TECNIACUSTICA Cáceres · 2011

42º CONGRESO ESPAÑOL DE ACÚSTICA
ENCUENTRO IBÉRICO DE ACÚSTICA
EUROPEAN SYMPOSIUM ON ENVIRONMENTAL ACOUSTICS
AND ON BUILDINGS ACOUSTICALLY SUSTAINABLE

ACÚSTICA FORENSE BASADA EN RELACIONES DE VEROSIMILITUD: REPRESENTACIONES PARAMÉTRICAS DE LAS TRAYECTORIAS FORMÁNTICAS DE ALGUNAS COMBINACIONES VOCÁLICAS DEL ESPAÑOL PENINSULAR

PACS: 43.72.Fx

Eugenia San Segundo Fernández
Laboratorio de Fonética. Consejo Superior de Investigaciones Científicas
Calle Albasanz 26-28
28037 Madrid. España
Tel: +34 916 022 940
E-mail: eugenia.sansegundo@cchs.csic.es

ABSTRACT

Some parametric curves (polynomials and discrete cosine transforms) were fitted to the formant trajectories of a series of Spanish vocalic sequences in order to analyze their suitability for forensic voice comparison. The estimated coefficient values from the parametric curves were used as input to a generative multivariate-kernel-density formula for calculating likelihood ratios expressing the probability of obtaining the observed difference between two speech samples under two opposing hypotheses: that the samples were produced by the same speaker and that the samples were produced by different speakers.

RESUMEN

Las trayectorias formánticas de algunas combinaciones vocálicas del español peninsular se han aproximado por dos tipos de curvas paramétricas (polinomiales cúbicas y transformadas discretas de cosenos). Los coeficientes estimados de dichas curvas se han utilizado para calcular relaciones de verosimilitud. Estas expresan la probabilidad de obtener la diferencia observada entre dos muestras de habla bajo la hipótesis de que las muestras han sido producidas por el mismo hablante y bajo la hipótesis de que las muestras han sido producidas por diferentes hablantes.

1. INTRODUCCIÓN

1.1. Presentación del tema

Son varios los autores [1-3] que sostienen que actualmente nos encontramos ante un cambio de paradigma en la mayoría de las Ciencias Forenses, tanto en lo que concierne a la evaluación como a la presentación de la evidencia científica. Dentro del amplio abanico de Ciencias Forenses, únicamente en el campo de la comparación de perfiles de ADN ya se ha producido este cambio de paradigma. El resto de Ciencias Forenses, entre ellas la

comparación forense de voces[1], ha empezado recientemente [3] a emular el modelo de la comparación de perfiles de ADN, si bien el nuevo paradigma todavía dista mucho de haber conseguido una aceptación completa entre los investigadores en ciencias de la voz [4].

El estudio experimental que presentamos a continuación se inscribe dentro de lo que se ha llamado "nuevo paradigma" en la medida en que presenta las siguientes características señaladas por [1]: 1) análisis probabilístico basado en datos, 2) uso de bases de datos con características muestrales de una población de referencia relevante[2] y 3) cuantificación de las limitaciones de la comparación forense llevada a cabo mediante la medida de índices de error.

Un cuarto y último componente del nuevo paradigma, que Morrison [1] considera implícito en [2] es la adopción del marco de relaciones de verosimilitud. Bajo esta perspectiva, también llamada marco bayesiano, la función del científico forense queda clara: ofrecer al juzgador de los hechos un informe con la fuerza de la evidencia en respuesta a la pregunta:

> *¿Cuánto más probable es que las diferencias observadas entre las muestras indubitada (muestra de origen conocido) y dubitada (muestra de origen desconocido) ocurran bajo la hipótesis de que ambas muestras tengan el mismo origen que bajo la hipótesis de que estas tengan un origen distinto?*

Para responder cuantitativamente a esta pregunta, el científico forense debe expresar los resultados de su análisis en forma de una relación de verosimilitud (ecuación 1):

$$LR = \frac{p(E|H_{so})}{p(E|H_{do})} \qquad (1)$$

En esta ecuación, *LR* (*Likelihood Ratio*) es la relación de verosimilitud, *E* es la evidencia científica, esto es, las diferencias medidas entre la muestra de origen conocido (indubitada) y la muestra de origen desconocido (dubitada), $p(E|H)$ es la "probabilidad de E dado H", H_{do} es la hipótesis de distinto origen y H_{so} es la hipótesis de mismo origen (*same origin*).

