Ingeniare. Revista Chilena de Ingeniería
ISSN: 0718-3291
facing@uta.cl
Universidad de Tarapacá
Chile

Mas Manchón, Lluís
Segmentación automática de noticias mediante procesamiento de formas prosódicas
Ingeniare. Revista Chilena de Ingeniería, vol. 22, núm. 3, septiembre, 2014, pp. 374-383
Universidad de Tarapacá
Arica, Chile

Segmentación automática de noticias mediante procesamiento de formas prosódicas

Automatic segmentation of news through prosodic forms processing

Lluís Mas Manchón[1]

Recibido 7 de octubre de 2013, aceptado 22 de abril de 2014
Received: October 7, 2013 Accepted: April 22, 2014

RESUMEN

La segmentación automatizada de noticias en tiempo real es un problema que ha tenido un abordaje fundamentalmente lingüístico y de procesamiento de la señal en los últimos años. El trabajo que presentamos tiene un enfoque sustancialmente diferente: desde una perspectiva comunicológica, se toman las formas prosódicas típicas del discurso informativo y se intentan programar mediante el procesamiento *cepstrum* en el entorno Labview. Después de numerosas pruebas se fijan los diferentes parámetros de procesamiento y estilización de la curva macromelódica basada en los máximos de *pitch* (parábolas). Se formula un algoritmo de segmentación automática de noticias a partir de la detección de pausas y el análisis de formas prosódicas de inicio y final de noticia. Sobre una muestra de cinco informativos reales emitidos en el canal español Cuatro en julio de 2009, el entorno detecta 98 pausas y etiqueta correctamente (corte/no corte) el 76% de las mismas. La totalidad de los cortes estaban contenidos en las 98 pausas localizadas, de los cuales el algoritmo pasa por alto 9 y genera 15 errores de cortes mal emplazados. Tanto la lingüística computacional como el procesamiento del habla pueden tener un importante margen de mejora si también se asume la perspectiva comunicológica: agrupar datos acústicos como variables complejas asociadas a cada acto de comunicación. Para futuros trabajos, este algoritmo puede ser mejorado si se complementa con sistemas de identificación de locutores, discriminación de ruidos y *word spotting*.

Palabras clave: Locución informativa, segmentación automática, formas prosódicas, *cepstrum*, estilización macromelódica.

ABSTRACT

The automatic segmentation of real-time news has been mainly researched by linguistics and signal processing disciplines in the last years. The piece of work presented here has a substantially different approach: from a communication perspective, the specific prosodic forms of news discourse are taken into consideration and programmed in Labview program by using the cepstrum processing. After several tests, the processing parameters are set to generate a macromelodic curve fitting based on pitch maximums (parabola). An algorithm of automatic news segmentation is designed by means of spotting pauses and analyzing prosodic forms for the beginning and the end of the piece of news. With a sample of five news programs broadcasted in the Spanish channel Cuatro in July 2009, the programming environment spots 98 pauses and labels correctly (cut/not cut) 76% of them. The totality of real cuts is contained in those 98, of which the algorithm misses 9 and finds 15 non-cuts. Both the computational linguistics and the spoken language processing have room for improvement by assuming a communication perspective too: group acoustic data as complex variables associated to the act of communication. For future pieces of research, this algorithm may be improved by complementing it with speakers' recognition systems, noise detection and word spotting.

Keywords: News Speaking, Automatic segmentation, Prosodic Forms, Cepstrum, Macromelodic curve fitting.

[1] Universidad Pompeu Fabra Barcelona. Departamento de Comunicación. Roc Boronat, 138, 08018. Barcelona, España. E-mail: Lluis.Mas@upf.edu

INTRODUCCIÓN

El habla de los informativos de televisión, planificada con el objetivo de parecer espontánea, responde a un tipo discursivo propio de gran estabilidad y estandarización. De forma específica, la unidad discursiva noticia tiene una caracterización supralingüística o prosódica que la hace reconocible más allá de su contenido: la parte sintáctica y léxica (idiomática) y la semántica del evento y su temática. La melodía del habla de los informativos los hace reconocibles en francés, portugués, catalán o español, sin necesidad de entenderlos, independientemente del hecho noticioso y su temática, y sin la ayuda de las imágenes o de otros elementos sonoros no vocales (músicas, efectos o rasgos no verbales del habla). Sobre este supuesto teórico se postuló, en el marco de un proyecto de investigación financiado por el CAC (Consell de l'Audiovisual de Catalunya) y la tesis doctoral del autor [1], que mediante las variaciones prosódicas suprasegmentales era posible la segmentación automática de las unidades noticia en el continuum del discurso informativo en televisión.

