UNIVERSIDAD DE ALCALÁ

Escuela Politécnica Superior

INGENIERÍA ..

Trabajo Fin de Carrera

"Título del TFC"

Nombre del Alumno

Año

UNIVERSIDAD DE ALCALÁ

Escuela Politécnica Superior

INGENIERÍA ...

Trabajo Fin de Carrera

"Título del TFC"

Alumno: Nombre del Alumno

Director: Nombre del Director

Tribunal:

Presidente: D. Nombre Presidente.

Vocal 1º: D. Nombre Vocal1.

Vocal 2º: D. Nombre Vocal2.

Calificación: .....

Fecha: .....

A mis dos Antonio Barra

"Empieza haciendo lo necesario, luego haz lo posible y de pronto empezarás a hacer lo imposible."
Francisco de Asís

Agradecimientos

"Más vale un minuto de ilusión que mil horas de razonamiento". Con esta frase acababan los agradecimientos de mi Proyecto Fin de Carrera, y con esta ideología empezó y acaba esta tesis. Durante todos estos años son demasiadas las personas a las que debo poder haber realizado esta tesis; siendo por ello, posiblemente más importante esta página que cualquiera del resto de capítulos de este libro.

Debo de estar agradecido a mi padre, Antonio Barra, por tratar de transmitirme su "ansia de saber": la valoración justa de lo aprendido y el eterno deseo insatisfecho por entender lo que todavía está por aprender. Creo que todo esto me ha permitido realizar las investigaciones recogidas en esta tesis, aunque también ha resultado algo más difícil decir "hasta aquí!".

Con estas herramientas heredadas en vida llegué al Grupo de Tecnología del Habla (GTH) allá por el año 2003. Con el paso de los años he comprobado que las universidades, centros y grupos de investigación son lo que son por las personas que trabajan y estudian en ellos. En este sentido, debo dar las gracias por haber gozado todos estos años de la elevada calidad humana, profesional y científica de toda la gente con la que he trabajado tanto en el Dpto. de Ingeniería Electrónica como especialmente en el GTH. Desde los ilusionados alumnos (Carmen, Lorena, Carlos, etc), hasta los apasionados y experimentados investigadores (Fernandito, Juanma, pero mira que os quiero; Manolo, LuisFer, Rick, Ferre, Fernando González (gracias por tus destilados consejos)), que a pesar de los años disfrutan de su trabajo como el primer día. Gracias a todos, sin olvidarme de compañeros como Syaheerah, Julián, Bea, Vero, etc. y de todos los alumnos que me han hecho crecer como profe un poquito más.

Esta tesis no hubiera sido posible sin las ayudas y proyectos proporcionados por el Ministerio de Educación y Ciencia y la Universidad Politécnica de Madrid al GTH. Dar las gracias a estos organismos por financiar las dos estancias de investigación realizadas en el extranjero durante esta tesis. Gracias a Bryan Pellom por instruirme en mis primeros pasos y enseñarme el reconocedor SONIC. Agradecer de forma muy especial a todos los miembros del CSTR por la increíble estancia en 2008 así como la continua y fructífera colaboración hasta hoy. Gracias especiales a Junichi Yamagishi por no escatimar su tiempo conmigo y darme la formación necesaria en síntesis de voz HMM, imprescindible para el desarrollo de la tesis. Dar las gracias a Rob Clark por su ayuda para adaptar Festival al Castellano. Dar las gracias a Simon King por escucharme, por su dirección y sus consejos tanto en lo profesional como en los personal y hacer que la estancia en CSTR fuera perfecta. Gracias a todos los amigos que hice en CSTR: Javi, Simon, María, Junichi, Mathew, Yolanda, Joao, Dong, Korin, Rob, Bela, Volker, Mike, Gregor, Blaisse, Mónica, Paula, etc. "You have a friend in Madrid".

Debo dar las gracias a dos compañeros especiales de profesión como son Rubén San Segundo y Sira Palazuelos. Rubén, gracias por trabajar en el GTH, con compañeros así, es mucho más fácil convertir el tiempo trabajando en experiencia fructífera. Sira, muchas gracias por compartir tu experiencia en la Universidad conmigo y por las múltiples conversaciones que hemos tenido y tanto me han ayudado. Gracias a ambos.

