MODELADO DE LA ENTONACIÓN EN EUSKERA UTILIZANDO EL MODELO DE FUJISAKI Y ÁRBOLES DE REGRESIÓN BINARIOS

Eva Navas, Inma Hernáez

(eva, inma@bips.bi.ehu.es)

Departamento de Electrónica y Telecomunicaciones

Euskal Herriko Unibertsitatea / Universidad del País Vasco

Resumen

En este trabajo se presenta el análisis realizado para la síntesis de la entonación vasca y los resultados obtenidos en una aplicación de conversión de texto a voz. Utilizando el modelo de Fujisaki, se han obtenido de forma manual y automática los parámetros de dicho modelo para dos bases de datos leídas. Los parámetros se han utilizado para entrenar árboles de regresión binarios. En el artículo se describen las variables utilizadas en el entrenamiento y su importancia en el modelado de la entonación. Los resultados obtenidos demuestran que el modelo de Fujisaki es un modelo válido para la lengua vasca.

1 Introducción

El conversor texto a voz aHoTTS utilizaba hasta el momento un sencillo modelo de entonación de "picos y valles" en el que a las sílabas acentuadas se les asignaban picos en la curva de entonación y entre ellos la curva evolucionaba linealmente formando los valles [1]. Las reglas utilizadas para determinar la declinación, así como para caracterizar las diferentes estructuras sintácticas eran también muy simples. Con objeto de mejorar la entonación resultante, actualmente estamos desarrollando el modelo de entonación de Fujisaki [2] que ya ha sido empleado para la síntesis de entonación en muchos otros idiomas [3][4][5][6][7][8][9]. En este artículo se describe el proceso completo seguido en la obtención de los modelos entonativos. En el siguiente apartado se describen las bases de datos utilizadas, junto con sus características más relevantes. El apartado 3 describe las herramientas utilizadas para el proceso de obtención de los parámetros del modelo. Los siguientes apartados están dedicados a la descripción de los resultados obtenidos en la predicción de los parámetros del modelo a partir de la información lingüística proporcionada por el texto, con las dos bases de datos utilizadas en nuestros experimentos. El apartado 5 resume las conclusiones del trabajo.

2 Corpus de voz

Para realizar los trabajos descritos en este artículo, se han utilizado dos bases de datos, de características muy diferentes. La tabla 1 resume las características más importantes de ambas bases de datos.

Con el objetivo principal de probar la validez del modelo de Fujisaki para el euskera, primeramente se utilizó el corpus *Bermeo* descrito en [10]. Este corpus está constituido por 348 frases simples enunciativas e interrogativas en las que el locutor era una mujer que hablaba en su propio dialecto. Esta base de datos había sido creada con el fin de realizar estudios de entonación, más concretamente para analizar la influencia de la posición del foco y las palabras con acento marcado en la curva de entonación total [11]. Todas las frases son cortas, están compuestas tan sólo por 2, 3 ó 4 sintagmas formados por palabras de 2, 3 ó 4 sílabas. En estas frases se combinaron diferentes posiciones del foco y de las palabras con y sin acento léxico.

El segundo corpus utilizado, que denominamos *Jokin* fue diseñado específicamente para este trabajo, y consta de 344 frases enunciativas, interrogativas y exclamativas, tanto simples como compuestas. En el diseño de este nuevo corpus se tuvo cuidado de incluir frases de diferente complejidad y longitud, que contuvieran pausas internas y todo

tipo de partículas que pudieran tener un efecto especial en la entonación de la frase. El vocabulario utilizado en su elaboración es muy amplio. Además y muy importante, este corpus fue escrito y leído en euskera estándar.

Las diferencias entre utilizar una base de datos dialectal – *Bermeo* – o una grabación de euskera estándar – *Jokin* –, y con las características citadas son muchas. Entre ellas podemos destacar las siguientes:

- El hablante en dialecto posee un modelo acentual y entonativo único y homogéneo, que aplica de forma uniforme y coherente a lo largo de toda la grabación, aún más si las frases enunciadas son simples. El hablante de euskera estándar aplica un modelo tanto acentual como entonativo desconocido, en el que por supuesto debe influir el modelo adquirido en casa (en el caso de nuestro locutor, los dialectos paterno y materno eran distintos...) pero en el que confluyen muchos otros factores, como el hecho de haber recibido educación en lengua vasca en diferentes centros (y por tanto muy variado), y la influencia de las lenguas mayoritarias próximas (el castellano), cuyo modelo se tiende a aplicar cuando aquello que se lee no se corresponde directamente con la expresión que se utilizaría en el dialecto. El locutor de la base de datos *Jokin* posee una entonación cuidada, y que puede considerarse *vasca*, pero no se corresponde con ningún dialecto específico.

- El dialecto de Bermeo (como muchos otros dialectos) ha sido estudiado intensamente por los lingüistas, y el modelo acentual utilizado es conocido. Este dialecto se clasifica entre las variedades conocidas como *pitch-accent* [12] y existe un conjunto de reglas que determinan sin ambigüedades las posiciones de los acentos, la formación de grupos acentuales (grupos de palabras que se encuentran bajo la influencia de un único acento), y la posición del sintagma focal, cuando los sintagmas y la estructura sintáctica de la frase son sencillas. Como las estructuras sintácticas de *Bermeo* son simples y homogéneas, el etiquetado de estas características lingüísticas, que se sabe que afectan a la entonación final, puede realizarse de forma totalmente automática.

