Data Mining applied to Forensic Speaker Identification

P. Univaso, J. M. Ale, *Member, IEEE* and J. A. Gurlekian

***Abstract*— In this paper we analyze the advantages of using data mining techniques and tools for data fusion in forensic speaker recognition. Segmental and suprasegmental features were employed in 28 different classifiers, in order to compare their performances. The selected classifiers have different learning techniques: lazy or instance-based, eager and ensemble. Two approaches were employed on the classification task: the use of all features and the use of a feature subset, selected with a gain ratio methodology. The best performances, with all features, were obtained by three classifiers: Logistic Model Tree (eager), LogitBoost (ensemble) and Multilayer Perceptron (eager). Support Vector Machine (eager) proved to be a good classifier if a Pearson VII function-based universal kernel was used. When low dimensional features were selected, ensemble classifiers exceeded the performance of all others classifiers. Segmental and tone features demonstrated the best speaker discrimination capabilities, followed by duration and quality voice features. Evaluation was performed on Argentine-Spanish voice samples from the Speech_Dat database recorded on a fixed telephone environment. Different recording sessions and channels for the test segments were added and the Z-norm procedure was applied for channel compensation.**

***Keywords*— Data Mining, Classifiers, Ensemble Methods, Speaker Recognition, Data Fusion.**

I. INTRODUCCIÓN

EL RECONOCIMIENTO de hablantes o reconocimiento de personas por su voz posee en la actualidad varias aplicaciones prácticas: Financieras, Legales y Forenses, Control de Acceso (Seguridad), Indexación de Audio y Video, Vigilancia, Teleconferencia, y Capacitación a distancia.

En el presente trabajo nos focalizamos en las aplicaciones forenses y las posibles mejoras que pueden introducirse a los sistemas actuales de identificación de hablantes a partir de herramientas y técnicas de minería de datos desarrolladas en otras áreas tecnológicas.

Se analizan las ventajas de incorporar diferentes fuentes de información suprasegmental o prosódica a la información segmental, base de los sistemas automáticos de reconocimiento de hablantes actual, y se comparan los resultados empleando diferentes clasificadores del tipo perezosos, ansiosos y de ensamble.

Los parámetros suprasegmentales empleados cubren los parámetros de tono, duración, acento y calidad de voz.

Se profundiza en la utilización de las Máquinas de Soporte Vectorial (*SVM*) como clasificadores, dada su extendida utilización en la literatura del reconocimiento de hablantes. Para eso se estudian sus configuraciones óptimas, y la influencia de las diversas funciones núcleo en su rendimiento.

Dada la importancia de reducir la cantidad de parámetros a ser empleados por un sistema de reconocimiento de hablantes práctico, se analizan diferentes técnicas para la selección de atributos. El objetivo del análisis es la selección de una técnica que optimice el rendimiento del clasificador y a la vez emplee la menor cantidad de parámetros posibles.

Finalmente se valoriza la importancia relativa de cada uno de los parámetros acústicos empleados en forma individual y agrupados según las categorías acústicas: segmental, tono, calidad de voz, duración y acento. El análisis de los parámetros empleados por hombres y mujeres, en forma discriminada, nos permite diferenciar los parámetros comunes de los parámetros particulares de cada género.

II. RECONOCIMIENTO DE HABLANTES

El área del "reconocimiento de hablantes" tiene como objetivo la determinación de la identidad de una persona a partir de su voz, pudiéndosela dividir en dos subáreas: "identificación" y "verificación " de hablantes.

En la "verificación" se pretende determinar si un hablante es quien dice ser mediante su voz, o detectarlo en una conversación, estableciendo si un segmento de habla fue emitido por él. En la verificación la respuesta del sistema es binaria: acepta o rechaza la identidad del hablante haciendo una sola comparación y utilizando un umbral que pesa el costo de aceptar un impostor o rechazar un hablante verdadero.

La "identificación" realiza la comparación de los rasgos de un hablante incógnita con respecto a un número de hablantes. Se elige el hablante con la mínima probabilidad de error. La probabilidad de error tiende a uno en la medida que el número de hablantes con que debe compararse aumenta. Al aumentar el número de hablantes crecen las probabilidades de que dos o más hablantes tengan distribuciones muy próximas entre sí. En esas circunstancias la identificación es una tarea difícil de resolver.

A. Aplicación forense

Podemos enmarcar la identificación de hablantes en el ámbito forense dentro de la fonética forense [81], y más ampliamente, dentro de la lingüística forense [82]. Las áreas relacionadas con la fonética forense son las siguientes: la identificación de hablantes -que es la abarcaremos en este trabajo-; la determinación de perfiles de hablantes (ante la falta de un sospechoso, dar información sobre particularidades de tipo socio-económicas); la construcción de ruedas de voces (para el reconocimiento de voces por parte de

P. Univaso, Laboratorio de Investigaciones Sensoriales, INIGEM,UBA-CONICET, Buenos Aires, Argentina, punivaso@yahoo.com.ar

J. M. Ale, Facultad de Ingeniería, UBA, Buenos Aires, Argentina, ale@acm.org

J. A. Gurlekian, Laboratorio de Investigaciones Sensoriales, INIGEM,UBA-CONICET, Buenos Aires, Argentina, jag@fmed.uba.ar

testigos o víctimas); la identificación de contenido (determinando lo que fue dicho cuando la grabación es de mala calidad, o cuando la voz es patológica o tiene un acento extranjero); y la autenticación de los registros de audio (determinando si una grabación ha sido manipulada).

La identificación de hablantes empleada en el ámbito forense generalmente parte de una grabación de una voz relacionada con un hecho delictivo (grabación dubitada, prueba o evidencia) la cual es comparada con otros registros atribuidos a una persona, normalmente conocida (grabación indubitada o del sospechoso).

En el caso de delitos de los secuestros extorsivos la grabación dubitada generalmente se obtiene de registros telefónicos (teléfonos fijos o celulares), mientras que la indubitada se realiza durante la toma de declaración del imputado. En este último caso resulta aconsejable cumplir con las formalidades que prevea el código procesal local, que para el caso de Argentina corresponde a la figura genérica del artículo 270 del Código Procesal Penal; con la particularidad de que, para la "rueda de voces", generalmente se utiliza un cuerpo o plana de voz correspondiente al imputado, y otras grabadas por terceros, en la que generalmente se les hace repetir generalmente las mismas frases, a fin de cumplir con el requisito de "condiciones semejantes" que exige dicho código en materia de reconocimientos.

Cuando esta última condición es factible, los peritos forenses aplican la identificación del hablante en la modalidad dependiente del texto. Dada la posibilidad, prevista ya en la ley, de que el imputado se rehúse a repetir las frases solicitadas, se realizan planas de voz cuyo texto difiere de las indubitadas, aplicándose en este caso la modalidad independiente del texto.

Otra aplicación de la identificación del hablante es la requerida por la justicia en casos civiles o comerciales. En estos casos la grabación dubitada puede provenir de un registro telefónico, pero también de grabaciones "en vivo" autorizadas por el juzgado.

En el caso forense y en las diversas formas de obtención de registros de voz, se posee o se tiene la capacidad de obtener las transcripciones de las grabaciones dubitadas e indubitadas. Esta posibilidad es aprovechada en el presente trabajo para incorporar, al reconocimiento independiente del texto (información segmental), información prosódica o suprasegmental.

B. Avances en el reconocimiento de hablantes

La historia de los primeros trabajos sobre reconocimiento de hablantes se remonta a los años 50 [1, 2], habiendo sido Pruzansky [3] en los 60s, de los Laboratorios Bell, el que desarrolló uno de los primeros sistemas, que empleaba bancos de filtros para la comparación de espectrogramas. Posteriormente Doddington en Texas Instruments (TI) reemplazó los bancos de filtros por el análisis de formantes [4].

En la década de los 70s las variaciones intra-hablantes fueron investigadas por Endres *et al.* [5] y Furui [6].

En la década de los 80s se introdujo el uso de los Modelos Ocultos de Markov (*HMM*), los cuales estaban siendo empleados en los primeros sistemas de reconocimiento de habla [7].

Al comienzo de la década de los 90s Rose y Reynolds [8] propusieron el uso de un *HMM* de un único estado, denominado Modelo de Mezclas Gaussianas (*GMM*) y que actualmente es el empleado en la mayoría de los sistemas de reconocimiento de hablantes. En esta década se profundizó en el reconocimiento robusto, que ataca las diferencias intra-hablante, el ruido y las diferencias de canal. También se introdujo el uso de modelos universales y cohortes de hablantes para el cálculo de relaciones de verosimilitud [9] y el modelado empleando redes neuronales [13]. A finales de los 90s el Instituto Nacional de Estándares y Tecnología (*NIST*) de los Estados Unidos de Norteamérica inició las evaluaciones de los sistemas de reconocimiento de hablantes, que continúan hasta la actualidad con una periodicidad bianual, para determinar el estado del arte de la tecnología.

En los 2000s se incorporaron diversas técnicas de normalización de modelos de hablantes basadas en la normalización cero (*z-norm*) desarrollada por Li y Porter a fines de los 90s [10], se introdujeron los primeros rasgos distintivos de largo plazo -en este caso rasgos idiolectales [11]-, y se comenzaron a emplear las Máquinas de Soporte Vectorial (*SVM*) como clasificadores [14]. Uno de los principales sistemas de reconocimiento, desarrollado en el instituto de investigación SRI International, incorpora rasgos suprasegmentales y prosódicos para mejorar la performance del sistema [12]. Otro trabajo [76] adiciona a los parámetros suprasegmentales y espectrales los correspondientes a la fuente de excitación.

En la década del 2010 están siendo investigados los métodos de representación del habla en subespacios vectoriales, como herramientas para remover o atenuar las características que no son propias del hablante, con técnicas de transformación como el Análisis Factorial Común (*JFA*) [24] y las aproximaciones por vectores de factor total (*iVectors*) [25].

La historia del reconocimiento de hablantes en el ámbito forense es mucho más antigua y se remonta al siglo XVII en la corte de Carlos I de Inglaterra [15]. La primera investigación científica fue realizada en 1937 en el caso Lindbergh [16]. La identificación de hablantes comenzó empleando técnicas de reconocimiento por escucha, haciendo uso de la discriminación de voces por parte de los humanos. La identificación de hablantes es practicada de forma sistemática por científicos forenses a finales de los 40s en la Unión de Repúblicas Socialistas Soviéticas y en los 50s en los Estados Unidos de Norteamérica.

Durante los años 60, y a la luz de la confiabilidad de los sistemas de huellas dactilares, desarrollados a fines del siglo XIX, se propuso un sistema similar denominado "huella de voz" [17], basado en el análisis visual de espectrogramas, el cual no alcanzó los resultados esperados dada la variabilidad intrínseca del habla. La controversia sobre la conveniencia de su uso duró más de 10 años y concluyó con el informe de Stevens *et al.* [18] que determinó que el método auditivo era más preciso que la inspección visual de espectrogramas.

