Síntesis de voz en el dialecto venezolano por medio de la concatenación de difonos

Speech synthesis of the venezuelan dialect via diphone concatenation

M. Rodríguez*, E. Mora**
*Departamento de Electrónica y Comunicaciones, Facultad de Ingeniería,
**Departamento de Lingüística, Facultad de Humanidades,
Universidad de Los Andes, Mérida, Venezuela

C. Cavé
Laboratoire de Parole et Langage, Université en Provence, Aix-en-Provence. France

Resumen

En este trabajo se presenta un sistema de síntesis de voz venezolana. Este sistema está basado en el método de concatenación de difonos para lo cual se utilizó una base de 794 difonos (Rodríguez et al. 2003) que permite generar cualquier enunciado en español venezolano. Se describe como generar la base de datos de difonos.

Palabras claves: tecnologías del habla, síntesis, concatenación de difonos.

Abstract

We present a system for synthesizing speech with a Venezuelan voice. This system is based on the method of diphone concatenation for which we used the 794 diphone data base (Rodriguez et al 2003), which allows for the generation of any sentence in Venezuelan Spanish. We describe how to generate the data base.

Key words: speech technology, synthesis, diphone concatenation.

1 Introducción

La síntesis del habla siempre ha merecido a lo largo de la historia un gran interés, por distintas motivaciones y aplicaciones potenciales. Desde hace aproximadamente 30 años, gracias al poder y bajo costo de procesamiento de las computadoras, legiones de investigadores a nivel mundial se han dedicado a la implementación y perfeccionamiento de los sistemas de síntesis de voz, en varios idiomas y dialectos. Un excelente tratamiento del tema se encuentra en Huang et al (Huang et al 2001).

Son muchas las aplicaciones de la voz sintética,, pero se pueden agrupar en a) como medio de transmitir información en forma auditiva desde un sistema al hombre, bien sea directamente o por vía telefónica, b) como ayuda a los minusválidos, dándole capacidad de voz a los sordos y capacidad de lectura asistida a los ciegos, y c) como una herramienta de investigación para realizar estudios de los mecanismos de percepción y producción del habla.

El objetivo es presentar algunos antecedentes de esta línea de trabajo, y a su vez entrar en detalles más concretos de un sistema que se ha desarrollado para sintetizar voz en el dialecto venezolano del idioma español.

Elaborar sistemas de síntesis para las diferentes variedades de una misma lengua se justifica por dos razones importantes: la primera es justamente tener en cuenta las variedades dialectales de una lengua dada que ciertamente reflejan una identidad particular, una forma de vida y realidad cultural. La segunda razón está ligada a la afirmación y reivindicación de una identificación cultural cuya lengua, en sus variedades dialectales es un elemento importante, por lo tanto, los usuarios de tecnologías del habla desean cada vez con mayor entusiasmo interactuar con un sistema que "hable como ellos" y no con un sistema que utilice una lengua "neutral" o "robotina". En Venezuela, la Universidad de Los Andes ha reunido especialistas tanto del campo de la

lingüística como de la ingeniería para hacer posible este logro y representa el único grupo del país que ha abordado el tema sistemáticamente y continuamente por varios años.

En la Fig. 1, se muestra el esquema general de un sistema de síntesis, cuyo objetivo es la generación automática por medio de un sistema informático o electrónico de una señal acústica que simule la voz humana.

Fig. 1. Esquema general de un sistema de síntesis

Los sistemas que entregan respuesta hablada se pueden clasificar en tres categorías: 1) Sistemas que graban y reproducen mensajes (correo de voz, contestadores, etc.). 2) Sistemas con mensajes grabados, normalmente palabras o frases, que se encadenan para confeccionar un número relativamente pequeño de mensajes diferentes. Los inconvenientes de estos sistemas son que el vocabulario es muy reducido, no es posible modificar la prosodia ni el locutor, y cada nueva palabra de vocabulario requiere ser grabada. Tiene aplicaciones en sistemas de información (relojes parlantes, aeropuertos, industria, sistemas de alarmas, etc.), juguetes, aparatos domésticos. Finalmente, 3) Sistemas de síntesis general, sin limitaciones (por ejemplo un lector de libros o de páginas web, en voz alta, para invidentes). El grado de dificultad y de sofisticación crece desde la categoría 1 a la 3. Así para la síntesis general se requiere reproducir la multiplicidad de sonidos existentes del habla y tener conocimiento sintáctico del idioma.