- Para el euskera estándar no existe un modelo acentual definido, y aunque existen diversas propuestas [13], no están extendidas en la práctica. Así, para realizar el etiquetado de los acentos, caben varias alternativas. Por ejemplo, podría pensarse en realizar un etiquetado manual de los acentos, escuchando y/o visualizando las curvas de F0 obtenidas. Sin embargo es un proceso costoso, en el que se debe considerar más de un etiquetador, y naturalmente se producen discrepancias en el etiquetado obtenido. La solución adoptada fue un compromiso: se tomó nota de las palabras que se realizaban como acentuadas en la primera sílaba (consideradas marcadas léxicamente), y en el resto de los casos, se realizó acentuación automática siguiendo la propuesta existente, con la implementación descrita en [14]. Esto proporciona buenos resultados para la acentuación de sintagmas simples, pero se producen muchos casos de duda (por ejemplo en palabras con sufijación larga, verbos declinados, partículas especiales etc.).

- Igualmente problemático resulta el etiquetado del foco o sintagma focal en la frase. Si en una frase simple la posición del foco queda fijada por la sintaxis en la mayoría de los casos, cuando la frase se complica, se producen muchas discrepancias al interpretar el texto escrito.

- Finalmente, el corpus *Jokin* posee frases largas, con pausas intermedias. Aunque se trató de colocar los signos ortográficos adecuadamente de modo que el locutor hiciera las pausas de forma controlada, quedó demostrado que no es posible controlar totalmente la ubicación de las pausas, si se pretende mantener un grado de naturalidad aceptable en la grabación. Este aspecto, junto con el anterior comentado, será objeto de un análisis más profundo en apartados posteriores.

Todas las frases se han segmentado automáticamente a nivel de frase, y manualmente a nivel de palabra. Posteriormente se segmentaron en fonemas empleando síntesis y un algoritmo de proyección dinámica [15][16]. El etiquetado de sílabas y otras unidades (palabras y grupo acentuales) se realizó de forma automática utilizando el módulo de análisis lingüístico del conversor de texto a voz.

Las curvas de entonación fueron calculadas utilizando un método basado en [17].

	Bermeo	Jokin
Dialecto	Bermeo	batua
Tipo	leída	leída
Grabación	casa	laboratorio
Propósito	entonación	entonación
Nº frases	348	344
Nº palabras	1376	2510
MB	7.5	64

Tabla 1: Características generales de las bases de datos utilizadas.

3 Obtención de los parámetros de entonación

El proceso de obtención de los parámetros de entonación se realizó primeramente de forma totalmente manual para la base de datos *Bermeo*. La parametrización manual se realizó con la ayuda de una herramienta gráfica diseñada específicamente para realizar esta tarea [10]. El programa representa dinámicamente la curva de entonación correspondiente a la combinación de parámetros seleccionada y permite la inmediata evaluación visual de los resultados. Además, lleva incorporado un *vocoder LPC*, de forma que la curva original y sintética de pitch pueden ser intercambiadas sobre la señal codificada y ambas señales codificadas pueden ser reproducidas, permitiendo una inmediata evaluación perceptual de los parámetros que se están editando. Por otro lado, se pueden representar también marcas de tiempo y etiquetas asociadas a las mismas.

3.1 Criterios para la parametrización manual

Después de parametrizar manualmente un número significativo de frases de cada tipo contenido en la base de datos *Bermeo*, se observó que las marcas de grupo acentual, las posiciones de los acentos y el tipo de frase eran especialmente significativos para fijar el rango de variación de cada parámetro. Teniendo esto en cuenta, se dedujeron un conjunto de reglas o criterios que se aplicaron para realizar la parametrización de la totalidad de las frases del corpus:

- Cada frase puede ser parametrizada empleando una única delta que proporciona la pendiente global de la curva de F0. La posición de esta delta se fijó en 320 ms antes del comienzo de la frase. Su amplitud era arbitraria.
- Cada grupo acentual puede ser modelado por un único pulso. La posición de este pulso depende fundamentalmente de la posición del acento dentro del grupo acentual:

  - Si el acento recae en la primera sílaba del grupo acentual, el pulso puede comenzar antes del comienzo del grupo acentual y acaba antes del final del grupo acentual.

  - Si el acento recae en la segunda sílaba del grupo acentual, el pulso comienza y termina dentro del grupo acentual.
- En el caso de las frases interrogativas y con el fin de modelar la pendiente positiva final de la curva de F0, se añade un pulso final extra.
- Existe una mínima distancia entre los pulsos (nunca se solapan), así como una duración mínima.

