



indisys:


Intelligent Dialogue Systems

DESARROLLO DE LAS VOCES SINTÉTICAS DE GUADALINEX

Created by: Guillermo Pérez	Reviewed by: Pilar Manchón	Accepted by
Date: 26/04/07	Date:	Date:
Signed:	Signed:	Signed:

Document History

Revision	Status	Date	Comment	Author
0.1	Draft	26/04/07	Creado	GPG, DMF, CSV
0.2	Draft	24/05/07	Modificado y revisado	PMP
0.3	Final	31/05/07	Revisado	PMP

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	2 (of 25)

ÍNDICE DE CONTENIDOS

1. INTRODUCCIÓN.....	5
1.1 Tecnología de síntesis.....	5
1.2 Motor de síntesis.....	5
2. INSTALACIÓN DE FESTIVAL.....	6
2.1 URL's de descarga.....	6
2.2 Preparación previa.....	7
2.3 Instalación de SPEECH TOOLS.....	7
2.4 Instalación de FESTVOX.....	7
2.5 Instalación de FESTIVAL.....	7
2.6 Instalación de voces.....	7
2.7 Prueba de la instalación.....	8
3. Grabación del corpus.....	9
3.1 Generación de un corpus de Difonemas.....	9
3.1.1 Introducción.....	9
3.1.2 Elección del conjunto de fonemas (o phoneset).....	9
3.1.3 Creación de la lista de difonemas.....	10
3.2 Elección de locutores.....	11
3.3 Grabación del corpus.....	11
3.4 Post-Procesado de las grabaciones.....	11
4. Generación de voces de difonemas.....	12
4.1 Proceso de generación de voz	12
4.1.1 Generación de voz.....	12
4.1.2 Actualización de la Addenda.....	12
4.1.3 Actualización de las letter-to-sound rules.....	12
4.1.4 Actualización de las token-to-word rules.....	12
4.1.4.1 Lectura de horas.....	13
4.1.4.2 Lectura de fechas.....	13
4.1.4.3 Lectura de mails.....	13
4.1.4.4 Lectura de líneas de símbolos.....	13
4.1.4.5 Lectura de URL's.....	14
4.1.4.6 Lectura de diferentes símbolos agrupados	14
4.1.4.7 Expansión de cadenas con caracteres/números y símbolos.....	14
4.1.5 Actualización del modelo de duración.....	15
4.1.6 Actualización del modelo F0.....	15
4.1.7 Actualización del Phrasing.....	15
4.2 Problemas encontrados y soluciones adoptadas.....	15
4.2.1 Logotomas de la nueva voz femenina.....	15
4.2.2 Picos de audio en frases sintetizadas.....	15
5. Aspectos generales.....	17
5.1 Codificación de caracteres.....	17
5.2 Uso en Guadalinux y a través de ORCA.....	17
6. Evaluación.....	18
6.1 Metodología.....	18
6.1.1 Selección de frases.....	18
6.1.2 Generación de archivos de audio.....	18
6.1.3 Criterios de evaluación.....	19

6.1.4 Proceso de evaluación.....	19
6.2 Resultados.....	19
7. MEJORAS SOBRE EL BASELINE (el_diphone).....	20
7.1 Fonemas en el Phoneset.....	20
7.1.1 Voz el_diphone.....	20
7.1.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	20
7.2 Locutores.....	20
7.2.1 Voz el_diphone.....	20
7.2.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	20
7.3 Entorno de grabación.....	20
7.3.1 Voz el_diphone.....	20
7.3.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	20
7.4 Addenda.....	21
7.4.1 Voz el_diphone.....	21
7.4.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	21
7.5 Letter-to-sound rules.....	21
7.5.1 Voz el_diphone.....	21
7.5.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	21
7.6 Modelo de duración.....	21
7.6.1 Voz el_diphone.....	21
7.6.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	21
7.7 Modelo de F0.....	21
7.7.1 Voz el_diphone.....	21
7.7.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	22
7.8 Token-to-word rules.....	22
7.8.1 Voz el_diphone.....	22
7.8.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	22
7.9 Símbolos especiales y de puntuación.....	22
7.9.1 Voz el_diphone.....	22
7.9.2 JuntaDeAndalucia_es_pa_diphone/es_sf_diphone.....	22
8. Bibliografía recomendada.....	24

1. INTRODUCCIÓN


1.1 Tecnología de síntesis

Las voces desarrolladas han sido diseñadas para un método de síntesis concatenativo, a partir de una base de datos de segmentos pregrabados por talentos de voz. La técnica de síntesis aplicada es la basada en difonemas.

1.2 Motor de síntesis

El motor de síntesis Festival fue elegido por los siguientes motivos:

1. Cumple el requisito de ser **código libre**, siendo distribuido con licencia tipo X-11.
2. Es un sintetizador consolidado en el mundo del código libre, estando ya integrado con las principales distribuciones de Linux, incluyendo *Ubuntu 6.10 "Edgy Eft"*.
3. Separa el nivel de ejecución del de especificación, estando el núcleo programado de manera robusta y eficiente en C/C++, y dejando así *Scheme* para el nivel de especificación.
4. Festival es una comunidad activa y en continuo desarrollo, lo que permite el aprovechamiento de posibles evoluciones de manera gratuita.
5. Festival es además la herramienta **disponible actualmente para Guadalinex**, siendo la voz actual en español manifiestamente mejorable.
6. No hay evidencias de que otros motores de síntesis que cumplan los requisitos anteriores proporcionen una mejor calidad final, o ventaja alguna con respecto a Festival.
7. Los miembros del equipo tienen amplia experiencia con Festival, incluyendo varios proyectos de desarrollo conjunto con la Universidad de Edimburgo.

