}$. Éstos son obtenidos mediante el procesamiento de la señal S_i descrito en la sección 2.3. De esta forma se obtienen 17 conjuntos de vectores de características de la misma cardinalidad. Los conjuntos de entrenamiento E_i, del *Background* B_i y de prueba P_i, satisfacen

$$E_i \cup P_i = \{x_{ik}\}_{k=1}^{l},\ E_i \cap P_i = \emptyset \tag{3.1}$$

$$B_i \subset \{x_{1k}\}_{k=1}^{l} \cup, \ldots, \cup \{x_{(i-1)k}\}_{k=1}^{l} \cup \{x_{(i+1)k}\}_{k=1}^{l} \cup, \ldots, \cup \{x_{17k}\}_{k=1}^{l}. \tag{3.2}$$

De esta forma el conjunto $\{x_{ik}\}_{k=1}^{l}$ es dividido para obtener el modelo M_i del usuario, proceso que se conoce como entrenamiento. El resto de los vectores es empleado para realizar pruebas que validen el modelo creado.

La base *Voces-MCyTI* consta de 17 registros, de ellos $S_1, \ldots, S_{15}$ serán considerados como usuarios válidos y los dos restantes como usuarios no incluidos en el sistema. El tiempo de los registros $S_i(n)_{voz}$, luego de la supresión de silencios, se ubicó en un intervalo de 33 a 39 segundos. Con el fin de uniformizar los registros de voz, se consideró procesar sólo los primeros 32.15 segundos de cada uno de ellos. De esta forma se obtuvieron un total de 1606 vectores para cada hablante, es decir $l = 1606$. La forma de selección y formación de los conjuntos E_i, P_i y B_i estuvo basada en los siguientes puntos:

1. Inicialmente se consideró la formación secuencial de los conjuntos E_i y P_i, es decir se formaron conjuntos de vectores asociados a tramas de voz, en el mismo orden en el que aparecen en los registros y con base en la ecuación 3.1. Sin embargo los resultados de clasificación obtenidos mostraron que el sistema se volvía dependiente del texto, al presentar un bajo porcentaje de clasificación para las frases del conjunto P_i.

2. Para evitar la dependencia de los modelos de los hablantes a las frases, se realizó un reordenamiento aleatorio de los vectores, respecto del obtenido originalmente, posterior al procesamiento.

3. La formación de los conjuntos E_i se hizo considerando un 78 % del total de los vectores. Con este porcentaje se obtiene un número de vectores aproximado al de otros problemas de clasificación resueltos mediante Máquinas de Soporte Vectorial [22]. Debido al reordenamiento aleatorio de los vectores, se consideró que el proceso de selección carecía de relevancia, por lo que se seleccionaron los primeros $1,252$ vectores.

4. A fin de evaluar el comportamiento de los modelos generados respecto al número de vectores de prueba, se consideró la formación de dos conjuntos de P_i^1 y P_i^2, con 252 y 102 vectores, respectivamente.

5. La tabla 3.1 sintetiza las características de los conjuntos de entrenamiento y prueba.

   El número de vectores de P_i^2 se asignó con base en el tiempo estimado requerido por el sistema para verificar a un hablante. De esta manera dicho valor corresponde a la señal $S_i(n)_{voz}$, sin intervalos de silencio. De forma real, el tiempo aproximado de grabación sería de 4s.

Conjunto	**No. de vectores**	**Tiempo de $S_j(n)_{voz}$**
E_i	$1,252$	25.05s
P_i^1	252	5.05s
P_i^2	102	2.05s

Tabla 3.1: Características de los conjuntos de entrenamiento y prueba.

6. Para la formación de los conjuntos B_i, se consideró que la cardinalidad de éste debía ser similar a la de los conjuntos E_i, pues al incrementar su valor se observó un incremento en el tiempo requerido por el sistema de clasificación. Por lo anterior y dado un usuario U_i válido, se seleccionaron los primeros 200 vectores de cada conjunto $\{x_{jk}\}_{k=1}^{l}$, $j = 1, \ldots, (i-1), (i+1), \ldots, 15$, sin considerar los registros de los usuarios no incluidos, es decir S_{16} y S_{17}. La selección se hizo sobre los conjuntos una vez aplicado el reorden aleatorio. De esta forma cada conjunto B_i tiene 2800 vectores de características. El objetivo de excluir los registros S_{16} y S_{17} en la formación de los B_i, es para observar, posteriormente, el desempeño del sistema ante peticiones de usuarios no incluidos en forma alguna.

3.1.1. Entrenamiento del Sistema

El entrenamiento del sistema de clasificación se realiza a partir de las parejas (E_i, B_i), lo cual genera un modelo M_i correspondiente al usuario. Para la generación de estos modelos se utilizó el software *Libsvm* [22] basado en Máquinas de Soporte Vectorial. Este software ha sido utilizado para mejorar porcentajes de clasificación en diferentes problemas y cuenta con herramientas que permiten un manejo adecuado de los datos. El procedimiento seguido para el entrenamiento del sistema de verificación se describe a continuación:

1. Se realiza la normalización de los conjuntos de vectores E_i y B_i con base en la siguiente ecuación:
$$N_i(x(n)) = \frac{1}{x^i_{max} - x^i_{min}} \left(2x(n) - (x^i_{max} + x^i_{min}) \right), \tag{3.3}$$
donde x^i_{max} y x^i_{min} son el máximo y mínimo del conjunto dado por $E_i \cup B_i$, considerando cada entrada de los vectores, y $x(n)$ es la n-ésima entrada del vector de características x. La ecuación 3.3 mapea cada conjunto de vectores $\{x_{ik}\}_{k=1}^{l}$ dentro del intervalo $[-1, 1]$. El registro de esta normalización se almacena en un archivo que se denomina rango_i.

2. Se procede al entrenamiento del sistema utilizando los nuevos conjuntos E_i, B_i normalizados. Las características del modelo M_i correspondiente se almacenan en un archivo con extensión *.model*. Éste contiene los datos respecto de las fronteras (vectores de soporte) determinadas por el sistema mediante Máquinas de Soporte Vectorial.

3. El núcleo utilizado para la clasificación es la función de base radial (Ecuación 2.36), cuyos parámetros c y γ pueden ser ajustados para mejorar los porcentajes de acierto en la clasificación. El software *Libsvm* cuenta con una herramienta para mejorar el porcentaje de acierto mediante la búsqueda de valores óptimos para c y γ en una matriz o rejilla de valores, dado un conjunto de vectores de prueba. En las pruebas realizadas con esta herramienta de búsqueda, se observó mejorías poco significativas en los porcentajes de acierto, a coste de un incremento en el tiempo de clasificación del orden de horas. En la sección siguiente se presentan los detalles de estas pruebas. Con base en esto, el entrenamiento se realizó con $c = 32$ y $\gamma = 1$, los cuales mostraron un mejor porcentaje de clasificación, para todos los usuarios.

La figura 3.1 esquematiza el procedimiento de generación de modelos para cada uno de los usuarios.

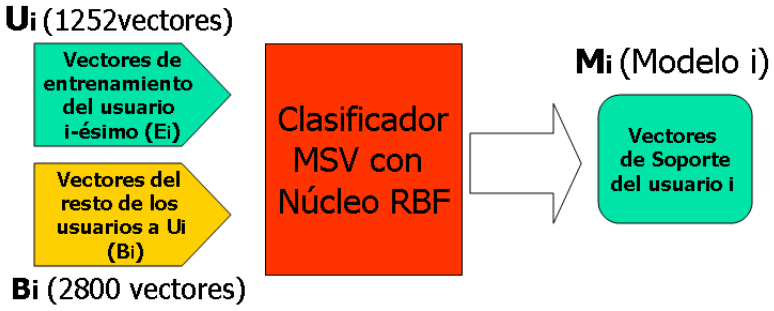

Figura 3.1: Procedimiento de generación de modelos M_i.

Una vez formados los 15 modelos M_i se procedió a realizar las pruebas descritas en la siguiente sección.

3.1.2. Pruebas y Resultados

Las primeras pruebas realizadas tuvieron el objetivo de determinar valores óptimos para los parámetros c y γ de la función de base radial, utilizada como núcleo en la clasificación. En estas pruebas se determina, dada (E_i, B_i), los valores de c_i y γ_i que maximizan el porcentaje de aciertos en la clasificación del conjunto P_i^1, al realizar una búsqueda de éstos en una matriz o rejilla de valores (c_j, γ_k). La búsqueda se inicia a partir de un intervalo dado para cada parámetro, por lo que el resultado obtenido no es necesariamente un óptimo global. Estos valores son utilizados posteriormente para obtener el porcentaje de clasificación con P_i^1 y P_i^2, mismos que fueron normalizados con la ecuación 3.3. Es importante recordar que una vez conformado el modelo de un usuario, el proceso de prueba para obtener un porcentaje de acierto consiste en ingresar al sistema un vector $x_{ik} \in P_i$, el cual es asignado al usuario

U_i, lo que es considerado como un acierto; o a su complemento $\{U_i\}^c$, lo que se considera un error. De esta forma, el porcentaje de acierto se calcula como el cociente del número de vectores clasificados correctamente entre la cardinalidad del conjunto de prueba.

$$Acierto(\,\%) = 100 \times \frac{No\_Aciertos}{No\_Total\_Vectores}. \tag{3.4}$$

La tabla 3.2 presenta los resultados de las pruebas descritas previamente.

Usuario	**Modelo**	c_i	γ_i	**Aciertos** P_i^1 **(%)**	**Aciertos** P_i^2 **(%)**
U_1	M_1	32	1.1	72.222	72.549
U_2	M_2	64	0.8	83.333	69.607
U_3	M_3	64	0.5	71.825	67.647
U_4	M_4	32	0.6	71.031	61.764
U_5	M_5	64	1.1	61.111	64.705
U_6	M_6	96	1.6	57.936	60.784
U_7	M_7	96	2.0	58.333	56.862
U_8	M_8	32	2.1	65.873	71.568
U_9	M_9	16	1.3	68.254	68.627
U_{10}	M_{10}	64	1.4	59.523	53.921
U_{11}	M_{11}	96	0.9	65.079	63.725
U_{12}	M_{12}	64	0.77	72.619	59.803
U_{13}	M_{13}	64	1.97	59.920	52.941
U_{14}	M_{14}	16	1.65	65.873	69.607
U_{15}	M_{15}	16	1.15	76.190	82.352

Tabla 3.2: Valores óptimos de c y γ para cada usuario.

De los resultados obtenidos puede observarse que hay una dependencia entre los valores c_i,γ_i y el conjunto P_i^1, a partir del cual se obtienen. Esto se deduce a partir de los porcentajes de acierto obtenidos con el conjunto P_i^2 que son, en general, menores que los de P_i^1. Esta disminución en el porcentaje de aciertos así como el tiempo requerido para la búsqueda, permiten excluir este procedimiento como parte del sistema de verificación, pues no existe un beneficio significativo en la clasificación para conjuntos de prueba distintos a P_i^1.

Como parte principal de las pruebas realizadas al sistema de verificación, se sometió cada conjunto de prueba $P_i^1,\ i = 1, \ldots, 15$, a ser clasificado utilizando cada uno de los modelos $M_i,\ i = 1, \ldots, 15$. Esta prueba puede entenderse de la siguiente forma:

Clasificación de P_i^1 con M_i. Se clasifica el primer conjunto de vectores de prueba del usuario U_i con su modelo correspondiente. En este caso el sistema trata la petición

de un usuario válido respecto de M_i, por lo que el porcentaje de acierto obtenido mide el desempeño o eficiencia del modelo generado para verificar al usuario.

Clasificación de P_j^1 con M_i, $i \neq j$. Se clasifica el primer conjunto de prueba del usuario U_j, con el modelo de otro usuario U_i distinto. En este caso el sistema trata con una petición de un usuario inválido respecto de M_i, por lo que el porcentaje obtenido mide el error cometido por el sistema ante este tipo de solicitudes.

Los modelos M_i utilizados para estas pruebas, fueron generados con los valores asignados a los parámetros c y γ. Los resultados obtenidos son reportados en las tablas 3.3 y 3.4. Los valores máximos obtenidos son resaltados en "negritas".

.	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
P_1^1	**71.428**	5.952	10.714	15.476	18.650	29.761	30.158	6.746
P_2^1	12.698	**82.142**	12.301	8.333	9.126	25.396	5.952	14.682
P_3^1	12.301	10.317	**70.634**	8.730	11.904	38.888	14.285	26.190
P_4^1	16.269	5.185	25.000	**68.254**	19.444	42.063	13.888	15.873
P_5^1	19.841	9.523	17.857	19.444	**58.333**	25.396	19.047	12.698
P_6^1	17.063	11.507	20.634	29.365	21.825	**55.555**	18.254	5.158
P_7^1	21.031	7.142	19.047	17.063	12.698	42.460	**51.873**	14.285
P_8^1	7.539	5.158	13.095	5.555	11.904	11.904	12.301	**62.301**
P_9^1	8.333	10.317	8.730	12.698	12.301	24.206	17.460	11.904
P_{10}^1	13.492	3.571	11.904	17.857	5.952	10.714	9.523	24.603
P_{11}^1	9.523	18.254	12.698	15.873	18.650	34.523	11.904	7.142
P_{12}^1	18.254	11.111	11.507	7.142	9.523	20.634	13.095	16.666
P_{13}^1	7.539	5.158	14.285	9.920	9.523	21.825	13.888	28.571
P_{14}^1	5.952	9.523	8.333	13.888	5.555	34.920	16.666	28.968
P_{15}^1	11.507	21.428	21.031	5.952	11.507	18.650	12.301	14.682

Tabla 3.3: Porcentajes de clasificación del sistema de verificación de los conjuntos P_j^1 con los modelos M_i.

Como se observa, los mejores porcentajes de clasificación son obtenidos, invariablemente, cuando se verifica el conjunto de prueba del usuario con su modelo respectivo. A pesar de que en algunos casos el porcentaje de error puede ser significativo (vgr. $P_{10}^1,\ M_9$), los porcentajes de clasificación obtenidos permiten siempre determinar de forma clara, a que modelo y usuario corresponde cada conjunto de prueba.

Para observar la estabilidad del sistema se realizaron pruebas con los conjuntos P_i^2, cuya cardinalidad es de 102 vectores (50 % menos que la de los conjuntos P_i^1).

.	M_9	M_{10}	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}
P_1^1	13.889	10.317	21.428	35.714	13.492	3.968	23.809
P_2^1	12.301	9.126	21.031	19.841	5.158	5.158	19.444
P_3^1	6.746	13.492	15.476	12.698	11.904	9.126	17.857
P_4^1	25.793	15.079	12.698	9.523	9.920	10.317	7.936
P_5^1	25.000	13.095	19.841	16.666	19.047	4.365	21.825
P_6^1	34.127	13.888	14.285	12.698	13.492	10.317	7.936
P_7^1	26.587	22.619	13.492	13.095	20.634	11.507	11.507
P_8^1	25.000	20.238	9.126	13.095	33.333	9.523	5.952
P_9^1	**67.857**	27.381	11.111	12.698	26.984	28.571	11.111
P_{10}^1	40.079	**58.333**	13.888	15.873	33.333	9.126	6.746
P_{11}^1	13.888	8.333	**63.095**	17.063	6.349	5.158	18.650
P_{12}^1	20.238	11.507	26.190	**70.634**	12.301	6.746	30.555
P_{13}^1	30.952	28.968	7.936	11.904	**55.158**	28.571	3.968
P_{14}^1	32.936	34.523	7.936	5.952	25.000	**62.698**	5.555
P_{15}^1	10.317	12.301	16.269	32.142	13.492	9.523	**74.603**

Tabla 3.4: (Continuación) Porcentajes de clasificación del sistema de verificación de los conjuntos P_j^1 con los modelos M_i.

Modelo	**Aciertos P_i^1 (%)**	**Aciertos P_i^2 (%)**	**Diferencia**
M_1	71.428	71.568	0.140
M_2	82.142	67.647	-14.495
M_3	70.634	63.725	-6.909
M_4	68.254	64.705	-3.549
M_5	58.330	62.745	4.415
M_6	55.555	53.921	-1.634
M_7	51.873	50.000	-1.873
M_8	62.301	71.568	9.267
M_9	67.857	69.607	1.750
M_{10}	58.333	57.843	-0.490
M_{11}	63.095	61.764	-1.331
M_{12}	70.634	56.862	-13.772
M_{13}	55.158	60.784	5.626
M_{14}	62.698	67.647	4.949
M_{15}	74.603	82.352	7.749

Tabla 3.5: Variabilidad del clasificador con diferentes conjuntos de prueba.

En este caso la clasificación se obtuvo únicamente operando el conjunto P_i^2 con el modelo M_i respectivo, cuyo modelo presenta el mayor porcentaje de clasificación. En la tabla 3.5 se reportan los resultados de esta prueba. Se observa que los porcentajes de acierto disminuyen, al reducir el número de vectores del conjunto de prueba. Sin embargo, aún con esta tendencia, los porcentajes de acierto se mantienen por encima del 50 % y en algunos casos son mayores con el conjunto P_i^2 que con P_i^1, este último de mayor cardinalidad. Estos resultados muestran que el sistema de verificación puede considerarse estable y sin dependencia de las frases pronunciadas. De esta forma los valores asignados a los parametros del núcleo, c y γ, permiten la operación adecuada del sistema. Adicionalmente se determina un tiempo estimado de 4 segundos de las muestras de voz que el sistema requiere para operar en tiempo real.

