Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 5-12 recibido 21-04-2005; aceptado 01-06-2005

Evaluación del modelado acústico y prosódico del sistema de conversión texto-voz Cotovía*

Francisco Campillo Díaz
Universidad de Vigo
ETSI Telecomunicación
campillo@gts.tsc.uvigo.es

Eduardo Rodríguez Banga
Universidad de Vigo
ETSI Telecomunicación
erbanga@gts.tsc.uvigo.es

Resumen: La elevada calidad de los sistemas de conversión texto-voz basados en corpus los ha convertido en el método de síntesis sobre el que más se investiga en la actualidad. En la literatura existen múltiples trabajos sobre los aspectos clave de esta tecnología, es decir, el diseño de las funciones de coste, la caracterización de los segmentos de voz y la estimación de la prosodia, pero hay muy poca claridad sobre las formas más adecuadas para evaluar la calidad de la voz sintética obtenida. En este trabajo se presentan las pruebas de evaluación, tanto subjetivas como objetivas, que se realizaron sobre nuestro conversor de voz Cotovía.
Palabras clave: Síntesis de voz, evaluación

Abstract: Unit selection speech synthesis has become the most researched area in speech technology, as a result of its high-quality performance. There are many approaches about the key points in this technology, that is, the cost functions design, speech parameterisation and prosody estimation, but there is very little research about suitable methods for evaluating the improvements in synthetic speech. In this work the subjective and objective tests performed on our TTS system Cotovía are presented.
Keywords: Speech synthesis, evaluation

1. Introducción

El desarrollo de las técnicas de síntesis de voz basadas en selección de unidades ha supuesto un importante avance en la comunicación entre hombre y máquina ((Black y Campbell, 1995), (Hunt y Black, 1996)). En este tipo de tecnología se renuncia a la comprensión del mecanismo humano de producción del habla y se genera la voz por medio de la concatenación de una serie de segmentos que se seleccionan de un corpus de unidades pregrabadas de un mismo locutor. En esta circunstancia, si se escogen unidades en contextos fonéticos y prosódicos similares a aquéllos en los que se van a aplicar, la voz sintética puede llegar a alcanzar una inteligibilidad y una naturalidad tales que en determinados ámbitos de aplicación es difícil de distinguir del habla humana.

Uno de los aspectos más problemáticos relacionados con cualquier disciplina vinculada con la interacción hombre-máquina es el de la evaluación de la calidad del servicio ofrecido. En el caso concreto de la síntesis de voz, resulta complicado comparar el rendimiento alcanzado con las diferentes formas de modelar los diversos factores que intervienen en el proceso de generación de voz, tales como la estimación de los correlatos prosódicos o la elección de los factores empleados para caracterizar los segmentos acústicos, ante la ausencia de métodos objetivos fiables y la propia complejidad de las pruebas de funcionamiento subjetivas.

En este trabajo se presentan las pruebas realizadas para evaluar el rendimiento de las principales partes del sistema de conversión texto-voz en gallego y castellano Cotovía, desarrollado con la colaboración de investigadores de las universidades de Vigo y Santiago de Compostela y del Centro Ramón Piñeiro para la Investigación en Humanidades. Para ello, se comienza con una serie de apartados dedicados a la descripción de las características más destacadas del conversor. En la sección 2 se menciona el diseño de las funciones de coste empleadas para la selección, mientras que en la sección 3 se explican los modelos de estimación de la prosodia de la versión actual del sintetizador. La sección 4 describe el método de selección combinada de unidades acústicas

* Este trabajo ha sido financiado parcialmente por el Ministerio de Ciencia y Tecnología, fondos FEDER y la Xunta de Galicia, dentro de los proyectos TIC2002-02208, PGIDT01PXI32205PN y PGIDT02PXI32201PR

y entonativas de Cotovía. Posteriormente, en la sección 5 se muestran las pruebas objetivas del funcionamiento de los modelos presentados en los apartados anteriores, y en la sección 6 se presentan las pruebas subjetivas realizadas. Por último, la sección 7 se dedica a las conclusiones extraídas a partir de este trabajo.