A partir de los 70s, los métodos basados en humanos, contaron con el aporte de los sistemas automáticos de reconocimiento de hablantes desarrollados para aplicaciones biométricas. A pesar de que dichos sistemas evolucionaron en aplicaciones comerciales, no se pudieron crear sistemas confiables a ser empleados en el ámbito forense, existiendo una permanente controversia en su uso por parte de los expertos forenses y los científicos del reconocimiento de hablantes [23].

En el 2003 Bonastre *et al.* [19] presentaron un informe alertando sobre la imposibilidad de identificar unívocamente a una persona por su voz con los avances científicos alcanzados hasta el momento, especialmente en el ámbito forense donde el entorno y los factores que afectan la performance pueden variar tremendamente con respecto al ámbito comercial.

Durante la primera década del 2000, un grupo de investigadores del Reino Unido [20, 21, 22] comenzó una discusión sobre las características que debe cumplir la comparación de hablantes en el ámbito forense, concluyendo en la importancia de la presentación de resultados cuantitativos en la forma de relaciones de verosimilitud en lugar de resultados binarios, de manera que sea el juez quien determine la identidad del sospechoso en base a las comparaciones realizadas por los peritos.

A pesar de los grandes avances en el reconocimiento automático de hablantes durante estos últimos años, todavía no se ha podido afianzar como una tecnología de amplia utilización, aunque algunas aplicaciones comerciales la emplean; pero es especialmente en el ámbito forense donde mayor resistencia encuentra en comparación con otras técnicas biométricas de mayor confiabilidad como las del ácido desoxirribonucleico (ADN) y las huellas dactilares.

C. Metodología GMM-UBM

La solución al problema del reconocimiento del hablante se aborda de acuerdo a las técnicas del estado del arte para la detección del hablante, mediante la construcción de una base de datos que se constituye en el universo de hablantes alternativos al hablante hipotético. Esta base de datos permite construir un modelo de referencia universal (*UBM*) a partir del cual se deriva el propio modelo del hablante –se la sesga de acuerdo con las características propias del hablante- y permite definir el cociente de verosimilitudes obtenido mediante la estimación de los parámetros de su representación probabilística por medio de los datos de entrenamiento del hablante y los obtenidos del conjunto de hablantes de la base de referencia.

Cada hablante puede ser representado por un Modelo de Mezclas Gaussianas (*GMM*), que incluye únicamente información segmental, y el cual puede ser visto como un Modelo Oculto de Markov (*HMM*) con un único estado. La adaptación de los modelos de cada hablante se realiza con una adaptación bayesiana, por medio de la estimación *maximum a posteriori* (*MAP*), y a partir del *UBM*.

El resultado de este proceso es el cociente entre las verosimilitudes entre la evidencia y el sospechoso, y entre la evidencia y la base universal. Dicho cociente se presenta normalmente como una diferencia de logaritmos, denominada cociente de verosimilitudes logarítmico (*LLR*).

Los sistemas que únicamente emplean la información segmental del *GMM* para reconocer a un hablante realizan la decisión final comparando el *LLR* resultante con un nivel de decisión determinado. Un esquema simplificado de esta metodología puede verse en la Fig. 1.

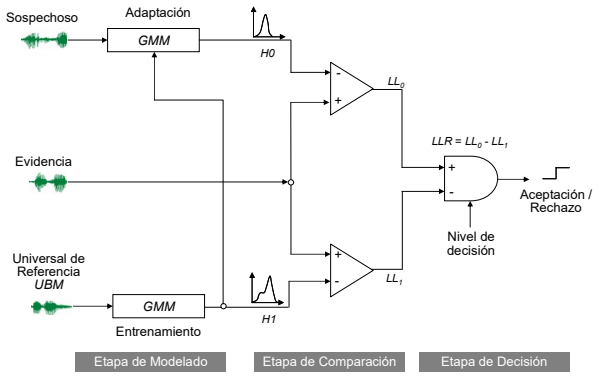

Figura 1. Esquema simplificado de la metodología de reconocimiento de hablante *GMM-UBM*.

D. Metodología HMM-UBM

Esta metodología consiste en el modelado de la información acústica presente en los fonemas por medio del empleo de un sistema de reconocimiento automático de habla (RAH). En la Fig. 2 se puede ver un esquema simplificado de la metodología. Para ello se emplea la metodología de alineación forzada entre las emisiones y sus transcripciones, aprovechando la posibilidad que nos brinda el entorno forense de poder realizar la transcripción fonética de las emisiones a comparar. Esta metodología ha logrado brindar mejores resultados que los obtenidos con la metodología *GMM-UBM* [26].

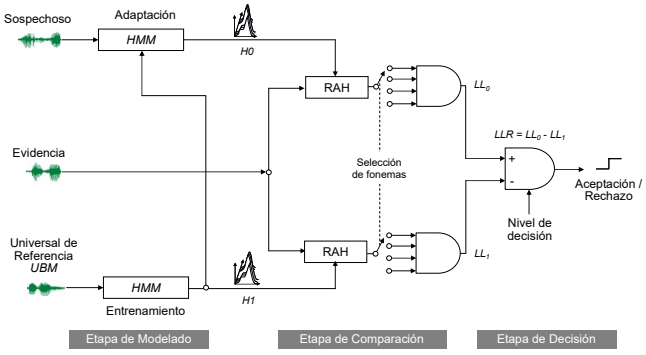

Figura 2. Esquema simplificado de la metodología de reconocimiento de hablante *HMM-UBM*.

Una variación a la metodología *HMM-UBM* es la denominada *DHMM-UBM* [27], que considera el aporte de cada fonema de acuerdo a su factor de discriminación. Dichos factores son calculados en base a la función de costo del cociente de verosimilitudes logarítmico [72], empleando la metodología *HMM-UBM* con un fonema a la vez.

III. CLASIFICADORES EMPLEADOS EN MINERÍA DE DATOS

El incremento continuo de los volúmenes y variedad de datos en el mundo actual ha llevado a los investigadores al desarrollo de nuevas técnicas y herramientas para intentar obtener información oculta dentro de los mismos, y poder extraer nuevo conocimiento a partir de su análisis. Esta rama de la ciencia, que se denomina minería de datos, extrae ideas del aprendizaje automático, la inteligencia artificial, el reconocimiento de patrones, la estadística y la tecnología de bases de datos. Podemos definir la minería de datos como: la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, desde grandes volúmenes de datos.

Dentro de las tareas de minería de datos encontramos la clasificación, el agrupamiento (*clustering*), el descubrimiento de reglas de asociación, el descubrimiento de patrones secuenciales, la regresión, y la detección de anomalías. La clasificación es un método predictivo que permite identificar el grupo de categorías (o clases) al cual pertenece una nueva observación, a partir de un conjunto de datos de entrenamiento que contienen observaciones (o instancias) cuya categoría es conocida. Usualmente, a los datos de entrenamiento se los divide en un conjunto de desarrollo, para crear el modelo, y un conjunto de datos de validación, usado para corroborar la validez del modelo generado.

Podemos agruparlos, desde el punto de vista del esquema del aprendizaje de los datos, en clasificadores perezosos (*lazy learners*) y clasificadores ansiosos (*eager learners*). Los primeros clasifican la nueva instancia de acuerdo a la clase del dato más cercano, para lo cual emplean funciones de distancia diversas. Se los denomina perezosos porque no generalizan hasta que no es necesario. En cambio, los clasificadores ansiosos generalizan antes de tener que clasificar; emplean los datos para construir previamente un modelo y luego lo aplican a la nueva instancia para su clasificación final.

Desde el punto de vista de la cantidad de clasificadores empleados, podemos encontrar métodos simples, que emplean un único clasificador, y métodos de ensamble que combinan múltiples algoritmos de aprendizaje para obtener mejores resultados de predicción que con esos algoritmos en forma individual.

En la Tabla I podemos ver un listado de los clasificadores incluidos en el paquete de algoritmos de minería de datos WEKA desarrollado por la universidad de Waikato (Nueva Zelanda) [34] y empleados en el presente trabajo. Se presenta un breve resumen de las características de cada uno y se adiciona una lista de referencias, para un detalle pormenorizado.

TABLA I. DESCRIPCIÓN DE LOS CLASIFICADORES INCLUIDOS EN EL PAQUETE DE ALGORITMOS DE MINERÍA DE DATOS WEKA Y REFERENCIAS BIBLIOGRÁFICAS AMPLIATORIAS. SE SEÑALAN LOS CLASIFICADORES PEREZOSOS(+), LOS ANSIOSOS (✓) Y LOS QUE EMPLEAN MÉTODOS DE ENSAMBLE (*).

Nº	Denominación del clasificador según WEKA	Descripción	Ref
1	*Ibk* +	k-Vecinos más cercanos	[36]
2	*K Star* +	Clasificador perezoso, con función de distancia basada en la entropía	[37]
3	*LWL* +	Aprendizaje ponderado localmente. Asigna ponderaciones con un clasificador perezoso y luego construye un clasificador a partir de ellas.	[38]
4	*Bayes Net* ✓	Clasificador basado en redes bayesianas.	[39]
5	*Classification Via Regression* ✓	Clasificador basado en métodos de regresión.	[40]
6	*Decision Stump* ✓	Árbol de decisión de un nivel.	[41]
7	*Decision Table* ✓	Tabla de decisión de simple mayoría.	[42]
8	*J48* ✓	Árbol de decisión empleando algoritmo *C4.5*.	[43]
9	*JRIP* ✓	Clasificador basado en reglas que emplea el algoritmo de inducción *RIPPER*.	[44]
10	*LMT* ✓	Árbol de decisión logístico. Genera funciones logísticas en las hojas del árbol.	[45]
11	*Logistic* ✓	Clasificador basado en un modelo de regresión logística multinomial.	[46]
12	*Multilayer Perceptron* ✓	Red neuronal artificial de perceptrón multicapa.	[47]
13	*Naive Bayes* ✓	Clasificador basado en el método *Naïve Bayes*.	[48]
14	*PART* ✓	Clasificador basado en reglas, obtenidas de árboles de decisión parciales, usando *C4.5*.	[49]
15	*REP Tree* ✓	Árbol de decisión rápido.	[50]
16	*SGD* ✓	Emplea el método de descenso de gradiente estocástico para la generación de modelos lineales.	[51]
17	*Simple Logistic* ✓	Clasificador basado en modelos de regresión logística.	[52]
18	*SMO* ✓	Máquina de soporte vectorial (*SVM*)	[53]
19	*Voted Perceptron* ✓	Red neuronal artificial de perceptrón con vectores pesados por votación.	[54]
20	*AdaBoost* *	Clasificador basado en el método de *boosting,* empleando el algoritmo *AdaBoost.M1*, que selecciona aleatoriamente subgrupos de instancias, y combina por votación la salida de varios clasificadores, dándole mayor peso a los más precisos.	[55]
21	*Bagging* *	Clasificador basado en el método de *bagging* (*bootstrap aggregating*), que selecciona aleatoriamente subgrupos de instancias con repetición, y combina por votación la salida de varios clasificadores, dándole el mismo peso a todos.	[56]
22	*LogitBoost* *	Algoritmo de *boosting,* que emplea regresión logística aditiva.	[57]
23	*Random Comittee* *	Construye un ensamble aleatorio de clasificadores. La predicción final es el promedio de las predicciones de cada uno.	[35]
24	*Random Forest* *	Construye un ensamble de árboles de decisión (bosque), seleccionando aleatoriamente subgrupos de instancias.	[58]
25	*Random Sub Space* *	Construye un ensamble de árboles de decisión (*RandomTree*), seleccionando aleatoriamente subgrupos de atributos (*bagging*).	[59]
26	*Random Tree* *	Árbol de decisión con atributos seleccionados aleatoriamente en cada nodo.	[35]
27	*Stacking* *	Combina varios, y en general diferentes, clasificadores empleando el método de apilado (*stacking*). En lugar del procedimiento de votación emplea un algoritmo (*meta-learner*) que combina los clasificadores en función de la confiabilidad de predicción individual.	[60]
28	*Vote* *	Combina varios, y en general diferentes, clasificadores empleando diferentes reglas de combinación en base a las estimaciones de probabilidad de predicción.	[61]

