Procesamiento del Lenguaje Natural, núm. 28 (2002), pp. 99-104 recibido 1-3-2002; aceptado 2-5-2002

Medición Cuantitativa de la Velocidad del Habla

Wainschenker Rubén
Doorn Jorge
Castro Marcela
INTIA - Facultad de Ciencias Exactas
Universidad Nacional del Centro de la
Provincia de Buenos Aires
Paraje Arroyo Seco - Campus Universitario,
(7000) Tandil – Argentina
{rfw, jdoorn, mcastro} @ exa.unicen.edu.ar

Resumen: La magnitud velocidad del habla no tiene una definición precisa, si bien es ampliamente utilizada tanto en actividades diarias como en tareas específicas como lo son la dactilografía, estenografía y taquigrafía, entre otras. Esta noción resulta fundamental en el contexto del estudio del comportamiento de los alófonos de un idioma, cuando se intenta realizar síntesis del habla con algún grado de control sobre la velocidad del sonido producido. No existe información precisa para el Castellano y, mucho menos, para una de sus variantes como lo es la hablada en Uruguay y en el centro y sur de Argentina llamada Castellano Rioplatense. En este artículo se presenta una caracterización cuantitativa de las nociones intuitivas de velocidad normal, lenta y rápida del habla. El mismo describe investigaciones realizadas una amplia base experimental, ya que sus conclusiones son obtenidas de mediciones llevadas a cabo sobre 120 textos emitidos por diferentes locutores a distinta velocidad, en un contexto completamente libre de condicionamientos. Más de la mitad de los textos fueron obtenidos desde fuentes públicas y de personas que nunca supieron que fueron especialmente grabadas para este estudio.
Palabras clave: síntesis del habla.

Abstract: The magnitude of the speed of speech does not have a precise definition, as it is widely used both in daily activities and in specific tasks. This notion is fundamental in the context of the study of the behavior of the speaker of a language, when it is tried to obtain a speech synthesis with some level of control on the speed of the speech. There is no knowledge about Spanish, and there is not on its variants. In this paper, it is presented a quantitative characterization of the intuitive notions of normal, low and high speed of speech.
Keywords: speech synthesis.

1 Introducción

El habla es una de las principales manifestaciones de la inteligencia humana naturalmente utilizada para dar y recibir información. La voz es la realización acústica del complejo mecanismo que representa el lenguaje humano. Actualmente, éste ha sido tomado como objeto de estudio de diversas disciplinas. Cada ciencia o arte lo aborda desde un punto de vista distinto de acuerdo a sus objetivos [1].

Acústicamente, la voz se puede ver como el resultado de una fuente de sonido alterada por un filtro selectivo representando el tracto vocal; las propiedades varían continuamente durante el proceso de producción del habla. Estas variaciones dependen de la forma del tracto vocal, el cual, a su vez, depende de la posición de los órganos articuladores como lengua y/o labios.

La habilidad para producir sonidos, no implica la habilidad de "hablar". Desde el punto de vista de la síntesis de voz, la simple concatenación de los sonidos elementales de un lenguaje conduce a generar un sonido poco natural e ininteligible. La producción de voz natural y continua es un desafío permanente en

ISSN 1135-5948

aplicaciones informáticas que necesitan sintetizar voz, ya que la mayoría de los enfoques conocidos sólo producen un sonido artificial [2,3].

Una característica, rara vez tenida en cuenta, es la velocidad de la emisión de la voz. Un modelo general, capaz de producir voz sintetizada, debería considerar la velocidad del habla como uno de sus parámetros. En otras palabras, conocer la velocidad de la voz es una etapa previa e ineludible para comprender cómo se comportan los diferentes alófonos cuando la misma varía. También, las técnicas corrientes usadas para contraer y expandir locuciones (tiempo de compresión y expansión o Speech Skimmer) deberían poner especial atención en cómo la velocidad global del habla impacta sobre cada alófono [4]. Es muy habitual notar la degradación producida en el mensaje al escuchar una locución que ha sido comprimida o expandida es debida al inadecuado tratamiento de la evolución de los alófonos con la velocidad del habla [5].