Nuestro objeto de estudio eran pues las variaciones de tono, ritmo e intensidad en el habla del presentador de informativos propias de la enunciación de la noticia como unidad discursiva. Se precisaba pues una revisión micro y macroestructural del discurso noticia que diera sustento teórico a la siguiente hipótesis: existe un patrón superestructural del discurso informativo que nos permite segmentar las unidades noticia a partir de las variaciones prosódicas de nivel pragmático. Fruto de un estudio manual de 90 fragmentos de locución [2], se localizaron unas formas prosódicas típicas del inicio y final de noticia. Dichas formas son patrones de variación de los datos de tono, ritmo e intensidad a partir de un contorno entonativo enfático [3]. De forma sintética, los inicios de noticia se caracterizan por grandes picos tonales y de intensidad con alargamiento de sílabas (prominencias), que se organizan en la locución en orden descendente (*downtrend*), con un ritmo locutivo alto (por encima de las 130 palabras por minuto), una tesitura tonal alta (alrededor de 200 Hz para hombres y 300 Hz para mujeres) y pausas breves (*reset,* menores de 0,3 segundos) que inician una nueva secuencia de prominencias en *downtrend*. Y los finales se caracterizan por cierta monotonía tonal (*plateau*) interrumpida con alguna prominencia y grupos de prominencias en sentido ascendente (*uptrend*), ritmo locutivo y tesitura bajos, y una última prominencia muy enfática que da lugar a una caída tonal acentuada, ralentización del ritmo y alargamiento de las dos últimas sílabas (coda). Estas formas responden a un estudio estructural y conceptual del discurso informativo ya publicado [4] y apoyado teóricamente en [1] y [2], que encauza el principal problema de esta línea: separar los determinismos estructurales del léxico, la sintaxis y la semántica de los determinismos exclusivamente superestructurales o pragmáticos [5].

El problema técnico que presentamos en este artículo para el que proponemos y probamos un algoritmo es análogo a este problema conceptual: la discriminación exclusiva en el procesamiento del lenguaje natural de los rasgos prosódicos superestructurales de la noticia [6]. La mayoría de trabajos en este sentido son herederos de las tradiciones de la lingüística clásica y de la nueva lingüística computacional, que ha intentado modelizar los rasgos micro y macrolingüísticos del habla espontánea [13-15]. Xu [21] fue de los primeros en establecer que el análisis acústico debía supeditarse al tipo de acto de habla. En otras palabras, las diferentes funciones comunicativas debían comandar el análisis prosódico, y no a la inversa, siempre y cuando el objeto de estudio fuera comunicológico o semántico y no lingüístico propiamente. En este marco, presentamos a continuación un procesamiento de la señal al servicio de las funciones localizadas en el discurso informativo y llamadas formas prosódicas.

Para ello hemos utilizado el entorno de programación Labview [17], con el que hemos captado el *pitch* de la locución informativa mediante el *cepstrum* de la señal [16]. Mediante la búsqueda de máximos y mínimos en diferentes partes de las unidades entonativas de inicio y final de noticia, convertidos en variaciones porcentuales (para desactivar la naturaleza logarítmica de la escala tonal) [7], se han localizado las parábolas de la curva entonativa y la tesitura tonal [8]. Mediante la intensidad o energía, cuya principal función es lingüística: marcación de sílabas, se obtiene el ritmo de habla; y finalmente, se ha programado la obtención de datos de la bajada de *pitch* en las dos sílabas finales de la unidad textual noticia [9].

Las formas prosódicas han sido definidas numéricamente mediante variables de procesamiento

del *pitch*, y se han diseñado dos sistemas y un algoritmo de segmentación. A partir de la detección de toda pausa superior a 0,5 segundos [18], el entorno debía evaluar si el fragmento de antes de la pausa tenía formas de final de noticia (sistema 1) y si el fragmento de después de la pausa tenía formas de inicio de noticia (sistema 2). Sobre una muestra de 98 pausas, nuestro entorno acertó en señalar cada pausa como "corte" o "no corte" en 74 casos (76%). De los 24 errores, 9 fueron cortes ignorados y 15 cortes mal emplazados, por lo que el porcentaje de acierto en los cortes sube hasta un 91%, lo que indica un buen funcionamiento del algoritmo.

CEPSTRUM

El tono o F0 es la frecuencia más baja de un espectro de frecuencias. Aunque su procesamiento en algunos programas de análisis acústico (Praat) comienza a estar resuelta, a nosotros nos interesan únicamente ciertas variaciones intencionales y estructurales del tono durante la locución del presentador de informativos. Estas son las propiedades que deben tener las variaciones de F0 que buscamos:

1. Vocales y silábicas [10]: solo hay tono cuando hay vibración de las cuerdas vocales (sonoridad); no es solo que las cinco vocales del español sean las grandes generadoras de vibración, sino que son las únicas cuya naturaleza permite controlar el tono con función discursiva.
2. Periódicas y estables: a cada movimiento articulatorio, al vibrar las cuerdas vocales, los ciclos o formas que toman las ondas sinusoidales en sus fases se repiten y generan sonoridad, por lo que la generación de grandes variaciones precisa de un tiempo mínimo de ajuste y control.
3. Melódicos: la energía de las vibraciones, o periodicidad de las ondas (cantidad de vibraciones por unidad de tiempo) deben ser coherentes a lo largo de una señal, según los rangos definidos por las capacidades articulatorias humanas y los ritmos internos de la locución informativa.

Debemos condicionar el tipo de representación entonativa de nuestro analizador a la toma de un dato por vocal, el tiempo necesario para hacer un pico de tono enfático propio de la noticia, los ritmos constantes de locución con los que deberá guardar coherencia la generación discursiva de tono, y tener rangos mínimos de estabilidad espectral del habla. Se decide emplear el *cepstrum* de la señal [11].

Este procedimiento se define como "la transformada de la transformada". Se trata de aplicar la transformada de Fourier (FT) dos veces consecutivas, o mejor aún: convolucionar la señal ya convolucionada:

Señal $\rightarrow$ FT $\rightarrow$ log $\rightarrow$ FT $\rightarrow$ *cepstrum*

Cepstrum de una señal = FT [log (FT (la señal)]

Este tipo de procesamiento desvirtúa la equivalencia perceptiva de la señal global. Sin embargo, nosotros tratamos con el tono deliberado de una locución experta en una grabación de calidad, por lo que esta segunda convolución estará aislando los efectos de sonoridad del resto de efectos sonoros.

