DISEÑO DE PRUEBAS PARA LA EVALUACIÓN DE HABLA SINTETIZADA EN ESPAÑOL Y SU APLICACIÓN A UN SISTEMA DE CONVERSIÓN DE TEXTO A HABLA

Lourdes Aguilar, Josep M. Fernández, Juan M. Garrido y Joaquim Llisterri
Departament de Filologia Espanyola, Facultat de Filosofia i Lletres, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona. Fax: (93) 581.16.86. Correo electrónico: Joaquim.Llisterri@cc.uab.es

Alejandro Macarrón, Luis Monzón y Miguel Angel Rodríguez
Servicios de Tratamiento del Habla, Telefónica I+D, Emilio Vargas 6, 28043 Madrid. Fax (91) 337.42.02. Correo electrónico: miguel@omega.tid.es

Resumen

En este trabajo se presenta un conjunto de pruebas para la evaluación de la inteligibilidad, comprensión y calidad global del habla generada por sistemas de conversión de texto a habla para el español, y su aplicación al sistema desarrollado por Telefónica I+D.

Las pruebas, desarrolladas en el *Departament de Filologia Espanyola* de la *Universitat Autònoma de Barcelona*, son básicamente adaptaciones al español de otras pruebas ya existentes para otras lenguas. Las pruebas están diseñadas para evaluar la inteligibilidad, comprensión y naturalidad del habla de los conversores texto a habla en diferentes niveles: la evaluación de la inteligibilidad se realiza por medio de tres pruebas diferentes y comprende, por un lado, la inteligibilidad segmental, y por otro, la inteligibilidad de palabras; la evaluación de la comprensión se realiza por medio de una prueba de comprensión de textos; finalmente, la última prueba sirve para establecer la calidad global, naturalidad y aceptación del sistema por parte del usuario.

Estas pruebas fueron empleadas, siguiendo la metodología desarrollada para tal fin, para evaluar la calidad del sistema de conversión de texto a habla desarrollado para el español por Telefónica I+D. Los resultados de esta evaluación han permitido obtener una primera valoración del sistema que servirá como punto de referencia para la comparación con versiones posteriores del mismo o con otros conversores desarrollados para el español.

1. Introducción

Los avances experimentados en el campo de la tecnología del habla permiten disponer actualmente de sistemas de conversión de texto a habla que alcanzan un elevado grado de inteligibilidad. Pisoni *et al.* (1985) apuntaban hace ya ocho años que tales sistemas llegarán pronto a los niveles de inteligibilidad propios del habla natural. Sin embargo, como afirmaba Allen (1985) por las mismas fechas - y su opinión sigue siendo válida hoy en día -, aún no disponemos de sistemas de síntesis que reproduzcan adecuadamente la variabilidad fonética observada en el habla natural.

Estos son también los dos aspectos que cabe considerar al plantearse la utilización de sistemas de síntesis en aplicaciones reales: la inteligibilidad por una parte y la naturalidad por otra. Si bien se han realizado notables avances en la primera, no se ha conseguido aún alcanzar el grado de naturalidad deseable para una difusión masiva de los sistemas de conversión de texto a habla. La relación entre naturalidad e inteligibilidad es, en cierto modo, complementaria y depende en parte de la aplicación del sistema de síntesis.

Tal como ha ido avanzando el desarrollo de los sistemas de conversión de texto a habla, haciendo posible su utilización práctica, se ha impuesto la necesidad de disponer de herramientas que permitan comparar las prestaciones de diversos sistemas de síntesis. Al mismo tiempo, se ha visto también el interés en estudiar la reacción al uso de la voz sintetizada y, muy especialmente, las diferencias en los mecanismos de percepción entre el habla sintetizada y el habla natural.

En concreto, siguiendo a Pisoni *et al.* (1985) pueden plantearse las cuestiones siguientes:
(1) ¿ Con qué grado de precisión se reconocen sonidos y palabras sintetizadas ?
(2) ¿ Con qué precisión se entiende el sentido de una frase en habla sintetizada ?
(3) ¿ Qué dificultades plantea la percepción y la comprensión del habla sintetizada ?

El desarrollo de estos temas de investigación ha llevado a la creación de un campo de trabajo conocido como *Speech Output Assessment* que se ocupa, en conjunto, de evaluar el resultado de la síntesis y el reconocimiento del habla. Una muestra del interés que han despertado estos aspectos entre la comunidad científica puede verse en las publicaciones monográficas que recientemente han aparecido sobre el tema (ESCA 1989; Pols (Ed) 1990; Castagneri (Ed) 1991).