IV. METODOLOGÍA

La metodología propuesta en el presente trabajo consiste en comparar los resultados obtenidos con diferentes clasificadores para la identificación de hablantes. El conjunto de datos a emplear está conformado por emisiones de diferentes hablantes (sospechosos) y compuesta por una serie de atributos (parámetros acústicos). La clase, a ser determinada por el clasificador, puede ser verdadera (la emisión del hablante coincide con la de la evidencia) o falsa (la emisión del hablante no coincide con la de la evidencia).

El atributo de cada parámetro acústico está conformado por el cociente entre la semejanza de dicho parámetro entre la emisión del sospechoso y la evidencia, y entre la semejanza del mismo parámetro, entre la evidencia y el *UBM*. Si la clase es verdadera se obtiene un atributo con valores elevados, y contrariamente si la clase resultante es falsa. Podemos representar al atributo correspondiente a cada parámetro acústico según la siguiente expresión:

$$Atributo\ (j) = \frac{f_j(s,e)}{f_j(e,UBM)} \tag{1}$$

Donde *Atributo* es el atributo de un parámetro acústico, *j* representa el parámetro acústico correspondiente, f_j la función de semejanza o similitud, *s* la emisión del sospechoso, *e* la de la evidencia y *UBM* la del modelo de referencia.

E. Atributos del conjunto de datos

En el presente trabajo se emplean los cocientes de verosimilitudes logarítmicas (LLR) de las metodologías GMM, HMM y DHMM, descritas anteriormente, como parte de los atributos del conjunto de datos. Dichos atributos representan los parámetros segmentales del habla, definidos como aquellas unidades discretas que pueden ser identificadas dentro en la secuencia de habla (fonemas) [30]. Cada LLR tiene asociado la condición de que la evidencia pertenece, o no, al sospechoso correspondiente. Es decir, cada LLR llevará asociado una clase binaria (verdadero o falso) que permitirá configurar el clasificador final.

Si se considera el atributo en valores logarítmicos, la ecuación (1) queda expresada de la siguiente manera:

$$Atributo(j) = \log\frac{L_0(j)}{L_1(j)} = LL_0(j) - LL_1(j) = LLR(j) \tag{2}$$

Donde *Atributo* es el atributo de un parámetro acústico, *j* representa el parámetro acústico segmental, L_0 la función de similitud, que en este caso es la verosimilitud entre el sospechoso y la evidencia, L_1 la verosimilitud entre la evidencia y el *UBM*, *LL* la verosimilitud logarítmica y *LLR* el cociente de verosimilitudes logarítmicas.

El resto de los atributos del conjunto de datos a emplear, quedó conformado por una serie de parámetros suprasegmentales, definidos como aquellos que extienden su dominio por más de un segmento de habla [31] (sílabas, palabras, frases, etc.). Lehiste [32] agrupa estos parámetros como: tonales o de tono, de cantidad o duración, y de acento; y propone otros parámetros, asociados con los anteriores, que agrupa dentro del paralenguaje, entre los que se encuentran la calidad de voz y la vocalización. En la Tabla II pueden verse todos los atributos considerados en el presente trabajo y su agrupación de acuerdo al tipo de parámetro acústico.

TABLA II. CLASIFICACIÓN DE LOS ATRIBUTOS EMPLEADOS PARA LA GENERACIÓN DEL CONJUNTO DE DATOS, DE ACUERDO AL TIPO DE PARÁMETRO ACÚSTICO.

Nº	Atributo	Parámetro Segmental	Parámetro Suprasegmental			
			Duración	Tono	Acento	Calidad de voz
1	*HMM*	✓				
2	*GMM*	✓				
3	*DHMM*	✓				
4	*F0 Medio*			✓		
5	*F0 Mediana*			✓		
6	*F0 mín*			✓		
7	*F0 Espectro*			✓		
8	*F0 pendiente*			✓		
9	*F0 desvío*			✓		
10	*F0 máx*			✓		
11	*F0 máx-mín*			✓		
12	*HNR*					✓
13	*Jitter*					✓
14	*Shimmer*					✓
15	*Grado*					✓
16	*varco ΔC*		✓			
17	*V%*		✓			
18	*ΔC*		✓			
19	*Dur. fonema*		✓			
20	*Dur. palabra*		✓			
21	*Acento*				✓	

El atributo correspondiente a cada parámetro suprasegmental se calculó a partir de la ecuación (2), y las siguientes funciones de semejanza o similitud $L_0(j)$ y $L_1(j)$:

$$L_0(j) = 1 - \left|\frac{p_e(j)}{p_s(j)} - 1\right| \qquad \forall\ p_s(j) \geq 0{,}5.p_e(j) \tag{3a}$$
$$L_0(j) = 0 \qquad \forall\ p_s(j) < 0{,}5.p_e(j)$$

$$L_1(j) = 1 - \left|\frac{p_e(j)}{p_{UBM}(j)} - 1\right| \qquad \forall\ p_{UBM}(j) \geq 0{,}5.p_e(j) \tag{3b}$$
$$L_1(j) = 0 \qquad \forall\ p_{UBM}(j) < 0{,}5.p_e(j)$$

Donde *Atributo* es el atributo de un parámetro acústico, *j* representa el parámetro acústico suprasegmental, y p_e , p_s y p_{UBM} los valores promedios de los parámetros correspondientes a la evidencia, el sospechoso y el *UBM*, respectivamente. La función de similitud así expresada, es el complemento del valor absoluto de la distancia entre el parámetro promedio de la evidencia y del sospechoso/UBM, normalizada con respecto a estos últimos. Estas funciones se encuentran acotadas entre [0;1], de igual manera que las funciones de similitud de los parámetros segmentales; aunque

no representan valores de probabilidad como es el caso de las verosimilitudes.

Los parámetros suprasegmentales empleados en el presente trabajo fueron los siguientes:

1) Tono

Se emplearon los siguientes parámetros tonales, relacionados con la frecuencia fundamental (*F0*), y calculados en los segmentos sonoros de las frases de cada hablante:
F0 Medio (Hz): valor medio de *F0*.
F0 Mediana (Hz): valor de la mediana de *F0* expresada en Hz.
F0 mín (Hz): valor mínimo de *F0* expresada en Hz.
F0 Espectro (%): distribución normalizada de *F0* en 15 bandas de frecuencias en escala de Mel, de 75 a 600 Hz.
F0 pendiente (Hz/mseg): pendiente promedio de la variación temporal de *F0*.
F0 desvío (Hz): desviación estándar de *F0*.
F0 máx (Hz): valor máximo de *F0*.
F0 máx-mín (Hz): diferencia entre *F0 máx* y *F0 mín*.

2) Calidad de voz

La calidad de voz incluye factores tales como: el rango y el control del tono, la resonancia, el control de la articulación, y la velocidad del habla (*tempo*).

El diagnóstico médico de las disfunciones de la voz requiere de la clasificación y valorización de las diferentes calidades de voz. Dicha medición se realiza por medio de juicios perceptuales y medidas objetivas, tales como características acústicas y aerodinámicas, para lo cual existe una gran variedad de técnicas. Por ejemplo, el índice de perturbación, que mide el riesgo vocal [77], emplea los parámetros *Jitter*, *Shimmer*, *HNR,* y la amplitud del cepstrum. Una de las técnicas perceptuales más empleadas es la propuesta por la Sociedad Japonesa de Logopedia y Foniatría (*Japan Society of Logopedics and Phoniatrics*), conocida como escala GRBAS (G de *Grade* o grado, R de *Roughness* o aspereza, B de *Breathiness* o soplido, A de *Astheny* o astenia, y S de *Strain* o forzada). La correlación entre las medidas objetivas y las perceptuales de la escala GRBAS ha sido profundamente estudiada, pero aún no se ha llegado a un acuerdo general entre los investigadores. Por ejemplo, Dejonckere *et al.* [28] determinaron que los parámetros con mayor correlación son: G con el cociente entre la perturbación de la amplitud (*Shimmer)* y la armonicidad *(HNR);* R con la perturbación de la frecuencia fundamental (*Jitter*); y B con el *Shimmer*. El trabajo de Martin *et al.* [74] relacionó R con *HNR* y *Shimmer*, y B con una combinación de *Jitter*, *Shimmer* y *HNR*. Por otra parte Bhuta *et al.* [75] encontraron que R estaba sólo correlacionado con HNR. Estas diferencias entre los distintos autores respecto a la asociación perceptual, muestran, sin embargo, una coincidencia entre los parámetros objetivos empleados. En base a estos resultados, en el presente trabajo, se emplearon los siguientes parámetros para representar la calidad de voz:

Jitter (%): fluctuación irregular del *F0*, calculada como la diferencia absoluta promedio entre períodos consecutivos, dividido por el período promedio.
Shimmer (%): fluctuación de la amplitud, calculada como la diferencia absoluta promedio entre las amplitudes de períodos consecutivos, dividido por la amplitud promedio.
Grado (%/dB): relación entre *Shimmer* y *HNR*.
HNR (dB): armonicidad o relación entre la periodicidad de la emisión y el ruido.

3) Duración

a) Ritmo

Los siguientes parámetros se seleccionaron de acuerdo a lo propuesto por Ramus *et al.* [29] para identificar las características acústicas confiables que pudieran representar el ritmo, basadas en la segmentación entre vocales y consonantes y al aporte de Dellwo [33] que adicionó un coeficiente de variación, independiente de la velocidad de habla.

