SÍNTESIS DE HABLA EMOCIONAL POR SELECCIÓN DE UNIDADES

Ignasi Esquerra

Centre de Recerca TALP
Departamento de Teoría de la Señal y Comunicaciones
Universitat Politècnica de Catalunya

RESUMEN

Una de las áreas de investigación en síntesis de habla más en boga actualmente es la síntesis de emociones y estilos de habla diferenciados del estilo por defecto habitual de los sistemas TTS. Principalmente existen dos maneras de atacar este problema: modelando adecuadamente la prosodia (entonación, duración e intensidad), o utilizando sistemas por concatenación con unidades grabadas con otros estilos. En el trabajo que se presenta a continuación se investiga la importancia relativa de la prosodia y de las unidades para sintetizar emociones y estilos con un sistema TTS por selección de unidades. Los experimentos realizados muestran diferencias entre las emociones y estilos estudiados, siendo unos más dependientes de la prosodia y menos de las características de la voz de las unidades.

1. LA SÍNTESIS DE HABLA EXPRESIVA

Gracias a la técnica de síntesis por selección de unidades a partir de grandes bases de datos se ha conseguido mejorar de forma substancial la calidad de los sistemas de conversión texto-habla. Sin embargo, en la mayoría de casos las voces sintéticas suenan todavía en cierto modo aburridas, con un mismo estilo de habla estándar, en especial cuando se sintetizan textos relativamente largos. Para mejorar la naturalidad de la voz sintética se debería considerar la variabilidad existente en la comunicación oral entre humanos. Así por ejemplo, hablamos más rápido o más lentamente, más fuerte o más flojo, según las condiciones del entorno, de nuestros interlocutores o incluso de nuestro estado de ánimo. A veces estos cambios en el estilo del habla no se reflejan solamente modificando la velocidad o intensidad del habla, sino también variando la entonación, el grado de articulación de los sonidos o las características tímbricas de la voz.

La síntesis de habla emocional cabría englobarla dentro del marco más general de la síntesis de voz con estilos de habla no neutros. Actualmente hay quien utiliza el término síntesis de habla ”expresiva” para significar que el habla humana es muy rica en matices, y que las emociones solamente son una parte más de la gran variabilidad del habla humana [1]. Históricamente el estudio de las emociones se ha realizado desde ámbitos muy diversos, como la psicología o la fonética acústica, y no ha sido hasta hace poco que también se ha abordado desde el punto de vista de las tecnologías del habla.

Desde el inicio del desarrollo de sintetizadores de voz ha habido diferentes intentos de dotarlos de la capacidad para generar voces con emoción [2,3]. Muchos de los primeros intentos de sintetizar habla emocionada se realizaron sobre sistemas basados en reglas, y en particular con sistemas de síntesis por formantes [4,5]. Éstos tienen la ventaja de poder actuar directamente sobre muchos de los parámetros que definen la prosodia y los sonidos en el modelo de síntesis, si bien la inteligibilidad y naturalidad es comparativamente peor que la de los sistemas más modernos basados en concatenación de unidades.

Actualmente la mayoría de sistemas texto-habla utilizan la concatenación de segmentos de voz extraídos de grandes bases de datos, seleccionando el segmento más apropiado en cada ocasión. Este hecho provoca que la voz generada mantiene las características propias del locutor y en cierta medida del estilo en que fueron grabadas. Para poder generar otros tipos de habla, es necesario disponer de corpus más extensos que recojan las variantes de habla deseadas [6].

En el caso del español hay pocos trabajos documentados de análisis o síntesis de emociones, entre los cuales cabe destacar los de A.Rodríguez [7] y J.M.Montero [8]. A destacar también el inicio de trabajos para el euskera de E.Navas [9], y para el caso del catalán de I.Iriondo [10].

En el presente trabajo se presenta un experimento de síntesis de habla con emociones realizadas con un sistema por selección y concatenación de unidades. Se ha utilizado una base de datos grabada con diferentes emociones y estilos para generar distintas bases de unidades, con las cuales se han sintetizado unas frases con la prosodia por defecto del sistema o con la prosodia analizada de las frases reales. Esto nos proporciona información sobre como afecta la prosodia o las características acústicas de las voces en la expresión de las emociones sintéticas.