Una comparación entre palabras articuladas en forma aislada y las mismas palabras habladas fluidamente demuestra que no solamente son más cortas sino que surgen cambios acústicos en la unión de las palabras debido al efecto de coarticulación y debido a reglas del idioma que cambian la pronunciación de las palabras en función del contexto. Además, la entonación, la energía y el ritmo apropiado para una frase no pueden sintetizarse si simplemente se concatenan palabras pregrabadas. Estas funciones prosódicas resultan ser extremadamente importantes para lograr una buena inteligibilidad y naturalidad de la frase (Quilis, 1983).

A fin de superar estas limitaciones de coarticulación, Rabiner, Schafer y Flanagan (1971) propusieron eliminar la ininteligibilidad de concatenar formas de onda por medio de un suavizamiento de las trayectorias de las características acústicas en las fronteras entre palabras. Este enfoque requiere un conocimiento profundo de estas características lo que a su vez implica un análisis de la voz para extraer las resonancias llamadas formantes.

Además se requiere un tipo de sintetizador basado en la reproducción de estas resonancias. Esto fue el foco de la investigación en los años 70 y llegó a su máxima expresión en idioma inglés con el sistema basado en el sintetizador Klatt (1980). Con este sistema salieron al mercado los primeros productos comerciales de conversión de texto a voz, donde se destaca el producto DecTalk de la empresa Digital Equipment Corporation que logró un buen nivel de inteligibilidad y de funcionalidad. Rodríguez et al (1984a) detalla los valores de los parámetros del sistema para todos los fonemas de la versión española. Sin embargo, la voz sintetizada tenía problemas de calidad; en algunos casos, por la dificultad en reproducir la transición de ciertos sonidos, sobre todo las consonantes, y también por pretender reducir a su mínima expresión la información inherente a los sonidos del habla, tratando su representación en base a los formantes, o por otras alternativas de codificación como la predicción lineal (LPC), coeficientes cepstrales (CC), coeficientes de área logarítmica, etc. cualquiera de los cuales agregaba un nivel de ruido o terminaba siendo una representación defectuosa.

Posteriormente, se fueron proponiendo alternativas y mejoras. Una de las estrategias era trabajar con unidades de síntesis más grandes que un fonema, de modo que esa parte difícil de describir analíticamente, las transiciones entre fonemas, fueran incorporadas a la unidad de sonidos del habla (Rodríguez et al, 1984b). Otra de las estrategias estaba basada en trabajar directamente con la señal en el tiempo, sin codificarla. Pero como la prosodia requiere variar la entonación y la duración de los fonemas en función del contexto, modificar una señal de voz en cuanto a su duración y su frecuencia fundamental, sin alterar el timbre de la voz, no es nada trivial.

A partir de allí, se fueron proponiendo técnicas de procesamiento de señal, gracias a las cuales se pudieron implementar estos cambios sin desmejorar la calidad del sonido.

El primer avance se logra con la técnica OLA, Overlap and Add, que se traduce Solapar y Sumar, y esas son las operaciones que se realizan con los sub-segmentos de la señal. Estas operaciones no modifican el espectro de un intervalo largo de la señal. A la vez, se logra variar la duración de una señal. Sea N el número de muestras un segmento de la señal. Se forman segmentos de 2N muestras, se multiplican por una ventana Hanning (una campana positiva), se desplazan y se suman distanciados N muestras. Resulta la mitad de la duración. Tiene el defecto de crear períodos de entonación irregulares. Al tener mayor cuidado en donde segmentar, se logra el segundo avance de consideración conocida como PSOLA (Pitch Synchronous Overlap and Add) (Moulines et al 1990), que se traduce como Solapar y Sumar Sincrónicamente con el Pulso Glotal. La meta es sintetizar una señal que tiene las mismas características espectrales de otra pero con una entonación y/o duración distintas. En principio se solapa y suma como antes, pero adicionalmente se tiene cuidado de que los segmentos sean síncronos con el pulso glotal de la voz. Las ventanas utilizadas se distancian a un periodo glotal, con una duración igual a 2 periodos glotales. De esta forma PSOLA permite modificar arbitrariamente la frecuencia fundamental y la duración del segmento