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	5 (of 25)

2. INSTALACIÓN DE FESTIVAL

Antes de pasar a detallar el proceso de desarrollo de las voces sintéticas con *Festival*, resultará útil dedicar esta sección a la instalación del propio motor de síntesis Festival.

La versión de **Festival** utilizada para la creación de las voces sintéticas ha sido la llamada **1.96-beta** (es decir, la versión 2.0, de Agosto de 2006). La elección de esta versión ha venido motivada por las recomendaciones y garantías ofrecidas por algunos de los miembros del grupo de desarrollo de Festival (CSTR¹) en la Universidad de Edimburgo. La versión 1.96-beta es la versión más actualizada del sintetizador.

Para la instalación del motor de síntesis Festival es necesario utilizar una serie de herramientas externas. Las más importantes son **FestVox** (en su versión 2.0) y **SpeechTools** (versión 1.2.96-beta). El objetivo del proyecto FestVox (con sede en la *Carnegie Mellon University*) es ofrecer toda la versatilidad posible para el desarrollo de voces con Festival (y ello, mediante *scripts*, documentación, bases de datos, etc.). Por otro lado, la librería SpeechTools (desarrolladas también por el CSRT), una colección de clases en C++ y funciones, ofrece varios programas para el tratamiento de los objetos más frecuentemente utilizados en el procesamiento del habla.

2.1 URL's de descarga

Los paquetes que hay que descargar obligatoriamente para poder instalar *Festival* (con la voz española *el_diphone*²) son los siguientes:

Nombre	URL	Descripción
festival-1.96-beta.tar.gz	http://www.cstr.ed.ac.uk/downloads/festival/	Distribución del código fuente de Festival
festvox-2.0-release.tar.gz	http://festvox.org/download.html	Distribución del código fuente de Festvox
speech_tools-1.2.96-beta.tar.gz	http://www.cstr.ed.ac.uk/downloads/festival/	Herramientas para Festival (Edimburgo)
multisyn_build-1.8.tgz	http://www.cstr.ed.ac.uk/downloads/festival/	Entorno para usar el engine MultiSyn
festvox_ellpc11k.tar.gz	http://www.cstr.ed.ac.uk/downloads/festival/	Voz española <i>el_diphone</i>

A continuación se indica cómo instalar tanto Festival como las herramientas especificadas anteriormente. Es importante saber que el orden en que se instalan cada uno de los paquetes es determinante. Por tanto, es conveniente no alterar los pasos enumerados en las siguientes secciones.

¹ The Center for Speech Technology Research

² La voz *el_diphone* ha sido necesaria para el desarrollo de las nuevas voces sintéticas en español.

2.2 Preparación previa

En primer lugar, es necesario instalar varias librerías específicas. Los nombres de estas librerías son **Lcurses**, **gcc** y **g++**. Es recomendable utilizar el gestor de paquetes **Synaptic** para la instalación de dichas librerías.

Por otro lado, en el disco local, se debe crear un directorio llamado `festival-1.96`. En él (a partir de ahora, directorio base) se instalarán todas las herramientas necesarias (además del propio motor de síntesis Festival).

2.3 Instalación de SPEECH TOOLS

En segundo lugar, se procederá a instalar las herramientas *SpeechTools*. Para ello, hay que extraer el paquete “**speech_tools-1.2.96-beta.tar.gz**” sobre el directorio que se ha creado en la sección anterior: el directorio base, es decir `/festival-1.96`.

La instalación del paquete se realiza mediante la ejecución de los siguientes comandos desde el terminal:

```
> cd festival-1.96/speech_tools
festival-1.96/speech_tools> ./configure
festival-1.96/speech_tools> make
```

2.4 Instalación de FESTVOX

A continuación, se instalará el paquete *FestVox*. De igual forma que en el paso anterior, se ha de extraer el paquete “**festvox-2.0-release.tar.gz**” sobre el directorio base: `/festival-1.96`.

La instalación se debe realizar mediante la ejecución de los comandos:

```
> cd festival-1.96/festvox
festival-1.96/festvox> ./configure
festival-1.96/festvox> make
```

2.5 Instalación de FESTIVAL


El siguiente paso viene dado por la instalación de *Festival*. Se extrae el código fuente de Festival (contenido en el paquete “**festival-1.96-beta.tar.gz**”) sobre el directorio `/festival-1.96`.

Después, se instala el paquete mediante las siguientes instrucciones:

```
> cd festival-1.96/festival
festival-1.96/festival> ./configure
festival-1.96/festival> make
```

2.6 Instalación de voces

Finalmente, se han de instalar las voces que sea necesario utilizar, mediante las descompresión del paquete “**festvox_ellpc11k.tar.gz**” sobre el directorio base `/festival-1.96`.


	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	7 (of 25)

2.7 Prueba de la instalación

Para comprobar que todo ha sido instalado correctamente se pueden ejecutar los siguientes comandos:

```
1 > cd festival-1.96/festival/bin
2 festival-1.96/festival/bin> ./festival
3 festival$ (voice_el_diphone)
4 festival$ (SayText "hola mundo")
```

En este ejemplo, en primer lugar (línea 2), se ha puesto en funcionamiento el motor de síntesis Festival (en su versión 1.96). Después, en la línea 3, se ha cargado la voz en español *el_diphone*. Finalmente, en la línea 4, mediante (*SayText* " "), se ha escrito un texto que será sintetizado por la voz *el_diphone*.