3.2. Curvas DET

Las curvas de comportamiento del error en la detección de usuarios (*Detection Error Trade-off, DET*), referidas como curvas DET, son un medio que permite observar el comportamiento del error de un sistema de verificación. El análisis se basa en determinar las probabilidades de los dos tipos de error que pueden suceder en el sistema.

Con base en la tabla 1.1, la cual se reproduce a continuación, podemos denotar las dos acciones posibles del sistema como **A** para aceptar y **R** para rechazar.

.	**Acción del Sistema**	
Propuesta	**Acepta**	**Rechaza**
H	**B**	E_1
$\neg H$	E_2	**B**

Tabla 3.6: Definición lógica de los errores posibles en un sistema de verificación.

La hipótesis **H={Muestra de voz de un usuario válido }** y la negación de ésta $\neg H$, conforman un espacio de eventos dado por

$$\Omega = \{(A, H),\ (R, H),\ (A, \neg H),\ (R, \neg H)\}, \tag{3.5}$$

de donde (R, H) y $(A, \neg H)$ son los dos eventos considerados como error, entonces la probabilidad de error total está dada por

$$p(Error) = p(R, H) + p(A, \neg H), \tag{3.6}$$

o de forma equivalente

$$p(Error) = p(R|H)p(H) + p(A|\neg H)p(\neg H). \tag{3.7}$$

A partir de la ecuación 3.7 se crea la función de costo de detección (*Detection Cost Function, DCF*):

$$DCF = w_1 p(R|H)p(H) + w_2 p(A|\neg H)p(\neg H), \tag{3.8}$$

donde w_1 y w_2 son pesos positivos que asignan un costo a cada tipo de error. Las curvas DET se generan a partir de la ecuación 3.8 al fijar w_1, w_2, $p(H)$ y $p(\neg H)$ y graficar $p(R|H)$ respecto de $p(A|\neg H)$. Los valores para estas variables son obtenidos a partir de los porcentajes de clasificación dados en las tablas 3.3 y 3.4. Estos porcentajes son separados en dos conjuntos, uno de ellos incluye los porcentajes del sistema para peticiones de usuarios válidos (P_i, M_i), es decir los que se encuentran en la diagonal. De esta manera el conjunto consta de 15 valores y se denota por *TrueScores*. El segundo conjunto incluye los porcentajes del sistema para peticiones inválidas $(P_j, M_i,\ i \neq j)$, es decir los que no pertenecen a la diagonal. Este conjunto se denota por *FalseScores* e incluye 210 valores. Para mejorar la resolución de las curvas DET se incrementa la cardinalidad de *TrueScores*. Para ello dado $v_i \in$ *TrueScores*, $i = 1, \ldots, 15$, se generán 500 valores con una distribución Normal, con media v_i y varianza σ^2. De esta forma el conjunto final *TrueScores* consta de 7500 valores.

En las figuras 3.2, 3.3 y 3.4 se esquematizan los resultados para diferentes valores de σ^2.

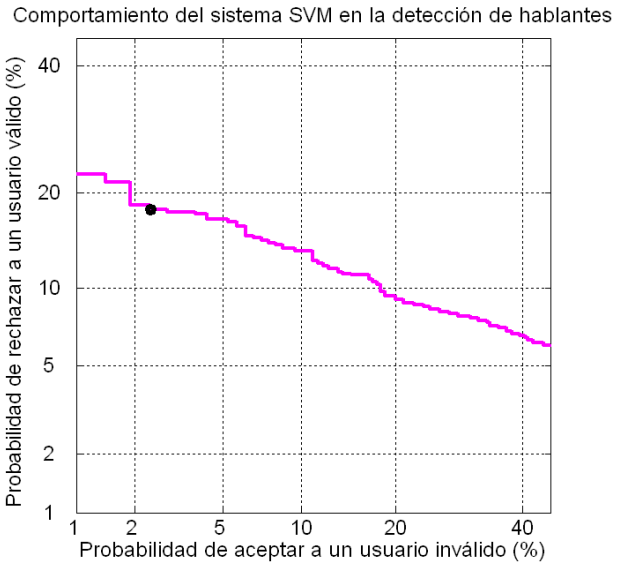

Figura 3.2: Curva DET del sistema de verificación con varianza σ^2 =10.0 en los porcentajes de acierto.

La tabla 3.7 presenta los valores dados a los parámetros de la función DCF para obtener las curvas de desempeño del sistema.

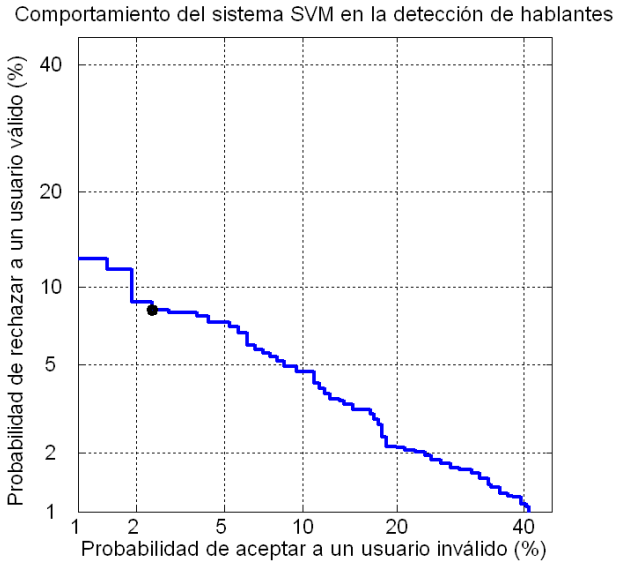

Figura 3.3: Curva DET del sistema de verificación con varianza σ^2 =4.0 en los porcentajes de acierto.

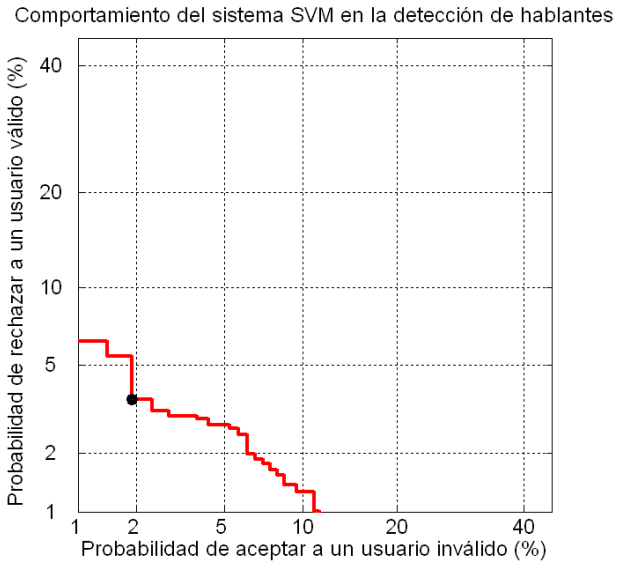

Figura 3.4: Curva DET del sistema de verificación con varianza σ^2 =2.0 en los porcentajes de acierto.

Parámetro	**Valor**	**Observaciones**
$p(H)$	0.015	Se tienen 15 usuarios válidos y un estimado total de 1000 posibles. $p(H) = \frac{15}{1000}$.
$p(\neg H)$	0.985	Se considera un total de 985 usuarios no válidos. $p(\neg H) = 1 - p(H)$.
(w_1, w_2)	(1,10)	Se penaliza en una relación 10 a 1 el aceptar a un usuario no válido.

Tabla 3.7: Parámetros asignados a la función DCF.

Es importante resaltar que los pesos asignados a w_1 y w_2 permiten reflejar condiciones reales de operacion del sistema. En este caso se considera que, por seguridad, es más grave la aceptación de un usuario inválido $(A|\neg H)$, que el rechazo de uno válido $(R|H)$.

Los valores óptimos encontrados por el algoritmo y que minimizan la función de costo DCF, son reportados en la tabla 3.8. Estos puntos también se encuentran graficados en las figuras 3.2, 3.3 y 3.4.

Curva	**Varianza**	$p(R\|H)$ (%)	$p(A\|\neg H)$ (%)	$p(Error)$	**Desempeño (%)**
fig. 3.2	10.0	17.773	2.381	0.0262	97.38
fig. 3.3	4.0	8.200	2.381	0.0247	97.53
fig. 3.4	2.0	3.506	1.905	0.0193	98.07

Tabla 3.8: Valores óptimos de error que minimizan DCF y error y desempeño asociados.

El algoritmo utilizado para obtener las curvas DET, se encuentra especificado en el apéndice A.2. Éste consiste en obtener las distribuciones de *TrueScores* y *FalseScores*. La intersección entre ambas distribuciones constituye el error en la detección. Las escalas de graficación son tales, que una distribución normal aparace como una recta con pendiente -1 y distancia al origen determinada por la media y la varianza. Es posible entonces determinar un valor de error óptimo para el sistema de verificación, mediante la ecuación 3.7, y así también una medida de su desempeño, mediante la siguiente expresión:

$$Desempeño = 100\big(1 - p(Error)\big) = 100\big(1 - p(R|H)p(H) - p(A|\neg H)p(\neg H)\big). \qquad (3.9)$$

Una descripción más detallada al respecto puede encontrarse en [2]. Esta medida nos permite observar que al disminuir la varianza en los porcentajes de acierto en la clasificación, se incrementa el desempeño del sistema. De esta forma se da un procedimiento de evaluación del sistema propuesto.

3.3. Comprobación con Voz en Tiempo Real

Una parte determinante en la validación de un sistema de verificación, consiste en la realización de pruebas con voz en tiempo real. Esto debido a que, en última instancia, el sistema implementado operará con muestras de voz de un usuario, las cuales serán grabadas, procesadas y evaluadas por el sistema, hasta decidir la aceptación o rechazo del solicitante. La diferencia esencial entre las pruebas realizadas en la sección 3.1.2 y las que aquí se muestran, es que los conjuntos $E_i,\ B_i$, y $P_i,\ i = 1, \ldots, 15$, son el resultado de una selección aleatoria de vectores del conjunto total $\{x_{ij}\}_{j=1}^{l}$, mientras que la comprobación en tiempo real utiliza muestras de secuencias de voz, es decir frases pronunciadas en forma secuencial y no necesariamente incluidas en el protocolo de grabación. Adicionalmente, éste tipo de pruebas permite evaluar el comportamiento, ante muestras de voz de hablantes que no cuenten con un modelo dentro del sistema de verificación. De esta forma se comprueba la independencia del sistema de las frases pronunciadas y se valida su operación ante peticiones de usuarios externos no válidos. En las secciones siguientes se presenta una descripción de las pruebas y resultados obtenidos para este fin, así como la evaluación del desempeño respectivo.

3.3.1. Pruebas y Resultados

La base de registros *Voces-MCyTI* se formó con 17 grabaciones de voz de 17 hablantes, de los cuales 15 de ellos fueron utilizados para la formación de modelos y la realización de pruebas, como se describió en la sección 3.2. Para la realización de pruebas, se seleccionó una muestra R_i de cada uno de los 17 registros *mcyti_i.wav*. Las muestras $R_i,\ i = 1, \ldots, 17$, se formaron extrayendo de cada registro, la tarea a) del protocolo de grabación (Sección 2.2.1), es decir el nombre completo del hablante. Dicha extracción se hizo de forma manual, de manera que el tiempo de duración de las muestras es variable, lo cual refleja una condición de operación real. La tabla 3.9 presenta las características de cada muestra R_i.

Una vez extraidas las muestras de voz, se sometieron al procesamiento referido en la sección 2.3 para extraer los vectores de características. Este procesamiento fue realizado en tres módulos: filtrado, supresión de silencios y obtención de vectores cepstrales. Con el objetivo de obtener un tiempo estimado requerido por el sistema para procesar las muestras R_i, se obtuvo el tiempo utilizado en cada uno de los módulos descritos. La tabla 3.10 presenta los tiempos de procesamiento de las muestras y también el número de vectores cepstrales obtenidos luego del procesamiento. Este número de vectores depende, principalmente, de la actividad del supresor de silencios, mismo que reduce la longitud de las muestras.

De esta manera se tiene que la duración promedio de las muestras R_i es de **2**,**695s** y que, en promedio, para procesar dicho periodo de grabación se utilizan **5**,**406s**, esto sin considerar el tiempo utilizado por el clasificador.

Muestra	**No. Puntos**	**Tiempo (s)**
R_1	19,573	2.446
R_2	19,774	2.471
R_3	19,828	2.478
R_4	19,771	2.471
R_5	19,996	2.499
R_6	20,005	2.500
R_7	19,355	2.419
R_8	19,674	2.459
R_9	35,841	4.480
R_{10}	23,840	2.980
R_{11}	19,963	2.495
R_{12}	20,030	2.503
R_{13}	20,761	2.595
R_{14}	19,971	2.496
R_{15}	19,976	2.497
R_{16}	25,142	3.142
R_{17}	23,147	2.893

Tabla 3.9: Características de las muestras de voz R_i.

Muestra	**t FIR (s)**	**t Supresor (s)**	**t MFCC (s)**	**Total (s)**	**Vectores**
R_1	3.594	0.359	0.218	4.171	56
R_2	3.704	0.375	0.250	4.329	77
R_3	3.703	0.344	0.266	4.313	85
R_4	3.688	0.375	0.187	4.250	59
R_5	5.422	0.360	0.266	6.048	85
R_6	4.016	0.375	0.203	4.594	68
R_7	5.484	0.375	0.125	5.984	39
R_8	5.359	0.265	0.094	5.718	29
R_9	6.641	0.437	0.297	7.375	92
R_{10}	4.609	0.438	0.172	5.219	58
R_{11}	5.484	0.266	0.297	6.047	94
R_{12}	3.735	0.218	0.125	4.078	36
R_{13}	4.187	0.391	0.296	4.874	97
R_{14}	5.422	0.375	0.156	5.953	51
R_{15}	5.609	0.453	0.219	6.281	71
R_{16}	4.844	0.469	0.188	5.501	59
R_{17}	6.422	0.500	0.250	7.172	81

Tabla 3.10: Tiempos de procesamiento de las muestras R_i. $c_i = 10$ (Ver ecuación 2.6).

La cantidad de vectores obtenidos varía en el intervalo de 29 a 97. Este intervalo muestra una gran variabilidad y difiere respecto del número de vectores que conforman los conjuntos de prueba (252 y 102 vectores). Se conviene denotar al conjunto de vectores de características asociado a una muestra R_i como V_{Ri}. Con la finalidad de evaluar el comportamiento del sistema con muestras de voz en tiempo real, se procedió a clasificar cada uno de los conjuntos $V_{Ri},\ i = 1, \ldots, 17$, con cada uno de los modelos $M_j,\ j = 1, \ldots, 15$, generados mediante Máquinas de Soporte Vectorial y presentados en la sección 3.1.1, con el objetivo de obtener los porcentajes de clasificación respectivos. Estas pruebas tienen la siguiente interpretación:

Clasificación de V_{Ri} con M_i, $1 \leq i \leq 15$. Se clasifica el conjunto de vectores de la muestra R_i, con el modelo de voz correspondiente al usuario U_i del cual proviene. Esto significa que el sistema trata la petición de un usuario válido respecto de M_i, por lo que el porcentaje de acierto obtenido mide el desempeño o eficiencia del modelo generado para verificar al usuario con voz en tiempo real.

Clasificación de V_{Ri} con M_j, $i \neq j$. Se clasifica el conjunto de vectores de la muestra R_i, con el modelo de voz de otro usuario U_j, distinto. Esto significa que el sistema trata con una petición de un usuario inválido respecto de M_j, por lo que el porcentaje obtenido mide el error cometido por el sistema ante peticiones inválidas en tiempo real. Debido a que en este caso $i = 1, \ldots, 17$ y $j = 1, \ldots, 15$, los conjuntos V_{R16} y V_{R17} no cuentan con un modelo de voz M_i dentro del sistema, lo cual significa que el sistema trata con peticiones de usuarios externos no incluidos.

Los resultados obtenidos de estas pruebas se encuentran reportados en las tablas 3.11 y 3.12.