Actualmente contamos con más de un sistema de conversión de texto a habla para el castellano, y parece por tanto que ha llegado el momento de dotarse de mecanismos que permitan evaluar tanto su inteligibilidad como su naturalidad y que favorezcan además la comparación objetiva entre sistemas distintos. Estos instrumentos se han desarrollado ya para otras lenguas - especialmente para el inglés - y están siendo aplicados con éxito a un gran número de sistemas de síntesis existentes en la actualidad, tanto comercializados como a prototipos de laboratorio.

En Telefónica I+D se ha desarrollado un conversor de texto a habla de alta calidad para el español, del que se dispone de una versión en voz masculina y de otra en voz femenina (Rodríguez *et al.*, 1993). Las evaluaciones realizadas hasta el momento reflejaban unos resultados muy alentadores, pero habían sido realizadas de una manera informal, y únicamente para una de las voces disponibles, la voz masculina de uno de los locutores utilizados en el desarrollo del sistema. Una vez que se consideró que la voz sintetizada producida por el conversor alcanzaba un nivel de calidad que hacía posible su uso para proporcionar servicios de información dirigidos a un público genérico, cobró importancia la necesidad de realizar una evaluación formal, con los siguientes objetivos:

(a) Comparar la inteligibilidad, naturalidad y aceptabilidad de la voz sintetizada con la voz natural, a la que se toma como límite de calidad máxima teórica para un conversor de texto a habla.

(b) Descubrir aquellos aspectos del conversor de texto a habla que precisan un mayor esfuerzo de mejora, y tener una base más o menos objetiva para cuantificar esa mejora a lo largo del tiempo, según se vayan desarrollando nuevas versiones del sistema.

(c) Disponer de datos que permitan comparar la calidad del conversor de texto a habla con otros sistemas similares disponibles, tanto para español como para otras lenguas.

(d) Evaluar la utilidad y la potencial utilización del conversor de texto a habla en aplicaciones dirigidas al público en general; básicamente, se trata de aplicaciones basadas en el suministro de información escrita como texto de ordenador, convertida en señal de habla sintetizada y enviada a través de la línea telefónica.

Para alcanzar estos objetivos, se han desarrollado las pruebas y la metodología que se describen en los apartados 2 y 3.1, respectivamente.

2. Presentación de las pruebas

Las pruebas desarrolladas son, en su mayor parte, adaptaciones españolas de pruebas clásicas utilizadas previamente en la evaluación de otros sistemas de conversión de texto a habla. Pretenden constituir una batería completa que pueda ser usada en el futuro con otros sistemas y que permita, a la vez, realizar un seguimiento de la evolución de las distintas versiones de un mismo sistema. Se optó por adaptar pruebas previamente utilizadas a fin de facilitar la comparación con datos publicados sobre sistemas en otras lenguas (véase, por ejemplo, el trabajo de Logan et al., 1989 donde se aplica el test de rimas modificado a ocho conversores), aunque en el futuro no se descarta una adaptación al español de las pruebas estandarizadas desarrolladas en el marco de los sucesivos proyectos SAM -*Multilingual Speech Input/Output Assessment, Methodology and Standardisation* - (SAM, 1992).

2.1. Pruebas de inteligibilidad segmental

La inteligibilidad de los elementos segmentales en los sistemas de conversión de texto a habla puede estudiarse mediante el Test de Rimas Modificado (*Modified Rhyme Test*, MRT), cuyo objetivo es evaluar la inteligibilidad del habla sintetizada en el nivel fonémico. La prueba consiste en listas de palabras monosilábicas con sentido, de estructura consonante - vocal - consonante (CVC), aunque en algunos casos puede usarse una estructura CV o VC. En la mitad de las opciones, las respuestas comparten el segmento vocal-consonante y en la otra mitad, la parte consonante-vocal. Otras restricciones impuestas a las palabras empleadas son la monosilabicidad y la representación ortográfica constante del núcleo vocálico dentro de un conjunto dado. Las listas no están fonéticamente equilibradas, pero se descartan las palabras poco frecuentes y se intenta incluir los alófonos representativos de cada una de las categorías fonéticas principales de la lengua. En la hoja de respuesta, el informante debe escoger entre un conjunto cerrado de alternativas que difieren únicamente en un alófono situado en posición inicial o final.