	Document title			Author		Page
	Voces en Guadalinex			GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status		
		0.3	31/05/2007	Validado		8 (of 25)

3. GRABACIÓN DEL CORPUS

3.1 Generación de un corpus de Difonemas

3.1.1 Introducción

El objetivo de la creación de un corpus de difonemas es tener un listado completo de todas las posibles transiciones entre los fonemas de un idioma.

Por lo general, el número de difonemas de un idioma es, aproximadamente, el cuadrado del total de los fonemas de ese mismo idioma. “Aproximadamente” significa que el conjunto de fonemas del que es necesario partir puede verse afectado por la inclusión de algunos alófonos (como la /B/ fricativa además de la /b/ oclusiva) e, incluso, de fonemas de otros idiomas (sobre todo, del inglés, debido a la gran cantidad de anglicismos que existen).

El corpus de difonemas se utiliza en el método de síntesis *UniSyn*.

3.1.2 Elección del conjunto de fonemas (o *phoneset*)


El conjunto de fonemas para las nuevas voces está compuesto por:

a) el total de fonemas del español

/p/	consonante, bilabial oclusiva sorda
/b/	consonante, bilabial oclusiva sonora
/t/	consonante, dental oclusiva sorda
/d/	consonante, dental oclusiva sonora
/k/	consonante, velar oclusiva sorda
/g/	consonante, velar oclusiva sonora
/f/	consonante, labiodental fricativa sorda
/th/	consonante, interdental fricativa sorda
/s/	consonante, alveolar fricativa sorda
/x/	consonante, velar fricativa sorda
/ch/	consonante, palatal africada sorda
/m/	consonante, bilabial nasal sonora
/n/	consonante, alveolar nasal sonora
/ny/	consonante, palatal nasal sonora
/l/	consonante, alveolar lateral sonora
/ll/	consonante, palatal lateral sonora
/r/	consonante, alveolar vibrante simple sonora
/rr/	consonante, alveolar vibrante múltiple sonora
/a/	vocal, central baja (sin acentuar)
/e/	vocal, anterior media (sin acentuar)
/i/	vocal, anterior alta (sin acentuar)
/o/	vocal, posterior media (sin acentuar)
/u/	vocal, posterior alta (sin acentuar)

b) algunas variantes alofónicas del español

/B/	consonante, bilabial fricativa sonora
/D/	consonante, interdental fricativa sonora
/G/	consonante, velar fricativa sonora
/a1/	vocal, central baja (acentuada)
/e1/	vocal, anterior media (acentuada)
/i1/	vocal, anterior alta (acentuada)
/o1/	vocal, posterior media (acentuada)
/u1/	vocal, posterior alta (acentuada)

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	9 (of 25)

c) ciertos fonemas del inglés

/hh/	consonante, glotal fricativa sorda (ej.: hardware)
/dh/	consonante, dental fricativa sonora (ej.: the gimp)
/zh/	consonante, post-alveolar africada sonora (ej.: digicam)
sh/	consonante, post-alveolar fricativa sorda (ej.: sawfish)
/z/	consonante, alveolar fricativa sonora (ej.: jaws)
/v/	consonante, labiodental fricativa sonora (ej.: evolution)
/ax/	vocal reducida (ej.: the)

Es importante saber que, a la hora de la síntesis, tanto los fonemas básicos del español (set A) como las variedades alofónicas que se han añadido (set B), pueden ser generados mediante las *letter-to-sound rules* definidas en el fichero de configuración "INST_es_VOX_lexicon.scm"¹. Sin embargo, los fonemas del inglés utilizados (set C) únicamente aparecerán cuando se decida sintetizar una palabra que se encuentre en la *addenda* definida (léxico de pequeñas dimensiones que contiene palabras del dominio informático, acrónimos, abreviaturas, etc.) y cuya transcripción fonética contenga dichos fonemas.

3.1.3 Creación de la lista de difonemas

a) Logotomas (también, *nonsense words*)

Un logotoma es una palabra sin significado, constituida, normalmente, por tres sílabas, lo cual permite que el difonema a tratar esté aislado, en la sílaba central.

La elaboración de un corpus de logotomas (frente a la elaboración de un corpus con palabras reales) tiene varias ventajas: asegura la cobertura de todos los difonemas posibles para un idioma (incluso la de aquellos que son extremadamente infrecuentes), permite consistencia en cuanto a la tonalidad y la duración de los difonemas. Además, los logotomas, por ser cadenas de sonidos, no dan lugar a duda sobre cómo deben ser pronunciados.

b) *Carriers*


La metodología básica detrás de la generación de un corpus de logotomas se centra en dos necesidades: por un lado, la definición de las clases de difonemas; por otro, la definición de los *carrier contexts* que rodearán a estos difonemas.

La idea es que el difonema aparezca en la sílaba central del logotoma, para que pueda ser debidamente articulado (y así minimizar los efectos que sobre la articulación tienen el comienzo y el final de la palabra). Aun así, debe recordarse que hay difonemas específicos del tipo consonante-silencio (C#), silencio-consonante (#C), vocal-silencio (V#), y silencio-vocal (#V), cuya posición en el logotoma nunca ocupa la sílaba central.

c) Delimitación silábica

La delimitación silábica permite establecer una distinción explícita entre los difonemas de la clase consonante-consonante (CC) y los difonemas conocidos como *consonant clusters* (también CC). Un difonema del tipo CC (ej.: /k-l/) aparece en un logotoma como "# t a **k** - l a l t a #". Además, el *grupo consonántico* /k_-l/ aparece en el logotoma "# t a - **k** l a l t a #".