V% (%): porcentaje de la duración de los intervalos vocálicos con respecto a la duración total de la emisión.
ΔC (mseg): desviación estándar de la duración del intervalo consonántico.
varco ΔC (%): coeficiente de variación de Pearson de la duración del intervalo consonántico o cociente entre *ΔC* y el valor promedio de la duración del intervalo consonántico.

b) Duración intrínseca

Se consideraron las duraciones promedio de los segmentos fonémicos y de las palabras dentro de la frases de cada hablante.

Dur. fonema (mseg): Duración promedio de los fonemas.
Dur. palabra (mseg): Duración promedio de las palabras.

4) Acento

Considerando que la percepción del acento léxico o prosódico (*stress*) es influenciada por las variaciones de la intensidad, la duración y el tono en la vocal acentuada con respecto a la no-acentuada [32], se empleó la siguiente ecuación para representar a este parámetro:

$$Acento(\%) = \frac{I_s}{I_u} \times \frac{t_s}{t_u} \times \frac{F0_s}{F0_u} \times 100 \quad (4)$$

Donde *I* es la intensidad promedio, *t* la duración promedio, y *F0* la frecuencia fundamental promedio. Y donde *s* corresponde a los segmentos vocálicos acentuados y *u* a los no-acentuados, dentro de los cuales se calculan los valores promedios anteriores.

F. Configuración de los clasificadores

Los 28 clasificadores empleados son los descritos en la sección III. Las configuraciones de los mismos fueron

seleccionadas, tras sucesivas pruebas, por ser las de mejor rendimiento, partiendo de las configuraciones estándar provistas por WEKA, y consideraron la totalidad de los parámetros. Los clasificadores de ensamble *LWL*, *AdaBoost*, *Bagging*, *RandomComittee* y *RandomSubSpace* utilizaron *RandomForest* como clasificador de base. Los clasificadores de ensamble *Stacking* y *Vote* emplearon como clasificadores de base una combinación de diferentes metodologías de aprendizaje: *SMO*, *Naive Bayes*, *LWL* y *J48*. Para el caso de *Stacking* el *meta-learner* también fue el clasificador *RandomForest*. El clasificador *SMO*, basado en *SVM*, empleó un función núcleo (*kernel*) universal basado en la función Pearson VII, modelos logísticos en las salidas y no se empleó el filtro de datos de entrenamiento (normalización / estandarización). Para el caso de *LogitBoost* se seleccionó como clasificador el árbol de decisión M5P, basado en el algoritmo M5 inventado por Quinlan [62].

G. Métrica de evaluación del rendimiento de los clasificadores

La métrica de evaluación a emplear en el presente trabajo es *ERR,* la tasa de igual error (*equal error rate*), que es la tasa de error en la que el porcentaje de falsas alarmas (*FA%*) es igual al porcentaje de casos perdidos (*CP%*). Se emplea esta métrica dada su extensiva utilización en los trabajos de reconocimiento e identificación de hablantes, y para poder comparar resultados.

El cálculo de *ERR* se realizó a partir de la curva de característica operativa del receptor *ROC* (*Receiver Operating Characteristic*), que es una representación gráfica de la sensibilidad de un clasificador binario según se varía el umbral de discriminación, y en el que se representa la tasa de verdaderos positivos (*VP%*) con respecto a la tasa de falsos positivos (*FP%*).

Partiendo de la matriz de confusión que representa los resultados de la clasificación:

		Predicción de clase: Si	Predicción de clase: No	
Clase real	Si	***VP***	***FN***	*c=VP+FN*
	No	***FP***	***VN***	*d=FP+VN*
		a=VP+FP	*b=FN+VN*	*c+d=a+b*

(5)

Donde *VP* son los verdaderos positivos, *FP* los falsos positivos, *FN* los falsos negativos y *VN* los verdaderos negativos. De los cuales se derivan las tasas *VP% = VP / c* y *FP% = FP / d* empleadas en las curvas *ROC.*

El porcentaje de falsas alarmas (*FA%*) empleadas en la definición de *ERR* se corresponde con *FP% = FP / d* y el de casos perdidos (*CP%*) con *FN% = FN / c* .

Partiendo de la definición de la tasa de igual error que plantea que *FA% = CP%* , y reemplazando las anteriores correspondencias se obtiene la siguiente ecuación:

$$\frac{FP}{d} = \frac{FN}{c} \tag{6}$$

Pero como por definición de la matriz de confusión (5):

$$c = VP + FN \tag{7}$$

Despejando *FN* de (7) y reemplazando en (6) nos queda:

$$\frac{FP}{d} = \frac{FN}{c} = 1 - \frac{VP}{c} \tag{8}$$

Reordenando y renombrando las variables llegamos a la ecuación final que relaciona la *ERR* con la curva *ROC*:

$$VP\% = 1 - FP\% \tag{9}$$

En la Fig. 3 podemos ver tres ejemplos de la aplicación desarrollada por Damasceno [63] para WEKA calcula el *EER* a partir de la curva *ROC* empleando la ecuación (9).

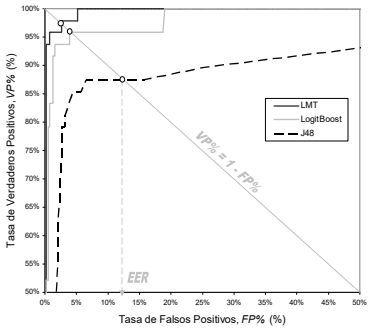

Figura 3. Cálculo del *EER* a partir de las curvas *ROC* para los clasificadores (a) *LMT*, (b) *LogitBoost*, y (c) *J48*, para la base de datos de hombres.

H. Método de estimación

El método de estimación empleado fue la validación cruzada de 10 particiones, la cual consiste en realizar particiones múltiples de los datos para luego estimar el error en base al promedio de las mismas.

I. Preprocesamiento de datos

Los datos fueron previamente normalizados empleando la técnica *z-norm*, de manera tal de reducir los efectos introducidos por las diferencias de canal entre las diferentes emisiones. Para ello se utilizó un grupo de instancias correspondientes a impostores, en el momento de crear los modelos de cada hablante, de las cuales se obtuvieron las distribuciones correspondientes a cada modelo, para finalmente modificar el resultado de cada atributo de acuerdo a la siguiente ecuación:

$$Atributo'(j) = \frac{Atributo(j) - \mu_i(j)}{\sigma_i(j)} \tag{10}$$

Donde *Atributo* es el atributo de un parámetro acústico, *j* representa el parámetro acústico correspondiente, μ_i el valor promedio del *Atributo* correspondiente al modelo del hablante de interés empleando datos de impostores, y σ_i es la desviación estándar de dicho *Atributo*, para los datos de impostores. *Atributo'* es el valor normalizado y que formará parte del conjunto de datos a ser empleado por los diferentes clasificadores.

V. BASE DE DATOS

La base de datos empleada es parte del Proyecto SALA I (*SpeechDat Across Latin America*) [64]. Las grabaciones fueron realizadas a través de la red fija de telefonía. El corpus de SALA I fue dividido en cinco regiones dialectales de Argentina. En el presente trabajo se emplearon exclusivamente emisiones de la región sur, las cuales incluyen 136 hablantes (47 hombres y 89 mujeres). Para el entrenamiento del *UBM* se utilizaron 118 hablantes, los restantes 18 hablantes (9 hombres y 9 mujeres) fueron empleados para construir los modelos de entrenamiento y constituir el corpus de prueba. Con la finalidad de incorporar diferentes sesiones y canales en el corpus de prueba –en forma similar a lo que ocurre generalmente en los casos forenses-, se adicionaron grabaciones de 6 hablantes (3 hombres y 3 mujeres) pertenecientes al grupo anterior. Las nuevas sesiones de grabación tuvieron lugar 10 años después de las originales y en esa ocasión se empleó un micrófono de laptop en lugar del canal telefónico empleado durante la grabación original de SALA I.

El corpus de prueba quedó finalmente conformado por 864 emisiones correspondientes a 18 hablantes (96 emisiones de los hablantes verdaderos y 768 emisiones de los impostores). En la Fig. 4 puede verse la distribución de los mismos en un plano de vectores reducidos, resultantes del análisis de componentes principales.

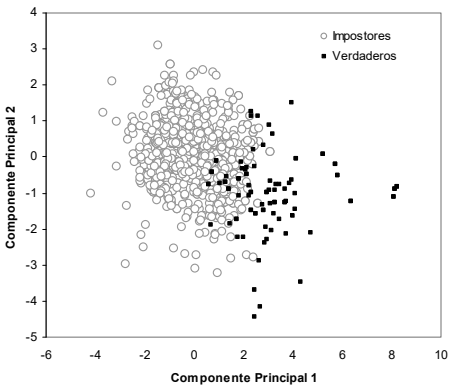

Figura 4. Distribución de los vectores reducidos, resultantes del análisis de componentes principales, para los casos de (a) impostores y (b) verdaderos.

VI. MODELOS SEGMENTALES Y SUPRASEGMENTALES

Los modelos segmentales (*GMM*, *HMM* y *DHMM*) se generaron con la herramienta, para la creación de sistemas automáticos de habla basados en modelos ocultos de Markov, desarrollada por la Universidad de Cambridge: HTK Toolkit ver. 3.4 [65]. La parametrización de la señal acústica se realizó empleando una frecuencia de muestreo de 8 KHz, 16 bits de resolución y sustracción de la media temporal, para eliminar cualquier nivel de continua proveniente de la etapa de adquisición. Se utilizaron ventanas de análisis del tipo Hamming de 25 mseg de duración y 10 mseg de avance, filtro de preénfasis con coeficiente 0.97, y normalización de la energía a nivel de frase. Se codificó cada ventana de la señal empleando 13 coeficientes cepstrales *MFCC* (*Mel-Frecuency Cepstral Coefficients*) a los cuales se les adicionó los coeficientes delta y aceleración (derivadas temporales de primer y segundo orden), conformando un total de 39 parámetros.

Los modelos *GMM* quedaron conformados por un único modelo oculto de Markov de un solo estado y para los modelos *HMM* y *DHMM* se crearon 30 modelos ocultos de Markov, de 3 estados cada uno, correspondientes a cada uno de los fonemas presentes en el español de Argentina, de acuerdo al alfabeto fonético SAMPA (*Speech Assesment Methods: Phonetic Alphabet)*, adaptado para el Español de Argentina [66].

A los modelos anteriores se les adicionó un modelo de silencio, presente generalmente entre frases, al cual se le asoció un modelo de pausa corta entre palabras. El mismo se utilizó para segmentar automáticamente los segmentos de habla de los de silencio.

