Uso de Entonación en Reconocimiento Automático de Locutor: Resultados Preliminares

David Escudero, Valentín Cardeñoso
Departamento de Informática
Universidad de Valladolid
{descuder,valen}@infor.uva.es

Juan María Sánchez, Eva Navas, Inma Hernáez
Departamento de Electrónica y Telecomunicaciones
Universidad del País Vasco
{ion,eva,inma}@bips.bi.ehu.es

RESUMEN

En esta comunicación se presenta una experiencia de reconocimiento de locutor basada en la entonación. A partir de un corpus multilocutor grabado con diferentes micrófonos y en diferentes sesiones, se plantea un experimento de verificación de locutor en la que se emplea la entonación como único rasgo distintivo. La entonación se parametriza empleando funciones de Bézier que aproximan los puntos de pitch en los grupos acentuales y se emplea un clasificador en árbol basado en el algoritmo C4.5 como sistema de decisión en el proceso de verificación de locutor Aunque los resultados muestran una dependencia notable del locutor que se está verificando, su robustez frente a cambios de micrófono y de sesión justifican el interés por emplear la entonación como rasgo biométrico en esta tarea. Como resultado adicional, se valora la notable incidencia de la elección del tipo de grupo acentual en los resultados finales.[1]

ABSTRACT

In this communication we present an experience on speaker recognition based in intonation. By using a multi-session corpus, recorded with different microphones and in different sessions, it is proposed an experiment of speaker verification where the intonation is used as the only acoustic feature. Intonation is parameterised by using Bézier functions that approximate the points of pitch in the stress groups. It is used a classifier based on the C.45 algorithm to implement the experiment of speaker verification. Results illustrate that some of the speakers are identified with low error rates. Robustness of the results when changes in the recording conditions (microphone or session) prove the benefits of the use of the features of intonation in real speaker recognition systems. One last experiment brings evidence to the fact that results can vary significantly in function of the type of stress group to be considered.

1. INTRODUCCIÓN

El principal argumento que hace pensar que la entonación puede aportar beneficios al ser empleada en reconocimiento de locutor es que la entonación lleva asociados dos tipos de información sociolingüística de especial relevancia: una relativa al grupo sociocultural a que

[1] Trabajo financiado parcialmente con cargo al proyecto C2000-1669-C0403 del Ministerio de Ciencia y Tecnología y VA083/03 de la Consejería de Educación y Cultura de la Junta de Castilla y León.

pertenece el locutor y otra asociada a la naturaleza psicosomática del propio locutor. En relación con el entorno sociocultural del hablante, la entonación puede informar del origen geográfico del individuo, el medio social al que pertenece, su grado de cultura, etc. Con respecto a la información sobre los rasgos personales del individuo, revela algunas características individuales: edad, sexo, temperamento, carácter, estado de ánimo, entre otros..

Un segundo argumento que justifica el interés por el uso de la entonación como característica biométrica en reconocimiento de locutor tiene que ver con la robustez del sistema de reconocimiento o identificación frente a cambios de los medios físicos con los que se realiza la grabación y del entorno en que se lleva a cabo.. La variación temporal de la frecuencua fundamental F0 de la señal hablada es la magnitud que se emplea habitualmente para caracterizar la entonación de una locución y su estabilidad frente a los cambios de medio o de entorno en la sesión de grabación se considera superior a la que proporcionan los parámetros de índole fonético habitualmente empleados en reconocimiento de locutor (e.g. coeficientes LPC o bandas de frecuencia).

Algunos trabajos de investigación recientes han puesto de manifiesto que la inclusión de información suprasegmental en sistemas de reconocimiento de locutor puede aportar mejoras relevantes en su funcionamiento[1]. Aunque puede encontrarse algún antecendente bibliográfico sobre la incorporación de prosodia y entonación en identificación de locutor[2], los trabajos realizados hasta la fecha se basan casi sistemáticamente en el uso de modelos de evolución del *pitch* excesivamente simplistas –estadísticas sencillas sobre los valores o sobre la evolución de las pendientes. En contra de esta tendencia, en este trabajo se propone un modelado estadístico de la entonación más sofisticado que pueda reflejar de forma más correcta la información prosódica relevante que acompaña a un mensaje determinado.