Se han desarrollado para el español tres tipos de pruebas dedicadas a evaluar la inteligibilidad de consonantes, de grupos consonánticos y de vocales en contacto. En la primera se evalúa, como en el MRT clásico, la inteligibilidad de consonantes en posición inicial y final de palabras monosilábicas de estructura CVC. En la segunda se utilizan como estímulos los grupos consonánticos obstruyente + líquida del español, ya que éstos no se incluyen en la prueba anterior, y en la tercera se estudia la inteligibilidad de las combinaciones de dos vocales y el efecto del acento en la síntesis de estas combinaciones.

La elección de los materiales en la adaptación del MRT al español[1] está sujeta a las características impuestas por House *et al.* (1965): monosilabicidad, estructura CVC (aunque se acepta también, como se ha mencionado antes, VC o CV) y representación ortográfica constante del núcleo vocálico. La última restricción, sin embargo, no es muy importante en una lengua como el castellano, cuyo sistema ortográfico mantiene una buena correspondencia con el sistema fonológico.

Por el contrario, el requisito de que las palabras sean monosílabas plantea problemas dada la tendencia mayoritaria del español a la estructura bisílaba. Por un lado, gran parte de las palabras monosílabas no son familiares para el hablante, aunque todas las que se han utilizado en la prueba se hallan recogidas en los diccionarios de la lengua y, por otra parte, para algunos conjuntos no se dispone de seis combinaciones tal como se propone en la versión original de la prueba. Para solucionar este último problema, se ha reducido el corpus de tal modo que las seis alternativas de respuesta se han reducido a cuatro, y los cincuenta grupos a cuarenta.

Aunque no hay un estricto control de la familiaridad de la palabra ni de la frecuencia relativa de aparición de los sonidos en la lengua, se ha intentado incluir sonidos representativos de todas las categorías fonéticas principales. La elección de los segmentos que aparecen en posición inicial o final se ajusta a las restricciones fonotácticas de la lengua española.

En lo que se refiere a la prueba de evaluación de la inteligibilidad de los grupos consonánticos, para encontrar un número suficiente de estímulos ha sido necesario relajar la condición de que las palabras que aparezcan tengan sentido. Por ello, la mayoría de las estímulos de esta prueba son palabras fonológicamente posibles pero sin sentido. En cambio, en la prueba de inteligibilidad de combinaciones de vocales, se han utilizado únicamente palabras con sentido en español.

2.2. Pruebas de inteligibilidad de palabras

Para evaluar la inteligibilidad de los sistemas de conversión de texto a habla mediante el reconocimiento de palabras en contexto oracional suelen utilizarse las llamadas Frases Psicoacústicas de Harvard (*Harvard Psycho-acoustic Sentences*; Egan, 1948). Se trata de oraciones con sentido, con estructura sintáctica

[1] La adaptación al español de esta prueba ha sido realizada por Lourdes Aguilar y se describe detalladamente en Aguilar (1991).

correcta y variada y fonéticamente equilibradas. La respuesta es abierta, de modo que el informante no elige entre distintas opciones sino que simplemente escribe o repite la frase que ha oído.

Este tipo de prueba presenta una limitación importante, que se refleja en un elevado porcentaje de respuestas acertadas. Ello se debe a que el individuo que realiza la prueba no utiliza únicamente información fonética, sino también semántica, ya que el apoyo del contexto ayuda a restituir los vacíos fonéticos que se pueden producir en el momento de la audición. Este fenómeno aumenta de manera muy significativa el grado de inteligibilidad.

La versión española de la prueba[2] consta de 10 grupos de 10 frases con sentido cada uno. Se trata de frases sintácticamente bien formadas, y aparecen tanto oraciones simples como coordinadas y subordinadas, así como todo tipo de tiempos y personas verbales. Con el fin de ejercer un cierto control sobre los aspectos suprasegmentales, no se han incluido frases exclamativas o interrogativas, que suponen una curva melódica diferente.

Las frases están fonéticamente equilibradas en grupos de 10, de forma que el equilibrio no se mantiene necesariamente si se toman las 100 frases en conjunto. Como referencia para la frecuencia de aparición de los fonemas del español se han utilizado los datos de Navarro Tomás (1946), cotejándolos con los de Alarcos (1950). Para conseguir el equilibrio fonético se ha utilizado el programa EQUIPHON[3]. Con él se compara la distribución teórica en la lengua - partiendo, en este caso, de los datos de Navarro Tomás ya mencionados- y la distribución de la muestra sobre la que se elabora la prueba.

Para resolver el problema de la información contextual que interviene al estudiar la inteligibilidad de palabras en frases con sentido completo, se ha desarrollado una prueba alternativa en la que las frases que se presentan para su identificación carecen de sentido. Esta prueba se basa en el corpus de frases conocido como las Frases Semánticamente Anómalas de Haskins (*Haskins Semantically Anomalous Sentences*, Nye - Gaitenby, 1974).