2. Las funciones de coste

El hecho de generar el habla sintética a partir de la concatenación de una secuencia de unidades disponibles en un corpus finito traslada la complejidad de la síntesis a dos puntos fundamentales: la caracterización del segmento de voz, y el diseño de las funciones de coste a partir de las cuales se efectuará la selección. Tradicionalmente se suelen emplear dos funciones ((Black y Campbell, 1995)): el coste de objetivo, que mide el parecido entre cada una de las unidades del corpus y aquélla cuyas características prosódicas y fonéticas se extraen de la frase de entrada, y el coste de concatenación, que da una idea de la distorsión potencial que se puede producir al unir dos unidades del corpus. Ambas funciones se suelen combinar con un típico algoritmo de programación dinámica, estilo Viterbi, para seleccionar la secuencia de segmentos acústicos disponibles en el corpus más próximos a las características deseadas. Los siguientes apartados se dedican a la exposición de los factores considerados en la versión actual de Cotovía.

2.1. El coste de objetivo

Aunque existen otras opciones (Black y Taylor, 1997), la función de coste de objetivo de Cotovía se basa en una suma ponderada de diversas subfunciones que consideran por separado las diferencias entre las características deseadas de la unidad objetivo y las disponibles de la candidata, para proporcionar un único valor que mide el parecido entre ambas. En concreto, se separan las contribuciones en dos partes diferenciadas (Campillo y Banga, 2003), el coste prosódico y el contextual, tal y como refleja la ecuación (1):

$$C_{obj} = \alpha \times C_{cont} + (1 - \alpha) \times C_{pros} \qquad (1)$$

Al igual que en la mayoría de los sistemas ((Black y Campbell, 1995), (Febrer, 2001)), el coste prosódico C_{pros} está constituido por las diferencias de frecuencia fundamental, duración y energía. En nuestro caso, además, se incluyen unos umbrales con los que se modela la capacidad de detectar diferencias del oído humano: cuando las diferencias están por debajo de dichos umbrales, los respectivos subcostes adoptan un valor nulo.

Por su parte, el coste contextual C_{cont} engloba todos los factores relacionados con el contexto fonético. En concreto, incluye las distancias Euclídeas entre los vectores mel-cepstrum de la unidad objetivo y la candidata (así como las de los fonemas circundantes), y los subcostes relacionados con las diferencias en el tipo de frase, la posición en la frase, el carácter tónico y la posición en la palabra. Para entrenar la influencia relativa de cada subcoste se consideran dos modelos. En el primero, se utiliza regresión lineal para aproximar la distancia espectral entre las unidades comparadas a partir de los valores de los subcostes (Campillo y Banga, 2003). Por su parte, en el segundo se emplea un perceptrón multicapa para aproximar igualmente dicha distancia espectral, con la diferencia de que se utilizan como entradas los valores concretos de los factores con los que se caracteriza el segmento de voz, siendo la propia red neuronal la encargada de aprender los subcostes más adecuados para cada uno de esos factores.

Finalmente, con el factor α de la ecuación (1) se regula la importancia de un subcoste frente al otro.

2.2. El coste de concatenación

Al igual que en la mayoría de los sistemas ((Hunt y Black, 1996), (Febrer, 2001)), en la versión actual de Cotovía se considera la continuidad de frecuencia fundamental, energía y envolvente espectral, modelada esta última por medio de la distancia Euclídea entre vectores de componentes mel-cepstrum. En este caso la mayor aportación son los índices de continuidad espectral (Campillo y Banga, 2004). De un estudio sobre la continuidad espectral en la voz natural se extrajo la conclusión de que ésta dependía de las clases de fonemas concatenados. Así, por ejemplo, en la tabla 1 se recogen las distancias medias de las uniones más frecuentes entre clases en nuestro corpus, ordenadas según su valor para el locutor Freire y con la posición que ocupan respectivamente en las distancias obtenidas del corpus del locutor Paulino. Como se puede observar, existe un cierto orden

de importancia en las voces de ambos locutores.