A la hora de interpretar los resultados, un LR mayor que 1 indica que es más probable que la evidencia ocurra bajo la hipótesis de mismo origen que bajo la hipótesis de diferente origen. Un LR menor que 1 indica que es más probable que la evidencia ocurra bajo la hipótesis de diferente origen que bajo la hipótesis de mismo origen. El tamaño del LR mide cuánto más probable es una hipótesis u otra. [3]

1.2. Revisión bibliográfica

Dado el carácter multidisciplinar de las perspectivas desde las que se ha abordado la cuestión de la comparación forense de voces, encontramos métodos de análisis muy dispares [5], así como una gran diversidad de parámetros acústicos [6] que pueden resultar útiles para los fines

[1] Diversos autores han propuesto diferentes nombres para denominar la técnica de comparación de voces con fines forenses. Algunos términos son muy generales (*Fonética Forense* o *Fonética Judicial* y *Acústica Forense*) y engloban otras actividades aparte de la comparación de voces. Otras denominaciones entre las que a veces existe confusión son *identificación* o *reconocimiento de hablantes*. Un buen resumen sobre estos problemas terminológicos se puede leer en [1] y [5].

[2] La población relevante es la población a la que pertenece el culpable (grabación indubitada), no el/los sospechoso/s.

[3] El numerador de un LR puede considerarse una medida de la similitud entre la muestra dubitada e indubitada mientras que el denominador sería una medida de la tipicidad de la muestra dubitada. De ahí que, a la hora de calcular la fuerza de la evidencia, sea necesario contar con una población de referencia relevante. Solo la similitud no conlleva un apoyo fuerte de la hipótesis del mismo origen [1].

forenses de esta disciplina. Por tanto, la cuestión de qué parámetros discriminan mejor entre hablantes sigue estando abierta a la investigación.

Los dos requisitos básicos para la elección de los parámetros fonético-acústicos son la *cantidad* (que exista un alto número de ocurrencias) y la *calidad* de esos parámetros. Más en concreto, se suelen mencionar [7] las siguientes cinco características que debería poseer el parámetro ideal: (1) presentar mucha variabilidad interhablante y poca variabilidad intrahablante, (2) ser resistente a los intentos de disimulo e imitación, (3) aparecer frecuentemente en el habla, (4), ser robusto a las posibles distorsiones debidas al canal de transmisión, y (5) ser fáciles de extraer y de medir.

Los formantes[4] vocálicos (en concreto, sus frecuencias centrales o promediadas), son considerados uno de los parámetros acústicos *tradicionales* en la comparación forense de voces, en oposición a los parámetros acústicos *automáticos*. La relevancia forense de los formantes es debida fundamentalmente al correlato anatómico-fisiológico de dichos parámetros con el tracto vocal de los hablantes, así como a su fácil extracción y medición.

Desde hace relativamente poco, son varios los estudios [8-12] que, lejos de centrarse en la utilidad forense de aspectos *estáticos* como las frecuencias centrales de los formantes vocálicos, han optado por el estudio de las propiedades *dinámicas* de dichos formantes, no solo en monoptongos (una sola vocal) [11,12], sino también en secuencias de dos vocales (diptongos) [8-10].

