TESTEO DE 3 PROCEDIMIENTOS DE OBTENCIÓN DEL PITCH PARA LA MODELIZACIÓN PROSÓDICA DEL DISCURSO NOTICIA

Lluís Mas Manchón (Lluis.Mas@uab.cat)
LAICOM (Laboratori d'Anàlisi Instrumental de la Comunicació)
Universitat Autònoma de Barcelona

Resumen

Con el objetivo general de conseguir formalizar numéricamente la entonación de las locuciones de los informativos, se necesita un modelo entonativo aplicado inserto en un modelo de la eficacia comunictiva. En el presente artículo se ha definido un protocolo de análisis y representación de la entonación a partir de tres modelos (*Melódico*, *MOMEL* y *TOBI*), y se ha testeado el primer paso de dicho protocolo: la selección de datos de pitch distribuidos en unidades sílaba para su representación. Para ello, se ha comparado el procedimiento del Análisis Melódico del Habla de Francisco José Cantero (selección del punto medio visual de la sílaba en el espectro) con otros dos procedimientos: la selección de la sílaba completa y obtención de la media aritmética de los datos, y la selección de todo la locución y obtención de las subsiguientes medias aritméticas de los datos correspondidos con la división en sílabas. La indiferencia de uso de cualquiera de los procedimientos nos permitirá utilizar el más conveniente en términos operativos o técnicos.

PALABRAS CLAVE: entonación, pitch, locución, análisis acústico, representación

Resum

Amb l'objectiu general d'aconseguir formalitzar numèricamen l'entonació de les locucions dels informatius, ens cal un model entonatiu aplicat insert en un model de l'eficàcia comunicativa. A present article hem definit un protocol d'anàlisi i representació de l'entonació a partir de tres models (*Melòdic*, *MOMEL* y *TOBI*), i hem testejat el primer pas d'aquest protocol: la selecció de dades de pitch distibuides en unitats sílaba per a la seva representació. Hem comparat el procediment de l'Anàlisi Melòdica de la Parla de Francisco José Cantero (selecció del punt mitjà visual de la sílaba a l'espectre) amb d'altres procediments: la selecció de la sílaba completa i obtenció de la mitjana aritmètica de les dades, i la selecció de tota la locució i obtenció de les mitjanes aritmètiques de les dades corresponents a la divisió de sílabes. La indiferència d'ús de qualsevol dels procediments ens permetrà utilitzar el més convenient en termes operatius o tècnics.

PARAULES CLAU: entonació, pitch, locució, anàlisi acústica, representació

Abstract

With the objective of formalizing numerically the intonation of the locutions in the News Broadcast, it is required an applied intonative model inserted in a model of efficacy in communication. In this article, a protocol of analysis and representation of intonation has been defined from three models (*Melódico*, *MOMEL* and *TOBI*) , and the first step of this protocol has been tested: selection of the pitch data distributed in units "syllable" for its representation. For doing so, we have compared the original proceeding from the Melodic Analysis of Intonation of Francisco José Cantero (selection of the middle visual datum in the syllable spectrum) with two other proceedings: selection of a complete syllable and acquisition of the data arithmetic mean, and selection of the whole locution and acquisition of the subsequent arithmetic means of the data associated to the syllable division. The indifference of use of any of those proceedings permits us to make use of the most convenient one from an operative or technical standpoint.

KEYWORDS: intonation, pitch, locution, acoustic analysis, representation

Introducción

El presente artículo presenta uno de los estadios de una investigación mucho más amplia: la búsqueda de parámetros sonoros para la segmentación automática de noticias de televisión y su clasificación automática en temas. La segunda cuestión ha quedado resuelta con la presentación de 1000 palabras claves, el reconocimiento de las cuales a partir de procesos de Word Spotting[1] permitirá clasificar cualquier noticia en 15 temas en tiempo real[2]. Por su parte, la segmentación automática de noticias está siendo investigada en estos momentos, de cuyo avance nace este artículo.

Por tanto, el objetivo general es muy claro: encontrar ciertos parámetros sonoros del discurso informativo o cierta variación de dichos parámetros, correlacionados con el cambio de noticia. En nuestro abordaje de la prosodia hemos encontrado grandes evidencias de parámetros claves para marcar el final de frases y temas en el discurso: variación de pitch, alargamiento de vocales, pausas, distribución de picos tonales… (Vaissiere, 1988, Ostendorf y Swerts, 1997, Sönmez, 1998, Hirschberg, 1999, Shriberg et al, 1999, 2000, 2001), por lo que estamos convencidos de trabajar en la buena dirección. Sin embargo, llegado el momento de considerar como variables propias del discurso informativo a estos rasgos prosódicos, nos encontramos huérfanos de una fenomenología, metodología y técnica de análisis y representación de la entonación aplicada al discurso de los informativos de televisión. Es decir, no existe un modelo

[1] Reconocimiento aproximativo de las formas sonoras aisladas en un discurso, previo entrenamiento del locutor, que según parece está experimentado un gran desarrollo en el modelado de la estructuras entonacionales también (Taylor, 1997).
Aunque nuestro objetivo es hacer un sistema de segmentación y clasificación independiente del locutor, el sistema de Word Spotting aun no ha resuelto esta cuestión. Esto no impedirá que nuestra investigación esté condicionada por el objetivo de que el reconocimiento sea algún día totalmente independiente del locutor.

[2] Esta parte de la investigación constituyó nuestro trabajo de investigación de doctorado ("tesina"), que puede ser consultado online en http//:laicom.uab.es

entonativo que se ajuste a este tipo de discurso y a nuestros objetivos de análisis. Por ejemplo, necesitamos un modelo que nos permita relacionar la prosodia con el ritmo de locución informativa, así como un mismo modelo que nos permita comparar contornos de larga duración y de diferentes locutores a partir de datos objetivos de Pitch.

En nuestra tarea de trazar las coordenadas de un modelo entonativo adaptado a nuestros intereses, nos encontramos con la problemática de elegir un procedimiento de obtención de datos de pitch. Los programas ofrecen curvas entrecortadas de elementos de sonoridad múltiples, y aunque las propuestas van desde la utilización de todos los datos de pitch hasta la elección periódica de x datos cada x tiempo, lo cierto es que el procedimiento debe someterse a pruebas de rigor científico, satisfacer ciertos condicionantes metodológicos y ser congruente con una filosofía de trabajo propia de la investigación aplicada en comunicación. En este artículo se comparará experimentalmente tres procedimientos de tratamiento de los datos de pitch obtenidos mediante el analizador Praat. Estos procedimientos buscan la generación de un modelo para el tipo de discurso informativo.