Su objetivo es evaluar, del mismo modo que en la prueba anterior, la inteligibilidad del habla sintetizada mediante el reconocimiento de palabras en contexto oracional. Las frases de la prueba tienen una estructura sintáctica correcta, pero en cambio carecen de sentido, por lo que la información predecible a partir del contexto disminuye radicalmente.

El formato de respuesta es abierto, de forma que el informante debe reconocer cuatro palabras clave por frase, sin que su elección se vea afectada por el significado de la frase. La información recogida refleja, por tanto, la comprensión de cada palabra según el grado de inteligibilidad alcanzado en la síntesis.

---

[2] La adaptación al español de esta prueba ha sido realizada por Anna Valero y se describe detalladamente en Valero (1991).

[3] Este programa ha sido desarrollado por el Dr. Bernard Harmegnies en la Universidad de Mons.

El corpus desarrollado en español[4] consta de 50 frases que mantienen las características originales de las Frases Semánticamente Anómalas de Haskins, aunque con algunas variaciones. El rasgo más importante que se ha tenido en cuenta, además de la falta de sentido de las frases, ha sido el equilibrio fonético. Por ello, se han introducido ciertas diferencias respecto a la versión inglesa.

Se han obtenido así 50 frases de estructura sintáctica fija, aunque se hayan introducido algunas variaciones categoriales en ciertas posiciones. La estructura sintáctica resultante es la siguiente: <Art> + [<Adj>/<Nom>/<Adv>] + <Nom> + <Verbo> + <Art> +<Nom>. La alternancia categorial sólo se da en una posición, y en la mayoría de los casos aparece <Adj>. El verbo de estas frases es siempre transitivo, y en forma simple (presente o pasado).

2.3. Prueba de comprensión de textos

Para llevar a cabo la evaluación de la comprensión del habla sintetizada suelen utilizarse textos cortos seguidos de una serie de preguntas sobre su contenido. Pisoni (1987) describe una prueba de este tipo, consistente en 15 textos narrativos y un conjunto de preguntas con un formato de respuesta de elección múltiple extraídos de diversas baterías de pruebas de comprensión de lectura para adultos. El número de preguntas varía según el texto (de 4 a 9), así como la duración (de 56 sg. a 135 sg.) y el número de palabras de cada texto (de 159 a 327).

En un principio se pensó llevar a cabo una prueba de comprensión tal como se ha descrito anteriormente. Para ello se utilizaron algunos textos de un programa de estimulación de comprensión lectora (Huerta - Matamala, 1989) y se diseñó un conjunto de 15 textos con sus correspondientes preguntas, de duración y estilo similar a los que se recogen en Pisoni (1987)[5].

Sin embargo, la realización de la prueba completa ocupa un tiempo excesivamente largo, por lo que, finalmente, se redujo la batería a tres textos, seguidos cada uno de 6 preguntas con 4 respuestas posibles entre las que el sujeto que realiza la prueba debe elegir [6].

En lo que respecta a la variedad estilística, el primer texto es un fragmento periodístico de 121 palabras, el segundo puede calificarse como de divulgación científica (169 palabras) y el tercero presenta todas las características del estilo narrativo (95 palabras). Se reflejan así tres tipos de textos con los que potencialmente se encontraría un usuario de un sistema de conversión de texto a habla, cada uno con un contenido y un estilo de redacción netamente diferenciados.

---

[4] La adaptación al español de esta prueba ha sido realizada por Anna Serra y se describe detalladamente en Serra (1991).

[5] La prueba se presenta detalladamente en Alberte (1991).

[6] Andreu (1991) presenta los textos seleccionados y describe detalladamente los criterios de diseño de esta prueba. Las modificaciones posteriores han sido realizadas por los autores del presente trabajo.

Se ha procurado también que al menos uno de los textos (el periodístico) contenga cifras y abreviaturas.

2.4. Prueba de evaluación de la calidad global

Con el fin de estudiar la calidad global de un sistema de conversión de texto a habla puede considerarse la prueba presentada por Robert *et al.* (1989), centrada en la evaluación subjetiva de la naturalidad y la aceptabilidad de tres sistemas de síntesis del francés.