Los porcentajes obtenidos demuestran que las muestras de voz en tiempo real seleccionadas así como el número de vectores, son suficientes para la correcta operación del sistema, pues en cualquier caso siempre es posible determinar, de forma correcta y con base en el mayor porcentaje, a que modelo pertenecen las muestras de voz. Los porcentajes obtenidos al clasificar los conjuntos V_{R16} y V_{R17} con todos los modelos, muestran que el sistema opera adecuadamente en cuanto al rechazo de usuarios externos no incluidos, pues en ningún caso los porcentajes obtenidos superan el 50 %. De esta manera pueden inferirse dos reglas para la aceptación de un solicitante, que ingrese una muestra de voz R al sistema:

1. **Se selecciona el porcentaje de acierto con máximo valor obtenido, en la clasificación del conjunto de vectores de características asociado, V_R, con cada uno de los modelos $M_j,\ j = 1, \ldots, 15$.**

2. **Se acepta al usuario con muestra de voz R, si el máximo porcentaje obtenido es mayor al 50 %. En caso contrario se rechaza.**

.	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$	$\mathbf{M_4}$	$\mathbf{M_5}$	$\mathbf{M_6}$	$\mathbf{M_7}$	$\mathbf{M_8}$
V_{R1}	**76.785**	1.785	14.285	12.500	12.500	30.357	30.357	12.500
V_{R2}	15.584	**75.324**	28.571	18.181	11.688	36.363	9.090	18.181
V_{R3}	9.411	11.764	**75.294**	9.411	10.588	37.647	15.294	21.176
V_{R4}	10.169	3.389	35.593	**71.186**	32.203	47.457	20.339	5.084
V_{R5}	16.470	8.235	17.647	22.352	**50.588**	28.235	25.882	14.117
V_{R6}	22.058	19.117	19.117	25.000	16.176	**69.117**	19.117	7.352
V_{R7}	17.948	0.000	25.641	23.076	15.384	51.282	**71.794**	23.076
V_{R8}	0.000	3.448	10.344	3.448	6.896	24.137	3.448	**72.413**
V_{R9}	8.695	6.521	9.782	5.434	15.217	26.087	7.608	9.782
V_{R10}	22.413	1.724	3.448	5.172	10.344	0.000	6.896	22.413
V_{R11}	9.574	19.148	17.021	15.957	15.957	32.978	14.893	4.255
V_{R12}	30.555	2.777	11.111	2.777	2.777	11.111	33.333	41.666
V_{R13}	9.278	5.154	15.463	9.278	12.371	26.804	6.185	25.773
V_{R14}	3.921	19.607	7.843	3.921	1.960	33.333	9.803	39.215
V_{R15}	28.169	21.126	9.859	5.633	25.352	15.493	18.309	22.535
V_{R16}	6.779	11.864	8.474	13.559	20.339	32.203	11.864	5.084
V_{R17}	3.703	14.814	19.753	12.345	9.876	32.098	18.518	24.691

Tabla 3.11: Comprobación del sistema SVM con muestras de voz en tiempo real.

.	$\mathbf{M_9}$	$\mathbf{M_{10}}$	$\mathbf{M_{11}}$	$\mathbf{M_{12}}$	$\mathbf{M_{13}}$	$\mathbf{M_{14}}$	$\mathbf{M_{15}}$
V_{R1}	10.714	12.500	37.500	37.500	25.000	5.357	32.142
V_{R2}	20.779	14.285	19.480	18.181	7.792	7.792	16.883
V_{R3}	8.235	12.941	11.764	12.941	12.941	15.294	22.352
V_{R4}	15.254	11.864	3.389	5.084	11.864	23.728	3.389
V_{R5}	31.764	17.647	16.470	20.000	27.058	5.882	18.823
V_{R6}	26.470	16.176	13.235	4.411	17.647	8.823	5.882
V_{R7}	33.333	25.641	20.512	17.948	30.769	20.512	10.256
V_{R8}	17.241	13.793	3.448	13.793	65.517	17.241	0.000
V_{R9}	**85.869**	8.695	6.521	15.217	39.130	34.782	14.130
V_{R10}	41.379	**72.413**	8.620	13.793	31.034	15.517	6.896
V_{R11}	10.638	9.574	**73.404**	13.829	12.766	8.510	11.702
V_{R12}	11.111	5.555	16.666	**77.777**	13.888	5.555	41.666
V_{R13}	35.051	25.773	7.216	19.587	**71.134**	27.835	6.185
V_{R14}	19.607	45.098	11.764	7.843	21.568	**80.392**	11.764
V_{R15}	8.450	18.309	28.169	46.478	15.493	4.225	**73.239**
V_{R16}	5.084	0.000	28.813	16.949	18.644	1.694	25.423
V_{R17}	12.345	16.049	12.345	12.345	44.444	16.049	14.814

Tabla 3.12: (Continuación) Comprobación del sistema SVM con muestras de voz en tiempo real.

Estas dos reglas establecen, de forma simple y con base en los porcentajes de clasificación obtenidos, un criterio de decisión para el sistema de verificación. Por supuesto que dicho criterio puede ser sujeto a comprobaciones más extensas.

3.3.2. Curvas DET con Voz en Tiempo Real

Como se hizo mención en la sección 3.2, las curvas DET son un instrumento que permite observar el comportamiento del error de un sistema, en este caso de verificación, con base en las distribuciones de porcentajes de acierto y error observados. Así los porcentajes de acierto, presentados en la diagonal de las tablas 3.11 y 3.12, conforman una distribución que es almacenada en un archivo llamado *TrueScores*. El resto de los valores de dichas tablas, conforman una distribución que es almacenada en el archivo *FalseScores*, por lo que este último conjunto consta de 230 valores. Nuevamente se mejora la resolución del conjunto *TrueScores* de la siguiente forma: para cada $v_i \in$ *TrueScores*, $i = 1, \ldots, 15$, se genera un conjunto de 500 valores con distribución normal, con media v_i y varianza σ^2. De esta forma el conjunto final *TrueScores* consta de 7500 valores. Las curvas DET son obtenidas a partir la función de costo dada en 3.7 y mediante el algoritmo dado en el apéndice A.2.

A modo de comparación, en las figuras 3.5, 3.6 y 3.7 se presenta el comportamiento del error del sistema SVM, mediante los conjuntos de prueba P_i^1 y mediante los conjuntos de vectores asociados a las muestras R_i, para tres valores diferentes de σ^2.

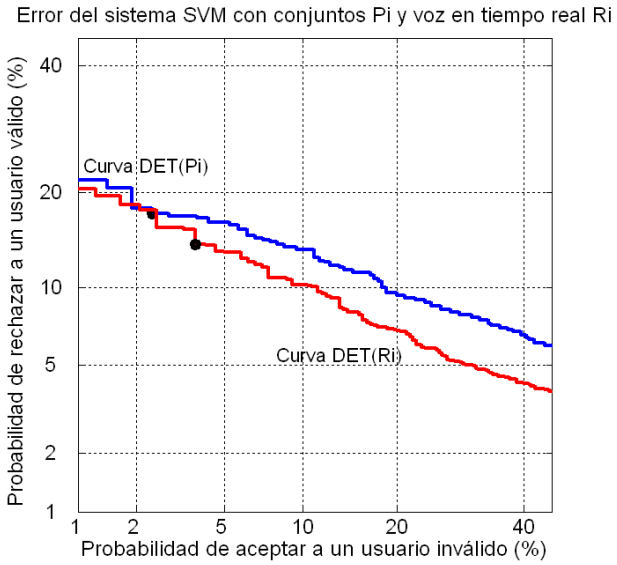

Figura 3.5: Comportamiento del error del sistema SVM con conjuntos de prueba P_i^1 y muestras de voz en tiempo real R_i. σ^2 =10.0.

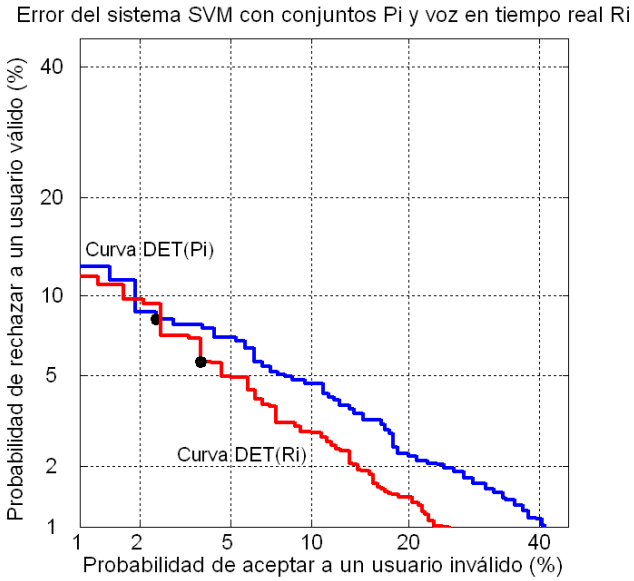

Figura 3.6: Comportamiento del error del sistema SVM con conjuntos de prueba P_i^1 y muestras de voz en tiempo real R_i. σ^2 =4.0.

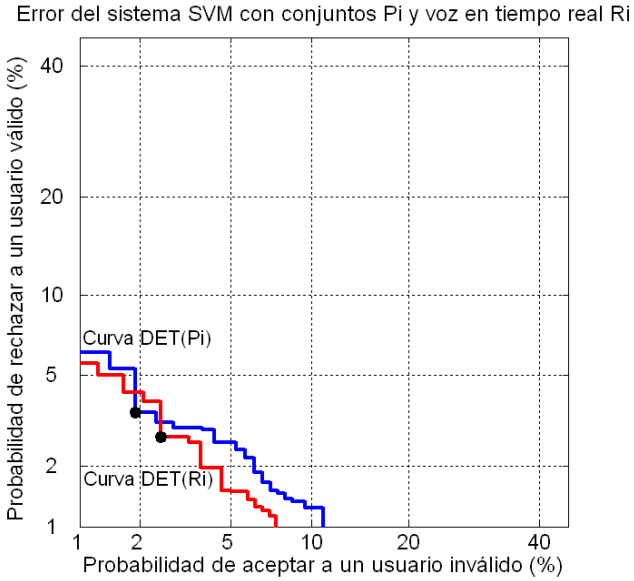

Figura 3.7: Comportamiento del error del sistema SVM con conjuntos de prueba P_i^1 y muestras de voz en tiempo real R_i. σ^2 =2.0.

Los parámetros asignados a la función DCF para la obtención de estas curvas, son los mismos que los reportados en la tabla 3.8. Los puntos óptimos de error encontrados y que minimizan DCF, son presentados en la tabla 3.13, donde también se presenta el desempeño asociado (Ecuación 3.9).

Curva	**Varianza**	$p(R\|H)$ (%)	$p(A\|\neg H)$ (%)	$p(Error)$	**Desempeño (%)**
fig. 3.5	10.0	13.906	3.750	0.0390	96.10
fig. 3.6	4.0	5.680	3.750	0.0378	96.22
fig. 3.7	2.0	2.733	2.500	0.0250	97.50

Tabla 3.13: Valores óptimos de error que minimizan DCF y error y desempeño asociados.

En la tabla 3.14 se comparan los valores de desempeño obtenidos en la tabla 3.8 con los que se reportan en la tabla 3.13.

.	**Desempeño (%)**	
Varianza	$\mathbf{P_i^1}$	$\mathbf{R_i}$
10.0	97.38	96.10
4.0	97.53	96.22
2.0	98.07	97.50

Tabla 3.14: Comparación de desempeño del sistema SVM con conjuntos de prueba P_i^1 y voz en tiempo real R_i.

La tendencia observada en estos resultados y en las gráficas obtenidas, muestra que el desempeño del sistema con las muestras de voz R_i, es menor al presentado con los conjuntos de prueba P_i^1 en 1 %, esto aún cuando las gráficas $p(R|H)$ vs $P(A|\neg H)$ en tiempo real, muestran un mejor comportamiento. Las diferencias en los porcentajes mostrados en la tabla 3.14, permiten afirmar que el sistema SVM implementado es independiente del texto. También se valida que los modelos M_i, pueden operar con muestras de voz en tiempo real o con frases pronunciadas secuencialmente. Adicionalmente, se comprueba que este sistema tiene un comportamiento estable ante peticiones de usuarios que no cuentan con un modelo de voz.

3.4. Resultados Recientes con SVM

En esta sección se presentan resultados de investigación publicados recientemente en dos artículos. En éstos se utilizan las Máquinas de Soporte Vectorial como sistema de clasificación aplicado a tareas de verificación por voz.

En ambos casos son presentados los detalles más relevantes de la implementación de forma sintética.

1. **"Speaker Verification Using Support Vector Machines" [45].**

   Datos: Todas las pruebas realizadas y reportadas están basadas en la base de voz *NIST 2003 Speaker Recognition Evaluation*, en idioma inglés.

   Método de procesamiento: Se utiliza la extracción de vectores mediante análisis MFCC normal o estándar. La dimensión de los vectores es 39. No se hace una descripción detallada del procedimiento de extracción.

   Conjuntos de entrenamiento y prueba: Todas las muestras para formar los conjuntos fueron de aproximadamente 2 minutos de duración. El conjunto de entrenamiento se obtuvo a partir de 60 pronunciaciones, mientras que el conjunto de prueba se obtuvo de 78. Ambos conjuntos se extrajeron de la base *Switchboard*. Para el conjunto de vectores de entrenamiento de un usuario particular, se tomaron todos los vectores obtenidos mediante el análisis cepstral. Para conformar los vectores del resto de los usuarios (*Background*) se eligieron aleatoriamente n vectores de cada conjunto de vectores de entrenamiento del resto de usuarios. El valor de n fue elegido de manera que fuera el doble de vectores obtenido para cada usuario. Esta decisión fue tomada a fin de que la aplicación fuera factible, ya que se sabe que el desempeño de sistemas de clasificación mediante *SVM* es adecuado sólo para conjuntos de entrenamiento pequeños.

   Metodología de clasificación: Se aplica directamente a los vectores de características de la voz, la metodología de las Máquinas de Soporte Vectorial. Esta metodología ha sido expuesta ampliamente por lo que no se requiere una descripción más detallada.

   Resultados: El primer paso en los experimentos fue obtener un valor óptimo para el parámetro γ del núcleo K (Ecuación 2.36). En este caso se obtuvo que no hay variación sustancial en los porcentajes para valores de γ entre 0.02 y 2.5 y que el valor óptimo es de γ =0.019. El valor mínimo que se obtiene de la función DCF es de **0.1406**. Adicionalmente se utilizan Modelos Ocultos de Markov con los que se reporta un valor mínimo de DCF de **0.2124**.

2. **"Support Vector Machines Using GMM Supervectors for Speaker Verification" [54].**

   Datos: La base de datos usada fue *2005 NIST SRE*. Esta base de voz fue grabada en inglés a partir de conversaciones telefónicas. La investigación se enfocó al uso de la grabación de entrenamiento No.8, individual y desde un mismo teléfono y al de

la grabación de prueba No.1, individual y desde un mismo teléfono. En promedio el tiempo de conversacin es de 5 minutos con 2.5 minutos de voz. Este esquema permitió la formación de 1672 muestras asociadas a los usuarios y 14406 muestras utilizadas como *Background*.

Método de procesamiento: Se obtienen los coeficientes MFCC de dimensión 19 de la señal preenfatizada. Se utiliza segmentación en ventanas de tiempo de 20ms y un traslape temporal de 10ms con la ventana anterior. En estos segmentos se aplica una ventana de Hamming. Posteriormente se aplica la Transformada Discreta de Fourier y al espectro resultante un Banco Triangular de Filtros Mel. El espectro de la señal es limitado en un rango de 300Hz a 3140Hz. Los vectores cepstrales son filtrados con *RASTA*. Los coeficientes cepstrales delta son calculados considerando 2 vectores anteriores y dos posteriores al de cada trama y agregados al vector considerado con lo que se forma un vector de dimensión 38. Los vectores obtenidos son sometidos a un detector de energía que descarta vectores con baja intensidad (silencios). Se aplica por último un mapeo de vectores para remover los efectos del canal y normalizar según la media y la varianza a cada vector.

Conjuntos de entrenamiento y prueba: El modelo para el resto de los usuarios, consiste en una Mezcla Gaussiana de 2048 componentes (*Gaussian Mixture Models, GMM*, véase sección 4.1). Para el entrenamiento se adaptó la media con un factor de relevancia de 16. El proceso de entrenamiento se llevó a cabo con el algoritmo *Expectation-Maximization EM*, con datos obtenidos de las siguientes bases: *Switchboard 2 fase 1*, *Switchboard 2 fase 4 (celular)* y *OGI National Cellular*. Se produjeron 2,326 supervectores GMM para formar el *Background* etiquetados con -1. Para el enrolado de usuarios se produjeron 8 supervectores GMM por cada uno, provenientes de 8 conversaciones. El entrenamiento de la Máquina de Soporte Vectorial se realizó con los 8 supervectores de cada usuario y el *Background*.

Metodología de clasificación: El método consiste en la generación de 8 supervectores provenientes de un Modelo de Mezclas de funciones de Gauss, a partir de la media conjunta de los componentes de la mezcla. Este supervector es usado para caracterizar al hablante y al canal mediante el uso de eigenvoces y eigencanales, respectivamente. También se extraen 2,326 supervectores para el *Background*. La Máquina de Soporte Vectorial clasifica a estos supervectores y genera el modelo final para cada hablante. Dos funciones núcleo son empleadas para llevar a cabo la clasificación.

Resultados: Dos de los cuatro esquemas de pruebas presentados utilizaron SVM: Supervetores GMM y clasificación SVM usando un núcleo lineal y supervectores GMM y clasificación SVM usando un núcleo no lineal. Los mejores resultados de clasificación se obtuvieron con el núcleo lineal con un valor óptimo de DCF de **0.0139**.

A modo de síntesis se presentan en la tabla 3.15 los valores de desempeño obtenidos de las aplicaciónes SVM descritas anteriormente.

Sistema	**minDCF**	**Desempeño**
SVM [45]	0.1406	85.94 %
GMM-SVM [54]	0.0139	98.61 %
SVM propuesto	0.0193	98.07 %

Tabla 3.15: Comparación de desempeño entre dos sistemas SVM publicados recientemente (2006) y el propuesto en este trabajo.

Es importante resaltar que el desempeño obtenido en el sistema SVM desarrollado en este trabajo, es muy próximo al obtenido mediante una técnica más elaborada como la que se presenta en [54].