En las secciones que siguen se introduce en primer lugar el sistema TTS y la base de datos utilizada. A continuación se presenta el método de generación de las bases de unidades, para seguidamente describir los experimentos de síntesis y sus resultados. Se finaliza con una discusión de las conclusiones del trabajo.

2. EL SISTEMA DE CONVERSIÓN TEXTO-HABLA

Una de las áreas principales de trabajo en el Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla (TALP) es la conversión texto-habla, es decir la generación de voz sintética a partir de texto. Estos sistemas, que se conocen habitualmente con las siglas TTS (del inglés, *text-to-speech*), analizan y procesan los textos de entrada, los convierten en su correspondiente representación en sonidos, determinan cómo se deben pronunciar, y finalmente generan acústicamente dichos sonidos. El sistema desarrollado en la UPC, como la mayoría de sintetizadores actuales, genera la voz mediante la técnica de selección de unidades [11].

El primer módulo del sistema TTS consiste en el procesado lingüístico. Su misión es capturar el texto que se quiere leer y prepararlo para los módulos siguientes. En particular, se debe identificar las frases y oraciones, y convertir los dígitos, abreviaturas y otros signos en sus correspondientes transcripciones ortográficas. A continuación, mediante el uso de diccionarios o de reglas de trascripción el texto se convierte a fonemas. El siguiente módulo es el de generación de la prosodia, que proporciona los valores de frecuencia fundamental, duración e intensidad de los sonidos.

El módulo de generación de señal consiste en el algoritmo de selección de las unidades, y en las técnicas de procesado de señal para unirlas y modificarlas convenientemente. La síntesis por selección de unidades permite escoger de entre todos los segmentos de la base de datos la unidad que mejor se adecue a lo que determinan los módulos anteriores, tanto los valores de prosodia, factores fonéticos o fonológicos, o las características acústicas de los segmentos previamente seleccionados.

Todos estos módulos trabajan sobre una estructura de datos multicapa, donde las informaciones se guardan en diferentes niveles (por ejemplo, palabras, fonemas, unidades de síntesis) y cuyos valores se relacionan entre sí y con los demás niveles mediante vínculos del tipo contenido-continente.

3. LA BASE DE DATOS EMOCIONAL

Las señales de voz utilizadas en este trabajo son una parte de la base de datos INTERFACE en español [12]. Dos locutores, uno masculino y otro femenino, grabaron un corpus de frases, párrafos y palabras simulando las seis emociones estándar de MPEG4 además del estilo neutro. También se grabó parte del corpus en otros cuatro estilos de habla (tabla 1). Para las emociones principales se grabaron dos sesiones del corpus completo en días diferentes lo que representa unos 12 minutos de voz por locutor, sesión y emoción. Para los estilos de habla se dispone de unos 10 minutos correspondientes a una única sesión de grabación.

Las señales fueron grabadas a 32kHz, aunque después se redujo la frecuencia de muestreo a 16kHz para generar las bases de unidades para el conversor texto-habla. De forma simultanea se adquirió la señal de un laringógrafo, que se utilizó para la extracción de los instantes de periodicidad glotal necesarios para determinar la frecuencia fundamental.

Enfado (A)	Alto (H)
Asco (D)	Bajo (L)
Miedo (F)	Lento (W)
Alegría (J)	Rápido (Z)
Sorpresa (S)	
Tristeza (T)	
Neutro (N)	

Tabla 1. *Emociones y estilos de habla de la base.*

Con la ayuda del sistema de reconocimiento del habla de la UPC basado en modelos de Markov se procedió a segmentar de forma automática los ficheros de voz [13]. Debido a la existencia de importantes diferencias en la duración de los fonemas y la presencia de muchas pausas en las locuciones, se realizó una segmentación en dos fases: una primera para detectar la posición de las pausas reales y obtener una segmentación aproximada, y una segunda fase para ajustar mejor las fronteras entre fonemas. Se usaron modelos entrenados de forma conjunta para todos los estilos. No se realizó una verificación manual de los resultados de la segmentación, salvo en algunos casos muy concretos detectados al hacer posteriormente la síntesis.