El *cepstrum* puede ser para valores reales o complejos. El primero solo contiene información de magnitud, mientras que el segundo tiene magnitud y fase, lo que permite "reconstruir" la señal. Nosotros adoptamos el *cepstrum* real porque solo nos interesa la magnitud de la sonoridad. Prueba del buen funcionamiento de este procedimiento de localización del tono dinámico es el gran uso que se hace de él en aplicaciones musicales. Así pues, haciendo uso del entorno Labview, optamos por la transformada de Fourier *Amplitude and Phase Spectrum* (VRMS en el Labview) y la aplicamos dos veces sobre la señal, según hemos descrito.

El *cepstrum* da como resultado unas bandas de frecuencia para cada momento muestreado de la señal. Esas frecuencias las llamaremos "quefrencys" ("quefrecuencias") en vez de "frequencies" ("frecuencias"). El quefrency es una medida de tiempo, pero no en el dominio del tiempo, sino en el dominio muestral del *cepstrum*. Es decir, para cada banda del espectro, calculamos la altura "temporal" (en muestras) del primer pico (F0), que es el que nos interesa (tono). Por ejemplo, en una señal de audio de 44.100 Hz, un primer pico de una serie de armónicos situado a 100 muestras de quefrency, daría como resultado un tono de 44.100/100 Hz = 441 Hz. De este modo se privilegia el procesamiento y representación del primer formante.

A partir de aquí, debemos hacer unos ajustes a la generación de tono adaptados a la curva macro

o supralingüística que pretendemos procesar. El proceso consiste en generar espectros sucesivos de las frecuencias y escoger aquellos valores de las frecuencias más bajas solo cuando haya gran sonoridad (armonía o periodicidad entre las frecuencias para cada momento espectral), por lo que debemos ajustar el procesamiento del *cepstrum* en función de:

- Privilegiar la representación de la energía en frecuencias bajas,
- Distinguir cuando esa representación es armónica si la energía articulatoria se concentra en 3, 4 o 5 picos y no se difumina en el espectro, y
- Elegir el primer pico (F0) de esa representación como valor definitivo de tono.

Los tres ajustes posibles que se manejan en el *cepstrum* son los siguientes:

1. El tramo: definimos los segmentos sobre los que se aplicará la transformada de Fourier y la transformada de Fourier del espectro resultante, es decir, el tamaño de la ventana de convolución. Para ello se deben considerar los tramos en que el tono de la voz es estacionario o los tramos en que nos conviene ignorar las variaciones dinámicas de tono. Por nuestra experiencia, el tracto vocal no es capaz de generar diferencias grandes en las modulaciones estables y voluntarias de tono en tiempos inferiores a 2 o 3 décimas de segundo. Por eso, para nuestro procesamiento, cogimos 512 puntos de la señal transformada.
2. El solapamiento y avance: los tramos que se elijan podrán ser sucesivos o empezar antes del fin del anterior, con lo que habría un solapamiento y un avance menor al de la longitud de 512 puntos. Esto se hace precisamente para tener una curva más suavizada. Al tratarse de una locución de gran calidad, diferentes pruebas nos indicaron que este solapamiento era innecesario para nuestro caso. Por lo que no se define solapamiento y el avance es de 512 puntos.
3. Filtraje de frecuencias: en el sonograma generado para cada "momento" espectral nos interesa encontrar la forma en que el pico F0 sea muy evidente y podamos medir su altura inequívocamente. Debido a que sabemos que la altura tonal de la voz tiene un rango que va de los 60 Hz a los 400 Hz, podemos eliminar o ponderar a la baja las quefrencys que se salgan de ese rango, de forma que el pico de F0 emerja claramente como el primero y más alto.

Con estos ajustes localizamos el pico máximo de quefrency que recae sobre el rango de 60-400 hz, y le sumamos 20 unidades para hacerlo aún más evidente.

ESTILIZACIÓN DE LA CURVA

Al tomar los máximos de esa banda espectral, y teniendo en cuenta que el *cepstrum* discrimina pero no ignora los fragmentos sin sonoridad, también estaremos tomando los máximos de las bandas sin sonoridad (como las consonantes fricativas o músicas y "ruidos" no vocales del habla). Esto es, de las series de datos de *pitch* que obtenemos, debemos priorizar las que se puedan constituir en las formas prosódicas macromelódicas presentadas en la Introducción. El criterio para distinguir estos datos de sonoridad suprasegmental de la noticia es estilizar la curva resultante en algunos casos paradigmáticos de los 90 estudiados e intentar correlacionar las curvas del sistema con nuestras curvas manuales. Este método ensayo-error nos lleva a detectar un doble problema.

En primer lugar, recordemos que la naturaleza del tono es logarítmica, y que los campos tonales son intersubjetivos al sexo, edad, estado emocional.

En el modelo propuesto [2] optamos por transformar cada dato de *pitch* como variación porcentual del anterior, comenzando la serie siempre por el valor neutro 100 [7]; esto será incorporado a la definición del algoritmo. En segundo lugar, nos interesan las parábolas macroestructurales [8] que definen los datos de *pitch* microestructurales generados, por lo que la curva melódica que buscamos se inscribe en los puntos medios de las variaciones micromelódicas.