Los autores que han trabajado en este ámbito [8, 9] parten de la idea de que la señal acústica contiene puntos (llamados *targets*) determinantes lingüísticamente para el reconocimiento de un fonema, unidos por las llamadas *transiciones*, que no estarían tan constreñidas por el sistema lingüístico del hablante y no serían claves para la percepción de fonemas. Por tanto, en estas últimas habría cabida para la idiosincrasia de los hablantes, ya que la solución que cada hablante encontraría para realizar dichas transiciones dependería de la fisiología de su tracto vocal y de su habilidad motora aprendida para moverse entre dichos *targets* fonéticos.

El español tiene la peculiaridad de que muchas combinaciones de dos vocales, como *ia* se pueden pronunciar en una misma sílaba, como en *seria* [∀se.4ja] o en dos sílabas diferentes, como en *sería* [se.∀4i.a]. En otras palabras, existen pronunciaciones diptongadas e hiáticas. Si bien existen unas reglas generales para conocer la distribución de los hiatos y los diptongos en español, lo cierto es que la aparición de unos y otros no siempre es predecible, ya que los hablantes no siempre pronuncian del mismo modo las secuencias vocálicas [13]. Influyen al respecto factores dialectales, sociolingüísticos e idiolectales, entre otros. Por todo esto, creemos que en el estudio de este tipo de parámetros podemos encontrar mayor variación interlocutor que intralocutor, y consideramos, por tanto, que serían parámetros de potencial interés forense.

1.2. Formulación de la hipótesis

El objetivo que nos planteamos en este estudio es contestar a la siguiente pregunta: "¿Hasta qué punto las trayectorias formánticas de ciertas secuencias vocálicas del español son parámetros útiles para la comparación forense de voces?" A partir de esta pregunta general, surgen otra serie de hipótesis:

- ¿Algunas representaciones paramétricas son más útiles que otras para ajustar las trayectorias formánticas de una secuencia vocálica en español?

[4] Picos de intensidad en el espectro de un sonido; se corresponden con la forma que adopta la cavidad oral en la producción de un sonido.

- ¿Considerar el diptongo [ja] y el hiato [ia] por separado ofrece mejores resultados que considerarlos conjuntamente?

- ¿Algunas secuencias vocálicas son más útiles que otras para la comparación forense de voces?

2. MÉTODO

2.1. Corpus e informantes

Se han seleccionado las grabaciones de 29 hombres adultos, hablantes de castellano (variedad centropeninsular), procedentes del Corpus Ahumada [14]. Las grabaciones seleccionadas pertenecen a un estilo de habla espontáneo y se han utilizado dos sesiones no contemporáneas por hablante, de aproximadamente 90 segundos cada una.

2.2. Procedimiento de análisis

2.2.1. Análisis acústico y métodos de ajuste paramétrico

De las 58 grabaciones (29 informantes x 2 sesiones cada uno), se han analizado las secuencias vocálicas [we], [je], [ja] y [ia], previamente etiquetadas [15]. Para cada secuencia vocálica, se midieron [16] los valores frecuencias de las trayectorias formánticas de F1, F2 y F3. Dos tipos de curvas paramétricas (polinomiales cúbicas y transformadas discretas de cosenos) se han ajustado a las tres trayectorias formánticas extraídas anteriormente.

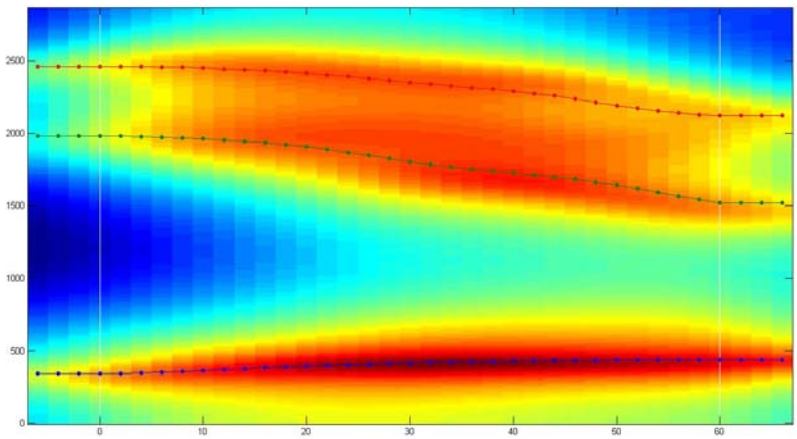


Figura 1. Trayectorias formánticas de [je] para el hablante 12.

