Habla emocional mediante métodos de re-síntesis y selección de unidades

Ignasi Esquerra, Antonio Bonafonte
Centro TALP – Dep. Teoría de la Señal y Comunicaciones
Universitat Politècnica de Catalunya
e-mail: ignasi@talp.upc.es, antonio@talp.upc.es

***Abstract*— The topic of emotional speech synthesis has received lately a lot of interest within the research community, as shown by the number of papers presented at conferences and workshops. In the beginning, synthesizers used rules to perform prosodic and voice quality changes in order to produce different styles of speaking to the synthetic voice. Nowadays, most of the TTS systems use unit-selection methods which provide high quality voices, but they rarely allow a modification of the acoustic and prosodic features of the segments. The present paper investigates different ways to generate emotional speech with TTS systems. The first is an experiment of copy-synthesis which takes the prosody of an emotional sentence and copies it onto the same sentence with a neutral-speaking style. Next, a set of unit databases are build for each emotion, and converted to virtual different voices for the TTS system. Using XML commands the text can be edited in order to provide different emotions to specific parts of the text. Both experiments show that the generation of emotional synthetic speech is possible and that the combination of prosodic modelling and signal processing techniques would be the key to improve TTS systems quality.**

I. SÍNTESIS DE HABLA EXPRESIVA

En estos últimos años se han conseguido avances importantes en la calidad de los sistemas de síntesis de voz, en especial gracias a las nuevas técnicas por selección de unidades. Existen sistemas con un alto grado de inteligibilidad y de naturalidad aunque sus voces suenan algo monótonas. Desde el inicio del desarrollo de los sintetizadores de voz ha habido diferentes intentos de dotarlos de la capacidad de generar voz sintética con una cierta emoción. Sin embargo, últimamente este tema de investigación ha retomado impulso y se ha convertido en uno de los más populares como se comprueba por las comunicaciones presentadas en recientes congresos y workshops.

La mejora en la naturalidad pasa sobretodo por añadir variabilidad a la voz sintética de manera que se tengan en cuenta las diferentes situaciones o contextos en el que se produce la comunicación oral. Así, los seres humanos hablamos más rápido o más lentamente en función del grado de conocimiento del idioma, o hablamos más fuerte o más flojo según las condiciones acústicas del entorno. 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. El estado anímico o emocional de los locutores también afecta a la manera de hablar y añaden otra dimensión más a la gran variabilidad presente en el habla humana [1].

Históricamente el estudio de las emociones se ha realizado desde ámbitos muy diversos, como la fonética acústica o la psicología, y no ha sido hasta hace poco que también se ha abordado desde el punto de vista de las tecnologías del habla. Esto hace que los trabajos de investigación sean muchas veces muy parciales y no se contemplen de forma pluridisciplinar [2]. La síntesis de de habla emocional cabría englobarla dentro del marco más general de la síntesis con estilos de habla no neutrales. Actualmente se tiende a sustituir el término de síntesis de ”emociones” por el de síntesis de habla ”expressiva”, o incluso ”afectiva” [3]. Con esto se quiere expresar, valga la redundancia, que el habla humana es muy rica en matices, y que las emociones solamente son una parte de la gran variabilidad del habla.

En general la síntesis con estilos de habla no neutrales se basa en la modificación de reglas o de las unidades con las que se genera la voz sintética. Conviene actuar sobre varios parámetros de la voz, algunos relacionados con la prosodia (entonación, duración de los segmentos, localización de las pausas, etc.), otros relacionados con las características espectrales de la voz (harmonicidad, brillantez, forma del pulso glotal, etc.) [4].

Muchos de los primeros intentos de síntetizar habla emocionada se realizaron sobre sistemas basados en reglas, y en particular con síntesis por formantes [5], [6]. Si bien la inteligibilidad y naturalidad de estos sistemas es comparativamente peor a los sistemas basados en concatenación de unidades, éstos tienen la ventaja de poder actuar directamente sobre muchos de los parámetros que definen la prosodia y la excitación en el modelo de síntesis. Mediante variaciones de las reglas, o con nuevas reglas, se pueden definir otros estilos de habla o tipos de voces.