Los modelos suprasegmentales se crearon a partir del software de análisis y síntesis de señales de habla PRATT [67], empleándose para todas las mediciones la configuración estándar. Para la determinación de segmentos vocálicos y consonánticos (parámetros de ritmo), vocales acentuadas y no-acentuadas (parámetros de acento) y fonemas y palabras (parámetros de duración) se aprovechó la segmentación creada por el sistema de reconocimiento automático de habla, utilizado para la creación de los modelos segmentales. Para mejorar la performance de dicha segmentación, se empleó la metodología de alineamiento forzado. Previo al alineamiento automático, se realizaron transcripciones manuales a nivel de palabra de las emisiones de cada hablante, para finalmente ser codificadas por un transcriptor fonético automático, basado en una serie de reglas de conversión de grafemas a fonemas [68].

VII. RESULTADOS

En el primer experimento realizado se emplearon los parámetros en forma individual, para analizar su influencia como discriminadores de hablantes. Para ello se empleó el clasificador *Naive Bayes*, que emplea un único nivel de decisión para la clasificación y es el que generalmente se utiliza en los sistemas de reconocimiento de hablantes basados en *GMM* (Ver Fig. 1). Los resultados obtenidos son los que se muestran en la Tabla III.

Como puede verse en dicha tabla, y corroborando resultados de otros trabajos [12], los parámetros segmentales utilizados en forma individual, son los de mejor rendimiento, seguidos por los parámetros suprasegmentales de tono.

Algunos de estos parámetros poseen un alto factor de correlación de Pearson ($fc > 0{,}8$): *GMM*/*HMM*/*DHMM*, *Grado*/*Jitter*/*Shimmer*, y $F0_{Medio}$/$F0_{Mediana}$/$F0_{Espectro}$, y por lo tanto dependencia estocástica entre dichos parámetros. Igualmente se los empleó a todos, permitiendo que los algoritmos de los clasificadores determinen su uso. Por otra parte, que los parámetros sean dependientes no excluye la posibilidad de que puedan brindar información complementaria para algunos casos particulares (aunque no estadísticamente comprobable), incrementando de esta manera la tasa de reconocimiento de hablantes.

TABLA III. TASA DE IGUAL ERROR PARA UN CLASIFICADOR *NAIVE BAYES*, EMPLEANDO LOS PARÁMETROS EN FORMA INDIVIDUAL, Y LA BASE DE DATOS DE HOMBRES Y MUJERES EN FORMA DISCRIMINADA.

Nro.	Parámetro acústico	Tasa de igual error, *EER* (%)		
		Hombres	Mujeres	Promedio
1	*DHMM*	5,4%	21,9%	13,7%
2	*GMM*	8,8%	20,8%	14,8%
3	*HMM*	10,2%	20,5%	15,4%
4	*F0 Espectro*	23,5%	22,9%	23,2%
5	*F0 Medio*	29,4%	25,8%	27,6%
6	*F0 Mediana*	26,3%	31,4%	28,9%
7	*F0 mín*	30,0%	31,5%	30,7%
8	*HNR*	37,5%	39,6%	38,5%
9	*Grado*	42,9%	41,7%	42,3%
10	*ΔC*	53,1%	32,9%	43,0%
11	*Acento*	40,0%	46,3%	43,2%
12	*Jitter*	35,4%	51,0%	43,2%
13	*varco ΔC*	42,3%	46,5%	44,4%
14	*Dur. Fonema*	50,0%	41,5%	45,7%
15	*Shimmer*	43,8%	47,9%	45,8%
16	*F0 máx*	45,5%	47,5%	46,5%
17	*F0 máx-mín*	44,8%	49,9%	47,3%
18	*Dur. Palabra*	45,8%	51,5%	48,7%
19	*F0 pendiente*	48,1%	50,4%	49,3%
20	*F0 desvío*	54,7%	51,7%	53,2%
21	*V%*	53,9%	58,3%	56,1%

En la Tabla IV se muestran los resultados de la aplicación de los diferentes clasificadores, discriminados por género, y empleando la totalidad de los parámetros.

Comparando las Tablas III y IV puede verse que el empleo de la totalidad de los parámetros mejora sustancialmente el rendimiento (excepto para el clasificador DecisionStump, que es un clasificador muy simple, del tipo: árbol de decisión de un nivel).

TABLA IV. TASA DE IGUAL ERROR PARA CADA UNO DE LOS CLASIFICADORES, EMPLEANDO LA TOTALIDAD DE LOS PARÁMETROS, Y LA BASE DE DATOS DE HOMBRES Y MUJERES EN FORMA DISCRIMINADA.

Nro.	Clasificador	Tasa de igual error, *EER* (%)		
		Hombres	Mujeres	Promedio
1	*LMT*	2,1%	2,1%	2,1%
2	*LogitBoost*	4,2%	0,1%	2,1%
3	*MultilayerPerceptron*	0,1%	4,2%	2,1%
4	*LWL*	4,2%	1,5%	2,9%
5	*Bagging*	4,2%	2,1%	3,1%
6	*SMO*	4,2%	2,1%	3,1%
7	*RandomComittee*	4,2%	2,1%	3,1%
8	*RandomForest*	4,2%	2,1%	3,1%
9	*RandomSubSpace*	4,2%	2,1%	3,1%
10	*SimpleLogistic*	2,1%	4,2%	3,1%
11	*Vote*	4,2%	2,1%	3,1%
12	*Stacking*	4,2%	2,3%	3,2%
13	*RandomTree*	5,4%	3,2%	4,3%
14	*AdaBoost*	6,3%	3,2%	4,7%
15	*SGD*	4,2%	5,2%	4,7%
16	*ClassificationVia Regression*	6,3%	4,2%	5,2%
17	*K Star*	6,3%	4,2%	5,2%
18	*Logistic*	4,2%	8,3%	6,3%
19	*BayesNet*	3,9%	10,4%	7,2%
20	*JRIP*	14,0%	1,3%	7,6%
21	*PART*	9,1%	6,3%	7,7%
22	*VotedPerceptron*	6,3%	10,4%	8,3%
23	*NaiveBayes*	6,3%	11,7%	9,0%
24	*J48*	12,5%	6,3%	9,4%
25	*Ibk*	14,6%	6,3%	10,4%
26	*REPTree*	16,7%	8,5%	12,6%
27	*DecisionTable*	10,4%	16,7%	13,5%
28	*DecisionStump*	8,3%	31,1%	19,7%

Para los casos reales es bastante poco frecuente poder contar con la cantidad de parámetros empleados en el presente trabajo (21 parámetros). Con lo cual, para analizar la influencia de la cantidad de parámetros a emplear, y poder valorizar la importancia relativa de cada uno de ellos, se realizó un nuevo experimento que consistió en la selección de atributos para luego utilizarlos en cantidades variables en un clasificador *SVM*. Las técnicas empleadas para la selección de atributos fueron: *Tasa de Ganancia*, *Correlación* y *Relief*.

La técnica de *Tasa de Ganancia* evalúa el valor de un atributo midiendo la tasa de ganancia con respecto a la clase, la de *Correlación* midiendo el factor de correlación de Pearson entre el atributo y la clase, y la de *Relief* de acuerdo a qué tan bien distingue entre instancias que están cerca unas de otras [69].

Sabiendo que el rendimiento de los clasificadores *SVM* se ve afectado por el *kernel* utilizado [71], se consideraron varias funciones núcleo, para conocer su influencia en un sistema de reconocimiento de hablantes. Se utilizaron los siguientes *kernels*, implementados en el paquete de algoritmos de minería de datos WEKA: *kernel* polinómico normalizado (*NPK*), *kernel* polinómico (*PK*), *kernel* radial (*RBF*), y *kernel* universal basado en la función Pearson VII (*puk*) [70].

En las Fig. 5, 6 y 7 se muestran los resultados para distintas cantidades de atributos, técnicas de selección y funciones núcleo.

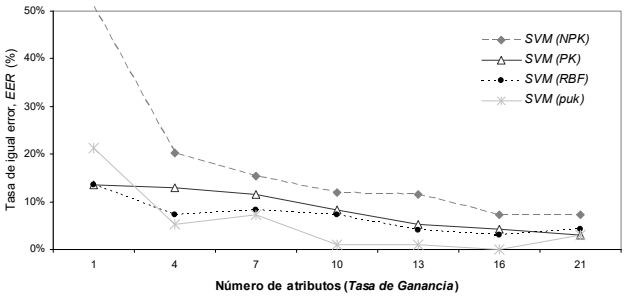

Figura 5. Tasa de igual error promedio (hombres y mujeres) en función del número de atributos seleccionados con la metodología de *Tasa de Ganancia*. Los clasificadores empleados fueron máquinas de soporte vectorial (*SVM*) con diferentes *kernels*: (a) polinomial normalizada (*NPK*), (b) polinomial (*PK*), (c) radial (*RBF*), y (d) Pearson VII (*puk*).

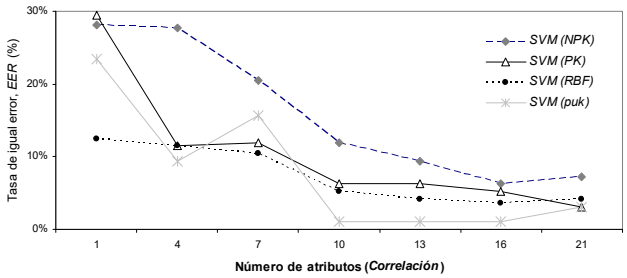

Figura 6. Tasa de igual error promedio (hombres y mujeres) en función del número de atributos seleccionados con la metodología de *Correlación*. Los clasificadores empleados fueron máquinas de soporte vectorial (*SVM*) con diferentes *kernels*: (a) polinomial normalizada (*NPK*), (b) polinomial (*PK*), (c) radial (*RBF*), y (d) Pearson VII (*puk*).

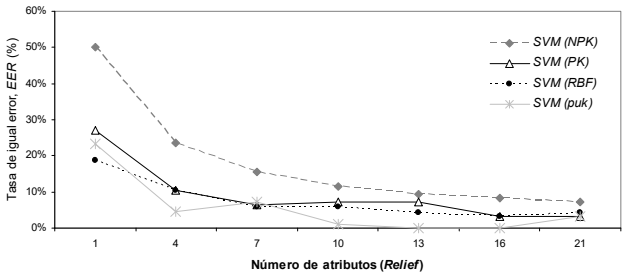

Figura 7. Tasa de igual error promedio (hombres y mujeres) en función del número de atributos seleccionados con la metodología *Relief*. Los clasificadores empleados fueron máquinas de soporte vectorial (*SVM*) con diferentes funciones *kernel*: (a) polinomial normalizada (*NPK*), (b) polinomial (*PK*), (c) radial (*RBF*), y (d) Pearson VII (*puk*).

