Káñina, Rev. Artes y Letras, Univ. Costa Rica XL (Extraordinario): 13-19, 2016 / ISSN:2215-2636

USO DE LA ENTONACIÓN PARA IDENTIFICAR CUÁNDO USAR LA TILDE DIACRÍTICA EN EL RECONOCIMIENTO AUTOMÁTICO DEL HABLA

USE OF INTONATION TO IDENTIFY WHEN TO USE THE DIACRITICAL ACCENT MARK IN AUTOMATIC SPEECH RECOGNITION

*Constantino Bolaños Araya**

*Arturo Camacho Lozano***

*Ximena del Río Urrutia****

RESUMEN

Los métodos y técnicas computacionales de reconocimiento automático del habla son herramientas que poco a poco se incorporan en la vida cotidiana. Una de sus principales ventajas es que permiten a las personas registrar texto rápidamente usando uno de los medios que mejor saben usar: su voz. Desafortunadamente, esta tecnología aún no es perfecta y los errores de transcripción son comunes. En el idioma español, uno de los errores más comunes de esta tecnología, es la omisión de la tilde diacrítica. Esto se debe en gran medida a que las técnicas utilizadas en el reconocimiento automático del habla ignoran el acento, es decir, la sílaba acentuada de una palabra, que en el idioma español no tiene un patrón fijo, como en otras lenguas. Esto se debe a que estas técnicas fueron desarrolladas inicialmente para el idioma inglés, en el cual no hay tildes y el acento juega un papel menor en la diferenciación de las palabras. Nuestra propuesta es incorporar el análisis del tono en el reconocimiento automático del habla para mejorar la marcación de tildes diacríticas en un texto. Ensayos previos han mostrado que la creencia extendida de que la sílaba acentuada es siempre la más fuerte (la más intensa) es falsa. Por tanto, la intensidad, por sí sola, no es un buen indicador de la ubicación o presencia de la sílaba acentuada que debe ser tildada, sino que el tono también ha de ser considerado. En esta investigación se muestra que el tono puede ayudar a determinar la sílaba que debe ser tildada según nuestra convención gráfica.

Palabras clave: acento diacrítico, reconocimiento automático del habla, procesamiento del lenguaje natural.

ABSTRACT

Tools based on computational methods and techniques for automatic speech recognition are slowly being integrated in everyday life. One of their main advantages is allowing people to easily record text, using their own voice. Unfortunately, this technology is not perfect, and transcription errors are common. In Spanish,

* Universidad de Costa Rica. Centro de Investigaciones en Tecnologías de la Información y Comunicación. Costa Rica. Correo electrónico: tinoba92@gmail.com

** Universidad de Costa Rica. Profesor Asociado de la Escuela de Ciencias de la Computación e Informática. Costa Rica. Correo electrónico: arturo.camacho@ecci.ucr.ac.cr

*** Universidad de Costa Rica. Profesora, Escuela de Filología, Lingüística y Literatura. Costa Rica. Correo electrónico: xdelrio@gmail.com

Recepción: 15/1/2016. Aceptación: 16/3/2016.

one of the most common errors is the omission of diacritical marks, since these techniques overlook stress. The recognition of the stressed syllable is vital in Spanish, since stress pattern is not fixed, as in other languages. This is to a great extent due to the fact that these techniques were developed originally for English, which lacks diacritical marks for stress, since it is not distinctive for word differentiation. Our proposal is to incorporate pitch analysis in automatic speech recognition techniques to improve writing of diacritical marks. Previous research has proven false the extended belief that the stressed syllable is always the loudest or most intense. Thus, intensity by itself is not the best indicator of the location of the stressed syllable in a word, i.e., the one that by graphic convention should carry a diacritical mark, and pitch should also be considered. Our research shows that pitch analysis helps to overcome this problem.
Keywords: diacritical accent, automatic speech recognition, natural language processing.

1. Introducción

La tilde diacrítica es un signo que se usa en el español para distinguir palabras con patrones de entonación y significados distintos, que de otra forma coincidirían en su forma escrita. Dicha marca es una convención gráfica y un hablante que conozca la norma va a marcarla. A la hora de transcribir una locución, un sistema de reconocimiento automático del habla debería colocar las tildes de la forma más aceptada, la cual es dictada por la Real Academia Española. Sin embargo, el empleo u omisión de estas puede perderse en la transcripción, debido a la incapacidad del sistema de reconocimiento del habla para reconocer tal detalle. La importancia de evitar estos errores radica en la asociación entre la validez del mensaje y la conformidad de su presentación con la norma establecida. Frases como «a sí mismo», «así mismo» y «asimismo» constituyen un ejemplo en el cual cada una se diferencia de las otras al ser pronunciadas en voz alta mediante el intercambio en la posición de los acentos. Ignorar tal acentuación puede conducir a una transcripción errónea.