Además, este segundo problema se solventa con una herramienta matemática capaz de aunar ambos

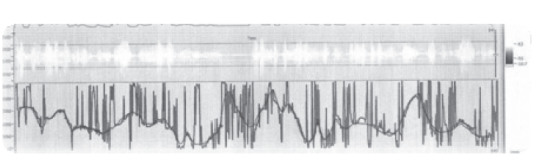

Figura 1. Media móvil.
Fuente propia.

criterios de interpolación y autoescala (Figura 1): la media móvil y la estilización mediante aproximación lineal (*curve fitting*):

Efectivamente, al aplicar una media móvil sobre la serie de datos, conseguimos:

1. Contrarrestar los valores extremos de las series: los valores de sonoridad muy bajos se anulan con los muy altos.
2. Maximizar las variaciones entre los datos relevantes *pitch*: se exagera la variación relativa entre los datos medios porque ya no hay un fondo escala de los valores altos y bajos.
3. Relativizar el sentido logarítmico de la curva (primer problema), pues al maximizar las variaciones de *pitch*, las pequeñas pero importantes variaciones intersubjetivas se diluyen.

Por otra parte, es muy común que los sistemas de procesamiento generen frecuencias-ruido, lo que puede llevar a equívoco en la coda del discurso y condicionar la media móvil. Sin embargo, esas frecuencias sonoras no vocales son tan exageradamente altas que se puede programar un criterio que limita las variaciones de un tono a otro a un rango máximo de 50% de variación.

Al mismo tiempo que se aplican estos dos criterios fundamentales de estilización –media móvil y variación interdato–, se hacen pruebas iterativas de las condiciones de convolución del *cepstrum* (del tramo y solapamiento) con el objetivo de encontrar el equilibrio entre la toma de datos de *pitch* más ponderados o más detallados. El equilibrio entre corte, solapamiento y avance en el análisis de la señal perseguía ajustar al máximo los valores de sonoridad vocal intercalados con los valores de sonoridad no vocal (Figura 2), obteniendo la curva entonativa y de intensidad final:

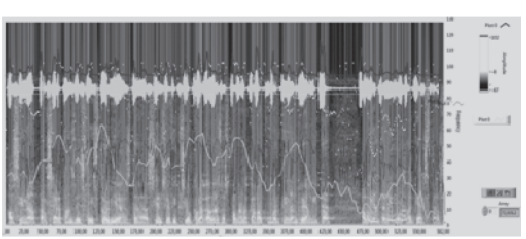

Figura 2. Labview.
Fuente Propia.

A partir de aquí, calculamos el tiempo como la inversa de la frecuencia y la intensidad como la integral de la energía de sonoridad al cuadrado (*Square Power Spectra*).

El resto de formas prosódicas son operaciones a partir de estos datos de *pitch*, tiempo y energía. El más complejo de los cuales fue el cálculo del ritmo de locución, entendido como la cantidad de sílabas por unidad de tiempo; en este caso, ya que la sílaba se define como un golpe de voz (energía), este ritmo será la intensidad microlingüística en los tramos temporales.

ALGORITMO

Nuestro entorno va a funcionar a partir de la localización de pausas candidatas a separar noticias. A partir de las mismas, se aplicará un algoritmo para la localización de formas prosódicas de inicio y final de noticia en fragmentos de 7 segundos antes y después de la pausa (unidades entonativas) y así decidir si efectivamente se trata de una pausa segmental.

Así, en primer lugar, definimos un protocolo para la detección de pausas. Definimos una pausa como una caída de la intensidad por un tiempo mínimo de 0,5 segundos. Porque nuestro sistema prioriza los datos producidos por el tracto vocal, resolvemos que la caída de intensidad es ausencia de voz.

En segundo lugar, nuestro algoritmo tendrá dos sistemas:

- Sistema 1: formas prosódicas que sirven para distinguir el final de noticia antes de una pausa segmental del fragmento de habla antes de una simple pausa no segmental.
- Sistema 2: formas prosódicas que sirven para distinguir el inicio de noticia después de una pausa segmental de la continuación del habla después de una simple pausa no segmental.

Estos sistemas tienen algunas formas prosódicas comunes y otras particulares. En general, ambos sistemas tendrán presencia de picos tonales, prominencias y ritmo de locución.

En primer lugar, los picos se toman como máximos absolutos en las 4 partes iguales de cada unidad

entonativa, y los mínimos como valores mínimos en esas mismas cuatro partes. Para fenómenos de *downtrend*, nos interesa una subdivisión mayor (1/8) de la unidad para localizar máximos de aparición frecuente y periódica:

► Forma Pico 1 = Variable Máx. 1 = $f_{Máx.\ 1}$ = Máx. $\{f_0, f_1, f_2,\ldots, f_{(n/4)-1}\}$
► Forma Pico 2 = Variable Máx. 2 = $f_{Máx.\ 2}$ = Máx. $\{f_{n/4}, f_{(n/4)+1},\ldots, f_{(2n/4)-1}\}$
► Forma Pico 3 = Variable Máx. 3 = $f_{Máx.\ 3}$ = Máx. $\{f_{2n/4}, f_{(n/3)+1},\ldots, f_{(3n/4)-1}\}$
► Forma Pico 4 = Variable Máx. 4 = $f_{Máx.\ 4}$ = Máx. $\{f_{3n/4}, f_{(3n/4)+1},\ldots, f_{(4n/4)-1}\}$
(Número mágico: 1/4)