En cuanto al primer ajuste paramétrico, la trayectoria de cada formante se ajustó a una curva descrita por una ecuación polinomial que describe la frecuencia en función del tiempo, de forma que, utilizando cuatro coeficientes, la diferencia con la curva original sea mínima, de acuerdo con la siguiente fórmula (ecuación 2):

$$ax^3 + bx^2 + cx + d = 0 \quad (2)$$

En lo que respecta a la aproximación por Transformadas Discretas de Cosenos (*Discrete Cosine Transform, DCT*), al vector de valores se le aplica una transformada DCT unidimensional, que lo convierte en otro vector de valores correspondientes a los coeficientes que definen dicha transformación (ecuación 3).

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \quad k = 0,\ldots,N-1 \tag{3}$$

2.2.2. Cálculo de las relaciones de verosimilitud

Para el cálculo de relaciones de verosimilitud hemos usado la formula MVKD (*Multivariate Kernel Density*) descrita en [17] e implementada en [10]. Las variables que se introdujeron en dicha fórmula son los valores numéricos de los coeficientes resultantes de la aproximación paramétrica de las trayectorias formánticas. Esta fórmula para la determinación del peso de la evidencia cuando existen múltiples variables tiene en cuenta la existencia de posibles correlaciones entre dichas variables. Además, este método modela la variabilidad interlocutor mediante una suma de *kernels* de tipo gaussiano.

3. RESULTADOS

Para definir el rendimiento de los LRs finales generados por los distintos sistemas de comparación forense[5] hemos utilizado los valores C_{llr} (*log-likelihood-ratio cost*) propuestos en [18] (ecuación 4):

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}}\log_2\left(1+\frac{1}{LR_{so_i}}\right)+\frac{1}{N_{do}}\sum_{j=1}^{N_{do}}\log_2\left(1+LR_{do_j}\right)\right) \tag{4}$$

La medida de precisión C_{llr} es una función continua con valores pequeños para LRs correctos y que se vuelve asintótica hacia cero a medida que los LRs divergen de 1, pero con valores grandes para LRs incorrectos y que crece exponencialmente a medida que los LRs divergen de 1. La idea básica es que las comparaciones de un mismo hablante deberían arrojar valores de Log-LR muy altos, mientras que los log-LRs resultantes de las comparaciones entre distintos hablantes deberían ser muy bajos, es decir, negativos. Las desviaciones de este principio básico se castigan de forma más dura cuanto mayor es la desviación. Todo esto redunda en valores de C_{llr} altos. En definitiva, cuanto más bajo sea el valor del C_{llr}, mejor rendimiento tiene el sistema que estemos evaluando.

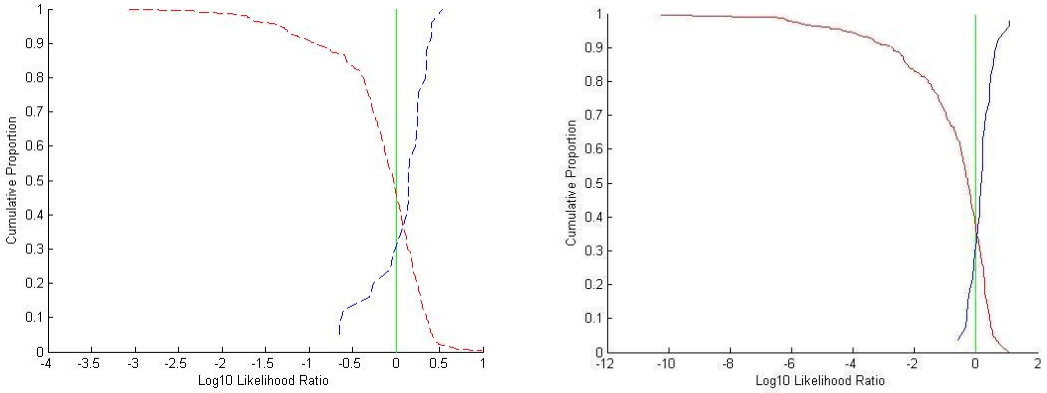
En la tabla 1 se muestran los resultados obtenidos en el cálculo de valores C_{llr} a partir de las relaciones de verosimilitud obtenidas en cada sistema.