Actualmente la mayoría de sistemas de conversión texto-habla utilizan la concatenación de unidades previamente grabadas, extraídas de grandes bases de datos y seleccionando el segmento más apropiado para cada ocasión. Esto ocasiona que la voz que se genera tiene las características propias del locutor y del estilo en que fueron grabadas, normalmente en un modo de habla normal. Para poder generar otros tipos de habla, es necesario disponer de corpus más extensos que recojan todas las variantes de habla deseadas.

En este trabajo se presentan diversos proyectos de investigación entorno a la conversión texto-voz mediante síntesis por selección de unidades. En primer lugar, se presenta un estudio sobre la viabilidad de sintetizar emociones modificando solamente las características de entonación y duración. Con esto se consiguió valorar la importancia de las componentes no prosódicas de la voz en la producción oral de las emociones. A continuación, se describe el desarrollo de una serie de bases de unidades específicas de emociones, y el resultado de sintetizar con ellas mediante el conversor texto-voz. Finalmente, se concluye el presente artículo con un resumen de las conclusiones más relevantes.

II. La base de datos

Las señales de voz utilizadas en los diferentes proyectos presentados son una parte de la base datos INTERFACE para el español [7]. Dos locutores, uno masculino y otro femenino, grabaron un corpus de frases, párrafos y palabras simulando las seis emociones recogidas en MPEG4, además del neutro y de cuatro estilos de habla (Tabla 1). Para las emociones principales se grabaron dos sesiones del corpus completo en días diferentes. Se dispone de unos 12 minutos de voz para cada emoción (y sesión). 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-voz. De forma simultanea se grabó la señal de un laringógrafo, la cual se utilizó para la extracción de los instantes de periodicidad glotal.

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

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

Con la ayuda del sistema de reconocimiento del habla mediante modelos de Markov de nuestro departamento se procedió a segmentar automáticamente los ficheros de voz. Debido a la existencia de importantes variaciones en la duración de los ficheros, por la diferencia de velocidades del habla y sobretodo por la presencia de pausas según de qué emoción se trate, se tuvo que realizar una verificación manual de las marcas de segmentación.

III. Conversión por re-síntesis

Como primera aproximación a la síntesis de habla emocional, se llevó a cabo un experimento de transformación de prosodias. Mediante un procedimiento de re-síntesis se modificaron un conjunto de frases pronunciadas de forma neutra y con emoción para obtener cuatro tipos de transformaciones según la procedencia de prosodia y muestras de señal. El objetivo principal del experimento era el de verificar hasta qué punto es importante la prosodia, las características acústicas no prosódicas de la señal, o ambas, para transmitir las emociones en la voz.

Para cada una de las transformaciones se requieren dos señales de voz: una de la cual servirá para extraer los segmentos de señal (”ficheroN”) y la otra para la prosodia (”ficheroE”). Esta denominación se corresponde con el experimento principal que se describe más adelante. A parte de los ficheros de señal, son necesarias las marcas de segmentación que permiten el alineado entre frases. El proceso bàsico consiste en el transplante de la prosodia del ”ficheroE” sobre las muestras de señal del ”ficheroN” (Fig.1). Después de alinear temporalmente cada par de ficheros, se obtuvieron sendas funciones de transformación de la F0 y de la duración para el posterior proceso de re-síntesis mediante la técnica TD-PSOLA (con el programa Praat).