¹festvox/src/vox_files/es/INST_es_VOX_lexicon.scm

	Document title			Author		Page
	Voces en Guadalinex			GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status		
		0.3	31/05/2007	Validado		10 (of 25)

La diferencia entre ambos difonemas reside en el hecho de que /k-l/ aparece en un contexto intersilábico (entre dos sílabas), mientras que /k_-l/ aparece en un contexto intrasilábico (dentro de una misma sílaba). La articulación de la /l/ en el contexto /k-l/ (intersilábico) no es igual que la articulación de la /l/ producida en el contexto /k_-l/ (intrasilábico). Esta diferenciación ha sido tenida en cuenta en la elaboración de nuestro corpus (cf. Black & Lenzo, 2007).

d) Generación de la lista de difonemas

La lista de difonemas se genera utilizando las herramientas disponibles en Festvox; no sin antes haber definido las clases de difonemas, los *carrier contexts* para cada una de estas clases, y las excepciones y peculiaridades relativas a cada uno de los fonemas del *phoneset* definido.

3.2 Elección de locutores

Los locutores fueron elegidos mediante un proceso de selección exhaustivo. Tras una criba inicial, se seleccionaron 4 locutores masculinos y 4 femeninos. Después, a partir de una grabación de prueba, los locutores finales fueron seleccionados por la Junta de Andalucía.

3.3 Grabación del corpus


La grabación del corpus de difonemas se realizó en un estudio de grabación profesional.

La señal grabada procedía de dos fuentes principales: el audio (grabado mediante un micrófono *Neumann* con su filtro POP) y la señal del electroglotógrafo. La salida de audio analógica del electroglotógrafo fue conectada al centro de control del estudio.

3.4 Post-Procesado de las grabaciones

El *post*-procesado aplicado sobre los ficheros de audio de ambos locutores consistió en dos acciones fundamentales:

- a) **Compresión** leve sobre las señales de audio para disminuir el rango dinámico de la señal y aumentar más la ganancia (sin llegar a saturar).
- b) **Normalización** de las señales de audio para que todas tuviesen la misma intensidad.

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	11 (of 25)
	0.3	31/05/2007	Validado		

4. GENERACIÓN DE VOCES DE DIFONEMAS

4.1 Proceso de generación de voz

4.1.1 Generación de voz

Éste es el proceso mediante el cual se genera la voz base a partir de la que se obtendrá la versión final de la misma. Para ello, se utilizan todos los archivos fuente instalados anteriormente, así como las herramientas de Festvox.

4.1.2 Actualización de la *Addenda*

La primera mejora de las voces desarrolladas con respecto a los ficheros de configuración de la voz *el_diphone* es la inclusión de una *addenda*. Por “addenda” se entiende un conjunto de palabras (y sus transcripciones) que quedan fuera del léxico compilado. Este conjunto de palabras, por tanto, puede ser modificado dinámicamente (cf. Black & Lenzo, 2007).

El motor de síntesis comprueba primero que las palabras a sintetizar se encuentren en el léxico compilado. Si no se encuentran ahí (o bien, sencillamente, no existe ningún léxico compilado, como es el caso generalizado de las voces en español), el siguiente lugar donde busca dichas palabras es la *addenda*. Si éstas aparecen en la *addenda*, entonces utiliza las transcripciones que tengan asignadas.

La *addenda* desarrollada para las nuevas voces incluye palabras del dominio específico de Guadalinux (aplicaciones, términos frecuentes en los menús, etc.). También contiene términos del dominio de la informática en general (abreviaturas, acrónimos, anglicismos, etc.).

4.1.3 Actualización de las *letter-to-sound rules*


Las *letter-to-sound (LTS) rules* son reglas fonéticas dependientes del contexto. Cuando las palabras a sintetizar no se encuentran ni en el léxico compilado ni en la *addenda* definida, Festival genera una transcripción fonética (a partir de estas reglas) con la que construye la *utterance* correspondiente (cf. Black et al., 2002).

El conjunto de reglas fonéticas incluido en las nuevas voces está compuesto por una amplia base procedente de las reglas de la voz para *el_diphone*, así como por un set de reglas añadidas (para el tratamiento de nuevos fenómenos observados en las voces desarrolladas). Las reglas añadidas establecen la diferencia entre los alófonos “b” (explosiva) y “B” (fricativa), “d” (explosiva) y “D” (fricativa), y “g” (explosiva) y “G” (fricativa).

4.1.4 Actualización de las *token-to-word rules*

Se han modificado las reglas *token-to-words* originales (cf. Black et al., 2002) para así permitir la lectura de horas, fechas, correos electrónicos, URL's, caracteres más símbolos, líneas de símbolos, etc. En principio, las reglas originales sólo permitían expandir los números, es decir, transformaban una entrada como “1234” al texto (hablado) “mil doscientos treinta y cuatro”, no a “uno dos tres cuatro”; sin embargo, no eran capaces de tratar entradas como “fernando.alonso@maclaren.com” o “http://www.yahoo.es/grupos”.

A continuación, se detalla la funcionalidad de las *token-to-words* agregadas:

	Document title		Author		Page
	Voces en Guadalinux		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	12 (of 25)

4.1.4.1 Lectura de horas

Se pueden leer tres tipos de formato.

- a) Hora **con am/pm** y con “:” o “.” como separador

Ejemplos:

21:33 am --> (“veintiuna” “treintaitrés” “a” “m”)

09.22 am --> (“nueve” “veintidós” “a” “m”)

- b) Hora **sin am/pm**

Ejemplo: 21:33 --> (“veintiuna” “treintaitrés”)

- c) Hora **exacta**

Ejemplo: 21:33:22 --> (“veintiuna” “horas” “treintaitrés” “minutos” “veintidos” “segundos”).