En esta prueba, los oyentes deben valorar una serie de pares de adjetivos semánticamente contrarios en una escala de cinco puntos (-2, -1, 0, +1, +2). Tal metodología tiene su origen en la técnica conocida como diferencial semántico. Una de las ventajas de la prueba es que no requiere conocimientos fonéticos específicos para realizarla, ya que las respuestas son de carácter eminentemente subjetivo. Precisamente en esto consiste también su mayor inconveniente, ya que la interpretación de los pares de adjetivos propuestos depende de cada uno de los individuos que realizan la prueba.

Los diferentes pares de adjetivos contrarios que configuran el test diseñado para el español[7] se han obtenido a partir de las listas presentadas por diversos autores que han tratado de evaluar de modo subjetivo la calidad global de la síntesis. Se partió de los 25 pares de adjetivos propuestos por Robert *et al.* (1989) para el francés; también se utilizó el listado propuesto por Nusbaum *et al.* (1984) para el inglés tal como lo reproducen Robert *et al.* (1989:642). Asímismo, se consultó la lista de 13 adjetivos propuesta por A.A. Rodríguez (1989) para la evaluación de las características de la voz radiofónica.

Considerando estas tres listas, se seleccionaron los pares de adjetivos que, en la versión castellana de la prueba, sirven principalmente para la evaluación de la calidad de la voz.

Sin embargo, la percepción de la calidad global de un sistema no depende únicamente de la calidad de la voz sintetizada, sino también de aspectos como la calidad de los elementos segmentales y, muy especialmente, de los suprasegmentales. Por ello se incluyeron en la prueba apartados con pares de adjetivos - en algunos casos afirmaciones con las que el sujeto debe mostrar su grado de acuerdo o desacuerdo - dedicados específicamente a tales cuestiones.

Considerando también que un sistema de conversión de texto a habla debe utilizarse en aplicaciones concretas, se incorporó a la prueba un apartado dedicado a evaluar la adecuación del conversor a diversos usos - servicios telefónicos de información general, lectura en alta voz de textos y enseñanza de la lengua a extranjeros - y a valorar la posible frecuencia de utilización por parte de futuros usuarios.

La prueba así diseñada contempla pues la evaluación de los siguientes aspectos:

(1) Calidad de la voz

---

[7] La adaptación al español de esta prueba ha sido realizada por Natividad Fernández y se presenta detalladamente en Fernández (1992). Las modificaciones posteriores han sido realizadas por los autores del presente trabajo.

(2) Estilo de lectura

(3) Variedad geográfica

(4) Calidad segmental y suprasegmental

(5) Calidad global

(6) Adecuación a aplicaciones concretas

(7) Frecuencia de uso

En cuanto al método de respuesta, se optó finalmente por una escala del 1 al 5, de modo que 1 y 5 representan los extremos negativo y positivo respectivamente, mientras que 3 representa el valor medio.

3. Evaluación del conversor de texto a habla de Telefónica I+D

3.1. Metodología

Tal como se ha expuesto en el apartado 2, la batería de evaluación consiste en cinco pruebas, cada una de las cuales se realizó en cuatro condiciones diferentes: voz masculina y voz femenina, y habla sintetizada y habla natural. La prueba de calidad global es la única que se llevó a cabo solamente con habla sintetizada, aunque dado que se trata de voz filtrada telefónicamente, se introducirá en el futuro una evaluación de la calidad global del habla natural.

La tabla 1 resume los siguientes datos para cada prueba: el número de respuestas a la prueba, el número de sujetos que la realizaron, el número total de respuestas que se obtuvieron para cada condición y el número total de respuestas que se obtuvieron para cada prueba en cada condición.

	Prueba de intelig.segmental: cons.	**Prueba de intelig. segmental: grupos cons. y combin. de vocales**	**Prueba de intelig. de palabras en frases con sentido**	**Prueba de intel, de palabras en frases sin sentido**	**Prueba de compr. de textos**	**Prueba de eval.de la calidad global**
Número de respuestas de la prueba	160	40	400	200	18	38
Número de sujetos	25	25	20	20	20	30
Número total de respuestas por condición	160 x25= 4000	40 x25= 1000	400 x20=8000	200 x20= 4000	18 x20= 360	38 x30= 1140

Número total de respuestas	4000 x2 tipos de voz x2 tipos de habla= 16000	1000 x2 tipos de voz x2 tipos de habla= 4000	8000 x2tipos de voz x2tipos de habla= 32000	4000 x2 tipos de voz x2tipos de habla= 16000	360 x2 tipos de voz x2tipos de habla= 1440	1140 x 2tipos de voz= 2280

Tabla 1: Número de respuestas obtenidas para cada prueba en función del número de sujetos que la realizan y de las condiciones de realización

El número total de sujetos seleccionados para la evaluación fue de 280; 93 hombres (33.3%) y 187 mujeres (66.6%), de edades comprendidas entre 18 y 30 años, aunque la mayoría se sitúa en torno a los 20- 23 años.