Los sistemas de reconocimiento del habla modernos han avanzado enormemente desde sus inicios. Mediante recursos computacionales ubicados en la nube y captando retroalimentación proveniente de grabaciones de una vasta cantidad de hablantes, la precisión de estos sistemas permite un sinfín de aplicaciones cotidianas. Desafortunadamente, aún no son suficientemente precisos al identificar los acentos diacríticos del español. Tales faltas se atribuyen en parte a los métodos de codificación y etiquetado de las palabras empleadas para construir las bases de datos, que dan soporte a los sistemas de reconocimiento. La representación más empleada es la de coeficientes de *cepstrum* en la frecuencia *mel (MFCC)*, que da como resultado una serie de valores que indican la distribución espectral de una muestra de audio. El *cepstrum* de una señal de voz permite separar las ondas de sonido producidas por las cuerdas vocales, de los filtros aplicados a ellas, según la disposición de los demás órganos del habla (la lengua, el paladar, los labios, etc.). La utilidad de realizar dicha separación es ubicar las frecuencia de los formantes. En fonética, los sonidos vocales se caracterizan entre otras cosas por sus formantes. Desafortunadamente, la información sobre la sonoridad y altura no está contenida en los MFCC, siendo esta vital para decidir si se debe añadir un acento diacrítico a una palabra.

Una estrategia común para resolver este problema es añadir una etapa de postprocesamiento al sistema de reconocimiento del habla. Utilizando algoritmos de análisis gramatical, se puede detecta la función que cada palabra tiene en la oración, con lo cual se determina si debe escribirse una tilde. Hay literatura extensa sobre cómo realizar este proceso; la presente propuesta

de investigación sugiere más bien examinar la información de altura de las frases para detectar directamente la presencia de acentos diacríticos en el análisis de la señal.

Formalmente, la tilde diacrítica se define como: «el acento gráfico que permite distinguir palabras con idéntica forma, pero que pertenecen a categorías gramaticales diferentes. En general, llevan tilde diacrítica las formas tónicas (...) y no la llevan las formas átonas (...)» (Real Academia Española, 2005)1. Se aplica la tilde diacrítica cuando, sin ella, la entonación y el significado de una palabra resultan ambiguos.

En el presente trabajo, se empleará el tono en distintos instantes de la señal de un hablante para tratar de decidir si debe añadirse o no una tilde diacrítica. El tono es un atributo perceptual del sonido, el cual permite a un oyente ordenar sonidos en una escala dependiente de la frecuencia. Este atributo permite comparar dos sonidos entre sí y decir si uno es más agudo o más grave que el otro. Desde un punto de vista de la física, el tono queda determinado por el período de la onda (Klapuri y Davy, 2007).

El concepto físico de tono se vincula con el de contorno de la entonación, que es la línea melódica que constituye una unidad estructural o significativa en la entonación del discurso (Cantero, 2002). Para una frase u oración dada (como el ya mencionado ejemplo de «a sí mismo», «así mismo» y «asimismo"), si se varía el contorno de maneras predeterminadas, se varía el significado del conjunto.

Hay una serie de aspectos que dificultan caracterizar el contorno entonativo de manera inequívoca, lo que se conoce como entonación no lingüística. Cantero (2002) explica que hay una serie de rasgos fónicos emocionales y volitivos, que tienen un efecto tangible en la forma en la cual una misma frase es pronunciada por diversos hablantes. La fonoestilística «estudia aquellos rasgos fónicos del habla no relevantes lingüísticamente o, cuando menos, no relevantes en el mismo sentido en que lo son los rasgos distintivos, que distinguen fonemas» (Cantero, 2002: 188). Los rasgos en cuestión aportan diversos caracteres, entre los que se mencionan información dialectal, sociolingüística, personal y expresiva. Una de las consecuencias más importantes de la entonación no lingüística para el desarrollo de este trabajo es la reubicación a voluntad del hablante de la declinación en el contorno entonativo, lo cual obliga al método de reconocimiento a considerar que los acentos en los segmentos tonales no pueden ser ubicados con total certeza.