4.1.4.2 Lectura de fechas

Se permite la lectura de fechas escritas con la siguiente estructura:
día/mes/año.

Ejemplo: 15/2/2007 --> (“quince” “de” “febrero” “del” “dos” “mil” “siete”)

4.1.4.3 Lectura de mails

Se permite la lectura de *mails* con dos formatos distintos.

En los dos casos, aquellas cadenas que contengan caracteres, símbolos y números serán expandidas de la siguiente forma:

prueba_simbolo --> (“prueba” “guión bajo” “símbolo”)

prueba_1234 --> (“prueba” “guión bajo” “mil” “doscientos” “treintaicuatro”)

prueba_sil234 --> (“prueba” “guión bajo” “ese” “i” “mil” “doscientos” “treintaicuatro”)

Para la lectura de correos electrónicos, se expandirán todos los símbolos identificados en la siguiente lista:

- Símbolos que se dirán: *%&\$€@#+=/_><[]{}\'\".,;:;!¿?-ao()


Los formatos de correo electrónico que pasarán por esta regla son:

- a) **Mails citados textualmente**

Ejemplo: <prueba@prueba.com> --> (“prueba” “arroba” “prueba” “punto” “com”)

- b) **Mails normales**

Ejemplo: prueba@prueba.com --> (“prueba” “arroba” “prueba” “punto” “com”)

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	13 (of 25)
	0.3	31/05/2007	Validado		

4.1.4.4 Lectura de líneas de símbolos

Se permite la lectura simplificada de líneas de símbolos.

- a) **Línea de guiones:** si se intenta sintetizar una línea con más de cinco guiones bajos la locución será ("línea" "de" "guiones" "bajos").
- b) **Línea de iguales:** si se intenta sintetizar una línea con más de cinco símbolos = se sintetizará ("línea" "de" "iguales")
- c) **Línea de guiones medios:** cinco ó más guiones medios producirá la salida ("línea" "de" "guiones" "medios")
- d) **Línea de asteriscos:** cinco ó mas asteriscos producirán la salida ("línea" "de" "asteriscos").

4.1.4.5 Lectura de URL's

Se permite la lectura de URL's. Se expanden las cadenas que contengan símbolos, tal y como se indica en el apartado "Lectura de mails".

- Símbolos que se expandirán: *%&\$€@#+-~=\/_|><[]{}\"'`.,;:;!¿? -a0()

Se tratarán dos formatos de URL's:

- a) **Formato http:** [Ejemplo: `http://www.prueba.com`]

Todas aquellas cadenas que comiencen con "**http://**" accederán a esta regla.

Ejemplo: `http://www.prueba_123.com/` --> ("hache" "te" "te" "pe" "barra" "barra" "uve doble" "uve doble" "uve doble" "punto" "prueba" "guión bajo" "ciento" "veinti" "tres" "punto" "com" "barra").

- b) **Formato www:** [Ejemplo: `www.prueba.com`]

Aquellas cadenas que comiencen por "**www.**" se considerarán dentro de esta regla.

Ejemplo: `www.prueba_123.com/` --> ("uve doble" "uve doble" "uve doble" "punto" "prueba" "guión bajo" "ciento" "veinti" "tres" "punto" "com" "barra").

4.1.4.6 Lectura de diferentes símbolos agrupados


En caso de sintetizar una cadena compuesta por uno ó mas símbolos (sin caracteres ni números a su alrededor), cada uno de aquellos será expandido según la lista siguiente:

- Símbolos a decir: *%&\$€@#+-~=\/_|><[]{}\"'`.,;:;!¿? -a0()

Ejemplo: `*_/` --> ("asterisco" "guión bajo" "barra")

4.1.4.7 Expansión de cadenas con caracteres/números y símbolos

En caso de recibir cadenas de texto compuestas por caracteres (o números) junto con algún símbolo, se tomará dicho símbolo como punto de referencia y se expandirán cada una de las cadenas de caracteres existentes a la izquierda y a la derecha de este símbolo de referencia.

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	14 (of 25)
	0.3	31/05/2007	Validado		

- Símbolos que no se dirán: " ` .,:;!¿? -ªº[]{}()"
- Símbolos a decir: *%&\$€@#+=\/_|><

Ejemplo:

```
> SayText("Esto es una prueba. El 10% de la empresa T+T está vendida")
("esto" "es" "una" "prueba") ("el" "diez" "porciento" "de" "la" "empresa" "te"
"más" "te" "está" "vendida")
```

4.1.5 Actualización del modelo de duración

Este modelo es el encargado de asignar valores de duración a cada uno de los segmentos del *phoneset* que aparecen en una locución.

El modelo de duración desarrollado para las nuevas voces masculina y femenina es un modelo basado en entranamiento, el cual utiliza un *CART tree* que contiene información de los factores *ZScores* para un segmento determinado (en función de su contexto).

4.1.6 Actualización del modelo F0

El modelo de contorno F0 añadido en las voces desarrolladas, generado también en forma de *CART Tree*, contiene valores de frecuencia fundamental para el fonema inicial y final de cada sílaba, así como para las vocales que constituyen el núcleo de las mismas (en función del contexto en el que aparecen: tanto dentro de la sílaba como dentro de la locución).