3.5. Validación del Supresor de Silencios

La supresión de silencios se presentó como parte importante del procesamiento de la señal previo a la obtención de los vectores de características (sección 2.3.1). En esta sección se presentan las pruebas y resultados que validan y justifican su aplicación.

Se debe considerar que las pruebas y resultados de desempeño reportadas previamente en este capítulo, fueron realizadas a partir de vectores extraidos de señales sometidas al supresor de silencios. Por ello, para establecer una validación es necesario clasificar conjuntos de vectores provenientes de señales que no hayan sido sometidas a este procesamiento. Por simplicidad se eligieron las muestras de voz R_i presentadas en la sección 3.3, que incluyen el nombre del hablante. De esta forma los resultados que se obtengan pueden ser comparados con los correspondientes. Ello permitirá formular conclusiones sobre los efectos de aplicar el supresor de silencios propuesto.

Las pruebas consistieron en procesar los registros de voz, excluyendo la etapa de supresión de silencios, y clasificar cada conjunto de vectores denotado por Rs_i, con cada uno de los modelos M_i creados en la sección 3.1.1. Los resultados de la clasificación son presentados en la tablas 3.16 y 3.17. Los mayores porcentajes de clasificación son resaltados en "negritas" para cada uno de los conjuntos de prueba.

De los resultados obtenidos en ambas tablas pueden establecerse los siguientes puntos que validan la etapa de supresión de silencios propuesta:

.	$\mathbf{M_1}$	$\mathbf{M_2}$	$\mathbf{M_3}$	$\mathbf{M_4}$	$\mathbf{M_5}$	$\mathbf{M_6}$	$\mathbf{M_7}$	$\mathbf{M_8}$
Rs_1	**60.833**	3.333	26.666	20.000	15.000	27.500	34.166	7.500
Rs_2	12.396	**63.636**	23.966	19.834	11.570	39.669	18.181	18.181
Rs_3	12.295	12.295	**68.032**	13.114	9.016	45.082	17.213	20.491
Rs_4	9.016	5.737	31.967	40.163	28.688	**50.000**	18.032	7.377
Rs_5	22.764	12.195	23.577	22.764	**50.406**	40.650	20.325	11.382
Rs_6	33.333	14.634	31.707	21.951	28.455	**65.040**	14.634	11.382
Rs_7	15.966	11.764	31.092	13.445	17.647	**52.100**	41.176	13.445
Rs_8	11.570	7.438	23.140	9.090	9.090	28.925	17.355	33.057
Rs_9	6.306	16.216	26.126	11.261	30.630	47.747	6.306	4.954
Rs_{10}	18.367	12.244	19.047	10.884	7.482	5.442	13.605	13.605
Rs_{11}	8.130	17.073	21.951	21.138	21.951	35.772	20.325	9.756
Rs_{12}	15.447	7.317	13.008	6.504	14.634	32.520	19.512	13.821
Rs_{13}	4.687	10.937	14.843	10.156	16.406	32.031	14.062	19.531
Rs_{14}	4.065	13.821	8.130	12.195	9.756	39.024	18.699	38.211
Rs_{15}	17.073	17.886	8.943	2.439	12.195	31.703	16.260	26.016
Rs_{16}	8.387	24.516	10.322	18.709	20.645	**29.032**	16.129	3.870
Rs_{17}	4.895	6.293	12.587	27.272	18.881	**39.160**	10.489	16.083

Tabla 3.16: Porcentajes de clasificación mediante SVM con muestras de voz R_i procesadas sin suprimir silencios.

.	**M₉**	**M₁₀**	**M₁₁**	**M₁₂**	**M₁₃**	**M₁₄**	**M₁₅**
Rs_1	25.833	21.666	23.333	24.166	14.166	2.500	24.166
Rs_2	30.578	15.702	17.355	10.743	9.917	8.264	10.743
Rs_3	12.295	18.852	10.655	12.951	11.475	9.836	10.852
Rs_4	25.409	16.393	16.393	4.098	10.655	19.672	8.196
Rs_5	36.585	15.447	16.260	13.821	21.138	7.317	15.447
Rs_6	34.959	10.569	13.821	4.878	8.943	6.504	8.943
Rs_7	31.092	18.487	21.008	15.126	10.084	10.084	12.605
Rs_8	**37.190**	23.966	10.743	14.876	27.272	12.396	9.090
Rs_9	**59.009**	17.567	10.810	20.720	26.126	27.027	12.162
Rs_{10}	47.619	**53.061**	14.966	8.163	27.210	11.564	5.442
Rs_{11}	13.008	15.447	**56.910**	17.886	12.195	7.317	13.008
Rs_{12}	30.081	8.943	27.642	**65.853**	17.073	4.878	35.772
Rs_{13}	31.250	21.093	7.812	16.406	**57.812**	39.843	7.031
Rs_{14}	22.764	34.959	22.764	8.943	24.390	**60.162**	8.943
Rs_{15}	24.390	17.886	17.886	35.772	20.325	16.260	**44.715**
Rs_{16}	14.838	10.967	23.871	15.483	20.645	3.870	26.451
Rs_{17}	16.083	9.090	12.587	6.993	25.174	23.076	11.888

Tabla 3.17: (Continuación) Porcentajes de clasificación mediante SVM con muestras de voz R_i procesadas sin suprimir silencios.

1. Los porcentajes más altos en la clasificación para Rs_4 y Rs_7 son con el modelo M_6.

2. El mejor porcentaje de clasificación para Rs_8 se obtiene con el modelo M_9.

3. Con base en los puntos 1 y 2 los modelos M_4, M_7 y M_8 quedan sin asignación de porcentaje máximo. En una tarea de reconocimiento esto implicaría la aceptación de un usario válido a través de una verificación erronea.

4. Debido a que el mejor porcentaje de acierto para Rs_8 es de 37.190, el porcentaje para rechazar un usuario debería estar por debajo de este nivel. Bajo este criterio el usuario Rs_{17} se asignaría al modelo M_6 lo que implicaría la aceptación de un usario inválido.

5. Bajo los criterios de decisión derivados en la sección 3.3.1 para aceptación o rechazo de hablantes (pag. 62), se hace patente el rechazo de usuarios válidos, como sucedería con Rs_8 y Rs_{15}.

6. Comparando los porcentajes obtenidos con los de las tablas 3.11 y 3.12 correspondientes, se concluye que los valores *TrueScores* se reducen hasta en 39 puntos (Rs_8 y M_8), mientras que los correspondientes a *FalseScores* se incrementan hasta en 11 puntos (Rs_6 y M_1).

7. Se infiere que las distribuciones para *TrueScores* y *FalseScores* son más próximas, es decir se disminuye la resolución entre ambos conjuntos.

De esta forma se valida la etapa de supresión de silencios, pues al omitirla se demuestra el incremento de las probabilidades de error del sistema e incluso su disfuncionalidad. Debido a que se presenta rechazo de usuarios válidos y aceptación de usuarios inválidos (errores de tipo I y II), se infiere un comportamiento insuficiente para implementar un sistema de verificación y la necesidad de contar con una etapa que suprima los silencios de los registros de voz.

Capítulo 4

Otros Sistemas de Clasificación

En este capítulo se presentan dos sistemas de clasificación adicionales, mismos que fueron usados para comparar el desempeño obtenido mediante Máquinas de Soporte Vectorial. Para ello se utilizaron los mismos conjuntos E_i y P_i^1, descritos en el capítulo 3 para la generación de modelos y la realización de pruebas. El procedimiento utilizado para obtener el desempeño de los sistemas propuestos fue expuesto en la sección 3.2. En las secciones siguientes se presenta una descripción breve de cada sistema, así como los resultados de clasificación al aplicarlos a la base de registros *Voces-MCyTI*.

4.1. Modelos de Mezclas Gaussianas

Una mezcla de Gaussianas o mezcla de funciones de Gauss (*Gaussian Mixture Models, GMM*) para un modelo λ_i, es una combinación de funciones de probabilidad denotada por

$$p(x|\lambda_i) = \sum_{j=1}^{M_i} w_{ij} p_{ij}(x), \tag{4.1}$$

donde $x \in \mathbb{R}^n$, p_{ij} es una función de probabilidad de Gauss en $\mathbb{R}^n$, con media $\mu_{ij} \in \mathbb{R}^n$ y matriz de covarianzas $\Sigma_{ij} \in \mathbb{R}^n \times \mathbb{R}^n$. $w_{ij} \in \mathbb{R},\ j = 1 \ldots M_i$, es un conjunto de valores positivos que satisface

$$\sum_{j=1}^{M_i} w_{ij} = 1. \tag{4.2}$$

De esta forma se obtiene que la función 4.1 es también una función de probabilidad. Así, se puede denotar a cada una de las funciones p_{ij} como

$$p_{ij}(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_{ij}|^{1/2}} exp\big(-\frac{1}{2}(x-\mu_{ij})\Sigma_{ij}^{-1}(x-\mu_{ij})^T\big),\ j = 1 \ldots M_i. \tag{4.3}$$

Por lo anterior el modelo λ_i queda determinado por

$$\lambda_i = (\{w_{ij}\}, \{\mu_{ij}\}, \{\Sigma_{ij}\}, M_i). \tag{4.4}$$

La función 4.1 puede ser empleada para implementar un sistema de clasificación para verificación por voz. Si suponemos la existencia de conjuntos de vectores de características para entrenamiento y de *Background*, denotados por E_i y B_i respectivamente, correspondientes a un usuario U_i, entonces es posible generar un modelo λ_i asociado a E_i y un modelo λ_{Bi} asociado a B_i. Cada uno de estos modelos se obtiene ajustando los valores dados en 4.4, de forma que se maximice la función 4.1, para cada uno de los vectores del conjunto considerado. El algoritmo utilizado para este fin es denotado por *EM*, por sus siglas en inglés *Expectation Maximization*. Este algoritmo incrementa de forma monótona el valor de probabilidad de la función 4.1, para cada uno de los vectores de entrenamiento (de E_i ó B_i) en cada iteración [14]. Las ecuaciones del algoritmo *EM* pueden ser encontradas en [13].

Una vez obtenidos los modelos λ_i y λ_{Bi} de un usuario se obtienen también los valores de probabilidad asociados a dichos modelos así como su media y su varianza. Dado un conjunto de prueba P_i, se pueden validar dichos modelos, al determinar la pertenencia de cada $x \in P_i$ al modelo al que muestre una mayor proximidad, en términos de sus parámetros de media y varianza. Dado que se sabe a priori si x proviene de un usuario válido o de uno inválido, se pueden determinar porcentajes de error ante ambos tipos de peticiones.

Es importante precisar que existen condiciones que afectan la representatividad de B_i, principalmente porque B_i proviene de un subconjunto propio del conjunto complemento $\{U_i\}^c$, es decir B_i no representa al total de los hablantes distintos a U_i. Este hecho, esquematizado en la figura 4.1, significa que $x \notin E_i$ no necesariamente implica que $x \in B_i$. De forma análoga $x \notin B_i$ no necesariamente implica que $x \in E_i$ y viceversa.

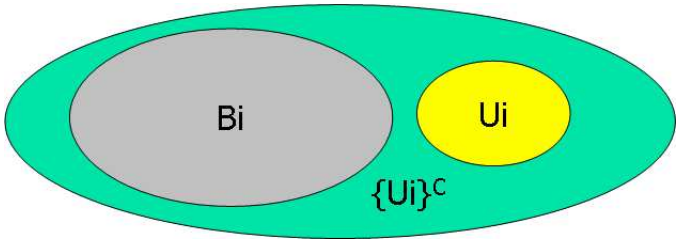

Figura 4.1: Representatividad del *Background* de un usuario respecto del universo de hablantes.

Si denotamos por $\neg\lambda_i$ al modelo del resto de los usuarios a U_i se tiene el siguiente cociente:

$$\frac{p(x|\lambda_i)}{p(x|\neg\lambda_i)}, \tag{4.5}$$

al que se denomina *Cociente de Probabilidades* y es denotado por LR por sus siglas en inglés *Likelihood Ratio*. Debido a que el modelo λ_i se obtiene maximizando $p(x|\lambda_i)$ para un conjunto de vectores $x \in E_i$, la función $p(x|\neg\lambda_i)$ es entonces minimizada. De esta manera el cociente LR toma también un valor máximo para el conjunto E_i, el cual puede puede denotarse por

$$LR_{max}(E_i). \tag{4.6}$$

Mediante el cociente LR y un valor de umbral θ_i es posible determinar si un vector x corresponde al usuario U_i o no y así obtener porcentajes de acierto y error:

$$LR(x) = \frac{p(x|\lambda_i)}{p(x|\neg\lambda_i)} \begin{cases} \geq \theta_i & x \text{ pertenece a } \lambda_i. \\ < \theta_i & x \text{ pertenece a } \neg\lambda_i. \end{cases} \tag{4.7}$$

El establecimiento de un valor de umbral θ_i es una labor experimental que depende de los conjuntos de entrenamiento y prueba utilizados, pues su valor depende directamente de $LR_{max}(E_i)$, es decir que

$$0 \leq \theta_i < LR_{max}(E_i) \tag{4.8}$$

Es común que se utilice el logaritmo natural del cociente LR en lugar de este último [14]:

$$log(LR(x)) = log\left(\frac{p(x|\lambda_i)}{p(x|\neg\lambda_i)}\right) = log\ p(x|\lambda_i) - log\ p(x|\neg\lambda_i), \tag{4.9}$$

el cual se denota por LLR por sus siglas en inglés *Logarithmic Likelihood Ratio*. Así, es común la programación de algoritmos EM utilizando el cociente LLR, por ejemplo el desarrollado por el *IDIAP Research Institute* llamado *Torch* (http://www.idiap.ch). Este software permite obtener, para un conjunto de vectores dado, el valor de $-log(LR_{max})$ referido en 4.6. Para el caso de clasificación se generan dos modelos; λ_i y λ_{B_i} correspondientes a los conjuntos E_i, del usuario U_i, y B_i, que representa el resto de los usuarios a U_i, respectivamente. De esta forma, dado un conjunto de prueba P_i y un vector $x \in P_i$ del cual se conoce su origen, se pueden obtener los siguientes valores:

$$-LLR_{\lambda_i}(x) = -log\left(\frac{p(x|\lambda_i)}{p(x|\neg\lambda_i)}\right), \tag{4.10}$$

$$-LLR_{\lambda_{Bi}}(x) = -log\left(\frac{p(x|\lambda_{Bi})}{p(x|\neg\lambda_{Bi})}\right). \tag{4.11}$$

Para determinar, con base en las ecuaciones 4.10 y 4.11, a cual de los dos modelos pertenece x con mayor probabilidad, es suficiente elegir el modelo cuyo valor de $LLR(x)$ sea mayor. Supongamos por ejemplo que

$$LLR_{\lambda_i}(x) > LLR_{\lambda_{Bi}}(x), \tag{4.12}$$

entonces se da el siguiente desarrollo:

$$log\left(\frac{p(x|\lambda_i)}{p(x|\lambda_i)}\right) > log\left(\frac{p(x|\lambda_{Bi})}{p(x|\neg\lambda_{Bi})}\right) \Rightarrow$$

$$\frac{p(x|\lambda_i)}{p(x|\neg\lambda_i)} > \frac{p(x|\lambda_{Bi})}{p(x|\neg\lambda_{Bi})} \Rightarrow$$

$$p(x|\lambda_i)p(x|\neg\lambda_{Bi}) > p(x|\lambda_{Bi})p(x|\neg\lambda_i). \tag{4.13}$$

La desigualdad 4.13, significa que es más probable que x pertenezca a λ_i y no pertenezca a λ_{Bi}, a que pertenezca a λ_{Bi} y no pertenezca a λ_i. De este silogismo se deduce que x corresponde al usuario U_i con mayor probabilidad y se considera como una petición válida. Al invertir la desigualdad en 4.12 se obtiene, de manera análoga, que x corresponde al complemento del usuario U_i con mayor probabilidad y es por ello considerado como una petición inválida.

De esta forma se establece el uso de Modelos de Mezclas de Gaussianas como un sistema de clasificación y el procedimiento para la validación de los modelos generados.

4.1.1. Pruebas y Resultados

Para las pruebas con GMM se consideraron los mismos conjuntos de vectores E_i, B_i y P_i^1 descritos en el capítulo 3 y que corresponden a cada uno de los 15 usuarios U_i de la base *Voces-MCyTI*. Los conjuntos E_i y B_i fueron utilizados para generar los modelos λ_i del usuario U_i y del *Background* del usuario λ_{Bi}, respectivamente. Los parámetros asignados a *Torch* para la generación de modelos se muestran en la tabla 4.1.

Parámetro	**Valor**
Número de Gaussianas	10
Iteraciones matriz de covarianzas	25
Iteraciones vector de promedios	25
Normalización de datos	No

Tabla 4.1: Parámetros asignados para la generación de modelos con GMM.

Estos parámetros fueron establecidos luego de realizar pruebas para obtener mejores porcentajes de acierto en la clasificación. El número de iteraciones referido en la tabla 4.1 es un valor máximo. Al incrementar estos valores no se observa una mejoría en los porcentajes de acierto. También se observó que la normalización de los conjuntos de vectores no incrementa los porcentajes de acierto, sino al contrario, por lo que dichos conjuntos fueron usados sin normalización.

El procedimiento seguido para evaluar un conjunto de vectores de prueba P_j respecto de los modelos λ_i y λ_{Bi} y obtener porcentajes de acierto, se encuentra esquematizado en la figura 4.2.