		Freire		Paulino	
Fon izquierdo	Fon derecho	D_{ij}	Pos	D_{ij}	Pos
Vocal media	Vocal cerrada	1.44	1	1.05	1
Vocal media	Aproximante	1.70	2	1.27	2
Aproximante	Vocal media	1.72	3	1.34	3
Aproximante	Vocal abierta	1.76	4	1.38	5
Vocal abierta	Aproximante	1.80	5	1.36	4
Vibrante	Vocal media	1.84	6	1.89	12
Vocal media	Nasal	1.86	7	1.49	6
Vocal cerrada	Nasal	1.88	8	1.50	7
Nasal	Vocal media	1.90	9	1.73	11
Vibrante	Vocal abierta	1.91	10	1.93	13
Vocal abierta	Nasal	1.96	11	1.56	9
Nasal	Oclusiva sor	2.09	12	1.55	8
Vocal media	Vibrante	2.09	13	2.01	15
Nasal	Vocal abierta	2.14	14	1.71	10
Vocal abierta	Vibrante	2.17	15	1.93	14

Cuadro 1: Distancias medias de las uniones más frecuentes por la zona de transición

Para modelar esta característica de la voz natural se introdujeron los índices de continuidad de la zona estacionaria y de transición (ecuaciones (2) y (3)):

$$I_i^{espectral} = \frac{\text{mín}_{j \in C_{esp}} D_j}{D_i} \quad (2)$$

$$I_{ij}^{espectral} = \frac{\text{mín}_{k,l \in C_{esp}} D_{kl}}{D_{ij}} \quad (3)$$

donde C_{esp} representa el conjunto de las clases espectrales de los fonemas y D_i y D_{kl} denotan las distancias por las zonas estacionaria y de transición, respectivamente. El índice de continuidad espectral se introduce como un factor multiplicativo del subcoste de continuidad espectral, de tal forma que modula la importancia que se le da a éste según las clases de las unidades concatenadas.

3. Estimación de la prosodia

Independientemente de la tecnología concreta en la que se englobe la técnica de conversión de voz, un punto fundamental es el de la estimación de la prosodia, ya que afecta en gran medida a la naturalidad de la voz sintética. En las siguientes secciones se describen los modelos de estimación de la duración, la energía y la entonación. Dado que ésta última está reconocida comúnmente como la más influyente de todas ellas, se le dedica un apartado completo.

3.1. Duración y energía

El modelo de duración de la versión actual se basa en regresión lineal multivariable. Considera factores como la identidad del fonema y de los que lo rodean, el acento, el tipo de frase, la posición en la palabra y en el grupo fónico, la distribución de sílabas tónicas y el tiempo transcurrido desde la pausa anterior.

En cuanto al modelo de energía, incluye parámetros como la identidad del fonema, los circundantes, la posición dentro de la palabra y en el grupo fónico, el acento y el tipo de proposición. Para aproximar los valores de energía se consideraron dos posibilidades: regresión lineal y redes neuronales.

3.2. Entonación

El modelo actual de entonación ((Campillo y Banga, 2002), (Banga et al., 2002)) se basa en la concatenación de contornos de grupos acentuales, extraídos de un corpus prosódico. De esta forma, al igual que en la selección de unidades acústicas, se dispone de múltiples contornos para cada posición en la frase, entre los que se escoge con un típico algoritmo de programación dinámica y diseñando de forma adecuada las funciones de coste.

Al igual que sucede en la selección de unidades acústicas, no todo contorno de grupo acentual es válido para ocupar cualquier posición en la frase. En concreto, en nuestro caso se realiza una clasificación en base a los siguientes factores:

- Tipo de frase: enunciativa, interrogativa, exclamativa e inacabada.
- Posición en el grupo fónico: inicial (antes del primer acento léxico, incluido), final (después del último acento, incluido), intermedia e inicial y final.
- Posición del acento: agudo, llano y esdrújulo.

Análogamente, se escoge la mejor secuencia de grupos acentuales mediante la combinación de una función de coste de objetivo y otra de concatenación. En el coste de objetivo se incluyen factores como el coste de posición del grupo acentual en el grupo fónico, el coste de tipo de proposición (con una clasificación más refinada que el tipo de frase, como grupo parentético, entre comillas...), el número de sílabas, la posición del grupo fónico en la frase, la duración temporal, el fin del grupo (coma, punto, puntos suspensivos, sin pausa...), y la pendiente final del contorno. En cuanto al coste de concatenación, pena-

liza básicamente la diferencia de frecuencia fundamental en el punto de unión.