En este trabajo pretendemos demostrar, en primer lugar, que la entonación puede ser un rasgo biométrico valioso en tareas de reconocimiento de locutor. En segundo lugar, aportar resultados que demuestren que las tasas de reconocimiento que pueden alcanzarse cuando se incorpora información suprasegmental de entonación son muy razonables, habida cuenta la porción reducida de locución que se emplea. Finalmente, mostraremos que dichas tasas se mantienen estables frente a cambios de sesión y de micrófono, lo cual puede resultar muy interesante desde el punto de vista de la robustez del reconocedor en entornos reales.

En este estudio planteamos un experimento de verificación de locutor en el que se emplee la entonación como único rasgo biométrico. Multiples sesiones de grabación con micrófonos diferentes de un mismo párrafo leído por diversos locutores constituyen la base material de este experimento. La entonación se describe empleando un modelo en el que cada uno de los grupos acentuales de una secuencia hablada se representa por medio de un conjunto fijo de parámetros obtenidos a partir de una función de Bézier[3]. Con los parámetros de entonación así obtenidos se elabora, en una fase de aprendizaje en la que se emplean árboles de decisión, un modelo de locutor y de impostor para cada uno de los locutores analizados y para cada sesión de grabación. Con estos modelos, se evalúan los resultados que se presentan y que sirven para sustentar las afirmaciones realizadas en el párrafo anterior.

En las secciones que siguen se describe esquemáticamente la técnica de parametrización de F0 empleada en este trabajo, el corpus a que se refiere la tarea de identificación de locutor que hemos abordado, la técnica de reconocimiento aplicada, los resultados obtenidos y las consecuencias que podemos extraer de los mismos y, para terminar, un resumen de conclusiones y posibilidades de trabajo futuro.

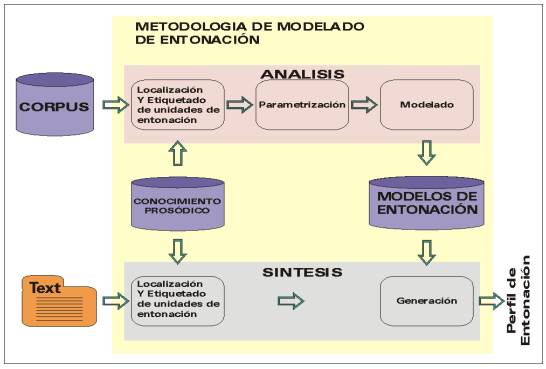

Figura 1. Modelado y generación de entonación.

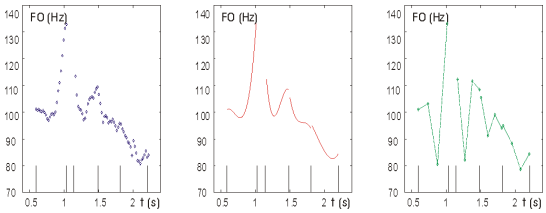

Figura 2. Parametrización de los perfiles de entonación. Izquierda: perfil de entonación original con marcas de frontera entre grupos acentuales; Centro: Función de Bézier aproximante; Derecha: puntos de control de la función de Bézier, parámetros de entonación empleados.

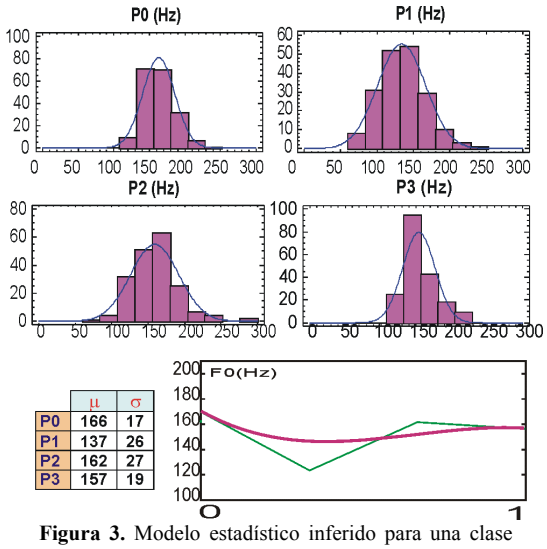

Figura 3. Modelo estadístico inferido para una clase dada. Arriba los histogramas que muestran la distribución de valores cada parámetro. Abajo a la izquierda los estadísticos de primer orden para la clase. Abajo a la derecha las representación gráfica del patrón característico de la clase.