4.1.7 Actualización del *Phrasing*

La función de *phrasing* desarrollada para las nuevas voces establece una distinción clara entre *Break* (B), *Big Break* (BB) y *No Break* (NB). Se caracteriza, fundamentalmente, por dos peculiaridades:

- a) el modelo que define se conoce como un simple modelo de predicción ***CART tree***, y además,
- b) elimina el *bug* existente en la voz *el_diphone*: ahora dos marcas de puntuación pueden aparecer de forma contigua (antes se anulaban entre sí).

4.2 Problemas encontrados y soluciones adoptadas


4.2.1 Logotomas de la nueva voz femenina

Las vocales producidas por la locutora femenina tienen una cualidad demasiado específica. En concreto, las vocales "o" y "u" (especialmente en sílabas átonas) son demasiado cortas.


Tras un análisis exhaustivo y varias pruebas fallidas, el problema fue finalmente solventado con éxito.

4.2.2 Picos de audio en frases sintetizadas

Algunos picos de audio han aparecido en ciertas locuciones, sobre todo en palabras que contienen combinaciones del segmento /s/ con otras consonantes y/o vocales.

	Document title Voces en Guadalinex		Author GPG, DMF, CSV, PMP		Page 15 (of 25)
	Document number	Revision 0.3	Date 31/05/2007	Status Validado	

Se han llevado a cabo cambios relativos a distintas líneas de investigación, los cuales han reducido este problema significativamente; sin embargo, estas mejoras no han alcanzado un resultado óptimo.

	Document title			Author		Page
	Voces en Guadalinex			GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status		
		0.3	31/05/2007	Validado		16 (of 25)

5. ASPECTOS GENERALES

5.1 Codificación de caracteres

FESTIVAL sólo acepta ficheros de configuración en **ASCII** o en **ISO-8859-1**. Este hecho es muy importante, ya que la plataforma de desarrollo *Ubuntu* 6.10 “Edgy Eft” utiliza por defecto *Unicode*. Para solventar estos problemas pueden cambiarse los esquemas de caracteres tanto del terminal (*gnome-terminal*) como del editor (*gnome-editor*).

Para sintetizar voz con FESTIVAL es necesario cambiar la codificación de la terminal, o si la síntesis se realiza a través de un archivo, éste debe estar en ISO-8859-1.


El uso de otras codificaciones puede dar lugar a errores de pronunciación y otros problemas. Esta limitación no proviene de las nuevas voces, sino del motor de síntesis.

5.2 Uso en Guadalinex y a través de ORCA

Orca es la aplicación que hace de interfaz entre el escritorio y el sintetizador de voz. Se encarga de recolectar los elementos de texto que aparecen en pantalla, procesarlos y mandarlos a los distintos elementos que facilitan la accesibilidad, como un teclado *braille* o el propio sintetizador de voz.

Para usar las voces en Guadalinex, basta con instalar los paquetes .DEB y seleccionar en ORCA, la voz deseada:

- JuntaDeAndalucia_es_pa_diphone --> voz masculina
- JuntaDeAndalucia_es_sf_diphone --> voz femenina

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	17 (of 25)
	0.3	31/05/2007	Validado		

6. EVALUACIÓN

6.1 Metodología

6.1.1 Selección de frases

La selección de las frases para la evaluación se hizo de forma que el conjunto total de las mismas agrupase locuciones de dominios diferentes (genérico, numérico, informático, de Guadalinx, y con direcciones de correo electrónico y URLs). Esta distribución dio lugar a un conjunto de 12 frases:

Introducción (genérico):

1. Hola, soy una de las voces sintéticas que participan en la siguiente evaluación.
2. El objetivo es evaluar la calidad relativa de cuatro voces sintéticas diferentes.

Numérico:

3. 7, 8, 9, 10, 11, 12, 1567, 35142
4. 907 34 56 78

Informático:

5. Antes de acceder a una red **wireless** comprobaremos que tenemos el **firewall** activado.
6. Las páginas **web** se desarrollan en **HTML**, un lenguaje de marcación para navegadores como Internet **Explorer** o **Mozilla Firefox**.

Guadalinx:

7. Bienvenido al centro de ayuda del gestor de paquetes **Synaptic**.
8. Pulse aquí para ocultar todas las ventanas y mostrar el escritorio.

Genérico (aleatorio):

9. Aguirre quería que yo fuese a las reuniones.
10. Cuando tuve mi primera cinta de vídeo sentí algo muy extraño.


Correo electrónico - URL:

11. fernando.alonso@maclaren.com
12. http://www.yahoo.es/grupos

- Las frases **1-6**, **11** y **12** fueron generadas con objeto de cubrir varios dominios;
- las frases **7** y **8** fueron extraídas del entorno del escritorio de Guadalinx;
- las frases **9** y **10** fueron seleccionadas aleatoriamente de entre 202 frases de un dominio genérico.

6.1.2 Generación de archivos de audio

Las voces sintéticas que participaron en la evaluación fueron 4 (todas ellas españolas):

	Document title		Author		Page
	Voces en Guadalinx		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	18 (of 25)

1. *voice_JuntaDeAndalucia_es_pa_diphone* (masculina para Festival)
2. *voice_JuntaDeAndalucia_es_sf_diphone* (femenina para Festival)
3. *voice_el_diphone* (antigua voz masculina para Festival)
4. *eSpeak* (masculina para eSpeak)

La voz *el_diphone* sirve como *baseline* para el análisis de los resultados. La voz de eSpeak es la voz española para el sintetizador (también de libre distribución) eSpeak. Se trata de una voz masculina, desarrollada para un motor de síntesis basado en la detección de formantes. Actualmente, eSpeak ha pasado a ser el sintetizador incluido por defecto en la distribución de *Ubuntu* "Feisty Fawn".