Una alta calidad de síntesis se logra con el sistema MBROLA de Dutoit (1997), donde se trabaja con el difono, descrito mas adelante, como unidad de síntesis. A su vez el grupo de trabajo de Dutoit forma parte de una propuesta internacional de facilitar herramientas a otros grupos de investigación en el desarrollo de sistemas de síntesis para otros idiomas y dialectos (Dutoit et al, 1996). Con este enfoque es que el grupo de la Universidad de Los Andes en unión con el Laboratoire de Parole et Langage de la Université en Provence emprende este proyecto de síntesis de voz Mora et al (2001) proponen un conjunto especial de difonos para el dialecto venezolano del español.

Después, los mayores adelantos se han basado en corpus y selección de unidades, es decir, utilizando unidades de tamaño variado, desde difonos hasta frases completas, e incluso con varios ejemplares de cada unidad cada una con una entonación distinta, para evitar en lo posible la necesidad de variar la entonación original de la frase (Lee et al 2003). Implica disponer de un corpus sumamente grande. Es el enfoque detrás de los sistemas comerciales de empresas como AT&T, Loquendo y ScanSoft que han logrado la mayor calidad.

2 Consideraciones de diseño

2.1 Unidades de síntesis

Hay varias alternativas de unidades de síntesis sublexicales, incluyendo fonemas, sílabas, demisílabas, y difonos, entre otros. En el sintetizador Klatt, la unidad por excelencia es el fonema, que consiste de el mínimo número de unidades necesarias para conformar el inventario de sonidos del habla. Sin embargo, como se ha indicado anteriormente, el fonema trae los siguientes defectos: es difícil sintetizar los fonemas no estacionarios, y la coarticulación de fonemas se debe generar por reglas, que en el mejor de los casos apenas logra una aproximación a la verdadera transición.

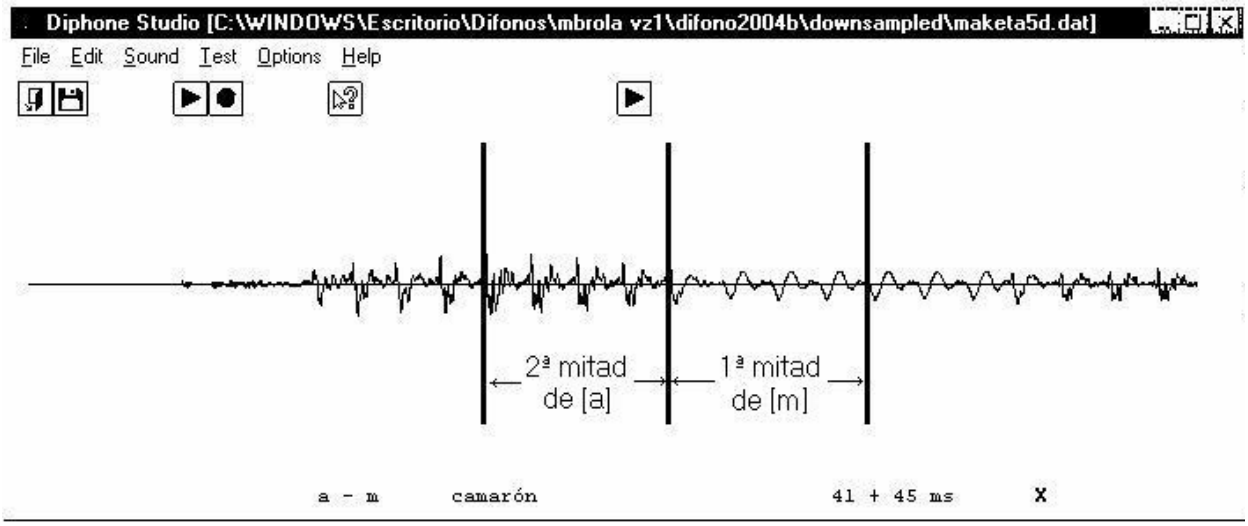


Fig. 2.. Segmentación del difono "a-m", extraído de la palabra "camarón" – incluye la segunda mitad de [a] y la primera mitad de [m].