3.2 Proceso automático de parametrización

Para obtener un buen modelo de entonación para el euskera, es decir uno que sea lo suficientemente representativo, se deben analizar y parametrizar un número elevado de frases. Por ello era necesario desarrollar un método automático que acelerara el proceso de parametrización. La extracción automática de los parámetros se realiza mediante *Análisis por Síntesis*. Este proceso automático trata de obtener F0min y los parámetros que definen los comandos de frase y de acento del modelo de Fujisaki que generan la mejor curva sintética según el criterio de menor error cuadrático medio (mínimo *MSE Mean Squared Error*) con la curva natural. En el proceso de cálculo del error no se consideran las tramas sordas.

Para obtener la mejor aproximación y encontrar la solución óptima, se realiza una búsqueda exhaustiva. Este proceso implica probar muchos valores para cada parámetro y el tiempo necesario para llevarlo a cabo resulta prohibitivo si no se aplican determinadas restricciones y se limita la búsqueda únicamente a las combinaciones de parámetros que estén permitidas. Las restricciones aplicadas se dedujeron del proceso de parametrización manual, y son las siguientes:

- Cada frase o sentencia queda parametrizada empleando una única delta que proporciona la pendiente global de la curva de F0. La posición de esta delta se fijó en 320 ms antes del comienzo de la frase. Su amplitud varía únicamente con el tipo de frase, y los valores obtenidos en la parametrización manual eran tan uniformes, que en *Bermeo* se dejó fija una vez determinado el tipo de frase.
- Cada grupo acentual es modelado por un único pulso. La posición de este pulso depende únicamente de la posición del acento dentro del grupo acentual:

  - Si el acento recae en la primera sílaba del grupo acentual, el pulso puede comenzar antes del comienzo del grupo acentual y acaba antes del final del grupo acentual.

  - Si el acento recae en la segunda sílaba del grupo acentual, el pulso comienza y termina dentro del grupo acentual.
- En el caso de las frases interrogativas, se añadirá un pulso final extra, con el fin de modelar la pendiente positiva final de la curva de F0. Este pulso comienza una vez ha transcurrido el 60% del último grupo acentual.
- Se establece una mínima distancia entre los pulsos de 20ms, así como una duración mínima de los mismos de 100ms.
- Se limita el dominio de búsqueda aplicando cuantificación a los valores de los parámetros. Así, las duraciones se varían por pasos de 30ms, y las amplitudes de los pulsos por pasos de valor 0,05. Además, se proporcionan valores máximos y mínimos para todos los parámetros. Esta cuantificación se hizo necesaria para reducir los tiempos de ejecución del programa (aproximadamente 9 minutos de CPU en un Pentium II a 300MHz, para parametrizar 1 segundo de voz, con las restricciones mencionadas).

La figura 1 muestra la estructura del proceso automático de extracción de parámetros.

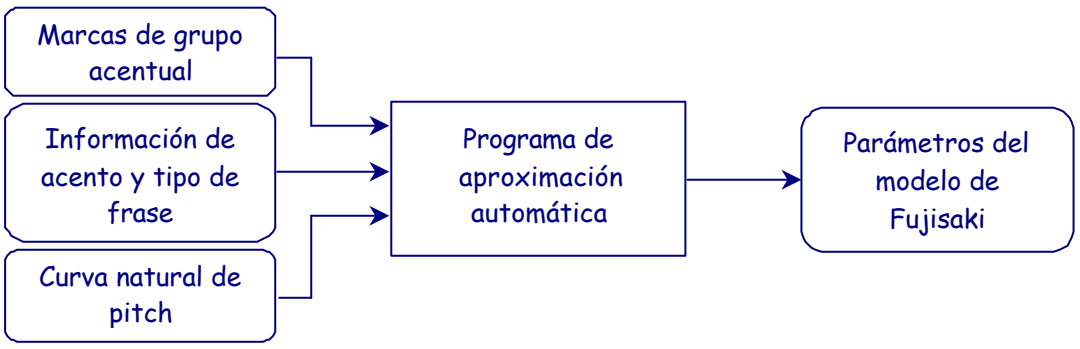

Figura 1: Estructura del proceso de etiquetado automático.

La valoración de la calidad de la parametrización obtenida de forma automática no es sencilla. En un intento de obtener una valoración numérica de los resultados, se ha medido la diferencia media cuadrática entre los valores de los parámetros obtenidos de forma manual y los obtenidos automáticamente. La tabla 2 muestra los resultados de dicha medida:

	Amplitud del pulso	Posición del pulso	Duración del pulso
Error medio	0.1087	109.155 ms	58.027 ms
Error medio relativo	0.414	0.8782	0.2762

Tabla 2: Diferencia entre los valores de los parámetros obtenidos manual y automáticamente.

Como puede verse, el error relativo cometido toma un valor muy alto (de ¡hasta un 90% en las posiciones de los pulsos!). Sin embargo, debemos considerar la influencia de las restricciones impuestas al sistema automático. Así por ejemplo, el hecho de fijar los valores de las amplitudes de las deltas, debe de ser compensado por el sistema modificando el resto de los parámetros que puede variar.

