Elvis resucita

¿Nunca te has engominado el tupé y has imitado a Elvis frente al espejo, lamentando que tu voz no se parezca a la del Rey? Ahora, con proyectos como el que un equipo del MTG desarrolló en Barcelona, lo tienes más fácil

Elvis Presley nació en una familia de granjeros humildes en la ciudad de Tupelo, Mississipi, en enero de 1935, y murió en agosto de 1977 en Graceland, habiendo sido coronado como el rey del rock'n'roll. El mito sobrevivió y aparecieron miles de imitadores que se dispersaron por todo el mundo copiando su manera de peinarse, de vestirse, de mover la pelvis, y lo más importante, también su manera de cantar. Japón, por motivos poco estudiados, concentra gran cantidad de estos imitadores. En la ciudad de Tokyo, por ejemplo, Mori Yasumasa (conocido también como J'Elvis) sube habitualmente al escenario a imitar a su cantante favorito.

Sin movernos de país, aunque veinte años atrás, en una ciudad llamada Kobe sucedió que el propietario de un bar, ante la noticia de que el guitarrista de la banda que venía a tocar aquella noche había enfermado, decidió sustituir al músico por algunas cintas grabadas que tenía en el bar. El cantante de la banda actuó esa noche acompañado de esas cintas y el público se entusiasmó con la idea. Poco más tarde el karaoke arrasó Japón.

Es a partir del conocimiento de estas dos historias que uno puede empezar a entender por qué Yamaha, una de las empresas con mayor dominio del sector del karaoke en Japón, decidió hace cosa de cuatro años iniciar un proyecto sobre transformación de la voz cantada para que los usuarios de sus máquinas de karaoke pudieran imitar la voz y la expresión de sus cantantes favoritos con ayuda de las nuevas tecnologías. Lo que no deja sorprender en esta historia es que la empresa nipona encargara el proyecto a un grupo de investigación y desarrollo de tecnología musical ubicado en Barcelona (ver cuadro "El Grupo de Tecnología Musical (MTG) de la Universitat Pompeu Fabra").

El objetivo del proyecto Elvis era desarrollar un sistema de transformación de voz pensado para integrarse en los sistemas comerciales ya existentes de karaoke que permitiera al usuario convertir su voz en la de su cantante favorito. La declaración de tal propósito levanta muchas dudas sobre cuáles son las posibilidades del sistema. Por eso quizás lo mejor sea hacer un repaso a las preguntas más frecuentes que se han hecho al respecto.

¿Puedo cantar *Satisfaction* tal y como lo hubiera hecho Bing Crosby?

No. Para poder imitar la voz y expresión de un cantante son condición indispensable dos cosas. La primera es disponer de una grabación del cantante al que se quiere imitar interpretando la canción que el usuario del karaoke haya elegido. La segunda es que esta grabación debe ser "seca", es decir, sin ningún tipo de efecto (reverberación, compresión...) ni acompañamiento musical. Así pues, para poder hacerlo se debería disponer de una grabación totalmente seca de Bing Crosby interpretando Satisfaction.

¿Puedo cantar *Yesterday* como Paul McCartney?

Si consigues una grabación seca de su voz, sin guitarras ni violines, entonces sí puedes. Una opción es contactar con la

La interfaz del programa Elvis. El texto corresponde a la canción (en japonés, como habrás podido apreciar) que se está interpretando. A la derecha, el Elvis animado reacciona con sonrisas o muecas de tristeza según lo bien o mal que cantes.

Ilustración: Miquel Martín

discográfica y pedirles el master. Otra posibilidad, más real, es confiar en las nuevas técnicas de análisis de audio para, a partir de la mezcla final, separar cada uno de los instrumentos, obtener así la pista de voz y des-reverberarla si es necesario. No obstante, dado el estado actual de la cuestión, lo mejor que puedes hacer es hibernar una temporada y confiar que cuando te despiertes ya se habrá inventado un des-reverberador y un separador de pistas profesional.