Por tanto, de entrada decimos que aunque el objeto de estudio y las técnicas a utilizar pertenezcan al ámbito de la prosodia, en el campo de la fonética y fonología, la nuestra será una perspectiva eminentemente comunicológica

Bases teóricas del modelo

Partiendo de las opciones teóricas y técnicas más utilizadas para el análisis y representación de la entonación, hemos adoptado unas filosofías de trabajo y medidas concretas para iniciar el camino de creación de un Modelo de Análisis Melódico del Discurso Informativo. En esta primera etapa de la investigación, nuestra meta es encontrar el procedimiento de análisis más riguroso y las coordenadas de un algoritmo de modelado automático, ya que nuestros resultados deberán ser susceptibles de automatización.

Tanto el objetivo de automatización como el de creación de una Teoría de la entonación, exigen un modelo de correspondencia articulatoria, acústica y perceptiva, esto es, un modelo fonético y acústico de la entonación. Lo primero, por tanto, es tener en cuenta los rasgos propios de la locución informativa. Como hemos dicho, nuestra perspectiva es comunicológica o discursiva, lo que significa que, apoyándonos en nuestro conocimiento experto del mensaje y en la estabilidad del mismo (ha sido creado con unos objetivos muy concretos), iremos de las funciones del discurso a su física y no a la inversa. Esto responde a un enfoque que, como el modelo PENTA de XU, pretende reflejar la realidad, y asume la relatividad de la misma si no se parte de fenómenos empíricos. Muchos expertos en prosodia llevan algunos años reconociendo y desarrollando la importancia de establecer una buena base teórica del discurso como paso previo al estudio de la prosodia (Hirschberg, 1993), pero, quizás debido a la dificultad de encontrar discursos tan modelizables por homogéneos en sus objetivos como el informativo, el caso es que la perspectiva funcional no se ha consolidado.

Nosotros damos un paso adelante para buscar las teorías del discurso que se puedan a aplicar al Discurso Informativo de forma específica, y que guiarán la modelización de parámetros prosódicos.

A nivel teórico, partimos de la Teoría del Discurso en 3 niveles de Grosz y Sidner (1986), la de la estructura de la información (global y local) de Grosz y Hirschberg (1992)[3] y la Teoría de la información de Halliday (1973, 2004) y Prince (1992). La primera puede considerarse el punto de partida de una teoría general entonativa del discurso informativo, ya que dice que cualquier discurso está atravesado, además de por el nivel léxico, por la intención y la atención comunicativa y pragmática del emisor para con el receptor en determinadas circunstancias. Participa por tanto de una visión funcional de la comunicación, en la que prima la eficiencia comunicativa (eliminado). Así, debemos considerar discurso modelizable sólo aquella parte de la entonación a la que se le puedan atribuir efectos perceptivos emitidos voluntariamente por el emisor, de acuerdo a un discurso y sub-discurso particular, y que tengan su correlato acústico. Por tanto, la forma de la noticia, el tema, las palabras-clave, el tipo de locuciones, y tantos otros rasgos de expresión prosódica sólo podrán ser considerados si son controlados por el emisor y pueden ser objetivados acústicamente. Y, precisamente, ese control de la información por parte del emisor se basa en una categorización en niveles de la información: *given, mediated or new information* (información dada, mediada o nueva)[4]. Esta clasificación es tomada por Calhoun et al (2005), añadiendo nueve subentidades a la información mediada, que es aquella a la que se había hecho referencia anteriormente en el discurso. Además, en un segundo y tercer nivel, se distingue entre el tema y el rema del discurso, y las categorías *background* y *konstrast* de la información (Valldubí, Vilkuna, 1992; desarrollado por Steedman, 2000). La aplicabilidad de estas teorías en el discurso informativo está todavía por comprobar, pero, efectivamente, se intuye una suerte de estructura entonativa coherente de la unidad noticia si la tomamos como proceso de comunicación regido por la "noticialidad", esto es, la transmisión de hechos novedosos: tema homogéneo, hechos puntuales nuevos, agentes importantes implicados…, que junto a claros objetivos de atención e intención, puede derivar en una estructura entonativa determinada por: palabras-clave y falsas palabras-clave, palabras funcionales y palabras de contenido, empatía con el receptor, atracción del interés, límite de la información, etc. De hecho, recientemente se ha trabajado con muchos tipos de discurso (especialmente en conversación espontánea) para apuntalar la existencia de esa estructura intencional (Lochbaum, 1998)[5].

Como vemos, estas teorías nos llevan directamente a las técnicas de presentación de noticias, y nos permiten reconocer en ellas fenómenos articulatorios como los picos tonales periódicos, una entonación descendente en el avance temporal de la locución,

[3] También estudiada por Geluykens y Swerts (1993), el segundo de los cuales también ha trabajado con los rasgos prosódicos de la finalidad y el final del discurso (Swerts, 1993), estudio que ha servido para confirmar nuestras variables de trabajo.

[4] Ni que decir tiene que tomamos el concepto de "información nueva" desde el punto de vista del receptor (Prince, 1981, 1992).

[5] En realidad, en los últimos años se han multiplicado los estudios sobre prosodia que utilizan estas teorías del discurso, como algunos trabajos de Bates y Ostendorf (2001), lo que ocurre es que sus objetivos y su objeto de estudio son diferentes.

una caída tonal al final de noticia, una distribución estratégica de las pausas y los silencios, un alargamiento de vocales a final de noticia y en determinadas palabras, todo lo cual representa, en definitiva, un vislumbre de la estructura general entonativa para el Discurso Informativo. Habida cuenta que estos fenómenos provocan unos efectos deseados en el receptor, a nivel de comprensión y atención del mensaje, su correlato acústico, es decir, su modelización acústica, no sólo podría esconder las claves entonativas del final e inicio de la noticia, sino que podría configurar una teoría general de la entonación informativa que mejoraría la práctica de la profesión.