A pesar de estos valores, que pueden parecer desalentadores, la inspección visual y auditiva de los resultados es sorprendentemente positiva. Obsérvese por ejemplo el parecido en las curvas obtenidas en los ejemplos que se muestran en la figura 2. En ellas se observa que pese a la clara diferencia en el juego de parámetros elegido en cada caso: en la parametrización manual la delta tiene mayor amplitud, el primer pulso comienza más tarde y es más corto que en la parametrización automática y el último pulso es también bastante diferente, tanto la curva sintética generada a partir de los parámetros manuales, como la obtenida a partir de la parametrización automática se ajustan muy bien a la curva natural.

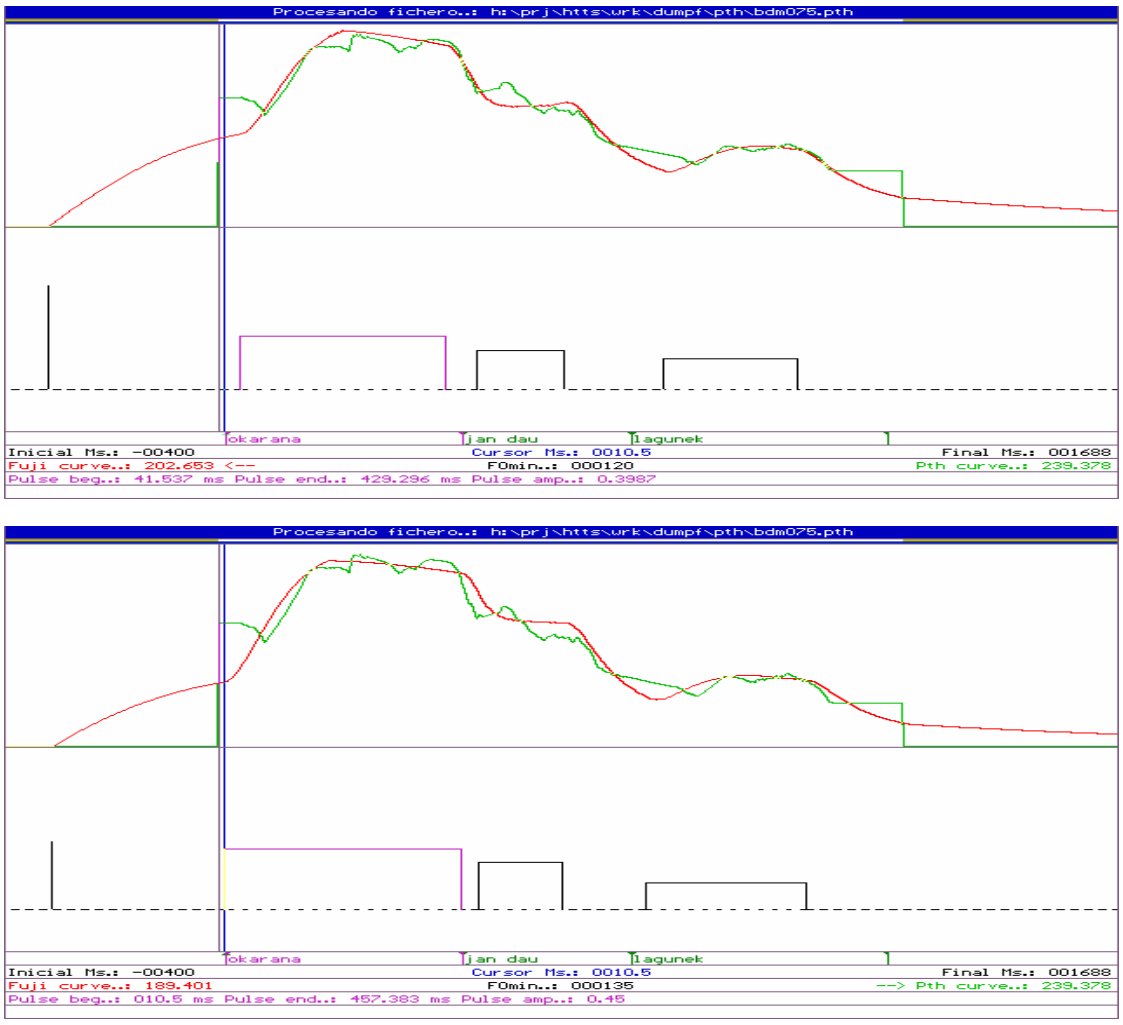

Figura 2: Comparación del etiquetado de Fujisaki manual (gráfica superior) y automático (gráfica inferior) de la frase *"Okarana jan dau lagunek"*.

3.3 Análisis para el euskera estándar

Tras el estudio realizado con la base de datos *Bermeo*, y una vez probada la validez del modelo de Fujisaki para el euskera, surge la necesidad de extender el estudio a frases compuestas, que incluyeran pausas y partículas que pueden tener una entonación especial, y especialmente, realizar el estudio con euskera estándar, y no con una variedad dialectal. Con este fin se diseñó el corpus *Jokin*.