¿Entonces por qué dices que podría cantar como mi cantante favorito?
No tenemos grabaciones de la voz seca y sin acompañamiento de, por ejemplo, Frank Sinatra. Pero sí tenemos este tipo de grabaciones tomadas de la voz de un buen imitador de Frank Sinatra. Así pues, en realidad, en una definición más precisa del sistema, deberíamos decir de éste que es capaz de hacerte cantar como el imitador de tu cantante favorito.

¿Y qué pasa si un chico quiere cantar como Aretha Franklin?
El sistema es capaz de hacer cantar un chico como una chica y viceversa. En la mayoría de estos casos el proceso de transformación comprenderá un cambio de octava, además de otras transformaciones un poco más complicadas.

¿Qué pasa si desafino?
Que cantas mal. El sistema es capaz de ajustar tu voz a la afinación del cantante profesional. Eso no sólo implica cantar con la afinación del cantante imitado sino también añadir vibrato a las notas largas o poder marcar las transiciones entre notas tal y como lo haría un buen cantante.

¿Qué pasa si no canto la letra de la canción?
El sistema incluye un sistema de reconocimiento del habla que ha sido entrenado para poder aplicar la transformación adecuada a cada uno de los fonemas de la voz de entrada. Si el usuario no canta la letra de la canción, el sistema de reconocimiento se confunde, se reconoce erróneamente el fonema y por lo tanto la transformación a aplicarle, y la voz sintética que resulta del proceso parece un trabalenguas. Desgraciadamente para la gente que no habla el japonés, el sistema de reconocimiento ha sido entrenado para trabajar exclusivamente con fonética japonesa.

¿Cómo funciona el sistema?
Los dos pilares que sostienen el algoritmo son:

- la alineación en tiempo real de los fonemas del usuario y del cantante profesional
- la interpolación de modelos, también en tiempo real, de las dos voces cantadas, la del usuario y la del cantante profesional.

Mientras el usuario canta, el sistema localiza el fragmento que éste está interpretando en la grabación del cantante profesional. Una vez localizado, se interpolan los modelos de las dos voces cantadas de una manera flexible y musical. Se podría definir como un sistema de análisis-*morph*-síntesis de voz cantada a tiempo real.

JORDI VIDAL

Álex Loscos, uno de los artífices de Elvis, haciendo una demostración del programa. Quizá no se aprecia en la pantalla del ordenador, pero el pobre Elvis está llorando a lágrima viva...

¿Qué significa morph?
El *morph* (abreviatura de *metamorphosis*, o sea, transformación no abrupta de la forma de un ser vivo, objeto, etc.) es una técnica que persigue, a partir de dos o más elementos, la generación de nuevos elementos con propiedades híbridas. En el campo del

OTROS PROYECTOS DEL MTG

El control de emisiones musicales en la radio convencional siempre se ha basado en las listas que las emisoras elaboran cada mes para que la Sociedad General de Autores determine a quién le corresponde percibir derechos de pública difusión, y en qué magnitud. Desgraciadamente, ese sistema no está exento de picaresca (y no daremos nombres ni pondremos ejemplos, claro). Con la proliferación de emisoras vía Internet el panorama se complica aún más, ya que ejercer algún tipo de control sobre las emisiones resulta muy difícil... ¿O no? Al fin y al cabo, bastaría con disponer de muchos "oyentes virtuales", cada uno sintonizado a una emisora (convencional o no), que fueran anotando qué canción es la que se está emitiendo en cada momento. Eso es, más o menos, lo que persigue el proyecto RAA (Recognition and Análisis of Audio). RAA es un proyecto parcialmente financiado por la Comunidad Europea, en el que participan, además del Grupo de Tecnología Musical de la UPF, Joanneum Research (Austria), HS-ART Digital Service (Austria), TaurusMediaTechnik, Kirch Media (Alemania), Filmkunst-Musikverlag (Alemania), Nederlands Omroepproduktie (Holanda), y Radio Flaixbac (España). En la actualidad se ha implementado un sistema informático experimental que es capaz de reconocer, a los pocos segundos de empezar a emitirse, prácticamente todas las canciones que emite la emisora participante en el proyecto. El sistema es tan robusto que poco importa que el locutor hable encima, que la canción se emita a partir de un punto diferente del comienzo, que se la haya comprimido en dinámica, o que se le haya hecho una compresión de tiempo (¡para conseguir meter más canciones en una hora, claro!).