Sin duda, con diferencia el mayor privilegio profesional ha sido poder contar como directores de tesis con Juan Manuel Montero (Juancho) y Javier Macías Guarasa (Maci): dos "titanes" de la docencia cuya combinación de aptitudes investigadoras y de gestión de equipos humanos se me antoja muy difícil de encontrar.

Juancho, el camino ha sido largo y no ha habido semana en todos estos años que no me hayas dedicado unas horas. Gracias por la dirección continua, por guiarme con éxito y maestría incluso salvando los caracteres diferentes de ambos (quizás por ello eres profesor de inteligencia emocional). Sobre todo gracias por el talento a la hora de potenciar el protagonismo y el trabajo de tu alumno, dejando en un segundo plano (a ojos del resto) tu labor de dirección. Gracias por dar muchísimo más de lo recibido.

He tenido la suerte de poder contar estos años con una exquisita referencia. Maci, como me dijiste un día: "ya esta todo dicho". Tan solo pedirte que cuando leas estos agradecimientos, leas el cuadro de tu salón. Sólo citar una frase: "Oh capitán, mi capitán!".

Soy de aquellas personas que les resulta imposible no establecer una sinergia entre la vida profesional y personal. Son los pequeños logros en el trabajo los que me han permitido disfrutar con tranquilidad e ilusión de las personas que quiero. Y no hubiera sido posible ofrecer lo máximo de mí mismo en mi labor investigadora sin el cariño de mi gente. Gracias a TODOS mis compañeros de trabajo por arroparme en mis momentos personales más difíciles y por disfrutar de vosotros en otros bien alegres.

Debo de dar las gracias a mi familia, a mis primos y amigos (Morante, Chema, Juan Pablo, Antonio, Jorge, Angela, etc), a la "familia" de amigos franciscanos (Ángel y Estrella, Gonzalo, Luis, etc), a los Balbacileros y a la familia de mi mujer (a Loli y Guzmán, Toñi, Óscar, Esther, Jimena, Gumán y Nuria). Sin todos vosotros esta etapa final hubiera sido infinitamente más dura.

Quiero y debo de dar las gracias a mi madre Mª de los Ángeles y a mi hermana Noemí. Por quererme como soy, por su apoyo incondicional y su cheque en blanco de cariño, tiempo y consuelo. Os quiero.

Pero sin duda, el logro de haber acabado esta tesis he de compartirlo con mi Compañera. Sin su paciencia, comprensión, ternura y apoyo no hubiera sido posible llegar hasta aquí. Moni, consigues que parezca más grande de lo pequeñito que realmente soy. Gracias infinitas más uno por donarme a fondo perdido todo el tiempo que os he dejado de dedicar a tí y a Antonio todos estos años. Gracias por perdonarme mis ausencias. ¡Os aseguro que vuelvo con fuerzas! sois mi siguiente tesis. Os Quiero.

Gracias a todas las personas que habéis formado parte de esta aventura durante todos estos años y por descuido omito en estas páginas, porque entre todos habéis hecho posible que esta tesis llegue a buen puerto.

Resumen

El trabajo realizado en esta tesis ha abordado diferentes estudios orientados a la mejora de un sistema de generación de respuesta mediante la incorporación de un sintetizador de habla con emociones en español. La tesis doctoral se ha abordado en tres fases fundamentales, cada una de las cuales está relacionada con una de las contribuciones científicas planteadas originalmente.

En primer lugar, y con el objetivo de obtener información sobre la relevancia de las distintas componentes de la señal de habla en los procesos de identificación de emociones, se ha realizado un estudio que demuestra la complementariedad entre los aspectos segmentales y suprasegmentales, caracterizando su importancia relativa para cada una de las emociones bajo estudio. Sobre una base de datos existente, se ha realizado un análisis de la naturaleza de las emociones en la voz mediante estrategias de identificación automática y la evaluación perceptual de estímulos generados mediante métodos de síntesis por copia. Adicionalmente, se ha realizado un estudio sobre la normalización de características acústicas con el fin de implementar sistemas de identificación de emociones multi-locutor y multi-idioma. Como complemento al análisis, se ha evaluado el comportamiento de un sistema automático de identificación basado en redes bayesianas dinámicas a la hora de identificar emociones reales (no actuadas), dicho sistema ha sido evaluado dentro de la primera competición internacional de reconocimiento automático de emociones.