De ahí, la necesidad de modelizar acústica y fonéticamente las noticias de televisión. Y de ahí la necesidad de partir de datos de pitch, de distribuirlos temporalmente como una curva general, de tener en cuenta los picos tonales, y, por supuesto, la necesidad de obtener una correspondencia total entre la articulación consciente del emisor, el mensaje de su articulación en forma de curva entonativa, y la percepción del receptor en forma de comprensión y atención (niveles intencionales y atencionales del discurso).

Por lo tanto, en resumen, con las teorías del discurso como fuente de inspiración epistemológica y fenomenológica, a continuación echamos mano de procesos de diferentes modelos de la entonación existentes para esbozar un modelo de análisis y representación adaptado a los rasgos del tipo de discurso.

En primer lugar, definimos la entonación como la variación de los datos de tono en el tiempo. Es evidentemente una definición operativa que irá completándose a lo largo del artículo, porque la hacemos teniendo en cuenta las bondades de las herramientas de análisis acústico (Praat, por ejemplo), pero es tan bien una definición que cumple requisitos básicos:

- son datos numéricos: son por tanto objetivos y estables; es por tanto un criterio científico sobre el que apoyarnos.

- son datos que se corresponden con la articulación y la percepción: su esencia es el número de vibraciones de las cuerdas vocales por unidad de tiempo. Esta es una capacidad humana universal con significaciones y funcionalidades muy extendidas. Es decir, está en la elección consciente, voluntaria y consecuente del emisor el generar unos datos de pitch y no otros, y esa decisión se toma en función de unos efectos perceptivos.

- son datos integrados en la teoría del discurso informativo: su incidencia en múltiples niveles es indiscutible. Como ya se ha constatado, provoca efectos a nivel pragmático.

En segundo lugar, esos datos de pitch se deben transformar en una curva. Al necesitar un modelo fonético, nos olvidamos de momento de codificaciones simbólicas (TiLT; TOBI, INSTINT...) y nos centramos en la necesidad de crear una función numérica a partir de puntos de pitch de discursos largos. Al respecto, es sintomático que uno de los modelos de más éxito en la actualidad, llamado “de la percepción” es el Fujisaki, se basa en funciones de representación de pitch cuadráticas y no lineales. Esto se debe al hecho de que el habla, y aún más la locución de informativos, no se articula ni se

percibe por niveles, por picos o de forma invariable…, sino de forma suavizada, con unas caídas y subidas que responden a un margen de variabilidad acústica y perceptiva, y, en todo caso, al servicio de un contorno global. Es decir, aunque nosotros "entonemos" por sonorizaciones y las herramientas lo traduzcan en señales de vibración, nuestra voluntad entonativa está, como no puede ser de otra manera, al servicio del discurso, compuesto por unidades entonativas, o al menos esa es la parte de decisión que atañe a la entonación que nosotros tomamos. Para ese paso de datos de pitch a un contorno macro-melódico de la locución tenemos el algoritmo MOMEL[6]: "consiste en la sustitución de la curva de F0 por una función numérica sencilla que conserva la información macro-prosódica original y deja de lado la información no relevante" (Baqué y Estruch, 2003: 130). Concretamente, este algoritmo consiste en encontrar los puntos de inflexión de cada parábola, para lo cual se definen unos grupos de datos de tono en función de unos parámetros de duración y variación de tono, a partir de los cuales se define el vértice de la parábola, que será creada como regresión cuadrática; esto se repite por grupos de datos de tono cuyos respectivos vértices crean parábolas unidas entre sí (Hirst, Di Cristo, Espesser, 2003, editor Horne).

En tercer lugar, la estilización realizada ha resuelto sólo una parte del problema. Digamos que ha resuelto la transformación acústica y horizontal del mensaje. Pero no ha resuelto la cuestión vertical: el espectro donde se sitúan los niveles de pitch en Hz provoca una percepción logarítmica de la sensación de tono. Y por si esto fuera poco, el campo tonal varía de persona en persona, por lo que la emisión de tono en función del alcance de unos objetivos sobre los receptores no se basa sobre parámetros universales, sino en su variación relativa (del tipo: "aumentar el tono en determinada palabra-clave del tema Política"). Y la única forma de modelizar variaciones relativas es mediante porcentajes. El Análisis Melódico del Habla (Cantero, 2002) consiste en el cálculo del porcentaje de subida o bajada de cada valor de pitch respecto del anterior. Así, situando el primer valor de cada discurso en 100, las subidas o bajadas serán más pequeñas conforme aumente la tonalidad absoluta (valor de pitch del Praat), y por tanto todas las curvas partirán de 100 y trazarán un mismo campo "tonal" sin haber alterado la variación tonal macro y micro melódica.

Por lo que respecta al carácter temporal del modelo, más allá de entrar o no en la definición de entonación, lo cierto es que es fundamental para el discurso Informativo[7]. Tanto es así que en nuestro modelo distribuiremos los puntos de pitch (estilizados en una curva, y estandarizados en porcentajes a partir de 100) a lo largo del eje x en función de la duración de la sílaba a la que pertenecen. La decisión de tomar datos de pitch por sílabas tiene tres razones de ser:

1. Teórica: D'Alessandro y Mertens (1995) persiguen la estilización entonativa del discursivo desde un modelo perceptivo, y la cuestión temporal la resuelven mediante la división y medición del discurso en sílabas.

[6] Se trata de un algoritmo de modelización fonética de la curva entonativa creado en el Laboratoire Parole et Langage de la Université de Provence por Robert Espesser.

[7] Prueba de ello es una investigación que utiliza rasgos temporales para distinguir entre habla, habla con música o música únicamente en el discurso Informativo (Johnson et al., 2001).

2. Operativa: son los golpes de voz que damos al hablar, y son las unidades básicas de la vocalización y la sonorización, especialmente en lenguas como el español y el catalán. No en vano, el modelo autosegmental y el modelo Cantero dividen el discurso en sílabas. Además, se está avanzando en el campo de la segmentación automática de sílabas a partir de la curva de intensidad (Shastri et al, 1999, Nagarasan et al., 2003, Jittiwarangkul et al., 1998, Mermelstein, 1975).