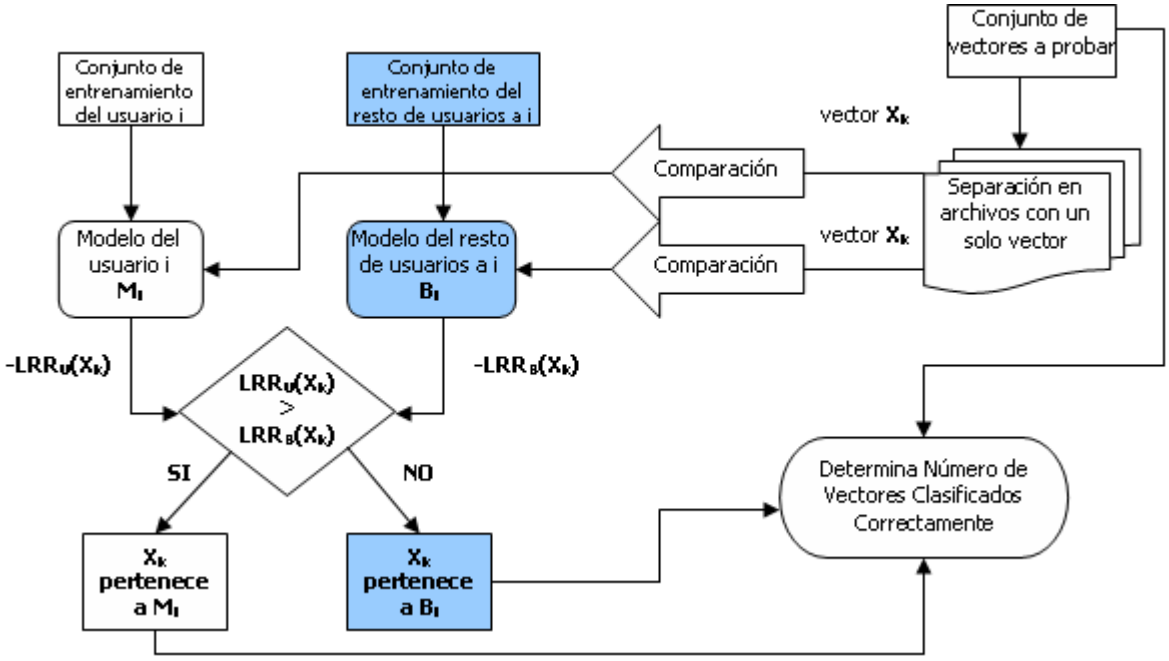

Figura 4.2: Procedimiento para obtención de porcentajes de acierto con GMM.

Con base en este procedimiento se realizaron pruebas al sistema de verificación basado en GMM. Para ello se sometió cada conjunto de prueba $P_i^1,\ i = 1, \ldots, 15$, a ser clasificado utilizando cada pareja de modelos $(\lambda_i, \lambda_{Bi}),\ i = 1, \ldots, 15$. Esta prueba debe entenderse de la siguiente forma:

Clasificación de P_i^1 con $(\lambda_i, \lambda_{Bi})$. Se clasifica el primer conjunto de vectores de prueba del usuario U_i con sus modelos asociados. Esto significa que el sistema trata la petición de un usuario válido respecto de λ_i, por lo que el porcentaje de acierto obtenido mide el desempeño o eficiencia del modelo generado para verificar a usuarios válidos.

Clasificación de P_j^1 con $(\lambda_i, \lambda_{Bi})$, $i \neq j$. Se clasifica el primer conjunto de prueba del usuario U_j, con los modelos de otro usuario U_i distinto. Esto significa que el sistema trata con una petición de un usuario inválido respecto de λ_i, por lo que el porcentaje obtenido mide el error cometido por el sistema ante este tipo de solicitudes.

Esta prueba es equivalente a la realizada con Máquinas de Soporte Vectorial, cuyos resultados están reportados en las tablas 3.3 y 3.4. Esencialmente se realiza una clasificación cruzada entre los conjuntos de prueba y los modelos de cada usuario. Los resultados obtenidos para este caso, se encuentran reportados en las tablas 4.2 y 4.3:

.	$\lambda_{\mathbf{1}}, \lambda_{\mathbf{B1}}$	$\lambda_{\mathbf{2}}, \lambda_{\mathbf{B2}}$	$\lambda_{\mathbf{3}}, \lambda_{\mathbf{B3}}$	$\lambda_{\mathbf{4}}, \lambda_{\mathbf{B4}}$	$\lambda_{\mathbf{5}}, \lambda_{\mathbf{B5}}$	$\lambda_{\mathbf{6}}, \lambda_{\mathbf{B6}}$	$\lambda_{\mathbf{7}}, \lambda_{\mathbf{B7}}$	$\lambda_{\mathbf{8}}, \lambda_{\mathbf{B8}}$
P_1^1	**79.365**	23.015	30.952	20.555	42.460	36.111	40.079	9.523
P_2^1	24.206	**82.936**	17.857	19.047	23.809	35.317	19.047	9.523
P_3^1	26.194	20.238	**78.571**	31.746	40.079	30.555	30.555	19.444
P_4^1	28.174	20.634	40.079	**73.015**	46.428	43.650	36.904	6.746
P_5^1	31.746	23.412	38.888	36.111	**62.698**	41.666	32.142	9.523
P_6^1	28.174	22.222	29.761	42.460	33.333	**71.428**	48.412	9.126
P_7^1	33.730	13.492	29.365	30.158	30.952	42.460	**66.666**	19.444
P_8^1	15.079	8.730	16.666	7.936	9.126	15.079	30.555	**79.365**
P_9^1	17.857	7.142	23.809	28.571	21.428	20.238	28.968	31.746
P_{10}^1	15.079	13.888	23.809	25.000	14.285	22.619	32.539	48.412
P_{11}^1	36.507	35.317	23.412	15.873	42.857	30.555	25.793	7.539
P_{12}^1	26.587	19.047	9.523	13.095	26.984	28.174	27.380	11.507
P_{13}^1	16.269	4.761	16.269	18.253	15.079	19.841	33.730	53.968
P_{14}^1	12.301	8.333	26.190	27.380	17.857	21.825	30.952	48.412
P_{15}^1	22.619	23.412	22.222	16.269	26.190	21.428	23.809	11.507

Tabla 4.2: Porcentajes del sistema de verificación con GMM de los conjuntos P_j^1 con los modelos $(\lambda_i, \lambda_{Bi})$.

Como puede observarse, los mejores porcentajes de clasificación se obtienen cuando se compara el conjunto de prueba con el modelo del usuario respectivo. Esto comprueba que es posible implementar un sistema de verificación basado en GMM como clasificador. En la siguiente sección se utilizan los porcentajes obtenidos para evaluar el desempeño de este sistema.

.	$\lambda_{\mathbf{9}}, \lambda_{\mathbf{B9}}$	$\lambda_{\mathbf{10}}, \lambda_{\mathbf{B10}}$	$\lambda_{\mathbf{11}}, \lambda_{\mathbf{B11}}$	$\lambda_{\mathbf{12}}, \lambda_{\mathbf{B12}}$	$\lambda_{\mathbf{13}}, \lambda_{\mathbf{B13}}$	$\lambda_{\mathbf{14}}, \lambda_{\mathbf{B14}}$	$\lambda_{\mathbf{15}}, \lambda_{\mathbf{B15}}$
P_1^1	16.67	13.095	42.857	34.920	11.904	12.698	25.000
P_2^1	8.730	11.507	45.634	33.333	12.301	13.095	36.904
P_3^1	20.634	24.206	21.825	15.873	14.682	21.031	21.031
P_4^1	26.587	23.412	23.412	19.047	11.111	22.619	11.507
P_5^1	21.031	17.063	36.111	25.793	12.698	14.285	27.777
P_6^1	18.253	18.650	25.000	27.380	15.079	24.603	16.269
P_7^1	20.634	26.587	19.444	28.571	30.555	28.571	14.682
P_8^1	25.793	47.222	6.349	15.873	54.761	42.857	6.349
P_9^1	**79.365**	53.968	10.714	17.857	48.412	47.619	11.904
P_{10}^1	53.968	**78.174**	10.317	20.238	51.984	52.777	7.539
P_{11}^1	10.714	13.095	**73.809**	43.650	10.714	7.539	32.936
P_{12}^1	14.682	18.650	48.015	**79.190**	20.238	10.317	40.873
P_{13}^1	44.047	54.365	12.301	20.238	**70.238**	49.603	10.317
P_{14}^1	48.412	60.714	9.920	9.920	55.158	**71.825**	5.555
P_{15}^1	9.126	9.920	38.492	44.841	15.079	11.507	**78.174**

Tabla 4.3: (Continuación) Porcentajes del sistema de verificación con GMM de los conjuntos P_j^1 con los modelos $(\lambda_i, \lambda_{Bi})$.

4.1.2. Curvas DET del Sistema GMM

Con el objetivo de realizar una comparación del desempeño de los sistemas propuestos, se utilizará el procedimiento descrito en la sección 3.2, así como los parámetros que allí se asignaron. De esta forma será posible, para ciertos valores significativos de la función DCF referida en 3.8, determinar qué sistema de clasificación presenta un mejor desempeño para la base de registros *Voces-MCyTI*.

Como referencia se reproduce la tabla 3.7 donde se asignan los valores de la función de costo de detección:

Parámetro	**Valor**	**Observaciones**
$p(H)$	0.015	Se tienen 15 usuarios válidos y un estimado total de 1000 posibles. $p(H) = \frac{15}{1000}$.
$p(\neg H)$	0.985	Se considera un total de 985 usuarios no válidos. $p(\neg H) = 1 - p(H)$.
(w_1, w_2)	(1,10)	Se penaliza en una relación 10 a 1 el aceptar a un usuario no válido.

Tabla 4.4: Parámetros asignados a la función DCF.

Es pertinente recordar que la curvas DET se forman utilizando los porcentajes de acierto que se encuentran en la diagonal de las tablas 4.2 y 4.3, considerados como una distribución de aciertos del sistema. Los valores de dichas tablas que no se encuentran en la diagonal son considerados como una distribución de errores del sistema, pues provienen de peticiones inválidas. Estas puntuaciones son almacenadas en conjuntos denominados *TrueScores* y *FalseScores*, respectivamente, a partir de los cuales y mediante el algoritmo del apéndice A.2, se obtienen las curvas DET. Al igual que en la evaluación de Máquinas de Soporte Vectorial, se mejoró la resolución de dichas curvas al incrementar la cardinalidad de *TrueScores*; para ello dado $v_i \in$ *TrueScores*, $i = 1, \ldots, 15$, se generaron 500 valores con una distribución normal con media v_i y varianza σ^2, por lo que la cardinalidad final de *TrueScores* es de 7500.

En las figuras 4.3, 4.4 y 4.5 se esquematizan las curvas DET obtenidas para valores de σ^2 de 10.0, 4.0 y 2.0 respectivamente.

La tabla 4.5 resume las medidas de desempeño obtenidas mediante los puntos óptimos de error de las curvas DET (Ver ecuaciones 3.7 y 3.9).

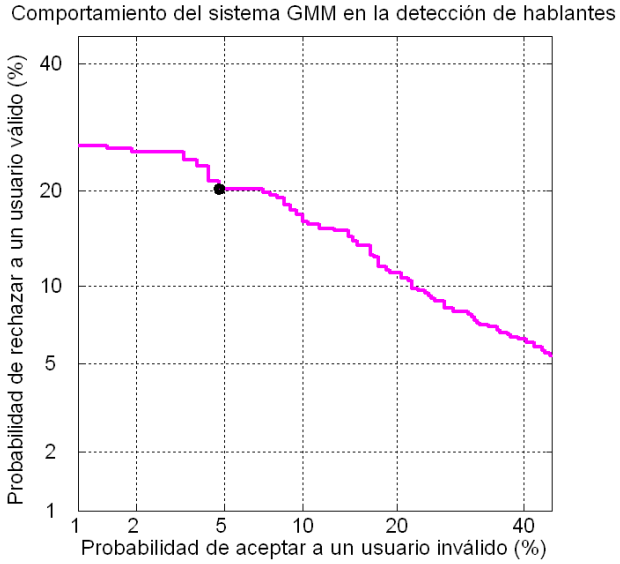

Figura 4.3: Curva DET del sistema de verificación GMM con varianza σ^2 =10.0 en los porcentajes de acierto.

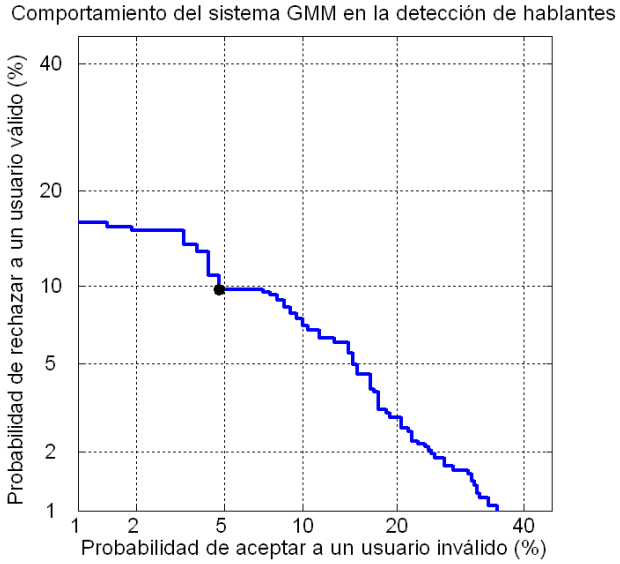

Figura 4.4: Curva DET del sistema de verificación GMM con varianza σ^2 =4.0 en los porcentajes de acierto.

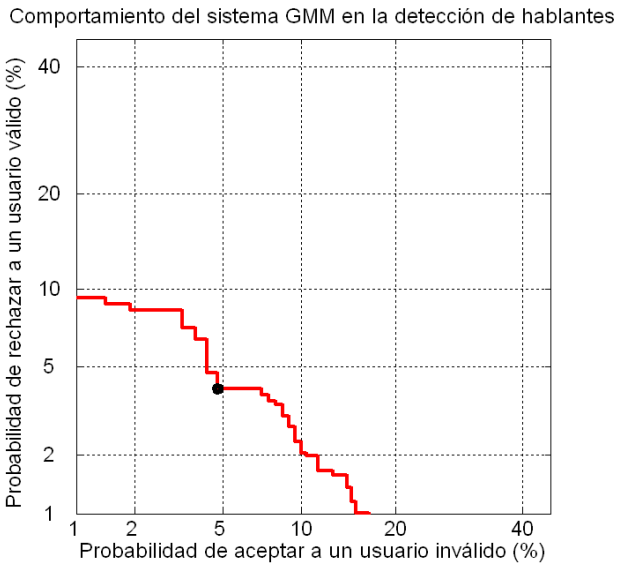

Figura 4.5: Curva DET del sistema de verificación GMM con varianza σ^2 =2.0 en los porcentajes de acierto.

Curva	**Varianza**	$p(R\|H)$ (%)	$p(A\|\neg H)$ (%)	$p(Error)$	**Desempeño (%)**
fig. 4.3	10.0	20.413	4.762	0.0499	95.01
fig. 4.4	4.0	9.573	4.762	0.0484	95.16
fig. 4.5	2.0	4.173	4.762	0.0475	95.25

Tabla 4.5: Valores óptimos de error que minimizan DCF y error y desempeño asociados para el sistema GMM.

Comparando los resultados de desempeño de la tabla 4.5 con los de la tabla 3.8, puede observarse que la clasificación mediante Máquinas de Soporte Vectorial proporciona, en cualquier caso, un mejor comportamiento que el obtenido mediante modelos de mezclas gaussianas. Sin embargo esta diferencia es, en el mejor de los casos, de tres puntos porcentuales, lo que no representa una diferencia determinante. También puede resaltarse que mientras en el sistema SVM se reducen las probabilidades de error óptimas $p(R|H)$ y $p(A|\neg H)$, al reducir la varianza σ^2, en el sistema GMM se mantiene constante $p(A|\neg H)$ y sólo se reduce $p(R|H)$. Esto significa que el comportamiento de ambos sistemas no es similar, lo cual se puede comprobar comparando punto a punto las curvas DET.

4.2. Redes Neuronales Artificiales

Una Red Neuronal Artificial (*Artificial Neural Networks, ANN*) es una estructura que puede ser representada matemáticamente [4]. En su forma más simple consta de sólo una unidad la cual cuenta con un conjunto de N entradas $x(i) \in \mathbb{R}, i = 1, \ldots, N$, y una salida $O \in \mathbb{R}$. La unidad de la red multiplica cada entrada $x(i)$ por un peso asociado $w_i \in \mathbb{R}$ y suma todos los productos, de forma que se obtiene una combinación lineal:

$$\sum_{i=1}^{N} w_i x(i). \tag{4.14}$$

La salida de la unidad está determinada por una función $f : D \subset \mathbb{R} \to \mathbb{R}$ de manera que

$$O = f\left(\sum_{i=1}^{N} w_i x(i)\right). \tag{4.15}$$

En la figura 4.6 se presenta un esquema de una red neuronal simple con una sola unidad. Observemos que el conjunto de entradas $x(i)$ puede ser representado como un vector $x \in \mathbb{R}^N$. De manera análoga el conjunto de pesos w_i puede ser representado por un vector $w \in \mathbb{R}^N$.