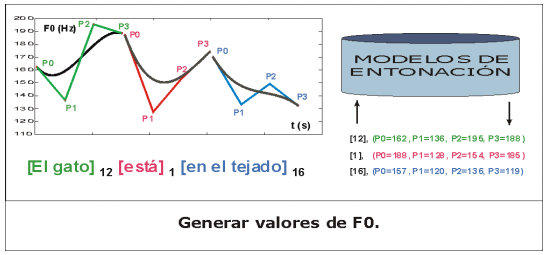

Figura 4. Esquema de generación de perfiles de entonación. Se genera un perfil de F0 para la frase de ejemplo.

2. MODELADO DE LA ENTONACIÓN

La metodología de modelado de entonación empleada ha sido ya descrita en [3][4][5] y se resume en la figura 1. Los modelos de entonación se generan a partir de un corpus. Primero se localizan en el corpus los grupos acentuales y se enriquecen con atributos prosódicos. Después, se parametrizan los perfiles de F0 asociados a cada grupo acentual y por medio de estos parámetros se modela la entonación de las muestras del corpus. Para la conversión texto-voz, se parte de una segmentación de la entrada en grupos acentuales y de entonación, se localizan los grupos acentuales con sus correspondientes etiquetas prosódicas en el corpus y se selecciona el modelo de perfil correspondiente para generar, usando técnicas de simulación, las curvas de F0.

Ya hemos apuntado que la parametrización de entonación se basa en el uso de funciones de Bézier. Cada unidad de entonación (grupo acentual) se parametriza de forma aislada y se toman como parámetros de entonación del grupo acentual los puntos de control de la función de Bézier que aproxima el perfil de entonación que se esté analizando. Se emplean cuatro puntos de control por grupo acentual y se aproximan los contornos de pitch usando una técnica de mínimos cuadrados. En [6] se describe el método de aproximación y las distintas opciones de

parametrización que fueron consideradas. El proceso de parametrización se ilustra en la figura 3.

Para modelar la entonación es necesario determinar el número de clases de grupos acentuales que se va a emplear, para lo que se consideran una serie de rasgos prosódicos bien reconocidos en la literatura. De esta forma, dos grupos acentuales pertenecerán a la misma clase si comparten los factores prosódicos que se estén empleando para la clasificación. En este trabajo se han considerado los factores prosódicos relevantes que fuesen ya estudiados en los trabajos de Garrido [7]. Para cada clase de rasgos prosódicos, el modelo de entonación es la distribución estadística de los puntos de control de las funciones de Bézier que aproximan las muestras del corpus que se han asignado a dicha clase. La razón de emplear como modelo de entonación una colección de distribuciones estadísticas etiquetadas por clase prosódica no es otra que modelar la regularidad de los patrones de entonación de una misma clase asegurando, a un tiempo, un grado de variabilidad suficiente que evite la monotonía en el proceso de conversión texto-voz. El proceso de modelado se describe con más detalle en [5]. Los histrogramas de la figura 3 ilustran la forma de las distribuciones que se obtienen como modelos de entonación para una clase particular de grupo acentual.

Los modelos estadísticos así obtenidos pueden emplearse fácilmente para generar nuevos perfiles sintéticos en un conversor texto-voz. Para ello, primero se localizan y clasifican los grupos acentuales del texto a sintetizar. Dada la clase, se extraen de la base de datos de modelos los parámetros correspondientes. Estos parámetros se emplean para generar el perfil de pitch correspondiente, que será la curva de Bézier asociada a los puntos de control que se obtengan por técnicas estadísticas a partir de los parámetros extraídos de la base de modelos, tal y como se describe en la figura 4. El lector interesado, puede encontrar en [6] un estudio comparativo detallado de varias técnicas de generación de estos puntos de control a partir de las distribuciones estadísticas.

La observación experimental más relevante desde el punto de vista de las aplicaciones de reconocimiento es que los modelos de entonación asociados a clases distintas de grupos acentuales son significativamente diferentes. Así, se ha podido constatar que los patrones característicos con grupos acentuales iniciales suelen ser muy diferentes a los patrones característicos de grupos acentuales finales. Este hace pensar que a la hora de emplear la entonación en reconocimiento se ha de tener en cuenta la información sobre el tipo de unidad de entonación que se esté empleando, ya que la forma de la entonación puede variar significativamente en función de dicho tipo. Estos aspectos se corroboran experimentalmente en los apartados siguientes.