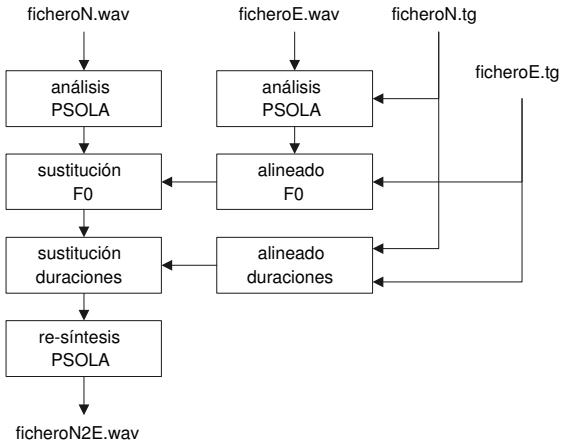

Fig.1 Esquema del proceso de transformación de prosodias

Se prepararon cuatro tests de dificultad creciente, según el grado de modificación sufrida por la señal, correspondientes a las cuatro posibles combinaciones de señal/prosodia:

- **Test N**: se utiliza la misma señal para ambos ficheros. La transformación consiste simplemente en un proceso de descomposición y reconstrucción.
- **Test E2E**: dos ficheros distintos con la misma emoción. Se produce un transplante de prosodias entre ficheros prosodicamente similares.
- **Test N2E**: transformación de neutro a emoción. El fichero de señal (neutro) y el fichero de prosodia (emoción) contienen emociones diferentes.
- **Test E2N**: transformación inversa, de emoción a neutro. El fichero de señal tiene emoción, mientras que el fichero de prosodia es neutro.

El primer y segundo test simplemente tenían como finalidad la verificación de que el procedimiento de descomposición de la señal y prosodia, y posterior re-síntesis funcionaban

correctamente. En el primer caso, se utiliza la misma señal para ambos ficheros de entrada, mientras que en el segundo test, les ficheros son diferentes aunque contienen la misma emoción. En este caso, se verifica que el alineamento entre ficheros se hace correctamente, siendo la transformación de prosodia pequeña. En general se utilizó los ficheros de la primera sesión para el "ficheroN" y los de la segunda sesión para el "ficheroE".

El test N2E constituye el experimento principal de este apartado. Fue concebido para mostrar la importancia de la prosodia en la percepción de una determinada emoción. Las muestras de señal provienen de frases neutras, mientras que la entonación y duración de los segmentos viene determinadas por la frase con emoción. Este test es por lo tanto un intento de "emocionalizar" una señal neutra forzando la entonación y duraciones de los fonemas a unos valores analizados de una frase con emoción. En cambio, el último test es el proceso contrario puesto que las muestras de señal provienen de la frase con emoción a la cual se le superpone una prosodia neutra. De alguna manera se está "desemocionalizando" prosódicamente la frase de entrada.

IV. Resultados de la transformación de prosodias

Veinte frases fueron procesadas de acuerdo con los cuatro tipos de transformación comentados y para cada una de las siete emociones. Los ficheros de voz resintetizados se presentaron al azar en un formulario web, con la petición a los participantes en el proceso de evaluación que identificaran las emociones que escuchaban. Los tests se presentaban en el mismo orden que se han descrito, a priori de mejor calidad a menor. Se trataba de un ejercicio de respuesta cerrada, es decir, que tenían que asignar una emoción de manera excluyente para cada test. Por ejemplo, para el primer test disponían de siete ficheros de audio, cada uno de ellos con una lista para seleccionar una entre las siete emociones. Podían escuchar tantas veces como quisieran las señales hasta completar las respuestas. Una macro se encargaba de verificar que no hubieran respuestas repetidas.

Los resultados para los tests N2E y E2N se presentan en forma de matrices de confusión, junto con la tasa de reconocimiento de la respuesta correcta (Tabla 2). No se muestran los resultados para los tests N y E2E cuyos valores son superiores al 80 %.

Como se puede observar, en todos los casos se supera el nivel de acierto al azar (14 %), excepto para la emoción "sorpresa"(S) en el test E2N. No se pueden apreciar pares de confusión claros entre emociones, seguramente debido a las pocas respuestas recibidas, aunque son parecidos en ambos tests. Las tasas más elevadas en el test N2E podrían indicar que la prosodia es más relevante que la propia señal a la hora de transmitir una determinada emoción. Aunque esto también podría ser debido al mecanismo, más o menos estereotipado, utilizado por los actores para simularlas.