Aunque la base de datos en su conjunto es relativamente grande, al repetirse las frases del corpus para cada emoción y estilo hace que el contenido en cuanto a número de unidades distintas sea pequeño. En la tabla 2, se compara el contenido de la base de datos emocional (IESSDB) con otra utilizada por el sistema TTS de mayor tamaño (ESMASE).

	IESSDB	ESMASE
palabras	1819	4411
fonemas	8681	22014
unidades	17362	44028
difonemas	494	539

Tabla 2. *Contenido de las bases de datos emocional y de síntesis.*

Se puede observar que la base IESSDB no está diseñada y suficientemente dimensionada como para cubrir todos los difonemas necesarios para la síntesis, aunque cabe decir que las unidades que faltan corresponden a combinaciones muy poco habituales entre fonemas y que, en caso de requerirse tal unidad, el sistema TTS las sustituye por unidades parecidas desde el punto de vista fonológico.

4. GENERACIÓN DE LAS BASES DE UNIDADES

Una de las ventajas, y a la vez inconveniente, de la síntesis basada en concatenación de unidades es que la voz sintética conserva las características del locutor o locutora que realizó la grabación de la base de datos oral. Si bien existen técnicas de procesado de señal (como por ejemplo TD-PSOLA) que permiten transformar las características temporales de las unidades que se concatenan, en general éstas no son suficientes para las grandes modificaciones prosódicas que se requieren para convertir una voz neutra en una voz emocionada. Por ejemplo, la frecuencia fundamental (F0) media de emociones como la alegría o la tristeza pueden estar bastante alejadas del valor medio normal del estilo neutro. En el caso de querer modificar además las características tímbricas de las unidades, sería necesario recurrir a otras técnicas de procesado de señal.

En un trabajo anterior [14] se vio cómo la frecuencia fundamental y las duraciones son factores importantes para reconocer las emociones. En particular se mostró cómo era posible transformar una frase prosódicamente neutra en una con emoción simplemente trasplantando su prosodia. Por otra parte, las frases con emoción a las cuales se les imponía un patrón de prosodia neutro continuaban siendo reconocibles por la calidad de la voz, aunque en menor medida.

4.1. Bases de unidades independientes por emoción

Mientras que en los sistemas basados en síntesis por formantes era relativamente fácil crear voces diferenciadas aplicando reglas y tablas de valores, en los sistemas basados en selección de unidades se necesita grabar bases diferentes para cada uno de los estilos de habla deseados [6].

Aprovechando que el sistema permite seleccionar las voces del sintetizador mediante comandos XML insertados en el texto de entrada, se pueden generar varias voces sintéticas del mismo locutor con diferentes estilos y conmutar entre ellas de la misma manera como se cambiaría de locutor (tabla 3).

<*SPEAKER="marta"*> Mientras encendía un cigarrillo <*SPEAKER="marta_surprise"*> me sorprendí a mí misma <*/SPEAKER*> pensando en cómo puede llegar cierta gente <*SPEAKER="marta_disgust"*> a ciertos puestos de responsabilidad <*/SPEAKER*> poseyendo una mentalidad a nivel de subcultura. <*/SPEAKER*>

Tabla 3. *Texto con selección de locutor/emoción.*

Partiendo de los ficheros de voz de la base IESSDB, se procedió a crear una base de unidades para cada emoción y locutor, y se incorporaron al sistema TTS como voces independientes. Es decir, cada locutor pasaba a tener hasta 11 voces diferentes seleccionables como se ha visto anteriormente. Al tener menos ficheros disponibles (entre 10 y 24 minutos), las bases de unidades resultantes son pequeñas y, aunque están presentes todas las unidades comunes del corpus, no existen muchas repeticiones de las mismas para conseguir una calidad óptima.