3. Discursiva: todos los manuales de teorías y técnicas de locución relacionan directamente el ritmo con el número de sílabas por unidad de tiempo. E incluyen el concepto de pausa como la ausencia de locución durante 0.1 segundos

Por último, ya se dijo que el énfasis era una de las mayores caracterizaciones del discurso informativo, y previsiblemente coincidente con las palabras-clave (ya definidas) y la parte de la información nueva. Es más, se prevé una correlación directa con un aumento en la duración de las sílabas (Vaissiere, 1991). Por esta razón, vemos la necesidad de marcar simbólicamente los niveles de énfasis de las siguientes unidades: unidades entonativas, párrafos, locuciones y noticias[8]. Porque una de nuestra principales hipótesis gira alrededor de una tendencia al downtrend de los picos tonales, las medidas medias y los valleys, no sólo de las unidades entonativas (primer nivel) sino incluso del párrafo (Suijter & Terken, 1992, 1993, citado por Garrido, 1996) y, quien sabe, si de la noticia. Está por descubrir, por tanto, si los niveles de la curva resultante del algoritmo MOMEL (T, M, B; alto, medio y bajo) existen en la unidad párrafo y/o noticia, y, ocurra o no, si esconden alguna relación estructural de la misma. En todo caso, nuestra modelización acústica parece acotar todos estos fenómenos y condicionantes articulatorios, perceptivos, discursivos y acústicos de la entonación informativa.

En resumen, la prosodia en nuestro caso no es únicamente una sucesión de categorías descriptivas. Ni siquiera es una sucesión de **datos de pitch**, sino una curva continua de los **datos estandarizados** de pitch entre vértices definidos fonéticamente, **distribuidos temporalmente** en **unidades sílaba** y **pausas**, y cuya especificidad se sitúa en la **prominencia marcada por picos tonales** distribuidos de más a menos altura tonal y con cierta periodicidad.

Por lo tanto, decimos que nuestro algoritmo tendría tres fuentes de inspiración:

1. Método del Laboratorio de Análisis Melódico del Habla (Cantero): este modelo nos permite hacer una ponderación de los datos de pitch de cada sílaba previa a la estilización. Esta ponderación libera al pitch de determinismos articulatorios y contextuales, tales como el sexo del presentador.

2. Algoritmo MOMEL de Aix-en-Provence (Espesser): este modelo nos permite hacer una estilización fonética, continua y macro-melódica de los datos discretos y ponderados de pitch. Es un paso fundamental, el de pasar de datos de tono a

[8] Corremos el riesgo de confundir la prominencia léxica con la expresiva, y puesto que el modelado de la primera resulta muy complicado (tal y como se comprueba en un experimento de Dogil et al, 1997), será la significatividad de una muestra extensiva sometida a análisis la que vislumbre una coherencia estructural de la prominencia.

una curva entonativa.

3. Método Autosegmental (Pierrehumbert y Beckman): la filosofía de este modelo nos lleva a tratar nuevamente la curva continua resultante para su análisis como un conjunto de niveles tonales sucesivos y discretos por sílabas, aunque ponderados, estilizados y temporales.

Y su creación se rige por los resultados de un Análisis instrumental del Discurso Informativo en el Laboratorio LAICOM de la Facultad de Ciencias de la Comunicación de la Universitat Autònoma de Barcelona. Este análisis parte de una muestra representativa del universo que se pretende estudiar, sigue por un análisis cualitativo del objeto de estudio para definir las variables dependientes, crea unos instrumentos de medición objetiva de esas variables (el modelo que resultará de las bases teóricas expuestas), y acabará en un análisis estadístico de significación en el que se relaciona las variables independientes (el cambio de noticia) con las variables dependientes definidas.

Planteamiento del problema de selección de datos de pitch y experimento

Pero antes de poder implementar estos pasos en un modelo de análisis extensivo de una muestra, conviene resolver un problema previo: la obtención de datos de pitch. Es bien sabido que los programas ofrecen datos de pitch de toda señal de sonoridad. A partir de aquí, muchos modelos que hemos nombrado han tendido a marcar o tener en cuenta únicamente los puntos de pitch que suponían una inflexión en el contorno de la curva, ignorando aquellos datos sucesivos que no implicaban una subida o bajada (“plateau”).

Pero nuestro caso, en el que debemos conjugar diferentes modelos, perspectivas y enfoques, no se puede dejar margen a asunciones contextuales de ningún tipo. El proceso de ponderación resultaba en niveles numéricos de una periodicidad de 0.1 segundos, así que debemos plantearnos su forma de organización como datos que el MOMEL pueda tratar para modelar la curva resultante: convertir los puntos en una curva. Pero al mismo tiempo, no debemos dejar de pensar en las unidades sílaba debido a su importancia rítmica. Así, con el condicionante de tener datos de pitch presentados por sílabas, y éstas situadas en el eje x en función de su duración, ¿cómo decidiremos el nivel tonal en Hz para cada sílaba?, ¿cuántos datos por sílaba asignaremos?, y ¿cuál es el criterio para asignar datos de pitch a las sílabas?

En nuestra posición, la respuesta debe conjugar los requerimientos fonéticos de los datos que tratará el MOMEL, y la necesidad puramente comunicológica de tener en cuenta las sílabas de la locución. Pero si nos disponemos a relacionar la unidad fonológica “silaba” con los datos fonéticos de pitch que nada tienen que ver con las sílabas, vemos la necesidad de definir el tipo de conexión que se establecerá entre dos visiones y dos metodologías “opuestas”. La filosofía de trabajo que regirá la recuperación de los datos de pitch para cada sílaba tiene que ver con la gran influencia del Modelo ToBi o autosegmental.

Este modelo asume unos niveles de tono para cada sílaba en virtud de dos principios:

- todas las sílabas tienen al menos una vocal (excepto cuando se presente la "y", Oropeza Rodríguez, 2006).