	[we]			[je]			[ja]			[ia]			[ja] & [ia]		
	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
DCT	1.046	1.075	0.999	1.003	0.971	0.951	1.043	1.087	1.047	1.027	1.168	1.007	0.978	1.060	0.998
POLINOMIAL	0.984	0.998	0.833	1.036	0.982	0.890	1.045	1.145	0.914	1.065	1.012	0.943	0.973	1.064	1.056

Tabla 1: Valores C_{llr} para cada diptongo/hiato y formante. En las tres últimas columnas se han considerado el diptongo [ja] y el hiato [ia] conjuntamente

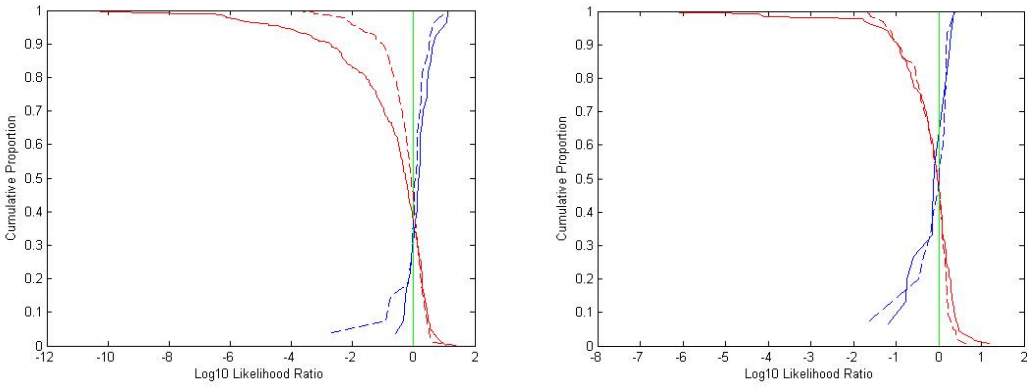
[5] Entendemos por sistema de comparación el conjunto de diptongo/hiato más formante (F1-F3) analizado en cada caso.

Los gráficos *Tippett*[6]de las figuras 2 y 3 muestran los resultados obtenidos tras ejecutar algunos de nuestros sistemas. En un gráfico *Tippett* se representan dos curvas. La azul, la que crece hacia la derecha, representa los resultados LR de las comparaciones para un mismo hablante, con la proporción cumulativa de log-LRs menores que, o iguales a, el valor indicado en el eje de abscisas. La curva roja, que crece hacia la izquierda, representa los resultados LR de las comparaciones entre distintos hablantes, con la proporción cumulativa de log-LRs mayores que, o iguales a, el valor indicado en el eje de abscisas. Un sistema de comparación forense ideal es el que devuelve valores de log-LRs muy grandes (positivos) para las comparaciones de un mismo hablante y log-LRs muy bajos (negativos) para comparaciones entre hablantes diferentes.



(a) (b)

Figura 1: Gráficos *Tippett* que representan el mejor sistema de comparación forense aplicando (a) un ajuste DCT: el sistema formado por el diptongo [je] y el F3 (C_{llr} de 0.951); y (b) un ajuste polinomial: el sistema formado por el diptongo [we] y el F3 (C_{llr} de 0.833).