Como alternativa se propuso en este trabajo el uso del difono, el cual comienza en la mitad de la parte estacionaria de un fono, un ejemplar físico de un fonema, y llega hasta la mitad de la parte estacionaria del próximo fono, como se ilustra en la Fig. 2 con el difono "a-m", segmentado de la palabra "camarón". Las virtudes de esta unidad son que la información sobre la coarticulación queda incluida dentro del segmento, mientras que el total de elementos es un número relativamente pequeño del orden de los fonemas al cuadrado. Como defecto, hay ciertas combinaciones, como los grupos consonánticos, que no contemplan los efectos de coarticulación.

2.2 Generación de la base de datos de unidades (difonos)

En el sistema de síntesis de fonemas por formantes, una gran cantidad de tiempo de los investigadores se invertía en determinar los parámetros acústicos de los fonemas. En este sistema basado en el difono como unidad de síntesis, igualmente se requiere una gran inversión de tiempo, pero de otra naturaleza, para generar una base de datos de difonos, tal como se indica en los pasos siguientes:

1. Determinación del conjunto de unidades fonéticas.
2. Determinación de la lista de unidades posibles (restricciones fonotácticas).
3. Generación de la lista de palabras o frases portadoras de las unidades.
4. Elección del locutor, preferiblemente con buena pronunciación del dialecto.Grabación del corpus de unidades de síntesis.
6. Preparación de las señales (eliminación de silencios).
7. Segmentación y extracción de las unidades de interés.

8. Verificación de la segmentación.
9. Codificación de la base de datos.
10. Evaluación acústica de la base de datos mediante la síntesis.

Por ser relativamente pocas unidades, lo que se traduce en una ventaja en aspectos computacionales, y a la vez incorporar la mayor parte de efectos de coarticulación, se decidió por el uso de la unidad de síntesis el difono. Para el dialecto venezolano del español, se trabajó con el conjunto de fonos propuesto por Mora et al (2001). A continuación se especifica este conjunto con su representación SAMPA, y en función de su descripción articulatoria:

- Vocales átonas: a, e, i, o, y u.
- Vocales tónicas: a*, e*, i*, o*, y u*.
- Oclusivas sordas: p, t, y k.
- Oclusivas sonoras: b, g, y d.
- Fricativas sonoras: B, G, y D.
- Africada: tS (correspondiente al "ch" de "chato").
- Fricativas sordas: f, s, s2 (un alófono en distensión de la "s"), y h (sonido de la "j" en "laja").
- Nasales: m, n, J (J corresponde a la "ñ").
- Laterales: l, L (corresponde a la "ll" de "llave" y a la "y" de "ayer"), r, y rr.
- Glides (denominados alternativamente semivocales): j, w (correspondientes respectivamente a las grafemas "i" y "u" en diptongos, como en "quieto" y "jueves") y
- Pausa: _.

De aquí se genera la matriz de la Tabla 1, donde se muestra el conjunto parcial de los difonos usados en este proyecto. En la fila superior se incluyen los fonos de la parte inicial del difono, y en la columna a la izquierda los fonos de la segunda parte del difono. En los cuadritos aparece un número representativo de la frase del corpus donde aparece el difono correspondiente. No todos los productos de fono por fono son difonos de la base de datos. En algunos cuadritos, en vez de un número aparece el símbolo $ seguido por un número del 1 al 7; esto corresponde a los casos de ausencia del difono para el banco de difonos; los motivos se describen al final de la Tabla 1.

Para la grabación de esta base de datos se elaboró un corpus basado en una palabra para cada uno de los difonos, a su vez se colocaba esta palabra en la siguiente frase portadora: "Él dijo _____. Yo sé que él dijo _____." El difono se extrae de la palabra repetida en la segunda frase. Por la repetición de información implícita en la frase, se logra una monotonía y una intensidad equilibrada para cada difono del conjunto, características deseables en los difonos de la base de datos a la hora de lograr alta calidad en la síntesis.