Usuario

ANÁLISIS

SRH

Reconocimiento y Alineamiento

MORPH Y SÍNTESIS

MORPH → SÍNTESIS

Voz híbrida

Parámetros extraídos de la voz del cantante profesional

Información musical de la canción

Análisis Alineamiento

Bloques del sistema de morph a tiempo real

procesamiento de vídeo, a diferencia del procesamiento de audio, el *morph* ha sido una técnica muy desarrollada y goza de gran popularidad en anuncios, videoclips o películas en las que vemos cómo los rostros de los personajes se transforman de unos a otros, o cómo una silla se transforma en un elefante. El sistema implementado en Elvis utiliza esta técnica para transformar la voz cantada de una persona en la de otra.

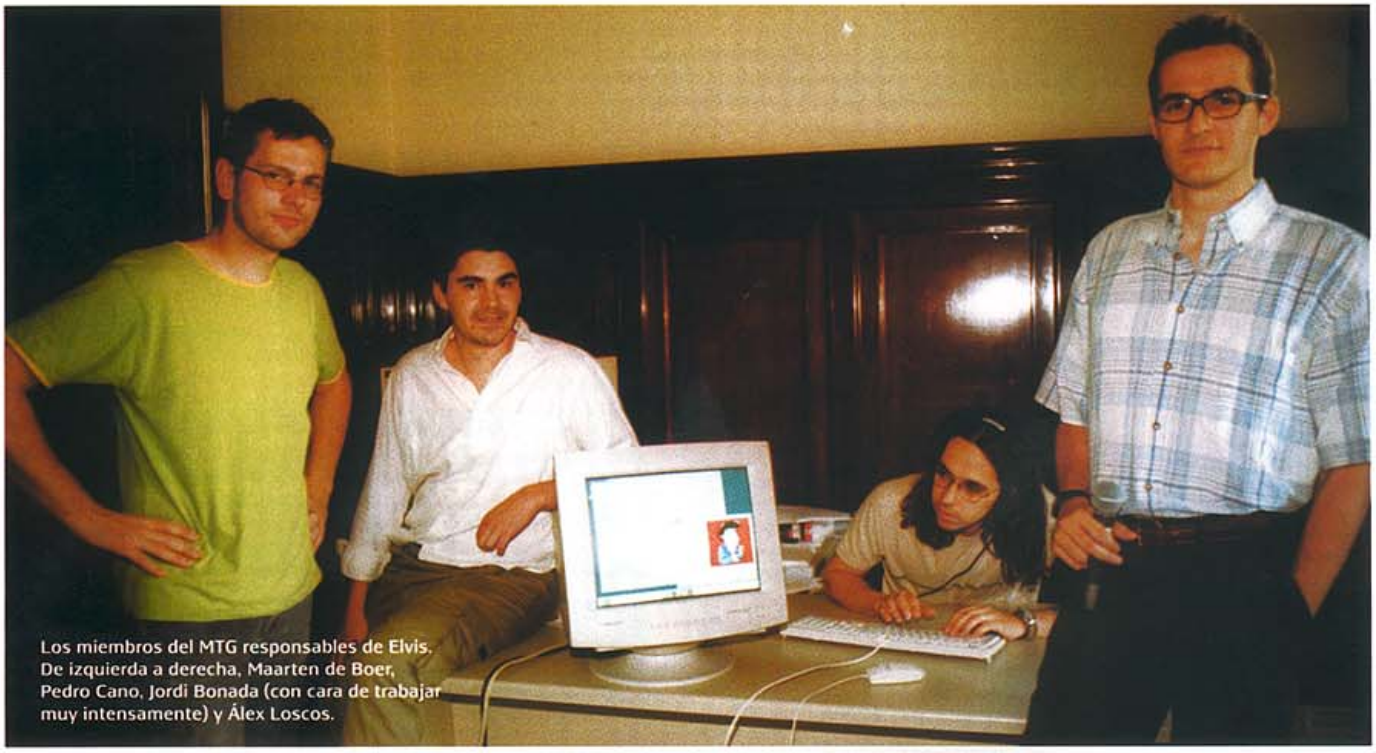
Durante estos dos últimos dos años, el equipo del MTG ha estado implementando las más de cien mil líneas de código que hacen posible el morph de dos voces cantadas a tiempo real sobre un ordenador personal en el entorno Linux, un sistema operativo que optimiza la captura y envío de audio. El programa, que incorpora una interfaz gráfica que emula el karaoke clásico, permite controlar la síntesis a tiempo real de manera que se puede pasar del timbre o la afinación de un cantante a la de otro mientras se está sintetizando la voz híbrida. En una de las ventanas de la interfaz, la afinación del usuario es puntuada a través de un Elvis animado.