Los sujetos escucharon las cuatro versiones de cada una de las frases de manera secuencial, evaluando las características solicitadas en cada caso. El orden de presentación de las versiones fue determinado de manera aleatoria para evitar posibles influencias positivas o negativas.

6.1.3 Criterios de evaluación

Los criterios de evaluación acordados pueden resumirse en los dos puntos siguientes:

1. Seleccionar la versión más NATURAL.
2. Seleccionar la versión mas INTELIGIBLE.
3. Puntuar cada una de las voces (entre 0 y 10) tras cada frase.

Las hojas de evaluación constaron de 12 tablas como la que aparece a continuación:

Frase 1	1ª voz	2ª voz	3ª voz	4ª voz
Naturalidad				
Inteligibilidad				
Puntuación				

6.1.4 Proceso de evaluación

Nº de Sujetos	10
Nº de Mujeres	6
Nº de Varones	4
Rango de Edad	21-33
Duración	10-15 minutos

6.2 Resultados Generales

VOTOS RECIBIDOS POR CADA VOZ PARA LA NATURALIDAD		%
PA		50,83
SF		23,33
EL		22,5
Espeak		3,33

VOTOS RECIBIDOS POR CADA VOZ PARA LA INTELIGIBILIDAD		%
PA		58,33
SF		25
EL		14,17
Espeak		2,5

PUNTUACION MEDIA POR VOZ		
PA		6,23
SF		5,45
EL		4,21
Espeak		2,56

Las nuevas voces desarrolladas para Festival son, holgadamente, las preferidas en cuanto a naturalidad, inteligibilidad y calidad general. La voz masculina ha sido evaluada como la mejor con diferencia. Las voces restantes (Espeak y Eduardo López) no han llegado al aprobado. Sin embargo las voces desarrolladas en el presente proyecto, no sólo aprueban en general sino que son altamente valoradas por algunos usuarios.

En un 99% de los casos los usuarios han seleccionado la nueva voz masculina como la mejor, seguida de la nueva voz femenina.

7. MEJORAS SOBRE EL BASELINE (*EL_DIPHONE*)

7.1 Fonemas en el *Phoneset*

7.1.1 Voz *el_diphone*

El conjunto de fonemas de la voz *el_diphone* no incluye variaciones alofónicas ni tampoco contrastes del tipo vocales tónicas frente a vocales átonas. El total de difonemas de *el_diphone* es **662**.

7.1.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*

El conjunto de fonemas de las nuevas voces masculina y femenina contiene los alófonos fricativos de los fonemas explosivos /b/, /d/ y /g/: es decir, /B/, /D/ y /G/. Además, incluye referencias explícitas en el fichero de corpus a los grupos consonánticos intrasilábicos “p_-l”, “b_-l”, “k_-l”, “g_-l”, “p_-r”, “b_-r”, “k_-r”, “g_-r”, “f_-l” y “f_-r”.

Por último, dado que el alcance principal de las voces generadas sería el entorno de Guadalínex, donde la mayoría de las aplicaciones tienen nombres que proceden del inglés, las nuevas voces contienen varios fonemas pertenecientes al *Phoneset* de este idioma: /hh/, /dh/, /zh/, /z/, /v/ y /ax/.

El total de difonemas tanto de *JuntaDeAndalucia_es_pa_diphone* como de *JuntaDeAndalucia_es_sf_diphone* es de **1090**.

7.2 Locutores

7.2.1 Voz *el_diphone*

Los fonemas que componen la voz *el_diphone* fueron grabados por Eduardo López, un estudiante de Doctorado.

7.2.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*

Los logotomas para las nuevas voces desarrolladas han sido grabados por dos locutores profesionales: Pedro Alonso para la voz masculina y Silvia Fernández para la voz femenina (ver sección “Grabación del corpus”).


7.3 Entorno de grabación

7.3.1 Voz *el_diphone*

Los logotomas para la voz *el_diphone* fueron grabados en un entorno no especializado.

7.3.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*

Las grabaciones de las nuevas voces se hicieron en un estudio de grabación profesional (ver sección “Grabación del corpus”), con material de grabación también especializado (sala, micrófonos, EGG, etc.).

	Document title		Author		Page
	Voces en Guadalínex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	21 (of 25)
	0.3	31/05/2007	Validado		

7.4 Addenda

7.4.1 Voz *el_diphone*

La *addenda* de la voz *el_diphone* sólo contiene determinados símbolos y caracteres especiales.

7.4.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*

La *addenda* para las dos voces desarrolladas contiene hasta **418** términos específicos (ver sección “Actualización de la *Addenda*”).

7.5 Letter-to-sound rules

7.5.1 Voz *el_diphone*

Las *letter-to-sound* rules de la voz *el_diphone* son bastante completas. La mayor parte de ellas han servido como base para el conjunto de *letter-to-sound* rules de las nuevas voces masculina y femenina.

7.5.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*

Las *letter-to-sound* rules de las voces *JuntaDeAndalucia_es_pa_diphone* y *JuntaDeAndalucia_es_sf_diphone* incluyen algunas reglas específicas para el tratamiento de los segmentos alofónicos /B/, /D/ y /G/ (ver sección “Actualización de las *letter-to-sound* rules”).

7.6 Modelo de duración

7.6.1 Voz *el_diphone*

La duración de cada uno de los fonemas en *el_diphone* viene determinada en una función llamada *spanish_el_phone_data*. La particularidad de estas duraciones es que sólo contienen información para el valor que señala la desviación estándar. Es decir, esta función no incluye información a cerca de la duración media de los fonemas.