Para la segmentación, se usa una de las herramientas propuestas por el grupo de MBROLA, llamada Diphone Studio. Un ejemplo de su uso se vió anteriormente en la Fig. 2. Los detalles de cómo debe segmentarse, sobre todo para casos especiales como las consonantes oclusivas y los grupos consonánticos, se describen en Rodríguez et al (2003). Una recomendación de ese trabajo es que el corpus debería

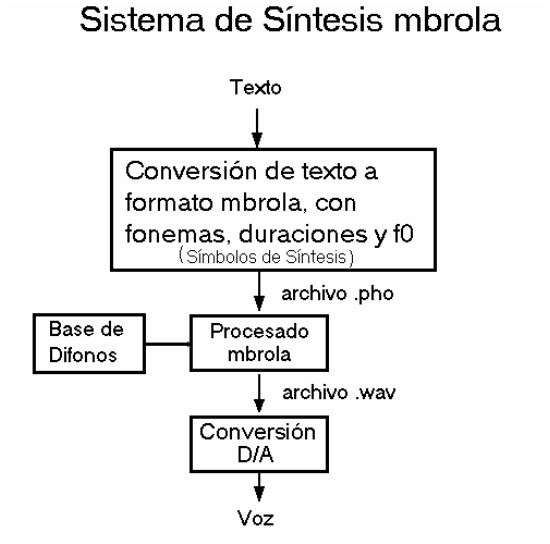


Fig. 3. Esquema de elementos del sintetizador de voz

contar con al menos 2 ejemplares de cada difono para permitir una selección, y evitar los inconvenientes de un difono defectuoso.

De esta forma se generó la base de datos para el español de Venezuela. Se conoce como vz1 y se puede bajar de Internet de la página web de MBROLA (tcts.fpms.ac.be/synthesis/**mbrola**.html). Consta de 794 difonos, y tienen una duración total de 120 segundos, es decir, una duración promedio de 150 ms para cada difono.

Para determinar la calidad de la base de datos, se ha sintetizado horas de voz con esta base de datos, incluyendo la síntesis de frases, párrafos y artículos completos, produciéndose una voz muy grata e inteligible, de una calidad aceptable.

3 Síntesis de voz con la técnica MBROLA

En la Fig. 3 se muestra el esquema del sintetizador de voz MBROLA. En el centro a la izquierda se muestra el bloque donde se almacena el banco de datos de difonos, el cual se describe anteriormente.

En la parte superior se muestra un bloque que se podría llamar el conversor ortográfico-fonético, que tiene como entrada un texto, y cuya salida es un archivo de parámetros que contiene la siguiente información: 1) La secuencia de fonemas, 2) la duración de cada uno de estos fonemas, y 3) los valores de entonación (f0) para algunos fonemas puntuales de la frase. Por ejemplo, la frase "Este atlas no es étnico.", se convierte en la siguiente secuencia de fonos: "_ e* s2 t e a* D l a s2 n o* e* s2 e* D n i k o _". En la Fig. 4 se muestra el archivo de parámetros correspondiente. La columna izquierda contiene la secuencia de fonemas a sintetizar. Cada fila a su vez contiene los datos acústicos de duración y entonación correspondientes a un fonema. Entonces

la segunda columna es la duración en milisegundos (ms) del fonema, y las columnas siguientes representan valores por pares para la frecuencia fundamental (f0) de vibración de las cuerdas vocales, precisando primero el % del intervalo del fonema donde se fija el segundo valor dado en Hz.

Entonces, para sintetizar una frase, hace falta armar este archivo de parámetros donde se incorpora toda la información necesaria y suficiente para las siguientes etapas. Se puede armar este archivo manualmente, a modo de prueba, con cualquier programa editor de texto. Con la información de este archivo, el programa MBROLA, correspondiente al bloque central de la Fig. 3 (**proceso MBROLA**) obtiene del banco de datos de difonos los segmentos requeridos, y además modifica tanto su duración como su entonación en función de los requerimientos de la frase, y los va concatenando, creando así un archivo de voz, hasta terminar la frase.

Finalmente el bloque de abajo de la Fig. 3 representa la reproducción de la señal por medio de una corneta de salida del computador.