5. SÍNTESIS Y RESULTADOS

Se realizaron unas primeras pruebas de síntesis con las bases de unidades independientes para cada emoción. El sistema TTS está desarrollado con único módulo de prosodia que proporciona un estilo de habla neutro, por lo que se asigna una prosodia igual para todas las emociones. Aun así, la voz resultante suena bastante con la emoción o estilo de habla de la base de unidades de la cual proviene.

Por otra parte, los valores de duración y F0 de las unidades seleccionadas por el sintetizador pueden estar muy alejados de los valores "objetivo" determinados por el módulo de generación de la prosodia, en especial para aquellas emociones con grandes variaciones de estos valores como son la alegría o la tristeza. Las técnicas para adaptar la prosodia no siempre consiguen alcanzar dichos valores puesto que se establecen límites a los factores de modificación, y en consecuencia las señales sintetizadas conservan en buena medida los valores de F0 y duración de las frases con emoción. Por esta razón es posible que las frases sintetizadas no lleguen a conseguir los valores de la prosodia neutra determinada por el sistema.

Es curioso observar como, aunque las frases que se sintetizan formen parte del propio corpus, un 8% de las unidades seleccionadas provienen de otras frases. Esto es debido a que en estos casos predomina el factor de los valores del coste "objetivo" por encima del coste de "concatenación" en el proceso de selección de las unidades.

5.1. Síntesis con prosodia natural forzada

Para comprobar el efecto de la prosodia por defecto del sistema, se procedió a sintetizar unas frases con su prosodia real. Estas frases de prueba se eliminaron en la fase de generación de las bases de unidades. Al forzar la duración y frecuencia fundamental de las unidades a los valores naturales, se consiguen frases muy parecidas a las locuciones originales, tanto en prosodia como en calidad de la voz.

Se procedió igualmente a sintetizar dichas frases con el sistema con voces estándar y forzando su prosodia a los valores analizados de las frases reales con emoción. Comparándolas con las anteriores, estas suenan mucho más naturales aunque es difícil percibir las emociones por la prosodia. Esto se contradice en parte con los resultados del trabajo de investigación previo de resíntesis con trasplante de prosodias [14], donde se mostraba que era posible reconocer las emociones en frases neutras a las cuales se les había forzado una prosodia con emoción.

Una posible causa de esta diferencia puede ser el hecho que al trasplantar la prosodia se preservaban todas las pequeñas variaciones de entonación (microprosodia) mientras que al hacer la síntesis por concatenación el patrón de prosodia se reduce a un único valor por fonema, y esto puede ser insuficiente para caracterizar la evolución de la entonación en las emociones.

5.2. Evaluación

Con el fin de conocer mejor la importancia relativa de disponer de una buena prosodia o de unas buenas unidades con emoción, se procedió a realizar un test perceptivo. En él se evaluaba la calidad y el grado de reconocimiento de las emociones y estilos en las señales sintetizadas mediante los tres métodos de síntesis descritos hasta ahora. Estos se denominaran de la siguiente manera:

- Método ES: bases de unidades independientes por emoción, y prosodia por defecto del sistema.
- Método EN: bases de unidades independientes por emoción, y prosodia natural.
- Método DN: base de unidades por defecto del sistema, y prosodia natural.

El test se realizó a través de Internet con 25 voluntarios, de manera no supervisada. Consistía en una introducción explicativa y tres formularios. En el primero se presentaban 12 señales de audio y se pedía identificar la emoción entre un conjunto de 7 posibilidades, e indicar la calidad subjetiva según una escala tipo MOS de 5 valores. La segunda parte del test era similar a la anterior para los estilos de habla, y se presentaban 6 señales. En la tercera parte, se comparaban los métodos de dos en dos mediante un test tipo AB; se presentaban dos ficheros de audio junto con una indicación de la emoción o estilo al que correspondían, y se tenía que marcar cuál de los dos ficheros era el que mejor expresaba dicha emoción o bien si se consideraba que eran igualmente válidos.