► Variable Mín. 1 = $f_{Mín.\ 1}$ = Mín. $\{f_0, f_1, f_2,\ldots, f_{(n/4)-1}\}$
► Variable Mín. 2 = $f_{Mín.\ 2}$ = Mín. $\{f_{n/4}, f_{(n/4)+1},\ldots, f_{(2n/4)-1}\}$
► Variable Mín. 3 = $f_{Mín.\ 3}$ = Mín. $\{f_{2n/4}, f_{(n/3)+1},\ldots, f_{(3n/4)-1}\}$
► Variable Mín. 4 = $f_{Mín.\ 4}$ = Mín. $\{f_{3n/4}, f_{(3n/4)+1},\ldots, f_{(4n/4)-1}\}$

► Pico 1.1 = Máx. 1.1. = $f_{Máx.\ 1.1}$ = Máx. $\{f_0, f_1, f_2, f_3,\ldots, f_{(5n/8)-1}\}$
► Pico 1.2. = Máx. 1.2. = $f_{Máx.\ 1.2}$ = Máx. $\{f_{(5n/8)}, f_{(5n/8)+1},\ldots, f_{(6n/8)-1}\}$
► Pico 1.3. = Máx. 1.3. = $f_{Máx.\ 1.3}$ = Máx. $\{f_{(6n/8)+1}, f_{(6n/8)+1},\ldots, f_{(7n/8)-1}\}$
► Pico 1.4. = Máx. 1.4. = $f_{Máx.\ 1.4}$ = Máx. $\{f_{(7n/8)}, f_{(7n/8)+1},\ldots, f_{(8n/8)-1}\}$
(Número mágico: 1/8)

►Máx. 3.1 = $f_{Máx.\ 3.1}$ = Máx. $\{f_{n/5+1}, f_{n/5+2},\ldots, f_{n/4}\}$
► Máx. 3.2 = $f_{Máx.\ 3.2}$ = Máx. $\{f_{n/4+1}, f_{n/4+2},\ldots, f_{n/3}\}$
► Máx. 4.1 = $f_{Máx.\ 4.1}$ = Máx. $\{f_{n/3+1}, f_{n/3+2},\ldots, f_{n/2}\}$
► Máx. 4.2 = $f_{Máx.\ 4.2}$ = Máx. $\{f_{n/2+1}, f_{n/2+2},\ldots, f_n\}$
► Mín. 4.2 = $f_{Mín.\ 4.2}$ = Mín. $\{f_{n/2+1}, f_{n/2+2},\ldots, f_n\}$

En segundo lugar, la prominencia es un gran énfasis tonal: un 24% de diferencia entre máximo y mínimo:

► Prominencia 1= / (($f_{Máx.\ 1}$ / $f_{Mín.\ 1}$) x 100)) – 100/ > 24
► Prominencia 2 = / (($f_{Máx.\ 2}$ / $f_{Mín.\ 2}$) x 100)) – 100/ > 24
► Prominencia 3 = / (($f_{Máx.\ 3}$ / $f_{Mín.\ 3}$) x 100)) – 100/ > 24
► Prominencia 4 = / (($f_{Máx.\ 4}$ / $f_{Mín.\ 4}$) x 100)) – 100/ > 24
(Número mágico: 24)

► Prominencias = Prominencia 1 + Prominencia 2
► Énfasis = / (($f_{Máx.\ 1}$ / $f_{Mín.\ 1}$) x 100)) – 100/ + / (($f_{Máx.\ 2}$ / $f_{min\ 2}$) x 100)) – 100/ > 25

Y en tercer lugar, el parámetro ritmo queda definido en función de la energía:

► Variable E- alargamiento: $\int y_{6000} - y_{1000}$
► Variable E- durante: $\int y$
► Variable E- final: $\int y_{6000} - y_{1000}$

► Variable Energía 1 = $\int \{f_0, f_1,\ldots, f_{n/3}\}^2$ / (n/3) x 100000
► Variable Energía 2 = $\int \{f_{n/3+1}, f_{n/3+2},\ldots, f_{n/2}\}^2$ / (n/3)) x 100000
► Variable Energía 3 = $\int y^2 \{f_{n/2+1}, f_{n/2+2},\ldots, f_n\}^2$ / (n/3)) x 100000

VARIABLES SISTEMA 2

El *downtrend* es una variable típica de principio de noticia, así como el *reset* al que se le asocia:

►Downtrend = (($f_{Máx.\ 1.3}$ / $f_{Máx.\ 1.1}$) x 100)) – 100; si $f_{Máx.\ 1.1} > f_{Máx.\ 1.2} > f_{Máx.\ 1.3}$

► Reset 1 = (($f_{Máx.\ 1.4}$ / $f_{Máx.\ 1.3}$) x 100)) – 100; si downtrend.
► Reset 2 = (($f_{Máx.\ 1.4}$ / $f_{Máx.\ 1.1}$) x 100)) – 100; si downtrend.

► Valle = $f_{Mín.\ 3} < f_{Mín.\ 2}$; $(f_{Mín.\ 3}+3) < f_{Mín.\ 1}$; $(f_{Mín.\ 3}+3) < f_{Mín.\ 4}$
E
$f_{Mín.\ 2} < f_{Mín.\ 3}$; $(f_{Mín.\ 2}+3) < f_{Mín.\ 1}$; $(f_{Mín.\ 2}+3) < f_{Mín.\ 4}$

► Ritmo inicio = +/- 2 (Energía 3 x 100) ≥ Energía 2 x 100 o Energía 2 x 100 ≤ +/- 2 (Energía 3 x 100) o
(Energía 1 x 100) = +/- 1 (Energía 2 x 100)

VARIABLES SISTEMA 1

El *uptrend* es una variable específica del final de noticia, consistente en generar algunos picos en aumento hasta una gran prominencia final:

► Uptrend =
Si
$f_{Máx.\ 1.1} < f_{Máx.\ 1.2}$
y
$f_{Máx.\ 1.2} < f_{Máx.\ 1.} = + 0{,}5$
Si
$f_{Máx.\ 1.1} < f_{Máx.\ 1.2}$
ó
$f_{Máx.\ 1.2} < f_{Máx.\ 1.3} = + 1$

► Prominencia retrasada =
$/ ((f_{max\ 3} / f_{min\ 3}) \times 100)) - 100/ > 20$; e $/ ((f_{max\ 4} / f_{min\ 4}) \times 100)) - 100/ > 20$

► Prominencia retrasada =
Si
$/ ((f_{max\ 3} / f_{min\ 3}) \times 100)) - 100/ > 24$; e $/ ((f_{max\ 4} / f_{min\ 4}) \times 100)) - 100/ > 24$
= 1
Si Ɇ
= 2

► Ritmo final = Energía 3 < energía 2; e Energía 2 ≤ Energía 1

► Ritmo final 2 = Energía 3 > Energía 2; e Energía 2 < Energía 1

ALGORITMO FINAL

Como hemos visto, cada forma prosódica se ha convertido en una o múltiples variables. A partir de aquí, solo nos queda hacer una ponderación de cada una de estas variables en función de su contribución para definir el "principio de noticia" en el Sistema 2 y el "final de noticia" en el Sistema 1. Para ello asignamos unos índices a cada variable. Por tanto, vemos en primer lugar los algoritmos de las variables que se derivan de las formas prosódicas para el Sistema 2 y 1, y después el algoritmo final como sumatorios de los índices de los algoritmos de cada sistema. A continuación listamos las formas y las correspondientes variables que las definen. Para el Sistema 2 contamos con 18 variables y para el Sistema 1 con 12.

En primer lugar, el sistema 2 distingue los "principios de noticia" de los "después de pausa" mediante la suma de los siguientes índices de las variables:

A. Forma prosódica "Ritmo":
   1. Si: E-Inicio[2] < 650 → índice +2
   2. Si: 3500 < E-Durante → índice +0.
   3. Si: E-Alargamiento < 220 → índice +3

B. Forma prosódica "tesitura":
   4. Si: Mean 4 ≥ (Mean 3 – 2) → índice +0,5
   5. Si: Mean 2 ≥ (Mean 3 – 2) → índice +0,5

C. Forma prosódica "Inicio":
   6. Si: Inicio > 0 → índice +0,5
   7. Si: [E-Máx 4.2. < E-Máx 4.1. y E-Máx 4.1. < E-Máx 3.2.] o E-Máx 4.2. < E-Máx 4.1. → índice +2
   8. Si: Inicio 2 > 0 → índice +0,5

D. Forma prosódica "Prominencia retrasada":
   9. Si: E-Máx 4.1. < E-Máx 3.2. y E-Máx 4.2. < E-Máx 4.1. → índice +0,5

E. Forma prosódica "Tesitura global":
   10. Si: [(Mean 1 + Mean 2) / 2 < (Mean 3 + Mean 4) /2] → índice +2
   11. Si: [(Mean 1 + Mean 2) / 2 > (Mean 3 + Mean 4) /2] → índice +3

F. Forma prosódica "Uprend" y "Downtrend":
   12. Si NO: Máx 1.1. > Máx 1.2. o Máx 1.2. > Máx 1.3. → índice +1

i. Si: Máx 1.1. > Máx 1.2. y Máx 1.2. > Máx 1.3. → índice +0,5

En segundo lugar, el sistema 1 distingue los "finales de noticia" de los "antes de pausa" mediante la suma de los siguientes 18 índices:

G. Forma prosódica "Inicio":
   1. Si: Dato 0 > 45 (número mágico) → índice +1
   2. Si: (Máx 4.1. > Máx 3.2. y Máx 4.2. > Máx 3.2.) o Máx 4.2. < Máx 4.1. → índice +2
   3. Si: Inicio > 7 (número mágico) → índice +1
   4. Si: Inicio 2 ≥ 10 (número mágico) → índice +1

H. Forma prosódica *Downtrend*:
   5. Si: Máx 1.1. > Máx 1.2. y Máx 1.2. > Máx 1.3. → índice +1
   6. Si: [(Máx 1 + 1 o Máx 1 – 1 o Máx 1) ≥ (Máx 2 +1 o Máx 2 –1 o Máx 2)] y [(Máx 2 + 1 o Máx 2 – 1 o Máx 2) ≥ (Máx 3 +1 o Máx 3 –1 o Máx 3)] y [(Máx 3

[2] "Energía de Inicio" = ritmo de inicio.