Tabla 1. Cuadro parcial de difonos con sus excepciones

Primer Fono	_	a	a*	e	e*	i	i*	o	o*	u	u*	p	b
Segundo Fono													
a	1	3	6	67	68	$6	123	188	189	$6	244	568	301
a*	2	5	4	69	70	$6	124	190	191	$6	247	569	302
e	61	7	8	63	66	$6	127	192	193	$6	248	570	308
e*	62	9	10	65	64	$6	130	194	195	$6	251	571	679
i	121	$6	$6	$6	$6	131	133	$6	$6	$3	$3	572	312
i*	122	11	12	71	72	132	787	197	198	$3	$3	573	314
o	182	15	16	75	76	$6	134	184	187	$6	254	576	322
o*	183	17	18	77	78	$6	137	186	185	$6	257	577	323
u	242	$6	$6	$6	$6	$3	140	$6	$6	258	260	581	332
u*	243	19	20	80	81	$3	141	201	202	259	261	582	334
p	567	47	48	107	108	168	169	228	229	286	287	785	$2
b	300	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2
B	$2	23	24	83	84	142	143	204	205	262	263	786	$7
t	656	57	58	117	118	178	179	238	239	296	297	580	$2
d	361	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2
D	$2	29	30	89	90	148	149	210	211	268	269	$7	$7
k	336	25	26	85	86	144	145	206	207	264	265	$7	$2
g	417	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2	$2
G	$2	33	34	93	94	152	153	214	215	272	273	$7	$7
tS	350	27	28	87	88	146	147	208	209	266	267	$7	$7
f	401	31	32	91	92	150	151	212	213	270	271	$7	$7
s	620	53	54	113	114	174	175	234	235	292	293	579	$7
s2	$4	55	56	115	116	176	177	236	237	294	295	$4	$7
h	447	35	36	95	96	154	155	216	217	274	275	$7	$7
m	511	41	42	101	102	162	163	222	223	280	281	$7	$7
n	531	43	44	103	104	164	165	224	225	282	283	575	$7
J	556	45	46	105	106	166	167	226	227	284	285	$7	$7
l	473	37	38	97	98	158	159	218	219	276	277	574	317

Primer Fono	_	a	a*	e	e*	i	i*	o	o*	u	u*	p	b
Segundo Fono													
L	500	39	40	99	100	160	161	220	221	278	279	$7	$7
r	$5	49	50	109	110	170	171	230	231	288	289	578	327
rr	609	51	52	111	112	172	173	232	233	290	291	$7	$7
j	745	14	13	74	73	$7	$7	196	199	793	252	676	680
w	744	22	21	79	82	792	791	200	203	$7	$7	677	681
_	$7	59	60	119	120	180	181	240	241	298	299	678	$7

Explicación de celdas: XXX: número de línea de corpus

Razones justificadas por ausencia de difono

$1: VC, donde C es oclusiva sonora, pues en esta posición se utiliza la fricativa sonora

$2: oclusiva sonora solo en posición inicial o después de nasal (m o n)

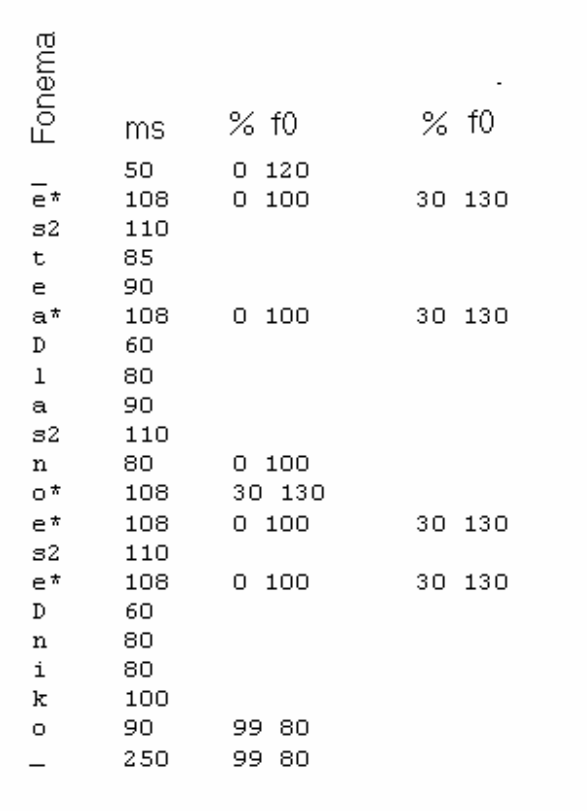
$3: dos vocales débiles en contacto crea diptongo

$4: contexto sC implica s en distensión se utiliza 's2', contexto Cs implica s en tension se utiliza 's'