De los 280 sujetos considerados, 271 (96,8%) estaban cursando estudios superiores y el resto tenía un nivel medio de estudios (Bachillerato, Formación Profesional o sus equivalentes).

Los sujetos eran mayoritariamente bilingües con dominancia castellana (189, que corresponde al 67.5%); en el caso de que fueran bilingües con dominancia catalana (91, es decir, el 32.5%) - lengua del padre y de la madre catalana, lengua de uso habitual catalán- se evaluaron sus problemas de producción en lengua castellana mediante una corta entrevista con las personas encargadas de seleccionar a los sujetos. No se aceptaron para las pruebas todas aquellas personas que presentaran algún tipo de dificultad durante esta entrevista. Sin embargo, teniendo en cuenta el alto nivel cultural de los sujetos de la prueba y la situación sociolingüística de Cataluña, puede asegurarse que todas las personas que participaron en el experimento tenían un contacto suficiente con el español, además de ser usuarios potenciales del sistema que se evaluaba.

Para evitar el efecto de habituación, cada sujeto respondió únicamente a una prueba con habla sintetizada. Por otra parte, cada sujeto respondió a una prueba con habla natural, que no coincidía con la que había respondido en habla sintetizada. La prueba de habla natural se pasó siempre en primer lugar, con el fin de que el sujeto utilizase el habla natural como punto de referencia y no a la inversa.

Para cada prueba se incluyeron estímulos de entrenamiento a los que los sujetos debían contestar. Las respuestas a estos estímulos no formaron parte de la valoración final.

Las pruebas se realizaron en una cabina semianecoica situada en las dependencias del *Laboratori de Fonètica* de la *Universitat Autònoma de Barcelona.*

Los estímulos de las pruebas se presentaron, tanto en habla natural como en habla sintetizada, filtrados con un filtro paso-banda de 300Hz - 3400Hz que modela la banda telefónica. La grabación a partir de la salida del conversor de texto a habla se llevó a cabo en cassettes Sony (Normal) con el sistema Dolby B. Las pruebas se realizaron con una platina Tascam 112 y una tabla de mezclas Tascam 106, que

incorporaban también el sistema Dolby B. Los sujetos recibieron los estímulos a través de unos auriculares binaurales Sennheiser HD-25-1 a un nivel confortable de intensidad.

3.2. Resultados

3.2.1. Pruebas de inteligibilidad segmental

En la tabla 2 se resumen los resultados obtenidos en las pruebas de identificación de elementos segmentales para cada una de las cuatro condiciones:

Prueba	masc. nat.	fem. nat.	masc. sint.	fem. sint.
Consonantes iniciales	98%	98%	93.2%	90.6%
Consonantes finales	94.6%	96.8%	79.7%	79.3%
Media consonantes	93.3%	97.4%	86.45%	84.95%
Grupos consonánticos	95%	93.5%	84%	59.66%
Comb. de vocales	97.07%	94.2%	79.4%	79.28%
Media global	96.16%	95.62%	84.07%	77.21%

Tabla 2: Porcentaje de identificaciones correctas obtenidas en cada una de las pruebas de identificación de elementos segmentales en las cuatro condiciones.

Puede observarse que, para el habla sintetizada con voz masculina, los mejores resultados se obtuvieron en la identificación de consonantes en posición inicial y los más bajos en la identificación de combinaciones de vocales y de consonantes en posición final. En el habla sintetizada con voz femenina, los resultados más bajos se presentaron en la prueba de identificación de grupos consonánticos y los mejores en la identificación de consonantes iniciales.

Las diferencias globales entre el habla natural y la sintetizada se situaron en un 12.09% para la voz masculina y en un 18.41% para la voz femenina en lo que a la inteligibilidad segmental se refiere.

3.2.2. Pruebas de inteligibilidad de palabras

En la siguiente tabla se presentan los porcentajes de error en la identificación de palabras en frases con sentido obtenidos en las cuatro condiciones de la prueba:

Condición	% errores
Voz natural masculina	0.67%
Voz natural femenina	0.67%
Natural	*0.67%*
Voz sintetizada masculina	3.35%
Voz sintetizada femenina	5.31%
Sintetizada	*4.33%*

Tabla 3: Porcentaje de errores en la prueba de inteligibilidad de palabras en frases con sentido

La identificación de palabras en frases con sentido presentó más errores en el habla sintetizada que en la natural, y el mayor número de errores se localizó en el habla sintetizada con voz femenina: el porcentaje de inteligibilidad fue del 96.65% en el habla sintetizada masculina y del 94.69% en la femenina.