4. Selección combinada de unidades acústicas y entonativas

Una de las características más destacadas de la voz natural consiste en que se puede transmitir un mismo enunciado con contornos entonativos diferentes, sin alterar su significado. La mayoría de los sintetizadores actuales obvian este hecho, ya que la fase de generación de la voz acepta los datos que le llegan de las etapas anteriores como si fueran los únicos posibles. En el caso concreto de la síntesis basada en selección esto supone que en muchos casos no se consiga la voz sintética de mayor calidad que el conjunto finito de unidades que es el corpus podría generar. Visto de otra forma, la posibilidad de considerar contornos alternativos de frecuencia fundamental proporciona al algoritmo de búsqueda un grado más de libertad, aumentando la probabilidad de encontrar una secuencia de unidades acústicas más adecuada.

En la figura 1 se muestra el esquema de la selección combinada de unidades acústicas y entonativas (Campillo y Banga, 2002). Como se puede observar, para cada grupo fónico de la frase de entrada se realiza una búsqueda de unidades entonativas, de la que se escogen los N mejores caminos. Este conjunto se reduce a M caminos, con el criterio de eliminar aquellos demasiado parecidos y efectuar la selección de unidades acústicas con los que ofrezcan más alternativas, a la vez que se disminuye la carga computacional. Finalmente, se escoge la combinación de unidades acústicas y entonativas que tenga un mejor coste. Hay que destacar que la propia naturaleza del método aumenta la variabilidad de la voz sintética, puesto que la entonación no depende solamente de la estructura de la frase, sino también de la secuencia de fonemas que la componen.

5. Pruebas objetivas

Evaluar un algoritmo de selección de unidades, o lo que es lo mismo, las funciones de coste diseñadas, es un tema complejo. En el caso concreto de la síntesis por corpus, dado que se basa en escoger unidades en los contextos adecuados, puede resultar útil un estudio acerca de en qué medida dichas funciones logran aproximarse a cada una de las

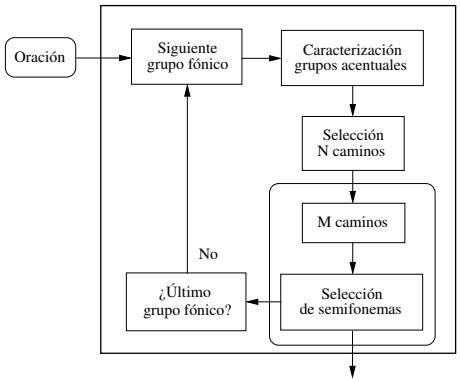

Figura 1: Selección combinada de grupos acentuales y semifonemas

características buscadas como objetivo. Para ello, se realizaron una serie de pruebas sobre un fichero (denominado **FrasesPrueba**) de 857 frases extraídas de un periódico en lengua gallega, con una media de 26.1 palabras por frase y totalmente independiente del corpus de unidades.

En la tabla 2 se muestran los errores cometidos en la frecuencia fundamental y la duración. Evidentemente, cuanto menores sean, menor será la modificación prosódica a la que habrá que someter finalmente a las unidades escogidas, disminuyendo así la distorsión inherente a este tipo de procesos.

	Frecuencia		Duración	
	Media (Hz)	Var	Media (ms)	Var
FrasesPrueba	0.77	37.8	-3.57	107.10

Cuadro 2: Diferencias en frecuencia y duración

Por otra parte, la figura 2 muestra la frecuencia fundamental deseada frente a la de la unidad acústica escogida[1]. Como se puede observar, se aproxima bastante al caso ideal, sobre todo si tenemos en cuenta, como se dijo en el apartado 2.1, que en el coste prosódico se incluyen unos umbrales de no percepción. En concreto, para la frecuencia el umbral es de 5 Hz, y 60 ms para la duración.