- todas las vocales son sonoras y por tanto tienen altura tonal

No hace falta hacer referencia a todas las excepciones que en verdad se producen a estos dos principios, porque ya sabemos que no existe un método de recogida de datos de pitch no aproximativo[9], pero como preceptos teóricos, su validez es incuestionable. De hecho, sus problemas de exactitud práctica son intrínsecos al estudio de la entonación, ya que el estudio de la misma parte de la falacia de no existir un sistema exacto de obtención natural de la curva entonativa sino de unos datos individuales de periodicidad sinuidal, por lo que toda manipulación de esos datos está sesgada de principio. Ahora bien, si nuestras necesidades son una organización de los datos de pitch en sílabas, ¿de qué forma se atribuye una altura tonal a un ente que en sí misma no la tiene? La solución más fácil y que pudiera parecer más libre de manipulación, esto es, la elección del nivel tonal de uno de los datos de tono centrales y estables de la parte visual de la sílaba en el espectro (la utilizada por Cantero en su modelo), nos sugiere problemas de operatividad metodológica (en vistas a su automatización) y cuestiones de rigurosidad científica[10] para su aplicación a nuestro especial objeto de estudio:

- su exactitud relativa: los programas de análisis acústico ofrecen datos de pitch con una frecuencia temporal muy alta, de forma que una análisis con el Praat puede resultar en 4 o más datos de pitch por sílaba[11]. Teniendo en cuenta que el punto fuerte de nuestro modelo es su independencia respecto de las decisiones arbitrarias y su supeditación a la acústica del discurso, ya que la única dificultad a la que se exponía esta perspectiva era la estilización universal de la curva, nos vemos en la necesidad de evaluar el costo de esta intervención subjetiva sobre la curva respecto de una curva no intervenida, es decir, una curva fruto de todos los datos de pitch del Praat distribuidos por sílabas.

- su aplicabilidad automática posterior en caso de éxito: sea o no sea importante el costo de la intervención subjetiva sobre la curva, nuestro modelo debe poder ser implementado de forma rigurosa y exacta en una máquina, y creemos que el algoritmo de "coger el valor del punto medio visual sonoro de cada sílaba" es más difícil de formular (en caso de que sea posible) que el de procesar todos los datos de pitch extraídos por el analizador y obtener una curva global. Que nuestro modelo primario de análisis haya tenido en cuenta las sílabas y los tonos, y demás elementos discretos, responde a una estrategia metodológica que

[9] Recordemos todos los problemas de un modelo fonético; menores en número, en todo caso, a los de cualquier modelo fonológico.

[10] No decimos que no sea un procedimiento científico, sino que necesitamos saber hasta qué punto son despreciables o relativos todos los datos de pitch por sílabas (dados en el PRAAT) para poder evaluar el costo de cualquier propuesta aún menos "científica", pero más útil, para la aplicación de nuestros resultados de investigación y para contornear curvas de varios minutos.

[11] Es más, aunque se pudiera hacer una distribución temporal media de datos de pitch, cada sílaba tiene un duración particular, por lo que irremediablemente habrían sílabas sin datos, y otras con más de un dato.

toma la perspectiva del discurso informativo, pero también se ha tenido en cuenta su aplicación informática posterior, y de hecho nuestros resultados darán respuestas a nivel de contorno entonativo global que deberán ser implementados y procesados a partir de la generación automática de la curva entonativa de los informativos de televisión. Por eso, debemos asegurarnos que la generación de esa curva a partir de datos de pitch elegidos "a dedo" por sílabas u otros procedimientos no contaminan sustancialmente de subjetividad las curvas.

- su exactitud general: esta es una cuestión un tanto más general y esotérica, porque tiene que ver con el adulterio que nuestra mano hace con el cursor a la curva entonativa real de un discurso; pero si hemos visto que de momento no existen modelos mejores, sino aproximaciones desde diferentes perspectivas, sólo podemos poner el modelo al servicio de nuestros objetivos (bajo el paraguas de las bases teóricas) y en última instancia comprobar las curvas resultantes entre un modelo híbrido fonético-fonológico y un modelo fonético puramente como el MOMEL original.

Por tanto, el procedimiento Cantero, en su integración a nuestro modelo ad-hoc, ha planteado tres dudas que deben ser resueltas. Puesto que el contorno resultante sí se ha rebelado válido, nos encontramos ante la tesitura de encontrar un procedimiento que dé un resultado parecido o lo suficientemente parecido sin caer en los problemas de procedimiento identificados. Así, proponemos otros dos procedimientos, el primero de los cuales sería el más fiel a los preceptos teóricos-universales del modelo autosegmental, y el segundo el más operativo en términos de manejabilidad y automatización. Así, estas son las tres opciones de atribución de datos de pitch a las sílabas:

1. Procedimiento Cantero-perceptivo (caso 1 en anexos): se escoge el dato de pitch del punto medio visual de la vocal. (PRAAT: situar el cursor en medio del espectro de la vocal de la sílaba y anotar el índice).

2. Procedimiento Laicom-perceptivo-acústico (caso 2 en anexos): se escoge el segmento visual de la vocal y se obtiene la media aritmética de los datos de pitch del segmento. (PRAAT: selección con el cursor de la vocal-"Pitch"-"Get Pitch")

3. Procedimiento Praat-perceptivo-acústico (caso 3 en anexos): se hace uso de todos los datos de pitch distribuidos en las sílabas conforme suenan. Es un mecanismo que confía en la fonética como campo de estudio y en la herramienta informática. Se trata de seleccionar el espectro más amplio que se puede obtener con el Praat (6.2 segundos) y obtener todos los datos de pitch ("Pitch"-"Pitch listing"); la ventana obtenida relaciona los datos de Pitch con su cronología en unidades de 0.01, por lo que se copiarán ambas columnas de datos al Exel y se hará la media aritmética de todos los datos de pitch de cada silaba cuya relación cronológica coincida con la duración asignada en la segunda columna del Exel.

A continuación nos disponemos a contrastar los tres procedimientos, y así tomar una decisión consecuente.

Resultados

Tal y como hemos dejado claro, sólo se elegirá el procedimiento 2 o 3 si la curva resultante de los mismos es "equivalente" a la del primero. Es más, dado que existen programas capaces de segmentar sílabas automáticamente y por tanto sería idóneo utilizar el procedimiento 3, el criterio de elección del procedimiento 3 estará supeditado a una equivalencia gráfica y perceptiva de la curva resultante respecto de las de los datos del primer y segundo procedimiento. El criterio de evaluación de la conveniencia de los procedimientos es doble: representacional (simples datos de pitch situados en el tiempo y en las sílabas; eje x = tiempo, eje y = Hz) y sintética (síntesis de una locución de una noticia[12] mediante asignación de los diferentes valores de pitch de cada procedimiento a las sílabas en el PRAAT – "to manipulation")

El valor de este doble criterio es muy relativo, ya que por lógica, el rango de variación (valores absolutos) difiere mucho entre los procedimientos (estamos comparando la elección consciente y experta de un punto de pitch con todo el espectro de datos de pitch del Praat). Sin embargo, hay tres factores (de diseño metodológico en función del objeto de estudio) que abogan por la validez del criterio y dirigen su forma de interpretación:

- No hemos tenido en cuenta el campo tonal como variable de análisis.