En esta base de datos, aparecen pausas en el interior de las frases, circunstancia que no se daba en *Bermeo*. A pesar de que las frases habían sido cuidadosamente puntuadas, se daban los siguientes hechos:

a) No todas las pausas que se realizan corresponden a signos ortográficos del texto.

b) Un mismo signo ortográfico no produce siempre el mismo tipo de pausa. Así, una coma provoca por lo general una reinicialización (al menos parcial) del proceso entonativo, requiriendo por tanto una nueva 'delta' en el modelo, pero no es así en una enumeración, por ejemplo.

c) El locutor puede haber omitido pausas indicadas por signo ortográfico (por considerarlas innecesarias).

Para analizar las consecuencias de estos aspectos sobre el proceso de parametrización, se realizó un etiquetado y clasificación manual de las pausas realizadas sobre las frases, indicando así por un lado las pausas realizadas, y por otro lado, el tipo de las pausas entendido como *'pausas que requieren delta'* y *'pausas que NO requieren delta'*. El resultado más destacado de esta clasificación fue que un 22% de las 'comas' ortográficas no requerían delta y un 36% de las *pausas que requieren delta* no llevaban signo ortográfico. La tabla 3 muestra la distribución de pausas con y sin signo ortográfico, clasificadas según el criterio anterior.

	Con signo ortográfico	Sin signo ortográfico
Con delta	105	58
Sin delta	30	4

Tabla 3: Distribución de las pausas en la base de datos *Jokin*.

Se realizaron tres experimentos de parametrización automática, bajo las siguientes condiciones:

- **Experimento I**: Se introdujo una delta en todas las pausas 'declaradas' ortográficamente por un signo ortográfico de pausa (,) independientemente de cuál hubiera sido su realización.
- **Experimento II**: Se introdujo una delta en todas las pausas, estuvieran éstas indicadas por un signo ortográfico o no, realizando previamente un etiquetado de las mismas.
- **Experimento III**: Se introdujo una delta exclusivamente en aquéllas pausas que previamente habían sido clasificadas como 'pausas necesitadas de delta' (estuvieran o no indicadas por un signo ortográfico).

Como criterio para el establecimiento del 'mejor experimento' se tomó el valor medio del MSE cometido a lo largo de toda la base de datos. La gráfica 3 muestra los resultados.

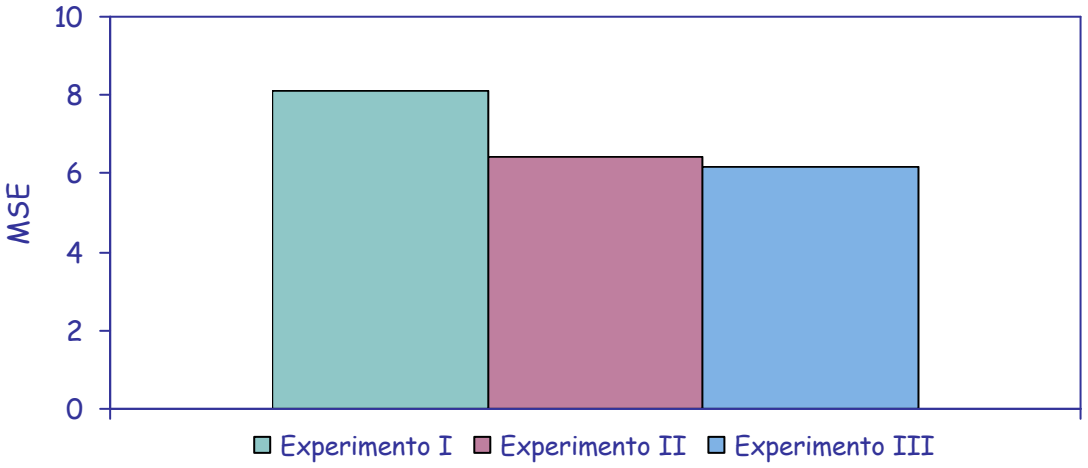

Figura 3: Comparación del error obtenido en cada experimento.

Como puede verse en la figura, efectivamente se obtienen curvas más ajustadas si las deltas se introducen en los puntos necesarios y no únicamente en las pausas indicadas por signo ortográfico. La poca diferencia entre los resultados de los casos II y III se debe por un lado al pequeño número de *pausas que no requieren delta* no indicadas por signo ortográfico existente en el corpus y por otro lado al hecho de que las pausas con signo ortográfico que no requieren delta corresponden en todos los casos a frases muy cortas, formadas por un solo sintagma (enumeraciones) y por tanto el efecto de insertar una delta puede compensarse con una disminución de la amplitud del pulso.

El experimento muestra claramente la importancia de disponer de un sistema automático de inserción de pausas en el texto. Por otro lado según estos resultados no resulta de interés clasificar las pausas. Sin embargo esto puede muy bien deberse a las características entonativas del locutor o al diseño del corpus, ya que el número de casos aunque significativo puede no ser representativo.

4 Estimación de los parámetros a partir del texto

El análisis estadístico de los parámetros se ha realizado utilizando un software comercial para entrenar árboles de regresión binarios [18]. Los árboles binarios de clasificación y regresión o CARTs (*Classification And Regression Trees*) presentan varias ventajas, entre ellas que seleccionan automáticamente los factores más influyentes en la predicción de cada variable y que proporcionan estimaciones bastante fiables. Además la estructura en árbol produce diagramas fáciles de interpretar.