7.6.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*

Las duraciones de los fonemas de las nuevas voces masculina y femenina han sido calculadas mediante entrenamiento (con resultados estocásticos) a partir de un conjunto de frases (1006, en concreto) grabadas por los dos locutores profesionales involucrados en este proyecto. Las duraciones medias calculadas (así como los valores de la desviación estándar) se completan con un modelo de duración del tipo *CART tree* (ver sección “Actualización del modelo de duración”).

7.7 Modelo de F0

7.7.1 Voz *el_diphone*

El modelo de frecuencia fundamental de la voz *el_diphone* es un modelo sencillo, basado en reglas.

	Document title		Author		Page
	Voces en Guadalínx		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	22 (of 25)

7.7.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*

El modelo de F0 de las voces *JuntaDeAndalucia_es_pa_diphone* y *JuntaDeAndalucia_es_sf_diphone* es un modelo entrenado estadísticamente a partir del mismo conjunto anterior de 1006 frases, grabadas por los locutores.

El modelo de contorno F0 asigna un valor de frecuencia fundamental para el fonema inicial y final de cada sílaba, así como para la vocal que se encuentra en mitad de una sílaba. El contenido se almacena también en un *CART tree*, con información del valor de frecuencia fundamental para dichos fonemas, en función del contexto de en que se encuentren (ver sección “Actualización del modelo F0”).

El resultado es apreciable en la entonación con que se realizan las frases que se sintetizan, puesto que resultan mucho más naturales.

7.8 *Token-to-word rules*

7.8.1 *Voz el_diphone*

Las *token-to-word rules* de la voz *el_diphone* sólo permiten la lectura de números. Ante una cadena con caracteres especiales, esta voz deletrea todos y cada uno de los elementos que componen dicha cadena.

7.8.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*

Las nuevas voces tanto masculina como femenina incluyen un conjunto más extenso de *token-to-word rules* (ver sección “Actualización de las *token-to-word rules*”). Esta extensión facilita la lectura de horas, URL's, direcciones de correo electrónico, líneas de símbolos, etc.

7.9 *Símbolos especiales y de puntuación*

7.9.1 *Voz el_diphone*


La voz *el_diphone* realiza la síntesis de determinados símbolos especiales, pero en condiciones bastante reducidas.

7.9.2 *JuntaDeAndalucia_es_pa_diphone/es_sf_diphone*


Las voces *JuntaDeAndalucia_es_pa_diphone* y *JuntaDeAndalucia_es_sf_diphone* sintetizan símbolos especiales tanto si éstos aparecen dentro de una cadena de texto (por ejemplo, “escritorio/documentos”) como si aparecen de forma aislada (por ejemplo, “escriba ; para separar los elementos de una lista”).

También con los símbolos de puntuación el comportamiento de las nuevas voces es más intuitivo. Símbolos como “¿” o “¡”, que no pueden incluirse dentro de un texto a sintetizar con la voz *el_diphone* (puesto que generan una síntesis errónea), sí pueden ser incorporados en los textos a sintetizar por las nuevas voces masculina y femenina.

Además, un *bug* identificado en la voz *el_diphone* ha sido resuelto en las voces desarrolladas. Este *bug* viene dado por el hecho de sintetizar un texto que contenga dos símbolos de puntuación seguidos (por ejemplo, “...vino desde Aranjuez (Madrid). Luego...”). Cuando algo como esto ocurre, la voz

	Document title		Author		Page
	Voces en Guadalínex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	23 (of 25)

original no realiza la pausa correspondiente en el punto. Los parámetros responsables de este *bug* han sido retocados para las nuevas voces: el resultado es que, con estas voces, la confluencia de símbolos de puntuación no perjudica a la asignación de pausas dentro del texto.

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	
		0.3	31/05/2007	Validado	24 (of 25)

8. BIBLIOGRAFÍA RECOMENDADA

Black, A. W. y K. A. Lenzo. 2007. *Building Synthetic Voices*. Language Technologies Institute, Carnegie Mellon University. (<http://festvox.org/bsv/>)


Black, A. W., P. Taylor y R. Caley. 2002. *The Festival Speech Synthesis System. Edition 1.4, for Festival Version 1.4.3*. Language Technologies Institute, Carnegie Mellon University. (<http://festvox.org/docs/manual-1.4.3/>)

Black, A. W. y K. A. Lenzo. 2000. *Limited Domain Synthesis*. ICSLP2000. Pekín, China. (http://www.cs.cmu.edu/~awb/papers/ICSLP2000_ldom.pdf)

Black, A. W., P. Taylor y M. Macon. *Speech Synthesis in Festival: A practical course on making computers talk. Edition 2.0, for Festival Version 1.4.1*. Language Technologies Institute, Carnegie Mellon University. (http://festvox.org/festtut/notes/festtut_toc.html)

Kominek, J. y A. W. Black. 2003. CMU ARCTIC Databases for Speech Synthesis, Ver. 0.95. Language Technologies Institute, Carnegie Mellon University. (http://festvox.org/cmu_arctic/cmu_arctic_report.pdf)

Ríos Mestre, A. 1999. "La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico". *Estudios de Lingüística Española*, Volumen 4. Universidad Autónoma de Barcelona, Barcelona. (<http://elies.rediris.es/elies4/index.htm>)

	Document title		Author		Page
	Voces en Guadalinex		GPG, DMF, CSV, PMP		
	Document number	Revision	Date	Status	25 (of 25)
	0.3	31/05/2007	Validado		