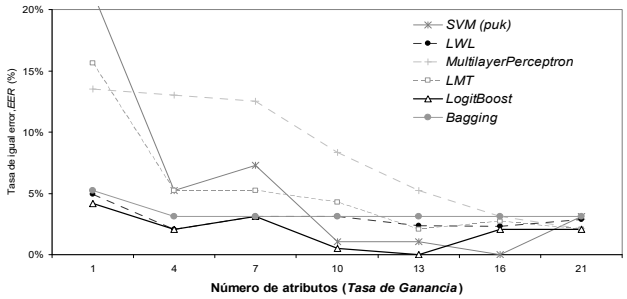

Figura 8. Tasa de igual error promedio (hombres y mujeres) en función del número de atributos seleccionados con la metodología *Tasa de Ganancia* para los principales clasificadores: (a) SVM, (b) LWL, (c) MultilayerPerceptron, (d) LMT, (e) LogitBoost, y (f) Bagging.

Como era de esperar, el uso de menor cantidad de parámetros degrada al sistema de reconocimiento de hablantes. El comportamiento del *kernel puk*, para todas las técnicas de selección de atributos es el que mejores resultados promedio brinda (Tabla V). Comparando las diferentes técnicas de selección de atributos, como puede verse en dicha tabla, la Tasa de Ganancia es la de mejor rendimiento entre las propuestas.

TABLA V. TASA DE IGUAL ERROR PROMEDIO PARA MÁQUINAS DE SOPORTE VECTORIAL (*SVM*), CON DIFERENTES *KERNELS* Y TÉCNICAS DE SELECCIÓN DE ATRIBUTOS. LOS VALORES DE *EER* SE PROMEDIARON PARA TODAS LAS CANTIDADES DE ATRIBUTOS EMPLEADAS (ENTRE 1 Y 21 ATRIBUTOS).

Técnica de selección de atributos	$EER_{promedio}$ **(%)**			
	NPK	***PK***	***RBF***	***puk***
Tasa de Ganancia	17,82%	8,43%	6,83%	5,58%
Correlación	15,89%	10,54%	7,37%	7,81%
Relief	17,95%	9,24%	7,50%	5,60%
Promedio	17,22%	9,40%	7,23%	6,33%

En base a los resultados obtenidos para el clasificador *SVM*, se empleó la técnica de selección de atributos de Tasa de Ganancia para analizar la influencia de la cantidad de parámetros a ser empleados por el resto de los clasificadores. Para ello, se compararon los seis clasificadores con mejores resultados promedio (hombres y mujeres) de la Tabla IV: *LMT* (árbol de decisión), *LogitBoost* (ensamble), MultiLayerPerceptron (red neuronal), *LWL* (clasificador perezoso con clasificador secundario tipo ensamble), *Bagging* (ensamble) y *SMO* (*SVM* con *kernel puk*). En la Fig. 8 pueden verse los resultados comparativos.

Es de notar que los clasificadores de ensamble (*LogitBoost*, *Bagging* y *LWL*) logran resultados muy acotados ($\Delta_{EER} < 5\%$), de acuerdo a la cantidad de atributos considerados, mientras que los clasificadores individuales (*LWL*, *MultilayerPerceptron* y *SVM*) poseen variaciones muy importantes ($\Delta_{EER} > 20\%$).

Para poder calificar el comportamiento de todos los clasificadores, se realizó un experimento adicional, empleando los primeros 7 atributos seleccionados con la técnica de *Tasa de Ganancia* (Tabla VI).

TABLA VI. TASA DE IGUAL ERROR PARA CADA UNO DE LOS CLASIFICADORES, EMPLEANDO LOS PRIMEROS 7 ATRIBUTOS SELECCIONADOS CON LA TÉCNICA DE *TASA DE GANANCIA*, Y LA BASE DE DATOS DE HOMBRES Y MUJERES EN FORMA DISCRIMINADA.

Nro.	**Clasificador**	**Tasa de igual error, *EER* (%)**		
		Hombres	**Mujeres**	**Promedio**
1	*Bagging*	4,2%	2,1%	3,1%
2	*LogitBoost*	4,2%	2,1%	3,1%
3	*LWL*	4,2%	2,1%	3,1%
4	*RandomComittee*	4,2%	2,1%	3,1%
5	*RandomForest*	4,2%	2,1%	3,1%
6	*RandomSubSpace*	4,2%	2,1%	3,1%
7	*Vote*	4,2%	2,1%	3,1%
8	*AdaBoost*	4,2%	2,1%	3,1%
9	*RandomTree*	5,6%	1,9%	3,8%
10	*Stacking*	4,2%	4,2%	4,2%
11	*K Star*	6,3%	3,1%	4,7%
12	*LMT*	6,3%	4,2%	5,2%
13	*BayesNet*	4,2%	9,1%	6,6%
14	*PART*	7,2%	6,3%	6,7%
15	*ClassificationVia Regression*	6,3%	8,3%	7,3%
16	*SMO*	12,5%	2,1%	7,3%
17	*J48*	9,1%	6,3%	7,7%
18	*Ibk*	10,4%	8,3%	9,4%
19	*NaiveBayes*	6,3%	12,5%	9,4%
20	*JRIP*	12,4%	7,1%	9,7%
21	*Logistic*	6,3%	16,7%	11,5%
22	*SGD*	6,3%	16,7%	11,5%
23	*VotedPerceptron*	6,3%	16,7%	11,5%
24	*SimpleLogistic*	4,2%	18,8%	11,5%
25	*Multilayer Perceptron*	8,3%	16,7%	12,5%
26	*Decision Table*	10,4%	18,8%	14,6%
27	*REPTree*	14,6%	15,3%	14,9%
28	*DecisionStump*	8,3%	31,3%	19,8%

En este caso, los clasificadores de mejor rendimiento fueron los que emplean la metodología de ensamble. Nótese que hasta el puesto décimo todos los clasificadores son de este tipo. El modelo generado con esta nueva metodología fue empleado para clasificar un conjunto de datos de validación, compuesto por emisiones no utilizadas para el diseño del mismo, y que incluyeron 48 emisiones correspondientes a 12 mujeres y 12 hombres, totalizando 576 instancias (48 emisiones verdaderas y 528 emisiones falsas). La precisión (*accuracy*) promedio obtenida con el nuevo modelo fue del 96,2%, frente a un 89,1% del sistema base (*GMM*). Quedando corroborada la validez del nuevo método descrito.

Para valorizar la importancia relativa de cada uno de los parámetros acústicos empleados se calculó la tasa de selección de los diferentes atributos (Fig. 9), como el promedio del valor asignado a cada uno de ellos por las técnicas de selección de atributos empleadas: *Tasa de Ganancia*, *Correlación* y *Relief*. En la Fig. 10 se agruparon los resultados de los atributos que pertenecen a las mismas categorías acústicas: segmental, tono, calidad de voz, duración y acento. La categoría suprasegmental quedaría conformada por estas últimas cuatro.

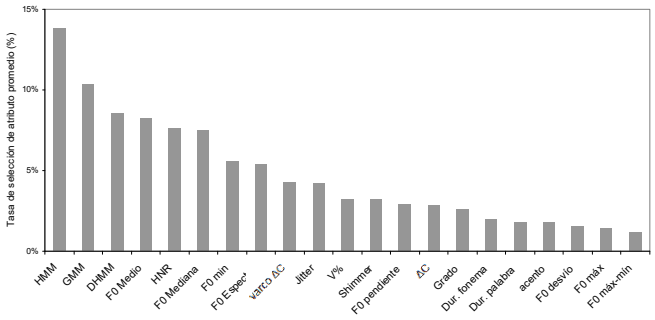

Figura 9. Tasa de selección de los diferentes atributos, promediada entre las metodologías de *Tasa de Ganancia*, *Correlación* y *Relief*.

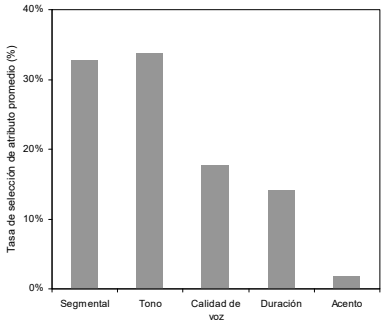

Figura 10. Tasa de selección de atributos, agrupados de acuerdo al tipo de parámetro acústico, promediada entre las metodologías de *Tasa de Ganancia*, *Correlación* y *Relief*.

Los parámetros empleados para identificar hombres y mujeres, como puede verse en la Fig. 11 (considerando los primeros 7 atributos seleccionados por medio de la metodología de *Tasa de Ganancia*), comparten los parámetros segmentales y difieren en los suprasegmentales. Nótese que los hombres emplean parámetros de tono, no utilizados por las mujeres, y comparten los mismos tipos de parámetros de duración y calidad de voz (aunque no los mismos parámetros en sí).

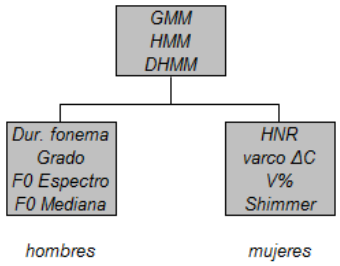

Figura 11. Esquema de los primeros siete atributos, seleccionados por la metodología de *Tasa de Ganancia*, agrupados por género.

VIII. DISCUSIÓN

Como puede verse en la Tabla III, los parámetros segmentales son los de mejor rendimiento, seguidos por los suprasegmentales de tono. A un resultado similar se ha llegado considerando la tasa de selección de atributos (Fig. 9). Analizando la Fig. 10 se puede ver, que en este caso, los parámetros segmentales y los suprasegmentales de tono son los de mayor poder discriminatorio, seguidos por los de calidad de voz y duración, mientras que el acento no realizaría un aporte importante al reconocimiento de hablantes.

De la comparación de las Tablas III y IV se concluye que el empleo de la totalidad de los parámetros produce una reducción de la tasa de igual error de hasta el 84,7 % con respecto al mejor resultado obtenido con un único parámetro (*DHMM*). Este resultado corrobora lo obtenido por Stolcke *et al.* [12] con el reconocedor de hablantes desarrollado en el laboratorio SRI.

La casi totalidad de los clasificadores empleados en el presente trabajo presentan resultados que superan al de un único parámetro, siendo la única excepción la del clasificador *DecisionStump*, que es un clasificador muy simple, del tipo: árbol de decisión de un único nivel.

Los clasificadores que mejor rendimiento obtuvieron, en estas condiciones, fueron: *LMT* (árbol de decisión), *LogitBoost* (ensamble), y *MultiLayerPerceptron* (red neuronal).