$5: 'r' solo aparece en posición interna tras vocal

$6: vocal fuerte en contacto con vocal débil crea diptongo

$7: una secuencia de fonos aparentemente inexistente en el español de Venezuela



Fonema	ms	% f0	% f0
_	50	0 120	
e*	108	0 100	30 130
s2	110		
t	85		
e	90		
a*	108	0 100	30 130
D	60		
l	80		
a	90		
s2	110		
n	80	0 100	
o*	108	30 130	
e*	108	0 100	30 130
s2	110		
e*	108	0 100	30 130
D	60		
n	80		
i	80		
k	100		
o	90	99 80	
_	250	99 80	

Fig. 4. Un ejemplo de un archivo de parámetros, para la frase"Este atlas no es étnico."

En la Fig. 5 se muestra la forma de onda de la frase sintetizada "Este atlas no es étnico.", y abajo el espectrograma y la función de f0 correspondiente.

4 Conclusiones y recomendaciones

Se ha creado una base de datos de difonos, vz1, para el dialecto venezolano que, conjuntamente con un conjunto de herramientas disponibles públicamente y gratuitamente por el grupo de MBROLA, permite sintetizar voz en el dialecto venezolano del español. Se ha podido realizar gran cantidad de pruebas para establecer la calidad de los difonos.

El mayor inconveniente es que hay algunas secuencias de fonos que no es capaz de generar, por ejemplo la secuencia "p-rr" en el nombre del Papa actual, Josep Ratzinger, que no aparece en esta base de datos, puesto que se limitó a los sonidos del habla venezolana. Una futura versión de una base datos debería incluir los fonos que permitan pronunciar hasta nombres extranjeros. También en una futura versión, cuando se grabe la voz, se debe procurar tener al menos 2 ejemplares de cada difono para no verse obligado a utilizar el único ejemplar, con el riesgo que éste puede ser defectuoso, cosa que ocasionalmente ocurre con esta base de datos.

Un segundo artículo (Rodríguez y Mora, 2006) extiende la utilidad de esta base de datos con la implementación de un conversor ortográfico fonético automático,-