Test N2E

	A	D	F	J	S	T	N	Rec.
A	13	2	2	1	1	0	1	65 %
D	0	12	1	2	3	2	0	60 %
F	3	0	14	1	0	1	1	70 %
J	3	1	2	10	4	0	0	50 %
S	0	2	0	6	12	0	0	60 %
T	0	2	1	0	0	17	0	85 %
N	1	1	0	0	0	0	18	90 %

Test E2N

	A	D	F	J	S	T	N	Rec.
A	8	3	2	2	3	0	2	40 %
D	4	7	3	0	4	2	0	35 %
F	3	3	6	2	5	0	1	30 %
J	0	0	0	13	4	0	3	65 %
S	4	5	4	3	2	1	1	10 %
T	0	0	3	0	0	16	1	80 %
N	1	2	2	0	2	1	12	60 %

Tabla 2. Resultados para los tests N2E y E2N

Se puede razonar igualmente que la calidad tímbrica de la voz que acompaña a las diferentes emociones también tiene su importancia, puesto que incluso transformando la prosodia a neutro todavía conserva la señal resintetizada rastros de la emoción original. Si no fuera así, los resultados para el test E2N habrían sido mucho más dispersos puesto que habría sido más difícil identificar la emoción.

En resumen, este experimento demuestra que es posible mediante transformación de la prosodia generar emociones a partir de habla neutra, tal como se tiene en las bases de unidades actuales en los sistemas de conversión texto-voz. Sin embargo, esto no es suficiente para conseguir una naturalidad total del habla sintética con emoción, ya que también es necesario tener en cuenta los cambios sufridos por el habla que se manifiestan en las características acústicas de la voz. Además, hay las limitaciones propias de las técnicas de procesado de señal para la transformación de la prosodia, que en mayor o menor medida introducen alguna distorsión en la señal.

V. Una voz para cada emoción

Uno de los problemas de la síntesis basada en concatenación de unidades es que la voz generada conserva las características del locutor o locutora que realizó la grabación. Aunque las técnicas de procesado de señal (por ejemplo TD-PSOLA o HNM) permiten transformar las características prosódicas de las unidades que se concatenan, en cualquier caso no son suficientes para las significativas modificaciones que se requieren para transformar una voz prosódicamente neutra en una voz emocionada.

La síntesis por selección de unidades (también llamada por corpus) permite seleccionar de entre todas las posibles unidades de la base la que mejor se adecue a los valores objetivo y al contexto. Inicialmente se consideraron sólo valores de prosodia, como la F0 o la duración, aunque más adelante se han ido incorporando otros factores fonéticos o fonológicos en las funciones de coste.

Aprovechando que el corpus oral disponible contiene un número elevado de frases, se ha generado una serie de bases de unidades específicas para cada emoción o estilo. Cada una de ellas se ha identificado con código compuesto por el nombre locutor y el tipo de emoción, y se ha incorporado al conversor texto-voz. El actual sistema de la UPC permite seleccionar entre diferentes voces mediante comandos XML (tabla 3). Editando el texto de entrada se puede definir la emoción con la que se sintetizará un trozo de texto, incluso palabra a palabra.

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

Tabla 3. Ejemplo de texto con selección de locutor/emoción

La audición de las frases sintetizadas permite identificar en la mayoría de casos la emoción a partir de la cual se ha generado. Se debe tener en cuenta que el módulo de prosodia proporciona una entonación y duraciones de estilo neutral puesto que todavía no se han desarrollado modelos de prosodia para las emociones. Aún así, las características acústicas de los segmentos son suficientes para aportar la sensación emocional en la voz sintética. Para los estilos de habla, sin embargo, este hecho es menos perceptible al poderse modelar como un cambio en la velocidad o en la intensidad del habla.