En cuanto a los modelos contextuales, en las tablas 3 y 4 se muestran la varianza explicada y el error cuadrático medio de la regresión lineal y la red neuronal, respectivamente, para cada clase de fonema considerado. Como se puede observar, con las redes neuronales se obtienen mejores resultados para cualquiera de las clases, lo cual es bastante lógico dada la dudosa naturaleza lineal del problema, y que son las propias redes las encargadas de

[1]Se obtuvieron gráficas similares para la duración.

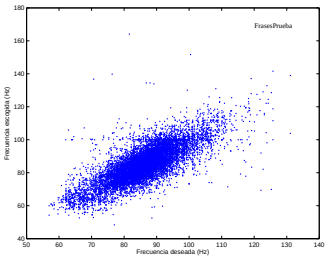

Figura 2: Frecuencia fundamental escogida frente a deseada en FrasesPrueba

aprender el mejor subcoste para cada característica de los segmentos de voz, a diferencia del modelo de regresión, donde se emplean directamente los subcostes como entradas.

Clase	R^2	Entrenamiento	Prueba
Silencio	0.229	1.12	1.08
Vocal abierta	0.238	1.27	1.23
Vocal media	0.242	1.23	1.27
Vocal cerrada	0.321	1.20	1.32
Oclusiva sorda	0.256	1.14	1.21
Oclusiva sonora	0.164	1.25	1.35
Aproximante sonora	0.528	0.92	0.94
Fricativa sorda	0.150	1.31	1.37
Lateral	0.365	0.97	0.94
Nasal	0.316	0.90	0.91
Vibrante	0.269	1.34	1.28

Cuadro 3: Resultados del modelo contextual basado en regresión lineal

Clase	R^2	Entrenamiento	Prueba
Silencio	0.452	0.728	0.699
Vocal abierta	0.480	0.830	0.840
Vocal media	0.398	0.914	0.897
Vocal cerrada	0.424	0.833	0.812
Oclusiva sorda	0.378	0.847	0.863
Oclusiva sonora	0.456	0.756	0.780
Aproximante sonora	0.539	0.801	0.778
Fricativa sorda	0.401	0.917	0.986
Lateral	0.531	0.642	0.650
Nasal	0.491	0.574	0.586
Vibrante	0.374	1.115	1.141

Cuadro 4: Resultados del modelo contextual basado en redes neuronales

En lo que respecta a la selección de unidades entonativas, la tabla 5 muestra la media del error (M), la media del valor absoluto del error (M (abs)) y la varianza del número de sílabas, la duración temporal, la posición del grupo acentual en el grupo fónico y del grupo fónico en la frase. Como aparece reflejado, el error cometido es muy pequeño en todos los casos.

En la figura 3 se muestra la diferencia de frecuencia fundamental en el punto de concatenación entre dos grupos acentuales consecutivos. La media es -0.07 Hz, mientras que

	Sílabas	Dur	Pos grupo$_{ac}$	Pos grupo$_{fon}$
M	0.17	-13.87	-0.05	0.06
M (Abs)	0.25	70.32	0.16	0.11
Var	0.31	9008	0.05	0.04

Cuadro 5: Estadísticos de las diferencias entre los parámetros buscados y los escogidos en la selección de unidades entonativas

la varianza es 3.50.

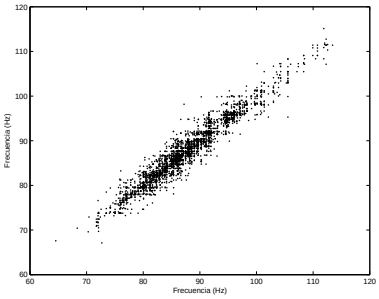

Figura 3: Diferencia de frecuencia en el punto de unión entre dos grupos acentuales consecutivos

En la sección 4 se mencionaba que al considerar diferentes contornos alternativos se le daba un grado más de libertad al algoritmo de selección, lo que permitía obtener mejores resultados. Para comprobar que esta afirmación es cierta se efectuó la selección de unidades variando el número de contornos considerados, y se comprobaron las diferencias entre los valores buscados y los de las unidades acústicas escogidas:

- C_{con}: coste de concatenación.
- C_{cont}: parte de la función de coste de objetivo relacionada con el coste contextual.
- C_{pros}: parte de la función de coste de objetivo relacionada con el coste prosódico.
- C_{obj}: coste de objetivo.