- La percepción de la entonación es relativa, en el sentido de "aprehender" la melodía en la relación diferencial entre valores tonales, antes que en los valores absolutos.

- Los parámetros-variables locutivos que pretendemos representar son las formas macro y los fenómenos relativos concretos.

Por tanto, para la evaluación de la conveniencia de adoptar el procedimiento 3, debemos priorizar la forma a escala de dicha representación y la percepción de la audiencia hacia la síntesis:

1. Representación gráfica de los datos de pitch ponderados obtenidos por los tres procedimientos (ver Anexo 1). Las conclusiones que extraemos de las tres curvas es una indiferencia del procedimiento que se utilice ya que las variables de cuya variación hipotizamos los patrones entonativos del Discurso Informativo se presentan iguales en las tres representaciones:
    a. Existe una coincidencia del contorno general entre las tres representaciones.

    b. Los diferentes niveles de T, M y B coinciden en las tres representaciones.

    c. El timing de aparición de los niveles por sílabas en los tres

[12] Locución de 29 segundos, perteneciente al informativo del mediodía de TV3 del 3 de enero de 2007.

procedimientos es exacto.

d. Los niveles de variación entre niveles equivalentes de pitch son prácticamente idénticos.

e. Las keywords están situadas en el mismo lugar, y en los tres casos presentan el mismo nivel de énfasis macro-melódico. Es cierto que pueden haber variaciones micro-melódicas del énfasis de las silabas de la misma keyword entre representaciones, pero son variaciones léxicas muy marginales (de hecho, sólo se aprecian en la primera, tercera y cuarta keyword de la representación por el 3º procedimiento) y que en ningún caso nos interesan.

2. Síntesis de los datos de pitch ponderados obtenidos por los tres procedimientos (ver Anexo 2):
   a. Las tres síntesis se reconocen como manipuladas, aunque sin dificultar en ningún caso la comprensión del mensaje.

   b. La voz de la síntesis mediante el procedimiento 1 parece ligeramente más natural, aunque es una diferencia tan pequeña que sólo es perceptible por comparación sucesiva de las diferentes versiones.

   c. La dirección y la proporción de las variaciones entre cada nivel tonal se mantiene igual en los tres casos respecto de la curva original, y sólo en el caso de la curva del primer procedimiento se puede decir que el rango de variación es un poco mayor.

   d. No se aprecian diferencias en la coherencia entonativa de las voces sintetizadas respecto de la original.

   e. La calidad de la voz en las curvas sintéticas está afectada por igual en los tres casos: voz metálica y áspera, y eco.

   f. Los receptores son incapaces de distinguir cualquiera de las tres curvas sintetizadas por sí mismas.

En conclusión, aunque el procedimiento 1 es más exacto, el 3 no es tan diferente como para afectar negativamente a los resultados finales de la curva que obtendríamos. Y en cambio sí resuelve los tres problemas identificados más arriba.

Los resultados de la comparación de las curvas obtenidas en cada procedimiento nos dicen que en el análisis de una muestra extensiva se podrá utilizar el procedimiento *Praat-perceptivo-acústico* con la seguridad de que las constancias prosódicas encontradas en el inicio, transcurso y final de las noticias, también lo serían con los procedimientos 1 o 2 (*"Laicom-perceptivo-acústico"* y *"Cantero-perceptivo"*). Y de esta forma, aparte de relativizar al mínimo los problemas antes identificados, contamos con las ventajas operativas del procedimiento 3, el rigor teórico y científico del

procedimiento 2 y las pruebas experimentales exitosas del procedimiento 1.

Ahora que sabemos qué conjunto de datos de pitch vamos a contar (uno por sílaba) y de qué forma se obtienen (procedimiento 3), se aplicará el modelo Cantero y sucesivamente el algoritmo MOMEL conforme vienen enunciados. Por tanto, los puntos de enlace de la curva, puntos de inflexión o vértices, serán el resultado de aplicar el MOMEL al conjunto de datos de pitch presentados y ponderados por sílabas. A partir de aquí, los datos por sílabas desaparecen (serán parte de cada parábola, o cercanos a la misma), y el único rastro de las unidades sílabas (que como hemos visto son esenciales para el ritmo de locución) será su distribución sobre el eje x en función de su duración y en relación a la curva final.

Conclusiones

En el contexto de la construcción de un modelo entonativo con la aplicación específica de definir el principio y final de una noticia, hemos tomado una perspectiva de estudio de la entonación esencialmente diferente: experimental, funcional, discursiva. Hemos priorizado el tipo de discurso antes que su forma, bajo dos condicionantes:

- La estabilidad del discurso: aun no existe una taxonomía clara de discursos, pero está claro que el discurso de los informativos tiene una identidad propia.

- El conocimiento experto del mismo: este es un trabajo desde una facultad de comunicación, lo que nos permite conocer todas las variables que el emisor pone en liza, a partir de las cuales investigar la física del mensaje.

A partir de aquí, hemos hecho un recorrido intentando objetivar, paso a paso, todas esas variables en un modelo de medición, representación y análisis: empezando por cuestiones de macro-melodía, por tratarse de locuciones largas y de variables a nivel macro, terminando por los picos tonales a gran escala, y pasando por la adecuación psicofísica e integración del parámetro temporal, fundamental en el ritmo locutivo de los informativos. Y para ello, hemos echado mano de muy diferentes modelos y aproximaciones.