VI. Conclusiones

Varios grupos de investigación en todo el mundo estan realizando estudios y experimentos en la síntesis de habla con emociones. Algunos parten del análisis de bases de datos orales de emociones (simuladas o espontáneas) para generar modelos de entonación y duración, los cuales son introducidos después en el sistema de conversión texto-voz. Otros utilizan estas bases de datos para obtener bases de unidades con segmentos clasificados según sus características emocionales (además de las fonéticas y fonológicas). En nuestro grupo de investigación se ha iniciado una línea de trabajo en ambos sentidos; por una parte, en el estudio y modelado prosódico de las emociones (cuyos resultados esperamos poder publicar en breve), y por otra, en el desarrollo de bases de unidades y posterior síntesis por selección.

Por lo que respecta al trabajo que se presenta en este artículo, se han descrito dos experimentos de transformación de prosodias y síntesis por concatenación con el objetivo de generar voz sintética con características emocionales. En el primero de ellos, se ha evaluado cómo influye la prosodia en la percepción de las emociones mediante unos experimentos de transplante de prosodia entre señales neutras y señales con emoción. Los resultados no son concluyentes en cuanto a que no se puede derivar que la prosodia sea el único factor que contiene la información de las emociones, puesto que la calidad tímbrica (espectro) de la señal también influye. Sin embargo, también es cierto que una adecuada entonación, duración de los segmentos y pausas ayuda bastante a la identificación de las características emocionales. Así pues, cabe concluir que ambas componentes, prosodia y señal, son importantes a la hora de sintetizar habla emocional.

En referencia a la síntesis por selección, se han desarrollado un conjunto de bases de unidades específicas para cada locutor y emoción. En el sistema de conversión texto-voz se tratan como voces independientes, aunque fueran grabadas por la misma persona. La síntesis se realiza con un modelo de prosodia neutro, puesto que todavía no se ha desarrollado el módulo de generación de prosodia para las diferentes emociones y estilos de habla. Sin embargo, los resultados preliminares muestran que es posible este tipo de síntesis y que las emociones son fácilmente reconocibles.

Actualmente se está trabajando en el análisis y caracterización de la base de datos con el fin de obtener modelos de prosodia para cada una de las emociones. Asímismo, se está desarrollando una única base de unidades que contenga el conjunto de segmentos de la base oral, la cual será etiquetada según el estilo de habla. Mediante una adecuada selección de los pesos y funciones de coste, el conversor texto-voz dispondrá de un mayor número de unidades candidatas para la concatenación, con lo que seguro que mejorará la calidad de la síntesis.

Agradecimientos

Este trabajo ha sido parcialmente financiado por la Unión Europea bajo el contrato FP6-506738 (TC-STAR project, http://www.tc-star.org/) y por el Ministerio de Ciencia y Tecnología como proyecto coordinado TIC2002-04447-C02 (proyecto ALIADO, http://gps-tsc.upc.es/veu/aliado/).

Referencias

[1] Iain R. Murray and John L. Arnott. Synthezing emotions in speech: Is it time to get excited? *International Conference on Spoken Language Processing (ICSLP'96), Philadelphia (USA)*, 1996

[2] Mark Schröder. Emotional speech synthesis: A review. *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH'01), Aalborg (Denmark)*, 2001

[3] Nick Campbell. Towards synthesising expressive speech; designing and collecting expressive speech data. En *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH'2003), Geneva (Switzerland)*, 2003.

[4] Ailbhe Ní Chasaide and Christer Gobl. *Voice Quality and the Synthesis of Affect*. "Improvements in Speech Synthesis", E.Keller et al.(eds.) Wiley&Sons, 2002

[5] Janet E. Cahn. *Generating Expression in Synthesized Speech*. Masters Thesis, Massachusets Institute of Technology, 1990

[6] Iain R. Murray and John L. Arnott. *Implementation and testing of a system for producing emotion-by-rule in synthetic speech*. Speech Communication 16:369–390, 1995

[7] V.Hozjan, Z.Kacic, A.Moreno, A.Bonafonte, A.Nogueiras. Interface databases: Design and collection of a multilingual emotional speech database. *Proceedings of the Third Int. Conference on Language Resources and Evaluation (LREC'02), Las Palmas de Gran Canaria (España)*, 2002