En la tabla 6 se recogen los resultados. La columna **Ant** muestra, por comparación, la misma información con el modelo entonativo de la versión original de Cotovía. En la figura 4 se muestra la evolución del error cometido en la frecuencia fundamental de la zona de transición entre fonemas (un resultado análogo se obtuvo para la zona estacionaria del semifonema). Considerando que los errores siguen una distribución normal, la variación con el número de contornos supone pasar de una probabilidad de no modificar prosódicamente la unidad de un 50 % a un 64 %, con

el umbral de 5 Hz ya mencionado. Por comparación, con el modelo entonativo antiguo dicha probabilidad era de un 20 %. La diferencia de duración se mantiene prácticamente independiente del número de contornos (8 ms de media del valor absoluto de la diferencia, y 110 de varianza), lo que probablemente se debe a que se le concede una menor importancia en el coste de objetivo. Otros datos que se mantienen constantes son el número medio de semifonemas consecutivos en la grabación original que se escogen para la síntesis (2,30), y el porcentaje de uniones entre semifonemas no consecutivos en el corpus por la zona estacionaria (98 %).

			Número de contornos		
		Ant	1	5	20
C_{con}	M	0.48	0.37	0.36	0.35
	Var	0.86	0.42	0.41	0.39
C_{cont}	M	2.85	2.70	2.68	2.67
	Var	0.42	0.38	0.37	0.37
C_{pros}	M	6.61	3.93	3.52	3.27
	Var	96.12	50.60	41.36	36.04
C_{obj}	M	3.61	2.94	2.85	2.79
	Var	4.06	2.29	1.91	1.70

Cuadro 6: Variación con el número de contornos considerados (modelo contextual basado en redes neuronales)

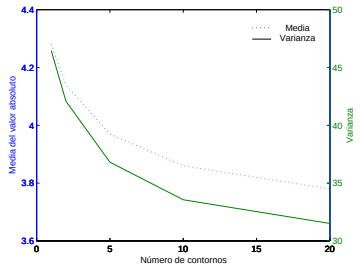

Figura 4: Evolución del error de frecuencia fundamental con el número de contornos

Cabe destacar que el nuevo método basado en selección de unidades entonativas obtiene mejores resultados que el de la versión anterior del conversor, incluso cuando se considera un único contorno. Esto refrenda la idea de que emplear los grupos acentuales del propio corpus de voz permite realizar una selección de unidades acústicas más eficiente, al mismo tiempo que se extrae más provecho de los recursos disponibles. En cuanto a la variación del número de contornos, se puede observar como las medias y varianzas van mejorando a medida que aumenta su número. El hecho de que las varianzas disminuyan muestra un comportamiento más regular del algoritmo de selección, lo que ayuda a mitigar en parte el problema de la síntesis por corpus, que alterna partes de calidad muy elevada con pequeños fallos claramente audibles.

6. *Pruebas subjetivas*

Para las pruebas de calidad subjetivas se recurrió a las típicas comparaciones de pares, en las que se le pide al usuario que dé una puntuación relativa a dos versiones sintéticas de una misma frase. En nuestro caso, se empleó la clasificación mostrada en la tabla 7. Dado que se quería conocer la opinión de gente no habituada a trabajar con conversores de voz, se decidió, por sencillez, que las pruebas fuesen de aceptación general.

1	Versión A mucho mejor que versión B
2	Versión A mejor que versión B
3	Iguales
4	Versión B mejor que versión A
5	Versión B mucho mejor que versión A

Cuadro 7: Puntuaciones de la comparación de pares

A continuación se describen las pruebas realizadas:

- Prueba 1:
  - Sistema A: Cotovía con el modelo entonativo antiguo, con modelo contextual de redes neuronales.
  - Sistema B: Cotovía con selección combinada de unidades entonativas y acústicas, modelo contextual de redes neuronales y un único contorno de frecuencia fundamental.
- Prueba 2:
  - Sistema A: Cotovía con selección combinada de unidades, con las funciones de coste entrenadas a mano, y un único contorno posible para la búsqueda acústica.
  - Sistema B: el mismo de la prueba 1.
- Prueba 3:
  - Sistema A: Cotovía con selección combinada de unidades, con las funciones de coste acústicas originales, en las cuales sólo se consideraban los subcostes prosódicos de frecuencia y duración, sin ningún tipo de umbral de percepción, ni índices de variación espectral, y con un coste de inteligibilidad que penalizaba a aquellas unidades cuyos fonemas por la

derecha e izquierda no eran exactamente los mismos que los de la unidad objetivo. Un único contorno posible de frecuencia fundamental.