El conjunto de procesos que constituyen el sistema global de *morph* se diferencia muy claramente en dos tipos: los procesos que tienen lugar antes de que el usuario del karaoke tome el micrófono (procesos fuera de tiempo real) y los que tienen lugar mientras el usuario canta por el micrófono (procesos a tiempo real).

Así pues, siguiendo un orden cronológico del *morph*, en primer lugar se debe obtener una grabación sin reverberación y sin acompañamiento musical de la voz del cantante que se quiere imitar. En nuestro caso, esto se traduce en grabar una serie de cantantes profesionales japoneses cantando los éxitos del karaoke en Japón. Todo este audio almacenado se utiliza para entrenar nuestro Sistema de Reconocimiento del Habla (SRH), adaptado a las peculiaridades específicas de la voz cantada. De este modo construimos los modelos de las 47 unidades fonéticas que constituyen el diccionario fonético japonés. El último paso previo al *morph* a tiempo real es analizar, parametrizar y segmentar al nivel de fonema las

¿QUÉ MÁS SE CUECE POR AHÍ?

La búsqueda de sonidos en una colección o "librería" se realiza en base al nombre que su creador le ha puesto. Pero explicar las características de un sonido por su nombre es una estrategia poco recomendable. Actualmente podemos analizar sonidos y extraer muchos "descriptores" significativos como por ejemplo qué nota es, de qué instrumento procede, qué tipo de escala se está interpretando, etc. A menudo utilizamos etiquetas que tienen que ver con esos descriptores de bajo nivel. Por ejemplo, cuando decimos que un sonido es brillante, eso significa que el centro de gravedad de su espectro (técnicamente llamado centroide) está desplazado hacia la zona de agudos. Un sistema que permita al usuario definir consistentemente algunas etiquetas, y usarlas a la hora de catalogar sonidos, puede facilitar mucho la recuperación de los mismos cuando se está orquestando una composición. El proyecto CUIDADO, financiado parcialmente por la Comunidad Europea y en el que participan también instituciones como el IRCAM de París, SONY CSL (París), Oracle España y Creamware, entre otros, busca consolidar el desarrollo de descriptores adecuados para caracterizar el contenido de un archivo sonoro, sin necesidad de que un humano se dedique a escribir anotaciones sobre él, ni que utilice términos excesivamente técnicos. Una vez disponemos de descriptores adecuados, podemos usarlos para generar transformaciones de los sonidos originales (por ejemplo, ¿qué tal sería un programa que, cuando estamos construyendo un *loop* de percusión nos buscara sonidos alternativos pero similares a los que ya tenemos? ¿O que nos hiciera el sonido más/menos brillante sin necesidad de usar los botones del ecualizador?). Eso es lo que persigue otro proyecto relacionado con éste, pero financiado por el Ministerio de Ciencia y Tecnología, que se denomina TABASCO (Transformación de Audio BASada en el COntenido), y en el que también participa el Instituto de Investigación en Inteligencia Artificial del CSIC.



Los miembros del MTG responsables de Elvis. De izquierda a derecha, Maarten de Boer, Pedro Cano, Jordi Bonada (con cara de trabajar muy intensamente) y Álex Loscos.