En segundo lugar, los conocimientos adquiridos de este análisis inicial han sido la base para la adquisición de un corpus pionero en el área de síntesis de emociones, dada la cobertura de su contenido emocional multimedia y multi-locutor. Este corpus ha sido imprescindible para adaptar y evaluar exhaustivamente la aplicación a la síntesis de habla emocional, de dos de las técnicas de alta calidad empleadas actualmente por la comunidad científica: síntesis por selección de unidades, dominante en la última década; y síntesis paramétrica basada en modelos ocultos de Markov, técnica emergente y base de las investigaciones futuras en síntesis de voz durante la próxima década. Tras un exhaustivo y novedoso análisis de los resultados obtenidos en una evaluación perceptual, se ha comprobado que ambas técnicas producen voz con emociones de la misma calidad. Sin embargo, a pesar de que las emociones se identifican mejor de forma global cuando sintetizamos voz mediante la técnica de selección de unidades, y que la intensidad emocional resultante es mayor al minimizar el modelado y el procesado de la señal de voz, es la síntesis de voz basada en modelos ocultos de Markov la que modela mejor la información prosódica, de máxima relevancia en cuanto a la expresión de emociones se refiere. El sistema basado en modelos ocultos de Markov adaptado al castellano ha sido galardonado con el premio al mejor sistema en la competición nacional de conversión texto a voz dentro de las Jornadas de Tecnología del Habla en 2008.

En tercer lugar, sobre las voces generadas utilizando una de las técnicas anteriores (concretamente las generadas exitosamente basándose en modelos ocultos de Markov, dada la flexibilidad en la manipulación de parámetros del modelo que ofrece esta técnica y los excelentes resultados obtenidos en la competición), se ha diseñado, implementado y evaluado una nueva estrategia de transformación de emociones independiente del locutor. Dicha estrategia está basada en la extrapolación de la emoción sobre aquellas características halladas como relevantes en el análisis inicial. De los resultados de la evaluación, se ha

comprobado que los patrones acústicos emocionales son extrapolados parcialmente a una locutora objetivo sin por ello perder similitud con la voz de dicha locutora, y que la intesidad de la emoción extrapolada puede ser modificada con éxito variando un coeficiente de extrapolación. Sin embargo, la intensidad con la que se extrapola la emoción tiene un impacto negativo en la calidad de la voz sintetizada, especialmente cuando dicha extrapolación se centra en la transformación de parámetros espectrales. Finalmente, se ha propuesto una nueva medida sobre la bondad de la extrapolación/transformación de emociones independiente del locutor, basándose en los resultados perceptuales en cuanto a calidad de voz, identificación de la emoción e identificación del locutor objetivo se refiere.

Abstract

The work carried out in this Thesis have been focused on the improvement of the response generation module by the incorporation of emotional speech synthesis in Spanish. This Thesis is divided in three stages, each one related with one of the defined scientific contributions.

Initially, in order to convey emotions through the speech signal, the relevance of each speech component has been studied. The complementary behaviour of segmental and supra-segmental rubrics has been demonstrated, by analysing its relevance for each of the studied emotions. The nature of the emotions, using an existing corpus, has been studied using automatic identification strategies and a perceptual evaluation of emotional stimuli synthesised by copy-synthesis. In addition to this, a speaker-independent modelling of emotional acoustic patterns has been studied by means of the implementation and evaluation of a multi-speaker and multi-language automatic emotion identification system. Additionally, the performance of a system for the automatic identification of real emotions (based on dynamic Bayesian networks) has been evaluated on the first international emotion recognition challenge.

Secondly, the conclusions obtained from the previous analysis have been the base for the acquisition of a novel emotional corpus in Spanish, due to its multimedia and multi-speaker content. This corpus has been essential for the adaptation and the exhaustive evaluation of two of the state-of-the-art high quality speech synthesis techniques to the synthesis of emotional speech: unit selection synthesis, the dominant technique during last decade; and HMM-based synthesis, an emerging technique and base of the future research in this area for the next decade. After, an exhaustive and novel analysis of the obtained results from a perceptual evaluation, it has been shown that both techniques synthesise emotional speech with the same quality. Although the emotions are best identified when they are synthesised using the unit selection technique and the resulting emotional strength with this technique is the highest , the HMM-based synthesis is the technique that best models the prosodic information, extremely important in expressive speech. The HMM-based system adapted to Spanish has been awarded as the best system in the text-to-speech challenge at the Jornadas de Tecnología del Habla in 2008.