+ 1 o Máx 3 – 1 o Máx 3) ≥ (Máx 4 +1 o Máx 4 –1 o Máx 4)] → índice +1

I. Forma prosódica "Valle":
   7. Si: Mín 2 < Mín 3 y [(Mín 2 + 3) < Mín 4 y (Mín 2 + 3) < Mín 4] → índice +1
J. Forma prosódica "Prominencia retrasada":
   8. Si [[(Máx 3 / Mín 3) x 100] – 100 > 20] y [[(Máx 4 / Mín. 4) x 100] – 100 > 20 (número mágico)] → índice +1
K. Forma prosódica "Ritmo":
   9. Si: E-inicio 2 > 900 y E-inicio 2 > E-durante 2 → índice +3
   10. Si: Energía 2 ≤ Energía 1 → índice +2
   11. Si: Energía 3 > Energía 2 y Energía 2 < Energía 1 → Ritmo final 2 = índice +1
   12. Si: [(Energía 3 x 100 + 1) o (Energía 3 x 100 +1+1) o (Energía 3 x 100 - 1) o (Energía 3 x 100 -1-1) ≥ Energía 2 x 100] o [(Energía 2 x 100 ≤ (Energía 3 x 100 + 1) o (Energía 3 x 100 +1+1) o (Energía 3 x 100 - 1) o (Energía 3 x 100 -1-1)] o (Energía 2 x 100 + 1 o Energía 2 x 100 - 1 = Energía 1 x 100) → índice +1
L. Forma prosódica "Coda":
   13. Si: E-alargamiento > 7 → índice +1
   14. Si: Coda 3 < 10 → índice +1
M. Forma prosódica "Tesitura":
   15. Si: [(Mean 1 + Mean 2) / 2 ≤ (Mean 3 + Mean 4) /2] → índice +1
   16. Si: Mean 1 > Mean 2 → índice +1
   17. Si: [(Mean 1 + Mean 2) / 2 ≤ (Mean 3 + Mean 4) /2] → Tesitura descendente = índice +1
   18. Si NO: [(Mean 1 + Mean 2) / 2 > (Mean 3 + Mean 4) /2] → Tesitura ascendente = índice +1

Por medio de sucesivas pruebas en muestras de 40 noticias, concluimos que el Sistema Integrado identificará un corte de noticia sumando los índices de cada sistema, de forma que definimos el siguiente algoritmo de segmentación:

Si:
(Índice Sistema 2 ≥ 11 y Índice Sistema 1 ≥ 5)
o
[Índice Sistema 1 ≥ 6 y (Índice Sistema 1 + Índice Sistema 2 ≥ 13,5)]
→ CORTE

RESULTADOS Y DISCUSIÓN

Se tomó una muestra de seis Informativos noche de la cadena Cuatro en los días 15, 16, 17, 20, 21 y 22 de julio de 2009. La elección fue totalmente casual. De todas las pausas mayores de 0,5 segundos localizadas automáticamente por nuestra plataforma, se filtraron 116 por presentar evidentes problemas extralingüísticos (carraspeos del locutor, solapamientos con gritos en reportajes o sonidos técnicos varios), así como aquellas pausas delimitadas por testimonios. De las 98 pausas restantes, candidatas a ser segmentales (separar noticias), el entorno creado acertó en la evaluación de "corte-no corte" en 74 casos, lo que supone 76% de acierto. De esos 24 errores, 9 fueron cortes ignorados y 15 cortes mal emplazados. En suma, de las 98 pausas superiores a 0,5 segundos, 50 eran cortes de noticia, por lo que se acertó en 72% de casos teniendo en cuenta esos 9 errores, si bien se generaron 15 cortes donde no los había. No hubo ningún corte de noticia que no estuviera en esas 98 pausas.

Téngase en cuenta que hemos trabajado con una muestra real no manipulada, esto es, con locuciones naturales y en directo y con la influencia del cambio de locutor. Bien es cierto que se han eliminado de la muestra aquellos elementos extralingüísticos claramente distorsionadores y todas las intervenciones de testimonios, pues escapan a nuestro modelo superestructural de la locución informativa. Por lo tanto, nuestra propuesta debe ser complementada con algoritmos de discriminación de las condiciones acústicas generales (músicas, ruidos, elementos extralingüísticos…), y con un algoritmo de procesamiento del lenguaje natural para los testimonios. Otra línea fértil podría ser trabajar en el *wordspotting* [12] de ciertas palabras-clave del discurso informativo, cuya presencia es claramente mayor a principio y final de noticia. Asimismo, se deberían hacer pruebas con un mínimo de solapamiento del enventanado del algoritmo, pues podría perderse información en los límites del procesamiento de la señal. En todo caso, un entorno con todos estos algoritmos complementarios tendría posibilidades de tener un funcionamiento robusto. Por otra parte, reconocemos cierta simplificación técnica como limitación de nuestra investigación.

CONCLUSIONES

Nuestro entorno tiene el siguiente funcionamiento general y secuencial:

1. Procesamiento de la señal sonora de un informativo,
2. Detección de pausas por bajadas de intensidad,
3. Análisis de la prosodia de la unidad anterior (sistema 2) y de la unidad posterior (sistema 1) a esa pausa,
4. Asignación de índices al análisis de las 12 variables prosódicas del sistema 1 y las 18 variables prosódicas del sistema 2,
5. Sumatorio de los Sistemas 1 y 2 y asignación de "corte" o no,
6. Generación final de una hoja Excel donde figuran todos los cortes y una etiqueta asociada a cada uno que diga si se trata de "corte" o "no corte".

El sistema funciona de forma razonable. En primer lugar, la detección de pausas es perfecta, pues no hubo ningún caso de cambio de noticia que no entrara como pausa candidata a ser segmental (corte). En segundo lugar, el tipo de procesamiento y algoritmo definidos para la localización de formas prosódicas ha permitido distinguir el cambio o no de noticia en el 76% de los casos. Estos resultados se asemejan a los obtenidos por diferentes autores en los últimos años [19], normalmente enfocados a la interacción hombre-máquina [20], si bien se trata de un enfoque semántico innovador que puede complementar los trabajos más técnicos y lingüísticos en este sentido. De hecho, mientras la ingeniería se ha alimentado en gran medida de los avances que desde la lingüística se hacían en muy diferentes lenguas y tipologías de lenguas en el mundo (según fueran de acento silábico o de frase), este enfoque puede servir de complemento a todas ellas, pues asume una prosodia discursiva hasta cierto punto independiente de la lengua y el contenido del tipo discursivo.