Con el fin de apreciar el comportamiento del contorno entonativo de una frase, es necesario estimar la altura de la señal para los segmentos sonoros. Una de las principales diferencias entre las emisiones sonoras y las sordas es la periodicidad de la señal, la cual resulta ausente para las sordas. Se habla de una función periódica cuando se repite su valor en períodos regulares. Para tales tipos de señales, la conformación armónica en un momento del tiempo es similar a momentos adyacentes a él en la señal, pero no en momentos distantes.

A las vocales y algunas consonantes les corresponden señales pseudoperiódicas (periódicas en intervalos pequeños del tiempo). Otras consonantes se asemejan más bien a un ruido, en el cual la señal se comporta de manera aleatoria y el oyente no puede percibir un tono.

Una de las tareas que debe realizar un reconocedor automático del habla es diferenciar fonemas. Para ello, necesita encontrar la disposición de sus formantes. Fant (1960) define los formantes como los picos espectrales en el espectro de la voz (cuando se habla de fonética, la resonancia acústica del tracto vocal también se asocia con el término). En otras palabras, los formantes son clave para reconocer vocales. Cuando un hablante pronuncia un fonema, lo hace configurando distintos componentes del aparato vocal en posiciones específicas y modificado su resonancia. Las cinco vocales del español se distinguen entre sí por la disposición de los formantes. Las frecuencias promedio para los dos primeros formantes se muestra en la tabla 1.

TABLA 1

Frecuencia promedio del primer y segundo formantes y su desviación estándar (Bradlow 1994)

Vocal	Frecuencia del primer formante (desv. est.)	Frecuencia del segundo formante (desv. est.)
i	286 (6)	2147 (131)
e	458 (42)	1814 (131)
a	638 (36)	1353 (84)
o	460 (19)	1019 (99)
u	322 (20)	992 (121)

Como se mencionó anteriormente, para poder obtener la información, referente a la posición en la frecuencia de los distintos formantes de una vocal, es necesario proveer un mecanismo en el cual se pueda separar la señal generada por la vibración de las cuerdas vocales del efecto que el aparato fonador tiene sobre la señal. El aparato vocal humano puede modelarse como la interacción de dos componentes: un oscilador pseudoperiódico, correspondiente a las cuerdas vocales; y un filtro, caracterizado por los formantes identificables en la envolvente espectral. En cuanto al *cepstrum* es una herramienta matemática propuesta por Bogart *et al.* (1963), empleada para separar las descripciones espectrales del oscilador y del filtro. Esta herramienta ha sido mejorada con la invención de los MFCC, que no solo usan un modelo del sistema auditivo para calcular el *cepstrum*, sino que también lo representan de una forma más compacta (que gasta menos recursos computacionales).

2. Metodología

Para efectos de este estudio, se procedió a recolectar ocurrencias de palabras monosílabas con tilde diacrítica en distintas locuciones, tomadas de diversos videos alojados en YouTube. Se consideraron distintos tipos de contenido, tales como noticieros de televisión, monólogos, y *sketches* de comedia. Los monosílabos considerados en el estudio fueron los siguientes: *de, dé, el, él, mas, más, mi, mí, se, sé, si, sí, te, té, tu* y *tú.* Cada muestra consiste en alguna de estas palabras, acompañadas de su contexto (por ejemplo, una de las muestras consiste de la frase «una sustancia *que te* va a ayudar a acelerar *tu* metabolismo celular y orgánico»). En total, se recopilaron 185 muestras y pronunciadas por personas de ambos sexos; además, provenientes de Latinoamérica y España. Hay 102 monosílabos átonos en su contexto y 83 monosílabos tónicos.

Antes de analizar las muestras, se consideraron las siguientes dos hipótesis:

1. Cuando se presenta una palabra que no deba escribirse con acento diacrítico, se apreciará una caída de al menos un 6% con respecto a la altura de la sílaba o segmento tónico de la palabra precedente, antecedente o ambas.
2. Cuando se presenta una palabra que deba escribirse con acento diacrítico, se apreciará un incremento de al menos un 6% con respecto a la altura de la sílaba o segmento tónico de la palabra precedente, antecedente o ambas.