3. DESCRIPCIÓN DEL CORPUS

El corpus empleado fue grabado en la Escuela Universitaria Politécnica de Mataró de la Universidad Politécnica de Cataluña (EUPMT) [8]. Los experimentos que presentamos en este trabajo se basan sólo en la parte del mismo en que varios locutores leen el mismo párrafo de texto. De todos los locutores disponibles, se tomaron las muestras de 16 de ellos. Se llevaron a cabo 6 sesiones de grabación diferentes, con 3 micrófonos distintos, para cada uno de los locutores del corpus. Cada párrafo se compone de 11 frases diferentes y contiene 106 grupos acentuales, lo que permite disponer de un total de 3816 muestras multilocutor, multi-micrófono de unidades de entonación diferentes.

Los grupos acentuales fueron segmentados manualmente. Se empleó un algoritmo de extracción del pitch robusto basado en [9] para calcular la frecuencia fundamental característica de los grupos acentuales. Luego, los grupos acentuales fueron parametrizados calculando una función de Bézier aproximante de tercer grado. Con ello, cada una de las unidades de entonación del corpus se representa mediante un vector de cuatro componentes.

Todos los locutores del corpus son varones y pertenecen al mismo ámbito sociocultural: estudiantes y profesores de la EUPMT naturales de la ciudad de Mataró (Barcelona) o alrededores. En consecuencia, la tarea de reconocimiento de locutor que se aborda en este estudio es especialmente difícil debido a que no se van a apreciar diferencias entre los locutores debidas a localismos o nivel cultural. Es importante tener esto en cuenta sobre todo a la hora de poner en comparación los resultados que se obtengan con los que aporten otros trabajos.

4. CONFIGURACIÓN DEL EXPERIMENTO

Nos referiremos a cada locutor como *Li* y al impostor del locutor *Li* como *Ii*. El corpus de trabajo se ha dividido en dos bloques: un bloque de entrenamiento en el que se selecciona un conjunto de muestras de locutor y de impostor para modelado, y otro bloque de prueba, que contiene un conjunto de muestras de impostor y de locutor, y que es sobre el que se computan las tasas de reconocimiento. Con objeto de valorar la robustez del reconocedor ante cambios de sesión y de micrófono, se elegirán muestras de diferentes micrófonos.

Como hemos apuntado, para cada locutor se dispone de 6 sesiones de grabación realizadas con 3 micrófonos diferentes. Se seleccionan las muestras de 5 de las sesiones como muestras de entrenamiento para modelar el locutor *Li* y su impostor (modelo único en verificación). En la fase de entrenamiento, se toman las muestras de entrenamiento de los locutores diferentes al *Li* se como muestras de entrada asociadas al impostor *Ii*.

Con respecto a las muestras de prueba, se reservan las muestras de la sesión no empleada en entrenamiento para calcular las tasas de reconocimiento. Las muestras de prueba correspondientes al locutor *Li* son las muestras de dicha sesión y las del impostor *Ii* se obtienen reuniendo las muestras de prueba del resto de locutores.

Esta selección de los conjuntos de entrenamiento y de prueba se hace para cada una de las sesiones. Así, para cada locutor, se dispone de seis conjuntos de muestras (uno por sesión-micrófono) compuestos de :

- Muestras de entrenamiento del locutor *Li*;
- Muestras de entrenamiento del impostor *Ii*;
- Muestras de prueba del locutor *Li*
- Muestras de prueba del impostor *Ii*

lo que permite valorar la influencia del cambio de sesión y micrófono en los tasas de reconocimiento que presentamos más adelante.

Como reconocedor, empleamos la versión de un clasificador C4.5[10] que proporciona la librería gratuita WEKA[11]. Las muestras de entrenamiento de cada locutor y de su respectivo impostor se combinan en un único fichero. Este fichero será el fichero de entrada a WEKA-C4.5 para entrenar el clasificador. Las muestras de prueba de cada locutor y de su impostor respectivo se combinan en un único fichero de entrada al clasificador que, de acuerdo con las especificaciones del algoritmo C4.5, implementa un árbol de decisión con poda.

En la fase de entrenamiento y prueba descrita se han considerado conjuntamente todos los grupos acentuales de un locutor. Para valorar la influencia del tipo de grupo acentual (y por

tanto de la información prosódica particular a un grupo de rasgos prosódicos concretos), se han repetido los experimentos anteriores considerando sólo grupos acentuales del mismo tipo, para tres clases de grupos diferentes: iniciales, centrales y finales.