Debido a que la síntesis de voz humana trata de imitar la voz natural, un requisito primario es conocer a ésta tanto como sea posible. Los valores conocidos de la velocidad del habla para el castellano y utilizados antes de este estudio cuantifican, solamente, la velocidad producida en condiciones normales. El objetivo de este artículo es producir una contribución en este sentido.

La parte experimental del estudio fue realizado para el Castellano Rioplatense, hablado en el Centro y Sur de Argentina (incluido Uruguay). Si bien, éste tiene pequeñas variaciones en relación con otras versiones del Castellano, la mayoría de las conclusiones obtenidas son aplicables al Castellano en general.

2 Velocidad del Habla

El concepto "velocidad del habla" juega un papel trascendente dentro del comportamiento de los alófonos de un lenguaje. Esta importancia se incrementa, en el área de voz humana procesada por computadoras, cuando se trata de desarrollar un algoritmo robusto capaz de producir voz sintetizada de buena calidad [2,3]. Por otro lado, este tipo de datos es útil también cuando se trata de diagnosticar enfermedades que provocan problemas en el habla [6].

Navarro Thomas [7] define como velocidad normal a aquella que se obtiene cuando se emiten 205 palabras por minuto (ppm) mientras que, Loprete [1] establece esta misma noción entre 120 y 150 ppm. Los contextos de ambas afirmaciones difieren, ya que la primera corresponde al Castellano Hispano hablado a mediados del siglo XX y la segunda al Castellano Rioplatense de fines del mismo siglo.

Otras fuentes [8] parecen confirmar al menos el límite inferior del habla normal. Por ejemplo, en los cursos de taquigrafía consideran apto un estudiante que maneje 120 ppm. En otros lenguajes, la información disponible no contiene datos preciso para la velocidad del habla. Por ejemplo, en el inglés los informes califican como velocidad normal a la producida entre 130 y 200 palabras por minuto [9].

La falta de precisión persiste cuando la unidad de medición es tenida en cuenta, puesto que diferentes textos contienen diferentes números de alófonos por palabra. El sentido común y las experiencias que se relatan en las siguientes secciones, indican que la unidad alófono por unidad de tiempo es casi invariante con la naturaleza del texto, aún en textos cortos.

Para evaluar lo apropiado de una técnica de medición es conveniente recordar los principios básicos de la teoría de mediciones, ampliamente difundida en otras disciplinas. Habitualmente cinco categorías de mediciones se aceptan [10,11]:

1. Cualitativa
2. Ordinal
3. Por intervalos
4. Proporcional
5. Cardinal

La calidad de la información obtenida crece desde la medición cualitativa a la cardinal. La primer mejora respecto de la medición cualitativa establecida por la medición ordinal consiste en garantizar un orden parcial entre los objetos medidos. La segunda mejora consiste en asegurar un orden total mientras que, la medición proporcional se caracteriza por tener un punto de referencia absoluto o cero de la escala. Finalmente, la medición cardinal requiere una unidad de medida precisa, repetitiva y confiable.

Todas las mediciones realizadas en el trabajo experimental reportado en este artículo fueron hechas en alófonos por segundo. La unidad alófono no cumple con la propiedad requerida en las mediciones cardinales, pero se

aproxima a ella mucho más que las mediciones basadas en palabras por minuto. Las mediciones cualitativas no deberían despreciarse ya que la extensión de su uso es muy grande, pero es conveniente que las mismas se precisen con una medición proporcional comparativa.

3 Proceso Experimental

Durante la experiencia, 30 personas diferentes leyeron 120 textos a distintas velocidades. La mitad de ellos leyó el texto conociendo el objetivo de la experiencia, mientras que la otra mitad fue grabada del fuentes del dominio público tales como radio y televisión.

De los 120 textos leídos, 40 fueron seleccionados enfatizando el objetivo de tener un conjunto de muestras lo más completo y uniforme posible reduciendo, tanto como sea posible, algún ruido producido por hablantes, fuente del sonido, y otros. El número de alófonos y la duración completa de la locución fue medido para cada texto seleccionado como se muestra en la Tabla 1.