- Sistema B: el mismo sistema B de la prueba 1.

- Prueba 4:
  - Sistema A: el mismo sistema B de la prueba 1.
  - Sistema B: Cotovía con selección combinada de unidades, con las funciones entrenadas mediante redes neuronales. 20 contornos posibles de frecuencia fundamental.

También se realizó una prueba comparando los modelos contextuales basados en regresión lineal y redes neuronales, pero no se apreciaron grandes diferencias. En las tres primeras pruebas se utilizaron cinco frases sencillas extraídas aleatoriamente de un texto periodístico. Para comprobar la influencia de la variación en el número de contornos entonativos se recurrió a un conjunto de frases diferente. El efecto de considerar un mayor número de contornos se suele apreciar en la corrección de pequeños errores aislados en las frases, así como en un aumento de la variabilidad entonativa que se hace patente al utilizar el sintetizador de forma continuada. Ello supone que las posibles mejoras no tengan por qué ser apreciables en cualquier frase, ya que podría darse el caso de que el contorno escogido fuese el mismo, o que contornos diferentes produjesen versiones sintéticas muy similares. Por esto, los autores seleccionaron un conjunto de frases extraídas de varios párrafos sintetizados variando el número de contornos entonativos, con el único criterio de que hubiese diferencias claramente reconocibles entre ambas versiones. Evidentemente, los resultados a los que se llegue deberán ser interpretados teniendo en cuenta esta forma de escoger el material de prueba.

En la tabla 8 se muestran las puntuaciones obtenidas considerando todos los evaluadores y sólo los habituados a trabajar en el ámbito de las tecnologías de voz, según la escala de la tabla 7. En total intervinieron 21 personas del ámbito universitario, de las cuales 9 fueron incluidas en el grupo de expertos por trabajar en síntesis o reconocimiento de voz.

A partir de estos resultados se pueden hacer las siguientes consideraciones:

	Evaluadores		
Prueba	Expertos	No Expertos	Todos
1	4.36	4.30	4.33
2	3.36	3.25	3.30
3	3.81	3.31	3.56
4	4.33	3.27	3.56

Cuadro 8: Resultados de las pruebas subjetivas

- Prueba 1:

  Queda patente la mejora que supone el nuevo modelo entonativo para la calidad global del habla sintética, incluso considerando un único contorno de frecuencia fundamental posible, lo cual concuerda con los resultados objetivos recogidos en la tabla 6.

- Prueba 2:

  Aunque, en principio, las pruebas subjetivas reflejan que las funciones entrenadas obtienen una calificación tan sólo de "ligeramente mejor", lo cual podría considerarse bastante decepcionante, no se trata de un mal resultado, dado que la comparación se realiza con unas funciones ajustadas manualmente a lo largo de las constantes pruebas de la frase de desarrollo del módulo de selección de unidades. Desde este punto de vista, el método de entrenamiento de los pesos del coste contextual descrito en el apartado 2.1 proporciona un conjunto de pesos automáticamente en muy poco tiempo y que genera una voz sintética de mayor calidad.

- Prueba 3:

  Hay una clara preferencia por el conjunto de frases sintetizadas a partir de las funciones de coste acústicas actuales. Esto demuestra que la caracterización del contexto fonético mediante los centroides de los vectores mel-cepstrum de los fonemas circundantes es bastante más apropiada que la simple consideración de los fonemas directamente, además de la adecuación de los índices de variación espectral.