Finally, a new strategy for the emotional speaker-independent transformation of synthetic speech has been designed, implemented and evaluated using the emotional voices generated with one of the previous techniques (specifically, the voices successfully generated using the HMM-based techniques, due to the flexibility and the controllability of the speech model parameters and the excellent results obtained in the challenge). This new strategy consists on the extrapolation of the emotions through the relevant speech components found in the initial analysis. From the results of the perceptual evaluation, it has been confirmed that the emotional acoustic patters have been partially extrapolated to the neutral voice of a target speaker, without extrapolating the identity of the source speaker. Additionally, the strength of the extrapolation can be successfully modified by using an extrapolation factor. However, the strength of the extrapolation has a negative impact in the quality of the synthesised speech, especially when the emotion extrapolation is focused on the transformation of the spectral parameters. Finally, a new metric for the evaluation of the goodness of the proposed new strategy has been defined, based on the speech

quality, emotion identification and speaker identification results.

Índice general

Índice de figuras

Índice de tablas

Acronyms

HMI *Human-Machine Interfaces*

HMI *Human-Machine Interfaces*

ETTS *Emotional Text To Speech*

TTS *Text To Speech*

PSOLA *Pitch Synchronous OverLap Add*

TD-PSOLA *Time Domain Pitch Synchronous OverLap Add*

AI *Artificial Intelligenge*

SPSS *Statistical Parametric Speech Synthesis*

VC *Voice Conversion*

US *Unit Selection*

HMM *Hidden Markov Model*

LSP *Line Spectral Pairs*

LPC *Linear Prediction Coefficiens*

LSF *Line Spectral Frequencies*

F0 *Fundamental Frequency*

MCEP *Mel Cepstral Coefficients*

MGCEP *Mel Generalized Cepstral Coefficients*

MFCC *Mel Frequency Cepstrum Coefficients*

ASR *Automatic Speech Recognition*

MDL *Minimum Description Length Criterion*

MSD *Multi Space Probability Distributions*

HSMM *Hidden Semi-Markov Models*

ML *Maximum Likelihood*

MLSA *Mel Log Spectrum Approximation*

MAP *Maximum A Posteriori*

MLLR *Maximum Likelihood Linear Regression*

CSMAPLR *Constrain Structural MAP Linear Regression*

AV *Average Voice*

ANN *Artificial Neural Network*

NIST *National Institute of Technology*

SES *Spanish Expressive Speech*

EMODB *Berlin Database of Emotional Speech*

FAU-AIBO *FAU AIBO Emotion Corpus*

SEV *Spanish Expressive Voices*

AER *Automatic Emotion Recognition*

UBEC *Universal Background Emotion Codebook*

DBN *Dynamic Bayesian Network*

SQ *Speech Quality*

EIR *Emotion Identification Rate*

SIR *Speaker Identification Rate*

ES *Emotional Strength*

STRAIGHT *Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum*

Bibliografía

[1] Información sobre gnu/linux en wikipedia. http://es.wikipedia.org/wiki/GNU/Linux.

[2] Página de la aplicación emacs. http://savannah.gnu.org/projects/emacs/.

[3] Página de la aplicación kdevelop. http://www.kdevelop.org.

[4] Leslie Lamport. *LaTeX: A Document Preparation System, 2nd edition.* Addison Wesley Professional, 1994.

[5] Página de la aplicación octave. http://www.octave.org.

[6] Página de la aplicación cvs. http://savannah.nongnu.org/projects/cvs/.

[7] Página de la aplicación gcc. http://savannah.gnu.org/projects/gcc/.

[8] Página de la aplicación make. http://savannah.gnu.org/projects/make/.

Apéndice A

Manual de usuario

A.1 Introducción

Introducción.

A.2 Manual

Pues eso.

Apéndice B

Herramientas y recursos

Las herramientas necesarias para la elaboración del proyecto han sido:

- PC compatible
- Sistema operativo GNU/Linux [1]
- Entorno de desarrollo Emacs [2]
- Entorno de desarrollo KDevelop [3]
- Procesador de textos LaTeX[4]
- Lenguaje de procesamiento matemático Octave [5]
- Control de versiones CVS [6]
- Compilador C/C++ gcc [7]
- Gestor de compilaciones make [8]