4.1 Resultados para la base de datos dialectal

En la base de datos *Bermeo* las posiciones y amplitudes obtenidas para los comandos de frase o deltas eran muy uniformes, así que no se construyó ningún árbol para predecirlas. Cada parámetro correspondiente a un pulso está directamente relacionado con las características del grupo acentual con el que se corresponde. Las características extraídas del texto son las siguientes:

- Posición que ocupa el grupo acentual actual desde el comienzo del enunciado. Este dato se proporciona al árbol tanto de forma absoluta como normalizado al número de grupos acentuales de la sentencia o frase y del enunciado completo.
- Instante de inicio del grupo acentual medido en milisegundos, tanto absoluto como normalizado a la duración de la sentencia y del enunciado completo.
- Instante de finalización del grupo acentual expresado en milisegundos, tanto absoluto como normalizado a la duración de la sentencia y del enunciado completo.
- Duración del grupo acentual en milisegundos, tanto absoluta como relativa a la duración de la sentencia y el enunciado.
- Distancia desde el comienzo del grupo acentual hasta el acento de ese grupo, medida en milisegundos. Este dato se da tanto de manera absoluta como relativa a la duración de la sentencia y del enunciado.
- Grupo acentual foco.
- Número de grupos acentuales entre el actual y el foco.
- Distancia en milisegundos entre el foco y el grupo acentual actual, tanto absoluta como normalizada a la duración de la sentencia y del enunciado.
- Tipo de acento del grupo acentual que puede ser normal (acento en la segunda sílaba del grupo) o marcado (acento en la primera sílaba del grupo).
- Posición en milisegundos del acento del grupo acentual.
- Distancia en milisegundos desde el comienzo del grupo acentual al acento del mismo, expresado tanto de forma absoluta como relativa a la duración del grupo acentual y a la de la sentencia.
- Tipo de enunciado, que puede tomar los valores enunciativo, interrogativo, exclamativo, neutro o de pausa.

- Tipo de pulso, que indica si es el último pulso de una interrogativa o exclamativa, el anteúltimo pulso de una interrogativa o exclamativa u otro pulso cualquiera.
- Índice del pulso: número decreciente que indica cuántos pulsos faltan de ser predichos en la frase.

Para la base de datos *Bermeo* se construyeron dos juegos de árboles, uno con los parámetros obtenidos del proceso de parametrización manual y otro con los parámetros resultantes de la parametrización automática. Las curvas predichas por estos árboles se pueden ver en la figura 4, que muestra también las curvas obtenidas en los procesos de aproximación manual y automáticas para la misma frase.

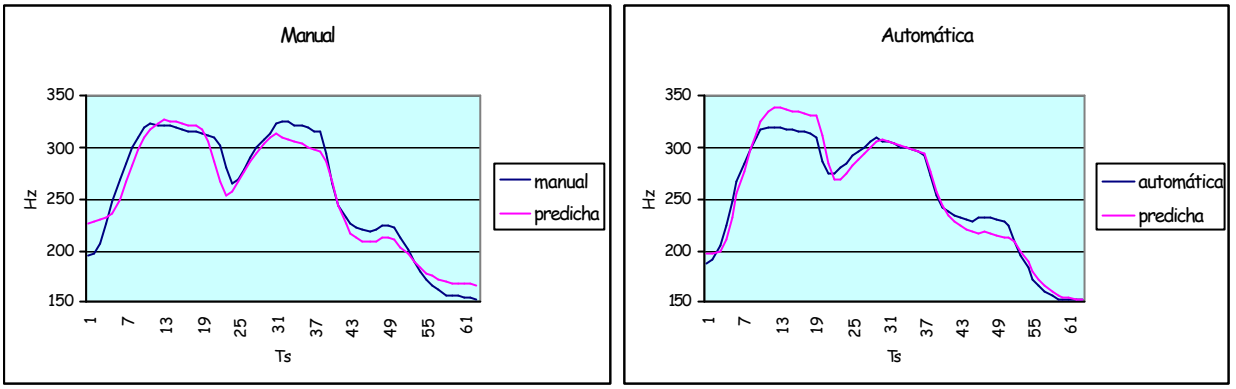

Figura 4: Comparación entre las curvas de F0 generadas con los parámetros predichos y con los calculados de forma manual o automática para una frase declarativa con tres grupos acentuales y sin palabras marcadas.

En la tabla 4 se muestran las variables utilizadas en la predicción, ordenadas por orden de importancia, en el árbol entrenado con parámetros resultantes del proceso de parametrización automática. Como puede observarse, para determinar la posición de los pulsos es esencial la información del acento, y con mucha menor importancia la posición del foco. Los árboles de amplitud muestran que este parámetro depende principalmente de la posición del foco y en los árboles de duración se ve que la duración del pulso depende de los instantes de comienzo y fin del grupo acentual, de la posición del grupo acentual en la frase y de la duración del grupo acentual.