Para las emociones, las matrices de confusión entre estímulo y respuesta muestran como para los métodos ES y EN no hay una preferencia clara hacia la emoción correcta respecto a las demás, aunque siempre es la obtiene la máxima valoración. Por el contrario, en el método DN las respuestas son mayoritariamente en el sentido de marcar la opción "neutral". Es decir, a pesar de haberse sintetizado con prosodia natural, ésta no es suficiente para expresar la emoción de forma convincente. Las tasas de reconocimiento de las emociones reflejan esta situación (tabla 4a) y corroboran las primeras impresiones comentadas anteriormente al principio de esta sección. Es curioso constatar que en algunos casos el método EN con prosodia natural da unos resultados peores que el método ES con la prosodia por defecto neutral.

En cuanto a los estilos de habla se observa que para el habla rápida o lenta se obtienen muy buenos resultados (tabla 4b) debido principalmente a que todas las unidades disponibles en las bases tienen una duración menor o mayor respectivamente al habla normal. No así, para los dos otros estilos los resultados son bastante malos. En particular se tiende a confundir el estilo "alto" por "rápido" y el "bajo" por "lento". Aquí cabe comentar que las etiquetas de estos dos estilos pudieron llevar a confusión entre los evaluadores puesto que no se especificó en las instrucciones si correspondían a altitud en entonación o en intensidad. De forma similar a las emociones, en el método DN se continua identificando mayoritariamente todas las señales como neutras.

	ES	EN	DN
A	68,8%	61,1%	0,0%
D	40,0%	29,4%	14,3%
F	35,3%	25,0%	0,0%
J	44,4%	36,8%	7,7%
S	33,3%	52,9%	0,0%
T	68,8%	86,7%	10,5%
N	83,3%	50,0%	78,6%

(a)

	ES	EN	DN
H	35,7%	46,2%	14,3%
L	20,0%	41,7%	25,0%
Z	100,0%	92,3%	53,8%
W	90,0%	93,3%	0,0%
N	33,3%	71,4%	81,8%

(b)

Tabla 4. *Resultados del test de reconocimiento de emociones (a) y estilos (b)*

Por lo que respecta a la valoración de la calidad, el método que recoge el valor MOS más alto es el DN, sin duda porque es el que utiliza una base de unidades más grande. En general, con el método EN se obtienen valores más uniformes para el conjunto de emociones, mientras que para el método ES se da un valor hay algún caso más dispares.

Finalmente, en la comparación de los métodos, hay una preferencia clara de los métodos con unidades con emoción (EN y ES) respecto al método por defecto (DN), excepto para el caso de la emoción "neutra" en que se prefiere el método DN porque además de llevar prosodia natural se ha sintetizado con una base de unidades mayor.

Respecto a la relación "prosodia sintética" versus "prosodia natural", se presuponía que el método EN daría mejores resultados que el método ES. Esto de verifica para la mayoría de emociones y estilos, con excepción de "asco" (D), "miedo" (F), "rápido" (Z) y

"lento" (W) donde la respuesta mayoritaria fue la indiferencia entre los dos métodos. Se debería investigar más a fondo si estas dos emociones tienen una caracterización basada en el rol de la duración.

Mediante un sistema de puntuación que asigna 3 puntos al método preferido, y un punto en caso de no preferencia por ninguno de los dos, se obtiene la siguiente gráfica (figura 1).

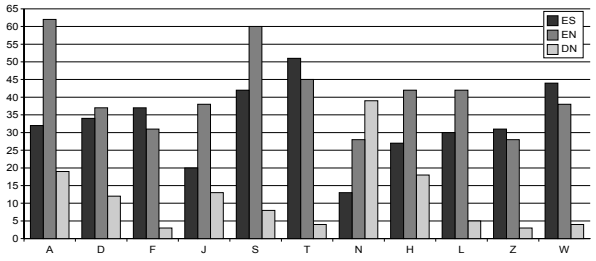

Figura 1. *Resultados de la comparación de métodos*

Para las emociones "enfado" (A), "alegría" (J), "sorpresa" (S), "alto" (H) y "bajo" (L) el método preferido es el de unidades con prosodia natural. Para los demás, no se aprecia mucha diferencia entre los métodos con prosodia natural o sintética, cosa que nos lleva a pensar que la expresión de las emociones se realiza principalmente por la calidad de la voz en lugar de por la prosodia.