La decisión de emplear un margen del 6% se fundamenta en que esa cantidad corresponde a medio *tono* o un *semitono* en la escala musical temperada (la distancia entre las notas do y do sostenido, por ejemplo). No se encontró en la literatura una unidad de medida que usaran los lingüistas para esto, ni tampoco un umbral que usaran para indicar la existencia de una caída de

tono, por lo que se recurrió a esta medida musical. La hipótesis es que, para marcar un cambio en la entonación, el hablante varía su tono en al menos un grado (un tono), por lo que un medio tono era un buen umbral para reconocer un cambio en la entonación.

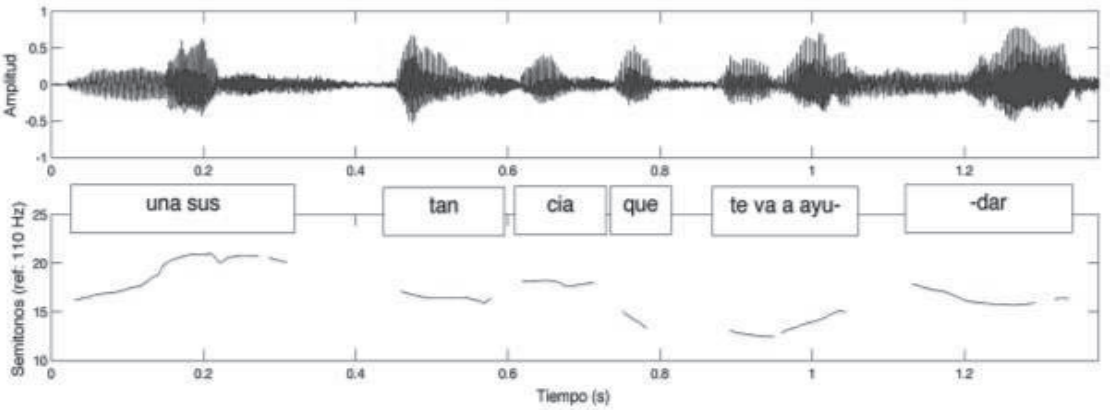

FIGURA 1

Ejemplo de una señal de audio correspondiente a la pronunciación de la frase "una sustancia que te va a ayudar" y su contorno entonativo.

Para cada una de las muestras, se consideró la palabra anterior y la siguiente en la oración donde se hallaron. Se empleó un algoritmo de estimación de altura para obtener el contorno entonativo de la muestra y las palabras descritas (ver figura 1). Luego, se calculó la mediana del tono de cada segmento para eliminar el efecto de valores extremos en la estimación de la altura (a veces causados por errores del algoritmo) y se obtuvo una descripción simplificada del contorno entonativo (ver figura 2). Las muestras correspondientes a monosílabos átonos se compararon con los segmentos tanto a su izquierda y como a su derecha para ubicar picos en el tono. Similarmente, los monosílabos tónicos fueron comparados con los segmentos donde se hallaron valles en el tono.

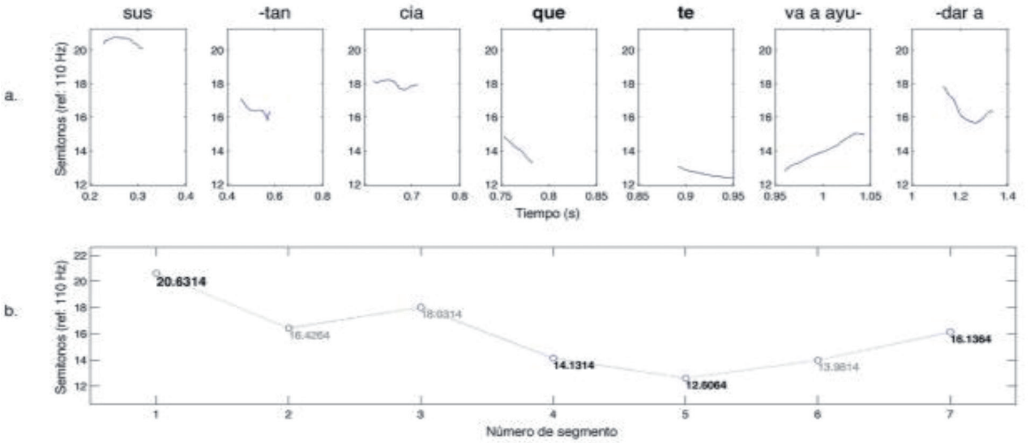

FIGURA 2

Contorno entonativo de la figura 1 y mediana del tono de cada una de los segmentos vocales.