- Prueba 4:

  Es en este caso en el que se aprecian más diferencias en las opiniones de los dos grupos de evaluadores. En el grupo de expertos se observa una clarísima preferencia (4.33) por las frases sintetizadas considerando 20 contornos alternativos de frecuencia fundamental, mientras

que en el otro grupo la diferencia es mucho menos marcada (3.27). Esto se debe probablemente a la mayor complejidad de las frases empleadas, que provoca la aparición de fallos no presentes en las pruebas anteriores, como los del módulo de inserción de pausas. Mientras que los evaluadores expertos fueron capaces de reconocer dichas fuentes de error en las dos versiones de cada frase y, por lo tanto, abstraerse de ellos para optar por una puntuación u otra, probablemente los no expertos acabaron realizando valoraciones más confusas al encontrarse con frases que tenían fallos en ambas versiones. De todas formas, no conviene olvidar que esta prueba se efectuó con frases especialmente seleccionadas por el autor, por lo que no pretende reflejar la diferencia real existente al sintetizar dos frases cualesquiera con los dos métodos comparados. Además, al haberse limitado la prueba a frases aisladas, tampoco deja evidencia de la mayor variabilidad que aporta a la conversión de voz la consideración de múltiples contornos.

7. Conclusiones

Este artículo se ha dedicado a la evaluación, tanto objetiva como subjetiva, del sistema de conversión texto voz Cotovía. Para ello se comenzó con una breve exposición de las principales características del sistema, como el diseño de las funciones de coste, la estimación de la prosodia y la selección combinada de unidades acústicas y entonativas.

Como el conversor se encuadra dentro de las técnicas de síntesis basadas en selección, tanto en lo referente a unidades acústicas como entonativas, en las que se genera la voz mediante la aplicación de unidades extraídas de contextos similares, las pruebas objetivas se encaminaron hacia la comprobación de en qué medida se escogían unidades con las características deseadas. Análogamente, se estudió la influencia en la selección de considerar un número mayor de contornos entonativos, llegando al resultado de que los errores cometidos van disminuyendo, al igual que sus varianzas.

En cuanto a las pruebas subjetivas, se recurrió a la comparación de pares, por su mayor sencillez de cara a los evaluadores, y porque aportan información para escoger los diseños más adecuados para las partes del sistema estudiadas. En este caso, se realizaron 5 pruebas: comparación del modelo entonativo nuevo con respecto al original, validez del método de entrenamiento de los pesos de las funciones de coste acústicas, comparación de dos configuraciones diferentes de dichas funciones de coste, comparación de los modelos contextuales basados en regresión lineal y redes neuronales y, por último, influencia de la variación del número de contornos alternativos de frecuencia fundamental considerados.

En http://www.gts.tsc.uvigo.es/cotovia se puede comprobar la calidad del sintetizador.

Bibliografía

Banga, E. R., F. Campillo, E. F. Rei, y F. Méndez. 2002. Sistema de conversión texto-voz en lengua gallega basado en selección combinada de unidades acústicas y prosódicas. *Procesamiento del lenguaje natural*, (29):153–158.

Black, A. y N. Campbell. 1995. Optimising selection of units from speech databases for concatenative synthesis. En *Actas de Eurospeech*, volumen 1, páginas 581–584, Madrid, España.

Black, A. y P. Taylor. 1997. Automatically clustering similar units for unit selection in speech synthesis. En *Actas de Eurospeech*, volumen 2, páginas 601–604.

Campillo, F. y E. R. Banga. 2002. Combined prosody and unit selections for Corpus-based text-to-speech systems. En *Actas de ICSLP*, volumen 1, páginas 141–144, Denver.

Campillo, F. y E. R. Banga. 2003. On the selection of the cost functions for a unit selection speech synthesis. En *Actas de Eurospeech*, volumen 1, páginas 289–292.

Campillo, F. y E. R. Banga. 2004. Diseño de la función de coste de concatenación en síntesis de voz basada en Corpus. En *Actas de URSI*, Barcelona.

Febrer, A. 2001. *Síntesi de la parla per concatenació basada en la selecció*. Ph.D. tesis, Universidad Politécnica de Calatuña.

Hunt, A. y A. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. En *Actas de ICASSP*, volumen 1, páginas 373–376.