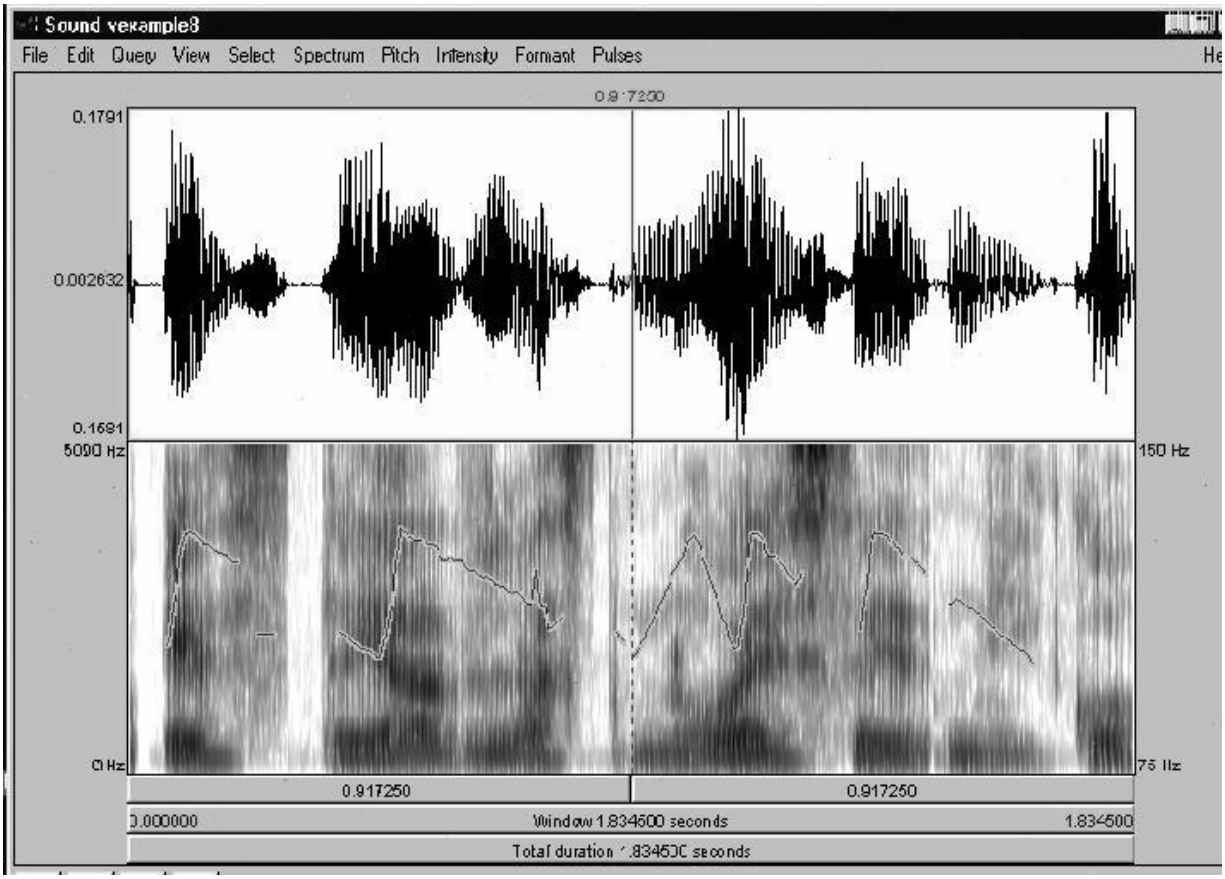


Fig. 5. Oscilograma de voz sintetizada de la frase "Este atlas no es étnico.", y abajo, espectrograma y función de f0

tico, permitiendo la creación así de un sistema de conversión de texto a voz automático.

Como trabajos futuros cercanos se piensa ampliar el atractivo del sistema al agregarle una voz femenina. Ya más a mediano plazo, se espera que pueda ser parte integral de una nueva línea de investigación de la prosodia del español hablado en Venezuela, permitiendo sintetizar frases con diferentes funciones de entonación, duración y ritmo para realizar pruebas de percepción.

Agradecimientos

Se agradece el financiamiento de este proyecto por parte de la agencia francesa ECOS-NORD (Action V99H01) y los institutos venezolanos CONICIT, CDCHT-ULA y Fundayacucho.

Referencias

Dutoit T, 1997, An introduction to text to speech synthesis, Dordrecht, Klewer.

Dutoit T, Pagel V, Pierret N, Bataille F y Van Der Vrecken O, 1996, The MBROLA project. Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. Proceedings ICSLP '96, Philadelphia, Vol. 3: pp. 1393-1396.

Huang X; Acero A y Hon H, 2001, Spoken language processing: a guide to theory, algorithm and system development, Pearson Edition.

Klatt D, 1980, Software for a cascade/parallel formant synthesizer, Journal of the Acoustical Society of America, JASA, Vol. 67, pp. 971-995.

Lee M, Lopresti D y Olive J, 2003, Speech platform for variable length optimal unit search using perception based cost funtions, Internatinal Journal of Speech Technologies, Vol. 6, No. 3.

Mora E, Hirst D y Cavé C, 2001, Développement et évaluation d'un système de synthèse pour l'espagnol vénézuélien: projet et état d'avancement, Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, Vol 19, pp. 91-98.

Moulines E y Charpentier F, 1990, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communications, Vol. 9: pp. 453-467.

Rabiner LR, Schafer RW y Flanagan JL, 1971, Computer synthesis of speech by concatenation of formant coded words, Bell Syst. Tech. Jour., Vol. 50, No. 5, May-Jun 1971, pp. 1541-1558.

Rodríguez M, Olabe, Santos, Muñoz P, Villaseca, Muñoz E y Martínez Q, 1984, Visión panorámica de la respuesta oral de máquinas, Mundo Electrónico, Nº 144, p. 57-66.

Rodríguez M, Iglesias E; Martínez R y Muñoz, E., 1984b Alternativas para síntesis de voz: aplicaciones de predicción lineal, Mundo Electrónico, Nº 144, p.p. 67-79.

Rodriguez M; Clairet S, Mora E; Cavé C y Hirst, D, 2003, Realización de una base de datos de difonos para el español hablado en Venezuela: Aplicación a la síntesis de voz TTS, Proceedings of VIII Simposio Internacional de Comunicación Social, p.p. 625-629.

Rodríguez M, Mora E, 2006, Conversor texto a voz en el dialecto venezolano por medio de la concatenación de difonos (por publicar).