De esta manera y con base en la notación del producto interno canónico, la ecuación 4.15 queda como

$$O = f(w \cdot x) \tag{4.16}$$

Supongamos la existencia de un conjunto de vectores $\{x_j\}_{j=1}^{l} \subset \mathbb{R}^N$ y un conjunto de valores $\{t_j\}_{j=1}^{l} \subset \mathbb{R}$, de forma que cada vector x_j está asociado con el valor t_j respectivo. Con esta estructura, se desea encontrar un vector de pesos $w = (w_1, \ldots, w_N) \in \mathbb{R}^N$ y una función f de forma que

$$f(w \cdot x_j) = t_j, \ \forall j = 1, \ldots, l. \tag{4.17}$$

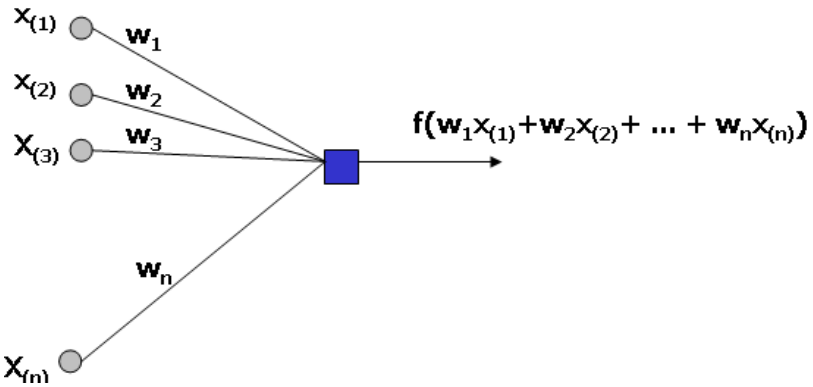

Figura 4.6: Esquema de una red neuronal con una sóla unidad de activación.

Dada una función propuesta f, al procedimiento para encontrar un vector de pesos w que satisfaga 4.17, se le denomina entrenamiento o aprendizaje de la red. A priori, la existencia de tal vector y función no está determinada, por lo que existen diferentes propuestas de funciones f, así como procedimientos para observar si es posible satisfacer 4.17.

Un procedimiento consiste en minimizar el error cuadrático medio (*Mean Squared Error*, MSE) dado por

$$MSE = \frac{1}{2}\sum_{j=1}^{l}\big(f(w \cdot x_j) - t_j\big)^2. \tag{4.18}$$

Derivando parcialmente la ecuación 4.18 se obtiene que

$$\frac{\partial MSE}{\partial w_k} = \frac{1}{2}\sum_{j=1}^{l} 2\big(f(w \cdot x_j) - t_j\big)\frac{\partial f(w \cdot x_j)}{\partial w \cdot x_j}\frac{\partial w \cdot x_j}{\partial w_k}$$

$$\frac{\partial MSE}{\partial w_k} = \sum_{j=1}^{l}\big(f(w \cdot x_j) - t_j\big)\frac{\partial f(w \cdot x_j)}{\partial w \cdot x_j}x_j(k). \tag{4.19}$$

En particular si la función f es la identidad, entonces la ecuación 4.19 toma la forma siguiente:

$$\frac{\partial MSE}{\partial w_k} = \sum_{j=1}^{l}\big(w \cdot x_j - t_j\big)x_j(k),\ k = 1, \ldots, N. \tag{4.20}$$

Igualando a cero las ecuaciones 4.20 se obtiene que

$$\sum_{j=1}^{l} w \cdot x_j x_j(k) = \sum_{j=1}^{l} t_j x_j(k),\ k = 1, \ldots, N,$$

$$w_1 \sum_{j=1}^{l} x_j(1)x_j(k) + \ldots + w_N \sum_{j=1}^{l} x_j(N)x_j(k) = \sum_{j=1}^{l} t_j x_j(k),\ k = 1, \ldots, N. \tag{4.21}$$

Las ecuaciones dadas en 4.21 conforman un sistema lineal de la forma $Aw = B$ en las variables $w_k,\ k = 1, \ldots, N$, con $A \in \mathbb{R}^N \times \mathbb{R}^N$ y $B \in \mathbb{R}^N$. Entonces el sistema tiene solución si la matriz A tiene inversa, en cuyo caso se garantiza la existencia del vector w que minimiza el error 4.18.

Es importante considerar que la ecuación 4.19 hace la suposición restrictiva de que la función f es derivable lo cual, en general, no se cumple. Un método que permite la obtención de un vector w sin hacer esta suposición, es el denominado Propagación Inversa (*Back Propagation*). Este método propone de forma aleatoria, un vector inicial w^0 cuyas entradas se acotan en un intervalo dado, posteriormente se realizan iteraciones de acuerdo a la siguiente regla de recurrencia, para cada una de las entradas w_k^0 del vector:

$$w_k^{m+1} = w_k^m + \Delta w_k^m, \tag{4.22}$$

donde

$$\Delta w_k^m = \eta \sum_{j=1}^{l} \big(f(w^m \cdot x_j) - t_j\big) x_j(k),\ k = 1, \ldots, N. \tag{4.23}$$

El parámetro η es llamado factor de aprendizaje y puede ser ajustado en función de los resultados experimentales obtenidos. Los términos Δw_k^m dados en 4.23, son similares a los elementos del gradiente del error dados en 4.19, por lo que se tiene que

$$w^{m+1} \approx w^m + \eta \nabla MSE. \tag{4.24}$$

Puesto que el gradiente de una función se dirije hacia su punto de máximo crecimiento o decremento, dado un vector inicial w^0 y un valor de η adecuados, se puede alcanzar un mínimo de MSE. Para establecer un criterio de paro para el algoritmo, se establece un valor mínimo de variación d_{min} entre los vectores w^m, en cada iteración:

$$d_{min} = \|w^{m+1} - w^m\|. \tag{4.25}$$

En general no es posible garantizar la convergencia del método ni que ésta sea a un óptimo global, pues esto depende de los valores iniciales. Sin embargo los resultados experimentales pueden validar dicho procedimiento.

Cuando la función f es derivable, el término Δw_k^m en 4.22 se sustituye por

$$\Delta w_k^m = \eta \sum_{j=1}^{l} \big(f(w^m \cdot x_j) - t_j\big) f'(w^m \cdot x_j) x_j(k),\ k = 1, \ldots, N, \tag{4.26}$$

en cuyo caso se da la igualdad en 4.24. Puede observarse que el factor de aprendizaje η realiza, en cualquier caso, un escalamiento del gradiente para ampliarlo o reducirlo. Para garantizar una aproximación adecuada a un óptimo se sugiere que $0 < \eta < 1$ [4]. La generalización de este desarrollo para redes neuronales multicapa es análoga.

Para utilizar redes neuronales artificiales como sistema de verificación para un usuario U_i, se consideran los conjuntos de vectores de características E_i, B_i y P_i asociados. E_i y B_i son utilizados para el entrenamiento de la red neuronal y P_i para la validación de dicha estructura. En este caso sabemos que E_i es un conjunto de vectores que corresponde a la voz del usuario, mientras que B_i corresponde al resto o complemento de usuarios a U_i. Se desea por tanto una red neuronal ANN_i cuya salida O_i satisfaga lo siguiente:

$$O_i(x) = \begin{cases} 1 & \text{si } x \in E_i \\ 0 & \text{si } x \in B_i \end{cases} \tag{4.27}$$

Una vez que se obtiene una red ANN_i que satisface 4.27, para alguna función f_i y valor η_i, se considera que tal red representa el modelo de la voz del usuario U_i. Para validar dicho modelo se utiliza el conjunto de prueba P_i, el cual se ingresa a la red neuronal ANN_i y se obtienen porcentajes de acierto y error ante peticiones válidas e inválidas.

4.2.1. Entrenamiento, Pruebas y Resultados

Las pruebas se realizaron utilizando el simulador de redes neuronales, basado en Java, desarrollado por la Universidad de Tübingen, en Alemania, llamado *Neural Network Simulator* y denotado por *JavaNNS* (http://www-ra.informatik.uni.tuebingen.de). Este software permite, entre otras, la creación de redes neuronales multicapa y provee de herramientas para su aprovechamiento en un ambiente gráfico.

La construcción de cada red neuronal ANN_i, $i = 1, \ldots, 15$, se basó en las características dadas en la tabla 4.6. Estos valores fueron asignados luego de realizar pruebas con diferentes estructuras, principalmente al variar el número de unidades ocultas de la red y la función de activación f. La figura 4.7 muestra una vista de la estructura final de cada una de estas redes.

Cada red neuronal ANN_i fue entrenada utilizando los conjuntos E_i y B_i respectivos estableciendo el patrón de salida dado en 4.27. Las características más relevantes de este procedimiento se encuentran especificadas en la tabla 4.7.

Parámetro	**Valor**
Unidades de entrada	13
Unidades de salida	1
Unidades ocultas	4
Valores de activación	Lógicos {0,1}
Función de salida (f)	Identidad
Conexiones de red	Feed-Forward

Tabla 4.6: Características estructurales de las redes nueuronales ANN_i.

Parámetro	**Valor**	**Excepciones**	**Valor**
η	0.2	ANN_{13}	0.5
d_{min}	0.1	ANN_{13}	0.35
Método de aprendizaje	Backpropagation		
Iteraciones	10,000	ANN_6, ANN_7 y ANN_{10}	20,000
Iniciación de w_k	Aleatoria en $[-1, 1]$		
Normalización conjuntos	En el rango $[-1, 1]$		
Máximo error	650		

Tabla 4.7: Valores de los principales parámetros de entrenamiento de las redes ANN_i.

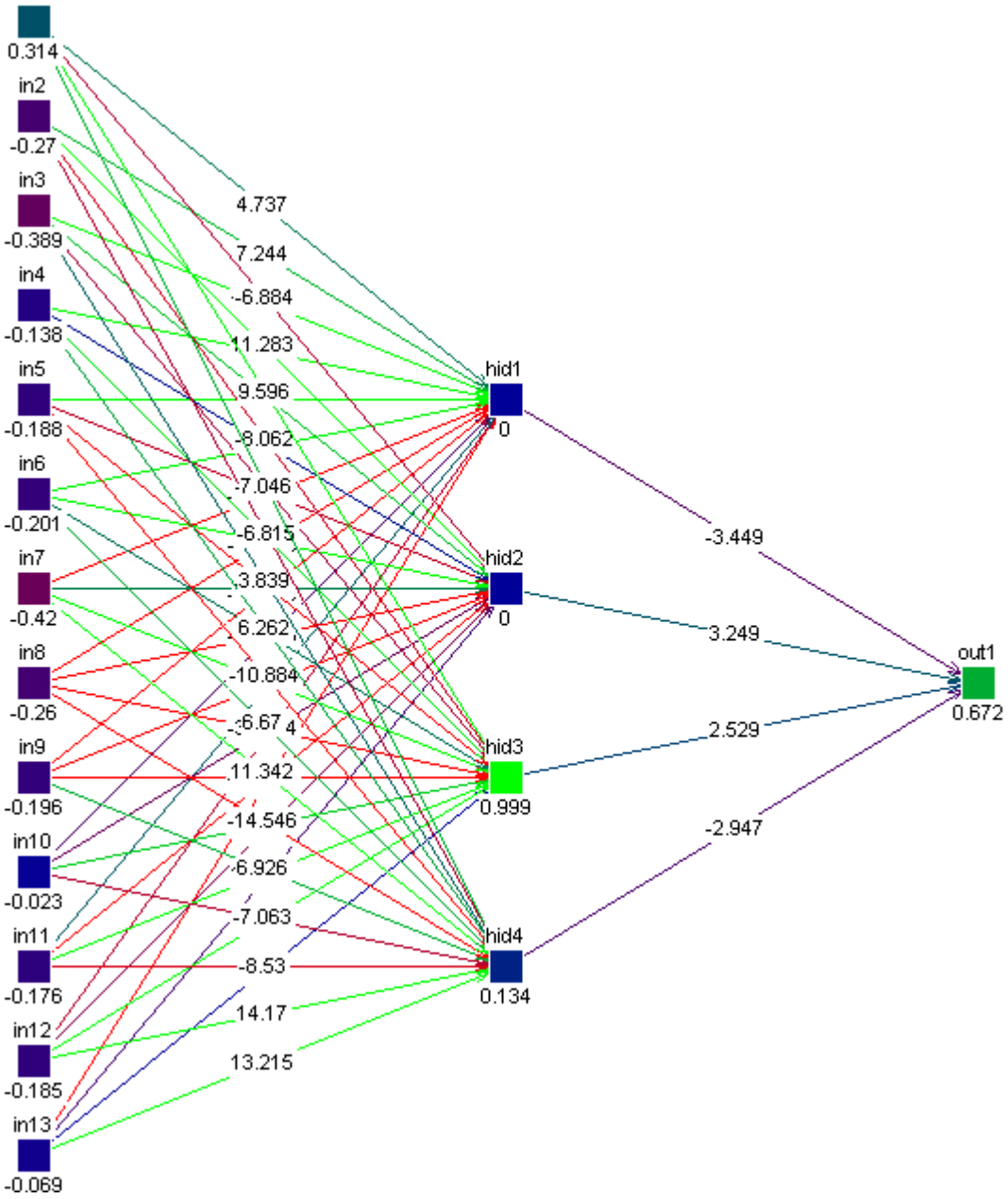

Figura 4.7: Vista de una red neuronal construida en JavaNNS para verificación de usuarios de la base *Voces-MCyTI*.

El número de iteraciones se asignó considerando que el aprendizaje de la red no es significativo si el MSE se mantiene constante. El valor máximo del error luego del entrenamiento se ubico en 650 unidades, para cualquier ANN_i. La figura 4.8 muestra el comportamiento típico del MSE observado durante el entrenamiento, para tres valores de iniciación del vector w_0 en una misma red.

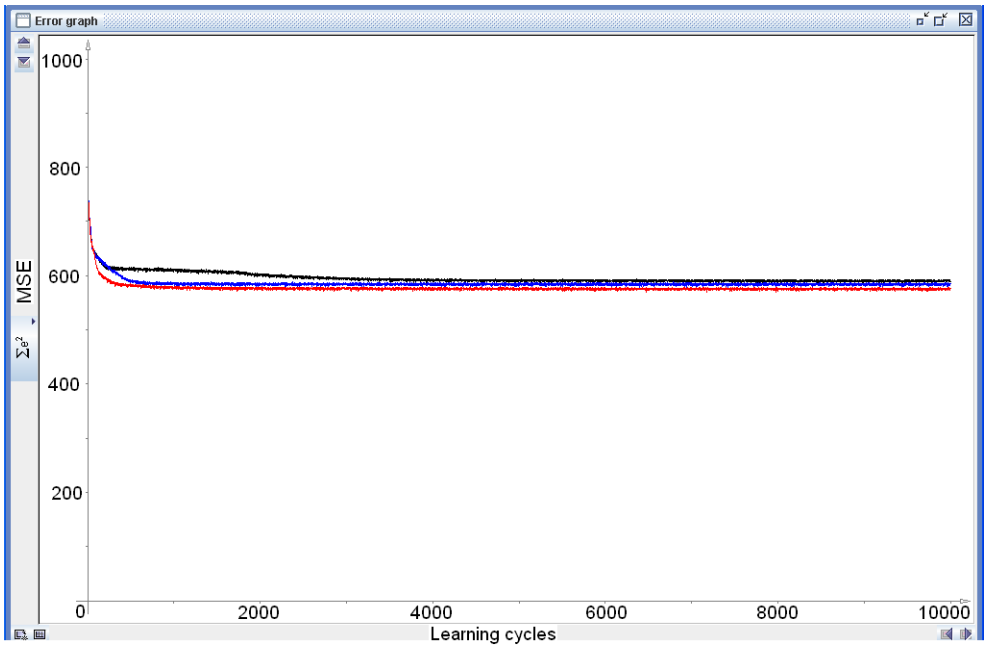

Figura 4.8: Comportamiento típico del MSE en el entrenamiento de las redes ANN_i. Cada curva representa una iniciación aleatoria del vector w^0

Una vez entrenadas las 15 redes ANN_i, se procedió a validar cada uno de estos modelos mediante los conjuntos de prueba P_i^1. Para ello se ingresó cada uno de estos conjuntos como entrada de cada una de las redes ANN_i. El significado de estas pruebas se da a continuación:

Clasificación de P_i^1 con ANN_i. Se clasifica el primer conjunto de vectores de prueba del usuario U_i con la red ANN_i asociada. Esto significa que el sistema trata la petición de un usuario válido respecto de ANN_i, por lo que el porcentaje de acierto obtenido mide el desempeño o eficiencia del modelo generado para verificar a usuarios válidos.

Clasificación de P_j^1 con ANN_i, $i \neq j$. Se clasifica el primer conjunto de prueba del usuario U_j, con la red de otro usuario U_i distinto. Esto significa que el sistema trata con una petición de un usuario inválido respecto de ANN_i, por lo que el porcentaje obtenido mide el error cometido por el sistema ante este tipo de solicitudes.