Admitimos en este momento errores de apreciación, consideración y precisión con respecto a la adopción de procedimientos de otros modelos, pero también se debe tener en cuenta que en todo momento hemos priorizado la perspectiva comunicativa, que es la del mensaje producido por personas subjetivas –de las que se busca lo intersubjetivo- e intencionales –que se representa a nivel macromelódico-. Y por eso hemos realizado una interpretación bastante libre de la entonación; de hecho, la selección del procedimiento 3 ignora principios de micromelodía, pero al mismo tiempo puede reportar ventajas a nivel macro prosódico, y por eso necesitaba ser testado. No descartamos por tanto la necesidad de una mayor rigurosidad en la aplicación de las medidas concretas de nuestro modelo, pero creemos estar en el buen camino. Confiamos pues en el inicio de un análisis extensivo de unidades noticia para perfeccionar de forma iterativa nuestro modelo. Hemos definido los objetivos y las variables de análisis, hemos hipotizado sobre su variación, y tenemos un instrumento de objetivación para su

medición, el cual cumple dos requisitos fundamentales para nuestro caso:

- discursivo del tipo de discurso: el modelo refuerza los principios de la práctica articulatoria en las locuciones informativas.

- cuantitativo de la macro-melodía: el modelo es congruente con propuestas generales de la acústica entonativa del habla (Campione et al., 1997)

Precisamente, en este artículo, en congruencia con las necesidades comunicológicas y los principios entonativos, hemos testeado el primer paso de nuestro instrumento de objetivación: selección de datos de pitch distribuidos en sílabas en el contexto de locuciones largas a partir de la señal sonora. Hemos respetado su rigurosidad científica en forma de constancia perceptiva y acústica, pero sin desvirtuar los principios básicos sobre los que se apoya el estudio de la prosodia. Por eso, estamos en condiciones de aplicar una metodología cuantitativista a una realidad tomada como subjetiva.

Referencias

BATES, R., OSTENDORF, R. (2001). "Modeling Pronunciation Variation in Conversational Speech using Syntax and Discourse". ISCA Archive. NJ. USA. Octubre 22-24.

CALHOUN, S. (2003). "The Nature of Rheme and Theme Accents". Centre for Speech Technology Research. Edimburg.

CALHOUN, S., NISSIM, M., STEEDMAN, M., BRENIER, J. (2005). "A Framework for Annotating Information Structure in Discourse". In Frontiers in Corpus Annotation II: Pie in the Sky, ACL2005 Conference Workshop .

CANTERO, F.J. (2002). "Teoría y análisis de la entonación". Edicions de la Universitat de Barcelona. Barcelona.

CAMPIONE, E., FLACHAIRE, E., HIRST, D., VÉRONIS, J. (1997). "Stylisation and symbolic coding of F0: A Quantitative Model". ISCA Archive. Athen (Greece). Septiembre 18-20.

D'ALESSANDRO, C., MERTENS, P. (1995). "Automatic pitch contour stylization using a model of tonal perception". LIMSI-CNRS, France. Department of Linguistics, K.U.Leuven, Belgium.

DOGIL, G., KUHN, J., MÖHLER, S., RAPP, S. (1997). "Prosody and discourse structure: issues and experiments". ISCA Archive. Athens (Greece). Septiembre 18-20.

GARRIDO, J.M. (1996). "Modeling Spanish Intonation for Text-to-Speech Applications". Tesis Doctoral, Universitat Autònoma de Barcelona.

GROSZ, B.J., SIDNER, C.L. (1986). "Attention, Intentions, and the Structure of Discourse". SRI Internation Menlo Park, California. BBN Laboratorios Inc, Cambridge.

HALLIDAY, M. A. K. (1973). "Explorations in the Functions of Language". Londres. Edward Arnold.

HALLIDAY, M. A. K., MATTHIESSEN, C. M. I. M (2004). "An Introduction to Functional Grammar". 3d ed. London. Arnold.

HIRSCHBERG, J. (1993). "Studies of Intonation and Discourse". ESCA Workshop on Prosody. Lund, Sweden. Septiembre 27-29, 1993.

HIRSCHBERG, J. (2000). "Intonational Variation in Spoken Dialogue Systems: Generation and Understanding". AT&T Labs-Research, USA.

HIRSCHBERG, J. (1999). "Communication and prosody: functional aspects of prosody". ETRW on Dialogue and Prosody. Veldhoven, The Netherlands. Septiembre 1-3, 1999.

JITTIWARANGKUL, N., JITAPUNKUL, S., LUKASANEEYANAVIN, N., AHKUPUTRA, V., WUTIWIWATCHAI, C. (1998). "Thai syllable segmentation for connected speech based on energy". IEEE APCCAS. The 1998 IEEE Asia-Pacific Conference.

JOHNSON, M. T., JAMIESON, L.H. (2001). "Temporal Features for Broadcast News Segmentation".

LOCHBAUM, K. E. (1998). "A Collaborative Planning Modelo f Intentional Structure". Association for Computational Linguistics.

MERMLESTEIN, P. (1975). "Automatic segmentation of speech into syllable units". Connecticut.

NAGARASAN, T., MURTHY, H. A., HEDGE, R.M. (2003). "Segmentation of speech into syllable-like units". Eurospeech, Geneve.

NAKATANI, C.H., HIRSCHBERG, J, GROSZ, B.J. (1995). "Discourse Structure in Spoken Language: Studies on Speech Corpora". Harvard University, CAmbridge. AT&T Bell Laboratorios, USA.

OROPEZA, J., SUÁREZ, S. (2006). "Algoritmos y Métodos para el reconocimiento de Voz en español mediante sílabas". Computación y sistemas Vol. 9 Núm. 3, pp. 270-286.

SHASTRI, L., CHANG S., GREENBERG, S. (1999). "Syllable detection adn segmentation using temporal flow neural Networks".

SHRIBERG, E., STOLCKE, A., TÜR, G., HAKKANI-TÜR, D. (1999). "Combining words and prosody for information extraction from speech". Speech Tecnology and Research Laboratory, Menlo Park, USA.

SHRIBERG, E., STOLCKE, A., TÜR, G., HAKKANI-TÜR, D. (2001). "Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation". Speech Tecnology and Research Laboratory, Menlo Park, USA.

SHRIBERG, E., STOLCKE, A., TÜR, G., HAKKANI-TÜR, D. (2000). "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics". Speech Communication 32 (1-2)

SHRIBERG, E., STOLCKE, A. (2001). "Prosody Modeling for Automatic Speech Understanding: An overview of Recent Research at SRI". Speech Technology and Research Laboratory, Menlo Park, California.