Otra conclusión que podemos destacar del análisis de la Tabla IV, es la diferencia en el comportamiento de los clasificadores de acuerdo al género de los hablantes, plasmado en un muy bajo factor de correlación de Pearson (fc = 0,29) entre hombres y mujeres. Desechando al clasificador *DecisionStump* (valor atípico, por lo expuesto anteriormente), el *ERR* promedio de los hombres fue de 6,2%, mientras que el de las mujeres se redujo al 4,9% (es decir un 21,5% mejor que el de los hombres). Lo cual nos estaría mostrando que la distribución de los vectores son diferentes para ambos géneros y más concentrados para el caso de las mujeres. Los parámetros empleados para identificar hombres y mujeres, como puede verse en la Fig. 11 (considerando los primeros 7 atributos seleccionados por medio de la metodología de *Tasa de Ganancia*), comparten los parámetros segmentales y difieren en los suprasegmentales. Los hombres emplean parámetros de tono, que no son utilizados por las mujeres, y

comparten los mismos tipos de parámetros de duración y calidad de voz (aunque no los mismos parámetros en sí).

Los clasificadores de ensamble, que han demostrado un muy buen desempeño en diferentes aplicaciones de minería de datos, también lo son para las tareas de reconocimiento de hablantes. De los clasificadores con un *EER* < 5,0 %, el 73,3% emplean la metodología de ensamble, siendo el total de los clasificadores de ensamble empleados en el presente trabajo sólo el 42,9% del total.

Para el caso de las máquinas de soporte vectorial (*SVM*), ampliamente utilizadas en el reconocimiento de hablantes, se puede ver en la Tabla V que el comportamiento del *kernel* universal basado en la función Pearson VII (*puk*) es el que brinda mejores resultados. Dentro de las diferentes técnicas de selección de atributos empleadas, la *Tasa de Ganancia* es la de mejor rendimiento.

Como era de esperar, el uso de menor cantidad de parámetros degrada al sistema de reconocimiento de hablantes. Pasándose de un $EER_{mín}$ de 2,1% a 3,1% cuando la cantidad se reduce a 7 parámetros (Tablas IV y VI respectivamente). Para estos casos, los clasificadores de ensamble son los de mejor rendimiento. Nótese que, hasta el puesto décimo de la Tabla VI, todos los clasificadores son de este tipo. A esta ventaja comparativa de los clasificadores de ensamble, podemos adicionar que logran resultados más acotados ante la variación de la cantidad de atributos (Δ_{EER} < 5%) que los clasificadores individuales (Δ_{EER} > 20%), como puede extraerse del análisis de la Fig. 8.

La hipótesis originalmente planteada sobre la importancia de la inclusión de parámetros estocásticamente dependientes quedó demostrada por el uso que hicieron los clasificadores de mejor rendimiento de algunos de ellos. Por ejemplo, para el caso del uso restringido de atributos, los 3 parámetros acústicos segmentales (calificados como dependientes en base a los factores de correlación) fueron empleados por todos los mejores clasificadores. La hipótesis anterior quedó corroborada al realizarse un experimento comparativo empleando solamente los parámetros independientes con los 3 mejores clasificadores de la Tabla IV (*LMT*, *LogitBoost* y *MultilayerPerceptron*). Para este caso la *EER* fue del 7,3%, muy por encima del 2,1% obtenido anteriormente con la totalidad de los parámetros.

IX. CONCLUSIONES

Hemos visto que, para las aplicaciones forenses de reconocimiento de hablantes, se pueden adicionar a los sistemas basados en modelos de mezclas gaussianas (*GMM*) otros parámetros segmentales y suprasegmentales, dada la posibilidad de contar con las transcripciones de las emisiones de los hablantes. Estos nuevos parámetros adicionan información relevante de los hablantes y permiten lograr una mejora sustancial en el sistema de reconocimiento.

Por otra parte, vimos que el empleo de clasificadores del tipo de ensamble permite hacer uso de menor cantidad de parámetros acústicos en el clasificador, manteniendo acotada la tasa de reconocimiento. Lo cual permitiría desarrollar sistemas de reconocimiento automático de hablantes más compactos y veloces.

Se ha corroborado que las máquinas de soporte vectorial (*SVM*), normalmente empleadas en el reconocimiento de hablantes, que emplean el *kernel* universal basado en la función Pearson VII, brindan resultados similares a los de los clasificadores de ensamble, cuando se emplean más de 10 parámetros acústicos.

Las redes neuronales (*MultilayerPerceptron*), también empleadas en los reconocedores de hablantes, poseen muy buen desempeño cuando se utilizan todos los parámetros disponibles. Pero su rendimiento disminuye drásticamente al reducirse en cantidad.

Los parámetros acústicos que han demostrado mayor poder discriminatorio han sido los segmentales y los suprasegmentales de tono, corroborando lo destacado en la literatura. La contribución inédita de este trabajo ha sido, el uso de parámetros extralingüísticos como la calidad de voz. Los cuales han demostrado tener un importante poder de discriminación entre hablantes.

La diferencia de parámetros empleados por hombres y mujeres nos lleva a pensar que los sistemas de reconocimiento automático de hablantes deberán, no solo manejar diferentes bases de datos para la construcción de modelos acústicos, sino también diferenciar los parámetros y clasificadores a emplear en cada caso.

X. TRABAJOS FUTUROS

Para contrastar o corroborar los resultados obtenidos para el español de Argentina se utilizarán datos de lengua inglesa para la realización de los experimentos presentados en el presente trabajo. Para ello se utilizarán las bases de datos utilizadas por *NIST* en las evaluaciones de los sistemas automáticos de reconocimiento de hablantes denominadas *SRE* (*Speaker Recognition Evaluation*). El empleo de las mismas permitirá, adicionalmente, comparar directamente los resultados de la metodología propuesta con el estado del arte. Otra alternativa a la base de datos *SRE* es emplear la base de datos, más limitada, pero que permite la intervención humana para generar las transcripciones, de una nueva evaluación de *NIST* denominada *HASR* (*Human Assisted Speaker Recognition*) [73]. La misma está especialmente diseñada para comparar sistemas de reconocimiento de hablantes con intervención humana, como es el caso de las aplicaciones forenses. En el 2010 se realizó una prueba piloto y en el 2012 se realizó una evaluación a la que se presentaron 13 sitios de diversas partes del mundo, entre los cuales participó el Laboratorio de Investigaciones Sensoriales con un sistema basado en la metodología HMM-UBM (La metodología HMM-UBM [27] fue presentada en la evaluación HASR 2012, totalizando 10 errores en los veinte pares de la evaluación, mientras que el mejor sistema presentado produjo un total de 6 errores.).

Dentro de los trabajos futuros pensamos también estudiar el comportamiento de cada parámetro ante diferencias de canales, ruido, sesiones, estados emotivos y envejecimiento del hablante, etc. con la finalidad de determinar parámetros robustos para el reconocimiento de hablantes.

También se incorporarán al sistema propuesto otros parámetros suprasegmentales como el índice de perturbación [77], el fonetograma [78], y los patrones prosódicos [79] y rítmicos [80], que podrían aportar características de discriminación complementarias.

REFERENCIAS

[1] I. Pollack, J. M. Pickett and W. H. Sumby, "On the Identification of Speakers by Voice", *Journal of the Acoustical Society of America*, vol. 26, pp. 403-406, 1954.

[2] J. N. Shearme and J. N. Holmes, "An Experiment Concerning the Recognition of Voices", *Language and Speech*, 2, pp. 123-131, July/September 1959.

[3] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition", *Journal of the Acoustical Society of America*, vol. 35, pp. 354-358, 1963.

[4] G. R. Doddington, "A method of speaker verification", *Journal of the Acoustical Society of America*, vol. 49, 139 (A), 1971.

[5] W. Endres, W. Bambach and G. Flösser, "Voice spectrograms as a function of age, voice disguise, and voice imitation", *Journal of the Acoustical Society of America*, vol. 49(6B), pp. 1842-1848, 1971.

[6] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition", *Electronics and Communications in Japan*, 57-A, pp. 34-41, 1974.

[7] J. M. Naik, L. P. Netsch and G. R. Doddington, "Speaker verification over long distance telephone lines", in *Proc. ICASSP Acoustics, Speech, and Signal Processing,* pp. 524-527, 1989.

[8] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation", in *Proc. ICASSP Acoustics, Speech, and Signal Processing,* pp. 293-296, 1990.

[9] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", in *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 27-30, 1994.

[10] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring", in *Proc. ICASSP*, *Acoustics, Speech, and Signal Processing,* vol. 1, pp. 595-598, 1998.

[11] G. R. Doddington, "Speaker recognition based on idiolectal differences between speakers", in *Proc. Eurospeech*, pp. 2521-2524, 2001.

[12] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez and G. Tur, "Speech Recognition as Feature Extraction for Speaker Recognition", in *Proc. of the IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, pp. 39-43, 2007.

[13] J. Oglesby and J. S. Mason, "Optimization of neural models for speaker identification", in *Proc. ICASSP Acoustics, Speech, and Signal Processing*, vol. 1, pp. 261-264, 1990.

[14] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification", in *Neural Networks Signal Process, Proc. 2000 IEEE Signal Processing Workshop*, Vol. 2, pp. 775-784, 2000.

[15] S. Furui, "Acoustic and Speech Engineering", *Kindai Kagaku-sha Publishing Company*, Tokyo, 1992.

[16] National Research Council, "On the theory and practice of voice identification", National Academy of Science, Washington, pp. 3-13, 1979.

[17] L. G. Kersta, "Voiceprint identification", *Nature*, vol. 196, no. 4861, pp. 1253-1257, 1962.

[18] K. N. Stevens, C. E. Williams, J. R. Carbonell and B. Woods, "Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material", *Journal of the Acoustical Society of America, vol.* 44, pp. 1596-1607, 1968.

[19] J. F. Bonastre, F. Bimbot, L. J. Boë, J. P. Campbell and D. A. Reynolds, I. Magrin-Chagnolleau, "Person Authentication by Voice: A Need for Caution", in *Proc. Eurospeech*, pp. 1-4, 2003.

[20] P. French and P. Harrison, "Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by Peter French & Philip Harrison", *International Journal of Speech Language and the Law*, vol. 14, no. 1, pp. 137-144, 2007.

[21] P. Rose and G. Morrison, "A response to the UK position statement on forensic speaker comparison", *International Journal of Speech Language and the Law*, vol. 16, no. 1, pp. 139, 2009.

[22] P. French, F. Nolan, P. Foulkes, P. Harrison and K. McDougall, "The UK position statement on forensic speaker comparison; a rejoinder to Rose and Morrison", *International Journal of Speech Language and the Law*, vol. 17, no. 1, pp. 143-152, 2010.

[23] B. Koenig, Federal Bureau of Investigation and US Dept of Justice, "Speaker Identification-Part 2-Results of the National Academy of Sciences'study", FBI Law Enforcement Bulletin, vol 49, no. 2, pp. 20-22, 1980.

[24] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A study of interspeaker variability in speaker verification", *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980-988, 2008.

[25] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification", *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788-798, 2011.