.	$\mathbf{ANN_1}$	$\mathbf{ANN_2}$	$\mathbf{ANN_3}$	$\mathbf{ANN_4}$	$\mathbf{ANN_5}$	$\mathbf{ANN_6}$	$\mathbf{ANN_7}$	$\mathbf{ANN_8}$
P_1^1	**82.936**	40.079	14.285	16.269	36.111	32.936	32.539	17.857
P_2^1	9.126	**84.920**	40.873	40.476	16.666	39.285	19.047	4.761
P_3^1	19.841	27.777	**75.793**	23.809	30.555	46.428	26.984	13.888
P_4^1	19.841	28.571	38.492	**68.650**	37.698	46.825	39.285	6.349
P_5^1	25.000	29.365	40.873	27.380	**61.904**	44.047	31.746	11.111
P_6^1	30.158	40.476	20.238	13.888	41.666	**66.269**	44.047	17.857
P_7^1	12.698	18.253	29.365	40.476	26.587	63.888	**71.428**	6.349
P_8^1	22.619	31.746	16.666	2.777	7.936	9.523	18.253	**84.126**
P_9^1	15.079	5.555	48.412	46.428	16.666	40.079	26.984	10.714
P_{10}^1	20.634	38.095	15.873	6.349	12.698	19.047	23.809	43.253
P_{11}^1	16.666	36.507	54.365	53.174	26.190	57.539	18.650	3.571
P_{12}^1	15.079	34.126	7.539	11.111	20.634	28.174	31.349	7.936
P_{13}^1	5.952	6.349	38.492	52.380	15.476	49.206	40.079	10.317
P_{14}^1	11.111	16.666	37.698	30.15	9.126	31.746	30.555	13.492
P_{15}^1	23.015	37.698	24.206	12.698	23.809	32.142	22.619	10.714

Tabla 4.8: Porcentajes del sistema de verificación con ANN.

.	$\mathbf{ANN_9}$	$\mathbf{ANN_{10}}$	$\mathbf{ANN_{11}}$	$\mathbf{ANN_{12}}$	$\mathbf{ANN_{13}}$	$\mathbf{ANN_{14}}$	$\mathbf{ANN_{15}}$
P_1^1	34.126	11.111	1.190	42.857	41.269	36.111	46.031
P_2^1	5.952	13.492	44.444	21.825	14.682	24.603	21.428
P_3^1	24.206	25.000	3.571	11.507	24.603	28.571	25.793
P_4^1	34.126	17.857	5.555	17.857	29.365	42.063	17.857
P_5^1	27.777	14.682	5.555	20.238	34.920	25.396	27.777
P_6^1	42.857	24.206	1.984	34.920	40.873	46.428	40.079
P_7^1	22.619	15.079	5.158	16.269	44.444	23.809	13.095
P_8^1	47.619	42.460	0.000	29.761	40.476	54.761	23.809
P_9^1	**66.666**	22.222	7.936	11.507	51.984	36.507	8.333
P_{10}^1	57.936	**74.603**	0.793	22.222	38.888	55.952	24.603
P_{11}^1	10.714	8.730	**73.015**	17.857	6.349	18.650	15.873
P_{12}^1	19.444	11.111	5.555	**71.031**	29.761	22.619	46.603
P_{13}^1	30.158	19.047	11.111	15.873	**65.079**	29.761	9.126
P_{14}^1	40.079	42.460	1.984	9.523	48.412	**73.809**	7.936
P_{15}^1	14.682	14.285	5.952	50.000	30.952	21.031	**78.968**

Tabla 4.9: (Continuación) Porcentajes del sistema de verificación con ANN.

Los resultados obtenidos de estas pruebas se encuentran reportados en las tablas 4.8 y 4.9.

Nuevamente, el mejor porcentaje de clasificación se obtiene cuando se compara el conjunto P_i^1 con su modelo respectivo ANN_i (Valores en negritas). La evaluación del desempeño del sistema ANN se da en la siguiente sección.

4.2.2. Curvas DET del sistema ANN

Los aspectos relevantes de la obtención de las curvas DET para un sistema de clasificación, han sido descritos en la seccion 3.2, por lo que únicamente se mencionan aquí los apectos principales.

Los parámetros asignados a la función DCF para la obtención de las curvas DET son los mismos que los establecidos en la tabla 4.4. Al igual que para la evaluación de Máquinas de Soporte Vectorial, se mejoró la resolución de dichas curvas al incrementar la cardinalidad de *TrueScores*; para ello dado $v_i \in$ *TrueScores*, $i = 1, \ldots, 15$, se generaron 500 valores con una distribución normal con media v_i y varianza σ^2, por lo que la cardinalidad final de *TrueScores* es de 7500. El conjunto *FalseScores* consistió de los 210 valores que se encuentran fuera de la diagonal en las tablas 4.8. En las figuras 4.9, 4.10 y 4.11 se esquematizan las curvas DET obtenidas para valores de σ^2 de 10.0, 4.0 y 2.0, respectivamente.

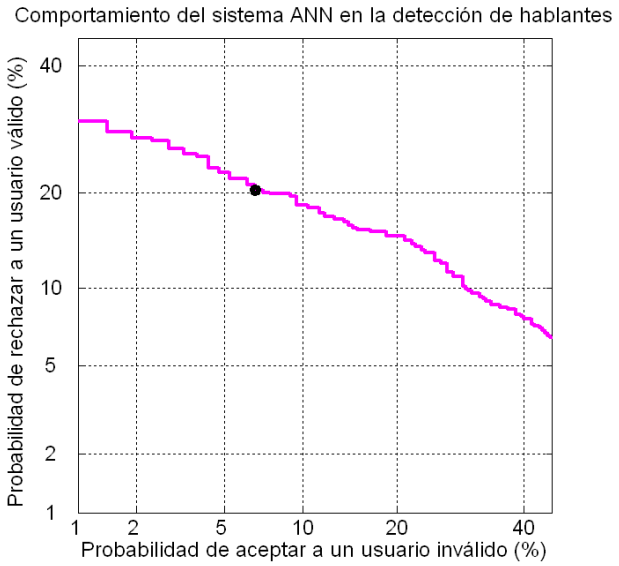

Figura 4.9: Curva DET para el sistema de verificación ANN con varianza σ^2 =10.0 en los porcentajes de acierto.

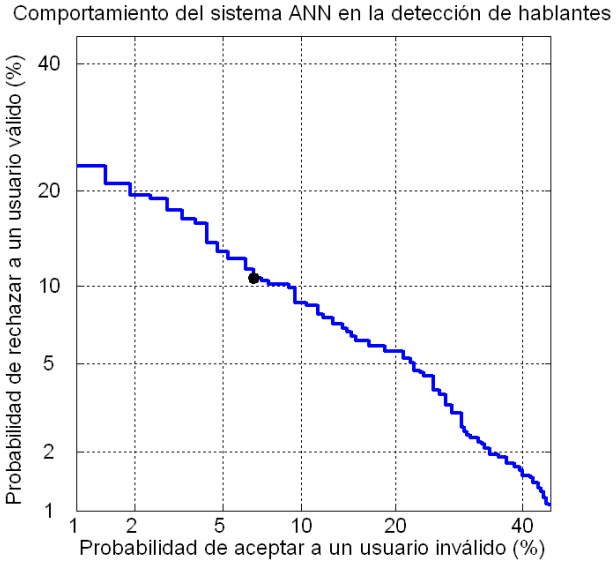

Figura 4.10: Curva DET para el sistema de verificación ANN con varianza σ^2 =4.0 en los porcentajes de acierto.

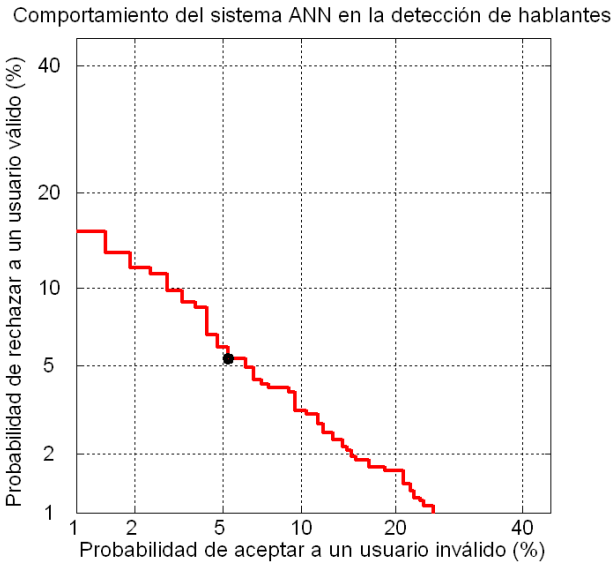

Figura 4.11: Curva DET para el sistema de verificación ANN con varianza σ^2 =2.0 en los porcentajes de acierto.

La tabla 4.10 resume las medidas de desempeño obtenidas mediante los puntos óptimos de error de las curvas DET (Ver ecuaciones 3.7 y 3.9).

Curva	**Varianza**	$p(R\|H)$ (%)	$p(A\|\neg H)$ (%)	$p(Error)$	**Desempeño (%)**
fig. 4.9	10.0	20.346	6.666	0.0687	93.13
fig. 4.10	4.0	10.613	6.666	0.0672	93.28
fig. 4.11	2.0	5.346	5.238	0.0524	94.76

Tabla 4.10: Valores óptimos de error que minimizan DCF y error y desempeño asociados para el sistema ANN.

Al comparar estos resultados de desempeño con los presentados en la tabla 3.8, se observa que nuevamente el sistema mediante Máquinas de Soporte Vectorial presenta un mejor comportamiento. Sin embargo las diferencias son, en cualquier caso, no mayores a 4.5 puntos porcentuales, lo cual representa una diferencia mínima.

A pesar de esto, el comportamiento punto a punto de las curvas DET es mejor en el sistema SVM que en ANN, lo cual significa que el sistema SVM tiene en cualquier caso, un mejor comportamiento y no sólo en los puntos óptimos.

Capítulo 5

Conclusiones y Trabajo Futuro

5.1. Conclusiones

El trabajo teórico y experimental realizado en esta investigación, permite establecer las siguientes conclusiones:

- Se estableció una metodología para crear un sistema automático de verificación de hablantes, basado en Máquinas de Soporte Vectorial como sistema de clasificación.
- Se validó la metodología propuesta implementando, a partir de ésta, un sistema automático de verificación por voz. Este sistema incluyó la creación de una base de registros de voz en español llamada *Voces-MCyTI*, módulos de procesamiento de los registros basados en el análisis MFCC y la aplicación del sistema *Libsvm*, como sistema de clasificación basado en Máquinas de Soporte Vectorial. En la etapa de procesamiento se presentó y validó un supresor de silencios o pausas entre palabras. Se comprobó que dicho proceso es necesario para un correcta operación del sistema de verificación.
- Los porcentajes de clasificación obtenidos en las pruebas realizadas al sistema (Tablas 3.3 y 3.4), permiten concluir que es posible crear un sistema de verificación e incluso de identificación de personas a partir de la voz.
- Los resultados presentados en la tabla 3.5, garantizan que el sistema es estable al disminuir el número de vectores del conjunto de prueba. Los porcentajes obtenidos para peticiones válidas con conjuntos de 102 vectores, seleccionados de forma aleatoria, están por arriba del 50 %.
- El desempeño mostrado por el sistema y representado en las curvas DET (figuras 3.2, 3.3 y 3.4), muestra que el error disminuye al disminuir la varianza en los porcentajes de acierto. Este comportamiento indica que el desempeño del sistema mejorará, si se logra minimizar la variabilidad en los vectores de voz asociados a un mismo usuario.

- El sistema implementado fue validado mediante su evaluación con registros de voz en tiempo real. Esta validación permitió comprobar que el sistema es independiente de las frases pronunciadas o si estas se hacen de forma secuencial. El desempeño obtenido es, en cualquier caso, muy próximo (menos de 1 %) al obtenido mediante los conjuntos de prueba P_i, formados mediante selección aleatoria. La validación incluyó muestras de voz de usuarios que no cuentan con un modelo dentro del sistema. Los porcentajes de acierto obtenidos se encuentran por encima del 50 %.

- Se presentó y validó un procedimiento para obtener el desempeño de sistemas automáticos de verificación. Mediante este procedimiento fue posible obtener los porcentajes para dos sistemas de clasificación adicionales; Modelos de Mezclas de Gaussianas y Redes Neuronales Artificiales. La tabla 5.1 muestra los resultados de desempeño para los tres sistemas propuestos en este trabajo.

σ^2	SVM	GMM	ANN
10.0	97.38 %	95.01 %	93.13 %
4.0	97.53 %	95.16 %	93.28 %
2.0	98.07 %	95.25 %	94.76 %

Tabla 5.1: Desempeño de tres sistemas de clasificación aplicados a la base *Voces-MCyTI*.

- Las gráficas dadas a continuación, presentan una comparación del error para los tres sistemas de clasificación evaluados.

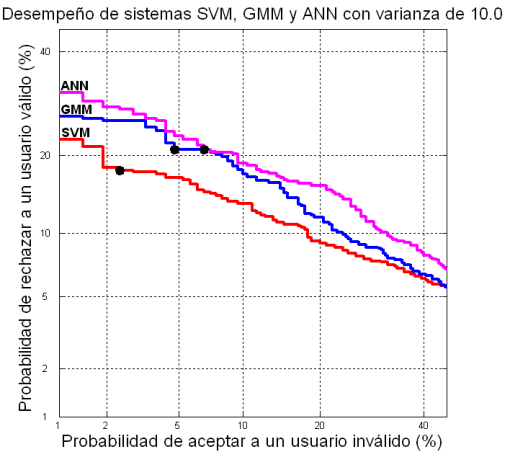

Figura 5.1: Error de tres sistemas de clasificación evaluados en la base *Voces-MCyTI*. σ^2 =10.0.

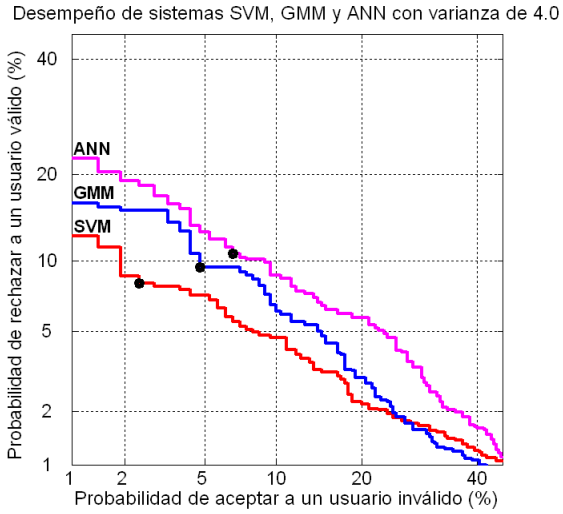

Figura 5.2: Comparación del error de tres sistemas de clasificación evaluados en la base *Voces-MCyTI*. σ^2 =4.0.

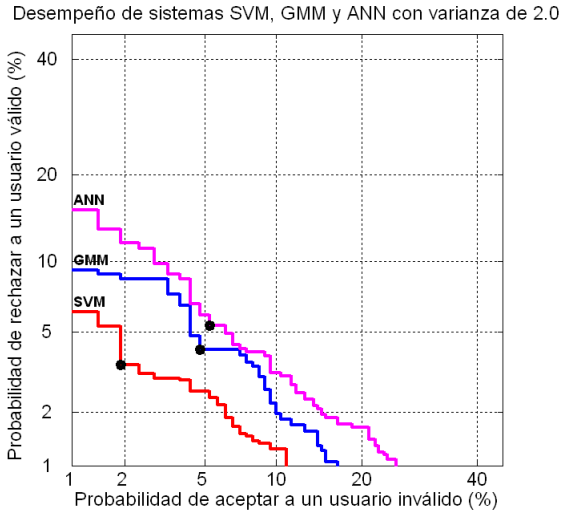

Figura 5.3: Comparación del error de tres sistemas de clasificación evaluados en la base *Voces-MCyTI*. σ^2 =2.0.

- De estos resultados se concluye que el sistema mediante Máquinas de Soporte Vectorial (SVM), tiene mejor desempeño que los sistemas GMM y ANN, aún cuando la diferencias encontradas se encuentren en un intervalo de 4.5 %.

- El error representado en las curvas DET de cada uno de los tres sistemas de clasificación, muestra que el sistema basado en Máquinas de Soporte Vectorial, tiene el mejor comportamiento de probabilidad de error. Esto significa que tiene un mejor manejo ante peticiones de usuarios válidos e inválidos.

Las conclusiones hechas permiten verificar el logro de los objetivos propuestos. También confirman que las Máquinas de Soporte Vectorial son una técnica exitosa que, con base en la metodología propuesta, puede mejorar el desempeño de otros sistemas de clasificación, como se mostró para los sistemas GMM y ANN. Por su relevancia en el contexto de la investigación, puede mencionarse la creación de una base de registros de voz en español, lo cual contribuye al acervo pues existen pocas referencias de este tipo de trabajos en nuestro lenguaje. En la parte de procesamiento, la propuesta y validación del supresor de silencios constituye una opción sencilla y viable que permite y mejora sustancialmente el desempeño global. Por último se resalta la contribución del trabajo en su conjunto, ya que permite una implementación directa mediante software, de un sistema automático de verificación de hablantes basado en SVM, lo cual resulta novedoso en México.