SÖNMEZ, K., SHRIBERG, E., HECK, L., WEINTRAUB, M. (1998). "Modeling dynamic prosodic variation for speaker verification". SRI Internacional, Nuance Communications, Menlo Park, California.

STEEDMAN, M. (2000). "Information structure and the syntax-phonology interface". Linguistic Inquiry, Vol. 31, Número 4 (649-689).

SWERTS, M. (1993). "On the Prosodic Prediction of Dicourse Finality". ESCA Workshop on Prosody. Lund, septiembre 27-29.

SWERTS, M., OSTENDORF, M. (1997). "Prosodic and lexical indications of discourse structure in human-machine interactions". Cnts, Belgium. IPO, Eindhoven, ECS Boston University, Boston.

SWERTS, M., GELUYKENS, R. (1993). "Local and global prosodic cues to discourse organization in dialogues". ESCA Workshop on Prosody. Lund, septiembre 27-29.

TAMBURINI, F. (2003). "Automatic Prosodic Prominence Detection in Speech using Acoustic Features: an Unsupervised System". University of Bologna, Italy.

TAYLOR, P., WRIGHT, H. (1997). "Modeling Intonational structure using Hidden Markov Models". ISCA Archive. Septiembre 18-20, 1997.

VAISSIERE, J. (1988). "The use of prosodic parameters in automatic speech recognition". Centre Nacional d'Etudes des Télécommunications, Lannion.

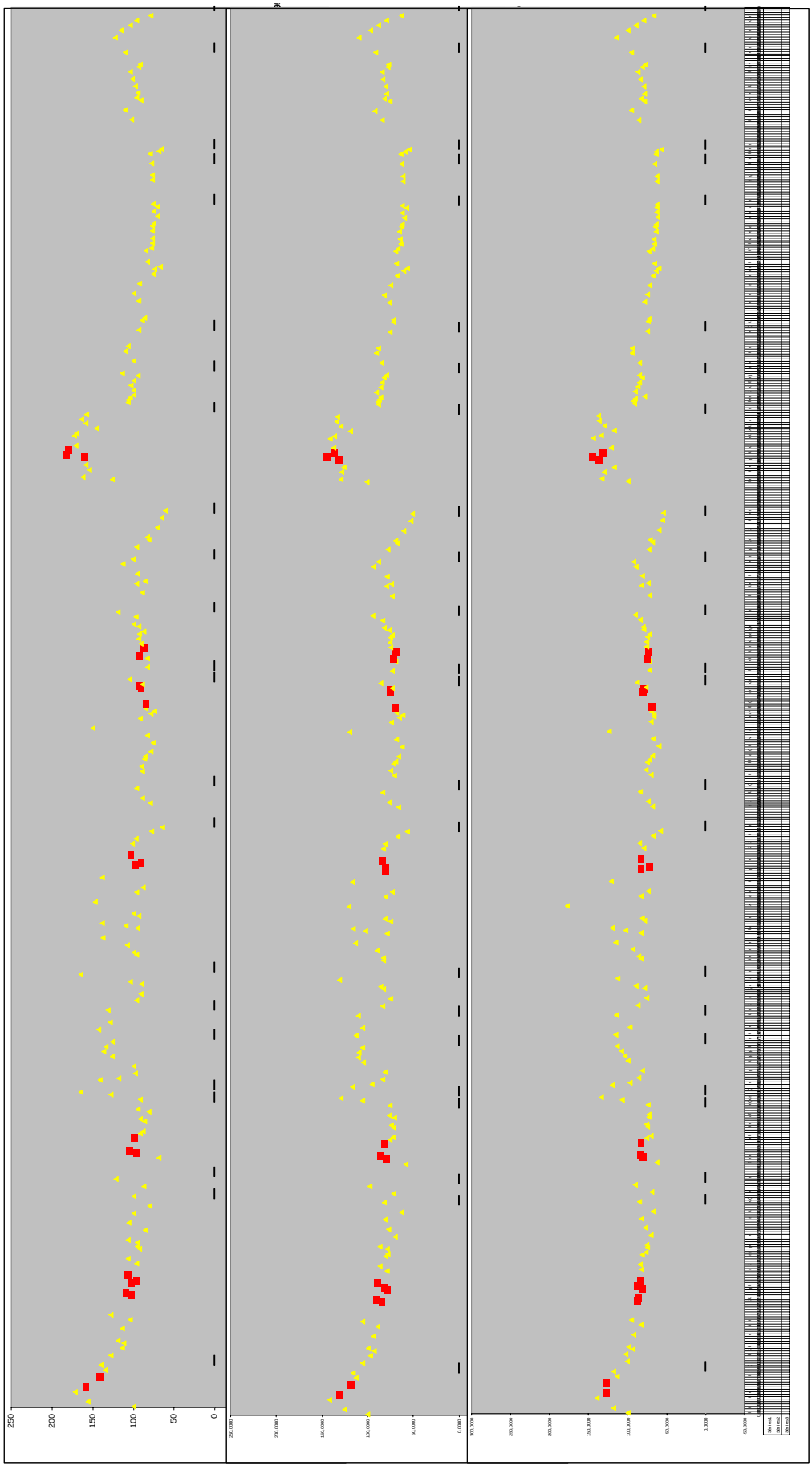
VAISSIERE, J. (1991). "Rhythm, accentuation and final lengthening in French".

VALLDUVÍ, E., VILKUNA, M. (1998). "On Rheme and Kontrast". Syntax and Semantics, 29:79-108.

XU, Y. (2004). "The Penta Model of Speech Melody: transmitting multiple communicative functions in parallel". Haskins Laboratorios, New Heaven, CT, USA.

ZIMBA, L.D., ROBIN, D.A. (1998). "The effects of varying signal intensity on the perceptual organization of rhythmic auditory patterns". The Nacional Center of Voice and Speech. University of Iowa.

ANEXO 1: COMPARATIVA GRÁFICA DE LA CURVA PONDERADA DE LAS 3 ALTERNATIVAS DE DISTRIBUCIÓN DE DATOS DE PITCH EN SILABAS



ANEXO 2: 30 PRIMEROS SEGUNDOS DE LA SÍNTESIS DE LA MELODÍA MEDIANTE LAS TRES FUENTES DE DATOS DE PITCH POR SÍLABAS