Por lo tanto, esta línea de trabajo desde una perspectiva comunicológica definiendo variables complejas con significación (formas prosódicas) vinculadas al tipo de acto de comunicación, como paso previo al procesamiento de datos, se debe complementar con un trabajo desde los niveles lingüísticos y desde el correspondiente procesamiento de la señal. De esta forma, podremos tener un abordaje menos intuitivo de la influencia de esas otras variaciones micro sobre las variables macro. Es más, este enfoque podría contribuir a solventar el mayor problema con el que se encuentran los estudios de lingüística: el análisis prosódico-acústico de los niveles semántico y pragmático de la prosodia (autores como [13-15], entre otros, han trabajado en este sentido), teniendo en cuenta los grandes avances en lingüística computacional en los niveles sintáctico, léxico y microlingüístico del mensaje.

AGRADECIMIENTOS

El autor agradece al profesor Dr. Antonio García Sánchez y a la Universidad Politécnica de Cartagena por su excelente acogida en la realización de este trabajo.

REFERENCIAS

[1] Ll.M. Manchón. "Modelo superestructural de la noticia en Televisión". Estudios sobre el Mensaje Periodístico. Vol. 17 N° 1, pp. 95-116. 2011. ISSN: 1134-1629

[2] Ll.M. Manchón. "Análisis prosódico discursivo de las fases de la noticia en televisión". Oralia. Vol. 15, pp. 205-239. 2012. ISSN: 1575-1430.

[3] A.W. Black and A. Raux. "Unit Selection Approach to F0 Modeling and Its Application to Emphasis". ASRU 2003. St Thomas. US Virgin Is. 2003.

[4] Ll.M. Manchón. "Formas Entonativas en las Fases del Discurso". En actas del Congreso XXXII Congresso Brasileiro de Ciências da Comunicação. Intercom, Curitiba, PR, Brasil. 4 al 7 de septiembre de 2009.

[5] K. Dusterhoff and A.W. Black. "Generating F_0 contours for speech synthesis using the TILT intonation theory". En Conference Intonation: Theory, Models, and Applications. Athens. Greece. September, 1997. ISCA Archive. URL: http://www.isca-speech.org/archive. Date of visit: March 12, 2008.

[6] H. Fujisaki. "Information, Prosody, and Modelling". In Conference Speech Prosody. Nara, Japan. March 23-26, 2004. ISCA ARCHIVE. Date of visit: June 16, 2010. URL: http://www.iscaspeech.org/archive/sp2004/sp04_001.pdf

[7] F.J. Cantero Serena y D. Font Rotchés. "Protocolo para el análisis melódico del habla". Estudios de fonética experimental. Vol. 18, pp. 18-32. 2009. ISSN: 1575-5533.

[8] R. Espesser and D. Hirst. "Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function". IPA, Travaux de l'Institut de Phonétique d'Aix. Vol. 15, pp. 75-85. 1993. ISSN: 0396-0978.

[9] I. Fonagy. "Prosody and syntax: cross-linguistic perspectives". John Benjamins. Amsterdam. 2006. ISBN: 90 272 3315 2.

[10] J.I. Hualde. "El modelo métrico y auto-segmental". En P. Prieto (coord.). Teorías de la entonación. Editorial Ariel. Barcelona, España. 2003. ISBN: 843448255X. DOI: 10.1017/S0305000912000359z.

[11] R.B. Randall. "Frequency Analysis". Bruël and Kjaer, Naerum, Denmark. 1987. ISBN: 88-7021-257-2.

[12] S. Renals, D. Abberley, D. Kirby and T. Robinson. "Indexing and Retrieval of Broadcast News". IEEE Signal Processing Society 1999 Workshop. September 13-15, 1999.

[13] M. Ostendorf and K. Ross. "A Multi-level model for recognition of intonation labels". Computing Prosody. Springer. Berlin, Germany. 1997.

[14] M. Swerts. "On the prosodic prediction of Discourse finality". En ESCA Workshop on Prosody. Lund. 1993. ISCA archive. URL: http://www.isca-speech.org/ archive_open/ prosody_93/pro3_096.html. Date of visit: April 21, 2014.

[15] J. Terken and D. Hermes. "The perception of prosodic prominence". En M. Horne, (ed.). Prosody: Theory and experiment. Studies presented to Gösta Bruce, pp. 89-127. Kluwer. Dordrecht, Holland. 2000.

[16] D. Jurafsky. "Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition". Pearson Prentice Hall. Upper Saddle River, N.J, London. 2009.

[17] M.L Chugani A.R. Samant and M. Cerna. "Labview Digital Signal Processing". Prentice Hall PTR. Upper Saddle River, NJ, London. 1998.

[18] J.J. Perona y A. Huertas. "Redacción y locución en medios audiovisuales: la radio". Bosch. Barcelona, España. 2008.

[19] E. Shriberg, E. Stolcke, A. Hakkani-tür and D. Tür. "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics". Speech Communication Vol. 32, Issue 1-2. Special Issue on Accessing Information in Spoken Audio. September, 2000.

[20] M. Swerts and M. Ostendorf. "Discourse prosody in human machine interactions". In ESCA Workshop. Vigso, Denmark. 1995. ISCA Archive. Date of visit: March, 2008. URL: http://www.isca-speech.org/archive

[21] Y. Xu. "The Penta Model of speech Melody: transmitting multiple communicative functions in parallel". Sound to Sence. MIT. June 11-13, 2004.