JORDI VIDAL

El Grupo de Tecnología Musical (MTG)

El Grupo de Tecnología Musical (MTG) desarrolla su actividad en el seno del Institut Universitari de l' Audiovisual (IUA) de la Universitat Pompeu Fabra. El Institut se creó en 1993 con la idea de complementar la carrera de Comunicación Audiovisual mediante la profundización en los aspectos más tecnológicos del ámbito audiovisual. Además de dedicar recursos a la investigación en tecnología audiovisual, el Institut también promueve proyectos de producción artística experimental, y colabora en Ingenierías, Masters y Posgrados relacionados con la tecnología y arte digital.

El Grupo de Tecnología Musical cuenta en la actualidad con más de 20 miembros, entre personal investigador, estudiantes de doctorado, becarios, y estudiantes que realizan su trabajo de final de carrera. El director del grupo es Xavier Serra, quien después de doctorarse en Computer Music en la prestigiosa universidad de Stanford (donde contribuyó a desarrollar una técnica de análisis y síntesis de sonido denominada "modelos espectrales"), trabajó durante un par de años en Yamaha antes de regresar en 1991 a Barcelona para hacerse cargo de la dirección de la Fundación Phonos, una de las instituciones pioneras de la música electrónica en España. La Fundación Phonos se integró en 1995 en el Institut Universitari de l'Audiovisual de la Universitat Pompeu Fabra. Y desde allí, con la ayuda de una infraestructura convenientemente preparada para la investigación, Xavier Serra fue organizando un equipo humano cuyo primer gran reto internacional vendría de la mano de sus antiguos jefes de Yamaha, quienes a finales de 1997 le encargaron el proyecto Elvis. Después de cuatro años, la colaboración prosigue a través de proyectos derivados de Elvis, así como de otros nuevos retos.

Para más información puedes dirigirte a la página web del MTG, www.iua.upf.es/mtg

grabaciones del cantante profesional de la canción que el usuario se dispone a cantar.

Los procesos a tiempo real empiezan con la digitalización de la señal de voz cantada que nos llega al ordenador a través de un micrófono conectado a una tarjeta de sonido. La señal digitalizada es analizada y parametrizada utilizando una técnica que la descompone en sus elementos resonantes y sus elementos ruidosos basándose en el espectro del sonido.

De los parámetros resultantes del análisis se pueden diferenciar dos grupos. Un primer grupo alimenta el SRH para saber en qué parte de la canción estamos; más concretamente, cuál es el fonema que estamos cantando en aquel momento. Una vez localizado el fonema, el otro grupo será utilizado para construir la voz sintética híbrida, interpolando los valores de los parámetros de la voz del usuario con los valores de los parámetros previamente extraídos y almacenados de la voz del cantante profesional. El proceso acaba cuando la voz sintética es enviada a la tarjeta de sonido y ésta la envía a su vez al sistema de escucha. La duración de este ciclo es de tan sólo 30 milisegundos.

En el CD de este número hallarás algunos ejemplos ilustrativos del funcionamiento del sistema (mira el cuadro "En el CD" para más detalles).

Después de 3 años de trabajar en el proyecto se obtuvo un prototipo susceptible de ser implementado en *hardware*. Desgraciadamente, por aquellas fechas la economía japonesa empezó a hacer aguas, los costes se dispararon, y el mercado potencial se deshinchó. Total, que los números para comercializarlo no salieron y hubo que poner a Elvis a dormir el sueño de los justos. No obstante, la relación del Grupo de Tecnología Musical con Yamaha no se acabó aquí, puesto que en la actualidad se están desarrollando otros proyectos conjuntos de los que esperamos poder hablaros en el futuro.

ÁLEX LOSCOS Y FECTOPER

Entre las pistas 3 y 14 del CD encontrarás ejemplos de las capacidades del programa. Las pistas 3-5 son la canción Nakinakala original, cantada por el usuario, por el cantante profesional y por ambos al mismo tiempo (con las voces separadas en los canales izquierdo y derecho). En las pistas 5-8, el usuario controla el tiempo. Y las pistas restantes (9-14) siguen el mismo esquema, esta vez con la canción Yokohama. En ellas, el usuario controla la afinación.