5. RESULTADOS

La tabla 1 muestra las tasas de aciertos obtenidas para los distintos locutores cuando se emplean como muestras de prueba los datos de los distintos micrófonos presentes en el corpus. En primer lugar, puede observarse que los resultados de reconocimiento dependen fuertemente del locutor. En este sentido, se puede apreciar que el locutor *L6* ofrece resultados que son claramente mejores que los que se obtienen con otros locutores como el locutor *L2* (80% frente a 50%). Estos resultados indican claramente que la entonación al leer es un rasgo que caracteriza fuertemente a algunos locutores, mientras que otros leen sus frases con una entonación difícil de distinguir de la que emplean otros locutores de su mismo grupo. Este resultado era, en cierta medida, previsible e indica que aunque la entonación es un rasgo útil en tareas de reconocimiento no puede considerarse de forma aislada a la hora de construir reconocedores competitivos.

	M1	M2	M3	M4	M5	M6	Media
L0	65,19	63,52	60,13	68,59	63,75	59,01	63,4
L1	71,60	72,67	65,82	69,28	67,70	66,25	68,9
L2	57,95	41,18	61,14	52,87	37,87	60,67	51,9
L3	51,72	53,61	58,28	40,35	59,64	50,89	52,4
L4	58,52	58,18	57,32	56,90	53,53	54,76	56,5
L5	73,24	74,81	78,03	72,54	71,13	71,43	73,5
L6	80,12	80,86	85,44	78,36	85,37	81,87	82,0
L7	65,50	56,25	59,88	63,16	61,85	58,72	60,9
L8	48,75	67,81	59,33	56,60	68,28	60,26	60,2
L9	64,77	67,74	68,03	62,57	60,13	65,56	64,8
L10	65,03	72,19	60,00	63,41	73,05	69,19	67,1
L11	62,50	72,81	64,12	69,84	64,75	66,42	66,7
L12	60,87	54,07	56,34	64,85	58,72	58,55	58,9
L13	54,97	68,64	59,52	66,28	61,05	63,64	62,4
L14	63,69	62,96	72,73	60,12	61,96	62,94	64,1

Tabla 1: **Tasas de Reconocimiento**. Filas: tasas para los distintos locutores (*Li, i*=0....15). Columnas: tasas empleando distintos datos de prueba. *Mj* (j=1...6) indica que se han empleado como datos de prueba los correspondientes a grabaciones realizadas con distintos micrófonos.

Por otro lado, los resultados de la tabla 1 muestran que los resultados de reconocimiento se ven influidos por cambios de sesión o de micrófono sólo ligeramente. Las distintas columnas se obtienen al aplicar diferentes conjuntos de modelado y de prueba asociados a diferentes micrófonos y sesiones de grabación. El resultado parece depender más del locutor que de la sesión considerada. Con ello se pone de manifiesto que la entonación puede ser más robusta que otras propiedades acústicas frente a cambios de micrófonos y o de sesión.

	L1	L5	L6	L10	L11
Total	68.89	73.53	82.00	67.15	66.74
Inicial	80.03	55.08	77.65	58.11	52.44
Central	65.02	72.95	80.86	65.22	69.01
Final	69.25	67.71	89.49	75.70	53.87

Tabla 2: Tasas de Reconocimiento: Columnas: las tasas para los cinco locutores con mayores tasas de acierto; Filas: las tasas empleando todos los grupos acentuales (*Total*), sólo los iniciales (*Inicial*), sólo los centrales (*Central*) o sólo los finales (*Final*).

Los resultados que se obtienen cuando se repiten los experimentos considerando las distintas clases de grupos acentuales se muestran en la Tabla 2. Por claridad, se presentan en ella sólo los resultados relativos a los locutores de la tabla 1 para los que se consiguió una mayor tasa de reconocimiento. Como se observa, los resultados cambian sensiblemente cuando se considera uno u otro tipo de grupo acentual como unidad de reconocimiento. Lamentablemente, no puede afirmarse que, de forma general, determinado tipo de grupo acentual aporte más información que otros sobre el tipo de locutor. Así, para el locutor, para el locutor L1, el tipo de grupo acentual más característico es el inicial, mientras que para el locutor L6 y L10 es tipo de grupo acentual más característicos es el final.

6. CONCLUSIONES Y TRABAJO FUTURO

Los resultados obtenidos ponen de manifiesto que la entonación puede aportar información relevante en aplicaciones de reconocimiento de locutor. En el experimento de verificación que se describe, se demuestra que determinados locutores pueden ser reconocidos con altas tasas de acierto empleando el perfil F0 como único rasgo biométrico.