El mismo grupo de 40 textos fue medido cualitativamente utilizando un grupo independiente de oyentes quienes, con criterio propio, debían adjudicar una y sólo una de las siguientes medidas: muy lento, lento, normal, rápido y muy rápido. Ninguno de estos oyentes tuvo acceso, en ningún momento, a ninguna de las opiniones de los otros, y cada uno escuchó los 40 textos en un orden diferente al de los demás. A partir de estos datos, se puede establecer el valor medio de velocidad normal en aproximadamente 13 alófonos por segundo, el cual corresponde a 150 ppm. Para esta comparación se usó un valor medio de 5,1 alófonos por palabra (ver próxima sección).

La experiencia también mostró datos cuantitativos para las velocidades del habla muy rápida, rápida, lenta y muy lenta, los cuales no podrían compararse con otras fuentes. Estos valores se muestran en la Tabla 2.

En locuciones prolongadas un factor distorsivo es introducido por las pausas entre palabras, especialmente en las locuciones muy lentas. En locuciones normales y rápidas estos lapsos se contraen y desaparecen.

	Alófonos (a)	Duración (s)	Velocidad (a/s)
T1	260	22.00	11.8
T2	298	30.00	9.9
T3	395	30.00	13.2
T4	570	30.00	19.00
T5	443	30.00	14.8
T6	397	30.00	13.2
T7	577	30.00	19.2
T8	335	20.00	16.7
T9	220	20.00	11.0
T10	218	21.00	10.4
T11	305	16.00	19.1
T12	95	28.00	3.4
T13	216	29.00	7.5
T14	413	28.00	14.8
T15	626	30.00	20.9
T16	903	31.70	28.5
T17	130	36.62	3.6
T18	427	29.70	14.4
T19	791	28.80	27.5
T20	61	31.00	2.0
T21	166	30.50	5.4
T22	364	30.00	12.1
T23	574	34.33	16.7
T24	826	35.34	23.4
T25	130	33.35	3.9
T26	381	31.08	12.3
T27	714	28.76	24.8
T28	445	31.36	14.2
T29	701	32.60	21.5
T30	488	33.00	14.8
T31	567	33.00	17.2
T32	356	31.53	11.3
T33	67	30.00	2.2
T34	543	40.00	13.6
T35	235	20.00	11.8
T36	310	25.00	12.4
T37	486	30.00	16.2
T38	216	30.00	7.2
T39	227	30.00	7.6
T40	708	30.00	23.6

Tabla 1. Número de alófonos, duración total y alófonos por segundo para el texto seleccionado

4 Resultados

Analizando los valores de la Tabla 2, en cada línea se muestra la identificación del texto, su velocidad medida en cantidad de alófonos por segundo (a/s) y en las siguientes cinco columnas la cantidad de oyentes que clasificaron la velocidad de locución con el mismo rótulo. La tabla está ordenada de acuerdo a la velocidad medida para resaltar el ajuste con la clasificación realizada por los oyentes.

La Tabla 3 contiene los valores medios y las desviaciones standard para cada grupo, expresadas en alófonos por segundo (a/s). Se establece entonces que al hablar a (3.3±1.3) a/s, (8.5±2.1) a/s, (12.9±2.0) a/s, (17.7±2.2) a/s y (24.4±3.0) a/s se habla: muy lentamente,

lentamente, normalmente, rápidamente y muy rápidamente.