(a) (b)

Figura 2: Gráficos *Tippett* que representan (a) la máxima diferencia hallada entre los valores de un sistema polinomial y uno DCT, correspondiente al sistema formado por el diptongo [we] y el F3 y (b) la mínima diferencia obtenida de la comparación entre un sistema polinomial y uno DCT, correspondiente al sistema formado por [ja] y F1.

[6] Los llamados *Tippett plots* son métodos gráficos de representar los resultados de la ejecución de un sistema de comparación forense basado en relaciones de verosimilitud. Introducidos por primera vez por [19], representan actualmente el método estándar de representación de los resultados en la investigación forense de voces basada en relaciones de verosimilitud.

4. ANALISIS DE LOS RESULTADOS

Los datos obtenidos, de acuerdo con la tabla 1, no son excesivamente buenos. Los resultados C_{llr} por debajo de 1 para las parametrizaciones con DCT son tan solo 5 de 15, mientras que para las parametrizaciones con curvas polinomiales son 8 de 15. Esto puede deberse al pequeño tamaño de la muestra (29 hablantes) o a la escasa incidencia de algunas secuencias vocálicas en algunos hablantes. Recordemos que las grabaciones usadas para la extracción de combinaciones vocálicas pertenecen a un estilo de habla espontáneo. Esto hace que el número de apariciones de los diptongos e hiatos usados en este corpus no sea predecible ni sea el mismo en todos los hablantes (como ocurriría en habla leída a partir de un mismo texto).

En general, no parece que existan diferencias entre el uso de curvas polinomiales y curvas DCT. Por otro lado, los resultados no parecen mejorar excesivamente si consideramos [ja] e [ia] conjuntamente, por lo que juzgamos que tener en cuenta la distinción entre hiato y diptongo es recomendable a la hora de analizar secuencias vocálicas que admiten ambas pronunciaciones. Sin embargo, habría que hacer una fusión de los tres formantes (F1, F2 y F3) para conocer si unas secuencias vocálicas resultan más útiles que otras para la comparación de voces. Lo que sí que podemos conocer, a partir de los resultados obtenidos, es qué combinación de formantes y de secuencias vocálicas ofrece mejores valores C_{llr}, y constituyen por tanto mejores sistemas de comparación forense. Por lo general, con los terceros formantes (F3) de cada combinación vocálica se obtienen mejores resultados que con el resto de formantes: F3 de [we] obtiene un C_{llr} de 0.833, F3 de [je] obtiene C_{llr} de 0.890, F3 de [ia] obtiene C_{llr} de 0.914 y F3 de [ja] obtiene un C_{llr} de 0.943. Finalmente, destacamos que los valores de C_{llr} son mejores, por lo general, para todos los formantes de los diptongos [we] y [je] que para los formantes de [ja] y [ia].

5. CONCLUSIONES

El hecho de que las propiedades dinámicas de los diptongos se hayan considerado parámetros con alta variabilidad interlocutor y baja variabilidad intralocutor en estudios fonético-forenses en otras lenguas, sumado al hecho de que las secuencias vocálicas [we], [je], [ja] y [ia] aparecen con relativa alta frecuencia en el habla espontánea en español, nos ha llevado a emprender el estudio de dichas secuencias como parámetros de potencial utilidad forense.

La aportación principal de este trabajo radica en que por primera vez se hace uso de la distinción lingüística entre diptongo e hiato a la hora de considerar los parámetros vocálicos con los que calcularemos las relaciones de verosimilitud. Además, se ha utilizado un corpus de habla espontánea, lo que supone una característica forense más realista que el uso de un corpus de lectura. En trabajos futuros pretendemos hacer uso de otras técnicas de cálculo de relaciones de verosimilitud, como el modelo GMM-UBM (*Gaussian Mixture Model – Universal Background Model*) descrito en [20]. Además, tenemos intención de utilizar técnicas de fusión y calibración, como las propuestas en [1] para conocer si algunas secuencias vocálicas (comprendiendo parte de sus formantes o todos ellos) son más útiles que otras para la comparación forense de voces, por ofrecer mayor capacidad discriminatoria entre hablantes y mejores valores C_{llr}.