[26] P. Univaso, M. Martínez Soler and J. A. Gurlekian, "A preliminary approach to forensic speaker recognition using phonemes", en *IberSPEECH 2012, VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*, Universidad Autónoma de Madrid, Madrid, España, pp. 21-23, 2012.

[27] P. Univaso, M. Martínez Soler and J. A. Gurlekian, "Human Assisted Speaker Recognition Using Forced Alignments on HMM", *International Journal of Engineering*, 2(9), 2013.

[28] P. H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements", *Revue de laryngologie-otologie-rhinologie*, 117(3), 219, 1996.

[29] F. Ramus, M. Nespor and J. Mehler, "Correlates of linguistic rhythm in the speech signal", *Cognition*, 73(3), pp. 265-292, 1999.

[30] D. Crystal, "A Dictionary of Linguistics & Phonetics", Oxford, England: *Blackwell*, 2003.

[31] E. Hamp, "A Glossary of American Technical Linguisitic Usage, 1925-1950", *Utrecht-Antwerp: Spectrum Publishers*, 1957.

[32] I. Lehiste, "Suprasegmentals", M.I.T. Press, *Cambridge, MA*, 1970.

[33] V. Dellwo, "Rhythm and Speech Rate: A variation coefficient for ΔC", *Language and language-processing*, pp. 231-241, 2006.

[34] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Volume 11, Issue 1, 2009.

[35] I. Witten, E. Frank and M. Hall, "Data Mining: Practical machine learning tools and techniques", *Morgan Kaufmann*, 3rd ed., 2011.

[36] D. Aha and D. Kibler, "Instance-based Learning", *Machine Learning*, 6, pp. 37-66, 1991.

[37] J. Cleary and L. Trigg, "K*: An Instance-based Learner Using an Entropic Distance Measure", in *12th International Conference of Machine Learning*, pp. 108-114, 1995.

[38] E. Frank, M. Hall and B. Pfahringer, "Locally Weighted Naïve Bayes", in *19th Conference of Uncertainty in Artificial Intelligence*, pp. 249-256, 2003.

[39] R. R. Bouckaert, "Bayesian network classifiers in Weka", *Department of Computer Science, University of Waikato*, 2004.

[40] E. Frank, Y. Wang, S. Inglis, G. Holmes and I. Witten, "Using Model Trees for Classification", *Machine Learning*, 32 (1), pp. 63-76, 1998.

[41] W. Iba and P. Langley, "Induction of One-Level Decision Trees", in *Machine Learning*, pp. 233-240, 1993.

[42] R. Kohavi, "The Power of Decision Table", in *8th European Conference of Machine Learning*, pp. 174-184, 1995.

[43] R. Quinlan, "C4.5: Programs for Machine Learning", *Morgan Kaufmann Publishers*, San Mateo California, 1993.

[44] W. Cohen, "Fast Effective Rule Induction", in *12th International Conference of Machine Learning*, pp. 115-123, 1995.

[45] N. Landwehr, M. Hall and E. Frank, "Logistic Models Trees", *Machine Learning*, 95(1-2), pp. 161-205, 2005.

[46] S. Le Cessie and J. C. van Houwelingen, "Ridge Estimators in Logistic Regression", *Applied Statistics*, 41(1), pp. 191-201, 1992.

[47] C. M. Bishop, "Neural Networks for Pattern Recognition", *Oxford University Press*, 1995.

[48] G. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers", in 11th *Conference of Uncertainty in Artificial Intelligence*, San Mateo, pp. 338-345, 1995.
[49] F. Eibe and I. Witten, "Generating Accurate Rule Sets Without Global Optimization", in *15th International Conference of Machine Learning*, pp. 144-151, 1998.
[50] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects", *Advances in Space Research*, 41(12), pp. 1955-1959, 2008.
[51] T. Zhang, "Solving large scale linear prediction problems using Stochastic Gradient Descent Algorithms", in *Proceedings of the 21th International Conference of Machine Learning,* p. 116, ACM, 2004.
[52] M. Summer, E. Frank and M. Hall, "Speeding up Logistic Model Induction", in *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 675-683, 2005.
[53] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization", in *B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernels Methods – Support Vector Learning*, 1998.
[54] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm", in *11th Annual Conference of Computational Learning Theory*, New York, NY, pp. 209-217, 1998.
[55] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm", in *13th International Conference on Machine Learning*, San Francisco, pp. 148-156, 1996.
[56] L. Breiman, "Bagging Predictors", *Machine Learning*, 24(2), pp. 123-140, 1996.
[57] J. Friedman, T. Hastie and R. Tibshirani,"Additive Logistic Regression: a Statistical View of Boosting", *Stanford University*, 1998.
[58] L. Breiman, "Random Forest", *Machine Learning*, 45(1), pp. 5-32, 2001.
[59] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), pp. 832-844, 1998.
[60] D. Wolpert, "Stacked Generalization", *Neural Networks*, 5, pp. 241-259, 1992.
[61] L. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms", *John Wiley and Sons, Inc.*, 2004.
[62] R. Quinlan, "Learning with Continuous Classes", in *5th Australian Joint Conference on Artificial Intelligence*, Singapore, pp. 343-348, 1992.
[63] M. Damasceno, "Calculate the Equal Error Rate (EER) using the ROC Curve", Ver 1.0.0, 10 de abril 2013, Disponible: https://github.com/marmundo/eer/blob/master/eer/EER.zip?raw=true, Consultado el 22 de febrero 2014.
[64] A. Moreno, "SALA: SpeechDat Across Latin America", in *Proceedings of the I Workshop on Very Large Databases*, Athens, Greece, 2000.
[65] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtech and P. Wooland, "The HTK Book", *Cambridge University Press*, 2006.
[66] J. A. Gurlekian, N. Colantoni, y H. Torres, "El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados", *Fonoaudiológica. Editorial: ASALFA*. Tomo:47, Número:3, pp 58-69, 2001.
[67] P. Boersma and D. Weenink, "Praat software", Ver. 5.2.01, 2005 *Amsterdam: University of Amsterdam*. Online: http://www.fon.hum.uva.nl/praat, Consultado el 11 de noviembre 2012.
[68] D. Evin, "Grapheme to Phoneme Conversion for Argentine-Spanish", Technical Report of LIS - Laboratorio de Investigaciones Sensoriales, Argentina, 2009.
[69] K. Kira and L. Rendell, "A Practical Approach to Features Selection", in *9th International Workshop on Machine Learning*, pp. 249-256, 1992.
[70] B. Uestuen, W. J. Melsse and L. M. C. Buydens, "Facilitating the application of Support Vector Regression by using universal Pearson VII function based kernel", *Chemometrics and Intelligence Laboratory Systems*, 81, pp. 29-40, 2006.
[71] S. K. Trivedi, S. Dey and P. Shikhar, "Effect of Various Kernels and Feature Selection Methods on SVM Performance for Detecting Email Spams", *International Journal of Computer Applications*, 66(21), 2013.
[72] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection", *Computer Speech Language*, 20(2-3), pp. 230-275, 2006.
[73] C. Greenberg, A. Martin, L. Brandschain, J. P. Campbell, C. Cieri, G. R. Doddington and J. Godfrey, "Human assisted speaker recognition in NIST SRE10", submitted to special session on Human Assisted Speaker Recognition, *Proceedings of IEEE ICASSP*, Prague, 2011.
[74] D. Martin, J. Fitch and V. Wolfe, "Pathologic voice type and the acoustic prediction of severity", *Journal of Speech, Language, and Hearing Research*, 38(4), pp. 765-771, 1995.
[75] T. Bhuta, L. Patrick and J. D. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements", *Journal of Voice*, 18(3), pp. 299-304, 2004.
[76] B. Yegnanarayana, S. Prasanna, J. M. Zachariah and C. S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system", in *Speech and Audio Processing, IEEE Transactions on*, 13(4), pp. 575-582, 2005.
[77] J. A. Gurlekian, y N. Molina, "Índice de perturbación, de precisión vocal y de grado de aprovechamiento de energía para la evaluación del riesgo vocal", *Revista de Logopedia, Foniatría y Audiología*, 32(4), pp. 156-163, 2012.
[78] J. Calvet and G. Malhiac, "Courbes vocales et mue de la voix", *J Fr Oto-Rhino-Laryngol*, 1, pp. 115-124, 1952.
[79] E. M. Albornoz, y D. H. Milone, "Construcción de patrones prosódicos para el reconocimiento automático del habla", en 34ta. *JAIIO*, pp. 225-236, 2005.
[80] M. Almeida, "Patrones rítmicos del español. Isocronía y alternancia", *Estudios Filológicos*, 29, pp. 7-14, 1994.
[81] P. Rose, "Forensic Speaker Identification", *London: Taylor & Francis*, 2002.
[82] J. Gibbons and M. T. Turell (Eds.), "Dimensions of forensic linguistics", *Amsterdam/Philadelphia: John Benjamins Publishing*, 2008.

Pedro Univaso nació en Buenos Aires, Argentina, el 4 de marzo de 1959. Se graduó en la Facultad de Ingeniería de la Universidad de Buenos Aires (UBA) como Ingeniero Electromecánico orientación Electrónica y es candidato al doctorado por la misma universidad. Es investigador invitado del Laboratorio de Investigaciones Sensoriales (LIS), perteneciente al Instituto de Inmunología, Genética y Metabolismo (INIGEM) del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y a la UBA. Sus temas de investigación son el reconocimiento automático de habla y hablantes y la identificación de hablantes en el ámbito forense.

Juan Maria Ale es Doctor en Ciencias, Orientación Computacion, de la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata (UNLP) y Licenciado en Ciencias de la Computacion de la FCEyN (UBA). Ha ocupado diversos cargos profesionales y de dirección, tanto en el área oficial como privada. Se desempeña como consultor independiente, con actuación en empresas privadas y organismos gubernamentales. Es profesor regular en la Facultad de Ingeniería (UBA) y profesor titular en la Facultad de Ingeniería de la Universidad Austral, donde se desempeña como Director de la Maestría en Explotación de Datos y Gestión del Conocimiento. Posee publicaciones científicas sobre bases de datos, OLAP, data warehousing y data mining, y ha participado en numerosos congresos nacionales e internacionales sobre estos temas.

Jorge A. Gurlekian nació en la ciudad de Buenos Aires, el 13 de Septiembre de 1949. Se graduó en la Universidad Tecnológica Nacional, Regional Buenos Aires como Ingeniero Electrónico y obtuvo su Doctorado en la Facultad de Medicina de la UBA en el tema de la evaluación de la inteligibilidad en la producción y percepción del habla. Es investigador principal del CONICET y director del Laboratorio de Investigaciones Sensoriales, perteneciente al INIGEM- CONICET y la UBA en el Hospital de Clínicas J. de San Martín. Su interés es la investigación en la comunicación verbal y su aplicación en el desarrollo de las tecnologías de habla en sistemas de comunicación hombre-máquina.