Texto	Veloc.	Muy lenta	Lenta	Normal	Rápida	Muy Rápida
20	2.0	31				
33	2.2	31				
12	3.4	31				
17	3.6	31				
25	3.9	31				
21	5.4	9	22			
38	7.2	4	24	3		
13	7.5	4	24	3		
39	7.6		22	9		
2	9.9		21	10		
10	10.4		7	24		
9	11.0		4	27		
32	11.3		5	26		
35	11.8		5	26		
1	11.8		5	26		
22	12.1		4	27		
26	12.3		5	26		
36	12.4			31		
3	13.2			28	3	
6	13.2			28	3	
34	13.6			28	3	
28	14.2			25	6	
18	14.4			25	6	
14	14.8			25	6	
5	14.8			20	11	
30	14.8			23	8	
37	16.2			8	23	
23	16.7			8	23	
8	16.8			5	26	
31	17.2			3	23	5
4	19.0				26	5
11	19.1			7	24	
7	19.2				21	10
15	20.9				21	10
29	21.5				20	11
24	23.4					31
40	23.6					31
27	24.8					31
19	27.5					31
16	28.5					31

Tabla 2. Comparación de la velocidad del habla determinada cuantitativamente y cualitativamente

Grupo cualitativo	Promedio (a/s)	Desviación Standard (a/s)
4.1.1 Muy lenta	3.3	1.3
Lenta	8.5	2.1
4.1.2 Normal	12.9	2.0
Rápida	17.7	2.2
Muy rápida	24.4	3.0

Tabla 3. Promedio y desviación standard de la velocidad del habla

Como se muestra en la Tabla 4, cada promedio con su desviación standard permiten definir un intervalo de confianza tal que es más probable que una locución arbitraria pertenezca a la categoría cualitativa correspondiente que a los grupos vecinos.

	Muy lenta	Lenta	Normal	Rápida	Muy rápida
4.1.2.1 romedio	3.3	8.5	12.9	17.7	24.4
Desviación	1.3	2.1	2.0	2.2	3.0
Límites	5.3	10.7	15.2	20.5	

Tabla 4. Límites entre grupos cualitativos

Otra información importante a tener en cuenta es la velocidad del habla para el cual la probabilidad de pertenecer a un grupo o al grupo vecino es la misma. En otras palabras, para qué velocidad teórica del habla se tendría que el 50 % de los encuestados la calificarían como rápida, y el otro 50 % como normal. Estos valores servirían para establecer los límites grupos cualitativos. La Tabla 4 muestra estos límites.

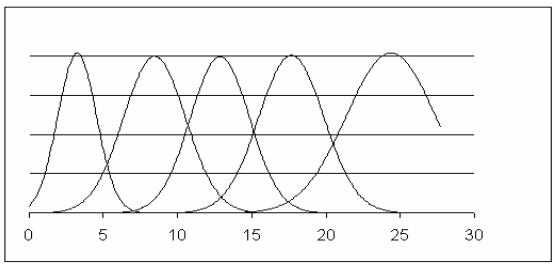

Figura 1. Distribución teórica de probabilidades para cada grupo cualitativo

La Figura 1 muestra las curvas de densidad de probabilidades correspondientes a las cinco categorías usadas a lo largo de este estudio. El eje vertical es presentado en unidades arbitrarias a fin de facilitar su comprensión.

A partir de la Tabla 4 se puede considerar que todo texto emitido a una velocidad inferior a 5,3 a/s entra en la categoría de muy lenta en tanto que si está entre 5,3 a/s y 10,7 a/s se la puede considerar como lenta; entre 10,7 a/s y 15,2 a/s se trata de una locución normal; será rápida si se emiten entre 15,2 a/s y 20,5 a/s y será muy rápida si supera los 20,5 a/s. Hablando en forma poco precisa los límites entre las categorías utilizadas son 5, 10, 15, y 20 a/s.

El hecho que la diferencia entre esos límites es siempre el mismo valor de 5 alófonos por segundo, da una sensación adicional de

confianza en los datos obtenidos independientemente de las pruebas estadísticas.

Todos los datos presentados en las Tablas 1 a 4 fueron recogidos incluyendo las pausas entre palabras. Esto fue realizado en forma tal que hacer las comparaciones con datos previos resulte más fácil. Se puede, entonces, argumentar que las pausas deforman la velocidad expresada en alófonos por segundo. El punto de vista soportado en este artículo es proveer datos acerca de la velocidad global del habla. Si las pausas fueran removidas, los límites dados en la Tabla 4 serían 6,7; 12,7; 16,9 y 21,6.