La tabla 4 resume los resultados de la prueba de inteligibilidad de palabras en frases sin sentido en las cuatro condiciones:

Condición	% de errores
Voz natural masculina	4.87%
Voz natural femenina	3.52%
Natural	*4.19%*
Voz sintetizada masculina	12.7%
Voz sintetizada femenina	15.8%
Sintetizada	*14.25%*

Tabla 4: Porcentaje de errores para las cuatro condiciones en la prueba de indentificación de palabras en frases sin sentido.

Como puede observarse, el porcentaje de inteligibilidad fue inferior en habla sintetizada: 87.3% en el habla sintetizada con voz masculina, frente al 95.13% en el habla natural. Cuando la síntesis se realizó con voz femenina, el mismo porcentaje descendió al 84.2%, mientras que en habla natural se situó en el 96.48%.

3.2.3. Prueba de comprensión de textos

Tal como puede observarse en la tabla 5, en habla sintetizada se obtuvo un 2.23% menos de respuestas acertadas que en habla natural en el caso de la voz masculina, y un 11.66% menos en la femenina:

Condición	Porcentaje de respuestas correctas
Voz natural masculina	76.11%
Voz natural femenina	83.88%
Natural	*79.99%*
Voz sintetizada masculina	73.88%
Voz sintetizada femenina	72.22%
Sintetizada	*73.05%*

Tabla 5: Porcentaje de respuestas correctas por sujeto obtenidas en la prueba de comprensión de textos en la cuatro condiciones.

En conjunto, los resultados de la prueba muestran que, si bien en habla natural parece ser mejor la comprensión cuando la voz es femenina, en habla sintetizada no existen grandes diferencias en la comprensión de textos en función de la voz utilizada. Sin embargo, es preciso destacar las diferencias que se encontraron entre la voz femenina natural y la sintetizada.

3.2.4. Prueba de evaluación de la calidad global

Se observa al analizar los resultados de esta prueba que la voz masculina sintetizada se percibía como más amable que la femenina, caracterizada como más agresiva, estridente, clara y aguda en relación con la masculina. En cambio, la voz sintetizada femenina se valoró, con respecto a la masculina, como una voz que ayuda más a la concentración, más fácil de entender, más rápida y más interesante.

En cuanto a la calidad global del sistema, se obtuvieron puntuaciones más altas en lo que respecta a la voz sintetizada femenina. El sistema se juzgó como más adecuado, eficaz, satisfactorio y aceptable en voz femenina. Esto es especialmente claro en los casos referidos a la utilización del sistema en un servicio telefónico de información general o de noticias de actualidad. En la misma línea, la voz femenina fue evaluada con valores más altos en lo que respecta a la frecuencia de uso.

4. Conclusiones

Como conclusión, puede decirse que el sistema de conversión de texto a habla desarrollado por Telefónica I+D es más adecuado en su versión masculina que en la versión femenina en lo que se refiere a la valoración objetiva, con una inteligibilidad de elementos segmentales cercana al 85%, una comprensión de palabras que supera ligeramente el 90% y que permite comprender más de un 70% en la lectura de un texto. La versión femenina proporciona unas prestaciones inferiores, pero tiene como ventaja una mayor aceptación por parte de los sujetos que participaron en las pruebas realizadas.

El análisis de los resultados de las pruebas ha permitido la localización de incorrecciones en la segmentación de algunas unidades. También ha hecho reconsiderar la caracterización de las semivocales y semiconsonantes en los diptongos. Hasta el momento, la única distinción entre semivocales,

semiconsonantes y las correspondientes vocales cerradas [i, u] se modela en el conversor mediante los parámetros prosódicos de duración y de frecuencia fundamental, lo cual se ha mostrado poco adecuado en algunos casos. Finalmente, ha permitido descubrir la necesidad de mejorar la calidad de la voz sintetizada femenina pues, aunque en cierto sentido ha sido mejor considerada que la voz sintetizada masculina en la prueba de evaluación de la calidad global, sus resultados en las pruebas de inteligibilidad son claramente inferiores.

Referencias

AGUILAR, L. (1991) *Propuesta de un test de evaluación segmental del habla para el castellano: el Test de Rimas Modificado.* Universidad Autónoma de Barcelona, Departamento de Filología Española. ms no publicado.