Los resultados de reconocimiento que se obtienen son poco sensibles a cambios de micrófono y sesión. La entonación aparece como una propiedad del habla robusta ante este tipo de cambios, lo que parece recomendar su uso en combinación con otras propiedades acústicas para conseguir mejores tasas de reconocimiento en escenarios reales. Como continuación de este primer experimento sería recomendable, por tanto, analizar las posibilidades de integración de este sistema de reconocimiento con un sistema de reconocimiento de locutor estándar basado en rasgos fonéticos y disponible en el grupo de investigación [12]. Los resultados de esos experimentos futuros deberían proporcionar ya información determinante sobre las posibilidades del uso de la entonación en sistemas reales.

El hecho de que las tasas de error dependan fuertemente del tipo de grupo acentual empleado, sugiere la necesidad de seguir experimentando técnicas de reconocimiento que saquen partido de esta variabilidad para mejorar el funcionamiento del sistema. Con el experimento limitado que presentamos en este trabajo no podemos aún concluir con firmeza que la incorporación de la información prosódica sea menos adecuada o más costosa computacionalmente que otras técnicas más establecidas. Puede parecer una debilidad del método que aquí se propone que para reconocer una determinada locución de un usuario sea preciso contar primero con información sobre la clasificación prosódica de la misma. De cara a integrar un sistema de reconocimiento prosódico con otro fonético, sin embargo, es una ventaja innegable que puedan proponerse alternativas ponderadas sobre la naturaleza de la entonación que ayuden a seleccionar el reconocedor de segundo nivel (en una arquitectura en cascada) que

mejor se adapte a la muestra. Por ello, consideramos interesante seguir experimentando en esta línea para valorar adecuadamente estos extremos, por sí mismos prometedores.

Finalmente, interesa señalar que sería deseable disponer de un corpus más amplio, que incluya, por ejemplo, locutores que pertenezcan a ámbitos socioculturales diferentes, para obtener resultados más firmes para evidenciar la utilidad de la entonación en estos casos. Por lo mismo, sería necesario disponer de un corpus con un recubrimiento adecuado de los diversos tipos grupos acentuales y de entonación para entrenar modelos de reconocimiento particulares para los diferentes tipos de unidades prosódicas y poder plantear sistemas en los que se fusionen los resultados de dichos modelos. Desafortunadamente, la disponibilidad de este tipo de recursos lingüísticos es muy limitada, en comparación con la de corpus etiquetados fonéticamente.

AGRADECIMIENTOS

Los autores quieren testimoniar su reconocimiento al Dr. Marcos Faundez Zanuy, que puso el corpus EUPMT a nuestra disposición para realizar este estudio.

BIBLIOGRAFÍA

[1] SuperSID Team, *Supersid final report, exploiting high-level information for high-performance speaker recognition* Tech. Rep. WS02 DAR 8/26/2002, Workshop 2002, An NSF Sponsored Event. The Centre for Language and Speech Processing, August 2002.

[2] K. Bartkova, D. L. Gac, D. Charlet, D. Jouvet. *Prosodic Parameter for Speaker Identification*. Proceedings ICSLP 2002

[3] D. Escudero and V. Cardeñoso, *Corpus based extraction of quantitative prosodic parameters of stress groups in Spanish*. Proceedings of ICASSP 2002, Mayo 2002.

[4] D. Escudero, C. González, and V. Cardeñoso, *Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis in Spanish.* Proceedings of ICSLP 2002 , Mayo 2002.

[5] V. Cardeñoso and D. Escudero, *Statistical modelling of stress groups in Spanish.* Proceedings of Prosody 2002 , Abril 2002

[6] D. Escudero, *Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión Texto Voz*, Ph.D. thesis, Dpto. de Informática, Universidad de Valladolid, España, 2002.

[7] J. M. Garrido, *Modelling Spanish Intonation for Text-to-Speech Applications*, Ph.D. Thesis, Facultat de Lletres, Universitat de Barcelona, España, 1996.

[8] C. Alonso, M. Faundez, *Speaker Identifications in Mismatch training and Testing Conditions*, ICASSP 2000.

[9] D. Griffin, J. S. Lim. Multiband excitation vocoder. IEEE Trans.ASSP. vol. 36, N 8. August 1988.

[10] Ross Quinlan *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA. 1993

[11] I. H. Witten, E. Frank. *Data Mining*. Morgan Kaufmann Publishers 1999

[12] C.E. Vivaracho, J.O. Ortega, L. Alonso, and Q.I. Moro, *A comparative study of mlp-based artificial neural networks in text independent speaker verification against gmm-based systems* Proceedings of Eurospeech2001 , 2001.