3. Resultados

A continuación se muestran los resultados para los monosílabos. La tabla 2 muestra los resultados obtenidos para los monosílabos átonos (en su contexto). En la muestra seleccionada, el 72,6% de los casos el tono está por debajo del tono máximo de alguno de los segmentos vocálicos que los rodean, por al menos un semitono, es decir, un 6%. Por su parte, la tabla 3 muestra los resultados para los monosílabos tónicos en su contexto. Además, en la muestra seleccionada, el 62,7% de los casos el tono está por encima del tono mínimo de alguno de los segmentos vocálicos que los rodean, por al menos un semitono.

TABLA 2

Resultados obtenidos con monosílabos átonos. P.I.: pico a la izquierda, P.D.: pico a la derecha, P.A.: pico a ambos lados

	N.º de muestras	P.I.	P.D.	P.A	Total
Átonos individuales	57	13	8	22	43
Átonos proseguidos por otro monosílabo	23	2	6	7	15
Átonos precedidos por otro monosílabo	15	4	6	0	10
Átonos al principio de una oración	7	6	-	-	6
Total	102	25	20	29	74

TABLA 3

Resultados obtenidos con monosílabos tónicos. V.I.: valle a la izquierda, V.D.: valle a la derecha, V.A.: valle a ambos lados

	N.º de muestras	V.I.	V.D.	V.A.	Total
Tónicos individuales	42	5	15	3	23
Tónicos proseguidos por otro monosílabo	10	2	4	0	6
Tónicos precedidos por otro monosílabo	20	6	4	6	16
Tónicos al principio de una oración	11	7	-	-	7
Total	83	20	23	9	52

Como se puede observar, los datos apoyan ligeramente más la hipótesis 1 (valle en los átonos) que la hipótesis 2 (pico en los tónicos). Sin embargo, ambas hipótesis están lejos de ocurrir en la totalidad de los casos.

Es interesante destacar que en las muestras utilizadas la gramática fue suficiente para resolver si se debe usar tilde o no. Por ejemplo, en «sustancia que *te* va a ayudar» está claro que *te* no es un sustantivo (la bebida), sino el pronombre posesivo de la segunda persona. Por tanto, el mensaje posiblemente hubiera quedado claro independientemente de que el hablante hubiera bajado el tono al pronunciar *te* o no.

4. Conclusión

Aunque, las expectativas iniciales no se cumplieron, los resultados obtenidos tampoco son despreciables. En relación con otros criterios, el tono puede ser de ayuda en casos donde los demás criterios no sean concluyentes. Por ejemplo, cuando

se disponga de una situación donde haya dos posibilidades, ambas gramaticalmente correctas (como «sí, como no» y «si como, no"). Por tanto, sería interesante averiguar si en tales casos el hablante es más cuidadoso con la entonación para evitar una interpretación errónea de su mensaje.

Nota

1. El acento diacrítico también puede emplearse para distinguir palabras que corresponden a la misma categoría gramatical, pero que de otra forma coincidirían gráficamente. Un ejemplo de este caso consiste en la pareja «tramite» y «tramité», donde el acento distingue únicamente tiempos verbales.

Bibliografía

Bogert, B. P., M. J. R. Healy y J. W. Tukey. 1963. «The quefrency analysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking». En: *Proceeding of the Symposium on Time Series Analysis:* 209-243.

Bradlow, A. R. 1994. «A comparative acoustic study of English and Spanish vowels». En: *Journal of the Acoustical Society of America* XCVII (3): 1916-1924.

Cantero, F. J. 2002. *Teoría y análisis de la entonación*. Barcelona: Edicions Universitat Barcelona.

Fant, G. 1960. *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations.* La Haya: Mouton Publishers.

Klapury, A. y M. Davy. 2007. *Signal Processing Methods for Music Transcription*. Londres: Springer-Verlag.

Real Academia Española. 2005. Diccionario panhispánico de dudas. http://lema.rae.es/dpd/?key=tilde. Consulta: 21 de julio de 2015.

Stevens, S. S., J. Volkmann y E. B. Newman. 1937. «A scale for the measurement of the psychological magnitude pitch». En: *Journal of the Acoustical Society of America* VIII (3): 185–190.