ALARCOS,E. (1950) *Fonología española.* Madrid: Gredos (Biblioteca Románica Hispánica, Manuales, 1) 1965 4a ed. aumentada y revisada.

ALBERTE, M. (1991) *Evaluación del habla sintetizada: test de comprensión auditiva.* Universidad Autónoma de Barcelona, Departamento de Filología Española. ms no publicado.

ALLEN, J. (1985) "A Perspective on Man-Machine Communication by Speech", *Proceedings of the IEEE* 73,11 : 1541-1550.

ANDREU, M. (1991) *Diseño de una prueba de comprensión de textos en habla sintetizada.* Universitat Autònoma de Barcelona, Departament de Filologia Espanyola. ms. no publicado.

CASTAGNERI, G. (Ed) (1991) *Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods.* Chiavari 26-28 September 1991. Organised by CSELT in cooperation with CEC DGXIII, ESCA, ESPRIT Project 2589 "SAM". Torino: CSELT.

EGAN,J.P. (1948) "Articulation testing methods", *Laryngoscope*, Vol.58, pp.955-991.

ESCA (1989) *Proceedings of the ESCA Tutorial Day and Workshop on Speech Input/Output Assessment and Speech Databases.* Noordwijkerhout, the Netherlands, 20-23 September 1989.

HOUSE, A.S.- WILLIAMS, C.E.- HECKER, M.H.L.- KRYTER, K.D. (1965) "Articulation Testing Methods: Consonantal Differenciation with a Closed- Response Set", *Journal of the Acoustical Society of America,* 37, 1: 158- 166.

HUERTA - MATAMALA (1990) *Programa de estimulación de Comprensión lectora (PCL)* Madrid: Distribuciones Visor.

LOGAN, J.S.- GREENE, B.G.- PISONI, D.B. (1989) " Segmental intelligibility of synthetic speech produced by rule ", *Journal of the Acoustical Society of America* 86,2: 566-581

NAVARRO TOMÁS, T. (1946) " Escala de frecuencia de fonemas españoles " in *Estudios de fonología española*. New York: Las Américas Publishing Company, 1966 2a ed. pp. 15-30

NUSBAUM, H.C.- SCHWAB, E.C.- PISONI, D. (1984) " Subjective evaluation of synthetic speech: Measuring Preference, Naturalness and Acceptability", *Research on Speech Perception*, Progress Report 10, Department of Psychology, Indiana University.

NYE, P.W.- GAITENBY, J. (1974) "The Intelligibility of Synthetic Monosyllable Words in Short, Syntactically Normal Sentences", *Haskins Laboratories Status Report on Speech Research* SR-37/38: 169-190

PISONI, D. B.- NUSBAUM, H. C.- GREENE, B. G. (1985) "Perception of Synthetic Speech Generated by Rule", *Proceedings of the IEEE* 73,11: 1665-1676.

PISONI, D.B. (1987) "Some measures of intelligibility and comprehension" in ALLEN, J.-HUNNICUT, M.S.-KLATT, D.H. *From text to Speech. The MITalk System.* Cambridge University Press. pp. 151-171.

POLS, L.C.W. (Ed) (1990) Speech Input / Output Assessment and Speech Databases, Special Issue, *Speech Communication* 9,4.

ROBERT, J.M.- CHOINIÈRE, A.- DESCOUT, R. (1989) " Subjective evaluation of the naturalness and acceptability of three text-to-speech systems in French " in TUBACH, J.P.- MARIANI, J.J. (Eds) *Eurospeech 89. European Conference on Speech Communication and Technology*. Paris, September 1989. Edinburgh: CEP Consultants Ltd. vol 2. pp. 640-643

RODRÍGUEZ, A.A. (1989) *La construcción de una voz radiofónica*, Universidad Autónoma de Barcelona, Facultat de Ciències de la Informació. Tesis Doctoral,ms. no publicado.

RODRÍGUEZ, M.A.- ESCALADA, J.G.- MACARRÓN, A.- MONZÓN, L. (1993) " AMIGO: Un conversor texto-voz para el español", *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural* 13: 389-400.

SAM (1992) *User Guide to Output Asessment.* ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation. Ref, SAM-UCL-G006.

SERRA, A. (1991) *Un test de evaluación de habla sintetizada para el castellano: las frases semánticamente anómalas de Haskins.* Universidad Autónoma de Barcelona, Departamento de Filología Española. ms no publicado.

VALERO, A. (1991) El *corpus de las frases psicoacústicas de Harvard: una adaptación al castellano.* Universidad Autónoma de Barcelona, Departamento de Filología Española. ms no publicado.