5 Comparación con información previa

Como ya se indicó más arriba, los datos de referencia disponibles están expresados en cantidad de palabras por minuto. Además, para comparar palabras por minuto con alófonos por segundo, es necesario establecer un valor representativo de cantidad de alófonos por palabra; entonces, un nuevo conjunto de 30 textos fue seleccionado. Nuevamente fueron seleccionadas diferentes fuentes para decrementar la influencia de posibles ruidos. Contando palabras y alófonos en aquellos textos se obtuvo el valor de 5,1 alófonos por palabra. La confiabilidad de este dato es excelente en relación con la precisión de los datos de referencia disponibles.

Convirtiendo los valores de la Tabla 4 a palabras por minuto para el caso del habla normal se puede expresar que la misma varía de 128 a 180 palabras por minuto, valor que concuerda razonablemente con C. A. Loprete [1] quien estima que la velocidad media de una conversación gira en torno a las 120-150 ppm. Análogamente, en su curso de taquigrafía, M. Vasallo y C. Fusca de Elías [6] proponen standards para tomar nota de conversaciones, comenzando por una velocidad de 20 ppm y aumentando hasta alcanzar las 102 ppm con lo que se cubriría todo el rango desde muy lento, lento y completando el curso cerca del umbral superior de una locución lenta.

No se disponen de datos acerca de locuciones diferentes de la normal, pero la razonable equidistancia entre las medias de las cinco categorías utilizadas hace pensar que los resultados obtenidos son confiables también en estos casos.

6 Conclusiones

Los resultados obtenidos muestran que:

Se puede medir con suficiente precisión y repetitividad la velocidad del habla utilizando como unidad la cantidad de alófonos emitidos por segundo (a/s)

La práctica diaria es lo suficientemente estable para categorizar la rapidez del habla al menos en las cinco categorías mencionadas

Se pueden establecer razonables relaciones entre la velocidad del habla y las categorías establecidas, las que se presentan en la Tabla 4 y la Figura 1.

Se desconoce cómo la información cognitiva del contexto influye sobre la percepción de la velocidad del habla y como afectó este fenómeno en el experimento realizado. Por ejemplo, es de conocimiento común que un locutor de fútbol relata hablando en forma muy rápida, quizás esta información haya afectado alguna opinión recogida en el experimento.

Bibliografía

[1] C. Loprete. ´El Lenguaje Oral: Fundamentos, Formas y Técnicas´, Ed: Plus Ultra, 1984.

[2] F. Casacuberta, E. Vidal. ´Reconocimiento Automático del Habla´. Ed: Marcombo, 1987, pag. 63-68.

[3] R. Sproat. ´Multilingual Text-to-Speech Synthesis: The Bell Labs Approach´, 1998, Lower Academic Publishers.

[4] B. Arons. Speech Skimer: A System for Interactively Skimming Recorded Speech, ACM Transaction on Computer-Human Interaction, 1997, Vol. 4:1, page 3-38.

[5] M. Covell, M. Withgott and M. Slaney. Mach1: Nonuniform Time-Scale Modification of Speech, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 1998, page 12-15.

[6] C. Ferrer Riego, M. Hernandez-Díaz Huici. Obtención de un Índice Objetivo de Razón Lenta, Actas del VII Simposio de Comunicación Social, 2001, pag. 390-394.

[7] T. Navarro. ´Manual de Pronunciación Española´, Consejo Superior de Investigaciones Científicas, Madrid, 1950.

[8] M. Vasallo, C. Fusca de Elias. ´Estenografía Vigente´, Ed: Kapeluz, 1992, Vol. 2, pag. 166-169.

[9] B. Arons. Techniques, Perception and Applications of Time-compressed Speech. Proceedings of 1992 Conference, American Voice I/O Society, 1992, page 169-177.

[10] B. Klaassen Klaas. Electronic Measurement and Instrumentation, Cambridge University Press.,1996, page 1-15.

[11] L. Briand, S. Morasca, V. Basili. Property-Based Software Engineering Measurement, IEEE Transactions on Software Engineering, 1996, Vol. 22:1, page 68-85.