AGRADECIMIENTOS

La elaboración de este estudio ha sido posible gracias a la concesión de una beca-contrato del Programa Nacional de Formación de Profesorado Universitario (FPU), otorgada por el Ministerio de Educación, con resolución del BOE del 11-07-2009. Quisiera asimismo agradecer a la Real Sociedad Española de Física la concesión de una beca para la asistencia a *Tecniacústica 2011*.

REFERENCIAS

[1] Morrison, G.S. (2010) Forensic Voice Comparison. In I. Freckelton, & H. Selby (Eds.), *Expert Evidence* (Ch. 99). Sydney, Australia: Thomson Reuters.

[2] Saks, M.J., Koehler, J.J. (2005) The coming paradigm shift in forensic identification science. *Science*, 309, 892–895.

[3] González-Rodríguez, J., Rose, P., Ramos-Castro, Toledano, D.T. and Ortega-García, J . (2007) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on audio, speech and language processing*, 15 (7): 2072-2084.

[4] Gold, E. (2011) Forensic Speaker Comparison Evidence: The International Picture. *International Association for Forensic Phonetics and Acoustics* (IAFPA) Conference, Vienna, 24-28 julio 2011.

[5] Rose, P. (2002) *Forensic Speaker Identification*, Taylor & Francis.

[6] Battaner, E., Carbó, C., Gil, J., Llisterri, J., Machuca, M.J., Madrigal, N., Marrero, V., de la Mota, C., Riera, M. and Ríos, A. (2007) VILE: Estudio acústico de la variación inter e intralocutor en español, *Actas do 3º Congreso Internacional de Fonética Experimenta*l, Santiago de Compostela, 24-26 octubre de 2005, pp.157-167.

[7] Nolan, F. (1983) *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

[8] McDougall, K. (2006) Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies, *International Journal of Speech, Language and the Law* 13 (1), pp.89-126.

[9] Rose, P. (2006) The Intrinsic Forensic Discriminatory Power of Diphthongs, In P. Warren and C. I. Watson (Eds.) *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, University of Auckland, New Zealand. December 6-8, 2006.

[10] Morrison, G.S. (2009) Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, 125 (4), 2387-2397.

[11] Morrison, G.S. and Kinoshita, Y. (2008) Automatic-Type Calibration of Traditionally Derived Lilkelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories. In *Proceedings of Interspeech 2008 incorporating SST'08*, pp. 1501-1504.

[12] López-Escobedo, F. (2010) *El análisis de las características dinámicas de la señal de habla como posible marca para la comparación e identificación forense de voz: un estudio para el español de México*. Tesis Doctoral. Universitat Pompeu Fabra.

[13] Aguilar, L. (2010) Vocales en grupo, *Cuadernos de Lengua Española*, Madrid: Arco Libros.

[14] Ortega, J., González, J. y Marrero, V. (2000) AHUMADA: A large corpus in Spanish for speaker characterization and identification, *Speech Communication* 31, 2-3: 255-264.

[15] Morrison, G. S. (2008b). SoundLabeller: Ergonomically designed software for marking and labelling portions of sound files (Release 2008-12-19). [Computer software. Available: http://geoff-morrison.net]

[16] Morrison, G. S., and Nearey, T. M. (2008) FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories (Release 2008-12-21). [Computer software. Available: http://geoff-morrison.net]

[17] Aitken, C.G.G. and Lucy, D (2004) Evaluation of trace evidence in the form of multivariate data. *Appl. Stat*, 53, pp.109–122.

[18] Brummer, N. and du Preez, J. (2006) Application independent evaluation of speaker detection, *Comp. Speech Lang*. 20, pp. 230–275.

[19] Meuwly, D. (2001) *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. PhD dissertation, University of Lausanne, Lausanne, Switzerland.

[20] Reynolds, D.A., Quatieri, T.F. and Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process*. 10, 19–41.