Árboles de posición		Árboles de amplitud		Árboles de duración	
variable	importancia	variable	importancia	variable	importancia
distancia comienzo GA-acento GA	100.00%	distancia entre foco y GA normalizada a duración sentencia	100.00%	instante fin GA normalizado a duración sentencia	100.00%
distancia comienzo GA-acento GA normalizada a duración GA	79.18%	distancia entre foco y GA	69.89%	posición GA normalizada a nº GAs sentencia	94.64%
distancia comienzo GA-acento GA normalizada a duración sentencia	76.71%	nº GAs hasta el foco	67.61%	instante inicio GA normalizado a duración sentencia	64.31%
distancia entre foco y GA normalizada a duración sentencia	18.99%	instante fin GA normalizado a duración sentencia	58.67%	tipo de pulso	19.86%
distancia entre foco y GA	17.82%	índice de pulso	44.66%	duración GA	16.91%
nº GAs hasta el foco	17.36%	posición acento	24.94%	duración GA normalizada a duración sentencia	11.44%

Tabla 4: Importancia de las variables en la predicción de los pulsos de la base de datos *Bermeo*.

4.2 Resultados para la base de datos en euskera estándar

En un primer experimento con la base de datos *Jokin*, se entrenaron los árboles binarios con el mismo juego de variables empleado para la base de datos dialectal. Los parámetros del modelo de Fujisaki empleados se corresponden con los obtenidos del Experimento I descrito en el apartado 3.

En la base de datos dialectal *Bermeo*, al tratarse de frases muy simples, la posición del sintagma focal en la frase era sencilla de obtener por regla (en estructuras sintácticas simples, el foco se sitúa siempre delante del verbo, o es el propio verbo). En frases más complejas, la localización del foco es también más complicada. En un experimento informal, se pidió a dos lingüistas que marcaran el sintagma focal sobre las frases escritas de la base de datos *Jokin*. En un porcentaje muy elevado, no había coincidencia total, además de no coincidir con la decisión tomada por el locutor (es decir, con el sintagma focal *entonado*). Además, uno de los etiquetadores sintió la necesidad de definir un foco secundario, lo cual añadía un factor más de complejidad.

Por otro lado, es poco realista pensar que en las condiciones en que se realiza actualmente el análisis sintáctico en nuestro sistema de conversión de texto a voz, dicho módulo sería capaz de definir el foco de la frase con la suficiente fiabilidad. Por ello, decidimos omitir totalmente en el entrenamiento de los árboles la información de foco.

Además, dado que en esta base de datos las frases eran más complejas y muchas de ellas incluían pausas, hubo que considerar variables que las tuvieran en cuenta. Los resultados obtenidos con este experimento se muestran en la tabla 5. En este caso para predecir la posición de los pulsos el factor fundamental es el tipo de pulso, ya que los pulsos finales de las exclamativas e interrogativas comienzan una vez ha transcurrido el 60% del último grupo acentual, y el resto lo hace en las proximidades del acento del grupo acentual (por ello el segundo factor es la posición del acento). En los árboles de amplitud la variable más significativa es el tipo de sentencia, siendo menores los pulsos de las enunciativas que los del resto de las sentencias. En la predicción de la duración de los pulsos es la posición del acento la que tiene más peso, seguida del tipo de pulso de que se trate y del tipo de sentencia.

Árboles de posición		**Árboles de amplitud**		**Árboles de duración**	
variable	**importancia**	**variable**	**importancia**	**variable**	**importancia**
tipo de pulso	100.00%	tipo de sentencia	100.00%	posición acento	100.00%
posición acento	32.61%	índice de pulso	48.01%	tipo de pulso	65.61%
índice de pulso	30.18%	duración GA normalizada a duración sentencia	22.13%	tipo de sentencia	56.13%
tipo de sentencia	25.28%	duración GA	13.00%	nº de AGs en la sentencia	39.55%
tipo de acento	16.85%			índice de pulso	34.42%
duración GA normalizada a duración sentencia	13.13%			duración GA	22.62%

Tabla 5: Importancia de las variables en la predicción de los pulsos de la base de datos *Jokin*, sin tener en cuenta la información de foco.

Las deltas en la base de datos de euskera estándar presentaban valores muy diferentes, por lo que en este caso era necesario también obtener un árbol para el cálculo sus amplitudes. Los resultados se muestran en la tabla 6, en la que puede apreciarse que la variable más importante en la predicción es el tipo de sentencia, siendo la amplitud mayor en las exclamativas e interrogativas que en el resto de sentencias, tal y como se había observado trabajando con la base de datos dialectal. Además la amplitud de las deltas de un mismo enunciado suele ser decreciente, por lo que el segundo factor que el árbol considera es el número de sentencia en el que se quiere colocar la delta.

Árboles de amplitud de la delta	
variable	**importancia**
tipo de sentencia	100.00%
nº de sentencia	46.58%
nº de AGs en la sentencia	11.06%
duración de la sentencia	7.98%

Tabla 6: Importancia de las variables en la predicción de la amplitud de la delta en la base de datos *Jokin*.