6. CONCLUSIONES

Con los experimentos realizados se ha visto como es posible sintetizar emociones con bases de unidades provenientes de frases grabadas con un diferentes estilos de habla. Las tasas de reconocimiento todavía no son lo suficientemente altas como sería de desear, seguramente debido a que las bases utilizadas no son lo bastante grandes y, por otra parte, los modelos de prosodia todavía están incorporados en el sistema TTS.

Actualmente se está trabajando en la incorporación de la característica de emoción/estilo dentro de las etiquetas de selección de unidades. Esto permitirá generar una única base y escoger las unidades no sólo en función de sus características acústicas, fonológicas o prosódicas, sino también en base a parámetros lingüísticos más abstractos como es el estilo de habla. Así mismo se trabaja en el desarrollo de módulos prosódicos que modelen las diferentes emociones.

7. AGRADECIMIENTOS

El autor quiere expresar su gratitud a las personas del departamento y externas que han participado en las pruebas de evaluación. Este trabajo ha sido parcialmente financiado por la Unión Europea bajo el proyecto TC-STAR (http://www.tc-star.org/).

8. REFERENCIAS

[1] N. Campbell, *"Towards synthesising expressive speech; designing and collecting expressive speech data"*. Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH), 2003

[2] I.R. Murria, J. L. Arnott, *"Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion"*. *JASA*, vol.2, pp. 1097-1108, 1993

[3] M. Schröder, *"Emotional Speech Synthesis: a Review"*, Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH), pp. 561-564, 2001

[4] J.E. Cahn, *"Generating expression in synthesized speech"*, Master's thesis, Massachusets Institute of Technology, 1990

[5] I. R. Murray, J. L. Arnott, *"Implementation and testing of a system for producing emotion-by-rule in synthetic speech"*, *Speech Communication*, vol. 16, pp. 369-390, 1995

[6] A. W. Black, *"Unit Selection and Emotional Speech"*, Proc. of 8th European Conference on Speech Communication and Technology (EUROSPEECH), 2003

[7] A. Rodríguez-Bravo, P. Lázaro, N. Montoya, J. M. Blanco, D. Bernadas, J. M. Oliver, y L. Longhi, *"Modelización acústica de la expresión emocional en el español"*. *Procesamiento del Lenguaje Natural*, Revista SEPLN, pp. 159-166, 1999

[8] J.M. Montero, *"Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano"*, Tesis doctoral, Universidad Politécnica de Madrid, 2003

[9] E. Navas, I. Hernáez, I. Luengo, J. Sánchez, *"Análisis Acústico de una Base de Datos de Habla Emocional"*, III Jornadas en Tecnologías del Habla (Valencia), pp. 233-236, nov. 2004

[10] I. Iriondo, F. Alías, J. Melenchón, M.A. Llorca, *"Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis"*, Proc. Workshop on Affective Dialogue Systems, (Kloster Irsee, Germany), pp. 197-208, 2004

[11] A. Bonafonte, P.D. Agüero, J. Adell, J. Pérez, A. Moreno, *"OGMIOS: The UPC Text-to-Speech Synthesis System for Spoken Translation"*, Proc. of TC-STAR Workshop on Speech-to-Speech Translation (Barcelona), pp.199-204, Junio 2006

[12] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, A. Nogueiras, *"INTERFACE Databases: Design and collection of a multilingual emotional speech database"*, Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC), pp. 2024-2028, 2002

[13] J. Adell, A. Bonafonte, *"Towards phone segmentation for concatenative speech synthesis"*. Proc. of 5th ISCA Speech Synthesis Workshop (Pittsburgh), 2004

[14] I. Esquerra, A. Bonafonte, *"Habla emocional mediante métodos de re-síntesis y selección de unidades"*, Actas del XIX Simposium URSI (Barcelona), 2004