5.2. Trabajo Futuro

Como parte principal del trabajo a futuro propuesto, se encuentra la variación de los parámetros utilizados en el procesamiento de los registros de voz; principalmente el tiempo de cada segmento o trama en el análisis Cepstral, la atenuación y resolución del banco de filtros Mel, así como el número de coeficientes cepstrales, mismo que determina la dimensión de los vectores de características. Estas variaciones tienen el objetivo de mejorar la resolución entre las distribuciones de los porcentajes de acierto y error, del sistema mediante Máquinas de Soporte Vectorial y aumentar así su desempeño. La base *Voces-MCyTI* puede complementarse incrementando el número de hablantes registrados. Algunos de éstos registros pueden ser utilizados para evaluar el sistema ante peticiones de usuarios que no cuenten con un modelo dentro del mismo, en una forma más amplia. Los vectores provenientes de registros no enrolados en el sistema, pueden complementar el *Background* de los usuarios válidos, lo cual puede mejorar el desempeño al disminuir errores de tipo $(A, \neg H)$. Por último se propone realizar pruebas con el sistema SVM utilizando otras funciones núcleo, a fin de observar su comportamiento en la clasificación.

Apéndice A

Anexos

A.1. Notación SAMPA del Español

Notación	Ejemplo	Ejemplo en SAMPA
p	**p**adre	"paDre
b	**v**ino	"bino
t	**t**omo	"tomo
d	**d**onde	"donde
k	**c**asa	"kasa
tS	mu**ch**o	"mutSo
jj	**hi**elo	"jjelo
f	**f**ácil	"faTil
B(=/b/)	ca**b**ra	"kaBra
T	**c**inco	"Tinko
D(=/d/)	na**d**a	"naDa
s	**s**ala	"sala
x	mu**j**er	mu"xer
G(=/g/)	lue**g**o	"lweGo
m	**m**is**m**o	"mismo
n	**n**u**n**ca	"nunka
J	a**ñ**o	"aJo
l	**l**ejos	"lexos
L	caba**ll**o	ka"baLo

Tabla A.1: Notación SAMPA del español.

Notación	Ejemplo	Ejemplo en SAMPA
r	pu**r**o	"puro
rr	to**rr**e	"torre
j	re**y**/ p**i**e	rrej/pje
w	de**u**da / m**u**y	"dewda / mwi
a	v**a**lle	"baLe
e	p**e**ro	"pero
i	p**i**co	"piko
o	t**o**r**o**	"toro
u	d**u**ro	"duro

Tabla A.2: (Continuación) Notación SAMPA del español.

A.2. Algoritmo de Cálculo de Curvas DET

function [Pmiss, Pfa]=Compute_DET(true_scores, false_scores)

```
SMAX = 9E99;
% Compute the miss/false_alarm error probabilities
num_true = max(size(true_scores));
num_false = max(size(false_scores));
total=num_true+num_false;
Pmiss = zeros(num_true+num_false+1, 1); %preallocate for speed
Pfa = zeros(num_true+num_false+1, 1); %preallocate for speed
scores(1:num_false,1) = false_scores;
scores(1:num_false,2) = 0;
scores(num_false+1:total,1) = true_scores;
scores(num_false+1:total,2) = 1;
scores=DETsort(scores);
sumtrue=cumsum(scores(:,2),1);
sumfalse=num_false - ([1:total]'-sumtrue);
Pmiss(1) = 0;
Pfa(1) = 1.0;
Pmiss(2:total+1) = sumtrue ./ num_true;
Pfa(2:total+1) = sumfalse ./ num_false;
return
```

function h = Plot_DET (Pmiss, Pfa, plot_code, opt_thickness)

```
Npts = max(size(Pmiss));
if Npts ~ = max(size(Pfa))
```

```
error ('vector size of Pmiss and Pfa not equal in call to Plot_DET');
end
%----------------------
% plot the DET
if ~ exist('plot_code')
plot_code = 'y';
end
if ~ exist('opt_thickness')
opt_thickness = 0.5;
end
Set_DET_limits;
h = thick(opt_thickness,plot(ppndf(Pfa), ppndf(Pmiss), plot_code));
Make_DET;
```

function Make_DET()

```
pticks = [0.00001 0.00002 0.00005 0.0001 0.0002 0.0005 ...
0.001 0.002 0.005 0.01 0.02 0.05 ...
0.1 0.2 0.4 0.6 0.8 0.9 ...
0.95 0.98 0.99 0.995 0.998 0.999 ...
0.9995 0.9998 0.9999 0.99995 0.99998 0.99999];

    xlabels = [' 0.001' ; ' 0.002' ; ' 0.005' ; ' 0.01 ' ; ' 0.02 ' ; ' 0.05 ' ; ...
' 0.1 ' ; ' 0.2 ' ; ' 0.5 ' ; ' 1 ' ; ' 2 ' ; ' 5 ' ; ...
' 10 ' ; ' 20 ' ; ' 40 ' ; ' 60 ' ; ' 80 ' ; ' 90 ' ; ...
' 95 ' ; ' 98 ' ; ' 99 ' ; ' 99.5 ' ; ' 99.8 ' ; ' 99.9 ' ; ...
' 99.95' ; ' 99.98' ; ' 99.99' ; '99.995' ; '99.998' ; '99.999'];
ylabels = xlabels;

    %----------------------
% Get the min/max values of Pmiss and Pfa to plot

    global DET_limits;
if isempty(DET_limits)
Set_DET_limits;
end
Pmiss_min = DET_limits(1);
Pmiss_max = DET_limits(2);
Pfa_min = DET_limits(3);
Pfa_max = DET_limits(4);
```

```
    %--------------------
% Find the subset of tick marks to plot

    ntick = max(size(pticks));
for (n=ntick:-1:1)
if (Pmiss_min <= pticks(n))
tmin_miss = n;
end
if (Pfa_min <= pticks(n))
tmin_fa = n;
end
end
for (n=1:ntick)
if (pticks(n) <= Pmiss_max)
tmax_miss = n;
end
if (pticks(n) <= Pfa_max) tmax_fa = n;
end
end

    %--------------------
% Plot the DET grid

    set (gca, 'xlim', ppndf([Pfa_min Pfa_max]));
set (gca, 'xtick', ppndf(pticks(tmin_fa:tmax_fa)));
set (gca, 'xticklabel', xlabels(tmin_fa:tmax_fa,:)); set (gca, 'xgrid', 'on');
xlabel ('Probabilidad de rechazar a un usuario válido (%)');
set (gca, 'ylim', ppndf([Pmiss_min Pmiss_max]));
set (gca, 'ytick', ppndf(pticks(tmin_miss:tmax_miss)));
set (gca, 'yticklabel', ylabels(tmin_miss:tmax_miss,:));
set (gca, 'ygrid', 'on')
ylabel ('Probabilidad de aceptar a un usuario invlido (%)')
set (gca, 'box', 'on');
axis('square');
axis(axis);
```

%function **ppndf (prob)** The input to this function is a cumulative probability. The output from this function is the Normal deviate that corresponds to that probability.

Referencias

[1] A. Higgins, L. Bahler and J. Porter, *Speaker Verification Using Randomized Phrase Prompting*, Digital Signal Processing, Vol. 1, No. 2, pp.89-106, 1991.

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, *The DET Curve in Assesment of Detection Task Performance*, National Institute of Standards and Technology, SRI International Department of Defense, Department of Defense, USA, http://www.nist.gov/speech/.

[3] A.L. Higgins and R.E. Wholford,*A New Method on Text-Independent Speaker Recognition*, IEEE Proceedings of The International Conference on Acoustic, Speech and Signal Processing, Tokyo, Japan, pp. 869-872, 1986.

[4] A.K. Jain, J. Mao and K.M. Mohiuddin, *Artificial Neural Networks: A Tutorial*, IEEE Transactions, Vol. 29, pp. 31-44, 1996.

[5] B.H. Juang, R. Perdue and D. Thompson, *Deployable Authomatic Speech Recognition Systems: Advances and Challenges*, AT&T Technical Journal, No 74(2), 1995.

[6] B.S. Atal, *Effectiveness of Linear Prediction Characteristics of Speech Wave for Automatic Speaker Identification and Verification*, Journal Acoustic of America, Vol. 55, No. 6, pp.1304-1312, 1974.

[7] C. Che and Q. Lin, *Speaker Recognition Using Hidden Markov Models with Experiments on The YOHO Data Base*, Proceedings EUROSPEECH, Madrid, Spain, pp.625-628, 1995.

[8] C.C. Tappert, N.R. Dixon, A.S. Rabinowitz and W.D. Chapma, *Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery*, Technical Report TR-71-146, Rome Air Devices Center, Rome, NY, 1971.

[9] C.J.C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, Microsoft Research, Data Meaning and Knowledge Discovery 2, pp. 121-167, 1998.

[10] D. Childers, R.V. Cox, R. DeMori, S. Furui, B.H. Juang, J.J. Mariani, P. Price, S. Sagayama, M.M. Sondhi, R. Weischedel, *The Past, Present and Future of Speech Processing*, IEEE Signal Processing Magazine, pp. 24-48, May 1998.

[11] D. Reynolds, *MIT Lincoln Laboratory Site Presentation*, Speaker Recognition Workshop, Maritime Institute of Technology, A. Martin, MD, 1996.

[12] D. Reynolds and B. Carlson, *Text-Dependent Speaker Verification Using Decoupled and Integrated Speaker and Speech Recognizers*, Proceedings EUROSPEECH, Madrid, Spain, pp. 647-650,1995.

[13] D. Reynolds and R. Rose, *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*, IEEE Transactions on Speech Audio Processing , Vol. 3, No. 1, pp. 72-83, 1995.

[14] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models*, Digital Signal Processing, Vol. 10, pp. 19-41, 2000.

[15] D.B. Fry, *Theoretical Aspects of Mechanical Speech Recognition*; and P. Denes, *The Design and Operation of the Mechanical Speech Recognizer*, University College London, Journal of British Institution Radio Engr., No. 19(4), pp. 211-229, 1959.

[16] D.R. Reddy, *An Approach to Computer Speech Recognition by Direct Analysis of The Speech Wave*, Technical Report C549, Computer Science Department, Stanford University, September 1966.

[17] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovzka-Delacrétaz, D.A. Reynolds, *A Tutorial on Text-Independent Speaker Verification*, EURASIP Journal on Applied Signal Processing 4, pp. 430-451, 2004.

[18] F. Itakura, *Minimum Prediction Residual Applied to Speech Recognition*, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-23(1), pp. 67-72, February 1975.

[19] F. Jelinek, L.R. Bahl and R.L. Mercer, *Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech*, IEEE Transactions on Information Theory, IT-21, pp.250-256, 1975.

[20] F.K. Soong, A.E. Rosemberg, L.R. Rabiner and B.H. Juang, *A Vector Quantization Approach to Speaker Recognition*, Proceedings of The International Conference on Acoustic, Speech and Signal Processing, Tampa, FL, pp. 387-390, 1985.

[21] G.R. Doddington, *Speaker Recognition, Identifying People by Their Voices*, IEEE Proceedings, Vol. 73, pp. 1651-1664, November 1985.

[22] H. Chi-Wei, Ch. Chih-Chung, L. Chih-Jen, *A Practical Guide to Support Vector Classification*, http://www.csie.ntu.edu.tw/∼cjlin/papers/guide.

[23] H. Fengei and W. Bingxi, *Text-Independent Speaker Verification Using Speaker Clustering and Support Vector Machines*, IEEE Proceedings of The International Conference on Signal Processing, 2002.

[24] H. Sakoe, *Two Level DP Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition*, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-27, pp. 588-595, December 1979.

[25] H. Sakoe and S. Chiba, *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-26(1), pp. 43-49, February 1978.

[26] H.F. Olson and H. Belar, *Phonetic Typewriter*, Journal of Acoustic Society of America, No. 28(6), pp. 1072-1081, 1956.

[27] J. Attili, M. Savic and J. Campbell, *A TM32020-Based Real Time, Text-Independent, Automatic, Speaker Verification System*, IEEE Proceedings of The International Conference on Acoustic, Speech and Signal Processing, New York, pp. 599-602, 1988.

[28] J. Colombi, D. Ruck, S. Rogers, M. Oxley and T. Anderson, *Cohort Selection and Word Grammar Effects for Speaker Recognition*, IEEE Proceedings of The International Conference on Acoustic, Speech and Signal Processing, Atlanta, GA, pp. 85-88, 1996.

[29] J. Naik, *Speaker Verification: A Tutorial*, IEEE Communication Magazine, Vol. 28, pp.42-48, January 1990.

[30] J. Ortega-García, J. González-Rodríguez, V. Marrero-Aguiar, *A Large Speech Corpus in Spanish for Speaker Characterization and Identification*, Speech Communication, 1999.

[31] J. Suzuki and K. Nakata, *Recognition of Japanese Vowels - Preliminary to Recognition of Speech*, Journal of Radio Research Lab, No. 37(8), pp. 193-212, 1961.

[32] J.D. Markel, S.B. Davis, *Text-Independent Speaker Recognition from a Large Linguistically Uncosntrained Time-Spaced Database*, IEEE Transactions on Acoustic Speech, Signal Processing, Vol. ASSP-27, No. 1, pp. 74-82, 1979.

[33] J.G. Proakis, *Digital Communications*, Ed. McGraw Hill, Fourth ed., pp. 239.

[34] J.P. Campbell Jr., *Speaker Recognition: A Tutorial*, IEEE Proceedings, Vol. 85, No. 9, September 1997.

[35] J.W. Forgie and C.D. Forgie, *Results Obtained from a Vowel Recognition Computer Program*, Journal of Acoustic Society of America, No. 31(11), pp. 211-229, 1959.

[36] K. Daoudi, J. Louradour, *A Novel Strategy for Speaker Verification Based on SVM Classification of Pairs of Sequences*, IRIT-CNRS, Toulouse, France, 2007.

[37] K.H. Davis, R. Biddulph and S. Balashek, *Automatic Recognition of Spoken Digits*, Journal of Acoustic Society of America, No. 24, pp. 637-642, 1952.

[38] K.P. Li and E.H. Wrench, Jr., *Text-Independent Speaker Recognition with Short Utterances*, IEEE Proceedings of The International Conference on Acoustic, Speech and Signal Processing, Boston, MA, pp. 555-558, 1983.

[39] L.R. Rabiner, S.E. Levinson, A.E. Rosemberg and J.G. Wilpon, *Speaker Independent Recognition of Isolated Words Using Clustering Techniques*, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-27, pp. 336-349, August 1979.

[40] M. Seltzer, *SPHINX III Signal Processing Front End Specification*, CMU Speech Group, 1999.

[41] N.Z. Tishby, *On The Application of Mixture Hidden Markov Models to Text Independent Speaker Recognition*, IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 39, No. 3, pp. 563-570, 1991.

[42] Q. Miao, H-Z. Huang and X. Fan, *A Novel Hybrid System with Neural Networks and Hidden Markov Models in Fault Diagnosis*, Lecture Notes on Artificial Intelligence 4293, pp. 513-521, MICAI-2006.

[43] R. Schwartz, S. Roucos and M. Beroutti, *The Application of Probability Density Estimation on Text Independent Speaker Identification*, Proceedings of The International Conference on Acoustic, Speech and Signal Processing, Paris, France, pp. 1649-1652, 1982.

[44] S. Furui, *Cepstral Analysis Technique for Automatic Speaker Verification*, IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. ASSP-29, pp. 254-272, 1981.

[45] S. Raghavan, G. Lazarou and J. Picone, *Speaker Verification Using Support Vector Machines*, Center for Advanced Vehicular Systems, Missisipi State University, 2006.

[46] T. Kinnunen, *Spectral Features for Automatic Text-Independent Speaker Recognition*, Licenciate's Thesis, University of Joensuu, Department of Computer Science, Finland, 2003.

[47] T. Sakai and S. Doshita. *The Phonetic Typewriter, Information Processing*, IFIP Congress, Munich, 1962.

[48] T.B. Martin, A.L. Nelson and H.J. Zadell, *Speech Recognition by Feature Abstraction Techniques*, Technical Report AL-TDR-64-176, Air Force Avionics Lab, 1964.

[49] T.K. Vintsyuk, *Speech Discrimination by Dynamic Programming*, Kibernetika, No. 4(2), pp.81-88, January-February 1968.

[50] V. Wan, *Speaker Verification Using Support Vector Machines*, Ph.D Thesis, University of Sheffield, Department of Computer Science, United Kingdom, 2003.

[51] V. Wan and S. Renals, *Speaker Verification Using Sequence Discriminant Support Vector Machines*, IEEE Transactions on Speech and Audio Processing, Vol 13, No. 2, 2005.

[52] V.M. Velichko and N.G. Zagoruyco, *Automatic Recognition of 200 Words*, Journal of Man-Machine Studies, No. 2, pp.223, June 1970.

[53] W.M. Campbell, *A SVM/HMM System for Speaker Recognition*, IEEE Proceedings of The International Conference on Acoustic, Speech and Signal Processing, 2003.

[54] W.M. Campbell, D.E. Sturim and D.A. Reynolds, *Support Vector Machines Using GMM Supervectors for Speaker Verification*, IEEE Signal Processing Letters, 2006.

[55] W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones and T.R. Leek, *High Level Speaker Verification with Support Vector Machines*, IEEE Proceedings of The International Conference on Acoustic, Speech and Signal Processing, 2004.

[56] X. Dong and W. Zhahoui, *Speaker Recognition Using Continuous Density Support Vector Machines*, IEEE Electronic Letters, Vol. 37, No. 17, pp. 1099-1101, 2001.