5 Conclusiones

En este trabajo hemos presentado los resultados obtenidos en el modelado de la entonación para la lengua vasca. Tras su realización queda probado que el modelo de entonación de Fujisaki es un modelo válido para la entonación del euskera, con el que se pueden conseguir unas curvas de entonación de calidad aceptable.

Para estudiar mejor el modelo se han desarrollado herramientas que permiten una fácil comprensión del significado de cada parámetro del modelo y una rápida evaluación de los cambios en los mismos. Además se ha desarrollado un sistema completamente automático para el etiquetado de la entonación según el modelo de Fujisaki.

Dicho modelo ha sido introducido en el conversor de texto a voz, mejorando notablemente la calidad de la entonación con respecto al modelo simple anteriormente existente. En particular, el modelo obtenido modelo ofrece una calidad satisfactoria para las frases enunciativas, presentando más problemas para el modelado de frases interrogativas y exclamativas, ya que éstas se encuentran menos representadas en la base de datos.

Todas las variables utilizadas para la predicción, han sido aquéllas que el módulo lingüístico de nuestro actual sistema de conversión de texto a voz es capaz de predecir. No cabe duda de que la mejora del sistema de etiquetado lingüístico mejoraría los resultados para la curva de entonación. En particular, el problema más acuciante es la adecuada ubicación de pausas en el texto, y su correcta clasificación.

6 Agradecimientos

Este trabajo se ha realizado con la subvención de la UPV/EHU, dentro del proyecto de código UPV 147.345-TA066/98.

Además queremos agradecer a Iñaki Gaminde el diseño y la grabación de la base de datos dialectal y a los alumnos del laboratorio su trabajo etiquetando ambas bases de datos.

7 Referencias

[1] I. Hernáez, J.C. Olabe, A. Cuesta, R. Gandarias, P. Etxeberria
Improving naturalness in a Text-to-Speech Conversion System for the Basque Language
7th Mediterranean Electrotechnical Conference, pp. 61-64, Antalya Turkiye 1994

[2] H. Fujisaki, K. Hirose
Analysis of voice fundamental frequency contours for declarative sentences of Japanese
Journal of Acoustic Society. Jpn. (E) 5, 4. 1984

[3] H. Fujisaki, S. Ohno
Analysis and modelling of fundamental frequency contours of English utterances
Proceedings of Eurospeech'95, pp. 985-988. Madrid, 1995.

[4] H. Mixdorff, H. Fujisaki
A scheme for a model-based synthesis by rule of F0 contours of German utterances
Proceedings of Eurospeech'95, pp. 1823-1826. Madrid, 1995.

[5] B. Möbius, M. Pätzold, W. Hess
Analysis and synthesis of German F0 contours by means of Fujisaki's model

Speech Communication vol. 13, pp. 53-61, 1993.

[6] H. Fujisaki, S. Ohno, T. Yagi
Analysis and modelling of fundamental frequency contours of Greek utterances
Proceedings of Eurospeech'97, pp. 465-468. Rhodes, 1997.

[7] C. Wang, H. Fujisaki, S. Ohno, T. Kodama
Analysis and synthesis of the four tones in connected speech of the standard Chinese based on a command-response model
Proceedings of Eurospeech'99, pp. 1655-1658. Budapest, 1999.

[8] M. Ljungqvist, H. Fujisaki
Generating intonation for Swedish text to speech conversion using a quantitative model for the F0 contour
Proceedings of Eurospeech'93, pp. 873-876. Berlín, 1993.

[9] H. Fujisaki
The fundamental frequency contour of speech – its modelling, underlying mechanism and application to multilingual speech
Proceedings of ICSP'99, pp. 19-26, Seoul 1999.

[10] B. Etxebarria, E. Navas, A. Armenta, I. Madariaga, I. Gaminde, I. Hernáez
Tools and Basque language databases developed in the AhoLab laboratory
Second International Conference on Language Resources and Evaluation. Workshop Proceedings, pp.62-70, Atenas 2000

[11] G. Elordieta, I. Gaminde, I. Hernáez, J. Salaberria, I. Matín de Vidales
Another step in the modelling of Basque intonation: Bermeo
Lecture Notes in Computer Science; Vol 1692: Lecture Notes in Artificial Intelligence pp 361-364, 1999

[12] I. Gaminde
Los tipos de acento del dialecto vizcaíno del euskera: aproximación acústica
Phonetica, Serie Lingüística, Vol. 6 pp. 11-42

[13] J.I. Hualde
Euskal azentuak eta euskara batua
Euskera XXXIX, pp. 1549-1568

[14] I. Hernáez
Conversión de texto a voz para el euskara basada en un sintetizador de formantes
Tesis doctoral, 1995

[15] H. Sakoe, S. Chiba
A dynamic programming approach to continuous speech recognition
Proc. Int. Congr. Acoust. Budapest, Hungary, Rep 20-C-13. 1971

[16] L.R. Rabiner / R.W Schafer
Digital Processing of Speech Signals
Prentice Hall, 1978

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

[18] L. Breiman, J.H. Friedman, R.A. Olsen, C. J. Stone
Classification and Regression Trees
ChapmanHall, 1984